From the holdings of the Department of Special Collections , Princeton University Library, Princeton, NJ

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyrightlaw. We do not charge any permission or usage fees for the publication of images of material inour collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff.For more information regarding copyright, credit, and citation please review the information onthe Copyright, Credit and Citation Guidelines page of our website.
https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. https://library.princeton.edu/special-collections/ask-us

---

Doctoral Dissertations, 2011-2026 Public and International Affairs

2015

Essays in Health Economics and Political Economy

Incerti, Devin

Princeton, NJ : Princeton University

---

http://arks.princeton.edu/ark:/88435/dsp010g354h60c

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. https://www.copyright.gov/title17/title17.pdf

ESSAYS IN HEALTH ECONOMICS AND POLITICAL ECONOMY

DEVIN INCERTI

A DISSERTATION

PRESENTED TO THE FACULTY

OF PRINCETON UNIVERSITY

IN CANDIDACY FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE

BY THE WOODROW WILSON SCHOOL OF PUBLIC AND

INTERNATIONAL AFFAIRS

ADVISER: JOHN LONDREGAN

NOVEMBER 2015

© Copyright by Devin Incerti, 2015.

All rights reserved.

Abstract

This dissertation examines issues in public policy—with an emphasis on health policy—from economic and political perspectives. It focuses on 1) how campaign donations influence party control in Congress, 2) racial/ethnic and educational inequalities in pharmaceutical spending, and 3) the effectiveness of cost containment policies in healthcare.

The first chapter looks at whether Democratic and Republican parties optimally allocate resources in House elections. To do so, it estimates a probabilistic voting model using a Bayesian approach and compares actual spending patterns to the amount that should have been spent under the model. The correlation between actual spending and the amount that should have been spent is over 0.5 in each non-redistricting election from 2000 to 2010 and has generally increased over time. Surprisingly, these correlations are consistent across different types of campaign donors including political parties, political action committees, and individuals.

Chapter two identifies differences in prescription drug utilization as a potential mechanism for the well known racial/ethnic and educational gradients in health. A two-part model predicts that, on average, blacks and Hispanics spend $350 and $560 less than whites respectively and that an additional 4 years of education increases prescription drug expenditures by $155. These documented disparities occur for two primary reasons: first, there are differences in the probability of being diagnosed with a disease; and second, there are gradients in expenditures conditional on diagnosis.

The final chapter uses a dynamic Bayesian model to study two quantities of interest—out-of-pocket expenditure inequality and the uncertainty of longterm spending—that are essential for evaluating cost containment reforms. It shows the distribution of spending is less concentrated over longer periods than in a single period, but that there is still substantial inequality in long-term spending. Out-of-pocket expenditures are determined more by permanent and transitory shocks than observed data so individuals face considerable uncertainty in future health costs. Implications for policy are twofold. First, reforms that increase

cost sharing must consider the risks of significant inequality and additional financial risk. Second, disease management programs have the potential to reduce costs significantly, but identifying high spenders will be difficult.

Acknowledgements

I am incredibly thankful to my parents for instilling in me the value of a good education and a desire to learn. Trevor, I enjoy our little chats about work, economics, public policy and the like; I also greatly appreciate your amazing ability to proofread my papers. I could not have completed this thesis without the support of my family. Although I was initially hesitant, I am ultimately glad that you convinced me to attend Princeton University even though my heart was back home in California. This brings me to my fiancée, Jennifer, who has been amazingly supportive during my time at Princeton even when we were over 3,000 miles apart. I greatly admire your work ethic and success as a young radiologist; you are truly an inspiration to me.

I am grateful to the two readers of my dissertation, John Londregan and Tom Vogl, for their patience and generous time commitments. I am incredibly thankful that they were both willing to speak to me on Skype (or on the phone) when I was in California or they were out of the country (Chile), or even to meet with me at another university (U.C. Berkeley). My primary adviser, John Londregan, has been incredibly supportive during the process. He has forced me to think differently about problems than I otherwise would have which has helped me produce better work. Tom Vogl has given me outstanding feedback on ways to improve my research. His comments have greatly improved this dissertation.

I would like to thank Janet Currie and Charles Cameron in advance for serving as examiners for my public defense. I also appreciate Janet's great comments on my work during a course on Health Economics.

This dissertation would not have been possible without Yuki Shirato and Gabriel López Moctezuma, who helped me get through the first few years of the program. I'll never forgot the late nights we spent together working on problem sets or studying for general exams.

Finally, I would like to thank Gary Jacobson for sharing his voting data for the United States House of Representatives which is used heavily in Chapter 1.

Contents

Abstract i

Acknowledgements iii

1 The Optimal Allocation of Campaign Funds in House Elections **1**

1.1 Introduction 1

1.2 Related Literature 3

1.3 Model 7

1.4 Estimation 13

1.5 Data 20

1.6 Forecasting the Two-Party Vote and Calculating $\mathbf{Q_i}$ 30

1.7 Relationship Between $\mathbf{Q_i}$ and District Spending 42

1.8 Discussion and Summary 51

Appendices **54**

1.A Stan Code for Hierarchical Model 54

1.B Gibbs Sampler for DLM 55

1.C Additional Figures 59

1.D Data Appendix 67

2 Racial, Ethnic and Educational Disparities in Pharmaceutical Expenditures **69**

2.1 Introduction 69

2.2 Theoretical Framework and Empirical Estimation 72
2.3 Data 75
2.4 Results 81
2.5 Discussion 94
2.6 Conclusion 97

Appendices **98**

2.A Predicting Positive Rx Expenditures 98
2.B Constructing the RX Disease Categories 101
2.C Additional Tables and Figures 102

3 An Assessment of Long-term Healthcare Expenditure Risk using a Bayesian Two-Part Model **107**

3.1 Introduction 107
3.2 Data 111
3.3 The Model 115
3.4 Bayesian Estimation 120
3.5 Long-term Expenditure Risk 130
3.6 Conclusion 142

Appendices **145**

3.A Out-of-Pocket Spending Data 145
3.B Sample Selection 146
3.C MCMC Algorithm 147
3.D Simulation Procedure 150
3.E Additional Tables and Figures 151
3.F Additional Model Checking 154

List of Figures

1.1 Average Daily Expenditures by PACs on Candidates by Month, 2000 - 2010 26

1.2 Summary Plots for District Polls 29

1.3 Forecasts of the Number of the Percentage of Democratic Seats 33

1.4 Forecasts of the 2010 National Vote by Date 36

1.5 Forecasts of the Democratic Fraction of Seats in the 2010 Election 38

1.6 Evaluations of 2010 Election Forecasts 39

1.7 Q_i versus the mean forecast of the Democratic vote in the 2008 election . 42

1.8 Correlation Between $\mathbf{Q_i}$ and Spending, by Contributor Type 44

1.9 Correlation Between Measures of Electoral Competitiveness and Campaign Spending 46

1.10 Correlation Between $\mathbf{Q_i}$ and PAC Spending by Date and Model, 2010 House Election 47

1.C.1 Posterior Densities for Parameters in Forecasting Model 60

1.C.2 Out-of-Sample Predicted Versus Actual Plot, House Elections 1990 - 2010 61

1.C.3 Estimates of National Opinion During the Election Campaign 62

1.C.4 Estimates of House Effects 63

1.C.5 RMSFE By Polling Availability and Forecast Date 64

1.C.6 MPE By Polling Availability and Forecast Date 64

1.C.7 Brier Score By Polling Availability and Forecast Date 65

1.C.8 Correlation Between Q and PAC Spending, by Party 66

2.1 Prescription Drug Use By Year and Race/Ethnicity 79
2.2 Prescription Drug Use By Year and Education 80
2.3 Predicted Disparities in Expenditures 85
2.4 Regression Estimates of Disparities in Specific Diagnoses 90
2.5 Regression Estimates of Disparities in Utilization for Common Drug Classes 93
2.6 Predicted Mean Expenditures by Education, Race/Ethnicity, and Drug Class 94
2.7 Results of Repeated 5-Fold Cross Validation 100
2.C.1 Histogram of Prescription Drug Expenditures Conditional on having an Rx Medical Condition 103
2.C.2 Histogram of Prescription Drug Expenditures 104
2.C.3 Density of Studentized Residual for Second Component of Two-Part Model for Continuous Component of Two-Part Model 105
2.C.4 Histogram of the Number of Rx Conditions 106

3.1 Nonzero Spending in the HRS 114
3.2 Mean Out-of-Pocket Expenditures by Wave 4 Quartile 115
3.3 Out-of-Pocket Expenditure Density Estimates 127
3.4 Survival Estimates 128
3.5 Distribution of Wave t + 4 Expenditures Conditional on Wave t Expenditures 129
3.6 Empirical Lorenz Curves for Out-of-Pocket Expenditures per Wave Alive 131
3.7 Simulated Lorenz Curves for Out-of-Pocket Expenditures per Period Alive, Age $\leq$ 60 in Period 1 133
3.8 Simulated Lorenz Curves for 14-Year Out-of-Pocket Expenditures per Period Alive, by Type of Risk 134

3.9 Empirical CDF of Out-of-Pocket Expenditure as a Percentage of Household Income by Income Quartile 136
3.10 Targeting High Spenders 139
3.11 Empirical CDF of Estimated Risk Premiums 142
3.E.1 Plot of $\ln y_{it}$ Against $d_{it-1} \times \ln y_{it-1}$ 152
3.E.2 Plot of $\ln y_{it}$ Against $d_{i0} \times \ln y_{i0}$ 153
3.F.1 Actual versus Predicted Plots 155
3.F.2 Posterior Predictive Checks for Summary Measures of Fit 156

List of Tables

1.1 Summary Statistics for Forecasting Variables 24

1.2 Average Post-August Spending by Political Parties per District in Non-Redistricting Years, 2000 - 2010 27

1.3 Posterior Inferences for Parameters in Forecasting Model, House Elections 1980 - 2008 31

1.4 Posterior Quantiles for Square Root of DLM Variance Paramters 35

1.5 OLS Regressions on Candidate Spending Shares (%) 49

1.6 Spending by the Financial Industry, Incumbent Races 2000 - 2010 50

2.1 Summary Statistics 76

2.2 Prescription Drug Utilization by Disease Status 78

2.3 Rx Conditions by Race/Ethnicity and Education 81

2.4 Regression Estimates of Disparities in Utilization 82

2.5 Regression Estimates of Disparities in Diagnoses 86

2.6 Prevalence of Most Common Rx Disease Groups 88

2.7 Regression Estimates of Disparities in Utilization Conditional on Rx Condition Categories 91

2.8 In-Sample Predictions of Nonzero Rx Expenditures 99

2.C.1 Regression Coefficients from Two-Part and Selection Models 102

3.1 The Distribution of Out-of-Pocket Health Expenditures in the HRS, Waves 4 - 11 113
3.2 Explanatory Variables Included in Model 119
3.3 Parameter Estimates for Expenditure Model 123
3.4 Parameter Estimates for Mortality Model 125
3.5 R-squared for 14-Year Out-of-Pocket Spending, by Simulation Type . . . 137
3.B.1 Sample selection in the HRS 146

Chapter 1

The Optimal Allocation of Campaign Funds in House Elections

1.1 Introduction

An important component of campaign strategies in legislative elections is the optimal allocation of resources across legislative districts. In this paper, I tackle this problem with a model that generates precise estimates of the marginal value of additional campaign resources in each district. I apply this model to U.S House elections and compare the amount of campaign funds that should have been spent on each candidate with actual spending.

The allocation of campaign resources is viewed as a competition for seats between Democrats and Republicans. This competition is analyzed using the probabilistic voting model of Strömberg (2008), which can be explicitly solved and directly estimated. Like Snyder (1989), I consider two different assumptions about goals: first, parties maximize the expected number of seats won, and second, parties maximize the probability of winning a majority of seats. In equilibrium, the parties spend the most on the closest elections under the first assumption, whereas under the second assumption, an additional factor—the probability that a seat is pivotal—impacts the marginal value of a dollar in a district.

The marginal value of additional spending under Stromberg's model depends on predicted

vote shares and accounts for uncertainty at the district and national levels. It is therefore easily integrated with forecasting methods that account for national swings. In this paper, I use two types of forecasting models. The first model is a Bayesian hierarchical model that uses information about districts and candidates using yearly data to provide a forecast as of September 1st of each election year. The second model is a state-space model that uses the hierarchical model as a prior and incorporates all available district and national polls. Unlike the first model, the latter is capable of providing real time forecasts at any date during a campaign. The hierarchical model is used to forecast each non-redistricting year election from 2000 to 2010 and the state-space model is used to analyze the 2010 election at various stages of the campaign.

The empirical analysis uses these forecasts to estimate Stromberg's model, which is then used to identify the districts that should have received the most contributions. The correlation between the amount that should have been spent according to the model and actual spending is over 0.5 in each non-redistricting election year from 2000 to 2010 and has generally increased over time. Surprisingly, these correlations are consistent across different types of campaign donors: the spending patterns of party committees, such as the Democratic Congressional Campaign Committee (DCCC) and the National Republican Congressional Committee (NRCC), whose strategies should be the most likely to be coordinated and concerned with electoral success, are *no* more consistent with the model's predictions than the spending patterns of political action committees (PACs) or individuals. There is also no evidence that spending for Republican candidates differs from spending for Democratic candidates. Finally, as one might expect, these correlations become even stronger when incorporating polling data using the state-space model: the correlation reaches a peak of over 0.8 when comparing spending in the final month of the 2010 campaign with equilibrium spending based on a forecast made using all information up until that date.

When elections are reasonably close, the parties' goals have a relatively minor effect on optimal spending strategies; when elections are lopsided, the opposite is true. Between

2000 and 2010, only one election—the 2008 election—was lopsided enough to yield spending strategies that were significantly sensitive to party goals. During that election, actual spending is highly correlated ($r \approx 0.664$) with a seat maximization based spending strategy but not ($r \approx 0.353$) one based on maximizing the probability of winning a majority of seats.[1] This provides evidence that parties are more concerned with whether an election is close than whether it is pivotal.

Overall, campaign spending seems to be consistent with a world in which the two major parties in the United States try to maximize expected seat share. However, this does not mean that other factors do not help explain observed differences in spending across districts as well. For example, incumbents and candidates running in open seats both raise more funds than challengers running against incumbents. Similarly, party leaders raise more funds than their counterparts and members of the House Committee on Financial Services raise considerably more funds from financial firms than other incumbents not on the committee. That being said, the explanatory power of these other factors pales in comparison to the explanatory power of Stromberg's model and it is only in cases in which actors have the largest incentives to consider other factors—like financial service firms buying access to candidates—that they are on close to equal footing.[2]

1.2 Related Literature

This paper builds on a number of other studies focusing on the strategic allocation of resources in campaigns. These studies date back to the work of Brams and Davis (1973; 1974), who look at how presidential campaigns should allocate campaign resources to maximize their expected electoral vote. Using a model in which each candidate has an equal probability of winning the popular vote in each state, they argue that presidential campaigns should allocate resources roughly in proportion to the 3/2 power of the number of electoral

[1] r refers to Pearson's correlation coefficient.

[2] Members of the House Committee on Financial Services receive a higher proportion of funding from the financial services industry than any other committee receives from a single industry.

votes in that state, which means that larger states should receive a disproportionate share of campaign resources. This result was then challenged by Colantoni, Levesque and Ordeshook (1975), who conclude that a modified proportional rule that accounts for the closeness of an election fits the data better.

One drawback of these earlier studies is that they assume that parties maximize the expected electoral vote rather than the probability of winning the election. Brams and Davis posited that the implications of this distinction would be relatively minor. Aranson, Hinich and Ordeshook (1974) support this notion by showing that the two goals are equivalent if the game is symmetric—which implies that the expected number of seats won by each party must be that same. However, as discussed by Snyder (1989), this symmetry assumption is strong and not realistic in real world settings. He shows that when this is not the case, the two goals yield different equilibria.[3]

More recently, Strömberg (2008) has developed a more general model that can reasonably be applied to actual campaigns. He uses the more reasonable assumption that presidential candidates maximize the probability of winning a majority of electoral college votes (i.e. winning the election) and then calculates the number of times that presidential candidates should have visited each state. He finds a very strong correlation (≈ 0.9) between these visits and observed visits.

The main difference between my paper and these other studies is that I focus on campaign spending in U.S. House elections rather than presidential campaigns. One consequence of this is that the parties' goals are more ambiguous because it is not clear whether they should maximize expected seat share or the probability of winning a majority of seats. In this sense, my substantive focus is closest to Snyder's 1989, although unlike his paper, I test my results empirically. My finding that parties maximize the expected number of seats and not the probability of winning a majority of seats is consistent with Jacobson and Kernell's (1985*b*) assertion that every congressional seat is valuable so parties should aim to maximize seats.

[3]These equilibria are consistent with the one's found in this paper.

This finding is also related to an interesting hypothesis set forth by Snyder (1989) that the leading party might want to be as certain as possible to try to win a majority while the trailing party might simply try to win as many seats as possible to improve future chances at controlling the legislature. Although interesting, I find no evidence that the parties play different strategies in U.S. House elections.

Methodologically, my paper is closest to Stromberg's since I use his model. The only major methodological differences between my paper and his are related to the forecasting methods. The two most significant of these differences are that first, I forecast House elections which are less predictable than presidential elections, and second, I use a technique that can update forecasts in real time as new polls become available. A third but relatively minor difference is that I utilize Bayesian techniques which account for uncertainty in the estimation of the district and national shocks, and allow for easier estimation of non-linear functions of the parameters.

The foundations of Stromberg's model date back to the probabilistic voting models of Lindbeck and Weibull (1987) and Dixit and Londregan (1996) used to analyze electoral competition. In these models, two competing political candidates must determine which interest groups should receive favors in order to maximize their probability of winning an election. When the political candidates do not differ in the efficiency in which they make transfers to various group, these models yield a "swing voter" equilibrium—similar to "close election equilibrium" in this paper—where interest groups that are the most politically central receive the most favors.[4]

The results here also shed some light on current debates regarding the motivations of campaign donors. These motivations are important because, as Stratmann (2005) notes, the predicted determinants of campaign contributions depend on the assumptions regarding contributor goals. Four objectives commonly cited in the literature are that contributions are a consumption good, an investment in policy, a means to buy access to a politician, or a

[4]See Johansson (2003) for an empirical test of the Lindbeck and Weibull and Dixit and Londregan models.

way to influence an election (e.g. Ansolabehere, Snyder and de Figueiredo 2003; Stratmann 2005).

This paper's finding that parties should (and do) contribute the most to districts that have the largest probability of being close is consistent with one of the more robust findings in the literature—that contributors spend more on close elections (Kau, Keenan and Rubin 1982; Jacobson 1985*a*; Poole and Romer 1985; Stratmann 1991).[5] In addition, the result that the financial industry donates more to members of the House Committee on Financial Services backs up research showing that candidates serving on congressional committees raise more money (Grier and Munger 1991; Romer and Snyder Jr 1994; Kroszner and Stratmann 1998). However, the explanatory strength of Stromberg's model relative to committee membership and other influence variables suggests that while both election-motivated and influence-motivated giving are important, election-motivated giving is likely to be more common.

The results in this paper depend on the assumption that additional contributions increase the probability that a candidate will win an election. While it seems difficult to believe that this would not be true, there is a large literature in political science and economics examining this question. Research findings are inconsistent, although more recent studies that have addressed biases have found that campaign spending does impact the vote.

The origins of this literature stem from Jacobson's (1978; 1980; 1985*a*) findings that campaign spending by challengers in congressional elections is very important but that incumbent spending has almost no effect on election outcomes. Subsequent work has refined these early studies by accounting for the endogeneity of candidate spending, which Jacobson's model does not address.[6] This is typically done by using instrumental variables (Gerber

[5]These studies have two main problems that this paper avoids. First, the closeness of an election is typically measured with an ex-post measure of the electoral margin or the lagged vote from the previous election. This does not mimic the decision of contributors who must make choices prior to election day and have considerably more information available to them then the vote in the previous election. Second, since they are not driven by theory, they do not provide any guidance the functional form of the relationship between the closeness of an election and spending, which should depend on the uncertainty (and probability distribution) of the predicted vote.

[6]Another explanation proposed by John Ferejohn and Morris Fiorina and mentioned in Jacobson (1985*a*) for Jacobson's finding that incumbent spending makes no difference is that there are almost no cases where incumbents do not respond to lavish spending by challengers with generous spending of their own.

1998; Green and Krasno 1988), including better control variables (Green and Krasno 1988) or addressing simultaneity biases (Erikson and Palfrey 2000). For instance, Green and Krasno (1988) instrument for incumbent spending in the previous election and control for candidate quality by creating an eight point scale based on various traits.[7] In contrast to Jacobson's earlier findings, they find that incumbent spending has a positive and statistically significant effect on the vote in House elections. Erikson and Palfrey (2000) reach similar conclusions by using a game theoretic model to show that simultaneity biases can be eliminated by only analyzing close races. After limiting analyses to these cases, they find the effect of both challenger and incumbent spending to be substantial.[8]

1.3 Model

This section describes a version of Stromberg's model suitable for analyzing U.S. House elections. The model is essentially the same as the one used in Strömberg (2008) to analyze presidential elections except for two minor changes that account for differences in the electoral settings. First, all districts are worth one seat while states in presidential elections are weighted according to their electoral votes. Second, parties maximize both the probability of winning a majority of seats (which is the equivalent of winning a majority of electoral college votes in presidential elections) and the expected number of seats. The second maximization problem is not considered in Strömberg (2008) because presidential candidates are concerned with winning the election, not maximizing their electoral votes.

[7] A related article is Gerber (1998), which uses challenger wealth as an instrument for challenger spending in Senate elections.

[8] One study that attempts to reduce biases but finds no effect of either challenger or incumbent expenditures on the vote is Levitt (1994). He controls for unobserved candidate characteristics by limiting his study to elections in which the same two candidates face each other more than once and taking first differences. The finding may lack external validity though since challenger quality is likely correlated with the probability of running in multiple elections. In addition, spending between candidates is unlikely to vary much from one election to the next which might lead to imprecise estimates.

1.3.1 Set Up

The model considers electoral competition between two parties, labeled Republican R and Democrat D.[9] During the campaign, each party must decide how to optimally allocate funds across the 435 Congressional districts. More formally, party $J = D, R$ must choose expenditures in district i, e_i^J, subject to the resource constraint,

$$\sum_{i=1}^{435} e_i^J \leq E^J, \tag{1.1}$$

where E^J is the amount of money party J has to spend on candidates.

The share of votes received by party D in district i is assumed to depend on four primary factors: spending by the national parties, predetermined characteristics of the district, the national political climate, and unknown shocks. The effect of the choice variable, e_i^J, is assumed to be an increasing concave function, $u(e_i^J)$, so that the effect of spending decreases with the amount of spending. The predetermined district characteristics and the national climate are known before the spending decision is made and can be used to make a prediction, V_i, of party D's vote share. Finally, there are two sources of uncertainty, a national error, δ, and a district specific error, ϵ_i. The national errors represent uncertain national swings that affect all districts equally and the district errors are unpredictable swings unique to each district. Both error terms are independently drawn from normal distributions,

$$h(\delta) = N(\delta|0, \sigma_\delta^2) \tag{1.2}$$

and

$$g_i(\epsilon_i) = N(\epsilon_i|0, \sigma^2). \tag{1.3}$$

[9] Third party candidates are ignored.

Letting $u(e_i^D) - u(e_i^R) = \Delta u_i$, party R will consequently win a district if,

$$\Delta u_i + V_i + \delta + \epsilon_i \leq 1/2. \tag{1.4}$$

The probability of a victory by party R conditional on expenditures, e_i^D and e_i^R, and the national swing, δ, is therefore,

$$G_i(1/2 - \Delta u_i - V_i - \delta),$$

where $G_i(\cdot)$ is the cumulative distribution function (CDF) of ϵ_i. It follows that if s_i is an indicator variable equal to 1 if party R wins a district and 0 if party D wins, then $s_i = 1$ with probability $G_i(\cdot)$ and $s_i = 0$ with probability $1 - G_i(\cdot)$. The total number of Republican seats is $S = \sum_{i=1}^{435} s_i$. Furthermore, since the $G_i(\cdot)$ are independently (but not identically) distributed conditional on δ, S follows a Poisson binomial distribution with mean

$$\mu_S = \mu_S(\Delta u, \delta) = \sum_{i=1}^{435} G_i(\cdot), \tag{1.5}$$

and variance,

$$\sigma_S = \sigma_S(\Delta u, \delta) = \sum_{i=1}^{435} G_i(\cdot)(1 - G_i(\cdot)), \tag{1.6}$$

where $\Delta\mu = (\Delta\mu_1, \Delta\mu_2, \ldots, \Delta\mu_{435})$ represents the utility differences resulting from any allocation of campaign resources across districts by the two parties.

1.3.2 Party Goals

Optimal strategies depend on the objective of the national parties. Unlike in presidential campaigns where the goal is clearly to win the election, the goals of the parties in House campaigns are less straightforward. As a result, I consider two plausible objective functions.

The first objective function assumes that parties simply maximize the expected number

of House seats. For party R, this is just the expectation of μ_S over the national shocks,

$$\mathbb{E}\left[S(\Delta\mu)\right] = \int \sum_{i=1}^{435} G_i(1/2 - \Delta u_i - V_i - \delta)h(\delta)d\delta. \tag{1.7}$$

A second possibility is that parties maximize the probability of winning a majority of seats. For party R this is,

$$P^R(\Delta\mu) = \int \Pr\left(\sum_{\mathrm{i}=1}^{435} \mathrm{s_i} > 218\right) \mathrm{h}(\delta)\mathrm{d}\delta. \tag{1.8}$$

This function is more difficult to maximize than equation 1.7 because it is not additively separable across districts; that is, party R's optimal strategy in one district depends on its strategy in all other districts. In order to solve the problem analytically, Strömberg (2008) suggests calculating the approximate probability of winning instead. Since the national and district shocks are independent, this can be done by using the Lyapunov Central Limit Theorem, which does not require random variables to be identically distributed. Using this approximation, the number of seats won by party R, $S = \sum_{i=1}^{435} s_i$, is asymptotically normally distributed with mean μ_S and variance σ_S. The approximate probability of party R winning the election is then,

$$P^R(\Delta\mu) = \int 1 - \Phi\left(\frac{218 - \mu_S(\Delta u, \delta)}{\sigma_S(\Delta u, \delta)}\right) h(\delta)d\delta, \tag{1.9}$$

where $\Phi(\cdot)$ is the standard normal cumulative density function.

1.3.3 Equilibrium

Let $e^j = (e_1^j, e_2^j, \ldots, e_{435}^j)$ be the allocation of campaign spending across districts by party j and $f^R(e^D, e^R)$ be the objective function used by party R. A Nash equilibrium in pure strategies (e^{D*}, e^{R*}) is then characterized by,

$$f^R(e^D, e^{R*}) \geq f^R(e^{D*}, e^{R*}) \geq f^R(e^{D*}, e^R), \tag{1.10}$$

where e^D and e^R satisfy the budget constraint in equation 1.1. The game has a unique interior NE which satisfies,

$$\frac{\partial f^J}{\partial e_i^J} = Q_i u'(e_i^{J*}) = \lambda^J, \tag{1.11}$$

where $Q_i = \partial f^J / \partial \Delta u_i$ and λ^J is the Lagrange multiplier for party j.[10] Since $u'(e^J)$ is decreasing in e^J, the parties allocate more resources to districts with higher values of Q_i. The value of Q_i will of course depend on the choice of the objective function.

When parties maximize the expected number of seats, Q_i is easy to calculate and is just,

$$Q_i = Q_i^{seats} = \int g_i(1/2 - \Delta u_i - V_i - \delta) h(\delta) d\delta. \tag{1.12}$$

Not unexpectedly, districts with a predicted vote share close to 1/2 should receive the most expenditures. Spending should also be more concentrated when the error terms, σ and δ, are smaller. If on the other hand, parties maximize the probability of winning a majority of seats, Q_i is more intricate. For party R, it is,

$$\begin{aligned} Q_i = Q_i^{maj} &= -\int \left(\frac{\partial \Phi(\cdot)}{\partial \mu_S} \frac{\partial \mu_S}{\partial \Delta u_i} + \frac{\partial \Phi(\cdot)}{\partial \sigma_S} \frac{\partial \sigma_S}{\partial \Delta u_i} \right) h(\delta) d\delta \qquad (1.13) \\ &= \int \frac{1}{\sigma_S} \phi(x(\delta)) g_i(\cdot) h(\delta) d\delta \\ &\quad + \int \frac{1}{\sigma_S} \phi(x(\delta)) x(k, \delta) \left(1 - 2G_i(\cdot)\right) g_i(\cdot) h(\delta) d\delta, \qquad (1.14) \end{aligned}$$

where $x(\delta) = (218 - \mu_S) / \sigma_S$. The first term is the effect of an increase in spending on the mean number of Republican seats while the second term is its effect on the variance. Parties have an incentive to influence the variance because they want to increase the probability of a desirable outcome. The trailing party will want to increase the variance to increase the probability of a major change in the election outcome while the leading party will want to do just the opposite. The trailing party can increase the variance by spending more on districts

[10] For a proof of uniqueness see Strömberg (2008).

in which its candidate it losing and the leading party can decrease the variance by spending more on districts in which it is winning. Intuitively, the leading party does not need to worry about districts in which it is losing because it only needs to make sure that it holds on to the one's it is leading in.

As shown by Strömberg (2008), an alternative interpretation of equation 1.13 is that it is the probability that a district is 1) decisive in whether or not a party wins a majority of seats and 2) the district is a swing district. Following Stromberg, I call such a district a "decisive swing district". The probability of being a swing district is the probability that an electoral race is tied (or at least very close), while a district is decisive if winning (or losing) that district would make the difference between winning (or losing) a majority of seats. The idea that parties should spend more money on swing districts is consistent with the first order condition in equation 1.12. The idea that parties should spend more on decisive (also known as pivotal) districts differentiates the two maximization problems.

1.3.4 Functional Form

In order to solve for equilibrium spending, e_i^{J*} and calculate Q_i, it is necessary to make an assumption about the functional form of $u(e_i^J)$. One functional form particularly amenable to empirical analysis is the logarithmic form, $u(e_i^J) = \theta \log(e_i^J)$, which results in the first order condition for district k and party J,

$$e_k^{J*} = \frac{Q_k}{\sum Q_i} E^J. \tag{1.15}$$

Each party spends the same fraction of the budget on district k, but e_i^{R*} only equals e_i^{D*} if both parties have identical budgets so that $E^R = E^D$.[11] Equation 1.15 implies that Q_i is evaluated at $\Delta u_i = \theta \log(E^D/E^R)$, which reduces to $\Delta u_i = 0$ when the budgets are equal. If θ, E^D or E^R are unknown (and time-varying), then this term will be incorporated into the national shock, δ, during estimation.

[11] For a formal proof that $e_i^{D*} = e_i^{R*}$ in equilibrium when $E^R = E^D$ see Strömberg (2008)

1.4 Estimation

To calculate Q_i for each district in each election, it is necessary to estimate the variances of the district and national shocks, σ^2 and σ_δ^2, as well as the two-party vote in each district, V_i. In the section I describe a Bayesian methodology that can estimate these parameters using historical political and economic information and when available, polling data. The historical information provides a forecast of the election as of September 1st in each election year. Polling data from September 1st up until election day is then used to update the forecasts as the campaign progresses. This allows for an examination of the relationship between post-August spending and Q_i at the election year level and whether campaign strategies responds to new polls (which change the values of Q_i).

1.4.1 A Bayesian Hierarchical Model

Previous research has shown that national elections are highly predictable from one year to the next using standard regression techniques (e.g. Campbell 1992; Gelman and King 1993; Kastellec, Gelman and Chandler 2008). These "historical" models are often said to use "fundamentals"— such as the political ideology of citizens in a political area or the state of the economy—to forecast the vote for a given election. In Stromberg's model, these fundamentals are predetermined characteristics and national trends that can be used to make a prediction, V_{iy}, of the Democratic share of the two-party vote in district i and election year y. To move from theory to empirics, I assume that this prediction is a linear function of a matrix of explanatory variables, X_{iy}. Using equation 1.4, the Democratic share of the two party vote, say v_{iy} is then,

$$v_{iy} = X_{iy}\beta + \delta_y + \epsilon_{iy}. \tag{1.16}$$

Since δ_y and ϵ_i are assumed to be normally distributed, this can be estimated using the Bayesian hierarchical model,

$$v_{iy} \sim N\left(X_{iy}\beta + \delta_y, \sigma^2\right) \tag{1.17}$$

$$\delta_y \sim N(0, \sigma_\delta^2), \tag{1.18}$$

where β is a vector of coefficients. δ_y is modeled as a random effect and is centered at 0 because X_{iy} includes an intercept. The Bayesian estimation is completed with uniform hyperprior distributions on β, σ and σ_δ.

1.4.2 Incorporating Polls with a Bayesian DLM

In recent elections, polling firms have begun polling certain races for the House of Representatives. Since these polls provide (hopefully) useful information about the likely vote in a district, it would be wise to include them in the forecasting model. The most straightforward way to do this would be to simply include them as additional columns in the data matrix X_{iy}. However, this is problematic because (1) polling data at the district level are a relatively recent phenomenon, (2) firms do not poll all districts, (3) polls are measured with error and may differ from the true state of public opinion at any given time and (4) districts that are polled are typically polled multiple times during the election campaign.

One strategy that can help alleviate these issues is to treat polls as additional data points, rather than as independent variables in a historical regression. This technique is commonly used by researchers and media outlets forecasting presidential and Senate elections.[12] The primary difficulty is that the polls are highly correlated and should not be treated as independent data. As a result, forecasters attempt to average the polls so that the most informative ones receive the most weight. This weighted average of polls can then, in turn,

[12]See, for instance, the Huffington Post's forecasts at http://elections.huffingtonpost.com/, which are based on Simon Jackman's poll-tracking model; Drew Linzer's forecasts at http://votamatic.org/; neuroscientist Sam Wang's website http://election.princeton.edu/; and Nate Silver's http://fivethirtyeight.com/.

be combined with information from a regression analysis, with weights that should be based on their respective variances.

As noted by the founder of http://fivethirtyeight.com/, Nate Silver, the variability of a poll's forecast can be thought of as a function of three major components: sampling error, temporal error and pollster-induced error. The first two terms are relatively straightforward: sampling error occurs because each poll is based on a sample from the electorate and temporal error is due to uncertainty about opinion shifts between the date a poll is taken and election day. The final term, pollster-induced error, can be thought of as the error left over after accounting for sampling and temporal error. A major part of this residual term can be attributed to house effects, or time-invariant biases specific to certain pollsters. However, this error can also occur due to other polling difficulties such as undecided voters, respondents who will not vote in the actual election or respondents that do not express their true voting intentions.

In this paper, I utilize a state-space framework that can sequentially adjust forecasts as new polls become available. This model-based poll averaging approach is very similar to the the state-space poll-tracking model employed by Simon Jackman (Jackman 2005, 2009) and the forecasting model based on reverse-random walks used by Drew Linzer (Linzer 2013)—which built on the idea used in (Strauss 2007).

One important feature of House elections that must be accounted for is that district polling is sporadic both over time and across districts. The incomplete nature of the data is important for calculating district and national errors because national errors based solely on district polling use less than half of all districts at any given date and the subset of districts available to calculate national errors changes over time (since different districts are polled at different times). One way around this is to to use the generic congressional vote as a measure of the national vote, since, unlike district polls, polling firms begin conducting polls of the generic congressional vote for the next election at a consistent rate almost as soon as election results are in.

To incorporate the national polls, I separate the vote in each district into the national vote and the district vote relative to the national vote as in Lock and Gelman (2010) and Strauss (2007). This separation allows me to use all available polling data to decompose national and local variation as required by the theoretical model.[13]

The model of the national vote follows the model employed in Jackman (2005) and Jackman (2009), which provides a framework for "pooling" the polls over the course of a campaign. To set notation, let $t = 1, \ldots, T$ represent days of the campaign where $t = 1$ corresponds to the first day of the campaign season and $t = T$ is election day. Furthermore, let k index a poll with sample size N_k so that the number of respondents, n_k, from poll k who support the Democratic candidate follows a binomial distribution,

$$n_k \sim \text{Bin}(\text{N}_\text{k}, \pi_\text{k}), \tag{1.19}$$

where π_k is the proportion of voters from poll k who intend (or report to intend) to vote for a (generic) Democratic candidate. Since the sample size of each poll is relatively large, the observed proportion of respondents who report that they intend to vote democratic, $y_k = n_k/N_k$ is approximated well by a normal distribution,

$$y_k \approx N(\pi_k, \sigma_k^2), \tag{1.20}$$

where $\sigma_k^2 = y_k(1 - y_k)/N_k$. However, the parameter of underlying interest is not π_k, but the actual state of national opinion at time t. The π_k are consequently modeled as a function of two components: the actual state of opinion, μ_t, and a house effect, λ_j, specific to polling firm $j = 1, \ldots, J$,

$$\pi_k = \mu_{t[k]} + \lambda_{j[k]}. \tag{1.21}$$

[13] Another strategy is to model the correlation between the national polls and the district polls in a multivariate time-series model (see Jackman (2012) for a brief explanation of a model that does this). This approach is not taken here because it is not consistent with the theoretical model used in this paper.

Equation 1.21 is not identified because one could shift $\mu_{t[k]}$ up/down and $\lambda_{j[k]}$ down/up by the same constant without changing the value of π_k. As a result, I use the identifying restriction that the house effects sum to zero, $\sum_j \lambda_j = 0$.

As currently specified, the model only provides a snapshot of national opinion on any given day. To forecast the election, it is necessary to estimate μ_T, which is an estimate for national opinion on the day of the election. Since forecasts are made on days $t' < T$, it is therefore necessary to make assumptions about the movement of μ_t from one day of the campaign to the next. Since there is no reason to expect there to be any trends in polling, it is reasonable to expect μ_t to follow a random walk, so that the full model can be written as,

$$y_k = \mu_t + \lambda_j + v_k,\ v_k \sim N(0, \sigma_k^2) \tag{1.22}$$

$$\mu_t = \mu_{t-1} + w_\mu,\ w_\mu \sim N(0, \sigma_\mu^2), \tag{1.23}$$

where σ_μ^2 is an estimate of the daily change in μ_t. As shown in Section 1.B.1, equation 1.22 and equation 1.23 form a state-space model, or more specifically, (since the model is linear and errors terms are Gaussian) a dynamic linear model (DLM). Equation 1.22 is known as the observation equation while equation 1.23 is known as the state equation.

A model for the district vote relative to the national vote proceeds in the same manner as the model for the national vote, but with a few additional differences. The first difference is that it is impractical to correct for house effects because there are only a few polls published per polling firm.[14] The second difference is that national opinion at time t is not actually observed, so the relative vote cannot be observed either. In practice, this is not a large problem because national opinion, μ_t, is estimated very precisely using equation 1.22 and equation 1.23 due to the abundance of large national polls.[15]

[14]Pollsters tend to focus on specific districts; there are consequently hundred of pollsters, but only a few polls from each one. This stands in contract to polls of the national vote, for which there are far fewer pollsters and many more polls published per pollster.

[15]The standard deviation of μ_t is typically around 0.004.

For the model of the relative district vote, let l index district polls and continue to let i index a district. Define the deviation of a district poll from the national vote as $d_l = y_l - \mu_{t[k]}$. The state-space model for the relative district vote is then,

$$d_l = \xi_{it} + v_l,\ v_l \sim N(0, \sigma_l^2) \tag{1.24}$$

$$\xi_{it} = \xi_{i,t-1} + w_\xi,\ w_\xi \sim N(0, \sigma_\xi^2), \tag{1.25}$$

where $\sigma_l^2 = y_l(1 - y_l)/N_l$, N_l is the sample size of the lth poll, $\xi_{i[l]t[l]}$ is an estimate of the deviation of opinion in state i from national opinion at time t, and σ_ξ^2 captures the variance of day to day movements in ξ. Equation 1.24 and equation 1.25 are just a simple multivariate extension of the model of the national vote that ignores house effects.

The overall forecast of the two-party vote for Democrats from each district is $\mu_T + \xi_{iT}$. Separate forecasts of μ_T and ξ_{iT} can basically be estimated using a Kalman filter[16], which the caveat that Bayesian MCMC techniques are needed to estimate the variance of the state equations and the house effects in the model for the national vote. The Kalman filter is instructive because it quantifies the relative weight attached to previous versions of the states and new polls. For example, using the Kalman filter, the mean and variance of the latent states in district i on day t given polls up to day t, $Y_{1:t}$ are,

$$m_{it} = E(\xi_{it}|Y_{1:t}) = \left[\frac{m_{i,t-1}}{C_{i,t-1} + \sigma_\xi^2} + \sum_{l \in \mathcal{P}_{it}} \frac{d_l}{\sigma_l^2}\right] \cdot C_{it}, \tag{1.26}$$

$$C_{it} = \text{Var}(\xi_{it}|Y_{1:t}) = \left[\frac{1}{C_{i,t-1} + \sigma_\xi^2} + \sum_{l \in \mathcal{P}_{it}} \frac{1}{\sigma_l^2}\right]^{-1}, \tag{1.27}$$

where $\mathcal{P}_{it}$ refers to the set of all polls published for district i on day t.[17] The mean of

[16]The Kalman filter is commonly used in engineering to track the movement of objects such as satellites or aircraft that are measured with noisy data. It is also frequently used in Macroeconomics and in political science to track public opinion.

[17]There is almost never more than one poll for a given district and time period of a reasonably short duration (such as two weeks). For national elections, there are almost always multiple polling firms surveying on a given day or time period.

ξ_{it} is thus a weighted average of its mean on the previous day, $m_{i,t-1}$ and the deviation of all new polls from national opinion, $\sum_{l \in \mathcal{P}_{it}} d_l$, with weights proportional to their respective precisions. The precision of each poll is equal to the inverse of its sampling error and the precision of m_{t-1} is the inverse of the sum of its variance and error in the movement of the states from one period to the next. The variance of ξ_{it} is just the inverse of the poll precision plus the prior (m_{t-1}) precision. The interpretation for μ_t is identical except that new estimates are a weighted average of prior states and new polls less the democratic bias of the polling firm (i.e. $y_k - \delta_{j[k]}$).[18]

To incorporate information from the historical regression, I treat forecasts from the hierarchical model described above as election day polls for both the national and district models. The two-party vote given to the national pseudo poll is the mean average district vote from the hierarchical model and the Democratic vote shares given to each district's pseudo poll is the mean of the district forecast from the regression less the mean of average district vote. The corresponding variances are calculating using the posterior predictive distributions of these quantities.[19] The regression forecasts in the national (district) models consequently receive weights proportional to the regression forecast errors of the average district vote and the relative district vote.

When estimating μ_T and ξ_T prior to the election, there will be gaps between the last published poll and the pseudo poll from the regression. The Kalman filter helps bridge this gap by pushing the latent states forward toward election day. The relative weights received by the regression analysis and the polls depends largely on the number of days until the election. For instance, due to the random walk assumption, the precision of μ_{T-1} for a forecast of the national vote made on day t' is equal to $1/\left[C_{i,t'} + (T - t') \cdot \sigma_\mu^2\right]$, which is linearly decreasing in time and in the day to day movements of the states. It follows that the regression analysis receives more weight when the election day is far away and when there is more movement in the polls.

[18] See Appendix 1.B for additional details.
[19] See Section 1.6.2 for more details.

The Kalman filter described in this section assumes that the variances of the states are known. Since, in practice, this is clearly not the case, I estimate all of the model parameters jointly using Bayesian methods. To do so, I assign the unknown variance parameters, σ_μ^2 and σ_ξ^2, inverse gamma priors. The house effects, λ_j, are given a normal prior centered at 0. The posterior density is simulated with a Gibbs sampler, which is described in Appendix 1.B.

1.5 Data

The analyses in this paper utilize four main sets of data: data on nationwide variables, data on representatives from the U.S House, campaign contributions data, and polling data. The main nationwide variables collected are the president's net approval rating and the Gallup generic Congressional ballot. Both the presidential approval rating and the Gallup generic ballot were obtained from the Roper Center Public Opinion Archives. The net approval ratings is the percentage of survey respondents who approve of the president minus the percentage who disapprove. The ratings are based on polls conducted by various polling organizations multiple times each month. The generic ballot is based on a survey question that asks voters whether they intend to vote for a generic (does not include candidate names) Republican or Democrat in the House election. Survey results for this question are more difficult to obtain than presidential approval ratings since each poll must be searched for individually so I consequently restrict the generic ballot questions to those asked by Gallup in August (the month the forecast is being made).

The second dataset consists of data from three sources: data on House elections from 1946 - 2012 obtained from Gary Jacobson; the committee assignments of members of Congress from each district (Stewart III and Woon 2015); and DW-Nominate scores from `http://voteview.com/`. The Gary Jacobson data provides information on a number of important characteristics of House elections at the district level. These include the Democrat's share of the two party vote in both House and presidential elections, whether an incumbent is running for reelection (and the party of the incumbent), and whether a challenger has previously held

office. The DW-nominate scores are the first dimension scores originally developed by Keith Poole and Howard Rosenthal, which can be interpreted as the liberal-conservative divide in modern politics (Poole and Rosenthal 1997, 2011).

The third dataset contains campaign finance data provided by the Center for Responsive Politics (CRP) at https://www.opensecrets.org/. CRP obtains the data from the Federal Election Commission and adds value to it by cleaning and categorizing the data. The data covers campaign contributions from individuals (above $200) and Political Action Committees (PACs).[20] The data can be separated by date, individual employer, and PAC type, which allows me to track contributions during different times of the campaign and to focus on spending by different types of contributors.

The final dataset covers district and national polls for the 2010 House election. The district polls are all of those used by the New York times to forecast the 2010 House election and the national polls are the polls listed for the generic congressional vote on the website http://www.realclearpolitics.com/. Each poll contains the date that the poll was taken, its sample size and the proportion of respondents favoring the Democrat and Republican candidate.

The rest of this section describes the specific uses of these datasets in more detail. For more detailed descriptions of the variables and information on sources see the data appendix (Appendix 1.D).

1.5.1 Forecasting Variables

The model used in this paper includes variables (and error terms) at the national and district level.[21] These variables were obtained from the nationwide data and the data on representatives from the U.S. House. Forecasts are made as of August of each election year using post 1980 data so all variables are measured before September 1st.

Recall that the primary dependent variable is the Democratic share of the two party

[20] Contributions from party committees are included in the PAC table.

[21] Regional variables were also considered but they did not improve the fit of the model and only increased model complexity.

vote. This is not the only possible choice for the response variable but it is frequently used in the literature. Other choice that are essentially identical include the incumbent party's share of the vote or the margin of victory for the incumbent candidate.

House elections are highly persistent from one election to the next so they can be predicted quite accurately using only the lag of the district vote. The lag is of course unavailable without substantial reaggregation in redistricting years so, as is common in the literature, the analysis excludes years ending in 2.[22] The lagged vote is tied in many respects to individual candidates and is not surprisingly a much stronger predictor of the vote when incumbents are running for re-election than in open-seats (Gelman and Huang 2008). My model consequently includes the lag of the presidential vote in each district in addition to the lag of the House vote. The presidential vote is subtracted from the nationwide presidential vote to control for national trends and can be interpreted as a measure of the degree to which a district leans toward one party or another.[23][24]

One of the largest and most consistent findings in the political science literature on American elections is that, all else equal, incumbent candidates receive more votes than challengers (Gelman and King 1990; Lee 2001; Ansolabehere and Snyder Jr 2002; Gelman and Huang 2008). To model this incumbency effect, I use the variable, *Incumbent*, which is equal to 1 if the incumbent is a Democrat, -1 if the incumbent is a Republican and 0 if there is no incumbent running. $+1, 0, -1$ variables are used repeatedly in the analysis and are always equal to +1 for Democrats and -1 for Republicans; they should be thought of as dummy variables that are constrained to have the same impact on the vote for both Democratic and Republican candidates (recall that the dependent variable is the Democratic candidate's share of the two-party vote).

[22] Examples of other studies that deal with redistricting years in this manner include Gelman and King (1990), Kastellec, Gelman and Chandler (2008), and Gelman and Huang (2008).

[23] This variable is very similar to the Cook Partisan Voting Index (Cook PVI), which compares the two-party vote in the past two presidential elections to the nation's average share of the same presidential vote. The main difference is that I only include the previous presidential election in order to deal with redistricting.

[24] Models that interacted both the lagged vote and the centered presidential vote with whether incumbents were running were also considered, but did not lead to improvements in the fit of the model.

Since more experienced candidates tend to do better at the polls, I include two $+1, 0, -1$ experience variables: *Freshman incumbent* and *Previous office holder*. *Freshman incumbent* is equal to +1(-1) if a Democratic (Republican) incumbent was elected for the first time in the previous election and 0 otherwise. *Previous office holder* is equal to +1 or -1 if a Democrat or Republican challenger had previously held office and was running against either an incumbent or a challenger without previous experience; it is equal to 0 in all other situations including open seats in which both candidates had previous political experience.

The model also accounts for the ideology of candidates relative to the ideology of voters in their districts. Candidate ideology is measured using the first dimension DW-nominate score—which measures the political ideology of candidates and ranges on a liberal-conservative scale from -1 to 1. A measure of district ideology on the same scale as the DW-nominate score was created by taking a weighted average of the DW-nominate scores of the most recent presidential candidates for each party, with weights equal to each party's share of the district two-party vote in the most recent presidential election.[25] The variable used in the model, *Relative 1st dimension DW-nominate score*, is the DW-nominate score minus the district ideology score.[26] In incumbent districts, this variable will have a positive sign if candidates do better at the polls when they have more moderate voting records (perhaps because they appeal more to independent voters), conditional on winning the primary election. In open seats, this variable will be positive if voters tend to elect a Democratic candidate when the previous candidate(s) had a conservative voting record and a Republican candidate when the previous candidate(s) had a liberal voting record.[27][28]

[25]Michael Dukakis does not have a DW-nominate score so his ideology was imputed using the mean score among Democratic presidential candidates between 1980 and 2010.

[26]This variable outperformed another variable which interacted the absolute value of the DW-nominate with the party controlling the seat, suggesting that voting moderation is more strongly associated with higher vote shares than a candidate's ideological distance from the ideology of his or her voters.

[27]In open districts, the DW-nominate score is equal to the average DW-nominate score among all representatives in the previous Congress. Most districts only have one representative per congressional session but certain events such as deaths create scenarios in which this is not the case.

[28]In OLS regressions, the coefficient on the interaction of the relative DW-nominate score variable with whether a seat is open is essentially zero, implying that the effect of the variable does not differ by whether a seat is open or not.

The three nationwide variables are the president's average August approval rating, the August generic ballot and an indicator variable for whether the election is being held in a midterm year.[29] The *August generic ballot* variable is an average of generic ballot polls in the month of August in each election year. Both the midterm variable and the presidential approval variable are multiplied by +1 if the president is a Democrat and -1 if Republican. The impetus for the *midterm election* variable is previous research showing that the party of the president usually loses seats during midterm election (e.g. Erikson 1988).

Summary statistics for these variables are presented in Table 1.1. Variables rarely exceed one in absolute value and are thus on a similar scale to the dependent variable.

Table 1.1: Summary Statistics for Forecasting Variables

Variable	Min	Median	Max
Dem. share of district vote in last election	0.077	0.536	0.971
Relative Dem. share of presidential vote in last election	−0.320	−0.019	0.538
Incumbent	−1.000	1.000	1.000
Relative 1st dimension DW-nominate score	−0.897	−0.149	1.100
Freshman incumbent	−1.000	0.000	1.000
Previous office holder	−1.000	0.000	1.000
August presidential net approval rating	−0.567	−0.030	0.369
August generic ballot	0.467	0.513	0.615
Midterm election	−1.000	0.000	1.000

Notes: *Relative Dem. share of presidential vote in last election* is the deviation of the Democratic share of the presidential vote in each district from the national vote in the most recent presidential election. *Incumbent* is equal to +1 or -1 depending on whether a Democrat or Republican is running for re-election and 0 otherwise. *Relative 1st dimension DW-nominate score* is the 1st dimension DW-nominate score minus a measure of district ideology. *Freshman incumbent* is defined in the same was as *Incumbent* but is only equal to +1 or -1 if the incumbent was elected for the first time in the previous election. *Previous office-holder* is equal to +1 (-1) if a non-incumbent Democrat (Republican) had previously held office and 0 otherwise (or if the seat was open and both candidates had previously held office). *August presidential net approval* is the president's average net approval rating in August multiplied by +1 if the current president is a Democrat and -1 if Republican. *August generic ballot* is an average of generic ballot polls in the month of August. *Midterm election* is equal to 0 in non-midterm years and +1 (-1) if the president is a Democrat (Republican) and it is a midterm election year. Other variables such as second quarter GDP growth and the party currently controlling a district were also considered, but were not included in the final model because they were not statistically significant and they did not improve the fit of the model.

1.5.2 Campaign Spending

House candidates receive funds from three primary sources: party committees, PACs, and individuals.[30] Campaign donors can, in turn, spend on the behalf of candidates by either

[29] A variable including second quarter GDP was also considered but it was not statistically or economically significant.

[30] Organizations—such as firms, unions or trade associations—that wish to spend money to influence federal elections must create a separate source of funds known as a PAC.

making a direct contribution, a coordinated expenditure or an indirect expenditure for or against a candidate.[31] Independent expenditures are not limited by law but cannot be coordinated with campaigns. By contrast, direct contributions and coordinated expenditures can be coordinated with a campaign but are limited by federal law.[32]

Although campaign fund-raising is a year-round endeavor for political candidates, it follows a cyclical pattern and naturally heats up as election day nears. This is shown in Figure 1.1, which plots average daily spending by PACs in election years by month across the 2000 - 2010 elections. Expenditures are fairly constant from January until August and then jump considerably in September before peaking at around $4 million dollars per day in October and then decreasing slightly in November.[33]

Political parties are the focus of this paper, so it is worth examining the spending patterns of party committees in detail. The official campaign arms of the Democratic and Republican party in the House are the Democratic Congressional Committee (DCCC) and the National Republican Congressional Committee (NRCC). The principal goal of these organizations is to assist House candidates and help elect them to Congress. Direct contributions and coordinated expenditure are severely limited by contributions limits, so almost all spending is made independently of candidates. The Democratic National Committee (DNC) and Republican National Committee (RNC) sometimes spend money on House candidates as well, but the amount is trivial compared to the amount spent by the DCCC and NRCC.

While the national party committees have clear incentives to maximize the electoral success of their respective parties, it is unclear whether other organization share this same goal. Herrnson (2009) has argued that the reach of political parties extends considerably further than with the national party committees. He views parties as multilayered coalitions that

[31] A common example of an independent expenditure is a TV advertisement praising a candidate or criticizing an opponent. Before 2010, corporations and unions could only fund independent expenditures from their PACs; however in 2010, the U.S. Supreme Court rule in *Citizens United v. Federal Election Commission* that corporations and unions could use their own treasuries to raise money for independent expenditures.

[32] Coordinate expenditures, are, as the name suggests, expenditures made on the behalf of campaigns that can be discussed with the campaign.

[33] November expenditures only include spending that occurred prior to election day. Between 2000 - 2010, the latest election occurred on November 7th.

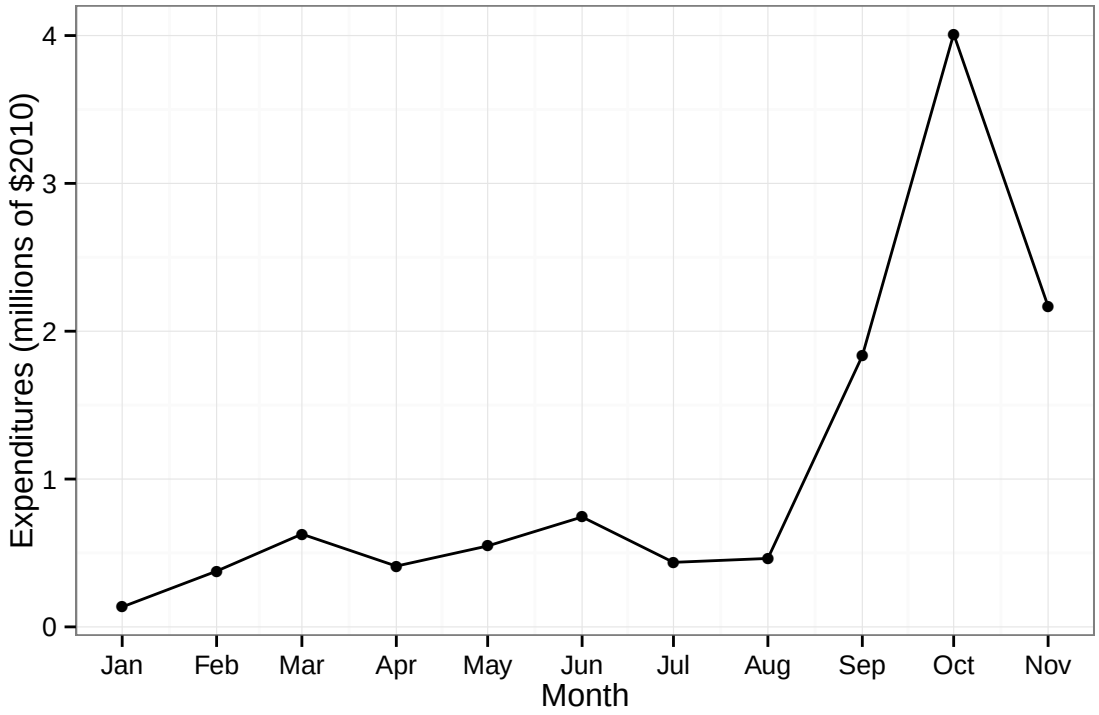


Figure 1.1: Average Daily Expenditures by PACs on Candidates by Month, 2000 - 2010

Notes: Expenditures are the average (across elections from 2000 to 2010) of the sum of expenditures on candidates by all PACS by month. Expenditures in non-election years are not included. Spending is in millions of 2010 dollars. November spending only include expenditures prior to election day.

include actors not usually considered as components of the party. I adopt this classification scheme and separate parties into three separate categories: national party committees, party-connected committees, and allied PACs.

The national party committees are the DCCC, NRCC, DNC and RNC. Party-connected committees consist of state and local party committees, candidate committees, and House candidate's leadership PACs. These committees have incentives to adopt spending strategies based on private or local goals, but Herrnson (2009) argues that they attempt to advance their party's collective goals as well. Spending by state and local party committees is also tied to the desires of the national parties because a significant portion of their receipts come from transfers from the DNC and RNC.[34] Finally, following Herrnson (2009), I consider

[34]The state and local party committees receive most of their funds from individual contributions and

allied PACs to be PACs that spend over 90% of their funds on one party's candidates.

Post-August spending for these groups and other non-aligned PACs are shown in Table 1.2. The table reports spending per district averaged over all non-redistricting year elections from 2000 - 2010. The national party committees spend considerably more than all other groups as national party committee spending is between 45 and 50 percent of all PAC and party spending for both the Democratic and Republican parties. Allied PACs and other non-aligned PACs are the next largest spenders. PACs are considerably more likely to be allied with Democrats than Republicans, although spending by non-aligned PACs is similar for both Republican and Democratic candidates. Spending by party-connected committees makes up less than 10% of total PAC and party committee spending, making them the least prolific spending group. In total, slightly more is spent on Democrats than Republicans.

Table 1.2: Average Post-August Spending by Political Parties per District in Non-Redistricting Years, 2000 - 2010

	Democratic party		Republican party	
	$	Fraction of total	$	Fraction of total
National party committees	114,773.40	0.457	107,813.99	0.501
Party-connected committees	21,567.06	0.086	20,115.03	0.093
Allied PACs	58,572.38	0.233	29,828.05	0.139
Other PACs	56,311.25	0.224	57,516.67	0.267

Notes: 2002 is omitted because it is a redistricting year. Spending consists of direct contributions, coordinated expenditures and independent expenditures for a canditate or against the opposing candidate in 2010 dollars. Spending figures are per district means averaged across the 2000 to 2010 non-redistricting year elections. Expenditures prior to September 1st are excluded. National party committees include the DNC, DCCC, RNC and RCCC. Party-connected committee are candidate committees or leadership PACs. Allied PAC consist of PACs that have spent 90% or more of their funds on one political party. Other PACs are all other PACs that do not fit into the other three categories.

Candidates also receive a large percentage of their funds from individuals. Contributions from individuals donating over $200 are itemized but contributions from smaller donors are not. During the 2009 - 2010 election cycle, only 0.26% of the U.S. population gave more than $200, but total contributions for these individuals summed to $1.9 billion while contributions from all other individuals were only $1.1 billion.[35] Individuals who donated over $200 spent $99,189.29 and $101,575.8 per district on Democratic and Republican candidates af-

transfers from affiliated party committees.

[35]See https://www.opensecrets.org/bigpicture/.

ter August 31st in non-redistricting years from 2000 - 2010 respectively. Since contributions for individuals who donated less than $200 are not itemized, I cannot make a comparable calculation for donations from these individuals. However, if one assumes that spending on House candidates mirrors total spending, then contributions from small donors should have totaled around $58,000.[36] Total PAC and party committee spending on House candidates is therefore larger than spending from individuals and the national party committees are the largest single source of funds for House candidates.

1.5.3 Polling Data

A steady stream of polls asking about the generic congressional vote during the 2010 election were released starting as early as December of 2008. All told, the website Real Clear Politics gathered 270 polls published by 30 different polling organizations over 670 days with the last poll conducted 4 days prior to the November 2nd election.[37] Most pollsters only published a few polls (median = 3.5), but the two most prolific pollsters, Rasmussen and Gallup, conducted 49% of all polls. The median sample size across surveys was 1600 and the average percentage of undecided respondents was 15%.

District polls were considerably less prolific. Polls were published by 113 different pollsters for 174 out of the 435 Congressional districts and the median number of polls per district in districts with at least 1 poll was 3. The median sample size for these polls was 400.

Figure 1.2 provides summaries of the frequencies at which district polls were fielded. As shown in panel (a) of the figure, polls were considerably more prevalent in close races (those in which the final vote share was close to 0.5), although few to no polls were published in some close races. This is important for forecasting the election because it means that the districts that are the most difficult to predict are the ones that have the most polling data.

[36] Individuals donating less than $200 spent $1.1 billion/$1.9 billion ≈ 0.58 as much as larger donors in 2010. These larger donors, in turn, spend around $100,000 per district so smaller donors likely spent close to $0.58 \cdot \$100,000 \approx \$58,000$.

[37] Polls are typically conducted over multiple days. I consider the poll date to be the midpoint between the start and end date of polling (when the poll is conducted over an even number of days I use the median day closest to the start day).

This matters for optimal spending as well since it reduces uncertainty about the closeness of a race and provides parties a better sense of which districts to target when attempting to influence the variance of the election.[38]

Panel (b) of the figure shows the cumulative number of polls during each day of the campaign. Nearly 2/3 of the polls were fielded after the start of September, when the rate of polling began accelerating quickly. This temporal pattern is consistent with the monthly pattern in contributions, suggesting that the final two months of campaigns capture the most crucial moments.

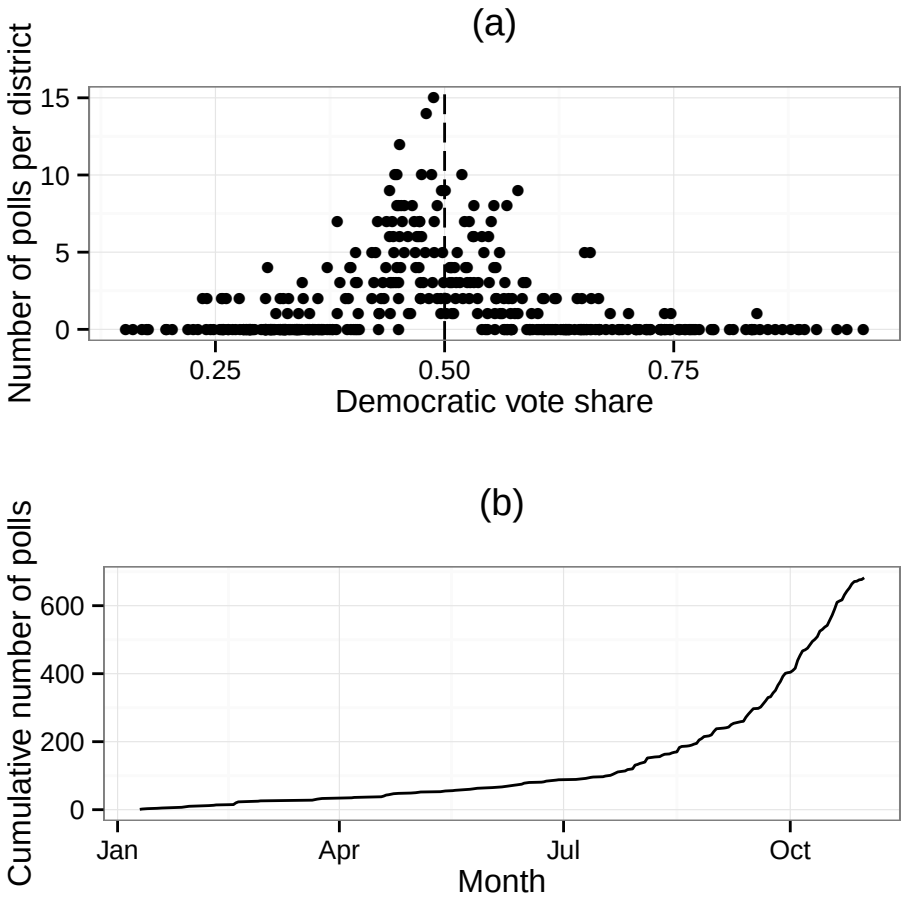


Figure 1.2: Summary Plots for District Polls

[38]Recall that the trailing (leading) party in total seats has an incentive to spend more on districts in which it is losing (winning) if parties are trying to win a majority of seats.

1.6 Forecasting the Two-Party Vote and Calculating Q_i

1.6.1 Results from Bayesian Hierarchical Model

The hierarchical model was implemented using Stan, which uses Hamiltonian Monte Carlo sampling (see Appendix 1.A for Stan code).[39] The model was fit 5 separate times using post 1980 data to forecast (out-of-sample) the 2000, 2004, 2006, 2008 and 2010 elections.[40] Uncontested seats were dropped from the model because parties would have known which districts were uncontested by September 1st.[41][42] The posterior distribution for each of the 5 models was simulated using the last half of 5 chains of $2,000$ iterations each. The Gelman and Rubin potential scale reduction factor $\hat{R}$ was approximately 1 for all model parameters in all 5 fits, which suggests that the Markov Chain successfully converged each time.

Posterior means and quantiles for the parameters using data from 1980 - 2008 (which was used to forecast the 2010 election) are shown in Table 1.3. All of the variables (except the August generic ballot) are precisely estimated and have 95% credibility intervals that do not include zero. The posterior densities for all parameters are all approximately normally distributed, which is consistent with the central limit theorem in a Bayesian context (see Figure 1.C.1 in Appendix 1.C).[43]

The signs of the variables are generally as expected. For example, the president's party does worse in midterm years, challengers with previous political experience receive more votes and voters tend to prefer representatives with less extreme DW-nominate scores. The

[39] Information about Stan is available at `http://mc-stan.org/`.

[40] For example, the 2000 election was fit using elections from 1980 to 1998; the 2004 election was fit using elections from 1980 to 2000 (recall that the model excludes years ending in 2); and so on.

[41] Signature-filing deadlines in districts are all before September 1st.

[42] I also dropped the few races with third party candidates. These do not impact the calculates of Q_i^{maj} because they all occurred before 2000. Independents Bernie Sanders of Vermont and Virgil Goode of Virginia are classified as a Democrat and a Republican respectively.

[43] The distribution on the national error is somewhat of an exception because it is somewhat right-skewed. This likely occurs because the number of national elections is small and the the error terms are truncated at zero.

Table 1.3: Posterior Inferences for Parameters in Forecasting Model, House Elections 1980 - 2008

		Posterior quantiles		
Variable	Mean	2.5%	Median	97.5%
Coefficients				
Intercept	0.343	0.196	0.344	0.489
Dem. share of district vote in last election	0.418	0.394	0.418	0.442
Relative Dem. share of presidential vote in last election	0.337	0.313	0.337	0.360
Incumbent	0.094	0.089	0.094	0.099
Relative 1st dimension DW-nominate score	0.060	0.049	0.060	0.070
Freshman incumbent	0.012	0.006	0.012	0.018
Previous office holder	0.032	0.028	0.032	0.037
August presidential net approval rating	0.061	0.026	0.061	0.095
August generic ballot	−0.079	−0.355	−0.080	0.199
Midterm election	−0.044	−0.059	−0.044	−0.029
Variance terms				
District error (σ)	0.061	0.060	0.061	0.062
National swing (σ_δ)	0.013	0.007	0.012	0.026

sign on the freshman incumbent variable may be somewhat of a surprise though, as less experienced incumbents are actually predicted to do better than their more experienced counterparts. Finally, the estimated incumbency effect is consistent with previous research, although it is a little closer to typical estimates of the incumbency effect in recent years (of around 10%) than the more involved method used by Gelman and Huang (2008) (which found the incumbency effect to be closer to 8%).[44]

Although parameters have only been reported for the model using data from the 1980 to 2008 elections, the model was actually fit 5 separate times in order to imitate 5 separate real time forecasts using data that would have been available to a political analyst in August *prior* to each election. Each model has its own unique posterior distribution for the parameters. In practice, the district parameters do not vary much from one election to the next but that national parameters (which are estimated using much less data) become more precise over time.

In the Bayesian context, the forecast for an election in year $\tilde{y}$ is a probability distribution

[44] Gelman and Huang (2008) estimate candidate level incumbency effects using a multilevel model. They also include a model for the party of the incumbent which depends on the district vote lagged two periods.

over $v_{i\tilde{y}}$, called a posterior predictive distribution, which uses data from election years $y < \tilde{y}$ to form a distribution for the unobserved (to the political analyst at the time) election in year $\tilde{y}$. Given the persistent nature of the data, it is not surprising that the model is able to forecast elections quite accurately. Indeed, if the forecast, f_{iy}, is the mean of the posterior predictive distribution of v_{iy}, then the root mean square forecast error (RMSFE), $\sqrt{E[(f_{iy} - v_{iy})^2]}$, is 0.056 and the winner is only picked incorrectly 7% of the time.

There is also evidence that the decomposition of the error into a district component and a national component captures uncertainty in elections quite well. To see this, consider Figure 1.3, which is a histogram of the posterior predictive distributions of the fraction of seats won by Democrats in each of the five non-redistricting year elections from 2000 to 2010.[45] The posterior predictive distribution for the number of seats is formed by simply counting the number of districts that had vote shares over 0.5 for each of the $5,000$ draws from the posterior predictive distributions of $v_{i\tilde{y}}$ for each forecast year $\tilde{y}$. The actual fraction of Democratic seats is usually within the 95% credible interval, but the intervals are not unreasonably wide either.[46] The 2010 election is a bit of an outlier since nearly all of the close elections broke toward the Republicans and large number of Democratic incumbents were defeated by Republican challengers.[47] The next part of this paper will examine whether polling data can help improve this forecast.

1.6.2 Results from the DLM

Although the DLM allows the states to vary on a daily basis, days are divided into two week periods in order to reduce the computational burden.[48] When a district (in the district

[45] Uncontested seats are not included in these percentages.

[46] If there was no national shock these 95% credibility intervals would be far to narrow because all of the errors would be independent.

[47] The majority party typically holds an electoral advantage because they tend to win a greater number of seats than votes (Tufte 1973). This Democrat's did not enjoy this partisan bias in 2010.

[48] For the district model using daily data, a separate parameter must be estimated for each district each day. For instance, if one started using polling data in July, then (since there are 174 districts with polling data), approximately $174 \times 90 = 15,660$ parameters (in addition to the variance parameter) would need to be estimated. This estimation would then need to be done each time a forecast was made so that if one wanted daily forecasts, $90 + 90 \times 15,660 = 1,409,490$ parameters would need to be estimated!

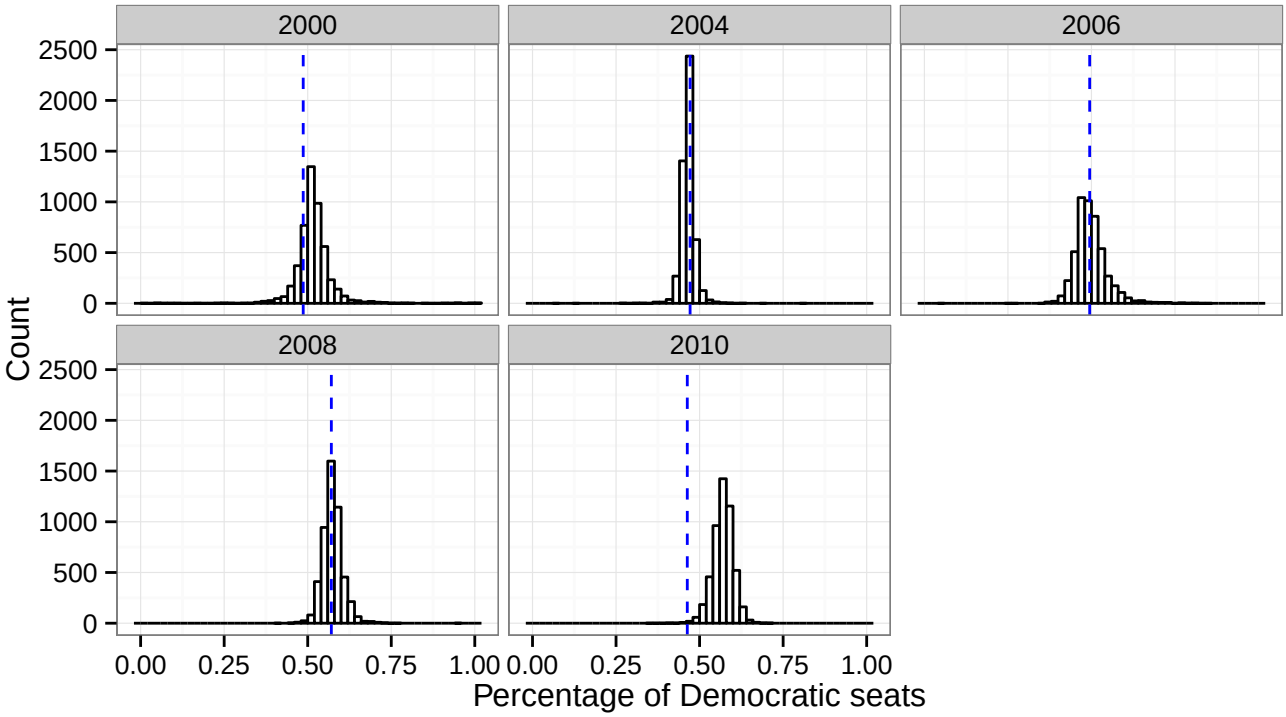


Figure 1.3: Forecasts of the Number of the Percentage of Democratic Seats

Notes: The histograms are from the posterior predictive distribution of the hierarchical model for each election year.

model) or polling firm (in the national model) has multiple polls during a period I take an average of the polls weighted by sample size. This average poll is then given a sample size equal to the sum of the sample sizes of the individual polls. Aggregating the polls in this manner does not have a substantial impact on the forecasts for a number of reasons. Firstly, academic research typically shows that frequent polling provides very little information (e.g. Gelman and King 1993; Lock and Gelman 2010). This is likely to be especially true in a window as short as two weeks. Secondly, there are very few cases in which a particular district was polled more than once in any two week span.

To remain relatively agnostic about biases, I set the prior variance for the House effects σ_λ^2, equal to 0.05^2. This allows for very biased polling firms because plus or minus 2 standard deviations from 0 is plus or minus 10 percentage points. In addition, the inverse gamma

priors for the unknown variance parameters, σ_μ^2 and σ_ξ^2, were assigned shape parameters equal to 9 and scale parameters equal to 0.02, corresponding to a mean of 0.05^2. The prior implies that period to period movements should be around 5 percentage points, but its effect is in practice dwarfed by the data. Election day polls were added to both the national and district models. The final poll in the national model is given a mean and standard deviation of 0.514 and 0.019 respectively, which are the mean and standard deviation of the posterior predictive distribution of the average district vote from the hierarchical model.[49] Similarly, the mean and standard deviation of the election day poll for the district model are the mean and standard deviation of the posterior predictive distribution of the difference between the vote in each district and the average district vote. For reference, the mean of this standard deviation is 0.061, with is nearly identical to the mean of the district level error, σ.

The posterior density of both the national and district DLM's were simulated using $6,000$ iterations of the Gibbs Sampler with a burn-in of $1,000$ iterations. The model is (retrospectively) estimated every two weeks during the month's of September and October. Polling data after a desired forecasting date is removed to ensure that only data that would have been available at the time is used. This yields four separate forecasts of the 2010 election in addition to the prior from the hierarchical model: a mid September (1.5 months prior to election day), late-September (1 month prior to election day), mid-October (2 weeks prior to election day) and late-October (just before the election) forecast. Trace plots of the parameters suggest that the Markov Chain converged each time and are available upon request.

Estimates of the square root of the variance parameters, σ_μ and σ_ξ are shown in Table 1.4. The table reports estimates listed by the date the model was estimated. The posterior quantiles for σ_μ are consistent across dates and yield 95% credible intervals ranging from

[49] Uncontested seats are assumed to be known with certainty. Uncontested Democratic (Republican) seats are given a vote share of 0.75 (0.25). These values are based on those reported in King and Gelman (1991) and Gelman and King (1994) and derived from vote shares received by districts in the last year before they became uncontested and the first election after they became uncontested. For another paper that uses this strategy see Kastellec, Gelman and Chandler (2008).

around 0.015 to 0.023 with a median of 0.018. Posterior estimates of σ_ξ are more variable but are generally pretty stable; the estimates do however increase slightly as election day nears.

Table 1.4: Posterior Quantiles for Square Root of DLM Variance Paramters

	National model (σ_μ)			District model (σ_ξ)		
Months to election	2.5%	Median	97.5%	2.5%	Median	97.5%
0.0	0.015	0.018	0.023	0.020	0.022	0.024
0.5	0.015	0.018	0.023	0.019	0.022	0.024
1.0	0.015	0.018	0.022	0.018	0.021	0.023
1.5	0.015	0.018	0.023	0.017	0.019	0.022

Notes: Months to election refers to the date that the model was estimated.

Summaries of model's forecasts are are presented in figures 1.4 and 1.5. Figure 1.4 focuses on the forecast of the national vote. The dark dotted line is the DLM forecast (mean of the posterior distribution of μ_T) and the error bars are 95% credible intervals; from top to bottom, the light dotted lines are the regression based forecast from the hierarchical model, the actual average district vote and an average of polls from the latest available two-week period.[50] Forecasts are reported for the standard model that include the regression estimate as a final pseudo poll (the "prior" model) and an alternative specification that does not (the "no prior" model). Standard errors are considerably smaller for the "prior" forecast than the "no prior" forecast although the difference shrinks as election day gets closer because there is less and less temporal error in the polls. Due to this temporal error, the forecast of the "prior" model does not deviate from the regression forecast until mid-October, but as temporal error decreases more weight is placed on the polls and the forecast is eventually nearly identical to the actual average district vote. On the other hand, the "no prior", or polls only model, hugs the polling average closely and predicts that the national vote will be considerably more Republican. The "no prior" model lies slightly above the polling average though because the model accounts for House effects (see Figure 1.C.4 in the appendix) and

[50]The actual average district vote includes uncontested seats. As noted in the footnote above, these calculations are made by imputing the vote share in uncontested Democratic seats to 0.75 and uncontested Republican seats to 0.25.

the most prolific pollsters (Rasmussen Reports and Gallup) leaned Republican during the campaign.

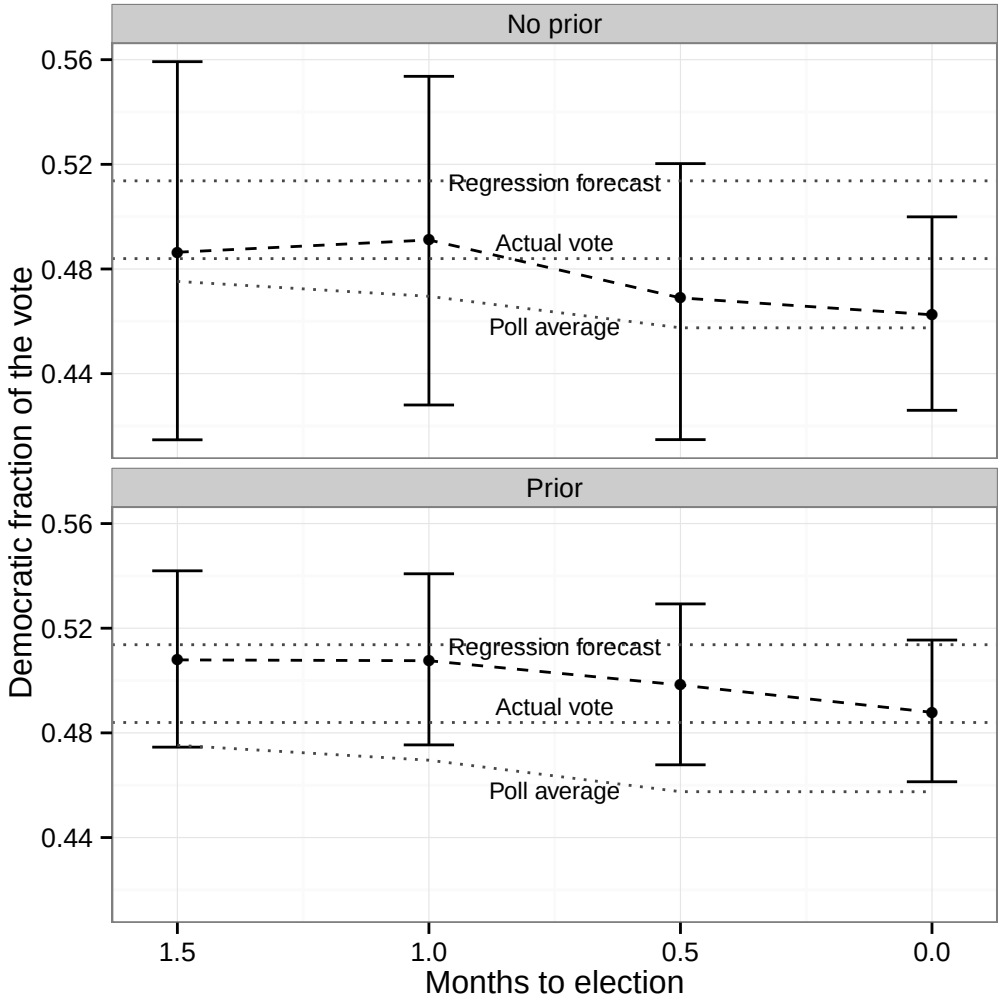


Figure 1.4: Forecasts of the 2010 National Vote by Date

Notes: The plot labeled "No prior" plots results from the DLM that does not include a final national pseudo poll based on the regression analysis; the plot labeled "Prior" includes this final pseudo poll. Vertical bars are 95% credible intervals for the forecast of the national vote from the DLM (points are median of the simulated vote). The upper dotted line is the point estimate of the regression based forecast (the mean of posterior distribution of the average district vote from the hierarchical model). The middle dotted line is the actual average district vote for Democrats. The lower dotted line is an average of all national polls during each time period (i.e. the poll average at the 2 month mark is all polls from days between 2 and 1.5 months before the election).

Figure 1.5 displays forecasts of the fraction of seats won by the Democrats after excluding uncontested seats. The upper light dotted line labeled "regression forecast" is identical to the

mean of the histogram for the 2010 election in Figure 1.3 and the reported fraction of actual seats won excludes uncontested seats as well. The forecast (and 95% credible intervals) are calculated by taking the fraction of seats with a predicted vote share greater than 0.5 for each simulated draw of $\mu_T + \xi_{iT}$.[51] Like the forecast of the national vote, the "prior" model does not begin to move away from the regression forecast until mid-October and the "no prior" forecast changes considerably more from period to period. The "no prior" model provides a forecast of the Democratic seat share that is very close to the actual Democratic seat share while the "prior" model overestimates the electoral success of the Democrats.

It is tempting to use this evidence to discount the regression based prior but it is worth remembering that i) 2010 is an outlier election, ii) nearly all of the close elections broke toward the Republicans and iii) comparing the average district vote to the actual district vote is only one of many possible ways to evaluate forecasts. Two alternative evaluation measures are shown in Figure 1.6.

The first summary measure is the standard root mean square forecast error (RMSFE) described earlier in the paper. Here the standard DLM (with a prior) does considerably better than DLM with no prior. Both however have considerably lower RMSFE's than the regression based forecast from the hierarchical model. Strangely, the RMSFE's do not decrease during the campaign.

The second summary measure is the Brier score, which is often used to evaluate the accuracy of binary probabilistic forecasts. For this paper, the Brier score at time t can be written as $BS_t = \frac{1}{408}\sum_{i=1}^{408}(f_{it} - a_{iT})^2$, where 408 is the number of contested House seats; f_{it} is the forecasted probability that the Democratic candidate would win district i at time t, and a_T is a binary variable equal to 1 if the Democratic candidate won the actual seat and 0 if the Republican candidate won. Higher Brier scores indicate worse forecasting performance. As shown in the figure, the Brier scores for both DLM's decline in a linear fashion over time

[51] Recall that the states, ξ_{it} are estimated from district polls relative to national opinion at time t and that national opinion is, in turn, estimated as the mean of the posterior distribution of μ_t. The posterior mean of μ_t from the model estimated for late-October is shown in Figure 1.C.3 in the appendix.

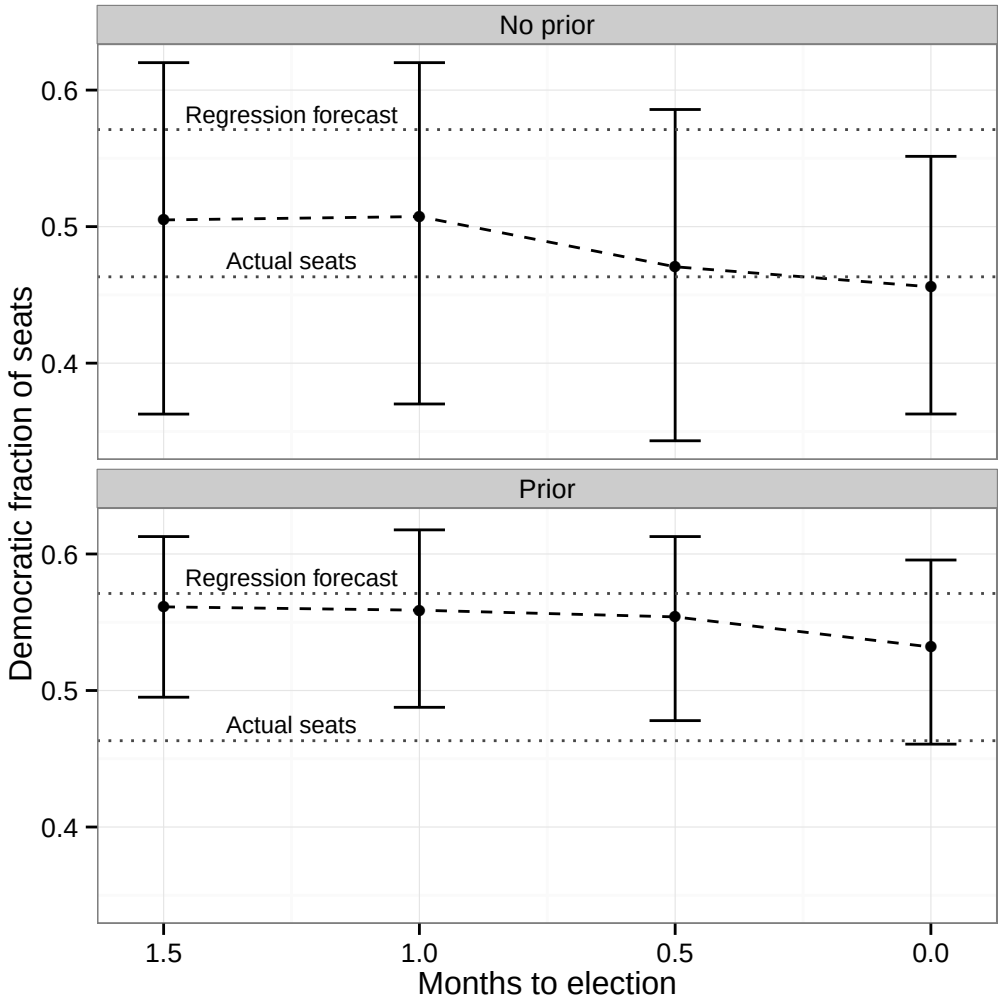


Figure 1.5: Forecasts of the Democratic Fraction of Seats in the 2010 Election

Notes: The plot labeled "No prior" plots results from the DLM that does not include a final pseudo polls based on the regression analysis; the plot labeled "Prior" includes these final pseudo polls. Vertical bars are 95% credible intervals for the forecast of the Democratic fraction of seats from the DLM (points are median of the simulated vote). The upper dotted line labeled "Regression forecast" is the mean of the posterior distribution of the Democratic seat share from the hierarchical model. The lower dotted line is the fraction of seats won by Democrats. All calculations omit uncontested seats.

and are considerably smaller than the Brier score of the hierarchical model. In contrast to the RMSFE, the Brier sore suggests that the DLM without a prior is a better predictor than the DLM with a prior.

The results suggest that forecasts with the DLM's are more accurate than forecast with the hierarchical model alone, but it is somewhat disconcerting that the RMSFEs increase as

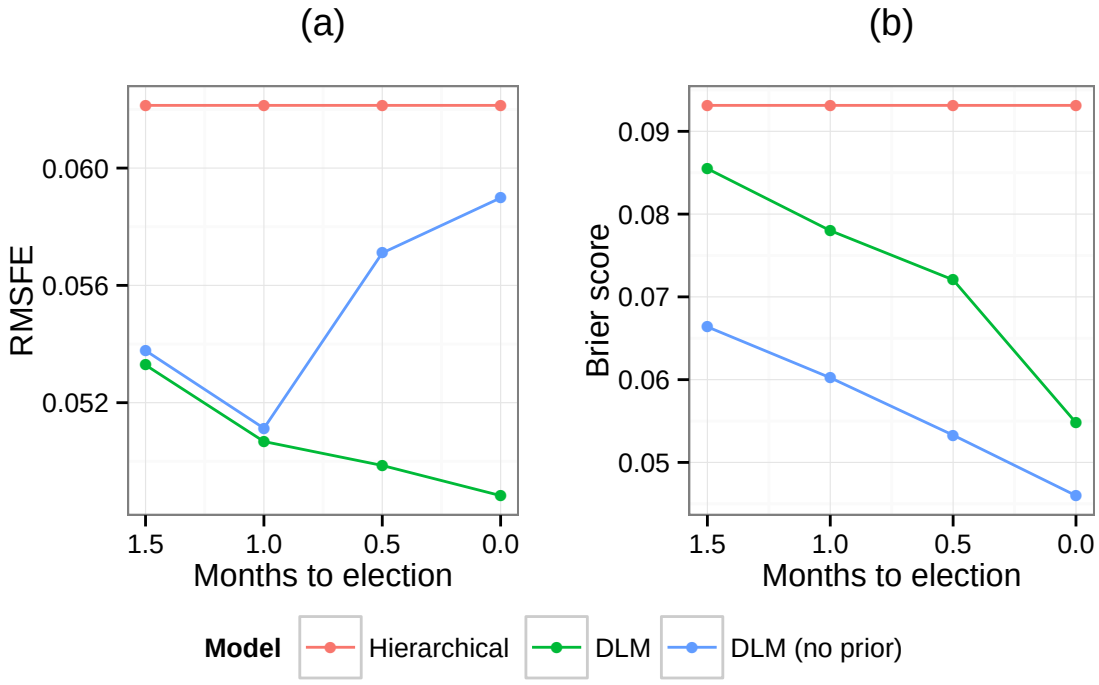


Figure 1.6: Evaluations of 2010 Election Forecasts

Notes: Figures (a) and (b) plot the root mean square forecast errors and brier scores respectively for the different forecast models by forecast date.

the election day gets closer. Figure 1.C.5 in the appendix shows that this is being caused by districts with no polling data: the RMSFEs decrease (increase) over time in districts with (without) at least one poll. The poor performance of the RMSFEs in districts without polls is an artifact of the downward trend in the forecast of the national Democratic vote (see Figure 1.4). This is illustrated in appendix Figure 1.C.6, which plots the mean prediction error (MPE)—defined as $E[f_{it} - v_{iT}]$ where v_{iT} is the observed vote on election day—by forecast date. The plot shows that forecasts become increasingly biased toward Republicans and that the bias is only severe in districts without polls (since the poll based forecasts in districts without polls are driven entirely by the biased national polls). The bias increases the overall RMSFE but not the overall Brier score because the Brier score is more sensitive to changes in forecasts of close elections (which are the most likely to have polling data) than lopsided ones while the RMSFE is equally sensitive to changes in both.

1.6.3 Calculating Q_i

Armed with parameter estimates from the hierarchical model and the DLM, one can use the estimates of V_i, σ and σ_δ to calculate the Q_i's from equations 1.12 and 1.13. In the theoretical model, these parameters are assumed to be known with certainty. But, when using a Bayesian approach this is not the case since the parameters have their own probability distributions.

Accounting for this additional uncertainty with the hierarchical model is straightforward: I simply calculate Q_i by averaging over all of the unknown parameters. In other words, I add extra uncertainty to the parties maximization problems by integrating the objective functions over $p(\beta, \delta_t, \sigma, \sigma_\delta)$ instead of just integrating over $h(\delta_t)$.

Q_i is calculated in a similar manner when the model is estimated using the DLM. Because I separate the national vote and the district vote relative to the national vote, there is a probability distribution for each component. Due to the normality assumptions, both the forecasts of the national vote, μ_T, and the district vote, ξ_{iT} are normally distributed. σ can consequently be estimated with the standard deviation of the posterior distribution of ξ_T. A posterior distribution for $V_i + \delta$ is simulated by summing the mean of the posterior distribution of ξ_T and each simulated draw from the posterior distribution of μ_T. In total, this provides a range of forecasts depending on the national error as well as an estimate of the district level error as required by the theory.

As discussed in Section 1.3.3, party goals can have an important impact on Q_i. The implication of these goals are illustrated in Figure 1.7, which shows how Q_i depends on forecasted Democratic vote shares. When parties want to maximize the expected number of seats, they should spend the most resources on the closest elections. This is clear in the leftmost plot in which Q_i follows a bell shaped pattern peaking when when the forecasted vote share is 0.5.

The strategies are more intricate when the parties maximize the probability of winning a majority of seats. In this case, the parties should spend the most resources on decisive swing districts, or districts that are most likely to be close when winning that district will

cause one party to win one more seat than the other party. This, in turn, implies that the trailing party should spend more resources in districts that they are losing in an effort to make the election more unpredictable. These incentives are best illustrated in the 2008 election when the Democratic party was expected to win a large majority of seats.[52] As shown in the plot, the Republican's optimal strategy (assuming they were only concerned with winning a majority of seats), was to to spend the most resources in districts with a predicted Democratic vote share over 0.5. Since the number of seats won by both parties would have only have been close if the national swing shifted drastically in the Republican's favor, Q_i is maximized for districts with a vote share close to 0.6. Interestingly, the cost of spending on districts that Republicans were expected to barely win has such an adverse effect on the variance of the election that Q_i is actually negative in these districts.

[52]Forecasts from the hierarchical model predict that the Republicans would have won, on average, 177 of the necessary 218 seats for a majority; in reality, they won 178 seats.

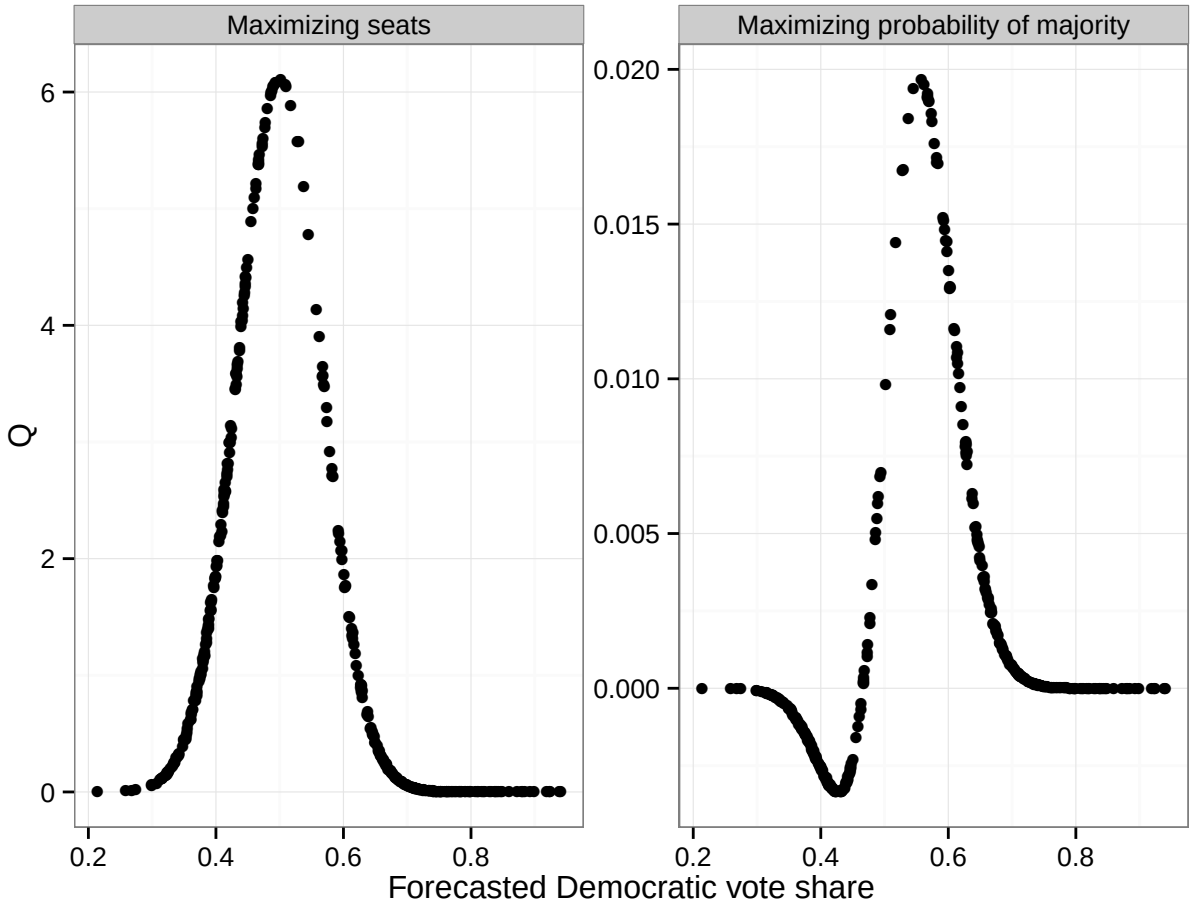


Figure 1.7: Q_i versus the mean forecast of the Democratic vote in the 2008 election

Notes: The forecasted Democratic vote share in each district is the mean of the posterior predictive distribution from the hierarchical model.

1.7 Relationship Between Q_i and District Spending

This section examines the empirical relationship between the estimates of Q_i and actual spending patterns by political parties, PACs and interest groups. The exact relationship between Q_i and spending depends on the functional form of $u(e_i^J)$. It would be preferable to estimate a utility function and its parameters using empirical estimates of the effect of spending on votes, but results from the empirical literature are very imprecise (see the discussion in Section 1.2). Instead, I use the logarithmic utility function analyzed in Section 1.3.4,

which yields an equilibrium in which $= e_i^{J*}/\sum e_i^{J*} = Q_i/\sum Q_i$. That is, I estimate the regression equation,

$$\frac{e_{iy}^J}{\sum e_{iy}^J} = \gamma \frac{Q_{iy}}{\sum Q_{iy}} + X\beta + \eta, \tag{1.28}$$

where e_{iy}^J is observed spending in district i and year y and X contains other covariates that affect district spending.

1.7.1 Correlations

Before moving to regression analyses, I examine simple correlations between the Q_i's and actual spending. Figure 1.8 plots the correlation between Q_i and spending by contributor type and the goal of the parties. The contributor types are broken down according to the party groups discussed in Section 1.5.2. The top and bottom panels show the correlation with spending between Q_i^{maj} and Q_i^{seats} respectively.

The plot highlights two primary aspects of the data that are worth discussing. First, there are few differences in spending patterns between contributor types. Surprisingly, the correlations for party affiliated groups are not significantly higher than the correlations for other groups. In addition, spending by individuals is actually the most correlated with the Q_i's among all groups.

Second, the correlations are not very sensitive to party goals except in 2008. In Section 1.6.3 I showed that assumptions about party goals impacted the values of Q_i greatly in 2008 because the Democrats were expected to win the House by almost 80 seats. The 2008 election therefore provides a natural experiment that can be used to identify the goals that the parties actually have. The huge dip in the correlation between spending and Q_i^{maj} but not between spending and Q_i^{seats} is telling. I view this as strong evidence that parties maximize the expected number of seats rather than the probability of winning a majority of seats. Of course, another possibility is that parties attempt to maximize seats when the election is lopsided but try to win a majority of seats when the election is closer. This is

plausible, but there is not enough evidence to support it because differences in correlations between the two goals are too small in election years other than 2008.[53]

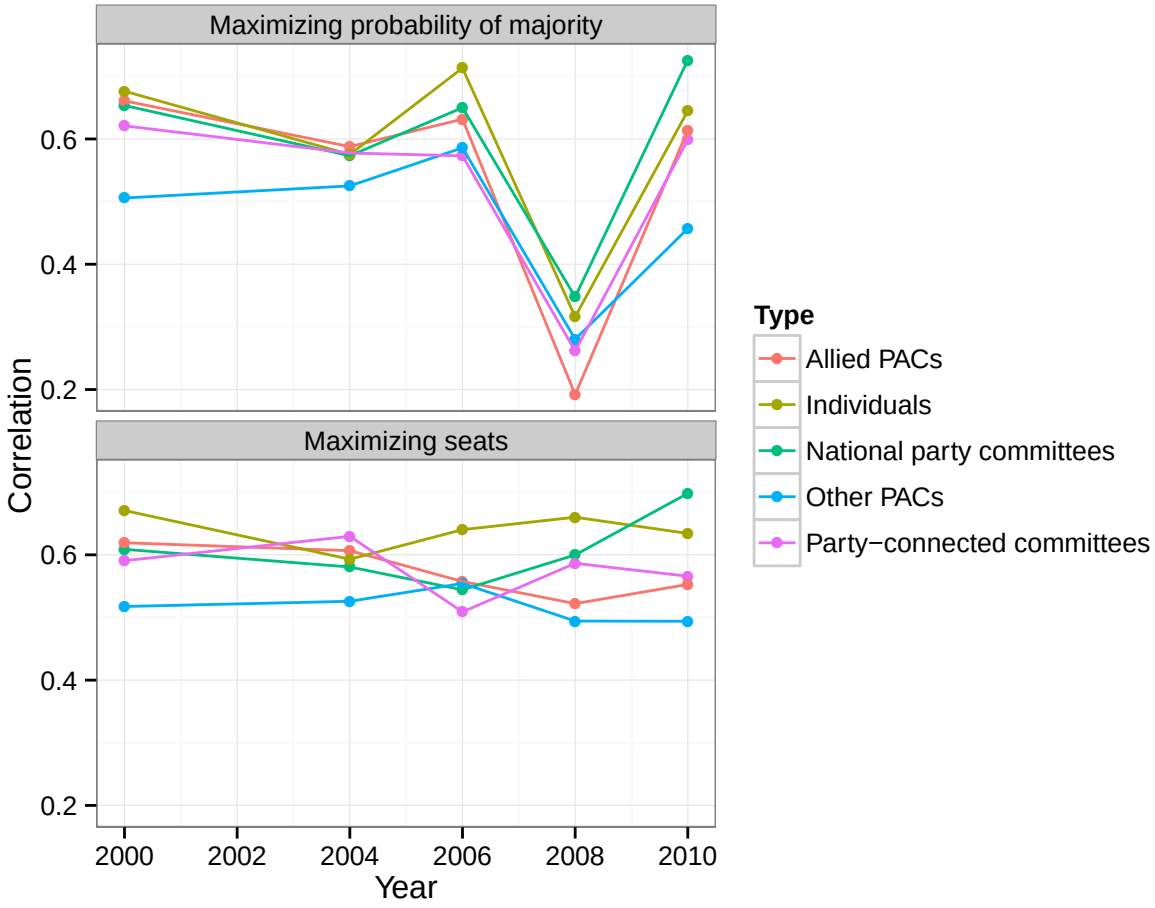


Figure 1.8: Correlation Between Q_i and Spending, by Contributor Type

Notes: Spending is the sum of spending for Republicans and Democratics in a district by a given contributor type. The top panel shows the correlation between Q_i^{maj} and spending while the bottom panel shows the correlation between Q_i^{seats} and spending.

Due to these results, I will now focus on total spending by PACs, party committees and individuals unless otherwise specified. According to Stromberg's model, if the effect of spending on vote share is of the log form, then both parties should spend the same proportion of their funds in each district. An examination of spending patterns by parties

[53]The correlation between actual spending and Q_i^{maj} is higher (by about 0.1) than the correlation between spending and Q_i^{seats} for all groups in 2000 and 2006. But in 2004 (when the predicted seat share is nearly identical to 2006) and 2010, the correlation with spending is actually slightly higher for Q_i^{seats} than Q_i^{maj}.

across districts is consistent with this notion. For instance, the simple correlation between spending on Democrats and spending on Republicans is 0.819. Moreover, Figure 1.C.8 in the appendix shows very little differences in the correlations between the Q_i's and spending by party, although Q_i^{seats} is slightly more correlated with spending on Republicans than spending on Democrats.

The finding that campaign contributions are higher in close elections is a common result in the political science and economics literatures. As mentioned earlier, one drawback of these studies is that they use the ex-post electoral margin as a measure of the closeness of the election and assume a linear relationship between this margin and spending.[54] Do the values of Q_i calculated using Stromberg's probabilistic voting model fit the data better?

Figure 1.9 takes a look at this question by plotting the correlation between various measure of electoral competitiveness. Based on Figure 1.8, I use Q_i^{seats} as the measure of electoral competitiveness derived from the probabilistic voting model. The electoral margin is measured as the absolute value of the Democratic vote share less 0.5.[55] Q_i^{seats} fits the data considerably better than the electoral margin for all five elections in the sample. There is also little difference between calculating electoral margin with the actual election outcome or a forecasted value, implying that Q_i^{seats} accounts for observed spending better because it depends on forecast uncertainty and that the normality assumptions on the error terms are reasonable.

The correlations presented thus far use the hierarchical model to calculate Q_i. One would expect that the DLM could improve the fit between the model and the observed data. Figure 1.10 examines this by comparing the correlation between Q_i^{seats} calculated using three different models and spending at different dates during the 2010 campaign.[56] Spending is the

[54]Some studies used the electoral margin from the previous election as a measure of closeness.

[55]Uncontested seats are treated as having an electoral margin equal to 0.5 to maintain consistency with the probabilistic voting model.

[56]The figure omits correlations made just prior to the election (late October) because the correlations are significantly lower and distort the figure: the correlations using the prior-informed DLM, no-prior DLM and hierarchical model are 0.422, 0.383 and 0.349 respectively. This likely occurs because there are only a couple of days between the final forecast date and election day.

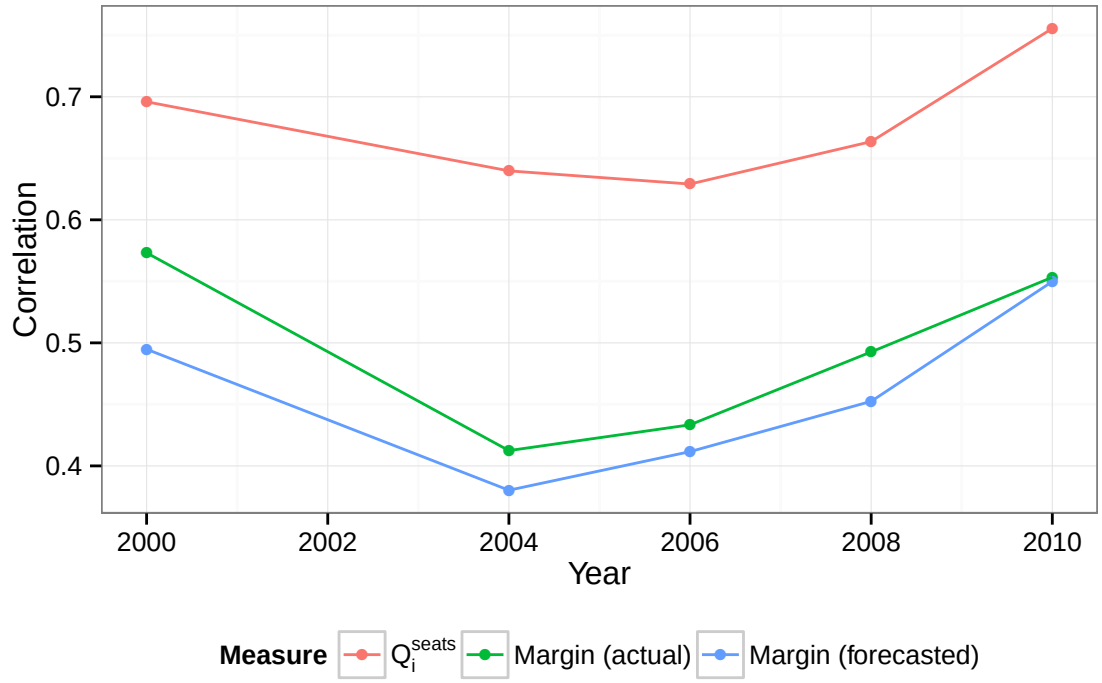


Figure 1.9: Correlation Between Measures of Electoral Competitiveness and Campaign Spending

Notes: Margin is the absolute value of the actual (forecasted) two-party vote share in a district minus 0.5. Campaign spending is the sum of all spending by PACs, party committees and individuals contributing over $200 in a given district and year.

sum of all spending in a district between election day and the the date the forecast is made. The three models are the hierarchical model, the prior informed DLM and the non-prior informed DLM.

In early September, or two months prior to the election, forecasts are only available from the hierarchical model so I assign all the models forecast values equal to the forecast from the hierarchical model. The figure shows that the Q_i^{seats} calculated using the DLM's match actual spending better than the Q_i^{seats} calculated using the hierarchical model, although the Q_i^{seats} from the hierarchical model are still pretty accurate. The correlations between the Q_i^{seats} estimated using the hierarchical model and spending decrease in a linear fashion over time while the correlations based on the DLM forecasts reach a peak of around 0.8 one month before the election. There is little difference between the performance of the "prior"

and "no prior" DLM's, although the correlations between the "no prior" DLM and spending are somewhat more variable. Overall, this figure suggests that campaign donors use polls to evaluate the competitiveness of a district and update these beliefs when new polls become available.

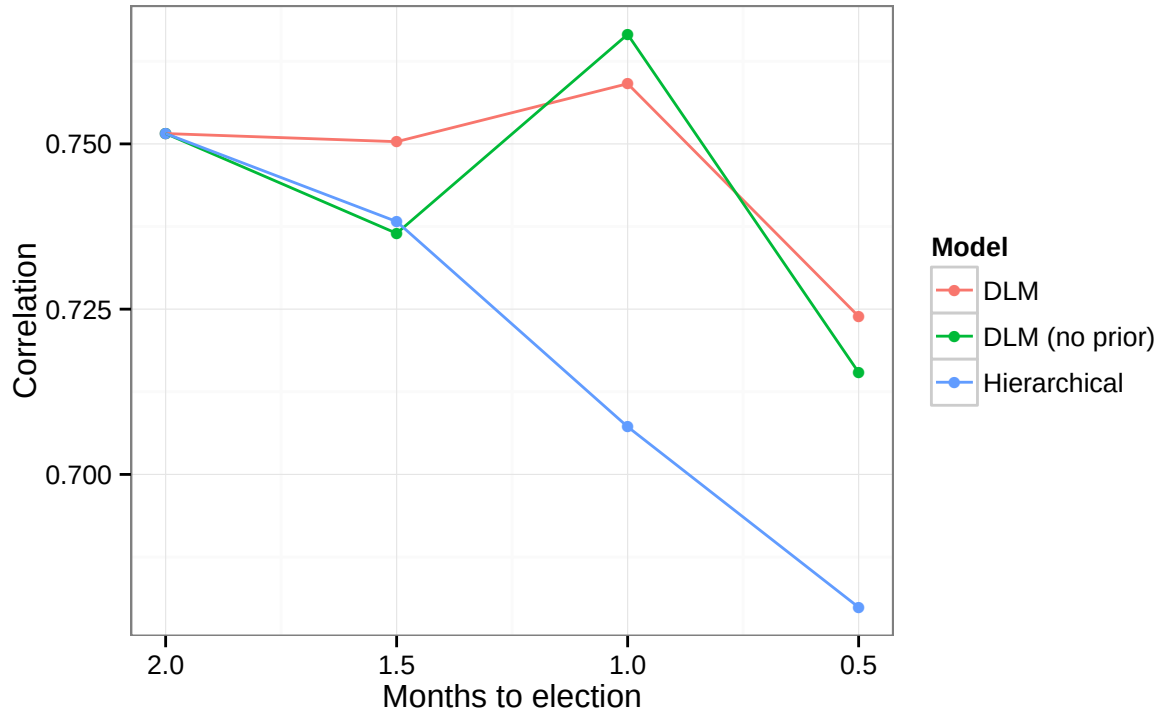


Figure 1.10: Correlation Between Q_i and PAC Spending by Date and Model, 2010 House Election

Notes: Spending refers to all spending by PACs, party committees and individuals contributing over $200 between a given date and election day.

1.7.2 Alternative Predictors of District Spending

I now move beyond simple correlations to analyses using regression equation 1.28. Table 1.5 reports regression estimates of γ and β when analyzing spending on Democratic candidates (panel A) and Republican candidates (panel B) separately. Equilibrium spending, $Q_{iy}^{seats} / \sum_i Q_{iy}^{seats}$, is estimated using the hierarchical model. Spending shares and equilibrium spending are multiplied by 100 so that they can be interpreted in percentage terms.

Mean and median spending shares for Democratic and Republican candidates are 0.23% and 0.085% respectively for both parties.

The first column reports an estimate of a simple linear regression (with an intercept) with equilibrium spending as the only explanatory variable. The equilibrium spending coefficients in both panels A and B are strongly positively associated with actual spending and significantly different than 0. The coefficients are somewhat inconsistent with the equilibrium spending conditions in equation 1.15 though because they are significantly different than 1. That said, the coefficients are fairly close to 1 and the R^2 values are high.

The second column adds two dummy variables indicating that a candidate is an incumbent or running in an open seat (the omitted category is a challenger in an incumbent district). Each variable is statistically significant and positively associated with spending in both panels. The R^2 in column 2 improves considerably in the Democratic specification and marginally in the Republican one.

Columns 3 and 4 add influence-motivated variables that campaign donors might consider when deciding which campaigns to contribute to. The influence variables in column 3 are three indicator variables for whether candidates are member of the Ways and Means Committee, party leaders or committee chairmen. In column 4, I include an estimate of the probability that a candidate will win the election (0 to 1 scale) since campaign donors concerned with establishing a relationship with a candidate should prefer to donate to a candidate who will be in office (see for instance Snyder Jr 1990). The influence related variables have a small impact on the R^2 and are often not statistically different than zero. Being a party leader is the most important influence variable and is positive and statistically significant in both panel A and panel B.

To help alleviate concerns that the regression estimates might be biased, column 6 adds district fixed effect that control for time invariant district specific characteristics. The number of observations is smaller in this specification because the 2000 election is dropped since district lines changed in 2002. Adding the fixed effects improves the model fit and only

Table 1.5: OLS Regressions on Candidate Spending Shares (%)

	(1)	(2)	(3)	(4)	(5)
Panel A. Democrats					
Q share (maximizing seats)	0.753	0.752	0.752	0.750	0.618
	(0.042)	(0.041)	(0.042)	(0.041)	(0.054)
Open seat		0.216	0.216	0.100	0.133
		(0.048)	(0.048)	(0.063)	(0.071)
Incumbent		0.165	0.165	−0.101	−0.099
		(0.012)	(0.013)	(0.084)	(0.094)
Ways and Means Committee			−0.006	−0.012	0.010
			(0.018)	(0.017)	(0.039)
Party Leadership			0.077	0.071	0.069
			(0.023)	(0.022)	(0.015)
Committee Chair			−0.029	−0.033	−0.008
			(0.018)	(0.018)	(0.030)
Probability of Victory				0.295	0.240
				(0.085)	(0.132)
Observations	2,175	2,175	2,175	2,175	1,740
Adjusted R-squared	0.363	0.406	0.406	0.416	0.499
District fixed effects?	No	No	No	No	Yes
Panel B. Republicans					
Q share (maximizing seats)	0.824	0.771	0.773	0.773	0.660
	(0.044)	(0.043)	(0.043)	(0.043)	(0.054)
Open seat		0.161	0.160	0.164	0.156
		(0.045)	(0.045)	(0.058)	(0.064)
Incumbent		0.081	0.070	0.077	0.083
		(0.012)	(0.013)	(0.085)	(0.098)
Ways and Means Committee			0.089	0.089	−0.080
			(0.050)	(0.050)	(0.116)
Party Leadership			0.313	0.313	0.131
			(0.067)	(0.067)	(0.143)
Committee Chair			−0.012	−0.012	−0.115
			(0.041)	(0.041)	(0.075)
Probability of Victory				0.008	−0.082
				(0.087)	(0.137)
Observations	2,175	2,175	2,175	2,175	1,740
Adjusted R-squared	0.415	0.428	0.431	0.431	0.510
District fixed effects?	No	No	No	No	Yes

Notes: The dependent variable is the share of yearly spending by all PACs in each district. The unit of analysis is a candidate-district-year. Robust standard errors are in parentheses. *Open seat*, *Incumbent*, *Ways and Means Committee*, *Ways and Means Committee*, and *Party Leadership* are indicator variables equal to 1 if candidates are running in open seats, incumbents, members of the Ways and Means Committee, chairs of a major committee or House party leaders. Party leaders include the speaker of the House, the majority/minority leaders and the majority/minority whip. *Probability of victory* is the probability that a candidate wins the election calculated using the posterior predictive distribution from the hierarchical model.

has a small impact on the coefficients. Importantly, equilibrium spending remains a highly significant predictor of actual spending.

As a whole, the regression estimates reported in Table 1.6 show that equilibrium spending explains a large amount of the variation in actual spending and that the relationship between

equilibrium and actual spending is robust to district fixed effects. A few other variables such as whether the candidate is an incumbent or the seat is open help explain the data a little bit better (especially for Democrats) but little is gained in terms of fit from adding influence related variables to the regression.

While spending in a district may be largely explained by equilibrium spending, this ignores some important heterogeneity across campaign donors. Organizations whose welfare depends heavily on policy decisions should have the largest incentives to influence policy. Table 1.6 looks at whether this is the case by analyzing the spending decisions of the financial industry, a group whose campaign goals might differ markedly from other organizations. Campaign contributions from the financial industry are defined as those coming from PACs classified as a member of the Finance, Insurance, and Real Estate industry by the CRP or from individuals employed in the same industry. The regression analysis examines the impact that being a member of the House Committee on Financial Services Committee has on contributions from these firms. The unit of analysis for the regressions is an incumbent candidate since challengers cannot serve on a congressional committee. The dependent variable is the district share of yearly spending by the financial industry on incumbents.

Table 1.6: Spending by the Financial Industry, Incumbent Races 2000 - 2010

Variable	(1)	(2)	(3)
Q share (maximizing seats)	0.553	0.533	0.295
	(0.084)	(0.084)	(0.088)
On Financial Services Committee		0.249	0.201
		(0.048)	(0.082)
Chair/ranking minority member of Finance Committee		0.574	0.127
		(0.212)	(0.198)
Observations	1,992	1,992	1,592
Adjusted R-squared	0.098	0.137	0.327
District fixed effects?	No	No	Yes

Notes: The dependent variable is the district share of yearly spending by the Finance, Insurance, and Real Estate industry on incumbents. Industry spending covers contributions from PACs and individuals (sorted by employer) donating at least $200. The unit of analysis a district-year. Open seat districts are excluded from the analysis. Robust standard errors are in parentheses.

Column 1 shows that equilibrium spending is positive and statistically significant. Column 2 includes an indicator variables equal to 1 if an incumbent is a member of the Financial

Services Committee and 0 otherwise. Since the chair and ranking minority members of the Financial Services Committee are likely to hold the most influence, I also include a dummy variable indicating such status. Both variables are statistically significant and have large positive coefficients. In column 3, which controls for district fixed effects, equilibrium spending and committee status remain statistically different than zero but being a chair or ranking minority member does not. This should not be seen as an evidence that a leadership position on the committee is unimportant, but rather a consequence of a fixed effect regression without enough variation in an explanatory variable.[57] In summary, the financial industry appears to pursue both election motivated and influence motivated spending strategies. The overall fit of the financial industry model is considerably worse than the fit reported in the previous table, but this is likely a byproduct of focusing on a single industry.

1.8 Discussion and Summary

This paper uses Stromberg's 2008 probabilistic voting model to quantify the amount that political parties should spend on districts in House elections. The calculations are made using two assumptions about goals: first, parties maximize the expected number of seats and second, parties maximize the probability of winning a majority of seats. The model is estimated using two types of Bayesian forecasting models, one of which is able to update forecasts in real time as new polls become available.

The assumptions regarding the parties' goals have important implications for the optimal allocation of resources. Under the first assumption, parties should spend the most on close districts. On the other hand, when the second assumption is true, parties should spend the most on decisive swing districts; that is, districts that are close when they are pivotal in whether a party wins or loses a majority of seats. In both cases, spending should be more concentrated when the forecasts are more precise.

The empirical results support the first assumption but not the second one. For instance,

[57] Members of Congress usually hold committee leadership positions for long periods of time.

the correlation between district spending and the amount that should have been spent if parties were maximizing the expected number of seats ranges from 0.5 to 0.8 depending on the date, party of the candidate and forecasting method. Conversely, during the 2008 election in which optimal strategies differed greatly because the Democrats were predicted to win a huge majority, observed spending is highly correlated with a spending strategy based on maximizing the expected number of seats but not with with a spending strategy based on maximizing the probability of winning a majority of seats. That said, the results here cannot rule out the possibility that the parties maximize total seats in lopsided elections in order to increase the probability of majority control in the future.

The empirical results also suggest that most political giving is done to affect elections rather than for other reasons like protecting incumbents or influencing candidates. That said, some political actors, like the financial industry, place close to equal weight on influencing elections and gaining access to politicians.

One curious finding that warrants discussion is that the correlation between equilibrium spending under the model and actual spending is nearly identical for political parties, individuals, and PACs. There are two competing explanations for this. First, it could suggest that the national parties are able to coordinate campaign efforts with individuals and groups that have similar political goals. This explanation is consistent with Herrnson's 2009 view of political parties as multilayered coalitions.

However, a second explanation that cannot be ruled out is that candidates have larger incentives to raise funds in tight races. This interpretation is consistent with the game theoretic model in Erikson and Palfrey (2000). Intuitively, this occurs because an additional dollar of spending has a larger impact on the probability that a candidate will win an election when the projected vote share is closest to 1/2. There is some empirical support for this as well. For example, Stein and Bickers (1994) show that the most vulnerable candidates for reelection (as measured by their vote share in the previous election) are the most likely to obtain new grant money for their constituents. Similarly, Bickers and Stein (1996) find

that vulnerable incumbents use the grant system to deter quality challengers from opposing them.

The analyses in this paper suggest a number of fruitful areas for future research. First, researchers should try to separate the competing theories of why close elections receive more campaign contributions. Interviews and surveys of party leaders, interest group leaders, individuals, political consultants and political candidates could be especially useful here. It is of course possible, and perhaps likely, that some campaign donors, like political parties, have the party's collective interest in mind, while others, like individuals, are influenced more by the fundraising efforts of candidates. Second, it would be beneficial to separate campaign spending into separate categories. An analysis of TV and radio advertisements might be particularly interesting, although it would require more detailed information about advertising prices. These ideas could help provide a better understanding of the role that parties play in influencing elections.

Appendix

1.A Stan Code for Hierarchical Model

```
data {
  int<lower=0> N; // number of observations
  int<lower=0> K; // number of predictors
  int<lower=0> T; // number of years
  int<lower=1,upper=T> year[N]; // year for district
  matrix[N,K] x; // district predictors
  vector[N] y; // outcomes
}
parameters {
  vector[T] delta; // intercept by year
  vector[K] beta; // district coefs
  real<lower=0,upper=1> sigma_delta; // year error
  real<lower=0,upper =1> sigma; // district error
}
model {
  beta ~ normal(0, 1);
  delta ~ normal(0, sigma_delta);
  for (n in 1:N)
    y[n] ~ normal(delta[year[n]] + x[n] * beta, sigma);
```

}

1.B Gibbs Sampler for DLM

1.B.1 DLM Form

National Model

Let K_t be the numbers of polls published on day t and J be the number of polling firms. Then, equations 1.22 and 1.23 can be written in the form,

$$\underset{(K_t\times 1)}{Y_t} = \underset{(K_t\times 1)(1\times 1)}{F_t}\ \underset{}{\theta_t} + \underset{(K_t\times J)(J\times 1)}{A_t}\ \underset{}{\Lambda} + \underset{(K_t\times 1)}{v_t}\ ,\ v_t \sim N(0, V_t) \tag{1.29}$$

$$\underset{(1\times 1)}{\theta_t} = \underset{(1\times 1)}{G}\ \underset{(1\times 1)}{\theta_{t-1}} + \underset{(1\times 1)}{w_t}\ ,\ w_t \sim N(0, W), \tag{1.30}$$

where Y_t is a column vector of all polls published at time t, F_t is a column vector of 1's, A_t is a matrix of binary indicators equal to 1 if polling firm j published the k'th poll and 0 otherwise, $G = 1$, $\theta_t = \mu_t$, $\Lambda = [\lambda_1\ \lambda_2 \cdots \lambda_J]^T$, V_t is a $K_t \times K_t$ diagonal matrix with σ_k^2 along the diagonal and $W = \sigma_\mu^2$.

District Model

Let m be the number of districts with polling data and L_t be the numbers of polls published on day t. Then, equations 1.24 and 1.25 can be written in the form,

$$\underset{(L_t\times 1)}{Y_t} = \underset{(L_t\times m)(m\times 1)}{F_t}\ \underset{}{\theta_t} + \underset{(L_t\times 1)}{v_t}\ ,\ v_t \sim N(0, V_t) \tag{1.31}$$

$$\underset{(m\times 1)}{\theta_t} = \underset{(m\times m)}{G}\ \underset{(m\times 1)}{\theta_{t-1}} + \underset{(m\times 1)}{w_t}\ ,\ w_t \sim N(0, W), \tag{1.32}$$

where Y_t is a column vector of all district polls less national opinion at time t, F_t is a matrix of binary indicators equal to 1 if poll l is from district m, $G = I_m$, $\theta_t = [\xi_{1t}\ \xi_{2t} \cdots \xi_{mt}]^T$, V_t is an $L_t \times L_t$ diagonal matrix with σ_l^2 along the diagonal and $W = \sigma_\xi^2$.

1.B.2 Gibbs Sampler

National Model

Using the DLM form, the joint posterior density of the states and variance parameters is,

$$p(\theta_{1:T}, \sigma_\mu^2, \Lambda | Y_{1:T}, A_t, V_{1:T}) = \prod_{t=1}^{T} p(Y_t|\theta_t, \Lambda, A_t, V_t) \prod_{t=1}^{T} p(\theta_t|\theta_{t-1}, \sigma_\mu^2) \times p(\theta_0)p(\sigma_\mu^2)p(\Lambda). \tag{1.33}$$

The variance parameter, σ_μ^2, is given an inverse gamma prior, $p(\sigma_\mu^2) \sim \text{IG}(\alpha_\mu, \beta_\mu)$ and the house effects, Λ, are given independent normal priors, $p(\lambda_j) \sim N(0, \sigma_\lambda^2)$. Using these priors and recalling that $\theta_t = \mu_t$ and $G_t = 1$, the Gibbs sampler then proceeds as follows by drawing from the conditional distributions of the unknown parameters.

1. σ_μ^2. The conditional distribution is,

$$p(\sigma_\mu^2|\ldots) \propto p(\sigma_\mu^2) \prod_{t=1}^{T} p(\theta_t|\theta_{t-1}, \sigma_\mu^2) \tag{1.34}$$

$$\propto (\sigma_\mu^2)^{(\alpha_\mu - 1)} \exp\left[\frac{\beta_\mu}{\sigma_\mu^2}\right] (\sigma_\mu^2)^{T/2} \prod_{t=1}^{T} \exp\left[\frac{(\mu_t - \mu_{t-1})^2}{2\sigma_\mu^2}\right] \tag{1.35}$$

$$\propto \sigma_\mu^{2(\alpha_\mu + T/2 - 1)} \exp\left\{\frac{1}{\sigma_\mu^2}\left[\beta_\mu + \frac{1}{2}\sum_{t=1}^{T}(\mu_t - \mu_{t-1})^2\right]\right\}, \tag{1.36}$$

   which is an inverse gamma distribution,

$$p(\sigma_\mu^2|\ldots) \sim \mathcal{IG}\left(\alpha_\mu + T/2, \beta_\mu + \frac{1}{2}\sum_{t=1}^{T}(\mu_t - \mu_{t-1})^2\right). \tag{1.37}$$

2. λ_j. The conditional distributions for the House effects are normally distributed since both the prior and the likelihood are normal. Letting $\mathcal{P}_j$ refer to the set of polls published by firm j during the campaign and using standard results for conjugate

normal models, λ_j is normal with mean,

$$\left[\sum_{k\in\mathcal{P}_j}\frac{y_k-\mu_{t[k]}}{\sigma_k^2}\right]\left[\sum_{k\in\mathcal{P}_j}\frac{1}{\sigma_k^2}+\frac{1}{\sigma_\lambda^2}\right]^{-1} \tag{1.38}$$

and variance,

$$\left[\sum_{k\in\mathcal{P}_j}\frac{1}{\sigma_k^2}+\frac{1}{\sigma_\lambda^2}\right]^{-1}. \tag{1.39}$$

3. $\theta_{1:T}$. The states can be sampled from their conditional distribution,

$$p(\theta_{1:T}|\ldots)=p(\theta_T|Y_{1:T},A_t,\Lambda,\sigma_\mu^2)\prod_{t=0}^{T-1}p(\theta_t|\theta_{t+1},Y_{1:t},A_t,\Lambda,\sigma_\mu^2). \tag{1.40}$$

Equation 1.40 suggests that if we could sample θ_T then we could recursively backwards sample through the rest of the states, $\theta_{1:T-1}$. The *forward filtering backwards sampling* (FFBS) is the standard method for generating this sample. The filtering portion of the algorithm is the Kalman filter discussed in the text and works by starting from an initial value θ_0 and then sequentially updating each θ_t as new information becomes available. The backwards sampling part then smooths the values of the states by first sampling from θ_T, and then from the conditional distribution of θ_{T-1}, and so on all the way through all of the remaining states $\theta_{1:T-2}$. To summarize, the FFBS algorithm proceeds in the following manner:

(a) Run Kalman filter.

(b) Draw $\theta_T\sim N(m_T,C_T)$ where $m_T=E(\theta_T|Y_{1:T},A_t,\Lambda,\sigma_\mu^2)$ and $C_T=\mathrm{Var}(\theta_T|Y_{1:T},A_t,\Lambda,\sigma_\mu^2)$.

(c) For $T-1,\ldots,0$ draw $\theta_t\sim N(h_t,H_T)$ where $h_t=E(\theta_t|\theta_{t+1},Y_{1:t},A_t,\Lambda,\sigma_\mu^2)$ and $H_t=\mathrm{Var}(\theta_t|\theta_{t+1},Y_{1:t},A_t,\Lambda,\sigma_\mu^2)$.

In this case, letting $\mathcal{P}_t$ refers to the set of all polls published on day t we have,

$$m_t = \left[\sum_{k\in\mathcal{P}_t} \frac{y_k - \lambda_{j[k]}}{\sigma_k^2} + \frac{M_{t-1}}{C_{t-1}+\sigma_\mu^2}\right] \cdot C_T, \tag{1.41}$$

$$C_t = \left[\sum_{k\in\mathcal{P}_t} \frac{1}{\sigma_k^2} + \frac{1}{C_{t-1}+\sigma_\mu^2}\right]^{-1}, \tag{1.42}$$

and,

$$h_t = \left[\frac{\theta_{t+1}}{\sigma_\mu^2} + \frac{m_t}{C_t}\right] H_t, \tag{1.43}$$

$$H_t = \left[\frac{1}{\sigma_\mu^2} + \frac{1}{C_t}\right]^{-1}. \tag{1.44}$$

For additional details on the Kalman filter and the smoothing algorithm see a text such as West and Harrison (1997).

District Model

Inference for the district model proceeds in a similar manner to the national model using the DLM form from equations 1.31 and 1.32. The posterior density is,

$$p(\theta_{1:T}, \sigma_\xi^2 | Y_{1:t}) = \prod_{t=1}^{T} p(Y_t|\theta_t, V_t) \prod_{t=1}^{T} p(\theta_t|\theta_{t-1}, \sigma_\xi^2) p(\theta_0) p(\sigma_\xi^2). \tag{1.45}$$

Again using an inverse gamma distribution for the variance parameter, $\sigma_\xi^2 \sim \text{IG}(\alpha_\xi, \beta_\xi)$, the Gibbs sampler is as follows.

1. σ_ξ^2. Draw from the conditional distribution,

$$p(\sigma_\xi^2|\ldots) \propto p(\sigma_\xi^2) \prod_{t=1}^{T} p(\theta_t|\theta_{t-1}, \sigma_\xi^2) \tag{1.46}$$

$$\propto (\sigma_\xi^2)^{(\alpha_\xi - 1)} \exp\left[\frac{\beta_\xi}{\sigma_\xi^2}\right] (\sigma_\xi^2)^{Tm/2} \prod_{t=1}^{T}\prod_{i=1}^{m} \exp\left[\frac{(\xi_{it} - \xi_{i,t-1})^2}{2\sigma_\xi^2}\right] \tag{1.47}$$

$$\propto \sigma_\xi^{2(\alpha_\xi + Tm/2 - 1)} \exp\left\{ \frac{1}{\sigma_\xi^2} \left[\beta_\xi + \frac{1}{2} \sum_{t=1}^{T} \sum_{i=1}^{m} (\xi_{it} - \xi_{i,t-1})^2 \right] \right\}, \tag{1.48}$$

which is an inverse gamma distribution,

$$p(\sigma_\xi^2 | \ldots) \sim \mathcal{IG}\left(\alpha_\xi + Tm/2, \beta_\xi + \frac{1}{2} \sum_{t=1}^{T} \sum_{i=1}^{m} (\xi_{it} - \xi_{i,t-1})^2 \right). \tag{1.49}$$

2. $\theta_{1:T}$. The states ξ_{it} are drawn using the FFBS algorithm.

1.C Additional Figures

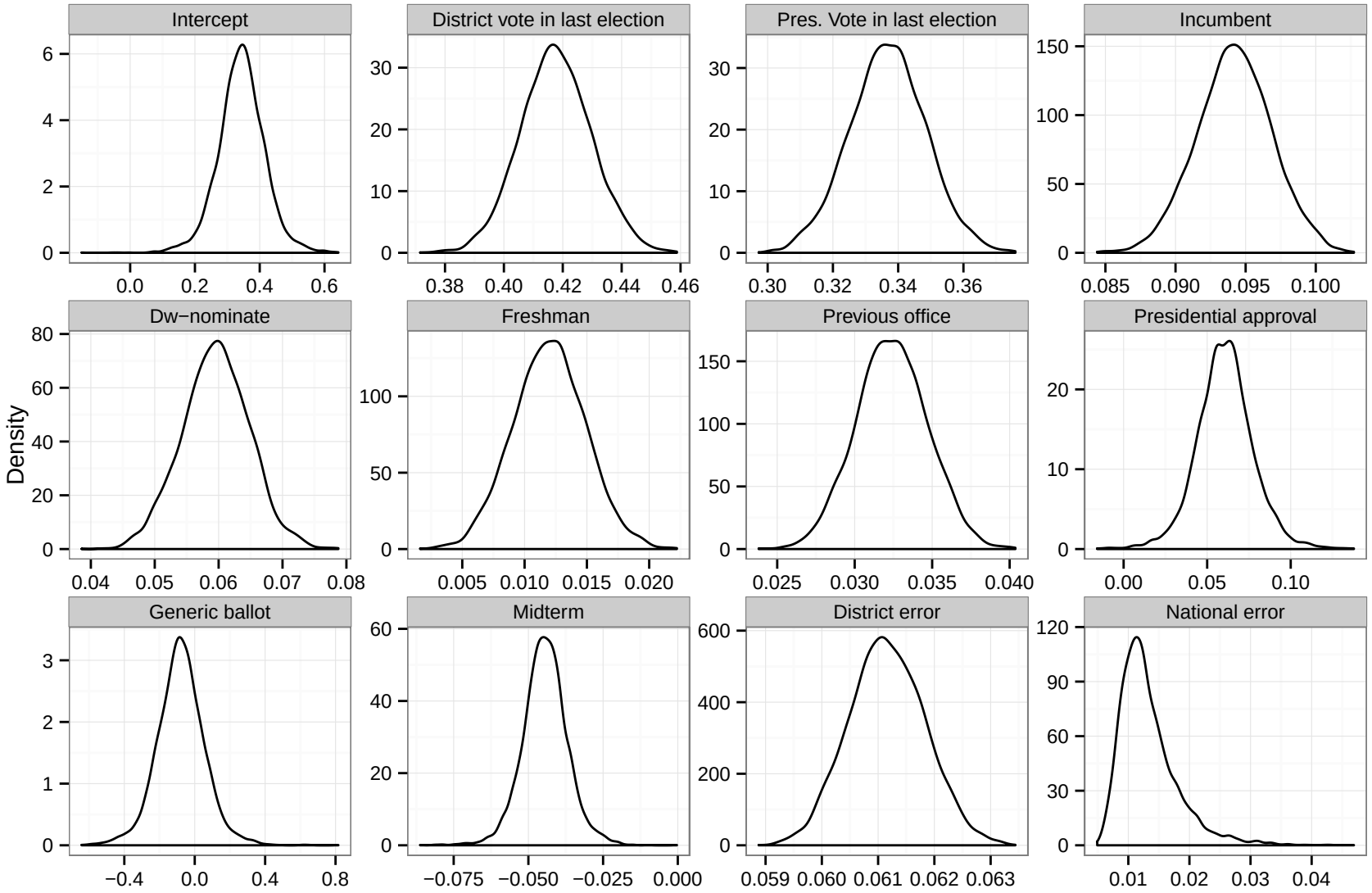


Figure 1.C.1: Posterior Densities for Parameters in Forecasting Model

Notes: Variables are the same as in Table 2.1 but names have been shortened for graphical purposes.

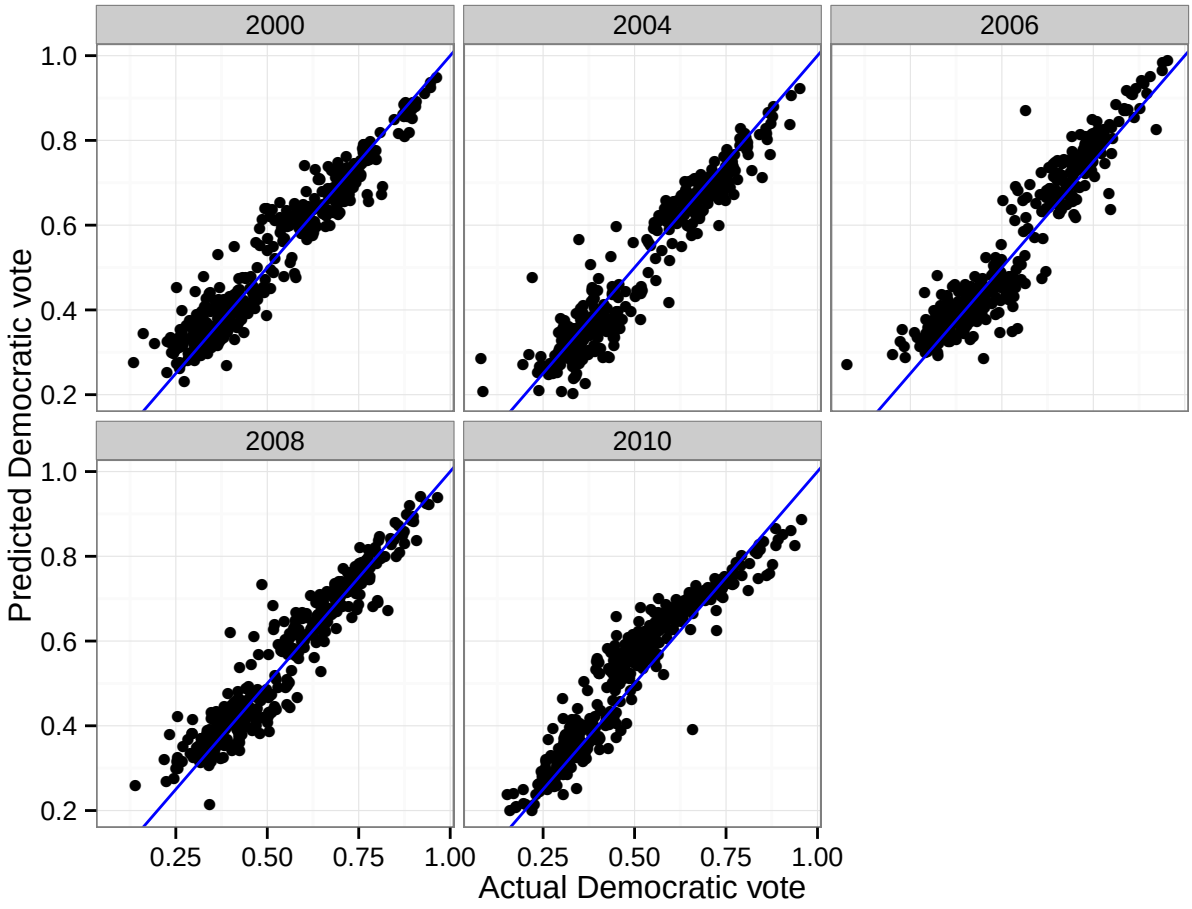


Figure 1.C.2: Out-of-Sample Predicted Versus Actual Plot, House Elections 1990 - 2010

Notes: Predicted value is the mean of the out-of-sample posterior predictive distribution. Model parameters are estimated using data from 1980 up until the year the prediction is being made (i.e. 1980 - 2008 when predicting the 2010 election). The blue line is a 45 degree line.

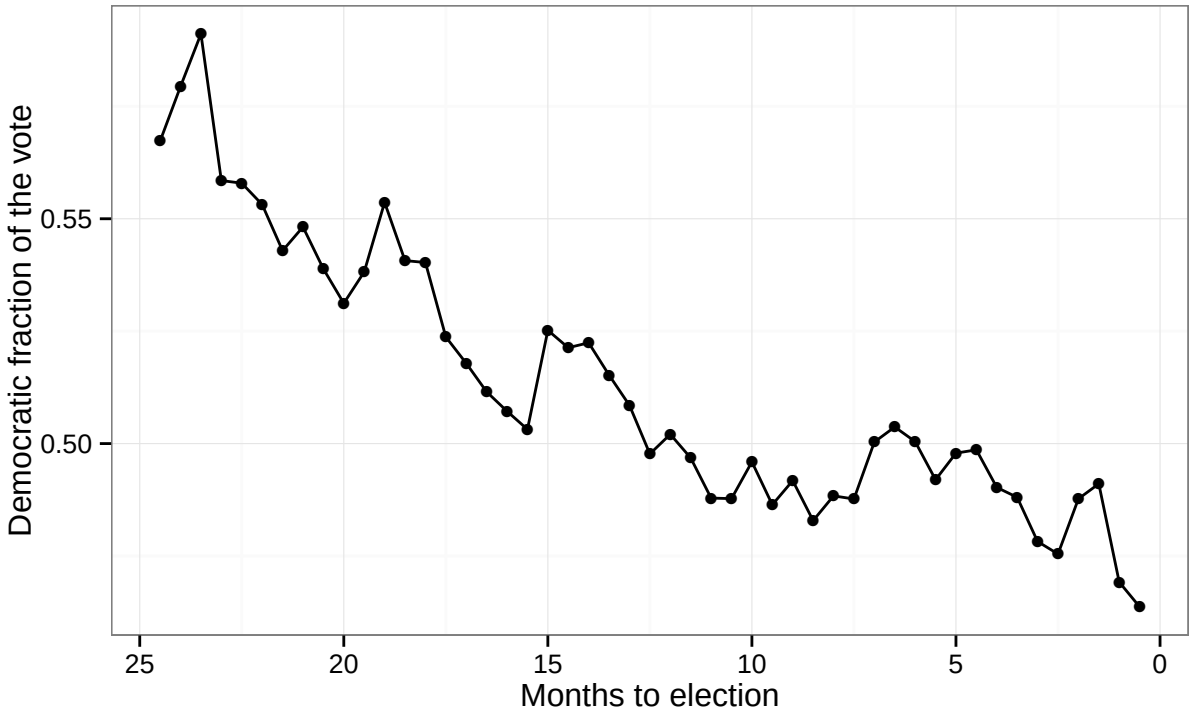


Figure 1.C.3: Estimates of National Opinion During the Election Campaign

Notes: Estimates are the mean of the posterior distribution of the states at each date from a DLM using 2010 generic ballot polling data up until election day.

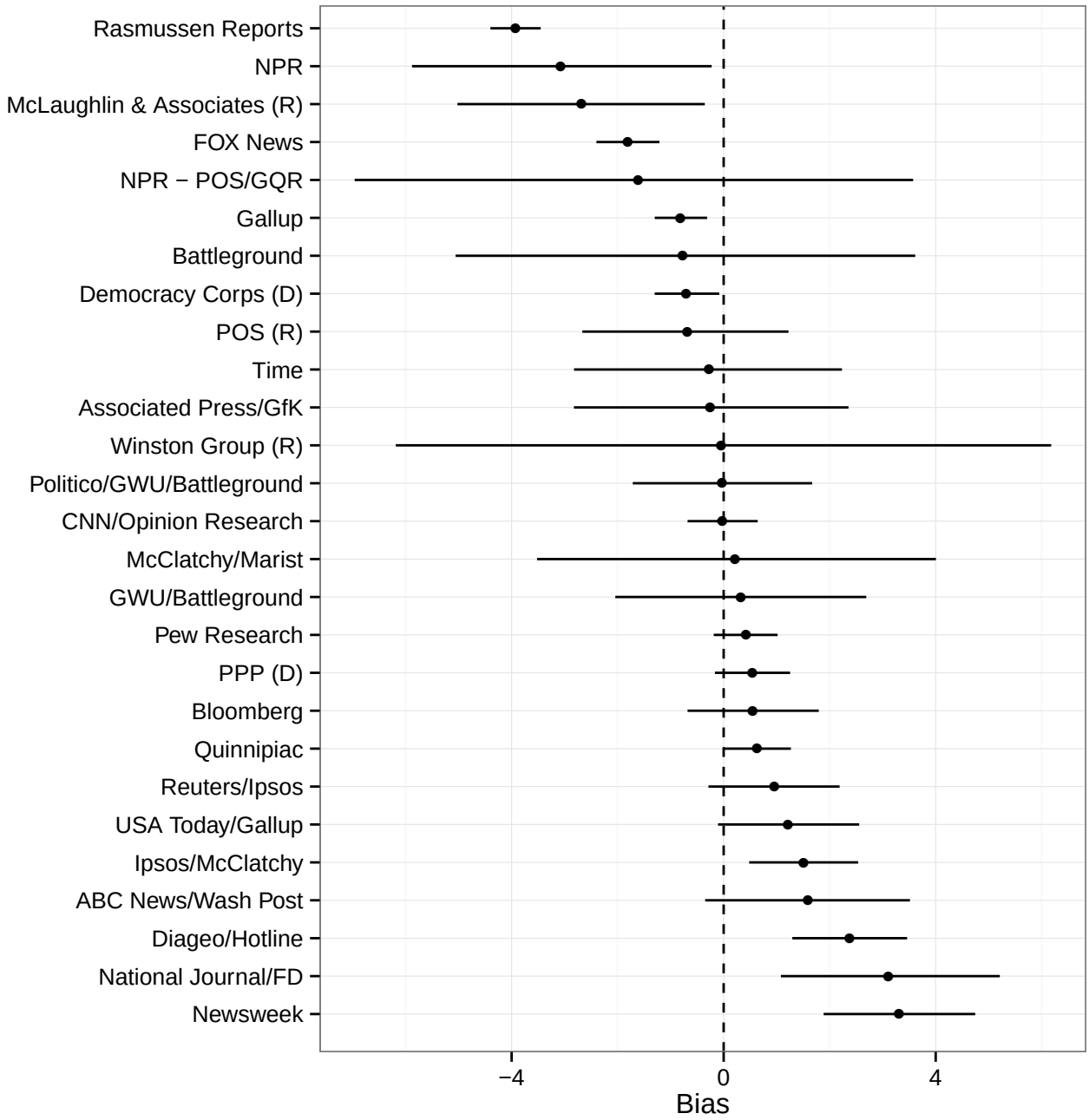


Figure 1.C.4: Estimates of House Effects

Notes: House effects are estimated using the DLM with 2010 generic ballot polling data up until election day. The bias across pollsters is assumed to sum to zero for identification purposes. The reported biases are from the late-October model just prior to election day.

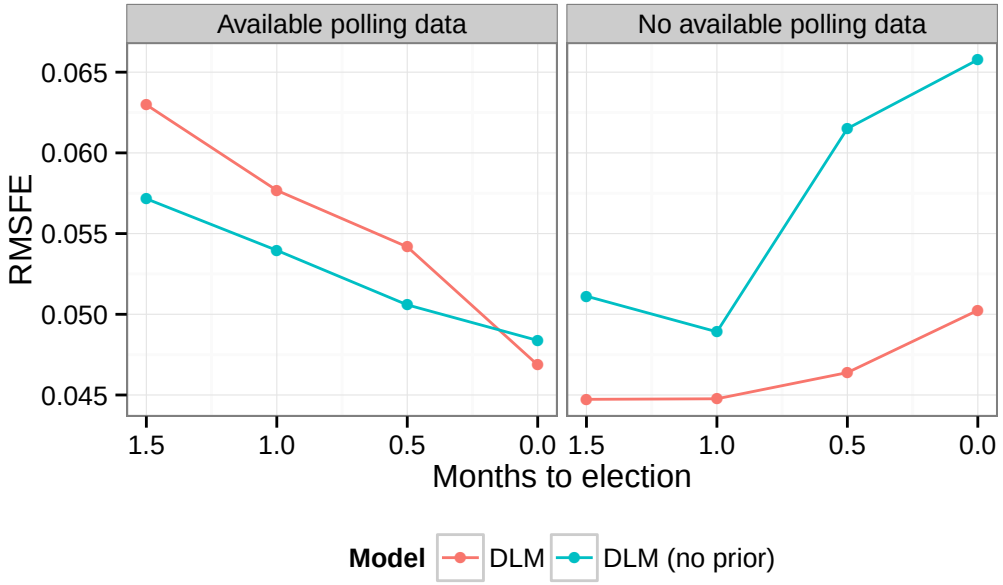


Figure 1.C.5: RMSFE By Polling Availability and Forecast Date

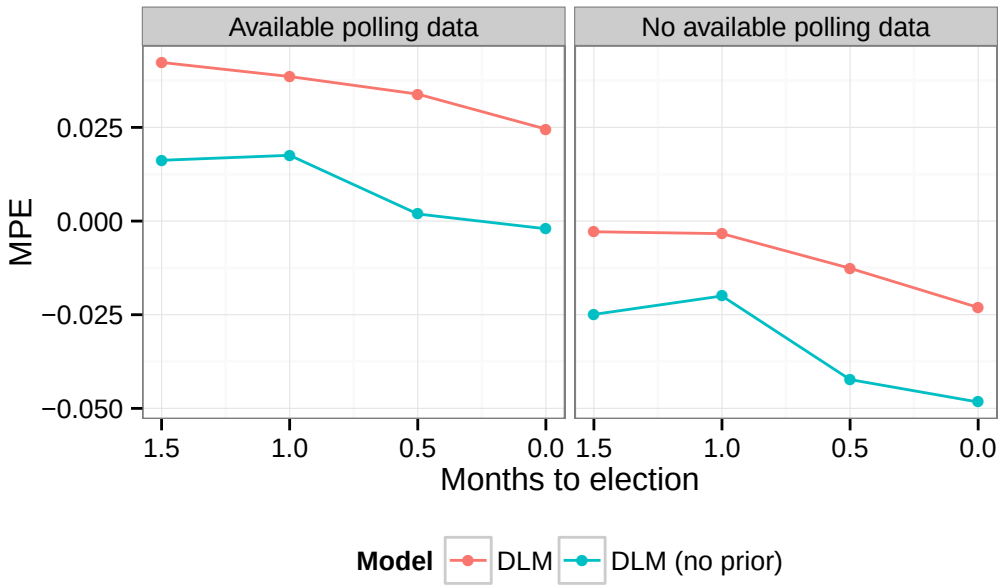


Figure 1.C.6: MPE By Polling Availability and Forecast Date

Notes: The MPE is the mean of the forecasted Democratic share of the two-party vote minus the Democratic share of the two-party vote in the actual election. Negative value reflect a bias toward the Republican party.

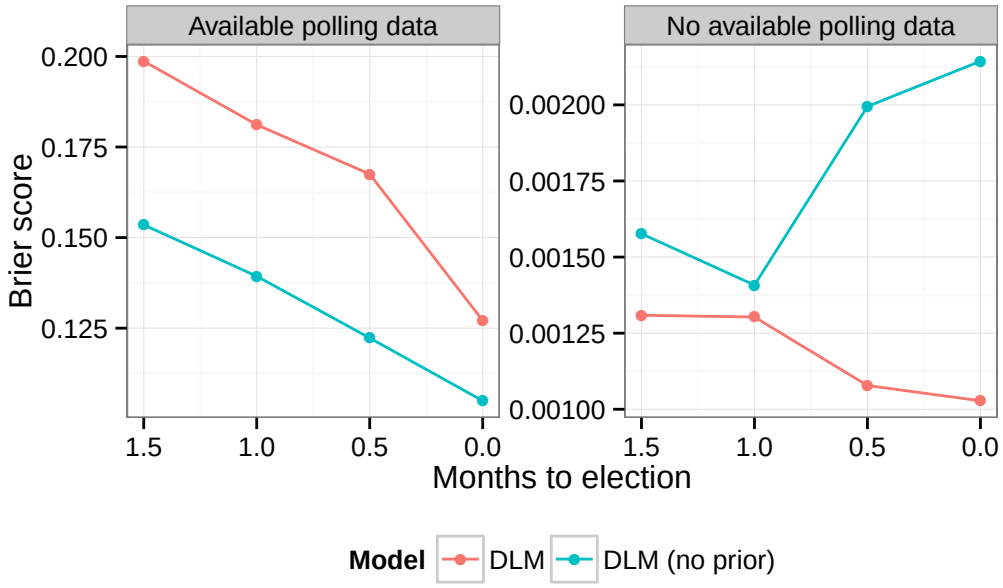


Figure 1.C.7: Brier Score By Polling Availability and Forecast Date

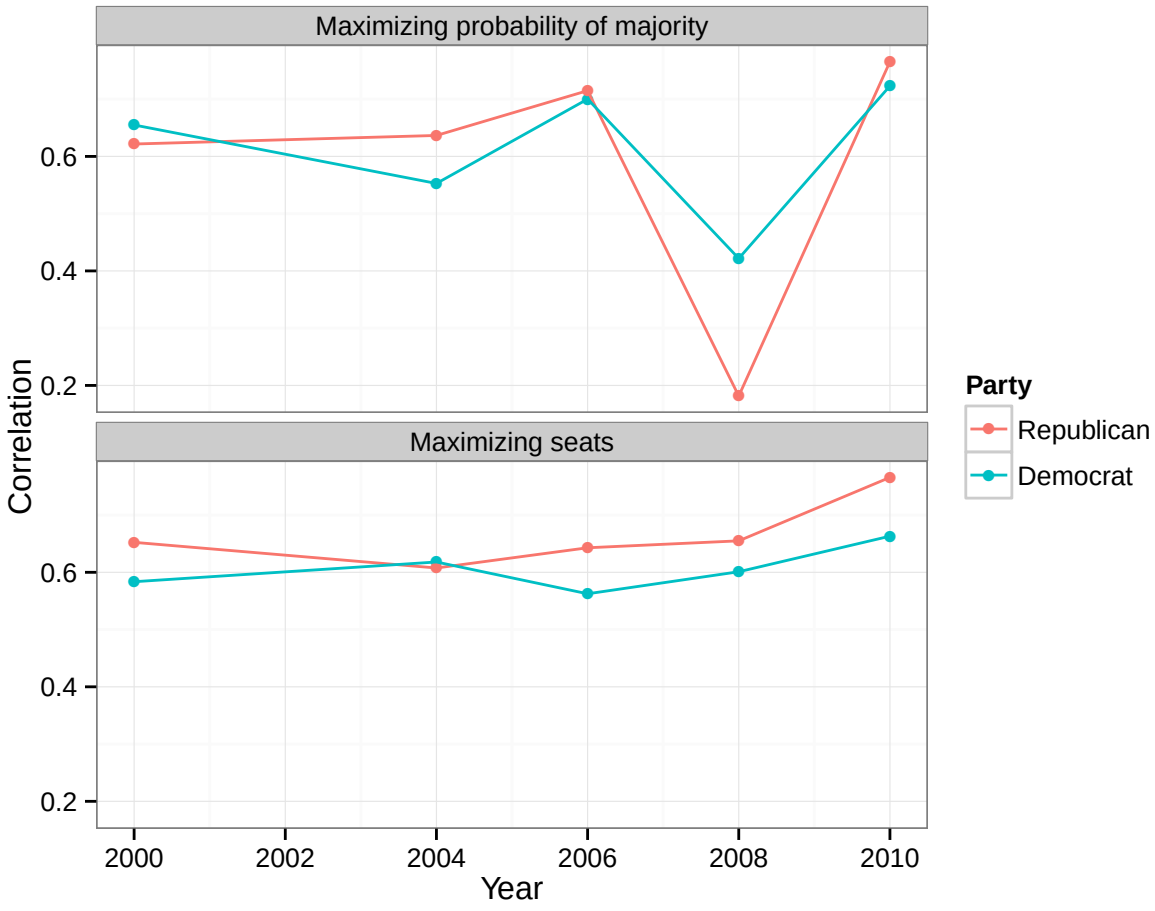


Figure 1.C.8: Correlation Between Q and PAC Spending, by Party

Notes: PAC spending for a party consists of a direct contribution by a PAC to that party or an indirect contribution by a PAC against the opposing party. The top panel shows the correlation between Q_i^{maj} and spending while the bottom panel shows the correlation between Q_i^{seats} and spending.

1.D Data Appendix

1.D.1 Gary Jacobson Data

Data were generously provided by Gary Jacobson.

1.D.2 DW-Nominate Score

The data is available for download at http://www.voteview.com/dwnominate.asp.

1.D.3 National Variables

Second quarter GDP: The percent change in U.S. GDP from one quarter to the next in chained 2009 dollars was considered but not used in the final hierarchical model. It is available for download from the Bureau of Economic Analysis at http://www.bea.gov/national/index.htm.

Presidential net approval rating: Approval ratings were scraped from the president specific webpages with links available at http://www.ropercenter.uconn.edu/CFIDE/roper/presidential/webroot/presidential_rating.cfm.

Generic Congressional Ballot: Gallup polls were searched for and downloaded using the iPOLL Databank at(http://www.ropercenter.uconn.edu/data_access/ipoll/ipoll.html. There were two main survey questions of interest that were typically worded as "If the elections for Congress were being held today, which party's candidate would you vote for in your Congressional district" and "As of today, do you lean more toward". I consider a respondent to favor a particular party if they say they would vote for that party or they lean more toward that party. Although a sample of likely voters is preferred, Gallup does not include the data used to estimate a likely voter sample, so I use a registered voter sample instead.

1.D.4 Spending Data

Data on campaign spending are from the campaign finance data provided by the Center for Responsive Politics (CRP) (https://www.opensecrets.org/). CRP obtains the campaign

finance data from the Federal Election Commission and adds value to it by cleaning and categorizing the data. Spending to specific candidates is from the "PAC" table and the "indivs" table which contain itemized spending data for PACs and individuals contributing over $200 respectively. Information on the districts of specific candidates is linked to these tables with the "Cands" table, which provides information on each candidate.

1.D.5 House Polls

Polls were scraped from the website http://www.realclearpolitics.com/epolls/other/2010_generic_congressional_vote-2171.html and the 2010 New York times House election forecasts (http://elections.nytimes.com/2010/forecasts/house.

Chapter 2

Racial, Ethnic and Educational Disparities in Pharmaceutical Expenditures

2.1 Introduction

A nearly ubiquitous finding in the health and social science literatures is that there are racial, ethnic and socioeconomic disparities in health status. Link and Phelan (1995) have gone so far as to call social conditions "fundamental causes of disease." Policymakers have tended to focus on access to medical services and differences in the quality of care when addressing these disparities. And while medical care may not be the most important component of an individual's overall health, a focus on medical care is warranted.

Its importance is best illustrated by considering the drastic increase in life expectancy and overall health during the 20th and 21st centuries. Much of the improvements during the latter half of the 20th century coincided with technological innovations and improvements in the quality of medical care. These included new surgical techniques to treat heart attacks, technologies such as special ventilators and artificial surfactants for low birthweight infants, and the use of serotonin reuptake inhibitors (SSRIs) to treat depression. Research studies

suggest that innovations such as these are the primary drivers of health improvements in the United States since the 1950s (Cutler and McClellan 2001; Cutler, Rosen and Vijan 2006). In addition, new studies that take advantage of natural experiments have shown that additional medical spending reduces mortality (e.g. Almond et al. 2010; Doyle and Joseph 2011).

It is perhaps no surprise then that disparities in access to care and utilization have been an important research area in the health policy literature. Early studies focused on access measures—such as physician office visits—and documented significant racial and ethic disparities (e.g. Collins et al. 1999; Mayberry, Mili and Ofili 2000; Weinick, Zuvekas and Cohen 2000), while more recent work has progressed toward disparities in use (e.g. Blanco et al. 2007; Cook, McGuire and Miranda 2007; Gross et al. 2008; Martinez et al. 2008; Cook and Manning 2009). The emphasis of these studies is either global (all medical use) or on particular conditions (i.e. treatments for cancers).

In this paper, I study a particular form of medical care that is influenced greatly by technological innovation but has received less attention from the disparities literature: prescription drugs. Like much of the recent literature, I focus on expenditures because it can capture differences in the intensity of care. I argue that disparities in the use of prescribed medicines should be thought of as a two-stage process that mimics an individual's demand problem. In the first stage, an individual chooses how much to invest in health–which depends on her unique environment, financial situation, preferences and biology. This choice, in turn, determines which (if any) medical conditions a provider has diagnosed her with. In the second stage, the individual has already been diagnosed with a medical condition and must decide—in what is really a joint decision with her provider and payer—whether to purchase prescribed medications to treat the disease. Disparities in prescription drug utilization can therefore exist because of differences in the probability of being diagnosed with a disease or because, conditional on having a diagnosis, individuals use fewer prescription drugs.

Following Cook, McGuire and Miranda (2007), I define disparities to be racial, ethnic, or

educational differences in care that are not due to heterogeneous preferences or clinical need. Consistent with the Institute of Medicine (IOM) (2003), patient preferences are defined as patient choices regarding utilization that are based on a full and accurate understanding of treatment options. This definition of disparities therefore recognizes that spending differences should adjust for differences in health and age but that disparities can be caused by access related factors that create barriers to care or from patient preferences that are based on incomplete knowledge. Typical access factors that might influence utilization patterns include income, health insurance, transportation, and proximity to healthcare providers or healthcare facilities. I focus on income and health insurance because they can be measured, but this does not mean that the other factors are unimportant. Moreover, although differences in treatment knowledge are important, they cannot be measured with the data used in this paper and will not be analyzed.

The empirical exercises begin by first documenting large disparities in prescription drug use. Prescription drug expenditures are modeled using a two-part model (e.g. Duan et al. 1983) that accounts for the large number nonspenders in expenditure data and the right skewed nature of nonzero values. On average, blacks and Hispanics are predicted to spend $350 and $560 less than whites respectively. Similarly, an additional 4 years of education is predicted to increase prescription drug spending by $155. A relatively small fraction of these expenditure differences—especially by race and ethnicity—can be explained by measurable access related factors: income and health insurance status explain approximately 3% of the black-white differential, 14% of the Hispanic-white differential, and 34% of the education gradient.

I then replicate the two stage decision process by first modeling the probability of being diagnosed with a disease and then modeling expenditures conditional on diagnosis. The first stage results suggest large racial/ethnic and educational gradients in diagnosis rates that are consistent across a range of disease groups. Overall, being black (Hispanic) decreases the probability of being diagnosed with a medical condition relative to whites by 6% (11%), and

an additional 4 years of education increases the probability of diagnosis by 4%.

The second stage results show that nearly all of the expenditure differentials remain even after conditioning on diagnoses. I estimate that only 10 to 30 percent of the overall disparities in both the probability of positive expenditures and total expenditures conditional on positive expenditures can be accounted for by disease groups. Furthermore, the racial/ethnic disparities are consistent across three drug classes—anti-cholesterol medications, anti-diabetics and antidepressants—although there is only an education gradient for antidepressants. All told, the results suggest that there are large racial/ethnic and educational disparities in prescription drug expenditures and that these disparities are caused by both differences in diagnosis rates and differences in utilization conditional on diagnosis.

2.2 Theoretical Framework and Empirical Estimation

The market for pharmaceutical products differs from the markets for other goods and services in a number of important ways. Chief among them are that prescription drugs are only demanded in response to medical need and that they are prescribed by providers—rather than consumers—who neither consume nor pay for them. This means that an individual will only use a prescription drug if three important steps are fulfilled: first, she must receive a diagnosis from a provider; second, the provider must write a prescription for that diagnosis; and third, she must fill the prescription. These three steps can be thought of as occurring within the context of a two-stage decision process, where in the the first stage, the individual much decide how much to invest in her health—which determines whether she is diagnosed with a medical condition—and in the second stage, she must make a decision about whether to purchase a prescription for her diagnosed condition. More formally, expected expenditures can be decomposed (approximately) as,

$$E(y) \approx \sum_{d} E(y|c_d)p(c_d), \tag{2.1}$$

where c_d and y denote medical condition d and expenditures on prescribed medications respectively. This relationship is only approximate because some conditions can be experienced at the same time.[1]

To stay consistent with the consumer's two stage decision process, I model $p(c_d)$ and $E(y|c_d)$ separately. In the first stage, the probability of being diagnosed with a medical condition depends on an individual's propensity to seek care, which can depend on either the demand for curative care or preventive care (or both). In some cases, the disease will cause enough discomfort to cause the individual to seek curative care to receive relief from symptoms. In other cases, the disease will remain undetected unless the patient uses preventive services such as a physical examination or a medical screening test.[2] The probability of being diagnosed with a disease is modeled with a simple logistic regression,

$$p(c_{idt} = 1) = \Lambda\left(z_{it}^T \gamma\right), \tag{2.2}$$

where $c_{idt} = 1$ is an indicator variable coded as 1 if individual i has condition d in year t and 0 otherwise, $\Lambda(\cdot)$ is the inverse logit function, and z_{it} is a column vector containing the primary variables of interest (black race, Hispanic ethnicity and years of education) as well as relevant control variables.

In the second stage, the consumer decides whether to purchase prescribed medications after having been diagnosed with a medical condition. Unlike in stage one, demand is only for curative care since the condition has already been diagnosed. Modeling medical expenditures is not straightforward due to the large number of observed zero's and the heavily right skewed distribution of the remaining values (see Figure 2.C.1 in the appendix). To address these issues, I model expenditure in two parts. First, the individual decides whether to use prescription drugs, and then, conditional on usage, decides how much to

[1] If medical conditions were mutually exclusive events, then this equation would hold exactly.

[2] This type of prevention, which reduces the severity of disease without affecting the probability that the disease occurs is called secondary prevention. Activities that reduce the probability of an illness occurring, such as exercise, are known as primary prevention.

spend. In addition to providing a better fit to the data, the two-part approach is useful conceptually because it captures both the extensive margin (whether individuals choose to consume a drug) and the intensive margin (how intensely individuals consume the drug). This distinction is important because the choice of whether to use a drug is predicated on prior beliefs about the efficacy of medications while the intensity of consumption depends on posterior beliefs obtained through experience.[3]

Formally, the model can be written as,

$$D_{it}^{*} = x_{1it}^{T}\alpha + \epsilon_{1it}, \tag{2.3}$$

$$\ln y_{it} | D_{it}^{*} > 0 = x_{2it}^{T}\beta + \epsilon_{2it}, \tag{2.4}$$

where $D_{it} \equiv I(D_{it}^{*} > 0) = I(y_{it} > 0)$ is a latent variable that describes whether expenditures are positive or zero. An important question is whether the error terms, ϵ_{1it} and ϵ_{2it}, are correlated. If they are correlated then a type 2 Tobit model is appropriate; otherwise, a two-part model can be used. This paper focuses on the two-part model because it has a simpler interpretation. That said, the appendix shows that both approaches generate similar results and that the two-part model actually fits the data better.

If ϵ_{1it} and ϵ_{2it} are independent, then the two equations are separable. The first equation can be modeled using a standard regression model for binary data—this paper uses a logistic regression. The second equation can be estimated using a simple linear regression model. Importantly, y_{it} is log transformed because the regression coefficients are more likely to be additive on the log scale and multiplicative on the raw scale.[4] As a result, the estimated gradients in prescription drug expenditures are easily interpreted in percentage terms. A disadvantage of the transformation is that it is more difficult to estimate the impact of the

[3]Prescription drugs are experience drugs because utility can only be ascertained upon consumption and continued use depends on satisfactory experiences. Indeed, Dranove (2009) notes that healthcare may be "the quintessential experience good."

[4]Another common approach is to model the second component using a generalized linear model (GLM). See Appendix 2.A for a brief discussion.

coefficients on expenditures in levels because predictions for y depend of the distribution of ϵ_{2it} when y_{it} is log transformed. Section 2.4 and Appendix 2.A examine this issue in more detail.

2.3 Data

This paper uses data from the Medical Expenditure Panel Survey (MEPS) from 1996 to 2012 for individuals age 25 and older. Each survey is a nationally representative sample of the U.S. population drawn from the National Health Interview Survey (NHIS). The MEPS is the most reliable source of nationally representative medical expenditure data available and is used to construct data for the National Health Accounts. The dataset is a rolling two-year panel with approximately 15,000 new respondents interviewed each year. Respondents are followed for two years and interviews are conducted three times each year.[5] I combine data from the Full-Year Consolidated Files, Prescribed Medicine Files and Medical Conditions Files to obtain detailed information on individual characteristics, medical conditions and prescription drug expenditures. The pharmaceutical data is especially reliable because the data is primarily based on pharmacy records rather than self reports.[6]

2.3.1 Variables

Summary statistics for the main variables are shown in Table 2.1. Since the MEPS does not use simple random sampling, the table compares summary statistics from the sample with population estimates using the survey weights provided by the MEPS. The differences in the sample and population means tend to be small, although there are significant differences in the distributions of the race, income and education variables because the MEPS oversamples certain "policy relevant" sub-groups such as racial minorities and low-income groups.

The primary dependent variable is prescription drug expenditures, which is normalized

[5]Respondents are actually interviewed five times over the two years but the third round of the survey overlaps two calendar years and is split into two distinct periods.

[6]If individuals refuse to release their pharmacy records, then expenditures are based on self-reports that are adjusted for outliers and item non-response based on imputations from the pharmacy data.

Table 2.1: Summary Statistics

	Sample Statistics				Population Estimates	
	Mean	SD	Min	Max	Mean	Median
Rx expenditures	959.412	4861.699	0	2,273,083	1005.175	155
Purchased a prescription	0.664	0.472	0	1	0.701	1
Rx medical condition	0.644	0.479	0	1	0.670	1
White	0.559	0.497	0	1	0.716	1
Other race	0.064	0.245	0	1	0.057	0
Black	0.156	0.363	0	1	0.110	0
Hispanic	0.221	0.415	0	1	0.118	0
Years of education	12.467	3.323	0	17	13.095	13
Health status excellent	0.221	0.415	0	1	0.240	0
Heath status very good	0.313	0.464	0	1	0.333	0
Health status good	0.302	0.459	0	1	0.286	0
Health status fair	0.120	0.325	0	1	0.103	0
Health status poor	0.044	0.204	0	1	0.038	0
Poor	0.150	0.357	0	1	0.101	0
Near poor	0.054	0.227	0	1	0.040	0
Low income	0.155	0.362	0	1	0.130	0
Middle income	0.306	0.461	0	1	0.310	0
High income	0.335	0.472	0	1	0.419	0
Insured	0.825	0.380	0	1	0.873	1
Age	48.751	15.854	25	90	49.351	47
Female	0.541	0.498	0	1	0.522	1

Notes: The sample consists of 330,580 individuals sampled by the MEPS from 1996 - 2012. Income categories are calculated by dividing family income by the applicable poverty line; the five categories are poor (less than 100%), near poor (100% to less than 125%), low income (125% to less than 200%), middle income (200% to less than 400$) and high income (greater than 400%). The population estimates are calculated using the survey weights provided by the MEPS. Rx expenditures are in 2012 dollars.

to 2012 dollars using the CPI. As shown in the table, expenditures are heavily right skewed with a mean that is considerably higher than the median (also see Figure 2.C.2). The two-part model seems well suited for modeling pharmaceutical expenditures since expenditures are zero for 34% of the observations. Maximum expenditures are over $2 million, which is more than 5 times as large as the next largest value. Although including this observation has little effect on the primary variables of interest it grossly inflates prediction errors so its is dropped from the analysis.[7]

The primary independent variables are the racial/ethnic variables and the years of education variable. The years of education variable ranges from 0 to 17; the average individ-

[7]The RMSE in the k-fold cross-validation reported in Appendix 2.A increases by a factor of nearly 4 when the observation is included in the validation set.

ual has obtained 12 years of schooling, which is approximately equivalent to a high school diploma. The race/ethnicity variables contain four mutually exclusive indicator variables for non-Hispanic whites, non-Hispanic blacks, Hispanics, and an "other" race group.

Disease groups included in the model are those that have proven to be the most important predictors of prescription drug expenditures. Specifically, I use the prescription drug condition categories used by the Center for Medicare & Medicaid Services (CMS) in its risk adjustment models for Medicare Part D, the prescription drug portion of Medicare (see Robst, Levy and Ingber 2007).[8] Each condition category consists of ICD-9 codes that are clinically homogeneous, similarly expensive to treat, and treated with similar drugs. I use the MEPS medical conditions files to create indicator variables for 70 condition categories based on the CMS model, whose construction is described in more detail in Appendix 2.B).[9] Not surprisingly, the proportion of individuals with at least one Rx medical condition, 0.64, is nearly identical to the proportion of individuals who have purchased at least one prescription, 0.66.

To get a sense of whether the condition categories are important predictors of pharmaceutical utilization, Table 2.2 shows three measures of utilization by whether an individual has at least one condition. As one would expect, those without a condition use far fewer prescribed medications than those with a condition. Those without a condition, do however, use some medications, which could be due to reporting error since the medical conditions are self-reported or because there are some ICD-9 codes that are not included in the 70 disease groups that are sometimes treated with a small amount of medication.

The remaining variables were chosen to be consistent with my definition of disparities. Variables used to adjust for clinical need are dummy variables for self-reported health status (five categories from poor to excellent) and single year of age. Additional covariates are

[8]Disease categories are only included in the final CMS model if they are deemed to be important predictors of prescription drug expenditures.

[9]Unlike the expenditure data, the medical conditions are based solely on self-reports and may therefore be measured with error due to misreporting. However, to the extent that purchasing prescribed medicines depends on knowledge of one's own medical conditions rather than actual diagnoses from providers, using self-reported medical conditions may be superior to actual diagnoses.

Table 2.2: Prescription Drug Utilization by Disease Status

	Rx condition	
	No	Yes
Proportion using prescibed medications	0.287	0.873
Number of prescriptions filled	1.161	18.954
Rx expenditures	64.841	1454.869

included to control for geographic region (west, midwest, south and northeast), marital status and calendar year. The clinical need variables and additional covariates are referred to as the basic controls. Access variables include dummy variables for income category (five categories from poor to high income) and insurance status (equal to 1 if insured).[10] I use a variable for health insurance rather than prescription drug insurance because the variables needed to construct a measure of prescription drug insurance contain a substantial number of missing observations.

The same basic controls and access variables are included in all models so in general, $z_{it} = x_{1it} = x_{2it}$. The one exception is when the design matrix for the expenditure regressions includes the full set of Rx disease group dummies. In practice, the Heckman 2-step estimator is problematic because there are no obvious variables that should be included in x_{2it} but not x_{1it}. Identification in the second stage regression comes solely from the nonlinearity of the inverse mills ratio which is approximately linear over a large range of its arguments (Puhani 2000).

2.3.2 Unadjusted Differences in Expenditures

Figure 2.1 shows prescription drug expenditures and the proportion of individuals who purchased at least one medication by year and race/ethnicity from 1996 to 2012. Although there are clear time trends in both expenditures (upward) and the proportion of individuals using (slightly downward), the racial gaps have remained fairly constant over time: whites

[10]Income categories are calculated by dividing family income by the applicable poverty line; the five categories are poor (less than 100%), near poor (100% to less than 125%), low income (125% to less than 200%), middle income (200% to less than 400$) and high income (greater than 400%)

have consistently used more prescription drugs than blacks and Hispanics.

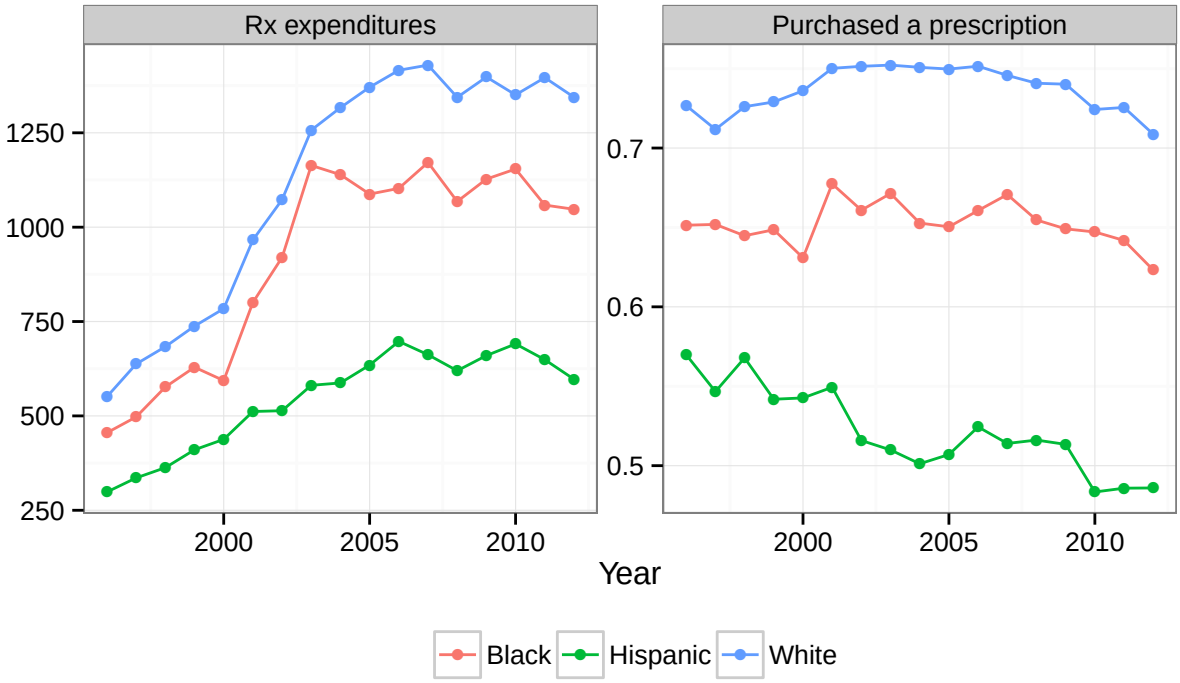


Figure 2.1: Prescription Drug Use By Year and Race/Ethnicity

Notes: Each point corresponds to mean drug use by race and year.

Figure 2.2 repeats this figure for years of education. The right panel is consistent with the race/ethnicity disparities as the less educated are less likely to have nonzero expenditures than the more educated. However, the opposite is true in the leftmost panel: there is actually a negative association between prescription drug spending and education. This finding is not entirely unexpected because better educated individuals tend to be in better health which should decrease demand for prescription drugs. Indeed, as shown in the next section, this result reverses after controlling for self-reported health status.

Table 2.3 reports unconditional results from the first stage of the decision process. In particular, it provides summary data on the Rx condition categories by race/ethnicity and education. Similar to the patterns in the expenditure data, there are clear racial and ethnic gradients in diagnoses but there is no evidence of an education gradient. Whites, on average,

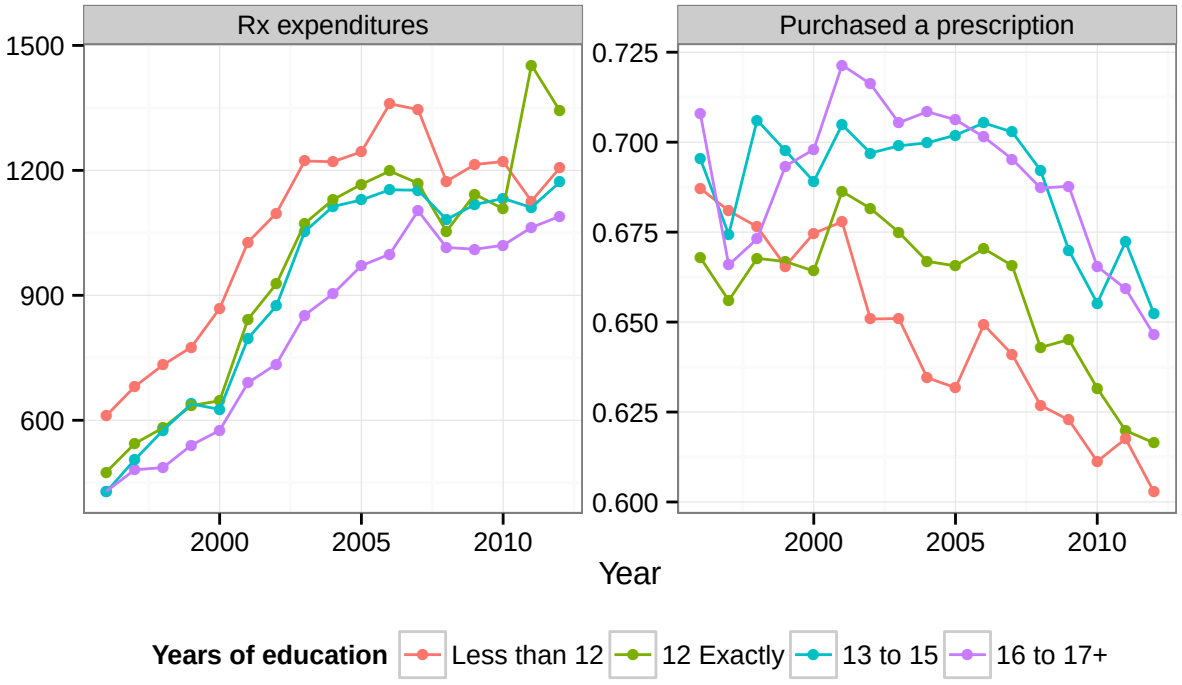


Figure 2.2: Prescription Drug Use By Year and Education

Notes: Each point corresponds to mean drug use by education category and year.

have more total medical conditions and are the most likely to have been diagnosed with at least one condition, while the Rx conditions are the least prevalent among Hispanics. There is, if anything, a negative association between years of education and diagnoses, although the pattern is not entirely linear. The racial and ethnic differentials are somewhat surprising since one would expect racial and ethnic minorities—who have been documented in study after study to be in worse health—to have more medical conditions. As with the expenditure data, Section 2.4 shows that the apparent negative relationship between education and the Rx conditions reverses after controlling for health status, which ultimately suggests that there is in fact an education gradient in diagnoses.

Table 2.3: Rx Conditions by Race/Ethnicity and Education

	Proportion with an Rx condition	Mean number of Rx conditions
Race/Ethnicity		
White	0.704	1.946
Hispanic	0.512	1.200
Black	0.643	1.701
Years of education		
Less than 12	0.651	1.901
12 Exactly	0.643	1.712
13 to 15	0.656	1.720
16 to 17+	0.639	1.517

2.4 Results

This section reports disparities in prescription drug spending after adjusting for covariates. It begins by providing regression adjusted estimates of disparities without conditioning on disease groups and then examines the role that diagnoses play in those estimates. More precisely, it first provides an estimate of overall disparities and then breaks the disparities down into the first and second stages of the consumer decision problem. All analyses are based on the $314,462$ observations with complete data for both the basic controls and access related variables.

Regression estimates of overall disparities in expenditures are shown in Table 2.4. Estimates are based on the two-part model described in Section 2.2. The first three columns report estimates from logistic regressions estimating the probability of positive expenditures. The final three columns report regression coefficients (and robust standard errors) from OLS regressions predicting the log of expenditures given that expenditures are positive. The table includes models with basic controls and models that add the access variables. The percent reduction in the coefficients after adding the access variables are computed for convenience.

The table provides the means of the dependent variables to help with the interpretation of the coefficients. The marginal effect of variable j in the logistic regression is approximately $\alpha_j \Lambda(x_{1it}^T \alpha)[1 - \Lambda(x_{1it}^T \alpha)] < \alpha_j/4$ since the inverse logit function is steepest at $\Lambda(0) = 1/2$. At

Table 2.4: Regression Estimates of Disparities in Utilization

	$1 \cdot (expenditures > 0)$			$\ln(expenditures) \mid expenditures > 0$		
	Basic controls	Adding access controls		Basic controls	Adding access controls	
	Coefficient	Coefficient	Reduction in coefficient	Coefficient	Coefficient	Reduction in coefficient
Black	−0.398	−0.372	7%	−0.310	−0.308	1%
	(0.012)	(0.013)		(0.010)	(0.010)	
Hispanic	−0.715	−0.543	24%	−0.528	−0.466	12%
	(0.011)	(0.012)		(0.011)	(0.011)	
Years of education	0.070	0.042	39%	0.026	0.018	30%
	(0.001)	(0.002)		(0.001)	(0.001)	
Mean of dv	0.670	0.670		1431.97	1431.97	
Observations	314,462	314,462		210,746	210,746	

Notes: The 1st-3rd and 4th-6th columns report results from logistic and OLS regressions respectively. Standard errors are in parentheses. Basic controls are indicator variables for "other race", single year of age, self-reported health status, marital status, sex and year. Additional access controls are indicator variables for five income categories and health insurance status.

the sample average, the impact of a coefficient is approximately $\alpha_j(0.67)(1-0.67) \approx 0.22\alpha_j$. For indicator variables, it is preferable to calculate an incremental effect—the change in the probability of positive expenditures when a variable is changed from 0 to 1— but the marginal effects are still a reasonable approximation.[11] For instance, at the sample average, the marginal effect of being Hispanic on the probability of positive expenditures in the first column is −0.158 and the incremental effect is −0.172. Both estimates suggest large effects: the model predicts that an Hispanic individual would have around a 52% chance of using prescription drugs when, all else equal, a white individual would have a 67% probability of usage.

The estimated disparities are consistent across both parts of the two-part model: blacks, Hispanics and the less educated are less likely to use prescribed medications and spend less on them when they do use them. The largest disparities are between Hispanics and whites in both components of the two-part model. The second component predicts that Hispanics spend $100 \cdot |\exp(-0.53) - 1| \approx 41\%$ less than whites without accounting for differences in

[11] Mathematically, the incremental effect for variable j is $\Lambda(x_{1it,-j}^T \alpha_{-j} + \alpha_j) - \Lambda(x_{1it,-j}^T \alpha_{-j})$ where $-j$ refers to all variables except variable j.

access to care. Disparities for blacks and the less educated are significant as well but not quite as large. For example, when the model includes basic controls, an additional 4 years of education is predicted to increase the probability of using prescription drugs by 6% at the sample average and to increase spending conditional on nonzero spending by approximately 10%. The education gradient in nonzero expenditures is consistent with previous work which has shown that, unadjusted, those with less education spend more on medical care than more educated individuals but that this result reverses after controlling for health status (Bhattacharya and Lakdawalla 2006).[12]

Table 2.C.1 in Appendix 2.C looks at whether these results change if ϵ_1 and ϵ_2 are allowed to be correlated. Regression coefficients are reported for three different estimators: the two-part model with a probit regression in the first component, a Heckman 2-step selection model, and a maximum likelihood type 2 Tobit model.[13][14] The coefficients between all methods are similar and none of the results are qualitatively different.

To assess the full two-part model, I examine the impact of race/ethnicity and education on predicted expenditures. Since the second part of the model is in logs, expected expenditures are $E(y|x_1, x_2) = \Lambda(x_1\alpha)\exp(x_2\beta)E(\exp(\epsilon_2)|x_2)$. The third term depends on both the distribution of the errors and whether the errors are heteroscedastic. Unfortunately, neither a normality assumption or a constant variance assumption is innocuous. I consequently use a Duan (1983) smearing factor that varies by single-year of age to estimate the distribution of the errors.[15] Appendix 2.A shows that predictions with this smearing factor are much more accurate than predictions that assume that the errors are normally distributed or assume a constant smearing factor.

Figure 2.3 uses the smearing factor to calculate predicted expenditures and average

[12] Bhattacharya and Lakdawalla (2006) control for insurance status as well.

[13] The Heckman 2-step and maximum likelihood models are often referred to as limited information maximum likelihood (LIML) and full information maximum likelihood (FIML) methods respectively.

[14] A probit specification is used in the first component of the two-part model to maintain consistency with the selection models.

[15] Duan's smearing factor is $s = E(\widehat{\exp(\epsilon_{2j})}|x_2) = \frac{1}{N_j}\exp(\ln y_j - x_{2j}\hat{\beta})$ where N_j is the number of individuals in age-group j.

marginal/incremental effects (AME's/AIE's) for the three primary variables of interest in models with basic controls. Panel (a) shows how varying an "average" individual's education or race/ethnicity would change his or her predicted expenditures. This "average" individual is initially assumed to be white with 12 years of schooling. Given these baseline characteristics, the probability of using prescription drugs, expenditures conditional on use, and the smearing factor are then set to sample averages.[16] A white individual with 12 years of education is consequently predicted to have expenditures equal to average expenditures in the estimation sample, or $960.

The figure shows that the racial/ethnic and educational gradients in utilization for this "average" individual are economically significant. A white individual with a college education (16 years of schooling) is predicted to spend $198 more than if he or she had a high school education (12 years of schooling). The effects are even larger by race/ethnicity: blacks and Hispanics with 12 years of schooling are predicted to spend spend $210 and $428 less than their white counterparts respectively. It is worth noting that the education gradient is slighter steeper for whites than for blacks and Hispanics because the covariates have multiplicative effects on non-transformed expenditures and predicted expenditures are higher for whites.

Panel (b) calculates AME's for education and AIE's for race/ethnicity. Calculations are made with both the basic controls and after adding the access variables. The AME is the marginal impact of education on (raw) expenditures averaged across all individuals in the data. An AIE is the effect of switching an indicator variable from 0 to 1 on expenditures, again averaged over survey respondents. Both calculations are made using a smearing factor that varies by age. Standard errors are calculated using a 1,100 boostrap replications in order to account for uncertainty in both the coefficients and the transformation of expenditures from the log to the raw scale.

The AME/AIE's are consistent with the predicted expenditures in panel (a). The coef-

[16] More specifically, I set $\Lambda(x_1\alpha) = \overline{D}$, $exp(x_2\beta)s = (\overline{y}|y > 0)$ and solve for $x_1\alpha$ and $x_2\beta$ when the smearing factor, s, is equal to its mean. Predictions are then made by changing $x_1\alpha$ and $x_2\beta$ by varying race, ethnicity and education.

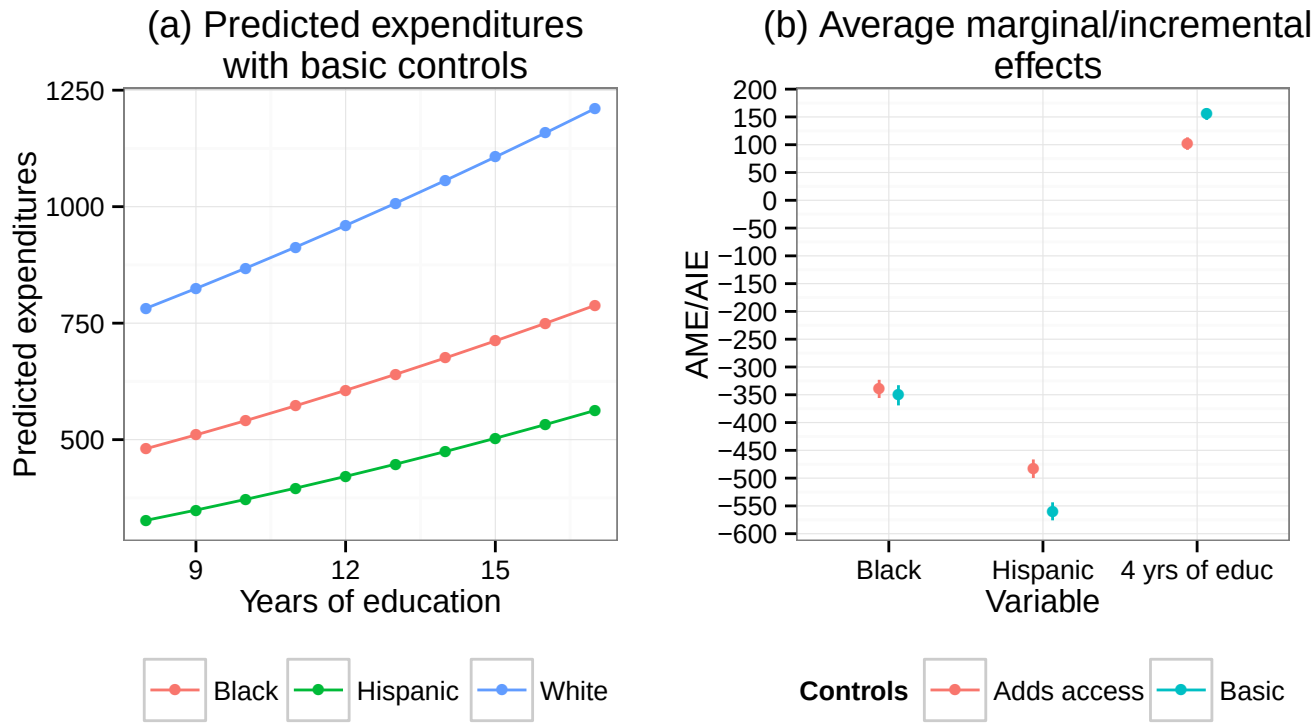


Figure 2.3: Predicted Disparities in Expenditures

Notes: (a) plots mean predicted expenditures by education and race/ethnicity for an "average" individual. (b) plots average marginal effects and bootstrapped 95% confidence intervals for each of the primary variables of interest. The AME for education is the original estimate multiplied by 4. Predicted expenditures and AME/AIEs are calculated using the two-part model and a Duan Smearing estimator that varies by single-year of age.

ficients are estimated very precisely since the sample size is so large. Blacks and Hispanics are predicted to spend, on average, $350 and $560 less than whites. An additional 4 years of education has a significant impact as well as it is predicted to increase expenditures by $155. Only a relatively small fraction of the racial/ethnic disparities in expenditures—3% of the black-white differences and 14% of Hispanic-white differences—are associated with access to care. Access factors explain a larger percentage of the education gradient ($\approx$ 34%) but a substantial fraction of the gradient still remains unaccounted for.

2.4.1 Disparities in Diagnoses

The outcome of a consumer's first stage decision problem is whether she has been diagnosed with a medical condition. Table 2.5 provides regression estimates of (1) the probability of being diagnosed with at least one of the Rx disease groups and (2) the number of disease groups that an individual is diagnosed with. The binary variable is modeled with a logistic regression and the count variable is modeled using a negative binomial regression.[17]

The table shows that the racial and ethnic disparities from the raw data remain after controlling for covariates, but that an education gradient emerges as well. There are racial and educational gradients in both the probability of having a condition and the number of total conditions. The largest disparities are again between whites and Hispanics, and the smallest are between education levels.

Table 2.5: Regression Estimates of Disparities in Diagnoses

	1·(# of conditions > 0)			# of conditions		
	Basic controls	Adding access controls		Basic controls	Adding access controls	
	Coefficient	Coefficient	Reduction in coefficient	Coefficient	Coefficient	Reduction in coefficient
Black	−0.357	−0.343	4%	−0.219	−0.221	-1%
	(0.013)	(0.013)		(0.005)	(0.005)	
Hispanic	−0.613	−0.494	19%	−0.351	−0.295	16%
	(0.012)	(0.012)		(0.005)	(0.005)	
Years of education	0.055	0.038	32%	0.018	0.014	23%
	(0.001)	(0.002)		(0.001)	(0.001)	
Mean of dv	0.647	0.647		1.713	1.713	
Observations	314,462	314,462		314,462	314,462	

Notes: The 1st-3rd and 4th-6th columns report results from logistic and negative binomial regressions respectively. Standard errors are in parentheses. Basic controls are indicator variables for "other race", single year of age, self-reported health status, marital status, sex and year. Additional access controls are indicator variables for five income categories and insurance status.

To interpret the coefficients, note that the marginal effect of covariate j in the negative binomial regression is $\kappa_j \exp(z_{it}^T \kappa)$ and the incremental effect of a one unit change in a variable is $\exp(z_{it}^T \kappa)[\exp(\kappa_j) - 1]$ where κ is the vector of covariates. The economic importance of the

[17] A regression based test of overdispersion, as suggested by Cameron and Trivedi (1990), indicated the presence of significant overdispersion.

covariates can then be quickly assessed by setting $\exp(z_{it}^T\kappa)$ and $\Lambda(z_{it}^T\gamma)$ equal to sample averages of the dependent variables in the negative binomial and logistic regressions respectively and calculating incremental and marginal effects for the appropriate model. For instance, at sample averages (using the basic controls), the probability that an Hispanic individual is diagnosed with a medical condition is 15% lower than the probability that a similar white individual would have been diagnosed. Likewise, the negative binomial regression predicts that an Hispanic individual is diagnosed with 0.51 fewer medical conditions that a white individual. The education gradient is smaller but not insignificant: an additional 4 years of education increases the probability of having a medical condition by 5% and the number of medical conditions by 0.12. AME's and AIE's are very similar to the effects at sample averages: averaged over the population, the black-white and Hispanic-white differences in the probability of having a medical condition are 6% and 11% respectively, and 4 additional years of education raises the probability of having a medical condition by 4%.

The effects of the access variables on the gradients in diagnoses are consistent with their effects on the gradients in expenditures. The access variables cannot account for almost any of the differences in diagnoses between blacks and whites but can account for a small fraction of the differences between Hispanics and whites. A larger fraction of the education gradient is associated with access factors, but most the education gradient remains unexplained even after controlling for income and health insurance status.

I next turn to more disaggregated evidence to examine whether there are specific disease groups that are driving these results. Table 2.6 lists the 20 most common diseases and the proportion with each disease by race/ethnicity. Whites are more likely to have been diagnosed with nearly all conditions than blacks and Hispanics. There are however some important exceptions such as diabetes and hypertension (for blacks but not Hispanics). Not surprisingly, the most commonly dispensed prescriptions in the United States as reported by IMS Health—which include antihypertensives, antidepressants, lipid regulators and antidiabetics—are used to treat the most prevalent disease groups listed in Table 2.6.

Table 2.6: Prevalence of Most Common Rx Disease Groups

	Proportion with disease			
Disease group	All	White	Black	Hispanic
Other Musculoskeletal and Connective Tissue Disorders	0.283	0.317	0.283	0.211
Hypertension	0.238	0.241	0.338	0.168
Disorders of Lipid Metabolism	0.146	0.167	0.134	0.100
Diabetes without Complication	0.096	0.082	0.137	0.102
Mild Depression	0.086	0.105	0.063	0.066
Allergic Rhinitis	0.077	0.089	0.056	0.059
Asthma and COPD	0.060	0.067	0.066	0.040
Thyroid Disorders	0.059	0.078	0.032	0.037
Anxiety Disorders	0.056	0.072	0.037	0.036
Headache (Including Migraine)	0.055	0.058	0.049	0.052
Esophagael Disorders	0.051	0.066	0.051	0.021
Coronary Artery Disease	0.035	0.042	0.035	0.020
Acute Bronchitis and Congenital Lung/Respiratory Anomaly	0.034	0.045	0.022	0.018
Disorders of the Vertebrae and Spinal Discs	0.029	0.039	0.023	0.013
Bone/Joint/Muscle Infections	0.028	0.037	0.013	0.017
Other Diseases of Upper Respiratory Tract	0.023	0.025	0.023	0.020
Peptic Ulcer and Gastrointestinal Hemorrhage	0.022	0.021	0.021	0.031
Urinary Obstruction and Retention	0.022	0.025	0.019	0.017
Menopausal Disorders	0.021	0.027	0.013	0.011
Other Cancers and Tumors	0.021	0.026	0.021	0.010

To see whether these disparities hold up after regression adjustment, I use logistic regression models to predict the probability of having each of the 70 Rx condition categories. Models are estimated using both the basic controls and after adding the access variables. Regression coefficients and 95% confidence intervals for the 20 most common diseases (those listed in Table 2.6) are shown in Figure 2.4. The coefficients on the black and Hispanic indicator variables are consistently negative and statistically distinguishable from 0. As with the unadjusted data, the two primary exceptions are diabetes and hypertension. The effects on each disease are relatively small, as the coefficients tend to be less than 0.05. Nonetheless, when summed across multiple diseases these small differences create significant disparities in the aggregate probability of being diagnosed with a disease.

The coefficient on the education variable is multiplied by 4 so that its economic importance can be more reasonably compared with the race and ethnicity variables. The coefficient tends to be small and in most cases (although still positive) close to 0. This is consistent

with the results shown in columns 1 and 2 of Table 2.5, which showed that additional years of education were associated with small, but positive, increases in the probability of having at least one medical condition.

As expected, the access variables tend to push the estimated coefficients toward zero. The effects are small although they are once again more pronounced for the education and Hispanic ethnicity variables than for the black race variable.

2.4.2 Disparities in Conditional Expenditures

The second stage of the consumer decision problem occurs after an individual is diagnosed with a medical condition. To examine whether there are disparities at this stage I add dummy variables for each of the 70 disease groups to the two-part expenditure model. The idea is to see if disparities remain even after conditioning on the disease groups that are the best predictors of prescription drug expenditures. The results are reported in Table 2.7.

To facilitate comparison, I re-report the regression coefficients from Table 2.4, which are from models that did not condition on the disease groups. The table displays the reduction in two gradients—the full gradient (i.e. models with basic controls) and the gradient unexplained by access (i.e. models with access controls)—after controlling for diagnoses. The size of the coefficients after controlling for the disease groups are measures of gradients in the consumer's second stage decision problem.

The coefficients in both models with and without access controls tend to decline by around 10 to 30 percent. The decline is due to the positive correlation between diagnoses and spending, and the negative correlation between diagnoses and being a member of a disadvantaged group. Even so, the coefficients remain highly statistically and economically significant. To illustrate, in models with basic controls (and when incremental effects are evaluated at sample averages), being Hispanic is still predicted to lower the probability of positive expenditures by 12% after conditioning on disease groups. Moreover, Hispanics are predicted to spend 31% less than whites in the second component of the two-part model

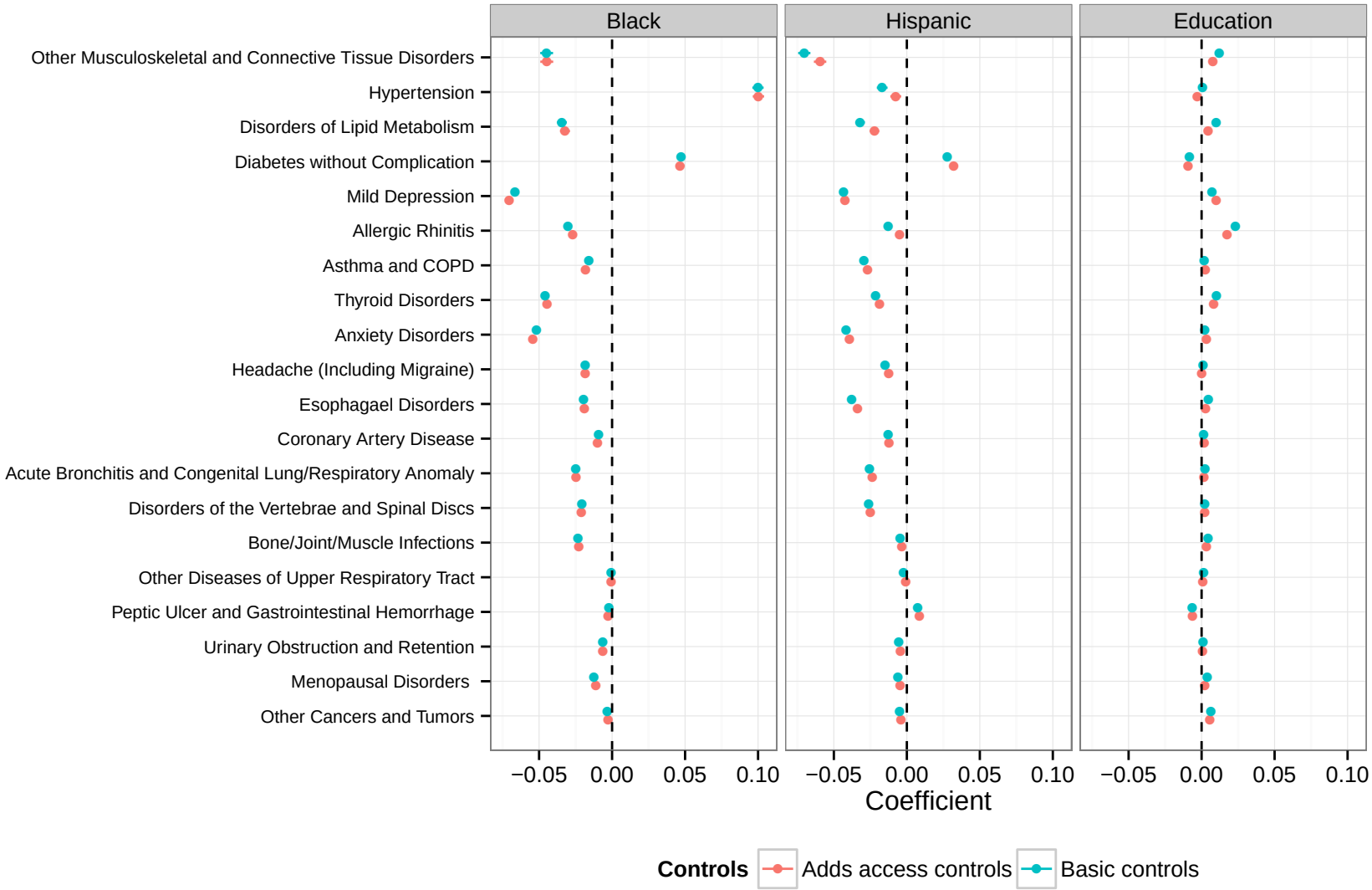


Figure 2.4: Regression Estimates of Disparities in Specific Diagnoses

Notes: The figure plots coefficients and 95% confidence intervals from logistic regression models. Coefficients are for the 20 most common disease groups. The reported coefficient for the education variable is the original estimate multiplied by 4. Basic controls are indicator variables for "other race", single year of age, self-reported health status, marital status, sex and year. Additional access controls are indicator variables for five income categories and insurance status.

Table 2.7: Regression Estimates of Disparities in Utilization Conditional on Rx Condition Categories

	Basic controls			Adds access controls		
	No Rx disease groups	Adds Rx disease groups		No Rx disease groups	Adds Rx disease groups	
	Coefficient	Coefficient	Reduction in coefficient	Coefficient	Coefficient	Reduction in coefficient
Panel A.	$p(expenditures > 0)$					
Black	−0.398	−0.333	16%	−0.372	−0.308	17%
	(0.012)	(0.015)		(0.013)	(0.015)	
Hispanic	−0.715	−0.520	27%	−0.543	−0.378	30%
	(0.011)	(0.014)		(0.012)	(0.014)	
Years of education	0.070	0.056	20%	0.042	0.031	27%
	(0.001)	(0.002)		(0.002)	(0.002)	
Mean of dv	0.670	0.670		0.670	0.670	
Observations	314,462	314,462		314,462	314,462	
Panel B.	$ln(expenditures)\|expenditures > 0$					
Black	−0.310	−0.241	22%	−0.308	−0.228	26%
	(0.010)	(0.009)		(0.010)	(0.009)	
Hispanic	−0.528	−0.378	28%	−0.466	−0.331	29%
	(0.011)	(0.009)		(0.011)	(0.009)	
Years of education	0.026	0.023	12%	0.018	0.014	23%
	(0.001)	(0.001)		(0.001)	(0.001)	
Mean of dv	1431.97	1431.97		1431.97	1431.97	
Observations	210,746	210,746		210,746	210,746	

Notes: Models in Panel A and Panel B are estimated using logistic regression and OLS respectively. The first three columns report regression results from models that include the basic controls (indicator variables for "other race", single year of age, self-reported health status, marital status, sex and year). The final three columns report regression coefficients from models that add the access controls (indicator variables for five income quintiles and health insurance status) to the basic controls. Models with Rx disease groups include 70 dummy variables based on the CMS risk adjustment model for Medicare Part D. Standard errors are in parentheses.

even after conditioning on diagnoses. This shows that the gradients in prescription drug expenditures are due to disparities in the likelihood of receiving proper diagnoses for diseases *and* from disparities in the use of prescription drugs after diagnosis.

Next, I look at whether these "conditional" disparities exist for a few of the most common therapeutic drug classes: anti-diabetics, antidepressants and anti-cholesterol medication. For each therapeutic class, I limit the sample to those diagnosed with a relevant disease.[18] The results are displayed in Figure 2.5. For reference, the proportion of survey respondents in the anti-diabetic, antidepressant and anti-cholesterol samples with positive expenditures are 0.83, 0.6 and 0.78 respectively.

The racial and ethnic disparities for each drug class are generally consistent with those in Table 2.7: blacks and Hispanics are less likely to have purchased prescribed medications for each drug class (except antidiabetics) and spend less when they do. The education gradient, on the other hand, is not: education only has a consistent positive effect on expenditures for anti-depressant drugs. The coefficients move toward zero after accounting for access but the size of the movement is generally small.

Estimates of the gradients in expenditures from the full two-part model are shown in Figure 2.6. The figure shows mean predicted expenditures for an "average" individual using the same procedure used for panel (a) of Figure 2.3. Mean expenditures for each drug class are fairly large—$806, $382, $564 for the anti-diabetic, antidepressant and anti-cholesterol samples respectively. The slope of education line is only significantly positive for antidepressants but expenditures at each level of education are considerably lower for blacks and Hispanics than whites. Differences in expenditures between whites and racial minorities tend to be on the order of $100 to $200. An "average" white individuals diagnosed with depression with a college level of education (16 years of schooling) is predicted to spend $70 more on antidepressants than an "average" white individuals with a high school education (12 years

[18]The anti-diabetic sample consists of individuals with diabetes with or without complications; the antidepressant sample contains individuals with mild depression, episodic mood disorders, personality disorders, or anxiety disorders; and individuals in the anti-cholesterol sample are those with disorders of lipid metabolism.

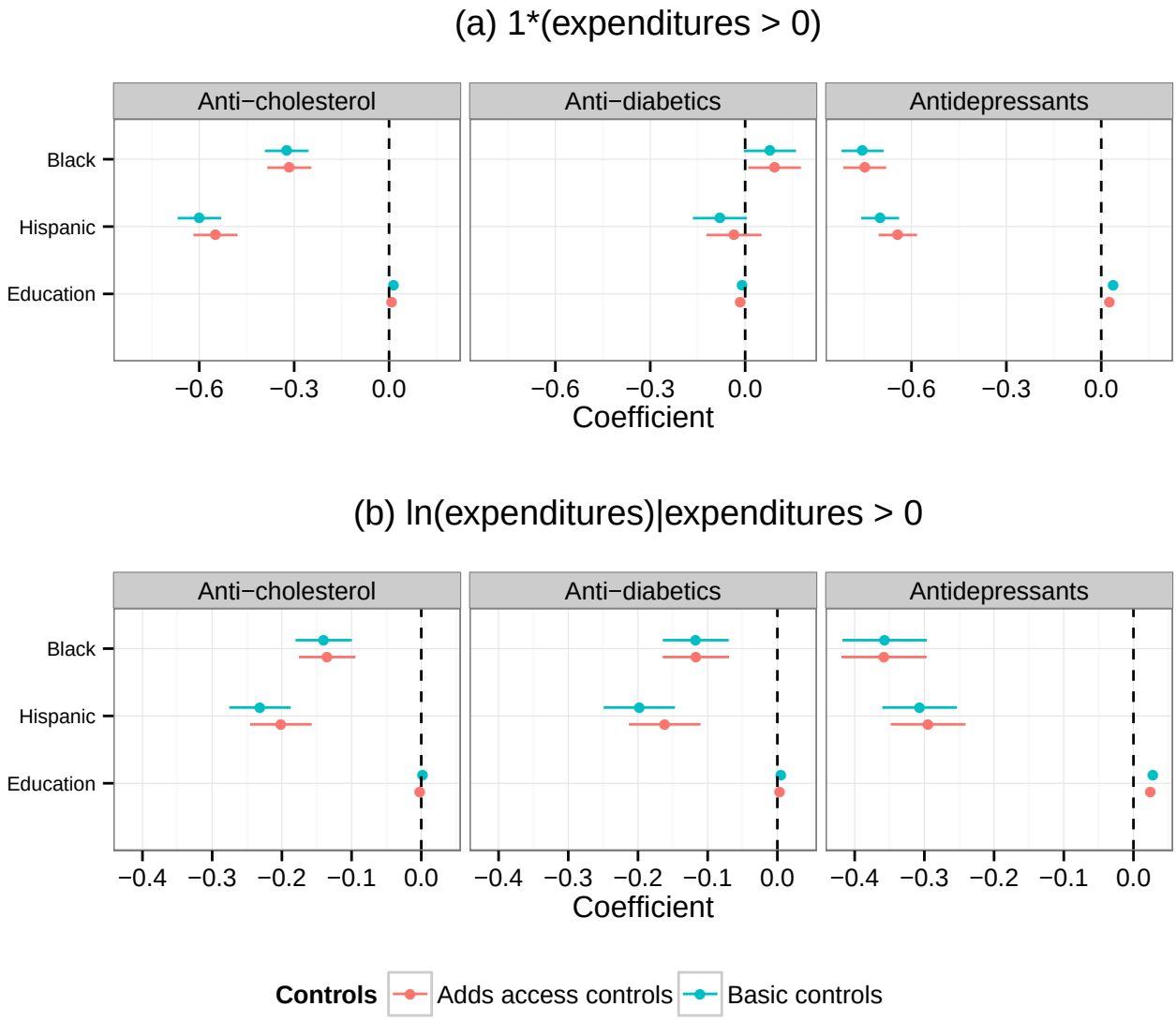


Figure 2.5: Regression Estimates of Disparities in Utilization for Common Drug Classes

Notes: Panel A and B are estimated with logistic and OLS regressions respectively. Basic controls are indicator variables for "other race", single year of age, self-reported health status, marital status, sex and year. Access controls are indicator variables for five income categories and health insurance status.

of schooling).

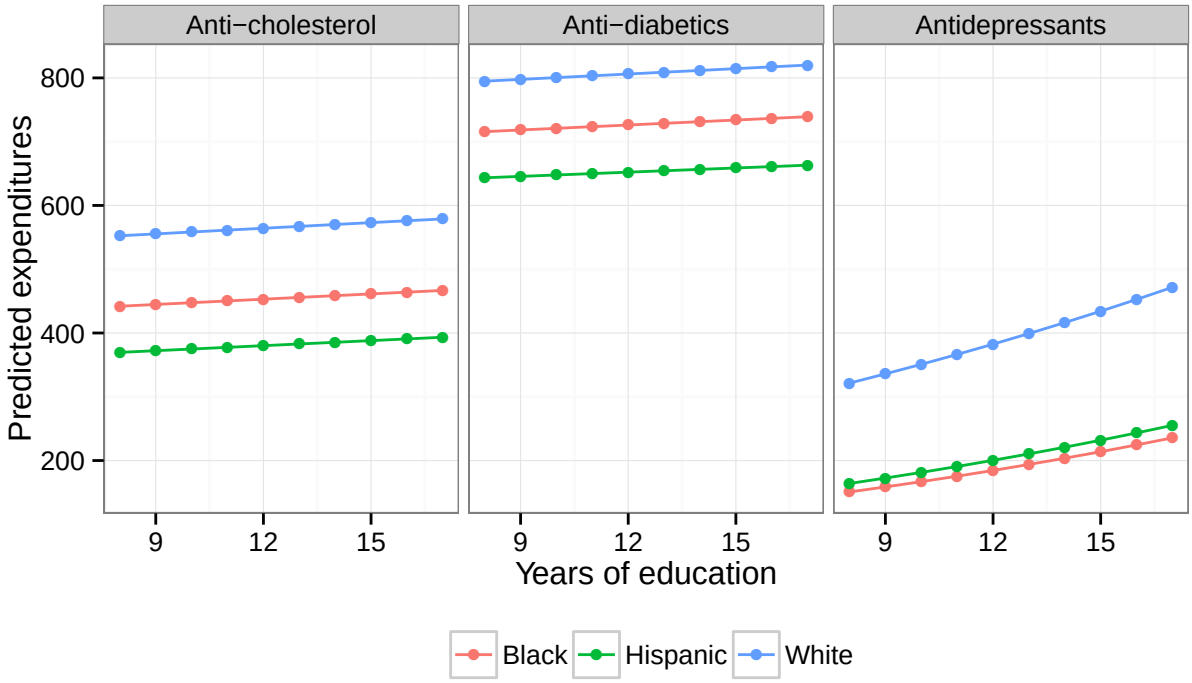


Figure 2.6: Predicted Mean Expenditures by Education, Race/Ethnicity, and Drug Class

Notes: Predicted expenditures are for an "average" individual. Predictions for each drug class are calculated using the two-part model and a Duan smearing estimator that varies by single year of age.

2.5 Discussion

2.5.1 Models for Specific Drug Classes

One drawback of this study is that it examines prescription drug expenditures on fairly aggregate levels in order to document larger patterns in the data. As a result, it cannot control all relevant clinical variables. One way forward is to focus on the utilization of drugs used to treat specific diseases and incorporate detailed clinical data if possible.

A more narrow focus would allow researchers to better understand the documented disparities in diagnoses. It would be useful to separate the gradients in diagnoses into (1) gradients in actual health conditions and (2) gradients in diagnoses conditional on those

health conditions. The second gradient is the true gradient in diagnoses and is likely underestimated in this paper because of well documented racial/ethnic and educational disparities in actual health conditions.

Analyses that focus on the treatment of specific diseases could also utilize discrete-choice models to analyze highly disaggregated consumption data. This approach would create a more uniform sample and allow researchers to detail heterogeneity in choices for drug types within a therapeutic class. For example, while this paper examines expenditures aggregated across diabetes drugs, a more disaggregated analysis could look at the specific drugs (biguanides, sulfonylureas, etc) that individuals use to treat diabetes.

2.5.2 Impacts on Health

The importance of the disparities documented in this paper depend on their impact on health, which is uncertain, but should be significant for a number of reasons. First, the disparities might have been caused by ineffective adherence to treatment protocols, which would be consistent with evidence from Goldman and Smith (2002) showing that more educated HIV and diabetes patients are more likely to adhere to therapy and less likely to switch to treatments that worsen health.[19] Second, if there are disparities in the timing of diagnoses (in addition to those at a given point in time), then racial/ethnic minorities and the less educated might begin taking prescribed medications at later stages of the disease process. This would likely increase disease severity and increase mortality rates, especially since research (Thomas et al. 1997) has shown that there are racial disparities in the onset of disease. Third, disadvantaged groups might adopt new medications more slowly, which is consistent with evidence showing that mortality rates for diseases treated with innovative technologies have declined more for the highly educated (Glied and Lleras-Muney 2008; Kinsey et al. 2008). There is also some evidence that these disparities are due to differences in learning (Lleras-Muney and Lichtenberg 2002) rather than knowledge, so educational

[19]There were also some racial differences in adherence, but they were less pronounced than the educational differences.

differences in the adoption of new technologies could be more pronounced than racial or ethnic differences.

2.5.3 Policy Implications

Provided that one is confident that the disparities found in this paper are not due to omitted variable bias and are actually detrimental to health, it is worth considering policy interventions that could reduce disparities. The success of these interventions will depend on identifying the mechanisms by which these disparities occur.

Differences in understanding about the benefits of treatment are one potential cause of the disparities documented in this paper. For instance, Whittle et al. (1997) has shown that race is not a significant predictor of the willingness to undergo revascularization procedures after controlling for familiarity with the procedures. A plausible way to improve familiarity is to improve communication between providers and patients, which pertains to another line of research suggesting that there are obstacles to effective communication between white physicians and black patients (Cooper-Patrick et al. 1999; Johnson et al. 2004). Another potential mechanism to improve understanding is by increasing health knowledge, which has been shown to be associated with disparities (Cutler and Lleras-Muney 2010).

Another reasonable cause of the observed disparities is that racial minorities have had negative experiences with healthcare in the past. Mistrust in the medical system is particularly plausible for African Americans, who were severely mistreated in the infamous Tuskegee Syphilis Study, in which researchers studied the natural progression of untreated syphilis in poor and mainly illiterate black men, but did not tell participants that they had syphilis or treat them for it despite known effective treatments.

Finally, there are other access related factors that could help explain the disparities in prescription drug use that this paper did not address. For instance, transportation costs might make it more costly for racial/ethnic minorities to visit providers, which could help account for the documented disparities in diagnoses.

More research is needed on each of these mechanisms so that policymakers and health professionals can optimize their efforts to help reduce disparities. To ensure that the policymakers target the correct mechanisms, researchers should analyze how these mechanisms vary with different medical treatments (i.e. prescription drugs vs surgery) and treatments for specific diseases (i.e. antidiabetics vs. anti-cholesterol medication). There is also a need for evaluations of specific interventions aimed at reducing racial and ethnic differences to determine both whether they are effective and which interventions are most effective.

2.6 Conclusion

This paper provides empirical evidence showing that there are racial/ethnic and educational disparities in prescription drug expenditures. The documented disparities occur for two primary reasons: first, there are differences in the probability of being diagnosed with a disease; and second, there are gradients in expenditures conditional on diagnosis. The racial/ethnic disparities exist both in the raw data and after controlling for variables proxying for clinical need. The education gradients exist after controlling for health variables, but not unadjusted. Access related factors account for almost none of the black-white differences in expenditures ($\approx 3\%$), a small fraction of the Hispanic-white differences ($\approx 14\%$), and a larger fraction of the education gradient ($\approx 34\%$). The disparities—especially those between racial and ethnic groups—are consistent across disease groups and the most common drug classes. Overall, a two-part model predicts that, on average, blacks and Hispanics spend \$350 and \$560 less than whites respectively and that an additional 4 years of education increases prescription drug expenditures by \$155.

Appendix

2.A Predicting Positive Rx Expenditures

Healthcare expenditures in the second component of two-part model's are often log transformed. This is done for a number of reasons. First, predictions in non-transformed models can be negative. Second, inferences are often sensitive to outliers. Third, the model is unlikely to be linear on the raw scale. Fourth, prediction intervals are more difficult to generate with heavily right skewed data.

Log transformed models are not without their own drawbacks though. The main problem is that predictions in levels are very sensitive to assumptions about the error term. If the error is not homoscedastic or normally distributed, then predictions can be very inaccurate. Table 2.8 illustrates this point. The table shows predicted mean expenditures and R^2 (measured as the square of the correlation between observed and predicted expenditures) for a number of different models. Each model includes the basic controls, the race/ethnicity indicator variables and the years of education variable.

The reference model is a simple OLS model on positive expenditures in levels. Predicted mean expenditures in this model are by definition equal to observed mean expenditures. The next three models are based on log transformed regression models. The first model assumes that the error term is normally distributed with a constant variance, σ^2, so that $E(y|x_2, y > 0) = \exp(x_2\beta + \sigma^2/2)$. The normal approximation is not unreasonable (see Figure 2.C.3), but predictions are still quite inaccurate. Mean expenditures are overpredicted by over \$600. Using the Duan smearing factor described in the text to allow for non-normality

Table 2.8: In-Sample Predictions of Nonzero Rx Expenditures

Model	Predicted mean expenditures	R^2
OLS on y	1431.97	0.09
OLS on ln(y) normal homoskedastic	2101.77	0.07
OLS on ln(y) Duan homoskedastic	1823.74	0.07
OLS on ln(y) Duan heteroskedastic	1559.78	0.08
Selection 2-Step on ln(y)	813.11	0.05
Selection MLE on ln(y)	2007.74	0.07

Notes: y refers to Rx expenditures. All models are estimated conditional on $y > 0$. R^2 is the squre of the correlation between y and $\hat{y}$ where $\hat{y}$ is predicted form the model. The selection 2-step assumes a normally distributed error in the second step. The selection MLE assumes a bivariate normal distribution for the error terms. All models include the basic control variables discussed in the text.

slightly improves predictions but mean expenditures are still overestimated. The largest improvement comes from a Duan smearing factor that varies by age: mean predictions only differ by around $100 from observed mean expenditures and the R^2 is higher than the R^2 of any of the other models except for the reference OLS model. In-sample predictions can be made more accurate by allowing the variance to vary by additional groups, but this increases the risk of overfitting. Predictions from the selection models are made assuming that the errors are normally distributed and homoscedastic.[20] Predictions from the selection model evaluated using maximum likelihood are similar to the predictions from the two-part model with normally distributed errors but the Heckman 2-step model is very inaccurate.

To ensure that the heteroscedastic Duan smearing estimator isn't overfitting the data, I used repeated 5-fold cross validation to test the out-of-sample performance of log OLS models on the non-transformed data. 5-fold cross validation partitions the data into 5 folds. Each round of cross validation estimates the model on 4 folds (the training data) and evaluates its performance on the remaining fold (the validation data). Each of the 5 folds is used as validation data one time so there are 5 total rounds. This procedure was repeated 10 times which yields a total of 50 out-of-sample tests of the model.

The out-of-sample prediction results mirror the in-sample results. The Duan smearing

[20] See, for instance, the formulas for the conditional and unconditional means in chapter 16 from Cameron et al. (2009).

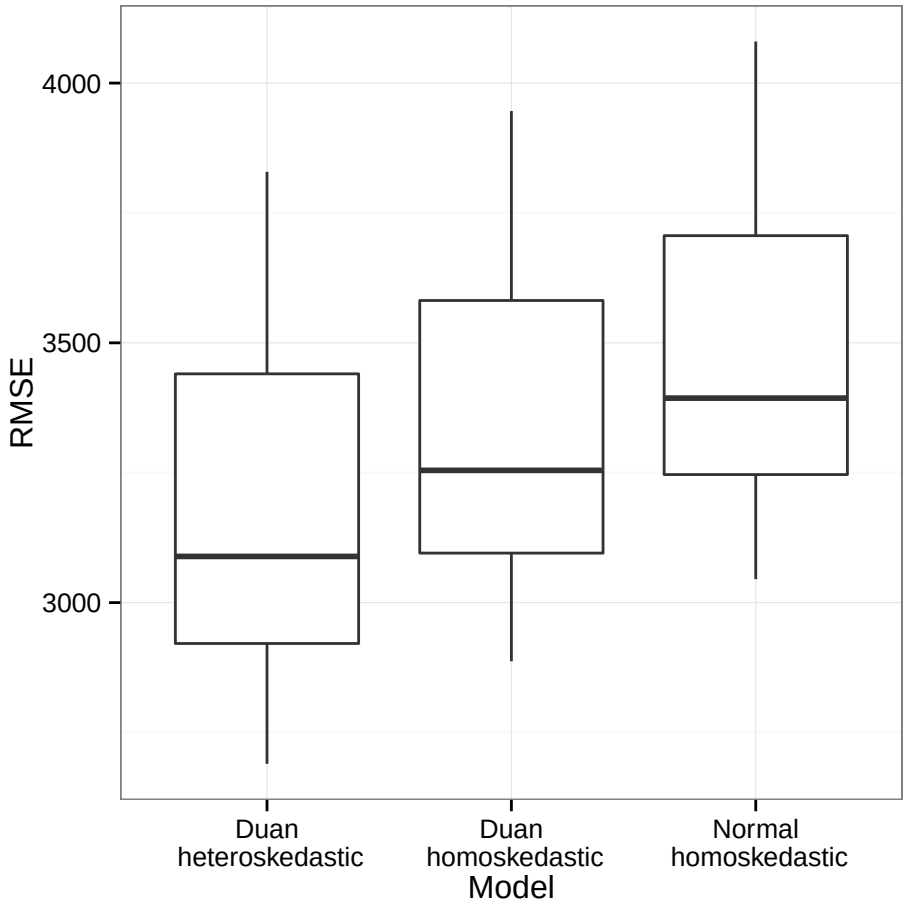


Figure 2.7: Results of Repeated 5-Fold Cross Validation

Notes: The figure is a box-and-whisker plot of the root-mean-square-error of 5-fold cross validation repeated 10 times. The RMSE is calculated for each model using 50 validation data-sets (5-fold cross validation and 10 repeats). All models include the basic control variables discussed in the text.

estimator with variances that vary by age has the lowest root-mean-square-error (RMSE).[21] Moreover, predictions with a Duan smearing estimator with constant errors continue to outperform predictions from a model assuming that the errors are normally distributed.

A number of recent studies have used generalized linear models (GLM's) to model health-care expenditure data (e.g. Blough, Madden and Hornbrook 1999; Manning, Basu and Mullahy 2005). A GLM with a gamma distribution and a log link function fit the data well when only basic controls were used. Predictions were generally similar to the OLS model

[21]The RMSE is $E\left[(y-\hat{y})^2\right]$ where y and $\hat{y}$ are observed and predicted expenditures respectively.

with a heteroscedastic Duan smearing factor. The gamma GLM assumes that the variance is proportional to the mean squared which is reasonable in some model specifications. However, after controlling for the 70 Rx condition categories the gamma model was no longer suitable. The errors become increasingly complicated as more clinical variables are added to the model, which is clearly an area for future research.

2.B Constructing the RX Disease Categories

The disease categories in this paper are based on the disease categories used by the CMS to risk adjust health plans participating in Medicare Part D. Each disease group consists of clinically and financially related five-digit ICD-9 codes that are treated with similar drugs. The CMS model does not include diagnostic categories when "the diagnoses were vague/nonspecific, discretionary in medical treatment or coding, not significant predictors of drug use, or transitory or not admitting of definitive treatment." The disease groups in the final CMS model are called hierarchical condition categories (RxHCCs) because related disease groups are clustered into hierarchies and ranked by severity (if an individual has multiple conditions within a hierarchy, then only the most severe disease group is included in the CMS model).

The disease classifications are taken from the MEPS Medical Condition Files. Unfortunately, only 3-digit ICD-9 codes are available in the public use data set for confidentially reasons, so I am unable to recreate the RxHCCs exactly. I consequently created disease groups in the following manner. First, I considered including all RxHCCs from the 2010 and all RxHCCs from 2014 (most recent CMS model) that were not in the 2010 model. Next, I examined the 5-digit ICD-9 codes within each RxHCC and determined whether it was sensible to classify them based on 3-digit ICD-9 codes.[22] In some cases the RxHCCs match the MEPS clinical classification codes (CCCs), which aggregate 5-digit ICD-9 codes into clinically meaningful categories, quite closely. In these cases I use the CCCs to approximate

[22] I considered it sensible to collapse 5-digit ICD-9 codes into 3-digit ones if a disease code contained the majority of the 5-digit codes within each 3-digit code.

the disease groups. For example, there is both a MEPS CCC and a RxHCC for congestive heart failure. The advantage of this approach is the CCCs are based on 5-digit, rather than 3-digit, ICD-9 codes. Finally, I considered adding additional CCCs that did not overlap with RxHCCs. These additional CCCs were only included it they were statistically significant in both components of the two-part model using basic controls.

Unlike the CMS model, I did not cluster the disease groups into hierarchies. I chose to add additional disease groups in the model even while their might have been significant overlap in treatments because the purpose of this paper is estimating racial/ethnic and educational gradients rather than prediction. The rational is that conditioning on additional diseases produces more conservative estimates of the gradients in expenditures conditional on diagnosis.

2.C Additional Tables and Figures

Table 2.C.1: Regression Coefficients from Two-Part and Selection Models

	Two-part		Selection two-step		Selection MLE	
	Part 1	Part 2	Part 1	Part 2	Part 1	Part 2
Black	−0.235	−0.310	−0.235	−0.184	−0.236	−0.287
	(0.007)	(0.010)	(0.007)	(0.012)	(0.007)	(0.011)
Hispanic	−0.428	−0.528	−0.428	−0.270	−0.427	−0.481
	(0.007)	(0.010)	(0.007)	(0.016)	(0.007)	(0.012)
Years of education	0.041	0.026	0.041	0.005	0.041	0.022
	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)

Notes: The two-part model replicates Table 2.4 using a probit regression for the first component rather than a logistic regression. The selection models use the same explanatory variables as the two-part model but allow the error terms to be correlated. Part 1 of the two-part and the two-step selection models are identical probit models so the coefficients are the same. Reported estimates are regression coefficients.

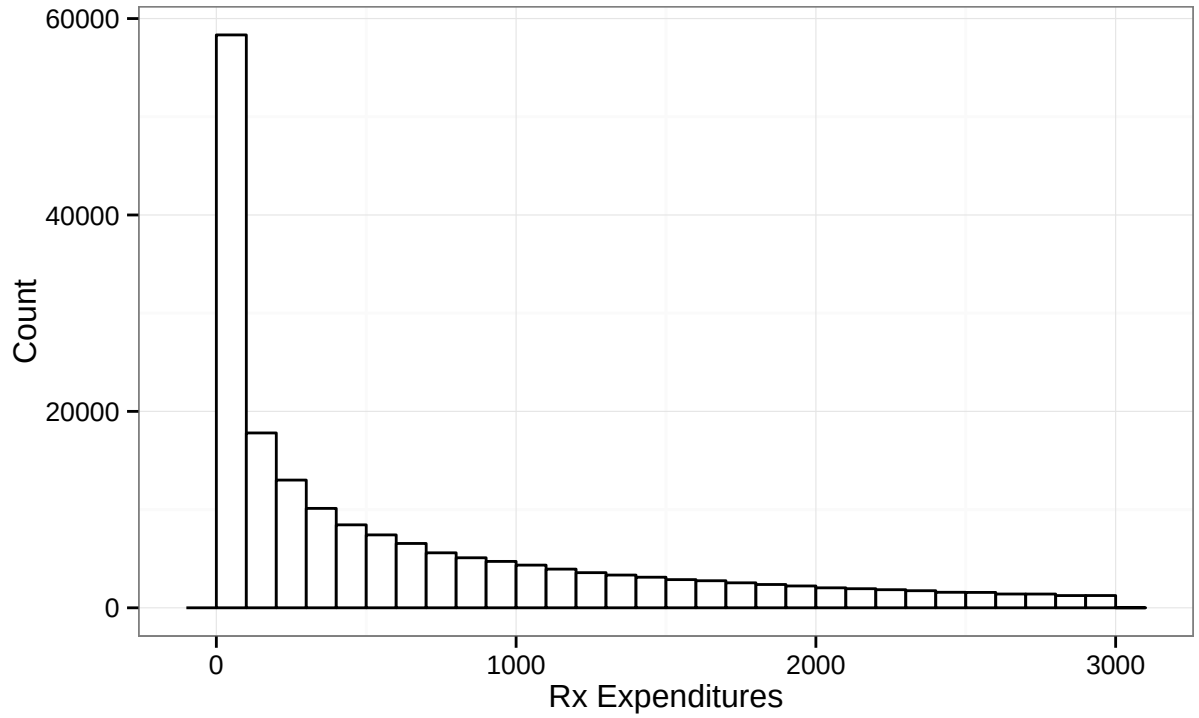


Figure 2.C.1: Histogram of Prescription Drug Expenditures Conditional on having an Rx Medical Condition

Notes: The distribution is truncated at $3000 for graphical purposes, although some expenditures are considerably higher.

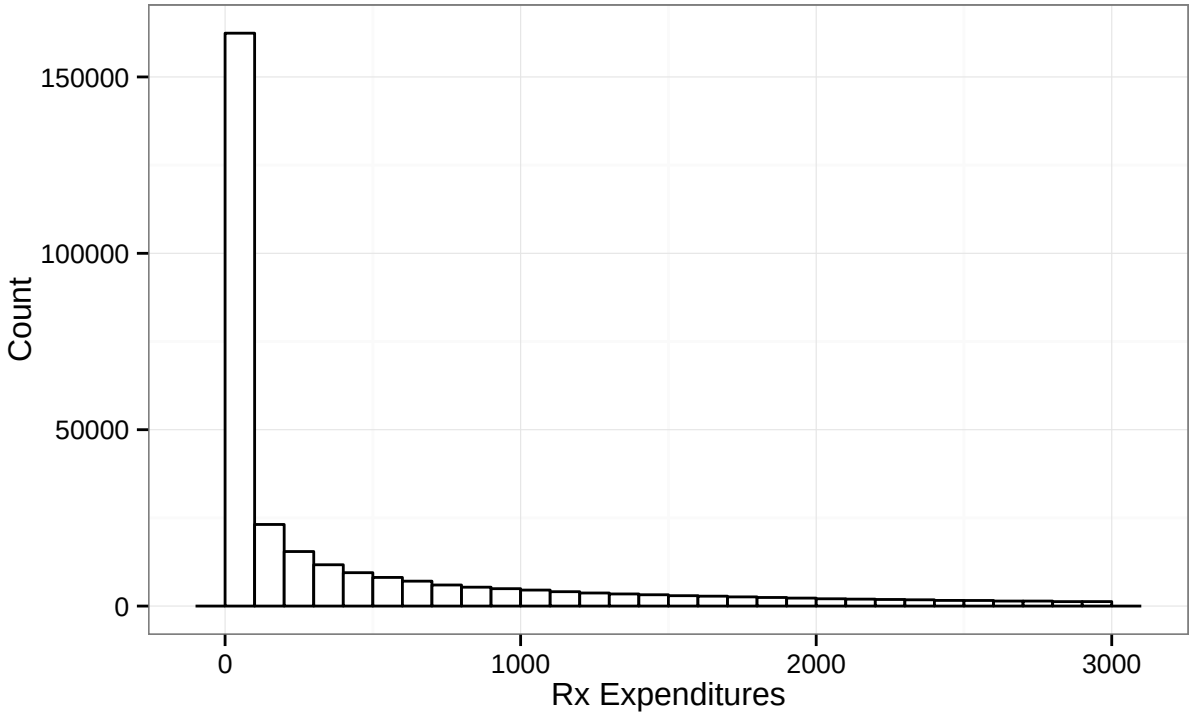


Figure 2.C.2: Histogram of Prescription Drug Expenditures

Notes: The distribution is truncated at $3000 for graphical purposes, although some expenditures are considerably higher.

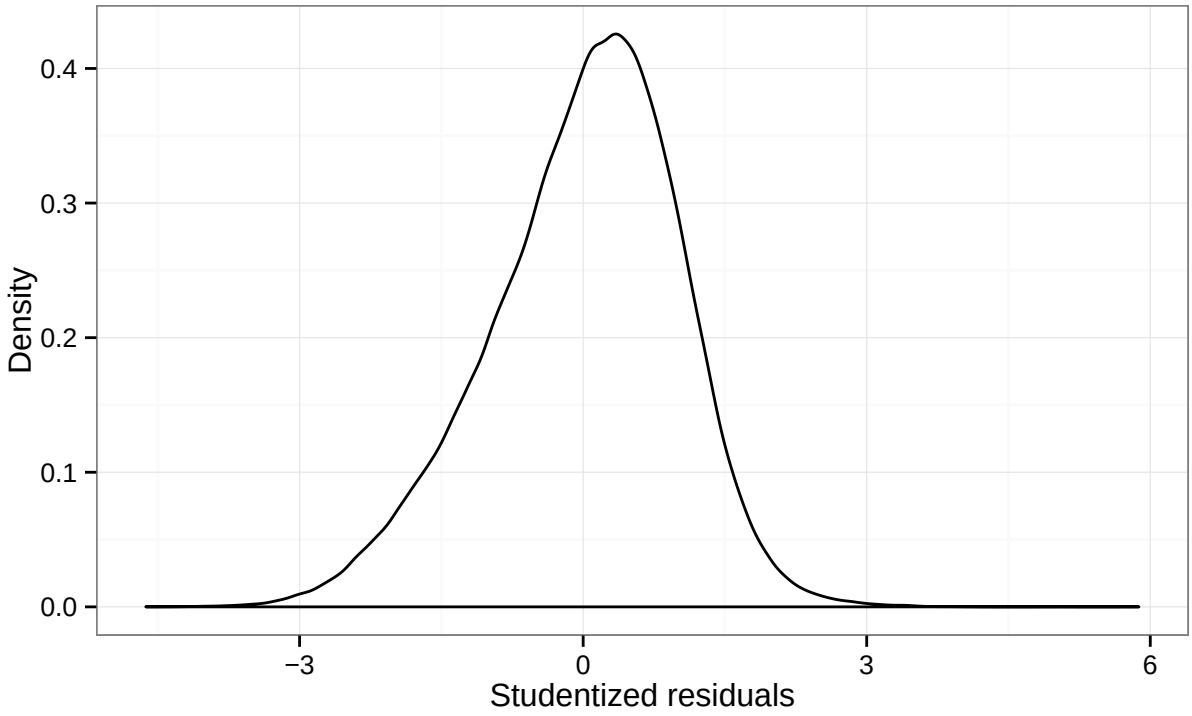


Figure 2.C.3: Density of Studentized Residual for Second Component of Two-Part Model for Continuous Component of Two-Part Model

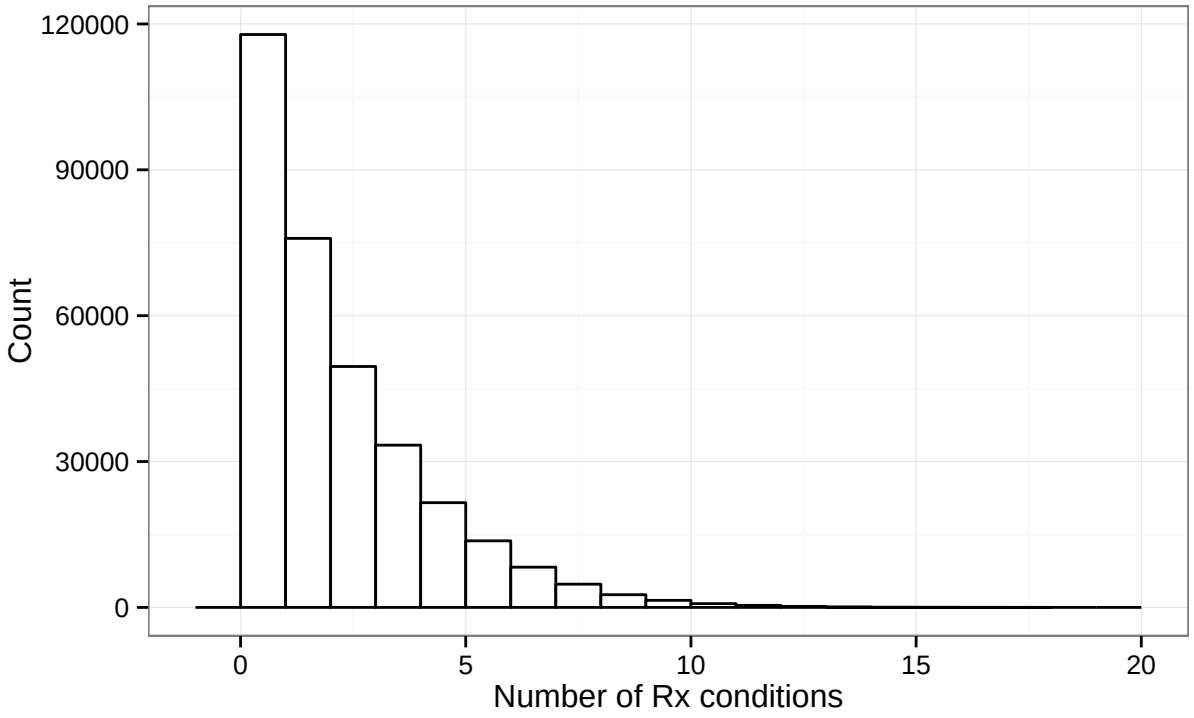


Figure 2.C.4: Histogram of the Number of Rx Conditions

Chapter 3

An Assessment of Long-term Healthcare Expenditure Risk using a Bayesian Two-Part Model

3.1 Introduction

Healthcare spending in the United States accounted for 17.4% of gross domestic product (GDP) in 2013 and is projected to reach 19.6% of GDP by 2024 (Centers for Medicare & Medicaid Services). These costs have placed increasing strains on U.S. firms–the primary sponsors of health insurance in the United States—as well as on government budgets. At the same time, individuals face rising premiums, lower wages and higher tax rates. Policy reforms that slow the rising cost of healthcare are consequently a top priority.

To shed light on cost containment efforts, I use dynamic Bayesian model to study 1) the long-term concentration of out-of-pocket healthcare spending and 2) whether out-of-pocket costs are predictable in the long run. This has important implications for two particular cost containment approaches. One approach is to use disease management programs, which provide organized, proactive care to patients with specific conditions, to help control costs associated with avoidable complications. These programs are the most likely to succeed

when costs are highly concentrated and when costs can be predicted accurately.[1] A second approach aims to reduce spending through consumer-directed health plans (CDHPs), which require consumers to use personal healthcare accounts to pay for non-catastrophic healthcare costs out-of-pocket. These plans have fewer tradeoffs when future costs are predictable because insuring against risk is less valuable. Furthermore, incentives for risk selection (by consumers and insurers) and the distributional implications of different cost sharing arrangements are directly related to the concentration of spending.

There is considerable evidence that spending varies significantly across individuals in a given year. Indeed, one of the most commonly cited statistics in American healthcare is that the top 5% of spenders account for approximately half of all U.S. healthcare spending (e.g. Zuvekas and Cohen 2007). There is also a large body of evidence suggesting that only a small percentage—at most 15 to 20 percent—of the variation in healthcare costs in a given year is predictable (e.g. Van Vliet 1992; Newhouse 2004; Pope et al. 2004; Breyer, Bundorf and Pauly 2012). Yet much less is known about the distribution of spending or the uncertainty of future health costs over extended periods. In this paper I use 8 waves (16 years) of the Health and Retirement Survey (HRS) to estimate these quantities over a significantly longer time period than in previous studies, which provides new evidence for evaluating the potential consequences of cost containment policies.

Modeling healthcare spending is not straightforward. For one, healthcare spending data is what is typically referred to as semicontinuous because a large portion of responses are equal to 0 while the remaining values follow a continuous, heavily right skewed distribution. Secondly, spending is persistent over time. To address these concerns I use a longitudinal two-part model (Olsen and Schafer 2001; Tooze, Grunwald and Jones 2002; Neelon, O'Malley and Normand 2011) with lagged dependent variables.[2] The two-part model addresses the

[1] Total spending—rather than out-of-pocket spending—is a more relevant cost measure for evaluating disease management programs. However, out-of-pocket and total costs are highly correlated so differences between the two measures should be small.

[2] Also see Zhang et al. (2006) for a Bayesian two-part hierarchical model in which patients are nested within physicians.

skewness of the data by modeling expenditures in two stages. In the first stage, a probit model is used to predict the probability of non-zero expenditures. Then, conditional on positive spending, a regression model predicts the log of expenditures. I account for the persistence in spending with past spending and unobserved individual effects. Since the probability of some use is likely correlated with the level of use, the individual effects are allowed to be correlated across the two equations.[3]

An additional complication is that long-term spending depends on life-expectancy. Since health spending tends to increase with age, mortality can have a large impact on future spending. Paradoxically, the sickest individuals can actually end up spending less over longer periods because they are more likely to die early in life. Thus, to forecast expenditures over part (or all) of an individual's lifetime it is necessary to forecast mortality as well. I consequently model expenditures and mortality jointly. The model is thus capable of simulating expenditures from any given date until death.[4]

The model highlights the importance of unobserved heterogeneity as it accounts for approximately 35 and 18 percent of the unexplained variation of the first and second stages of the two-part model. The correlation of the unobserved individual effects across equations is 0.516 which suggests that allowing the individual effects to be correlated is important. The model also exhibits strong state dependence as both lagged and initial values of the dependent variables are important predictors of future spending.

Out-of-pocket spending over longer periods is less skewed than spending in a single period, but a small fraction of the population still accounts for a large share of total health spending. For example, over a 16-year period, the top 5% of spenders still account for 38% of total spending. In simulations, the distributions become less and less skewed as the number of periods increase; however, even after 26 years the top 5% of spenders account for 28% of total

[3]Su, Tom and Farewell (2009) have shown that not accounting for this correlation can cause biased inferences.

[4]This makes the model suitable for state transition modeling (see Siebert et al. (2012) for an overview of state transition modeling and its use in cost-effectiveness analysis). The model can also be used to estimate lifetime out-of-pocket medical costs from a given age until death for households managing finances for retirement (see Webb and Zhivan (2010) for a related model used for this purpose).

spending. Most of the variation in spending is explained by random shocks and unobserved heterogeneity, which implies that individuals face substantial uncertainty in future health costs. As a result, it is difficult to predict high spenders and welfare calculations suggest large potential utility gains from using insurance to reduce financial risk.

This paper is related to a small collection of studies that have analyzed the dynamics of healthcare expenditures. Eichner, McClellan and Wise (1997) estimated a simple two-part model using 3 years of data and lagged dependent variables in order to gauge the feasibility of medical savings accounts. Feenberg, Skinner et al. (1994) and French and Jones (2004) estimate models with complex auto-correlated error structures on fairly long panels (6 and 8 years respectively) using GMM but do not account for the large point mass at zero in medical care data.[5]

My study differs from these articles in a number of important respects. First, I use a considerably longer panel than previous studies. Second, I use a model that captures both the cross-sectional (two-part model) and dynamic aspects (unobserved heterogeneity and lagged dependent variables) of medical care data. Third, I model both mortality and spending. Lastly, the model is estimated using Bayesian methods which allows me to form a complete probability distribution for out-of-pocket expenditures given any initial state or spending history.

The remainder of this article proceeds as follows. Section 3.2 describes the HRS data. Section 3.3 introduces the model. In Section 3.4, I outline the Bayesian algorithm used to estimate the model, summarize results, and check the fit of the model to the data. Section 3.5 analyzes the distribution and uncertainty of long-term spending, and then examines implications for cost-containment reforms. Finally, Section 3.6 concludes.

[5] Feenberg, Skinner et al. (1994) only observe individuals whose medical expenses are larger than 3% of income and attempt to correct this using a Tobit model; French and Jones (2004) recode all expenditures below $250 to $250.

3.2 Data

I use data from the 4th through 11th waves (1998 - 2012) of the Health and Retirement Survey (HRS). The HRS is a nationally representative longitudinal survey dataset of individuals over age 50. The main goal of the survey is to provide information for the study of health and retirement. It is the only nationally representative panel dataset containing information on health expenditures in the United States. I use the RAND HRS data files, which contain cleaned versions of the original variables that are renamed to allow for easier comparability across waves.[6] The main variable is total out-of-pocket expenditures which is only available from wave 4 onward.

To facilitate longitudinal analysis, the sample used in this study consists of all survey respondents who participated in the 4th wave, but not those who joined the survey during later waves. I also drop individuals from the sample if they have missing data due to non-response or dropout, which reduces the number of individuals in the sample from $20,569$ to $16,591$. An additional 8 individuals are dropped because information on race or ethnicity is unavailable. In Section 3.B, I show that restricting the data in this manner does not significantly alter the characteristics of the sample. Overall, this suggests that dropping observations has a minimal impact on the coefficient estimates and that the final sample is reasonably representative of the original sample.

The HRS data is self-reported so the reported expenditures are almost certainly measured with some error. Hurd and Rohwedder (2009) and Goldman, Zissimopoulos and Lu (2011) compare the HRS out-of-pocket spending estimates to those in the Medical Expenditure Panel Survey (MEPS) and the Medicare Current Beneficiary Survey (MCBS). Each survey spends considerable resources in order to collect high quality data: the MCBS combines survey data with administrative Medicare files while the MEPS combines survey data with data from the providers who provided care for the survey respondents. Both studies

[6] See Appendix 3.A for more details on RAND's cleaning of the data.

find that out-of-pocket spending estimates from the HRS are comparable (albeit slightly overestimated) to those in the MCBS but significantly higher than those in the MEPS, with discrepancies largest at the mean and the upper end of the distribution.

However, as noted by Marshall, McGarry and Skinner (2011), the HRS may capture expenses from sources—such as in-home care—that are not covered in other surveys. Furthermore, the MEPS is known to underreport expenditures relative to both the MCBS (Zuvekas and Olin 2009) and the National Health Expenditure Accounts (NHEA) (Keehan 2006; Bernard et al. 2012). In fact, as Marshall, McGarry and Skinner (2011) have pointed out, the per-year sample average of out-of-pocket spending in the 2004 HRS for the over age 65 population, or approximately $2100, is nearly equivalent to the 2002 and 2004 estimates of over age 65 out-of-pocket spending from the NHEA as reported in Hartman et al. (2008). In a similar fashion, French and Jones (2004) has shown that average expenditures from the HRS tend to match US averages closely.

Table 3.1 reports summary statistics on the distribution of out-of pocket expenditures in the HRS. Expenditures reported in the table and the rest of the paper are in $2012. The first row reports summary statistics from the HRS pooled across the 4th through 11th waves. Mean out-of-pocket expenditures are around $2100 per year, which is consistent with the NHEA estimates reported in Hartman et al. (2008) and discussed above. The distribution is heavily right skewed with a mean considerably larger than the median and a very large maximum value of over $1 million. The percentage of non-spenders is also quite high, which suggests that modeling out-of-pocket expenditures using a two-part model is warranted.

The second and third rows each look at the distribution of spending averaged (by individual) across the 4th through 11th waves. The second row is the sum of expenditures across all waves divided by the number of waves in the data, or 8. Out-of-pocket expenditures are coded as $0 in all years after death. This quantity is the quantity that a public payer such as Medicare or Medicaid (or perhaps a family saving for retirement) would be concerned with. The distribution is less skewed than the distribution of the pooled cross section: the mean

Table 3.1: The Distribution of Out-of-Pocket Health Expenditures in the HRS, Waves 4 - 11

Data	Sample size	% nonspenders	Mean	Quantile 25%	Quantile Median	Quantile 75%	Max
Pooled cross-section	99,564	10.93	4,261	456	1,518	3,840	1,539,869
Individual means	16,583	3.49	3,198	658	1,784	3,723	213,088
Individual means per wave alive	16,583	3.49	4,890	1,041	2,395	4,968	292,616

Notes: Estimates are two-year (or single wave) averages. Individual means are expenditures summed across all 8 waves divided by 8 (individuals are coded as having $0 expenditures in all waves after death). Individual means per wave alive are expenditures averaged across all waves in which an individual is alive.

is lower, the median is larger, the maximum is substantially smaller, and there are very few non-spenders. Nonetheless, the distribution is still quite skewed with a mean considerably higher than the median. At 95th percentile, an individual spends $\$86,361$ over 16 years. Spending at the median is much lower, but the median individual still pays an economically significant $\$14,274$ out-of-pocket.

Individuals might be more concerned with how much they would pay when they are alive. The third row consequently provides summary statistics for mean out-of-pocket expenditures per wave alive. Mean spending in the third row is higher than mean spending in either of the first two rows because individuals tend to spend more right before death. At the 95th percentile, individuals spend, on average, $\$130,226$, per wave alive. Spending at the 25th percentile and the median are also high, which suggests that a substantial fraction of the population experiences large out-of-pocket expenditures over their lifetimes later in life.

Table 3.1 showed that nonzero spending is quite common in a single wave but much less prevalent over multiple waves. Figure 3.1 investigates this more closely. The leftmost figure plots the cumulative share of individuals with nonzero spending by wave; that is, the fraction of (surviving) individuals with nonzero spending in either the current wave or any previous wave. In wave 4, 87% of individuals had positive expenditures. The cumulative share of individuals with nonzero expenditures then rose to 94% by wave 5 and steadily increased thereafter before surpassing 99% by wave 9. This means that nearly everyone who survived

at least 6 waves had positive medical spending during at least one of those two-year periods.[7]

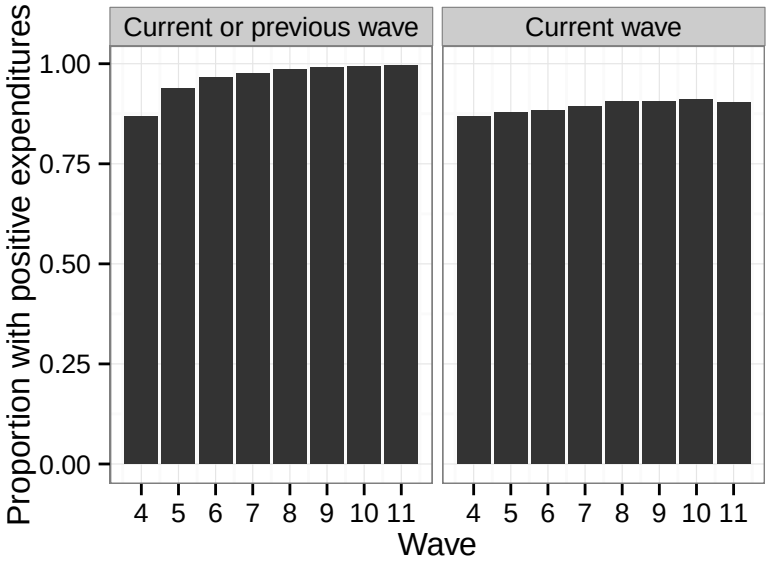


Figure 3.1: Nonzero Spending in the HRS

Notes: Each wave only includes survey respondents who survived to that wave.

The rightmost figure plots the fraction of nonzero spenders in each wave. This quantity increases slightly over time due to population aging. In wave 4, 87% of individuals had positive spending; by wave 11, 90% of survey respondents were nonzero spenders.

Table 3.1 showed that expenditures remain skewed even after averaging over 8 waves. This is consistent with a number of other studies that have shown that medical expenditures are highly persistent. Figure 3.2 examines persistence in the HRS data by looking at whether expenditures in wave 4 are correlated with expenditures in subsequent waves. More precisely, it plots mean out-of-pocket expenditures by expenditure quartile in wave 4.

During the initial wave an individual in the highest quartile spent, on average, almost $7483 while those in the lowest quartile spent, on average, only $96. Even mean spending in the third quartile was less than one-third mean spending in the highest quartile. But

[7]Note that the cumulative share of nonzero spending by wave 11 is higher than the fraction of individuals with nonzero spending reported in rows 2 and 3 of Table 3.1. This is because the statistic reported for wave 11 in Figure 3.1 is estimated using only those who survived from wave 4 to wave 11 while the statistic in Table 3.1 is based on all individuals regardless of their survival length.

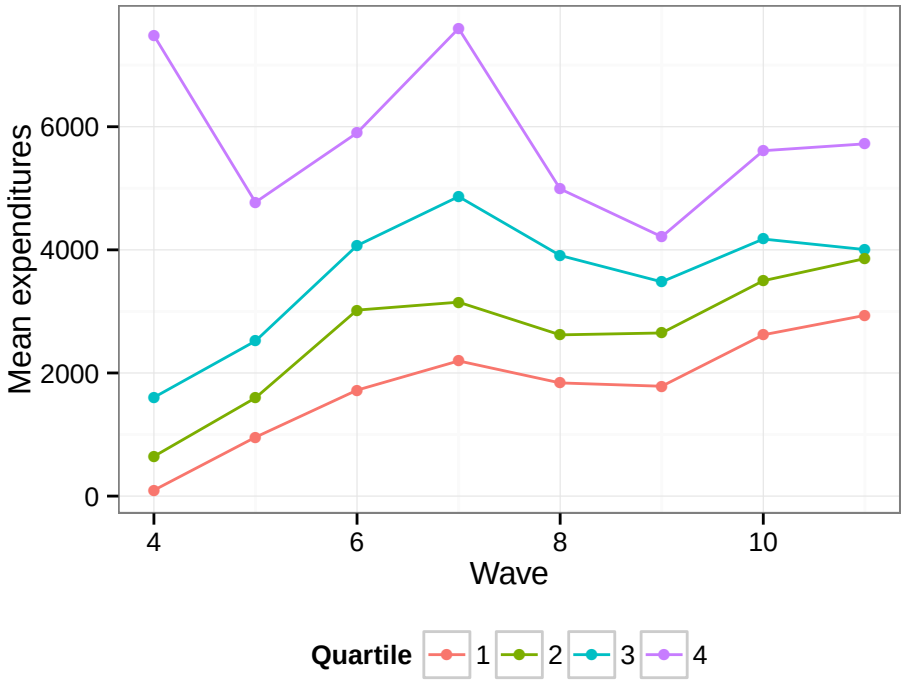


Figure 3.2: Mean Out-of-Pocket Expenditures by Wave 4 Quartile

expenditures tend to gravitate toward the mean in all quartiles. Spending in the bottom three quartiles rose over time while spending in the highest quartile declined. By the 11th wave mean spending in the lowest quartile rose to nearly one-half mean spending in the highest quartile. Nonetheless, the lines never cross which implies that, on average, those that had higher expenditures in wave 4 continued to have higher expenditures in subsequent waves. There is thus both significant mean reversion and persistence in the data.

3.3 The Model

This paper models out-of-pocket expenditures for individuals over age 50. Expenditures are modeled in two-parts with separate equations for nonzero expenditures and log expenditures conditional on positive expenditures. Spending is modeled as a function of exogenous personal characteristic, lagged spending, unobserved heterogeneity and transitory errors. Since

the model includes lagged variables, it is estimated on the sample of $15,273$ individuals who survived to at least wave 5.

In order to forecast expenditures over an individual's lifetime it is necessary to forecast their mortality as well. Mortality is modeled using a probit equation as a function of out-of-pocket spending in the previous period and exogenous personal characteristics. The probit model is estimated using the sample of $15,273$ individuals based on their survival from wave 6 until wave 11.

3.3.1 Expenditure Model

Let y_{it} denote out-of-pocket expenditures where the indices i and t refer to individual i and period t respectively—for clarity, the initial wave (wave 4) is referred to as period 0. Furthermore, let $d_{it} \equiv I(d_{it}^* > 0) = I(y_{it} > 0)$ so that the latent variable d_{it}^* describes whether expenditures are positive or zero. The dynamic longitudinal two-part model can then be written as,

$$d_{it}^* = \phi_1 d_{it-1} + z_{1it}^T \gamma_1 + c_{1i} + \epsilon_{1it}, \tag{3.1}$$

$$\ln y_{it} | d_{it}^* > 0 = g(y_{it-1})^T \phi_2 + z_{2it}^T \gamma_2 + c_{2i} + \epsilon_{2it}, \tag{3.2}$$

where γ_1 and γ_2 are the vectors of coefficients on the explanatory variables; c_{1i} and c_{2i} are individual random effects; and ϕ_1 is the coefficient on the lagged dependent variable in the first part of the model. The function $g(\cdot)$ is a column vector that allows the observed response in the second part of the model to depend on lagged values in a number of ways. ϕ_2 is an $m \times 1$ vector where m is the number of variables contained in $g(y_{it-1})$. As is typical of two-part models, it is assumed that $\epsilon_{1it} \sim N(0,1)$ so that the first equation is a probit model and that $\epsilon_{2it} \sim N(0, \sigma_\epsilon^2)$. As will be shown later, the data generating process is well approximated by a lognormal distribution for nonzero y_{it} so assuming that ϵ_{2it} is normally distributed is not problematic. If this were not the case, the model could be extended so that the distribution of ϵ_{2it} were a student-t distribution, a skew normal distribution or some

mixture of normals.[8]

A well documented problem with dynamic non-linear models with individual random effects is that the estimates depend on assumptions about the initial conditions [see Hsiao (2014, section 7.5)]. Standard assumptions are that the initial conditions are either fixed or random, but neither is without its limitations. Assuming that an initial condition is fixed implicitly assumes that it is uncorrelated with the individual effects. On the other hand, it is difficult to properly specify the density of the initial conditions given the individual effects. In this paper I use an alternative approach suggested by Wooldridge (2005) in which the distribution of the individual effects is modeled as a function of initial conditions and exogenous variables.[9] Specifically,

$$c_{1i} = \delta_{10} + \delta_{11} d_{i0} + z_{1i}^T \delta_{12} + b_{1i} \tag{3.3}$$

$$c_{2i} = \delta_{20} + h(y_{i0})^T \delta_{21} + z_{2i}^T \delta_{22} + b_{2i}, \tag{3.4}$$

where d_{i0} and y_{i0} are initial conditions and z_{1i} and z_{2i} are column vectors of time invariant explanatory variables for individual i. The role of the function $h(\cdot)$ is the same as the function $g(\cdot)$ above in that it allows the initial conditions to appear in a number of ways. Variables that can be included in z_{1i} and z_{2i} include time-constant variables such as race and sex and averages (over time) of the time-varying variables. Based on evidence from previous two-part longitudinal models, the individual specific error terms, b_{1i} and b_{2i}, are allowed to be correlated across the two equations. That is, they are given a bivariate normal distribution,

$$b_i = \begin{bmatrix} b_{1i} \\ b_{2i} \end{bmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_b = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right], \tag{3.5}$$

[8] See chapter 15 in Koop, Poirier and Tobias (2007) for details on estimating these types of non-normal models.

[9] My model is therefore similar to Li and Zheng's 2008 Bayesian semi-parametric dynamic type I Tobit panel data model which also follows Wooldridge's 2005 suggestion (their model is semiparametric because it models the distribution of unobserved individual effects as a mixture of normals.)

where ρ is the correlation between b_{1i} and b_{2i}. In this paper, I refer to b_{1i} and b_{2i} as unobserved heterogeneity or unobserved individual effects.

The model is assumed to be dynamically complete which implies that there is no serial correlation in the error terms after controlling for the lagged dependent variables. The observed persistence in out-of-pocket spending is consequently modeled through the lagged dependent variables and unobserved heterogeneity. The error terms are what is often referred to as equicorrelated because the co-variance in the errors between one wave and the next are equal to the unobserved individual effects and do not vary with the time difference between waves.

Variables included in the model are those that would be known to an individual at the initial wave and could be used to forecast spending over long periods of time given this initial information. The time-varying vectors z_{1it} and z_{2it} contain a quadratic function of age.[10] The time-constant vectors z_{1i} and z_{2i} consist of standard variables for demographics and socioeconomic status. Both age and a variable for years of education are centered to make the estimate of the intercept more meaningful. The functions $g(\cdot)$ and $h(\cdot)$ were chosen to best approximate the relationship between lagged (and initial) log spending and current log spending. As illustrated in figures 3.E.1 and 3.E.2 in the appendix, the effect of lagged (initial) expenditures should differ depending on whether previous (initial) expenditures were zero or strictly positive. Accordingly, I set $g(y_{it-1}) = (d_{it-1}, d_{it-1} \times \ln y_{it-1})^T$ and $h(y_{i0}) = (d_{i0}, d_{i0} \times \ln y_{i0})^T$ so that current log expenditures are predicted by an indicator variable if spending was zero in the previous (initial) period and lagged (initial) log expenditures if lagged (initial) spending was positive. Sample quantiles for the explanatory variables used in the model are provided in Table 3.2.

It is worth mentioning that the two-part model assumes that ϵ_{1it} and ϵ_{2it} are uncorrelated. An alternative approach is to use a type 2 Tobit or Heckman selection model to jointly model the errors. I believe that this would likely only add to the complexity of the current model

[10] Additional time-varying variables could be included but these would not be known to an individual at the initial wave and would consequently need to be forecasted.

Table 3.2: Explanatory Variables Included in Model

Variable	Variable included in?			Sample Quantiles		
	Expenditure model		Mortality model			
	Binary component	Continuous component		Min	Median	Max
Intercept	Yes	Yes	Yes	1.0	1.0	1.0
$D_{it-1} \times \ln y_{it-1}$	No	Yes	Yes	0.0	7.3	14.2
D_{it-1}	Yes	Yes	Yes	0.0	1.0	1.0
$D_{i0} \times \ln y_{i0}$	No	Yes	No	0.0	7.0	12.6
D_{i0}	Yes	Yes	No	0.0	1.0	1.0
$(Age - 65)/10$	Yes	Yes	Yes	-1.5	0.5	4.4
$[(Age - 65)/10]^2$	Yes	Yes	Yes	0.0	0.5	19.4
Female	Yes	Yes	Yes	0.0	1.0	1.0
Black	Yes	Yes	Yes	0.0	0.0	1.0
Other race	Yes	Yes	Yes	0.0	0.0	1.0
Hispanic	Yes	Yes	Yes	0.0	0.0	1.0
Years of education - 12	Yes	Yes	Yes	-12.0	0.0	5.0

with little added benefit, but future work might want to consider extending the model to allow for dependence in the error terms.

3.3.2 Mortality Model

One approach to modeling mortality is to generate survival curves using a hazard/survival modeling approach. Parametric proportional hazards models based on the Weibull or Gompertz distributions are natural candidates. However, in this paper, I use a model appropriate for discrete time since survival times are aggregated into two-year periods in the HRS.

Binary choice models are a simple and common choice for grouped survival data (Cox 1972; Lynch and Brown 2005; Cameron and Trivedi 2005). Here, I use a probit model but a logistic model could be used instead. It is convenient for Bayesian analysis to write the model in terms of a latent variable, say m_{it}^*, which can thought of as representing the risk of death. Then, letting $m_{it} = I(m_{it}^*) > 0$, the model can be written as,

$$m_{it}^* = x_{Mit}^T \kappa + \epsilon_{Mit}, \tag{3.6}$$

where x_{Mit} is the vector of explanatory variables for individual i, κ is the corresponding

coefficient vector and $\epsilon_{Mit} \sim N(0,1)$. The observed death indicator, m_{it}, is equal equal to 0 in all periods except the period of death in which case it is equal to 1. Individuals are dropped from the estimation sample following the period in which they died. The data vector x_{Mit} contains the same function of lagged expenditures from equation 3.2 and the same time constant variables from the expenditure model. All variables are summarized in Table 3.2.

3.4 Bayesian Estimation

3.4.1 Priors

In the Bayesian approach, prior distributions are assigned to all of the model parameters. To ensure a proper posterior that is determined almost entirely by the data, weakly informative prior distributions were chosen for the hyperparameters. The regression coefficients (γ_1, γ_2, δ_{10}, δ_{11}, δ_{12}, δ_{20}, δ_{21}, δ_{22}, κ) and the coefficients on the lagged dependent variables (ϕ_1, ϕ_2) are given diffuse normal priors. The precision of the second equation of the two-part model, σ_ϵ^{-2} is assumed to have a $\text{Ga}(\text{a}_0, \text{b}_0)$ conjugate prior. As suggested by Neelon, O'Malley and Normand (2010) and Neelon, O'Malley and Normand (2011) the covariance matrix of the unobserved individual effects, Σ_b, is assumed to have a conjugate inverse-Wishart $IW(S_0, v_0)$ distribution. To make the prior uninformative, I set the degrees of freedoms, v_0, equal to 3 and S_0 equal to the identity matrix.

3.4.2 Posterior Computation

In the expenditure model, substitute equations 3.3 and 3.4 into equations 3.1 and 3.2 respectively so that the *ith* individual during period t has the vectors of explanatory variables $x_{1it} = (d_{it-1}, z_{1it}, d_{i0})$ and $x_{2it} = (g(y_{it-1}), z_{2it}, h(y_{i0}))$.[11] Denote the corresponding coefficient vectors as α and β respectively. In addition let expenditure data be available through time T_i and $T_i^{'}$ be either the the period of death or the final wave of the survey. That is, $T_i^{'} = T_i$ for

[11]The vectors z_{1it} and z_{2it} therefore include both time-varying and time-invariant variables.

those who survived all 8 waves and $T_i^{'} = T_i + 1$ for those who died during the survey period. Letting $\theta = (\alpha^T, \beta^T, \sigma_\epsilon^2)^T$, the conditional density, $f(y_{i1}, y_{i2}, \ldots, y_{iT}|y_{i0}, b_i, \theta)$ for individual i can then be written as,

$$\begin{aligned}\prod_{t=1}^{T_i} f(y_{it}|y_{it-1}, b_i, \theta) &= \prod_{t=1}^{T_i}(1 - \Phi(\mu_{1it})^{1-d_{it}} \left[\Phi(\mu_{1it}) \times \mathrm{LN}(\mathrm{y_{it}}; \mu_{2\mathrm{it}}, \sigma_\epsilon^2)\right]^{d_{it}}, \\ \mu_{1it} &= x_{1it}^T\alpha + b_{1i}, \\ \mu_{2it} &= x_{2it}^T\beta + b_{2i}, \end{aligned} \tag{3.7}$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution and $LN(\cdot)$ is the lognormal distribution. For individual i, the conditional density for mortality is just the standard likelihood for a binary choice model,

$$\prod_{t=2}^{T_i^{'}} p(m_{it}|\kappa, y_{it-1}) = \prod_{t=2}^{T_i^{'}} \Phi(x_{Mit}^T\kappa)^{m_{it}} \left[1 - \Phi(x_{Mit}^T\kappa)\right]^{1-m_{it}}, \tag{3.8}$$

except that m_{it} can only equal 1 if $t = T_i^{'}$.

Given prior distributions $p(\theta)$, $p(\kappa)$ and $p(\Sigma_b)$, it follows that the full joint posterior is,

$$\begin{aligned} p(\theta, \kappa, b_i, \Sigma_b|y, m) \propto\ & p(\theta)p(\Sigma_b)\prod_{i=1}^{n}\prod_{t=1}^{T_i} f(y_{it}|y_{it-1}, b_i, \theta)p(b_i|\Sigma_b) \\ & \times p(\kappa)\prod_{i=1}^{n}\prod_{t=2}^{T_i^{'}} p(m_{it}|\kappa, y_{it-1}), \end{aligned} \tag{3.9}$$

where y is the the stacked vector of y_{it}, m is the stacked vector of m_{it}, and there are n individuals. A Gibbs sampling algorithm is used to estimate the parameters by partitioning the joint posterior into conditional distributions. The parameters of the expenditure and mortality models are simulated separately. κ can be sampled using the standard normal regression model by using the data augmentation technique described in Albert and Chib (1993).

Convenient priors were chosen for the expenditure model so sampling from β, σ_ϵ^2 and Σ_b is straightforward. α can be sampled using the same data augmentation technique used to sample κ. However, the conditional distribution for b_i is nonstandard so it is sampled using a random-walk Metropolis step. Full details of the Markov Chain Monte Carlo (MCMC) algorithm used to sample the parameters in the expenditure model are provided in Appendix 3.C.

The algorithms were implemented using R version 3.1.2. Inferences for the mortality equation were based on $2,100$ draws of κ after a burn-in period of $1,000$ iterations simulated using the *MCMCprobit* command in the R package *MCMCpack*.[12] Estimation of the expenditure equation was more time consuming. 7 initially dispersed chains of $65,000$ iterations were run in parallel. The first $50,000$ iterations were discarded and the sequence was thinned by only keeping every 50th draw after burn in, yielding $2,100$ random draws from the posterior distribution. Run-time on a ThinkPad W530 Mobile Workstation with 8 cores is approximately 7 hours. Convergence was assessed using the R package coda by visually inspecting traceplots of all the parameters (which are available upon request) and using the Gelman-Rubin diagnostic. The Gelman-Rubin potential scale reduction factor $\hat{R}$ was less than 1.1 for all parameters suggesting approximate convergence.

3.4.3 Parameter Estimates and Model Summary

Table 3.3 presents posterior means, standard deviations, and quantiles for the parameters of the expenditure model.[13] Unsurprisingly, the model predicts that women spend more out-of-pocket than men. Another expected result is that out-of-pocket costs are predicted to increase with age. However, the predicted effect of age on the probability of nonzero expenditures decreases as individuals get older while the predicted effect on positive out-of-pocket expenditures increases with age. The coefficients on the racial/ethnic and education variables are consistent with research on SES differentials in the utilization of healthcare

[12] Estimates were nearly identical to those from maximum likelihood estimation. Run time was less than 1 minute and traceplots suggested that the parameters successfully converged.

[13] The posterior distributions of the regression coefficients are all approximately normal so the posterior means are nearly identical to the posterior medians.

services (e.g. Blanco et al. 2007; Cook, McGuire and Miranda 2007; Gross et al. 2008; Martinez et al. 2008; Cook and Manning 2009), which typically finds that variables correlated with low socioeconomic status are associated with less medical spending.[14]

Table 3.3: Parameter Estimates for Expenditure Model

			Posterior Quantiles		
	Mean	SD	2.5%	Median	97.5%
Binary component covariates					
Intercept	0.49	0.03	0.43	0.49	0.56
D_{it-1}	0.60	0.02	0.56	0.60	0.65
D_{i0}	0.94	0.03	0.88	0.94	1.00
$(Age-65)/10$	0.08	0.02	0.05	0.08	0.11
$[(Age-65)/10]^2$	−0.07	0.01	−0.09	−0.07	−0.06
Female	0.09	0.02	0.05	0.09	0.13
Black	−0.43	0.03	−0.49	−0.43	−0.38
Other race	−0.27	0.05	−0.37	−0.27	−0.16
Hispanic	−0.44	0.04	−0.51	−0.44	−0.36
Years of education - 12	0.07	0.00	0.06	0.07	0.08
Continuous component covariates					
Intercept	6.49	0.03	6.44	6.49	6.54
$D_{it-1} \times \ln y_{it-1}$	0.24	0.00	0.23	0.24	0.25
D_{it-1}	−1.34	0.04	−1.42	−1.34	−1.26
$D_{i0} \times \ln y_{i0}$	0.20	0.00	0.19	0.20	0.21
D_{i0}	−0.94	0.04	−1.02	−0.94	−0.85
$(Age-65)/10$	0.10	0.01	0.08	0.10	0.12
$[(Age-65)/10]^2$	0.03	0.00	0.02	0.03	0.04
Female	0.04	0.01	0.01	0.04	0.06
Black	−0.15	0.02	−0.19	−0.15	−0.11
Other race	−0.13	0.04	−0.20	−0.13	−0.05
Hispanic	−0.08	0.03	−0.14	−0.08	−0.03
Years of education - 12	0.02	0.00	0.02	0.02	0.03
Variance parameters					
σ_ϵ^2 (lognormal variance)	1.30	0.01	1.29	1.30	1.32
σ_1^2 (var$[b_{1i}]$)	0.53	0.03	0.48	0.53	0.58
σ_2^2 (var$[b_{2i}]$)	0.29	0.01	0.27	0.29	0.30
ρ (corr$[b_{1i}, b_{2i}]$)	0.52	0.02	0.48	0.52	0.56

Notes: Parameter estimates are from the model described in equations 3.1 and 3.2. d_{it-1} and d_{i0} are indicator variables for whether there was positive spending in the previous and initial wave respectively. Likewise, $\ln y_{it-1}$ ($\ln y_0$) is the natural logarithm of expenditures in the previous (initial) wave.

The expenditure simulations described in the next section depend crucially on the coef-

[14]Mediating variables like insurance status and income are not included so economic factors are allowed to contribute to these differences. Health variables are not included in the reported results either but the SES differentials are even more pronounced after controlling for self-reported health status.

ficients on the lagged dependent variables and on the variance parameters. The impact of lagged spending in both parts of the model is fairly strong. In the binary component, d_{it-1} is predicted to increase d_{it}^* by 0.6 at its posterior mean. In the continuous component, each percentage point increase in lagged spending is predicted to increase current spending by 0.24 percentage points at its posterior mean when lagged expenditures are nonzero. Likewise, a 1% difference in nonzero initial spending between individuals is associated with a 0.20% difference in spending during period t. This strong association suggests that treating the initial conditions as fixed is inappropriate. The large negative coefficients on d_{it-1} and d_{i0} in the continuous component reflect the non-linear pattern between $d_{it-1} \times \ln y_{it-1}$ and $d_{i0} \times \ln y_{i0}$ created by the degenerate distribution at zero (see appendix figures 3.E.1 and 3.E.2).

The size of the variances of both b_{1i} and b_{2i} suggest that unobserved heterogeneity is important. For instance, the posterior mean of σ_2 is $\sqrt{0.29} = 0.53$, which means that conditional on positive expenditures, a person with a value of b_{2i} one standard deviation above its mean would have out-of-pocket expenditures approximately 53% above average, given personal characteristics. Similarly, the intraclass correlations in both the first and second part of the model are large: at posterior medians $\sigma_1^2/(\sigma_1^2+1) = 0.35$ and $\sigma_2^2/(\sigma_2^2+\sigma_\epsilon^2) = 0.18$.[15] In other words, unobserved heterogeneity explains about 35 and 18 percent of the unexplained variation in the binary and continuous components respectively. Finally, the high correlation between b_{1i} and b_{2i} of 0.516 highlights the importance of modeling the unobserved individual effects jointly.

Table 3.4 reports results for the mortality model. As in the expenditure model, the negative coefficient on d_{it-1} implies that there is a non-linear relationship between $d_{it-1} \times \ln y_{it-1}$ and the dependent variable (mortality) caused by the point mass at zero in medical expenditure data. Lagged log (nonzero) spending, $d_{it-1} \times y_{it-1}$, is fairly large in magnitude and precisely estimated. At posterior means and conditional on positive lagged expenditures,

[15] For $t \neq s$, the intraclass correlations in the binary and continuous components are $corr(\epsilon_{1it}+b_{1i}, \epsilon_{1is}+b_{1i})$ and $corr(\epsilon_{2it}+b_{2i}, \epsilon_{2is}+b_{2i})$ respectively.

a 1% increase in lagged (nonzero) expenditures is predicted to increase the probability of mortality by approximately 0.02% for an individual with a predicted mortality rate of 0.09 (the sample average). The other coefficients are as expected. At older ages, mortality is an upward sloping quadratic function of age. Furthermore, women, non-blacks, and those with more years of education are predicted to survive to older ages. Finally, the large negative coefficient on the Hispanic variable is consistent with the well known "Hispanic paradox" in epidemiology in which Hispanics live longer lives despite their socioeconomic disadvantages.

Table 3.4: Parameter Estimates for Mortality Model

			Posterior Quantiles		
	Mean	SD	2.5%	Median	97.5%
Intercept	−1.55	0.02	−1.59	−1.55	−1.51
$D_{it-1} \times \ln y_{it-1}$	0.12	0.00	0.11	0.12	0.13
D_{it-1}	−1.11	0.04	−1.19	−1.11	−1.02
$(Age - 65)/10$	0.29	0.01	0.26	0.29	0.32
$[(Age - 65)/10]^2$	0.07	0.00	0.06	0.07	0.08
Female	−0.21	0.01	−0.24	−0.21	−0.19
Black	0.11	0.02	0.07	0.11	0.15
Other race	0.04	0.04	−0.03	0.04	0.12
Hispanic	−0.20	0.03	−0.25	−0.20	−0.14
Years of education - 12	−0.03	0.00	−0.04	−0.03	−0.03

Notes: Parameter estimates are from the model described in equation 3.6. d_{it-1} is an indicator variable for whether there was positive spending in the previous wave. Likewise, $\ln y_{it-1}$ is the natural logarithm of expenditures in the previous wave.

In order to compare the predictive ability of the expenditure model to previous studies, I calculated conditional per period mean costs for each individual using the following formula,

$$y_{it}^{*}|\theta, b_i = \Phi(x_{1it}^T\alpha + b_{1i}) \exp(x_{2it}^T\beta + b_{2i} + \sigma_\epsilon^2/2), \tag{3.10}$$

where the second term is the mean of a lognormal distribution. To make predictions, I use the mean of the posterior distribution produced by equation 3.10, or $\hat{y}_{it}$. Predictions were made separately for a two-part model that did not include unobserved heterogeneity and the primary model with correlated unobserved individual effects. The fit of each model was summarized using the percent of the variation explained, or, $R^2 = 1 - \mathrm{Var}(\hat{\mathrm{y}}_{\mathrm{it}} - \mathrm{y}_{\mathrm{it}})/\mathrm{Var}(\mathrm{y}_{\mathrm{it}})$.

In the model with no unobserved heterogeneity, the R^2 of the model is only 0.07, which

suggests that only a very small fraction of out-of-pocket costs can be predicted with the observed data. Adding unobserved heterogeneity doubles the proportion of variation explained by the model as the R^2 increases to 0.16, which is around the maximum percentage—15 to 20 percent—of variation in health costs that previous studies have been able to predict.

3.4.4 Model Checking

To evaluate the fit of a Bayesian model, Gelman et al. (2013) suggest using posterior predictive checks in which observed data, y, is compared to data generated from the model, y^{rep}. One way to draw replicated data is to simply use the draws of the parameters from the simulated posterior. However, a better test of the model in this paper is whether it can simulate an entire earnings history—while simultaneously accounting for mortality—by only using information available during period 0. I therefore use a simulation procedure that first randomly draws the parameter values from the joint posterior distribution, then simulates b_i from $N_2(0, \Sigma_b)$ (rather than using actual draws from the posterior of b_i), and finally recursively simulates y_{it} and m_{it} conditional on spending in period 0. Overall, $1,000$ simulations, y^{sim}, were generated for each of the $15,273$ individuals who survived to period 1.[16] (For complete details of the simulation procedure, see Section 3.D.)

Figure 3.3 plots the density of a single 7-period simulation, y^{sim}, against the corresponding empirical density of y from waves 5 to 11.[17] The top-left figure is a simple plot of the cross-sectional distribution of out-of-pocket spending and the top-right figure is just the $log(x+1)$ transformation. The large spike at zero in the logarithm of spending plot illustrates the large proportion of nonspenders while the plot in levels is a testament to the skewed nature of the data. The simulated data captures the number of non-spenders accurately: the mean percentage of non-spenders across simulated datasets is 10.03% (95% CI: 9.35% to 10.73%) while 10.47% of individuals are non-spenders in the observed data. The log of positive spending is also approximately normally distributed which suggests that the normality

[16] Since the simulation accounts for mortality the length of the vector y^{sim} is random.

[17] There is very little variation in the distribution of spending from one simulation to the next.

assumption on the error term is valid. Finally, the simulated data matches the observed data very closely which implies that the model is able to simulate the cross-sectional distribution of spending.

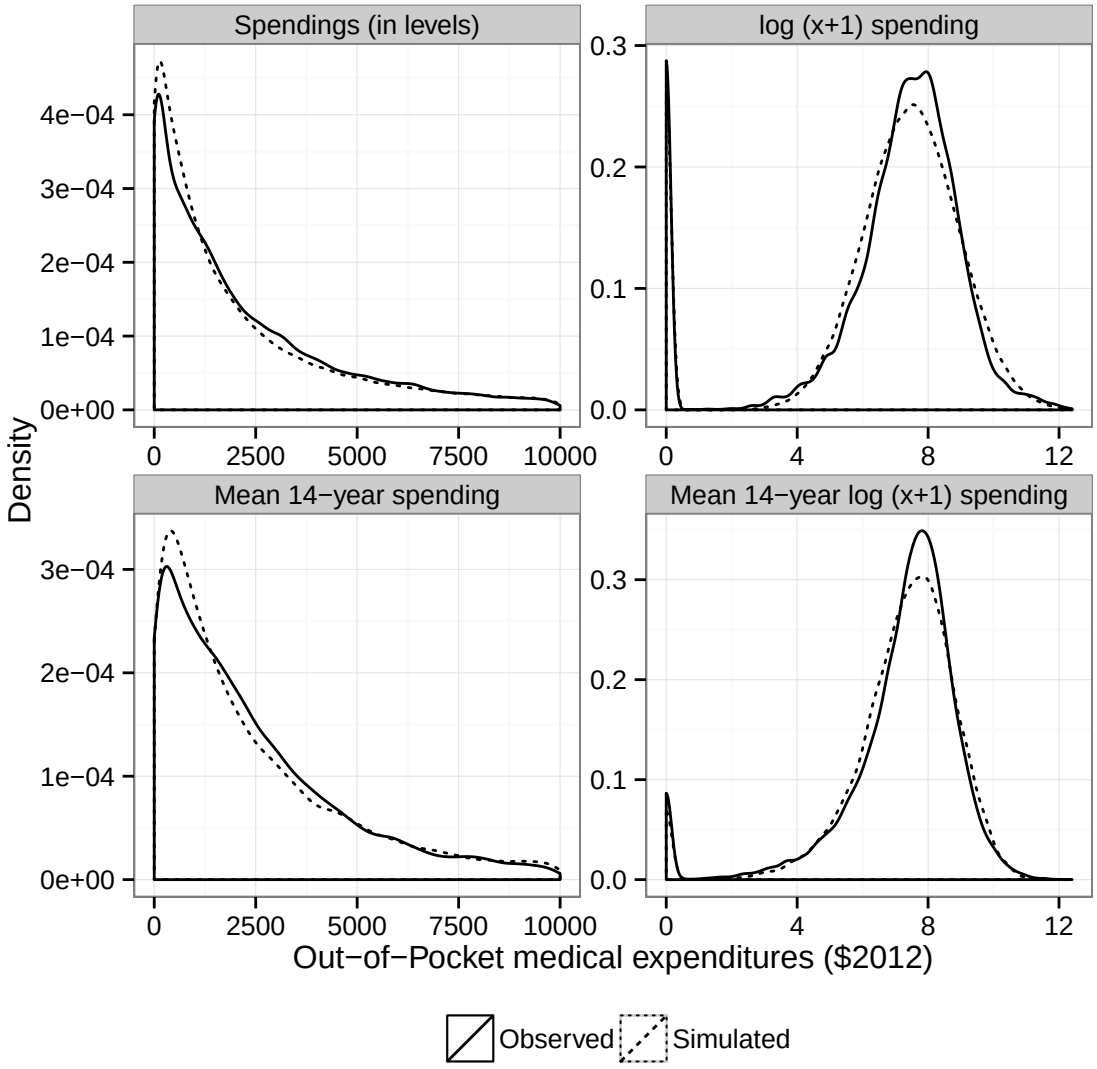


Figure 3.3: Out-of-Pocket Expenditure Density Estimates

Notes: Spending (in levels) is truncated at $10000 (approximately the 95th and 96th percentile of mean 2-year and 14-year spending respectively) for graphical purposes.

The bottom two figures are within individual spending averages across the 7 periods, with expenditures in periods after death coded as $0. Consistent with Table 3.1, the distribution

of the 14-year averages is somewhat less skewed than the distribution of cross-sectional spending pooled across waves. The distribution of y^{sim} is again similar to the distribution of y so it would seem that it is possible to simulate the distribution of expenditures accurately.

Figure 3.4 looks at whether the mortality model accurately predicts the number of survey respondents surviving from one wave to the next. The figure reports the proportion of the $15,273$ individuals alive during wave 4 still alive during each subsequent wave. The model tends to slightly underpredict survival in earlier waves but by wave 11 the simulated survival rate is nearly identical to the observed rate.

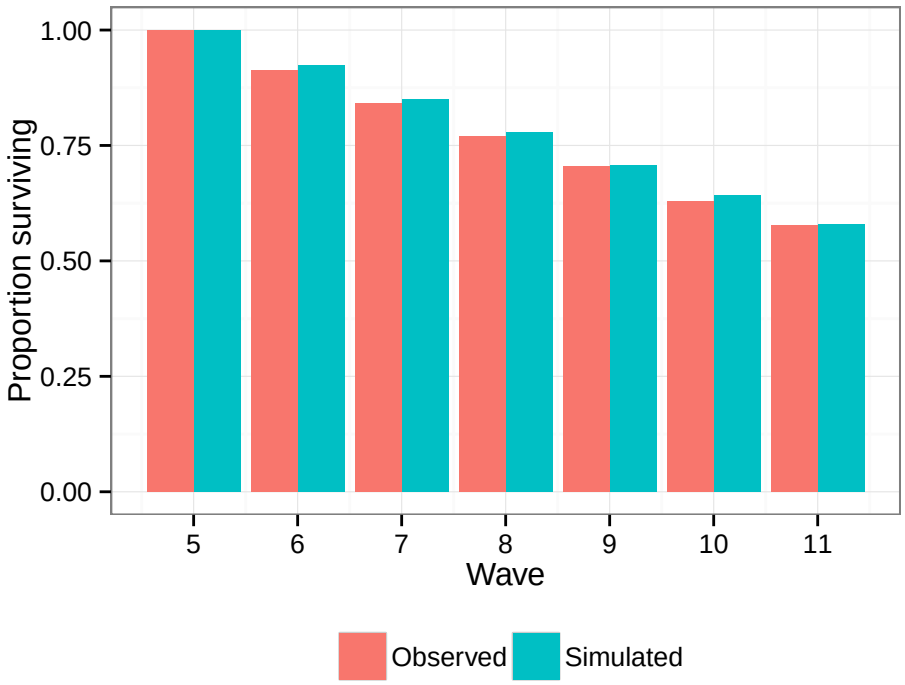


Figure 3.4: Survival Estimates

To asses whether the model adequately captures persistence in out-of-pocket expenditures, Figure 3.5 considers the distribution of expenditures in wave $t+4$ conditional on the

distribution of expenditures in wave t and survival until wave $t+4$.[18][19] The plot is similar to Figure 3.2 in that it presents the evolution of an individual's expenditures across waves. Each of the five panels in the plot shows the proportion of individuals in one of five spending quintiles at wave $t+4$ given their spending quintile at wave t.

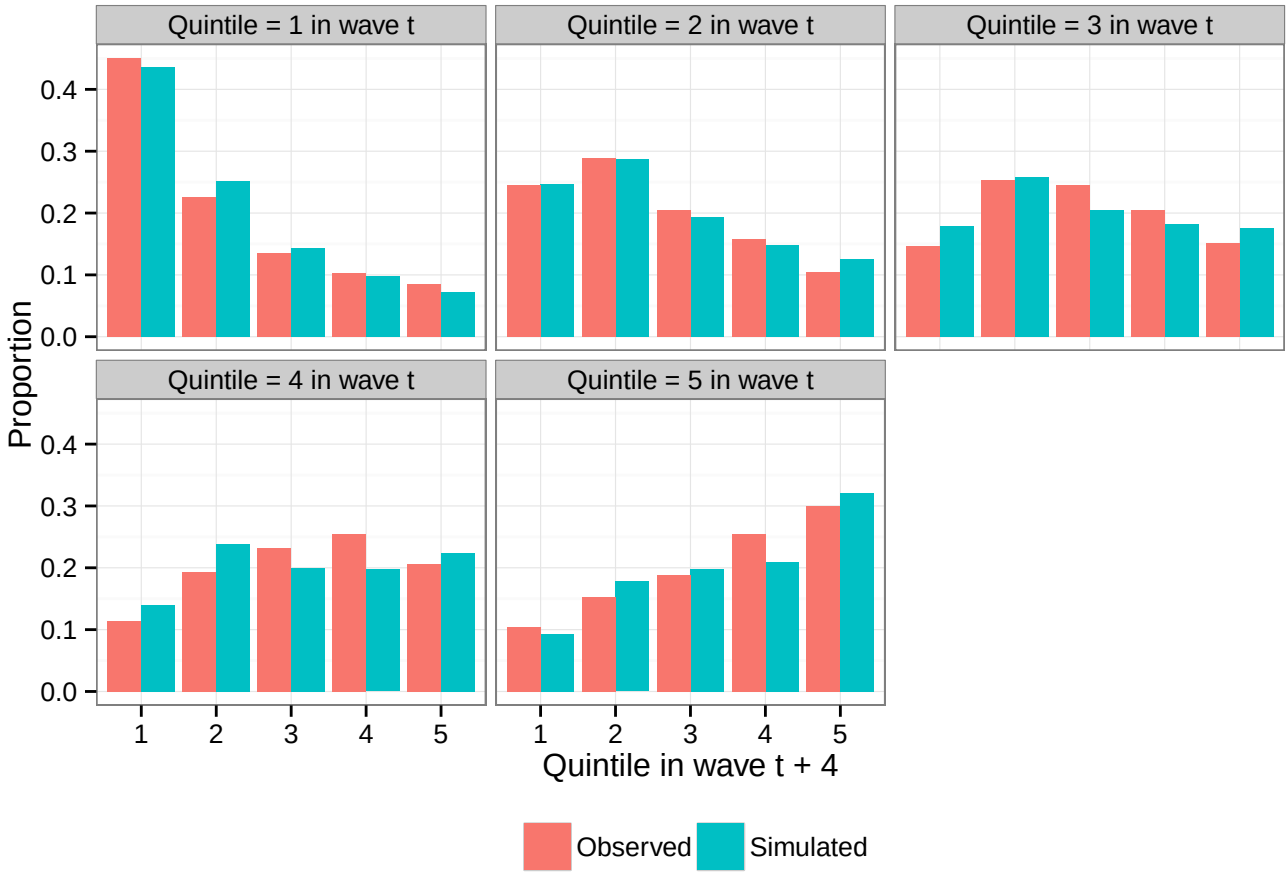


Figure 3.5: Distribution of Wave t + 4 Expenditures Conditional on Wave t Expenditures

Notes: The figure only contains expenditure data for those who survive to wave $t+4$.

The results show that there is significant mean reversion in spending but that spending is still somewhat persistent over time. Individuals in the first expenditure quintile in wave

[18]For example, this compares expenditures in wave 5 to wave 9, wave 6 to wave 10, and wave 7 to wave 11.

[19]Results from other wave differences (i.e. comparing waves t and $t+1$) are similar.

t are much more likely to remain in the first quintile 4 waves later. Likewise, individuals in the highest quintile in wave t are more likely to remain in the higher quintiles in wave $t+4$. That said, the majority of individuals move to new spending quintiles.

The simulated data tends to replicate these patterns but there are a few discrepancies. For instance, compared to the observed data, individuals simulated to be in the fourth quintile in wave t are slightly less likely to remain in the fourth quintile in wave $t+4$. That being said, the model does a good job of simulating transitions from one quintile to another. For additional model checks see Section 3.F.

3.5 Long-term Expenditure Risk

3.5.1 Equity

How unequal is the distribution of out-of-pocket expenditures? Most studies analyzing this question have focused on the distribution of expenditures over short periods of time (typically a single year) (e.g. Berk and Monheit 2001; Monheit 2003; Stanton and Rutherford 2006). Here, I compare the distribution of out-of-pocket expenditures in a single (two-year) period to the distribution of out-of-pocket expenditures over multiple periods.

One way to measure spending inequality is by using the approach developed in the income inequality literature. The standard graphical representation of inequality is a Lorenz curve which plots the cumulative share of spending on the y-axis and the cumulative portion of the population on the x-axis. The figure also typically contains a 45 degree line called "the line of perfect equality", which depicts the distribution of expenditures when everyone spends the same amount.

Lorenz curves for the raw HRS data are shown in Figure 3.6. A curve is plotted for both expenditures in a single wave and mean expenditures over waves 4 to 11 per wave alive. Estimates of single-wave spending are taken from each individual's midpoint period, or period $T_i/2$ rounded to the smallest integer not smaller $T_i/2$. The figure shows that the distribution becomes more equal when considering long-term spending patterns, but that

out-of-pocket spending is still highly concentrated. For example, in a particular two-year wave the top 5% of spenders account for 50% of total spending; over 8 waves the top 5% of spenders still account for 38% of total spending.

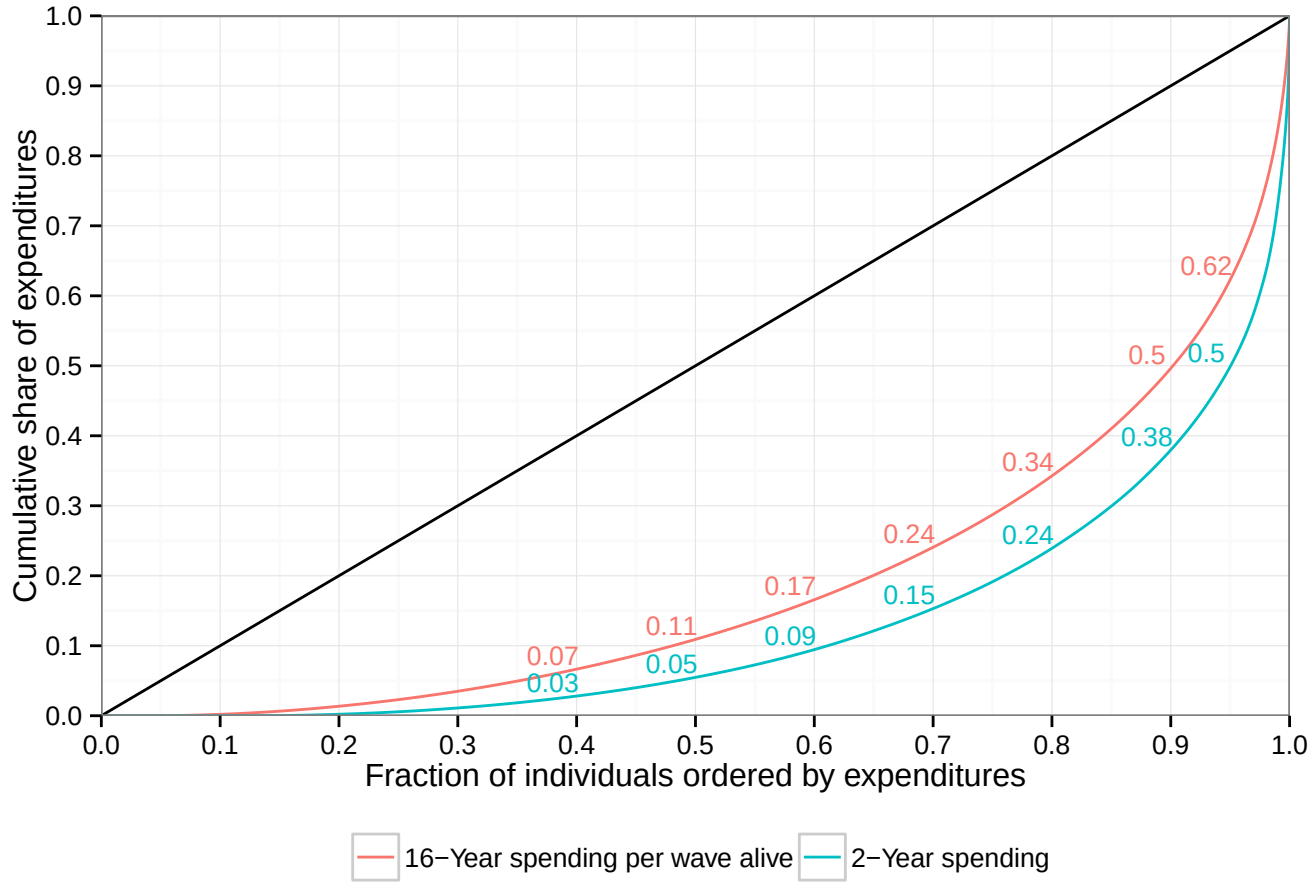


Figure 3.6: Empirical Lorenz Curves for Out-of-Pocket Expenditures per Wave Alive

Notes: The 45 degree line is the "line of perfect equality", which depicts the case where everyone spends the same amount out-of-pocket on healthcare.

One way to quantify inequality is with the Gini coefficient. The Gini coefficient is the ratio of the area between the line of perfect equality and the Lorenz curve to the total area under the line of perfect equality. Its value is 0 when everyone spends the same amount and 1 when only one person has positive expenditures. The Gini coefficient for 16-year spending

(0.63) is substantially lower than for single-wave spending (0.74). Yet, both are fairly close to 1 which suggests that spending inequality persists over long periods of time.

The long-term out-of-pocket spending distribution might be more equal if one examines total usage over the 16-year period since high cost users are more likely to die and become zero-cost users. However, coding expenditures after death as $0 has a negligible impact on the inequality estimates as the Gini coefficient only decreases from 0.63 to 0.60. Thus, long-term estimates of inequality are not very sensitive to whether spending is tracked until death or over a given period of time.

A related question is whether the distribution of spending over a given number of periods would becomes more equal as the number of periods increased. To examine this question, I simulated expenditures (and mortality) over 13 periods (26 years) for individuals age 60 or below during wave 5. Lorenz curves for spending per period alive were then calculated in a rolling fashion over different period lengths.

Figure 3.7 plots some of these simulated Lorenz curves. The figure shows that increasing the number of periods reduces inequality in out-of-pocket expenditures but that inequality decreases at diminishing rates as the number of periods increases. Once spending is measured over 6 periods additional increases in the number of periods have very little impact on the distribution of spending.

Differences in expenditures across individuals are driven by three primary factors: predictable variation, unobserved heterogeneity and random shocks. In order separate predictable variation from variation due to uncertainty, three cases were simulated. In the first case, out-of-pocket spending variation is driven entirely by predicted variation from observable individual characteristics. Predictions are made using equation 3.10, without accounting for state dependence, initial conditions, or unobserved individual effects. In the second case, expenditures are still driven entirely by predictable variation, but spending is persistent because the model includes lagged dependent variables and initial conditions. The third case simply estimates out-of-pocket expenditure variation using the simulation procedure used

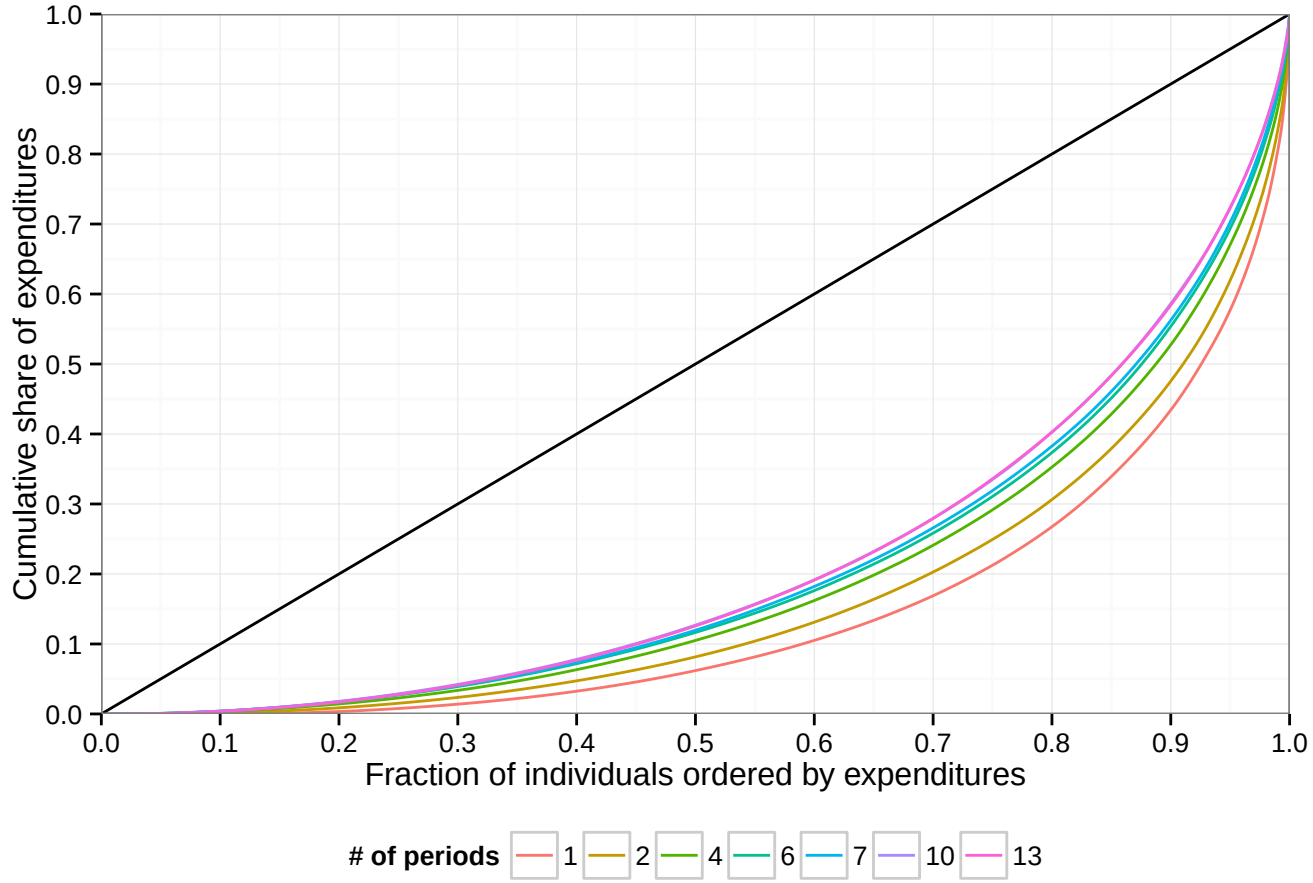


Figure 3.7: Simulated Lorenz Curves for Out-of-Pocket Expenditures per Period Alive, Age $\leq$ 60 in Period 1

Notes: Expenditures for individuals less than or equal to age 60 in wave 5 were simulated for various lengths of time using the simulation procedure described in the text. The Lorenz curves are calculated using spending averaged across the number of periods in which an in individual is simulated to be alive. The 45 degree line is the "line of perfect equality" in which every individuals spends the same amount out-of-pocket on healthcare.

throughout the paper.[20]

Figure 3.8 plots Lorenz curves from data generated from these three cases. Spending was simulated over 7 periods in each case and average spending per period alive was computed for each individual. The legend is ordered from left to right in decreasing order of the

[20] A separate model was estimated for each case. The first two cases use two-part models without unobserved heterogeneity and the third case uses the model described in the text.

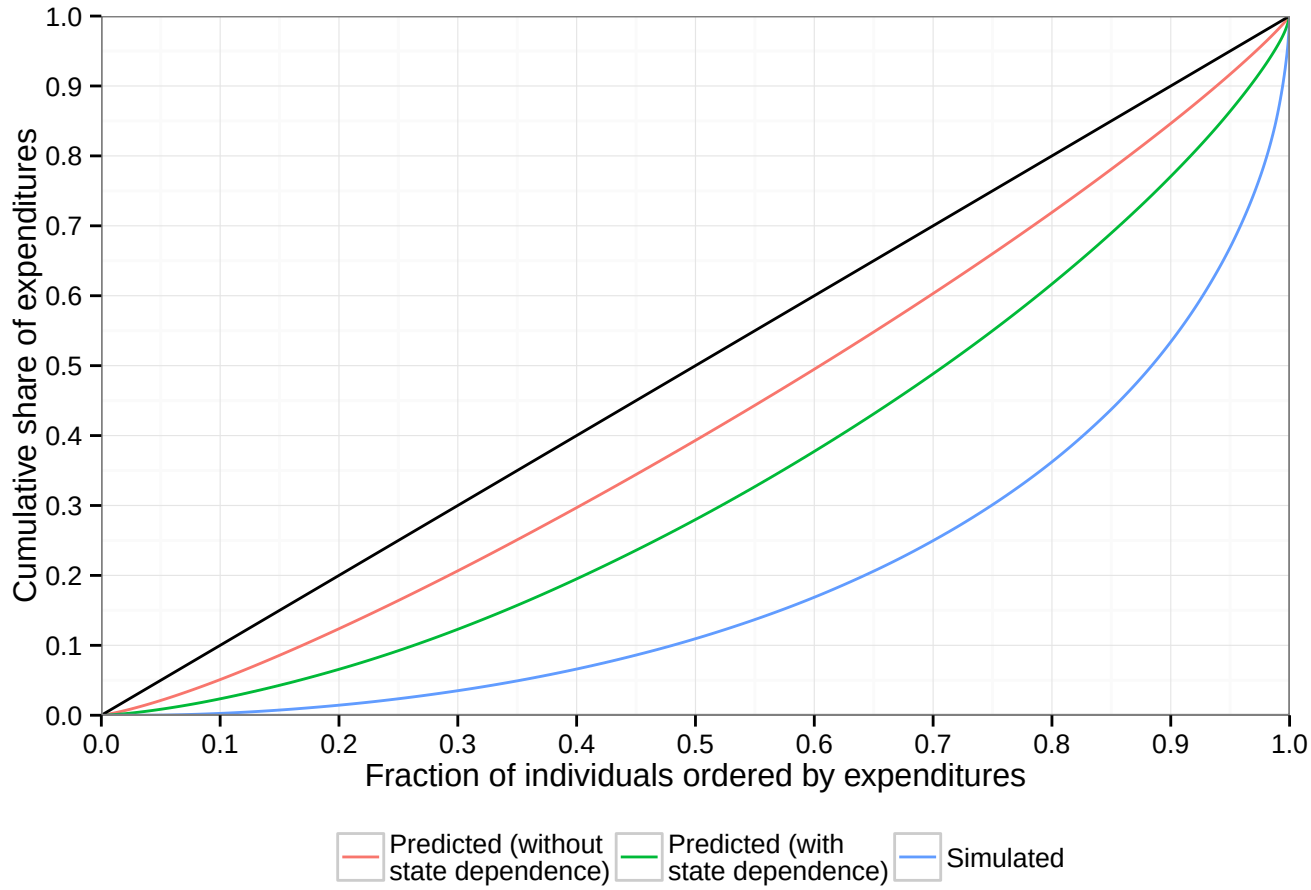


Figure 3.8: Simulated Lorenz Curves for 14-Year Out-of-Pocket Expenditures per Period Alive, by Type of Risk

Notes: The 45 degree line is the "line of perfect equality", which depicts the case where everyone spends the same amount out-of-pocket on healthcare.

Gini coefficient. The spending distribution is the most equal when there is no uncertainty or persistence, and variation is driven entirely by observable characteristics. The distribution of spending becomes more unequal when state dependence is incorporated into the simulation, but the transitory shocks and unobserved individual effects contribute the most to spending differences. It is no surprise that the model estimated in the paper generates a distribution that is considerably less equal than the distributions generated from predicted means.

One potential consequence of highly concentrated out-of-pocket spending is that certain

individuals might be more at risk for catastrophic healthcare expenditures. It is of course an open question as to what constitutes a catastrophic expenditure. To this end, researchers have come up with various ways to measure financial protection in health systems. One measure suggested by Waters, Anderson and Mays (2004) for analyses in the United States is whether out-of-pocket expenditures exceed 10% of family income. Using this (and two other) definitions of financial protection, Waters, Anderson and Mays (2004) argue that there is a lack of financial protection in the United States—particularly for poor families and those with chronic conditions. The results reported so far would seem to suggest that this risk remains significant over longer periods of times as well.

Figure 3.9 looks at whether this is the case by plotting the empirical CDF of out-of-pocket expenditures as a percentage of family income.[21] The CDF's are plotted for both a single 2-year wave (the midpoint wave) and over 16 years. The CDF's are then further divided into four (annualized) income quartiles.

The distribution of spending as a percentage of income is, in general, not very sensitive to whether the distribution is measured in a single wave or over 8 waves. The upper end of the distribution of out-of-pocket spending as a percentage of income is very similar regardless of income quartile or of the time period over which it is measured. For instance, 21% of the poor spend over 20% of income in a single wave; over 8 waves, 21% do as well. But at the far right tail, financial risk is somewhat less catastrophic over longer periods of time. For example, low-income individuals at the 95th percentile spend 96% of their income on out-of-pocket healthcare costs in a single wave; over 8 waves this number drops to 69%. Even so, healthcare costs are still a significant financial risk—especially for the poor— over long periods of time.

At the lower end of the distribution, healthcare spending as a percentage of income can actually be larger over long periods of time—again, especially for the poor. This occurs

[21]Some households reported earning $0 in some wave, which is unrealistic because they should have earned income from either welfare or social security. Incomes are consequently truncated from below at the 1st percentile, although this has no impact on the results.

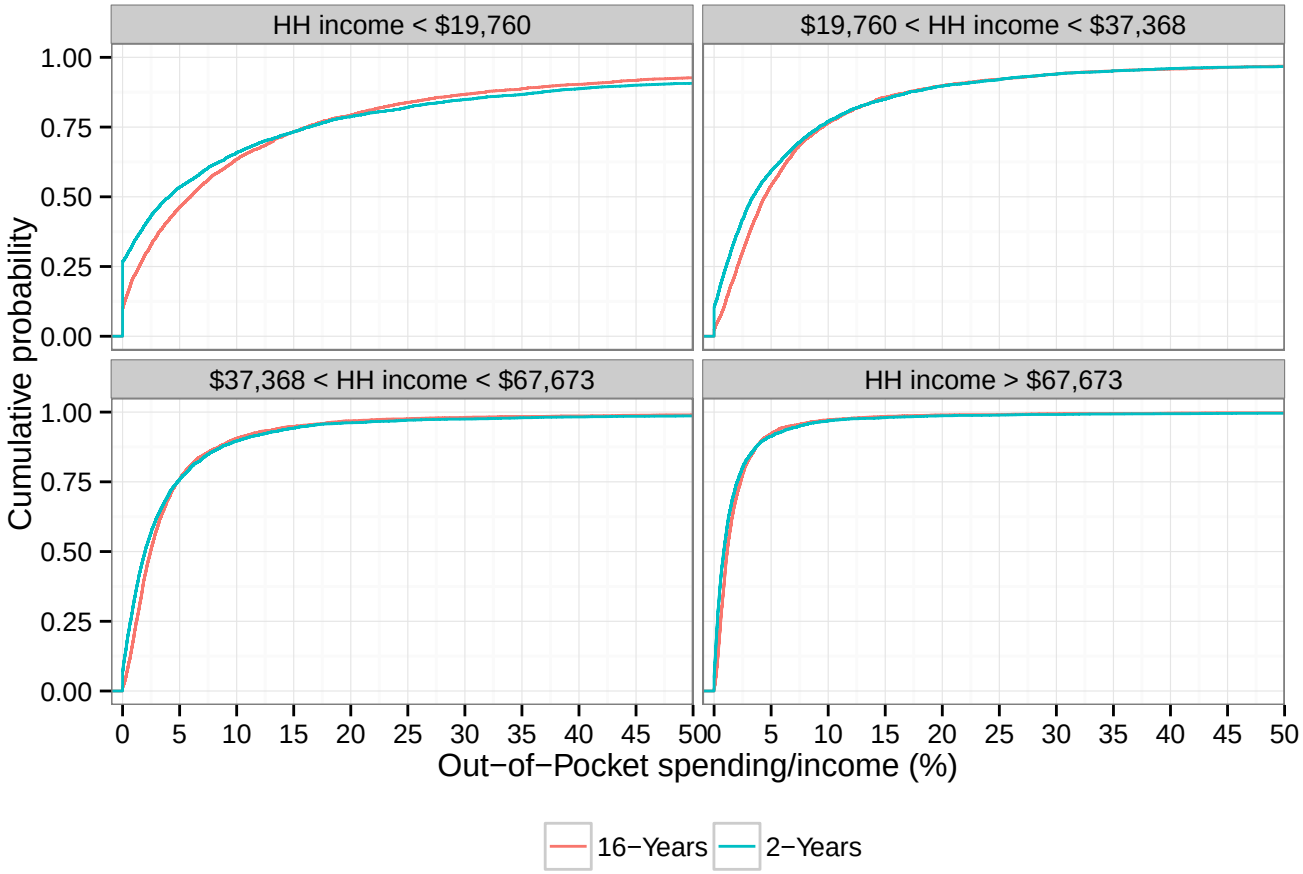


Figure 3.9: Empirical CDF of Out-of-Pocket Expenditure as a Percentage of Household Income by Income Quartile

Notes: Reported income quartiles are annualized.

because a smaller portion of individuals have very small healthcare costs over extended periods. For individuals in the lowest income quartile a larger fraction of survey respondents spend approximately 5% of their income on out-of-pocket healthcare when measured over 8 waves than in a single wave. Likewise, the median low-income individual spends 4% of their household income on on out-of-pocket health costs in a single wave and 6% over 8 waves.

3.5.2 Uncertainty and Welfare

So far this paper has shown that there is significant variation in the long-term distribution out-of-pocket expenditures and that out-of-pocket spending in a given two-year period is largely unpredictable. In this section, I provide evidence on the degree of uncertainty in long-term spending. I then examine the extent to which future spending uncertainty makes it difficult to target high spenders. Lastly, I use a stylized utility framework to provide a rough estimate of potential welfare gains from eliminating exposure to risk.

To investigate the predictability of long term spending I estimated expected out-of-pocket expenditures over a 14-year period for each individual conditional on period 0 information. This was done by calculating the mean of the $1,000$ simulations described in Section 3.4. I then repeated the simulations assuming that each individual knew the value of his or her b_i. More precisely, in the first simulation the b_i's were drawn from their $N_2(0, \Sigma_b)$ distribution and in the second simulation the actual posterior draws of b_i's were used.

Table 3.5 examines the amount of variation in actual 14-year spending that each simulation is able to predict. The figure reports the R^2 value described in Section 3.4 for total 14-year out-of-pocket expenditures and out-of-pocket expenditures per period alive. Estimates are then further sub divided by whether the simulations estimated mortality or the simulations used the observed life expectancy for each individual.

Table 3.5: R-squared for 14-Year Out-of-Pocket Spending, by Simulation Type

	Total out-of-pocket expenditures		Out-of-pocket expenditures per period alive	
Simulation type	Simulated mortality	Observed mortality	Simulated mortality	Observed mortality
Unknown b_i	0.05	0.07	0.10	0.10
Known b_i	0.35	0.41	0.37	0.36

The main takeaway from the table is that only a small fraction of the variation in spending can be predicted using period 0 information. The R^2 values are generally quite low when there is unobserved heterogeneity (row 1 in the table). The simulation can explain around

10% of the variation in spending per period alive but less than 7% in total expenditures. On the other hand, when observed values of b_i are used in the simulations, the percent of explained variation increases noticeably. The model can explain over 35% of the variation in 14-year spending per wave alive and in total expenditures. But even so, a large fraction of the variation in long-term spending remains unaccounted for.

The low R^2 values suggest that it may be difficult for disease management programs to target high spenders. Figure 3.10 investigates this by comparing observed spending by predicted high spenders to observed spending by actual high spenders. As in Table 3.5, predictions are made using expected 14-year expenditures from the simulation. A high spender is defined as someone with spending above a minimum spending threshold. Predicted expenditures are larger than the threshold for predicted high spenders while observed expenditures are larger than the threshold for actual high spenders.

The figure considers a number of thresholds ranging from spending at the 70th to 99th percentile. At each minimum spending percentile, the figure divides observed spending by predicted high spenders by observed spending by actual high spenders. This ratio is the fraction of spending by high spenders that the model can account for and is an upper bound on the fraction of spending that disease management programs could eliminate.

As expected, the model can account for a much larger fraction of out-of-pocket expenditures when heterogeneity is known. The spending ratio decreases at an increasing rate as the minimum spending percentile increases, which suggests that there is a tradeoff between using more resources per individual and correctly identifying high spenders. For example, when b_i is unknown, the spending ratio decreases from 0.62 when the minimum threshold is at the 70th percentile to 0.20 when it is at the 99th percentile.

Although uncertainty makes targeting high spenders difficult, it also makes reducing risk more valuable. To quantify these benefits, I estimate the potential welfare gains from completely eliminating uncertainty. My approach is similar to a number of other studies that have estimated the utility gains associated with specific insurance plans (McClellan

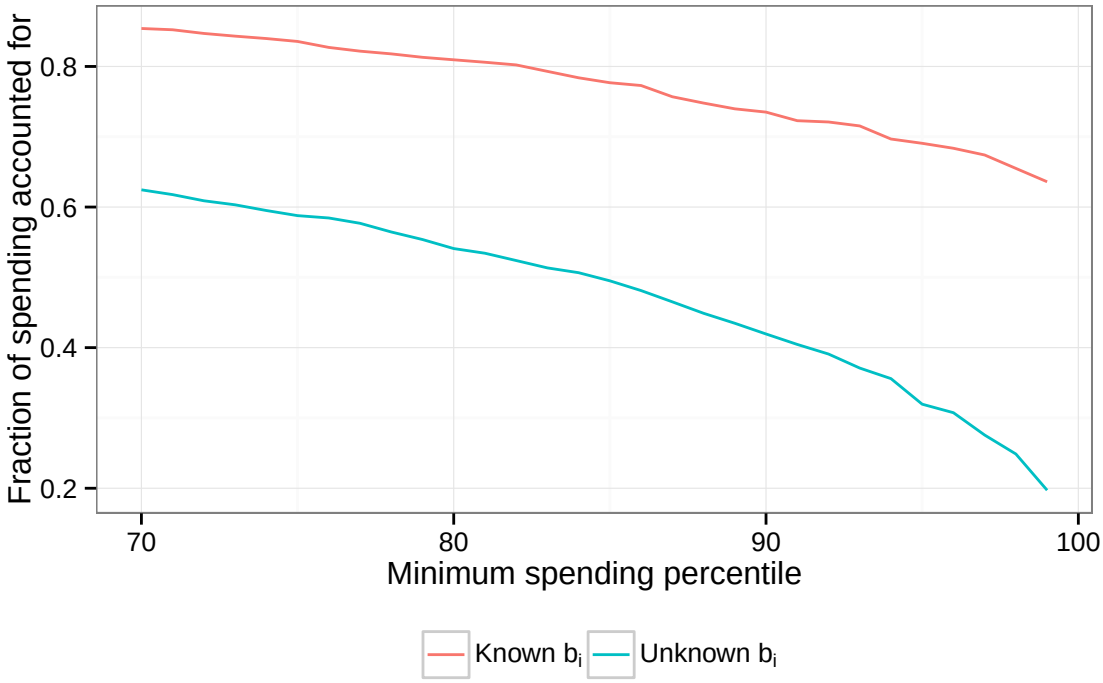


Figure 3.10: Targeting High Spenders

Notes: Predicted (actual) high spenders are individuals with predicted (actual) expenditures above a particular "minimum spending percentile." The "fraction of spending accounted for" is the sum of observed spending by predicted high spenders divided by the sum of observed spending by actual high spenders.

and Skinner 2006; Brown and Finkelstein 2008; Finkelstein and McKnight 2008; Engelhardt and Gruber 2011). In particular, I estimate welfare by calculating each individual's risk premium, or the maximum amount that they would be willing to pay at the beginning of period 1 to completely insure against future risk.

Following Finkelstein and McKnight (2008) and Engelhardt and Gruber (2011), individuals are assumed to receive utility from consumption, c_t, which is subject to a per period budget constraint, $c_t = I_t - y_t$, where I_t is income and y_t is out-of-pocket spending. Over T periods, individuals are assumed to receive utility from the present value of consumption, $\sum_{t=1}^{T} \delta^t c_t$, where δ_t is the discount rate. The risk premium, π, is the amount of money an individual would need to receive to become indifferent between a world in which out-of-pocket costs are uncertain each period and a world in which he or she would pay the

expected present value of future spending over T periods.[22] Given a utility function, $U(\cdot)$, and probability density function, $f(\cdot)$, π is the solution to the implicit equation,

$$U\left(\sum_{t=1}^{T}\delta_t[I_t - E(y_t)] - \pi\right) = \int U\left(\sum_{t=1}^{T}\delta_t[I_t - y_t]\right) f(y_t)dy_t. \tag{3.11}$$

This setup is obviously a simplification of reality since it does not allow for borrowing, saving, or non-income wealth. However, there are a number of reasons to believe that it might not be too unrealistic. First, a number of studies have shown that the relationship between consumption and age is humped shaped—consumption increases during young adulthood before peaking around age 45 and then declines until death—and tracks income closely (Gourinchas and Parker 2002; Fernández-Villaverde and Krueger 2007).[23] Second, it has been estimated that nearly four-fifths of net wealth is bequeathed (Kopczuk and Lupton 2007), so it is unreasonable to assume that individuals consume all of their wealth over their lifetime.

To solve the model, individuals are assumed to have a constant relative risk aversion (CRRA) utility function,

$$U(C) = \begin{cases} \frac{1}{1-\gamma}C^{1-\gamma} & \text{if } \gamma > 0, \gamma \neq 1 \\ \ln C & \text{if } \gamma = 1, \end{cases} \tag{3.12}$$

where γ is a risk aversion parameter. Future consumption is discounted at 3% per year; that is, $\delta_t = 1/(1+0.03)^{2t-2}$. The integral in the second term of equations 3.11 is evaluated by using the probability distribution of out-of-pocket expenditures estimated from the Bayesian

[22] Paying a fixed sum equal to the expected present value of future spending is equivalent to paying an actuarially fair insurance premium each period (a premium is actuarially fair if the expected present value of benefits is equal to the expected present value of premiums).

[23] Note that these empirical findings are inconsistent with the permanent income hypothesis. Common explanations within an expected utility maximization framework are that individuals have high discount rates, only save as a precaution against sharp falls in income, or that they are liquidity constrained. Shefrin and Thaler (1988) provide an alternative explanation in which individuals do not base their consumption choices on maximizing their lifetime utility. Instead, individuals follow a rule of thumb in which they they spend current income and only dip into their savings if absolutely necessary.

model. For instance, using the 1000 simulated values, y^{sim}, expected utility for the ith individual in period t is just $(1/1000)\sum_{s=1}^{1000} U(I_{it}-y_{it}^{sim,s})$. To remain consistent with previous studies (e.g. Finkelstein and McKnight 2008; Engelhardt and Gruber 2011), out-of-pocket spending is truncated from below at 0 and above at 80% of income. In addition, incomes are treated as fixed and truncated from below at the 1st percentile.[24] Household incomes are converted to individual incomes using the OECD equivalence scales.[25]

Figure 3.11 reports 14-year risk premiums for the entire sample of 15,273 survey respondents analyzed in this paper. The figure plots empirical CDF's of the risk premiums for $\gamma = 2$ and $\gamma = 3$, which are common estimates in the literature.

Variability across individuals stems from three primary factors. First, welfare gains are larger for poorer individuals because the CRRA utility function implies decreasing absolute risk aversion (DARA); that is, risk premiums are decreasing in income. Second, risk premiums are larger when individuals live longer lives. Third, risk premiums increase when the variance of out-of-pocket expenditures increases. This in turn implies that risk premiums are increasing in predicted (mean) out-of-pocket expenditures because the distribution of nonzero expenditures is assumed to have a lognormal distribution.

For a relative risk coefficient of 2, the long-term per period risk premium for the median individual is \$3,215 while the risk premium for an individual at the 90th percentile is \$8,349; for a risk premium of 3, these quantities increase to \$6,898 and \$17,103 respectively. These estimates are fairly large, which implies that there are potential welfare gains from additional insurance, especially for those at the upper end of the distribution. That said, any welfare gains from additional public insurance must be weighed against inefficient increases in spending due to moral hazard.

[24] Since incomes are set equal to observed values it is necessary to set mortality to observed values as well; otherwise income would need to be forecasted. See footnote 21 regarding truncating income at the 1st percentile.

[25] Equivalence scales measure the change in consumption needed to keep the welfare of a family constant when its size changes. OECD equivalence scales place a weight of 1 on the household head, 0.7 on subsequent adults over age 14, and 0.5 on each child under 14. Child ages are unavailable in the RAND HRS data so each household member other than the household head is assigned a weight of 0.7.

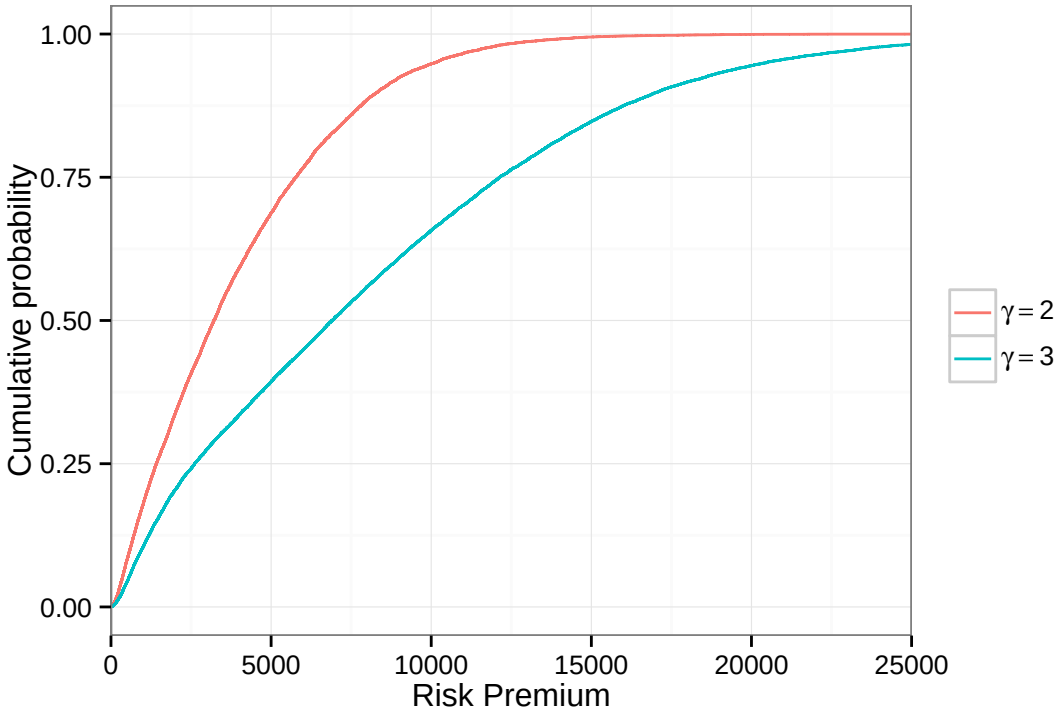


Figure 3.11: Empirical CDF of Estimated Risk Premiums

Notes: Risk premiums are calculated assuming a CRRA utility function. Individuals are assumed to be more risk averse when the risk aversion parameter, γ, is larger.

3.6 Conclusion

This paper analyzes out-of-pocket expenditures over a much longer time period than in previous studies. It develops a dynamic Bayesian models of out-of-pocket expenditures and mortality and applies it to the HRS. It finds that the data generating process for expenditures is well represented by a dynamic two-part longitudinal model with correlated unobserved individual effects, and that mortality can be modeled using a probit model that is conditional on previous expenditures. The model accounts for state dependence and three potential sources of error: parameter uncertainty, unobserved heterogeneity and a transitory shocks.

The analysis has some important implications for cost containment policies. Expenditures are difficult to predict using observed data alone: both unobserved heterogeneity and transitory shocks explain a large proportion of the variance in spending. This suggests that

it may be difficult for clinically oriented methods such as disease management programs to identify high spenders. At the same time, the uncertain nature of future costs makes health insurance valuable and suggests that individuals using financial instruments like health savings accounts may still be subject to significant financial risk. The distribution of spending also remains highly unequal—even over long periods of time—which means that increasing cost sharing would likely create policy pressure to redistribute healthcare costs in other ways. For consumer driven healthcare to succeed, it must successfully balance the benefits of reduced moral hazard against the costs of increased inequality in out-of-pocket spending and reduced welfare from risk reduction.[26]

There are many avenues for future research. First, the model could be extended in a number of ways. For example, analyses focusing on financial planning should model wealth alongside mortality and out-of-pocket expenditures. In addition, one could allow survival to depend on the unobserved individual effects in the expenditure model (e.g. Wulfsohn and Tsiatis 1997; Ibrahim, Chu and Chen 2010). Finally, the model should be applied to new data sources. There is, in general, a need for high quality longitudinal datasets of healthcare spending. Different mortality models might be appropriate if the data were less aggregated; other data may be less normal and require a relaxation of the lognormality assumption.

Second, it would be useful to analyze individuals with specific chronic conditions using longitudinal data. The analyses in this paper do not use detailed clinical data and certain chronic conditions might be more suitable for disease management programs than suggested in this paper. Indeed, Monheit (2003) has shown that spending is more persistent when individuals have certain medical conditions like diabetes or mental disorders. Allowing persistence to depend on health status would aid efforts aimed at identifying individuals in need of medical and financial management.

Third, researchers should let long-term out-of-pocket spending depend on cost-sharing mechanisms like copayments, coinsurance and deductibles since there is convincing evidence

[26] Another potential cost is that increased cost sharing could cause individuals to forgo necessary medical care.

that spending is sensitive to cost sharing (Manning et al. 1987; Aron-Dine, Einav and Finkelstein 2013). These behavioral responses could be incorporated into the simulation procedure used here and built into dynamic utility-based models (e.g. Brown and Finkelstein 2008) to analyze the long-term implications of specific health insurance reforms. This would, in turn, allow researchers to quantify the competing welfare effects of risk reduction and moral hazard. If possible, these simulations should allow the price elasticity of health expenditures to vary across individuals (e.g. Kowalski 2015) in order to properly analyze the distributional implications of different insurance policies.

Appendix

3.A Out-of-Pocket Spending Data

The HRS collects data on out-of-pocket spending in the following manner. First, each respondent is asked to provide a continuous estimate of spending in a number of medical categories. If the respondent is unsure they are then asked a series of unfolding bracket questions where they are asked whether spending falls within a number of ranges. According to careful analyses of the data by (Goldman, Zissimopoulos and Lu 2011), non-response on the continuous question is typically around 20%, although it can reach as high as 30% in some categories such as hospital and nursing home care. The unfolding bracket question reduces the number of non-responses to below 5% for most categories.

The RAND HRS data files impute out-of-pocket expenditure data for all individuals in the sample. RAND uses a matching imputation method. They first use a linear regression to model the inverse hyperbolic sine transformation of out-of-pocket expenditures and then use the regression to predict out-of-pocket expenditures for every individual in the survey. These predicted values are then used to impute continuous expenditure values for respondents who answered the unfolding bracket questions but did not report a continuous expenditure value. More specifically, each individual whose expenditures were reported to fall within a given bracket were assigned the expenditure value of the respondent with the closest predicted expenditure value whose actual (continuous) expenditures fell with that bracket. A slightly different technique was used to impute data for open ended brackets. For further details on the imputations see [see St Clair et al. (2014, section 7.5, pp. 18–21)].

3.B Sample Selection

Here, I discuss two potential problems with dropping observations from the sample. The first problem is that the sample may become less representative of the population, which limits external validity. Table 3.B.1 examines this by comparing three groups of survey respondents. Respondents in the first group have complete data until death or through wave 11. Members of the second group have missing data in some years (due to non-response), but still participated in the survey until either death or wave 11. Finally, survival times for individuals in the third group are right censored because they dropped out of the HRS completely.

Means of out-of-pocket expenditures, survival times, and time constant observed characteristics are reported for each group. The out-of-pocket expenditure statistics reported in the table are means of individual per wave averages from waves 4 to 11. Mean out-of-pocket expenditures for non-respondents and respondents with complete data are similar, although those with complete data spend a little bit more. Non-respondents were three years younger during the initial wave, so they predictably survived 1 wave longer. Observed characteristics are also similar between the two groups.

Table 3.B.1: Sample selection in the HRS

	Complete data	Non-respondents	Dropouts
Out-of-pocket expenditures	4889.79	4393.81	2919.36
Survival time (number of waves)	6.00	7.08	4.66
Initial age	67.41	64.32	63.47
Black	0.14	0.18	0.09
Female	0.57	0.58	0.55
Hispanic	0.07	0.09	0.08
Years of education	11.84	11.87	12.42

Notes: Means of observed characteristics are averages across individuals. Survival time refers to the number of waves in which a survey respondent was tracked by the HRS before either death or wave 11. Out-of-pocket expenditures are means of individual per wave averages.

Expenditures are lower for dropouts, but this is likely because they were not observed during the end of the 16-year survey period. This biases the out-of-pocket expenditure

estimate downward because spending increases with age and prior to death. In fact, there are a few reasons to believe that excluding dropouts will only have a minor impact on the sample. First, differences between dropouts and individuals in the other two groups decrease significantly if one restricts the sample to those who did not die during the survey: survivors with complete data spent $\$3,434$ while non-respondent survivors spent $\$3,826$. Second, observed characteristics are generally similar across all three groups, although dropouts tend to be slightly better educated, a little more likely to be male, and less likely to be black. Finally, excluding dropouts only eliminates 820 individuals. It is important to note however that dropping right censored observations will bias survival times downward, although this bias should be small given that only a small fraction of the sample is right censored.

The second problem with excluding observations is that it might bias the model parameters. However, the differences in expenditures in Table 3.B.1 are relatively small and are likely due to differences in observable characteristics between individuals. It seems reasonable to assume that the probability of non-response or dropout is unrelated to current spending after controlling for demographics, lagged and initial spending, and unobserved heterogeneity. In other words, the missing data can plausibly be thought of as missing at random (MAR).

3.C MCMC Algorithm

The prior distributions for the model are,

$$\begin{aligned}
\alpha &\sim N(\alpha_0, V_\alpha), \\
\beta &\sim N(\beta_0, V_\beta), \\
\sigma_\epsilon^{-2} &\sim \mathrm{Ga}(\mathrm{a}_0, \mathrm{b}_0), \\
\Sigma_b &\sim IW(S_0, v_0).
\end{aligned}$$

To facilitate posterior computations, it is convenient to augment the joint posterior distribution in equation 3.9 with the latent variable d^*_{it}. Denote n as the number of individuals in the data, $T_i - 1$ as the number of periods for which expenditure data is available for individual i, and N as the total number of observations. Furthermore, let y, d, and d^* be the $N \times 1$ stacked vectors of y_{it}, d_{it} and d^*_{it} respectively. The joint posterior distribution is then,

$$
\begin{aligned}
p(d^*, \alpha, \beta, b_i, \sigma_\epsilon^{-2}, \Sigma_b | y, d) &\propto p(\alpha)p(\beta)p(\sigma_\epsilon^{-2})p(\Sigma_b) \\
&\quad \prod_{i=1}^{n}\prod_{t=1}^{T_i-1} p(d_{it}|d^*_{it})p(d^*_{it}|\alpha)p(y_{it}|d^*_{it}, \alpha, \beta, \sigma_\epsilon^{-2}, b_i)p(b_i|\Sigma_b) \\
&= p(\alpha)p(\beta)p(\sigma_\epsilon^{-2})p(\Sigma_b) \\
&\quad \times \prod_{i=1}^{n}\prod_{t=1}^{T_i-1} [I(d_{it}=1)I(d^*_{it}>0) + I(d_{it}=0)I(d^*_{it}\leq 0] \\
&\quad \times N(d^*_{it}; x_{1it}^T\alpha + b_{1i}, 1) \\
&\quad \times \left[LN(y_{it}; x_{2it}^T\beta + b_{2i}, \sigma_\epsilon^2)I(d^*_{it}>0) + I(y_i=0)I(d^*_{it}\leq 0)\right] \\
&\quad \times N\left(\begin{bmatrix} b_{1i} \\ b_{2i} \end{bmatrix}; \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma_b\right),
\end{aligned}
$$

where $I(\cdot)$ is an indicator function equal to 1 if the statement in parentheses is true and 0 if it is false. Starting with initial values, the MCMC algorithm iterates through the following steps by drawing from the following conditional distributions:

1. **Draw α:**. To draw α, make use of the augmented data and first draw d^*_{it} from its full conditional,

$$
d^*_{it}|\cdot \sim \begin{cases} TN_{(-\infty,0)}(x_{1it}^T\alpha + b_{1i}, 1) & \text{if } d_{it}=0 \\ TN_{(0,\infty)}(x_{1it}^T\alpha + b_{1i}, 1) & \text{if } d_{it}=1, \end{cases}
$$

where "$\cdot$" is a simplifying notation used to denote the remaining parameters in the

model. α can then be drawn from its full conditional,

$$\alpha|\cdot \sim N(\overline{\alpha}, \overline{V}_\alpha),$$

where

$$\begin{aligned}\overline{V}_\alpha &= \left(V_\alpha^{-1} + X_1^T X_1\right)^{-1} \\ \overline{\alpha} &= \overline{V}_\alpha \left[V_\alpha^{-1}\alpha_0 + X_1^T(D^* - b_1)\right],\end{aligned}$$

and X_1 is an $N \times k_1$ design matrix and b_1 is an $N \times 1$ column vector of unobserved individual effects for the binary component (i.e. b_{1i} is repeated $T_i - 1$ times for individual i).

2. **Draw β:** Let $\ln y^+$ denote the $N^+ \times 1$ stacked vector of y_{it} for $y_{it} > 0$ where N^+ is the total number of positive observations. Similarly, let X_2^+ be the corresponding $N^+ \times k_2$ design matrix and b_2^+ be the $N^+ \times 1$ concatenated column vector of unobserved individual effects for the continuous component. Then β can be drawn from its conditional distribution using results for a standard normal linear regression with a normal prior,

$$\beta|\cdot \sim N(\overline{\beta}, \overline{V}_\beta),$$

where,

$$\begin{aligned}\overline{V}_\beta &= \left(V_\beta^{-1} + \sigma_\epsilon^{-2} {X_2^+}^T X_2^+\right)^{-1} \\ \overline{\beta} &= \overline{V}_\beta \left[V_\beta^{-1}\beta_0 + \sigma_\epsilon^{-2} {X_2^+}^T \left(\ln y^+ - b_2^+\right)\right]\end{aligned}$$

3. **Draw σ_ϵ^{-2}:** With the gamma prior, the conditional distribution is also a gamma

distribution,

$$\sigma_\epsilon^{-2}|\cdot \sim \text{Ga}\left(\text{a}_0 + \text{N}^+/2, \text{b}_0 + \frac{1}{2}\left[\ln \text{y}^+ - \text{X}_2^+ - \text{b}_2^+\right]^{\text{T}}\left[\ln \text{y}^+ - \text{X}_2^+ - \text{b}_2^+\right]\right)$$

4. **Draw Σ_b:** With the inverse Wishart prior, the conditional posterior is also inverse Wishart,

$$\Sigma_b \sim IW\left(S_0^{-1} + b^T b, n + v_0\right),$$

where b is an $n \times 2$ matrix with the first and second columns contain unobserved individual effects from the binary and continuous components respectively.

5. **Draw b_i:** The full conditional (marginalized over d_{it}^*) is,

$$p(b_i) \propto N_2\left(b_i; 0, \Sigma_b\right) \prod_{t=1}^{T_i - 1} f(y_{it}|y_{it-1}, \alpha, \beta, \sigma_\epsilon^2, b_i).$$

It is not possible to sample directly from this distribution so a random walk Metropolis algorithm is used to update b_i using a multivariate normal proposal density centered at the value from the previous iteration, b_i^{old}. The variance of the proposal density is tuned to achieve desired acceptance rates.

3.D Simulation Procedure

Each simulation is conditional on data in period 0 (wave 4). For a given individual i and a set of parameter values from the posterior distribution, a single simulation from period 1 until period k proceeds as follows:

1. At $t = 1$:

   (a) Draw unobserved effects: $b_i \sim N_2(0, \Sigma_b)$.

   (b) Draw expenditures:

i. $d_{i1}^* \sim N(x_{1i1}^T \alpha + b_{1i}, 1)$. If $d_{i1}^* \geq 0$, then $d_{i1} = 1$. Else $d_{i1} = 0$

ii. If $d_{i1} = 0$, then $y_{i1} = 0$. Else $y_{i1} \sim \text{LN}(x_{2i1}^T \beta + b_{2i}, \sigma_\epsilon^2)$

2. For $2 < t < k$:

   (a) For $j = 1, 2, M$, update x_{jit} to reflect y_{it-1}. Increment age by 2 and update x_{jit} accordingly.

   (b) Draw death indicators:

      i. $m_{it}^* \sim N(x_{Mit}^T \kappa, 1)$. If $m_{it}^* \geq 0$, then $m_{it} = 1$. Else $m_{it} = 0$

   (c) If $m_{it} = 0$, draw expenditures:

      i. $d_{it}^* \sim N(x_{1it}^T \alpha + b_{1i}, 1)$. If $d_{it}^* \geq 0$, then $d_{it} = 1$. Else $d_{it} = 0$.

      ii. If $d_{it} = 0$, then $y_{it} = 0$. Else $y_{it} \sim \text{LN}(x_{2it}^T \beta + b_{2i}, \sigma_\epsilon^2)$.

   (d) If $t < k$, increment t by 1. Else, stop simulation.

3.E Additional Tables and Figures

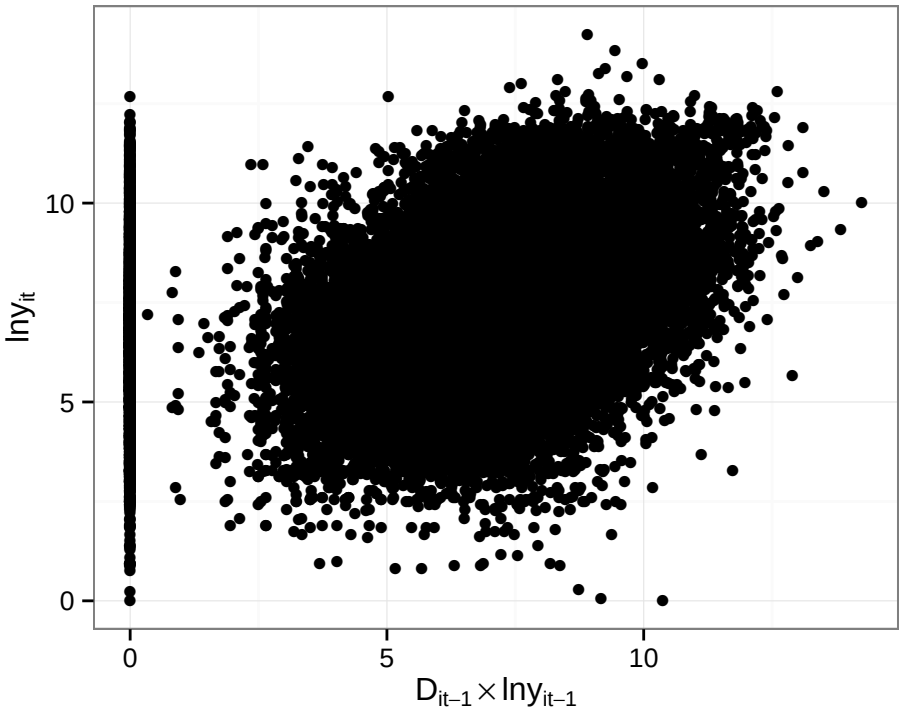


Figure 3.E.1: Plot of $\ln y_{it}$ Against $d_{it-1} \times \ln y_{it-1}$

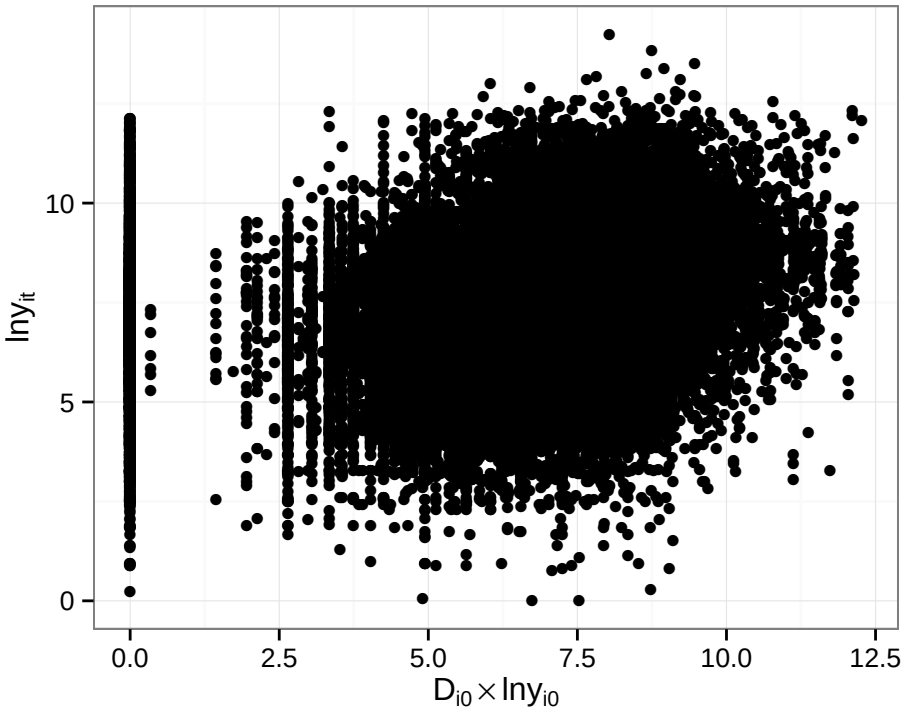


Figure 3.E.2: Plot of $\ln y_{it}$ Against $d_{i0} \times \ln y_{i0}$

3.F Additional Model Checking

A few diagnostics for longitudinal two-part models have been proposed in the literature. I use some of these to further check the fit of the model. First, as suggested by Olsen and Schafer (2001), I compare the observed and predicted logarithm of non-zero expenditures for each individual averaged across waves and the actual and predicted number of waves with positive expenditures. In both cases predictions are made using the posterior means of the relevant posterior predictive distributions. The predicted versus actual plots are shown in Figure 3.F.1. The plots don't show any major flaws with the model although the predictions tend to underestimate (overestimate) values at the upper (lower) end of the distribution. This is not unexpected since the hierarchical model partially pools the random effects.

I conducted two additional checks to ensure that the model was able to predict sample means accurately. The first test quantity is the observed proportion of nonzero observations. The second test quantity is the mean of nonzero expenditures, which is not always easy to predict because the log of expenditure needs to be transformed. This transformation can be sensitive to the lognormality assumption and the homoscedasticity assumption. Model fit is measured with a Bayesian p-value, which is just the proportion of simulations for which the test quantity is greater than or equal to observed value. P-values close to 0 or 1 indicate that the model does not capture a particular aspect of the data.

Figure 3.F.2 compares the observed test quantities, $T(y, \theta)$, to the simulated test quantities, $T(y^{sim}, \theta)$. Figure (a) graphs the simulated distribution of the percentage of observations with nonzero spending. The Bayesian p-value is equal to the shaded error below the density curve. The p-value is relatively close to 0.5 which indicates that the model predicts the number of nonzero observations adequately. (Note that this p-value is consistent with the 95% credible interval reported for the proportion of positive values of y^{sim} reported in Section 3.4.4).

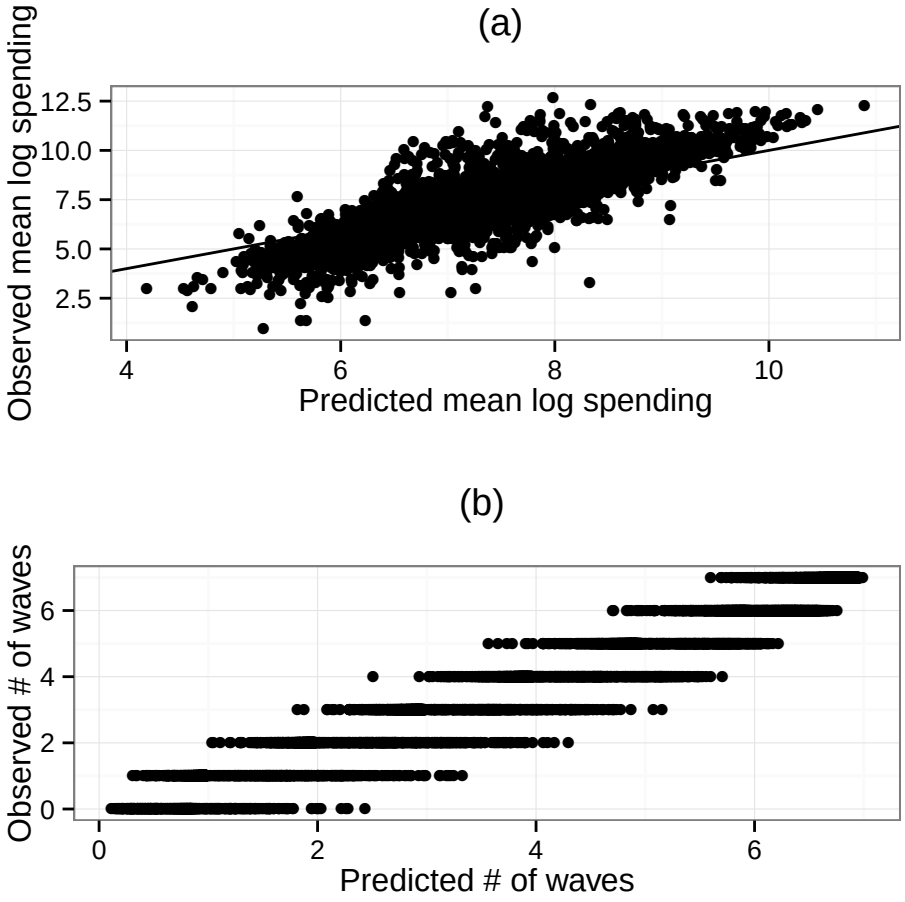


Figure 3.F.1: Actual versus Predicted Plots

Notes: Figure (a) compares observed versus expected average log-spending for each individual across waves 5 to 11. Figure (b) plots the observed number of waves with positive expenditures against expected number of waves with positive expenditures.

Figure (b) repeats figure (a) for mean expenditures. The p-value is again far from 0 or 1 which suggests that the continuous component of the expenditure model is able to simulate the sample mean fairly well.

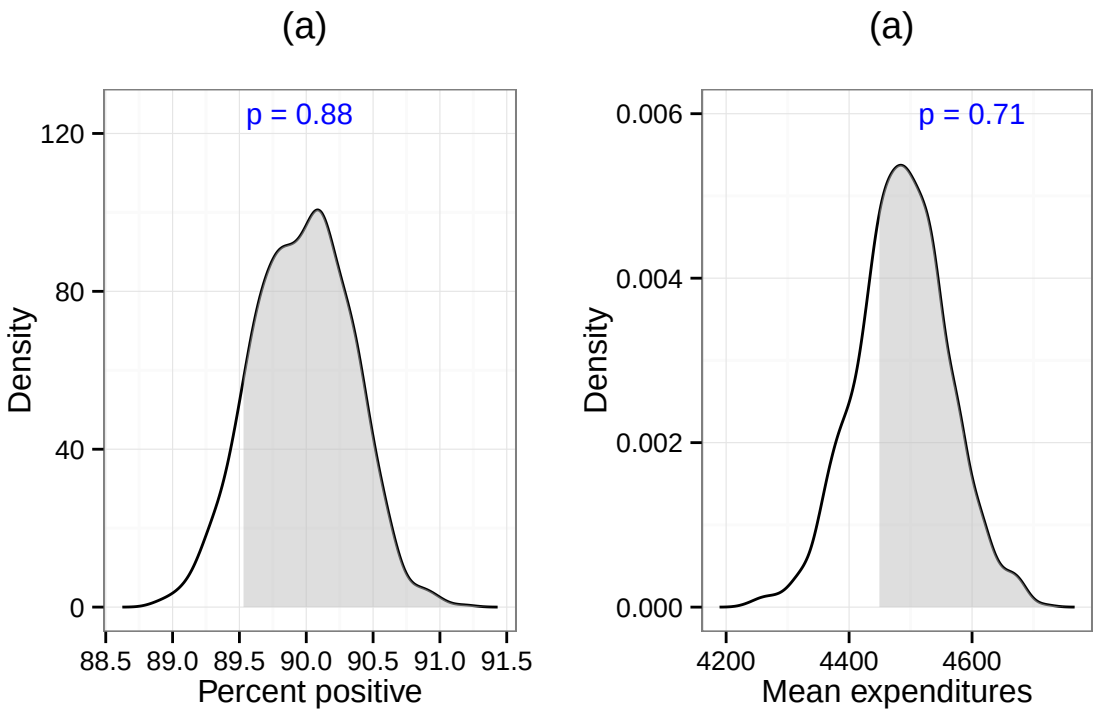


Figure 3.F.2: Posterior Predictive Checks for Summary Measures of Fit

Notes: Figure (a) plots the simulated distribution of nonzero observations and figure (b) plots simulated mean expenditures. The p-value in each figure is the proportion of simulations in which the simulated data is greater than the observed value.

References

Albert, James H and Siddhartha Chib. 1993. "Bayesian analysis of binary and polychotomous response data." *Journal of the American statistical Association* 88(422):669–679.

Almond, Douglas, Joseph J Doyle, Amanda E Kowalski and Heidi Williams. 2010. "Estimating marginal returns to medical care: Evidence from at-risk newborns." *The quarterly journal of economics* 125(2):591–634.

Ansolabehere, S., J. Snyder and J. de Figueiredo. 2003. "Why Is There So Little Money In US Elections?".

Ansolabehere, Stephen and James M Snyder Jr. 2002. "The incumbency advantage in US elections: An analysis of state and federal offices, 1942-2000." *Election Law Journal* 1(3):315–338.

Aranson, Peter H, Melvin J Hinich and Peter C Ordeshook. 1974. "Election goals and strategies: Equivalent and nonequivalent candidate objectives." *American Political Science Review* 68(01):135–152.

Aron-Dine, Aviva, Liran Einav and Amy Finkelstein. 2013. "The RAND Health Insurance Experiment, Three Decades Later." *The Journal of Economic Perspectives* pp. 197–222.

Berk, M.L. and A.C. Monheit. 2001. "The concentration of health care expenditures, revisited." *Health Affairs* 20(2):9–18.

Bernard, Didem, Cathy Cowan, Thomas Selden, Liming Cai, Aaron Catlin and Stephen Heffler. 2012. "Reconciling medical expenditure estimates from the MEPS and NHEA, 2007." *Medicare & medicaid research review* 2(4).

Bhattacharya, Jay and Darius Lakdawalla. 2006. "Does Medicare benefit the poor?" *Journal of Public Economics* 90(1):277–292.

Bickers, Kenneth N and Robert M Stein. 1996. "The electoral dynamics of the federal pork barrel." *American Journal of Political Science* pp. 1300–1326.

Blanco, Carlos, Sapana R Patel, Linxu Liu, Huiping Jiang, Roberto Lewis-Fernández, Andrew B Schmidt, Michael R Liebowitz and Mark Olfson. 2007. "National trends in ethnic disparities in mental health care." *Medical Care* 45(11):1012–1019.

Blough, David K, Carolyn W Madden and Mark C Hornbrook. 1999. "Modeling risk using generalized linear models." *Journal of health economics* 18(2):153–171.

Brams, Steven J and Morton D Davis. 1973. "Resource-Allocation Models in Presidential Campaigns: Implications for Democratic Representation." *Annals of the New York Academy of Sciences* 219(1):105–123.

Brams, Steven J and Morton D Davis. 1974. "The 3/2's rule in presidential campaigning." *American Political Science Review* 68(01):113–134.

Breyer, F., MK Bundorf and MV Pauly. 2012. "Health Care Spending Risk, Health Insurance, and Payment to Health Plans." *MV Pauly, TG McGuire, & PP Barros, Handbook of health economics* 2:691–762.

Brown, Jeffrey R and Amy Finkelstein. 2008. "The Interaction of Public and Private Insurance: Medicaid and the Long-Term Care Insurance Market." *The American Economic Review* 98(3):1083.

Cameron, A Colin, K Pravin, Ulrich Kohler, Frauke Kreuter and J Scott Long. 2009. "Microeconometrics Using Stata." *Atlanta* p. 376.

Cameron, A Colin and Pravin K Trivedi. 1990. "Regression-based tests for overdispersion in the Poisson model." *Journal of Econometrics* 46(3):347–364.

Cameron, A Colin and Pravin K Trivedi. 2005. *Microeconometrics: methods and applications*. Cambridge university press.

Campbell, James E. 1992. "Forecasting the presidential vote in the states." *American Journal of Political Science* pp. 386–407.

Centers for Medicare & Medicaid Services. N.d. "National Health Care Expenditure Data." https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/index.html. Accessed August 13, 2015.

Colantoni, Claude S, Terrence J Levesque and Peter C Ordeshook. 1975. "Campaign resource allocations under the Electoral College." *American Political Science Review* 69(01):141–154.

Collins, Karen Scott, Allyson G Hall, Charlotte Neuhaus and Commonwealth Fund. 1999. "US minority health: A chartbook.".

Cook, Benjamin Lê and Willard G Manning. 2009. "Measuring racial/ethnic disparities across the distribution of health care expenditures." *Health services research* 44(5p1):1603–1621.

Cook, Benjamin, Thomas McGuire and Jeanne Miranda. 2007. "Measuring trends in mental health care disparities, 2000–2004." *Psychiatric Services* 58(12):1533–1540.

Cooper-Patrick, Lisa, Joseph J Gallo, Junius J Gonzales, Hong Thi Vu, Neil R Powe, Christine Nelson and Daniel E Ford. 1999. "Race, gender, and partnership in the patient-physician relationship." *Jama* 282(6):583–589.

Cox, David R. 1972. "Regression models and life-tables." *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 187–220.

Cutler, David M and Adriana Lleras-Muney. 2010. "Understanding differences in health behaviors by education." *Journal of health economics* 29(1):1–28.

Cutler, D.M., A.B. Rosen and S. Vijan. 2006. "The value of medical spending in the United States, 1960–2000." *New England Journal of Medicine* 355(9):920–927.

Cutler, D.M. and M. McClellan. 2001. "Is technological change in medicine worth it?" *Health affairs* 20(5):11–29.

Dixit, Avinash and John Londregan. 1996. "The determinants of success of special interests in redistributive politics." *Journal of politics* 58:1132–1155.

Doyle, J. and J. Joseph. 2011. "Returns to Local-Area Health Care Spending: Evidence from Health Shocks to Patients Far From Home." *American Economic Journal: Applied Economics* 3(3):221–243.

Dranove, David. 2009. *The economic evolution of American health care: from Marcus Welby to managed care*. Princeton University Press.

Duan, Naihua. 1983. "Smearing estimate: a nonparametric retransformation method." *Journal of the American Statistical Association* 78(383):605–610.

Duan, Naihua, Willard G Manning, Carl N Morris and Joseph P Newhouse. 1983. "A comparison of alternative models for the demand for medical care." *Journal of business & economic statistics* 1(2):115–126.

Eichner, M.J., M.B. McClellan and D.A. Wise. 1997. Health expenditure persistence and the feasibility of medical savings accounts. In *Tax Policy and the Economy, Volume 11*. MIT Press pp. 91–128.

Engelhardt, Gary V and Jonathan Gruber. 2011. "Medicare Part D and the financial protection of the elderly." *American Economic Journal: Economic Policy* 3(4):77–102.

Erikson, Robert S. 1988. "The puzzle of midterm loss." *The Journal of Politics* 50(04):1011–1029.

Erikson, Robert S and Thomas R Palfrey. 2000. "Equilibria in campaign spending games: Theory and data." *American Political Science Review* 94(03):595–609.

Feenberg, Daniel, Jonathan Skinner et al. 1994. "The Risk and Duration of Catastrophic Health Care Expenditures." *The Review of Economics and Statistics* 76(4):633–47.

Fernández-Villaverde, Jesús and Dirk Krueger. 2007. "Consumption over the life cycle: Facts from consumer expenditure survey data." *The Review of Economics and Statistics* 89(3):552–565.

Finkelstein, Amy and Robin McKnight. 2008. "What did Medicare do? The initial impact of Medicare on mortality and out of pocket medical spending." *Journal of public economics* 92(7):1644–1668.

French, Eric and John Bailey Jones. 2004. "On the distribution and dynamics of health care costs." *Journal of Applied Econometrics* 19(6):705–721.

Gelman, Andrew and Gary King. 1990. "Estimating incumbency advantage without bias." *American Journal of Political Science* pp. 1142–1164.

Gelman, Andrew and Gary King. 1993. "Why are American presidential election campaign polls so variable when votes are so predictable?" *British Journal of Political Science* 23(04):409–451.

Gelman, Andrew and Gary King. 1994. "A unified method of evaluating electoral systems and redistricting plans." *American Journal of Political Science* pp. 514–554.

Gelman, Andrew, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari and Donald B Rubin. 2013. *Bayesian data analysis*. CRC press.

Gelman, Andrew and Zaiying Huang. 2008. "Estimating incumbency advantage and its variation, as an example of a before–after study." *Journal of the American Statistical Association* 103(482):437–446.

Gerber, Alan. 1998. "Estimating the effect of campaign spending on senate election outcomes using instrumental variables." *American Political Science Review* 92(02):401–411.

Glied, Sherry and Adriana Lleras-Muney. 2008. "Technological innovation and inequality in health." *Demography* 45(3):741–761.

Goldman, Dana P and James P Smith. 2002. "Can patient self-management help explain the SES health gradient?" *Proceedings of the National Academy of Sciences* 99(16):10929–10934.

Goldman, Dana P, Julie Zissimopoulos and Yang Lu. 2011. Medical Expenditure Measures in the Health and Retirement Study. In *Forum for health economics & policy*. Vol. 14.

Gourinchas, Pierre-Olivier and Jonathan A Parker. 2002. "Consumption over the life cycle." *Econometrica* 70(1):47–89.

Green, Donald Philip and Jonathan S Krasno. 1988. "Salvation for the spendthrift incumbent: Reestimating the effects of campaign spending in House elections." *American Journal of Political Science* pp. 884–907.

Grier, Kevin B and Michael C Munger. 1991. "Committee assignments, constituent preferences, and campaign contributions." *Economic Inquiry* 29(1):24–43.

Gross, Cary P, Benjamin D Smith, Elizabeth Wolf and Martin Andersen. 2008. "Racial disparities in cancer therapy." *Cancer* 112(4):900–908.

Hartman, Micah, Aaron Catlin, David Lassman, Jonathan Cylus and Stephen Heffler. 2008. "US health spending by age, selected years through 2004." *Health Affairs* 27(1):w1–w12.

Herrnson, Paul S. 2009. "The roles of party organizations, party-connected committees, and party allies in elections." *The Journal of Politics* 71(04):1207–1224.

Hsiao, Cheng. 2014. *Analysis of panel data*. Vol. 54 Cambridge university press.

Hurd, Michael D and Susann Rohwedder. 2009. "The level and risk of out-of-pocket health care spending." *Michigan Retirement Research Center Working Paper* 218.

Ibrahim, Joseph G, Haitao Chu and Liddy M Chen. 2010. "Basic concepts and methods for joint models of longitudinal and survival data." *Journal of Clinical Oncology* 28(16):2796–2801.

Jackman, Simon. 2005. "Pooling the polls over an election campaign." *Australian Journal of Political Science* 40(4):499–517.

Jackman, Simon. 2009. *Bayesian analysis for the social sciences*. Vol. 846 John Wiley & Sons.

Jackman, Simon. 2012. "Pollster Predictions: 91.4% Chance Obama Wins, 303 or 332 EVs." http://www.huffingtonpost.com/simon-jackman/pollster-predictions_b_2081013.html.

Jacobson, Gary C. 1978. "The effects of campaign spending in congressional elections." *American Political Science Review* 72(02):469–491.

Jacobson, Gary C. 1980. *Money in congressional elections*. Yale University Press.

Jacobson, Gary C. 1985*a*. "Money and votes reconsidered: Congressional elections, 1972–1982." *Public choice* 47(1):7–62.

Jacobson, Gary C. 1985*b*. "Party organization and distribution of campaign resources: Republicans and Democrats in 1982." *Political Science Quarterly* pp. 603–625.

Johansson, Eva. 2003. "Intergovernmental grants as a tactical instrument: empirical evidence from Swedish municipalities." *Journal of Public Economics* 87(5):883–915.

Johnson, Rachel L, Debra Roter, Neil R Powe and Lisa A Cooper. 2004. "Patient race/ethnicity and quality of patient-physician communication during medical visits." *American journal of public health* 94(12):2084–2090.

Kastellec, Jonathan P, Andrew Gelman and Jamie P Chandler. 2008. "Predicting and dissecting the seats-votes curve in the 2006 US House election." *PS: Political Science & Politics* 41(01):139–145.

Kau, James B, Donald Keenan and Paul H Rubin. 1982. "A general equilibrium model of congressional voting." *The Quarterly Journal of Economics* pp. 271–293.

Keehan, Sean P. 2006. "Reconciling medical expenditure estimates from the MEPS and NHEA, 2002." *Health Care Financing Review* 28(1):25.

King, Gary and Andrew Gelman. 1991. "Systemic consequences of incumbency advantage in US House elections." *American Journal of Political Science* pp. 110–138.

Kinsey, Tracy, Ahmedin Jemal, Jonathan Liff, Elizabeth Ward and Michael Thun. 2008. "Secular trends in mortality from common cancers in the United States by educational attainment, 1993–2001." *Journal of the National Cancer Institute* 100(14):1003–1012.

Koop, Gary, Dale J Poirier and Justin L Tobias. 2007. *Bayesian econometric methods*. Cambridge University Press.

Kopczuk, Wojciech and Joseph P Lupton. 2007. "To leave or not to leave: The distribution of bequest motives." *The Review of Economic Studies* 74(1):207–235.

Kowalski, Amanda. 2015. "Censored quantile instrumental variable estimates of the price elasticity of expenditure on medical care." *Journal of Business & Economic Statistics* (just-accepted):00–00.

Kroszner, Randall S and Thomas Stratmann. 1998. "Interest-group competition and the organization of congress: theory and evidence from financial services' political action committees." *American Economic Review* pp. 1163–1187.

Lee, David S. 2001. The Electoral Advantage to Incumbency and Voters' Valuation of Politicians' Experience: A Regression Discontinuity Analysis of Elections to the US.. Technical report National bureau of economic research.

Levitt, Steven D. 1994. "Using repeat challengers to estimate the effect of campaign spending on election outcomes in the US House." *Journal of Political Economy* pp. 777–798.

Li, Tong and Xiaoyong Zheng. 2008. "Semiparametric Bayesian inference for dynamic Tobit panel data models with unobserved heterogeneity." *Journal of Applied Econometrics* 23(6):699–728.

Lindbeck, Assar and Jörgen W Weibull. 1987. "Balanced-budget redistribution as the outcome of political competition." *Public choice* 52(3):273–297.

Link, B.G. and J. Phelan. 1995. "Social conditions as fundamental causes of disease." *Journal of health and social behavior* pp. 80–94.

Linzer, Drew A. 2013. "Dynamic Bayesian Forecasting of Presidential Elections in the States." *Journal of the American Statistical Association* 108(501):124–134.

Lleras-Muney, Adriana and Frank R Lichtenberg. 2002. The effect of education on medical

technology adoption: are the more educated more likely to use new drugs. Technical report National Bureau of Economic Research.

Lock, Kari and Andrew Gelman. 2010. "Bayesian combination of state polls and election forecasts." *Political Analysis* 18(3):337–348.

Lynch, Scott M and J Scott Brown. 2005. "A new approach to estimating life tables with covariates and constructing interval estimates of life table quantities." *Sociological Methodology* 35(1):177–225.

Manning, W.G., J.P. Newhouse, N. Duan, E.B. Keeler and A. Leibowitz. 1987. "Health insurance and the demand for medical care: evidence from a randomized experiment." *The American Economic Review* pp. 251–277.

Manning, Willard G, Anirban Basu and John Mullahy. 2005. "Generalized modeling approaches to risk adjustment of skewed outcomes data." *Journal of health economics* 24(3):465–488.

Marshall, Samuel, Kathleen McGarry and Jonathan S Skinner. 2011. The risk of out-of-pocket health care expenditure at the end of life. In *Explorations in the Economics of Aging*. University of Chicago Press pp. 101–128.

Martinez, Steve R, Anthony S Robbins, Frederick J Meyers, Richard J Bold, Vijay P Khatri and James E Goodnight. 2008. "Racial and ethnic differences in treatment and survival among adults with primary extremity soft-tissue sarcoma." *Cancer* 112(5):1162–1168.

Mayberry, Robert M, Fatima Mili and Elizabeth Ofili. 2000. "Racial and ethnic differences in access to medical care." *Medical Care Research and Review* 57(4 suppl):108–145.

McClellan, Mark and Jonathan Skinner. 2006. "The incidence of Medicare." *Journal of Public Economics* 90(1):257–276.

Monheit, A.C. 2003. "Persistence in health expenditures in the short run: prevalence and consequences." *Medical Care* 41(7):III.

Neelon, Brian, A James O'Malley and Sharon-Lise T Normand. 2011. "A Bayesian Two-Part Latent Class Model for Longitudinal Medical Expenditure Data: Assessing the Impact of Mental Health and Substance Abuse Parity." *Biometrics* 67(1):280–289.

Neelon, Brian H, A James O'Malley and Sharon-Lise T Normand. 2010. "A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use." *Statistical Modelling* 10(4):421–439.

Newhouse, J.P. 2004. *Pricing the priceless: a health care conundrum*. Vol. 2 MIT Press.

Olsen, Maren K and Joseph L Schafer. 2001. "A two-part random-effects model for semicontinuous longitudinal data." *Journal of the American Statistical Association* 96(454):730–745.

Poole, Keith T and Howard L Rosenthal. 2011. *Ideology and congress*. Vol. 1 Transaction Publishers.

Poole, Keith T and Howard Rosenthal. 1997. *Congress: A political-economic history of roll call voting*. Oxford University Press.

Poole, Keith T and Thomas Romer. 1985. "Patterns of political action committee contributions to the 1980 campaigns for the United States House of Representatives." *Public Choice* 47(1):63–111.

Pope, Gregory C, John Kautter, Randall P Ellis, Arlene S Ash, John Z Ayanian, Melvin J Ingber, Jesse M Levy, John Robst et al. 2004. "Risk adjustment of Medicare capitation payments using the CMS-HCC model.".

Puhani, Patrick. 2000. "The Heckman correction for sample selection and its critique." *Journal of economic surveys* 14(1):53–68.

Robst, John, Jesse M Levy and Melvin J Ingber. 2007. "Diagnosis-based risk adjustment for medicare prescription drug plan payments." *Health Care Financing Review* 28(4):15.

Romer, Thomas and James M Snyder Jr. 1994. "An empirical investigation of the dynamics of PAC contributions." *American journal of political science* pp. 745–769.

Shefrin, Hersh M and Richard H Thaler. 1988. "The behavioral life-cycle hypothesis." *Quasi Rational Economics* pp. 91–126.

Siebert, Uwe, Oguzhan Alagoz, Ahmed M Bayoumi, Beate Jahn, Douglas K Owens, David J Cohen, Karen M Kuntz, ISPOR-SMDM Modeling Good Research Practices Task Force et al. 2012. "State-transition modeling: a report of the ISPOR-SMDM modeling good research practices task force-3." *Value in Health* 15(6):812–820.

Smedley, Brian D, Adrienne Y Stith and Alan R Nelson. 2003. "Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care.".

Snyder, James M. 1989. "Election goals and the allocation of campaign resources." *Econometrica: Journal of the Econometric Society* pp. 637–660.

Snyder Jr, James M. 1990. "Campaign contributions as investments: The US House of Representatives, 1980-1986." *Journal of Political Economy* pp. 1195–1227.

St Clair, Patricia, Delia Bugliari, Philip Pantoja, Serhii Ilchuk, Gabriela Lopez, Steven Haider, Michael Hurd, David Loughran, Constantijn Panis, Monika Reti et al. 2014. "RAND HRS Data Documentation Version N." *RAND Corporation, Santa Monica* .

Stanton, Mark W and MK Rutherford. 2006. *The high concentration of US health care expenditures*. Agency for Healthcare Research and Quality Washington, DC.

Stein, Robert M and Kenneth N Bickers. 1994. "Congressional elections and the pork barrel." *The Journal of Politics* 56(02):377–399.

Stewart III, Charles and Jonathan Woon. 2015. "Congressional Committee Assignments, 103rd to 112th Congresses, 1993–2011.".

Stratmann, Thomas. 1991. "What do campaign contributions buy? Deciphering causal effects of money and votes." *Southern Economic Journal* pp. 606–620.

Stratmann, Thomas. 2005. "Some Talk: Money in Politics. A (Partial) Review of the Literature." *Public Choice* pp. 135–156.

Strauss, Aaron. 2007. Florida or Ohio? Forecasting presidential state outcomes using reverse random walks. In *Princeton University Political Methodology Seminar*.

Strömberg, David. 2008. "How the Electoral College influences campaigns and policy: the probability of being Florida." *The American Economic Review* 98(3):769–807.

Su, Li, Brian DM Tom and Vernon T Farewell. 2009. "Bias in 2-part mixed models for longitudinal semicontinuous data." *Biostatistics* 10(2):374–389.

Thomas, John, D Johniene Thomas, Thomas Pearson, Michael Klag and Lucy Mead. 1997. "Cardiovascular disease in African American and white physicians: The Meharry cohort and Meharry-Hopkins cohort studies." *Journal of health care for the poor and underserved* 8(3):270–283.

Tooze, Janet A, Gary K Grunwald and Richard H Jones. 2002. "Analysis of repeated measures data with clumping at zero." *Statistical methods in medical research* 11(4):341–355.

Tufte, Edward R. 1973. "The relationship between seats and votes in two-party systems." *American Political Science Review* 67(02):540–554.

Van Vliet, René CJA. 1992. "Predictability of individual health care expenditures." *Journal of Risk and Insurance* pp. 443–461.

Waters, H.R., G.F. Anderson and J. Mays. 2004. "Measuring financial protection in health in the United States." *Health Policy* 69(3):339–349.

Webb, Anthony and Natalia Zhivan. 2010. How Much Is Enough? The Distribution of Lifetime Health Care Costs. Technical report Center for Retirement Research.

Weinick, Robin M, Samuel H Zuvekas and Joel W Cohen. 2000. "Racial and ethnic differences in access to and use of health care services, 1977 to 1996." *Medical Care Research and Review* 57(suppl 1):36–54.

West, Mike and Jeff Harrison. 1997. *Bayesian forecasting and dynamic models*. Vol. 18 Springer New York.

Whittle, Jeff, Joseph Conigliaro, CB Good and Monica Joswiak. 1997. "Do patient preferences contribute to racial differences in cardiovascular procedure use?" *Journal of general internal medicine* 12(5):267–273.

Wooldridge, Jeffrey M. 2005. "Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity." *Journal of applied econometrics* 20(1):39–54.

Wulfsohn, Michael S and Anastasios A Tsiatis. 1997. "A joint model for survival and longitudinal data measured with error." *Biometrics* pp. 330–339.

Zhang, Min, Robert L Strawderman, Mark E Cowen and Martin T Wells. 2006. "Bayesian inference for a two-part hierarchical model: An application to profiling providers in managed health care." *Journal of the American Statistical Association* 101(475):934–945.

Zuvekas, Samuel H and Gary L Olin. 2009. "Validating household reports of health care use in the medical expenditure panel survey." *Health Services Research* 44(5p1):1679–1700.

Zuvekas, Samuel H and Joel W Cohen. 2007. "Prescription drugs and the changing concentration of health care expenditures." *Health Affairs* 26(1):249–257.