Essays on Medical Care Using Semiparametric and Structural Econometrics

by

Amanda Ellen Kowalski
A.B. Economics
Harvard College, 2003

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Economics
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June 2008

Author .....
Department of Economics
May 15, 2008

Certified by .....
Jonathan Gruber
Professor of Economics
Thesis Supervisor

Certified by .....
Jerry A. Hausman
John and Jennie S. MacDonald Professor
Thesis Supervisor

Accepted by .....
Peter Temin
Elisha Gray II Professor of Economics
Chairman, Department Committee on Graduate Students

Essays on Medical Care Using Semiparametric and Structural Econometrics

by

Amanda Ellen Kowalski

Submitted to the Department of Economics
on May 15, 2008, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Economics

Abstract

This dissertation consists of an empirical chapter, an econometrics chapter, and a theoretical chapter, all of which advance the study of the price elasticity of expenditure on medical care. In Chapter 1, I estimate the price elasticity of expenditure on medical care across the quantiles of the expenditure distribution. My identification strategy relies on family cost sharing provisions that generate differences in marginal prices between individuals who have injured family members and individuals who do not. I use a new censored quantile instrumental variables (CQIV) estimator, which allows me to examine variations in price responsiveness across the skewed distribution of medical expenditure. The CQIV estimator does not require any parametric assumptions to account for individuals who consume zero medical care. Using CQIV, as well as traditional estimators, I find elasticities that are an order of magnitude larger than those in the literature. My CQIV estimates suggest strong price responsiveness among people who spend the most. I find that the price elasticity of expenditure is approximately -2.3, which is stable across the .65 to .95 quantiles of the expenditure distribution.

In Chapter 2, Chernozhukov and Kowalski (2008), we develop a censored quantile instrumental variables (CQIV) estimator. The CQIV estimator handles censoring nonparametrically in the tradition of Powell (1986), and it generalizes standard censored quantile regression (CQR) methods to incorporate endogeneity. Our computational algorithm combines a control function approach with the Chernozhukov and Hong (2002) CQR algorithm. Through Monte-Carlo simulation, we show that CQIV performs well relative to Tobit IV in terms of median bias and interquartile range.

In Chapter 3, I develop a structural model to estimate the price elasticity of expenditure on medical care. The model relies on deductibles, coinsurance rates, and stoplosses that generate nonlinearities in consumer budget sets. The model generalizes existing nonlinear budget set models by allowing for more than one nonconvex kink. Furthermore, it incorporates censoring as a corner solution. Unlike reduced form models, the model utilizes identification from utility theory, it allows for preference heterogeneity, and it allows for the direct calculation of welfare effects.

Thesis Supervisor: Jonathan Gruber
Title: Professor of Economics

Thesis Supervisor: Jerry A. Hausman
Title: John and Jennie S. MacDonald Professor

Acknowledgments

I am very grateful to my advisors Jonathan Gruber, Jerry Hausman, and Amy Finkelstein. Working with them has been the highlight of my experience at MIT. My dissertation and my developing career have been greatly enriched by their passion for economics, their ability to see the big picture, and their accessibility. Victor Chernozhukov, who has coauthored the second paper of this dissertation, also deserves special thanks.

The ideas in this dissertation have been greatly enhanced by discussion with faculty members at several schools as well as fellow graduate students. Specifically, I would like to thank the following individuals for thoughtful comments and enjoyable conversation: Michael Anderson, Joshua Angrist, David Autor, John Beshears, Soren Blomquist, Tonja Bowen-Bishop, David Card, Amitabh Chandra, David Cutler, Ian Dew-Becker, Deepa Dhume, Peter Diamond, Joseph Doyle, Esther Duflo, Jesse Edgerton, Matthew Eichner, Brigham Frandsen, John Friedman, Michael Greenstone, Raymond Guiteras, Matthew Harding, Naomi Hausman, Panle Jia, Lisa Kahn, Jonathan Kolstad, Whitney Newey, Joseph Newhouse, Douglas Norton, Christopher Nosko, Matthew Notowidigdo, James Poterba, David Powell, Andrew Samwick, David Seif, Mark Showalter, Erin Strumpf, Heidi Williams, and the residents of Lowell House at Harvard. Participants at the MIT econometrics and public finance lunches, the MIT public finance seminar, and the NBER health and aging fellow lunch have provided helpful feedback. I would like to extend a special thanks to my fellow graduate student, Hui Shan, who has motivated me with her enthusiasm, diligence, and friendship.

I have been very fortunate to have conducted much of my dissertation research at the National Bureau of Economic Research. The computing and data expertise of Mohan Ramanujan and Jean Roth made the empirical work possible. Funding from the National Institute on Aging to the NBER, Grant Number T32-AG00186, is gratefully acknowledged.

My parents, Dr. Donald and Maryann Kowalski, and my sister, Dawna Kowalski, have been an unwavering source of support. I dedicate this dissertation to them.

Contents

List of Figures

List of Tables

Introduction

Evaluation of the latest wave of cost control and patient empowerment initiatives, broadly known as "consumer-directed care," depends critically on the underlying price elasticity of expenditure on medical care. The extent to which price responsiveness varies with expenditure also has important policy implications. This dissertation consists of an empirical chapter, an econometrics chapter, and a theoretical chapter, all of which advance the study of the price elasticity of expenditure on medical care. The first two chapters employ and develop semiparametric econometric methods, and the third chapter develops a structural econometric method.

In Chapter 1, I estimate the price elasticity of expenditure on medical care across the quantiles of the expenditure distribution using detailed claims and enrollment data on individuals in employer-sponsored health insurance plans. To examine the effect of marginal price on expenditure while avoiding mechanical relationships induced by cost sharing parameters, I use an instrumental variables strategy. Identification in my strategy relies on family-level cost sharing provisions, which generate differences in marginal prices between individuals who have injured family members and individuals who do not have injured family members. Because a large number of individuals consume no medical care, my strategy requires the use of estimators that account for censoring in the dependent variable. However, traditional censored estimators require strong distributional assumptions. In response to this concern, I use a new censored quantile instrumental variables (CQIV) estimator. The CQIV estimator handles censoring without any distributional assumptions, and it allows me to examine variations in price responsiveness across the skewed distribution of medical expenditure. Using CQIV, as well as several traditional censored estimators for comparative purposes,

I find elasticities that are an order of magnitude larger than those in the literature. In particular, my CQIV estimates suggest strong price responsiveness among people who spend the most. I find that the price elasticity of expenditure is approximately -2.3, which is stable across the .65 to .95 quantiles of the expenditure distribution.

In Chapter 2, Chernozhukov and Kowalski (2008), we develop a new censored quantile instrumental variables (CQIV) estimator and describe its properties and computation. The CQIV estimator handles censoring semi-parametrically in the tradition of the Powell (1986), and it generalizes standard censored quantile regression (CQR) methods to incorporate endogeneity in a manner that is computationally tractable. Our computational algorithm combines a control function approach with the CQR estimator developed by Chernozhukov and Hong (2002). Through Monte-Carlo simulation, we show that CQIV performs well relative to Tobit IV in terms of median bias and interquartile range in a model that satisfies the parametric assumptions required for Tobit IV to be efficient. Given the strong parametric assumptions and the constant coefficients required by Tobit IV, the gains to CQIV relative to Tobit IV are likely to be large in empirical applications.

In Chapter 3, using the theory of utility maximization subject to a nonlinear constraint, I develop a structural model to estimate the price elasticity of expenditure on medical care among people with traditional health insurance policies. The model relies on deductibles, coinsurance rates, and stoplosses that generate nonlinearities in consumer budget sets. Relative to reduced form demand models, the model utilizes additional identification from utility theory, it describes behavior consistent with the functional form of the demand function, it allows for preference heterogeneity across agents, and it allows for the direct calculation of the welfare effects of price changes. Furthermore, it incorporates censoring in a manner that is consistent with a corner solution decision to consume zero care. The model generalizes existing nonlinear budget set models by allowing for more than one nonconvex kink. Relative to other nonlinear budget set applications, the medical care application allows for a particularly tight link between the agent's actual budget set, the model, and the estimation strategy.

Chapter 1

Censored Quantile Instrumental Variables Estimates of the Price Elasticity of Expenditure on Medical Care

1.1 Introduction

Spending on medical care is increasing by almost any metric, and yet the effects of consumer price on medical care utilization are not well understood. Even so, the most recent wave of cost control initiatives in medical care depends on consumer responsiveness to price. An understanding of the price elasticity of expenditure on medical care and the extent to which it varies across the expenditure distribution is crucial in evaluating these initiatives.

The first wave of cost control in medical care, which began in the 1960's, was also predicated on consumer responsiveness to price. It used cost sharing mechanisms such as deductibles and coinsurance rates to limit spending on the demand side. In contrast, in the 1980's, the second wave of cost control, known as managed care, took the form of constraints on the supply side. Managed care limited patient choice to

certain provider networks and relied on doctors in those networks to manage spending. Although spending levels temporarily decreased, managed care did not curb spending growth, and it caused consumer outcry against limited access to procedures.

In response to managed care, a third wave of cost control is currently in its nascent stages. This wave of cost control falls broadly under the name of "consumer-directed care." Like the first wave, consumer-directed care imposes constraints on the demand side by encouraging individuals to purchase high deductible policies and pay for routine expenses out of pocket. Although the ideas of consumer-directed care have been circulating for several years (see Eichner, McClellan, and Wise (1997)), they have just recently gained traction. The Medicare Modernization Act, enacted in December 2003, included provisions for the establishment of health savings accounts (HSA's), tax-advantaged accounts that can only be opened by the holders of qualified high deductible policies.

The rationale behind the encouragement of high deductible policies is that when the deductible increases, some consumers will pay a higher marginal price for the last dollar of care that they consume. In a simple model of medical care consumption, if consumers know that they will meet the deductible by the end of the year, the amount of the deductible should not affect consumption because the amount paid toward the deductible will induce a pure income effect. Consumers should consume throughout the year based on the marginal price that they expect to face at the end of the year. However, if an increase in the deductible ensures that some individuals will no longer meet the deductible, making their expected year-end price higher, total yearly medical expenditures for these individuals should decrease. Given the rationale behind the encouragement of HSA's, the policy-relevant parameter is the responsiveness of total yearly medical expenditures to variation in the year-end marginal price of care, and this is precisely the parameter that I estimate.

However, the price elasticity of expenditure need not be constant. In particular, it could change with the quantiles of the expenditure distribution. Specifically, the effectiveness of the mechanisms that consumers use to manipulate medical expenditure could vary with the level of expenditure. For example, for minor ailments,

consumers could limit expenditure by deciding not to go to the doctor, but for more serious ailments, consumers could limit expenditure by choosing a less expensive hospital or a more minimalist treatment plan. If consumers become less price responsive as expenditure increases, policies aimed at making consumers more price responsive will have a limited impact because consumers with the largest expenditure are responsible for the most dollars of medical spending. However, if consumers remain price responsive as expenditure increases, policies that increase cost sharing for high spenders are likely to be more effective at reducing overall medical spending. It should be noted, though, that the optimality of such policies also depends on the health impact of foregone expenditure and the value of insurance as expenditure increases. In this paper, I focus only on measuring price responsiveness across the quantiles of the expenditure distribution.

In the existing economics literature, there is evidence of a modest price elasticity of expenditure on medical care, which is assumed constant across the expenditure distribution. This evidence comes from three generations of estimates: the RAND health insurance experiment of the 1970's, an approach using medical claims by Eichner (1997, 1998), and recent reduced form studies that focus on the price elasticity of demand for prescription drugs. My strategy builds on these three generations of estimates.

The RAND health insurance experiment, which took place between 1974 and 1982, randomized subjects into health insurance plans of varying generosity and estimated the price elasticity of expenditure on medical care to be around -.2. See Manning et al. (1987), Keeler and Rolph (1988), Keeler (1992), and Newhouse (1993) for a discussion of methods. However, to induce people to participate in the study, researchers set a very low cap on out-of-pocket costs. Since over 70% of subjects with inpatient spending exceeded this cap, it is difficult to assess their price sensitivity. Through the use of recent data on plans with less generous cost sharing parameters, I can observe meaningful price variation for the current policy environment.

Eichner (1997, 1998) used medical claims data to estimate the price elasticity of expenditure on medical care to be approximately -.3. The main innovation of his

approach was his instrument. I capture the spirit of his instrument here, but I modify it to take more family interactions into account and to increase the plausibility of the exclusion restriction. Because of data limitations, Eichner was not able to observe people who did not submit any claims, which could result in bias. In my data, I observe people with zero medical expenditure, and I include them in my analysis through the use of a new censored quantile instrumental variables (CQIV) estimator.

Recent reduced form studies such as Li et al. (2005), Hsu et al. (2006), and Chandra et al. (2006) use changes in cost sharing provisions over time to estimate the price elasticity of demand for prescription drugs and medical care among the elderly. Unlike these studies, I focus on the price elasticity of expenditure on all medical care in the non-elderly population. In their estimates, Li et al. (2005) use a price index for simplicity, but my strategy allows for a tighter link between expenditure and marginal price.

Though my study is most comparable to the aforementioned partial equilibrium analyses, it should also be noted that some general equilibrium evidence exists. Specifically, Finkelstein (2007) examines the price elasticity of expenditure on medical care in response to the establishment of Medicare. She finds much larger elasticities than those in the partial equilibrium literature.

The primary goal of my paper is to measure how consumers respond to the marginal year-end prices that they face for medical care within employer-sponsored health insurance plans. Specifically, I am interested in how this responsiveness varies with quantiles of the expenditure distribution. As a starting point, in Section 2, I explain how the cost sharing provisions of traditional employer-sponsored health insurance plans govern the marginal prices paid by individuals and families. Given the traditional cost sharing provisions and the specific provisions that govern family policies, I identify a subset of individuals who, because of injuries to family members, face lower prices than they would otherwise. My instrumental variables identification strategy, which I discuss in detail in Section 3, formalizes the comparison of expenditures between these individuals and other individuals whose marginal prices are unaffected.

To execute my strategy, I use recent, comprehensive, longitudinal data on medical claims, which I describe in Section 4. I focus on the plans offered by one employer because doing so allows me to isolate variation in cost sharing provisions from other plan attributes. Furthermore, one of the offered plans has a $1,000 deductible, which is coincidentally the initial qualifying amount for a plan to be considered eligible for use with a health savings account in the 2003 legislation. Because all plans have deductibles that are presumably large enough to be economically meaningful, and plans only differ in the deductible and stoploss, I can use data from all plans and isolate within-plan price variation for identification.

The underlying variation that I use for identification is so pronounced that I can illustrate it in simple graphs, which I discuss along with formal results in Section 5. The transformation of the underlying variation into an elasticity estimate requires a censored estimator because approximately 40% of individuals in my sample consume no medical care in the entire year. I present results from three standard censored estimators: a truncated model, a two-part model, and a Tobit model. However, these estimators require strong distributional assumptions, and the impact of these assumptions on the results is unclear.

In response to this concern, I use a new censored quantile instrumental variables estimator which is semiparametric in the sense that it does not require any distributional assumptions to handle censoring. The CQIV estimator is a generalization of Powell's censored quantile estimator that incorporates instrumental variables in a way that is computationally tractable. Chernozhukov and Kowalski (2008) develop the CQIV estimator, which has its first application here. Relative to mean estimators, CQIV is particularly advantageous in my application because medical expenditures are so skewed: 25% of individuals account for 94.5% of expenditures in my main sample. The CQIV estimator allows me to examine price responsiveness at the upper quantiles of the medical expenditure distribution.

My estimates of the price elasticity of expenditure on medical care are much larger than those in the literature. I find that the people who spend the most on medical care are particularly price responsive. Across the .65 to .95 quantiles of the

expenditure distribution, the price elasticity of expenditure is stable around -2.3, with a point-wise 95% confidence interval at the .80 quantile of -2.7 to -2.0. In Section 6, I perform several refinements to main specification, and my findings do not change. I also conduct two robustness tests in Section 7 that go beyond the main specification to examine the validity of the instrumental variables strategy, and I find that my results are robust. In an extension of my main results in Section 8, I find evidence of strong complementarity between prescription drug expenditures and expenditures on other types of medical care. In Section 9, I provide empirical and simulation evidence to explain the discrepancy between my estimates and the RAND estimates. In Section 10, I conclude and discuss implications for future research.

1.2 Background: Marginal Pricing for Medical Care

Traditional employer-sponsored health insurance plans have three major cost sharing parameters: a deductible, a coinsurance rate, and a stoploss. The "deductible" is the amount that the consumer must pay before the insurer makes any payments. Before reaching the deductible, the consumer pays one dollar for one dollar of care, so the marginal price is one. After meeting the deductible, the insurer pays a fractional amount for each dollar of care, and the consumer pays the rest. The marginal price that the consumer pays is known as the "coinsurance rate." After the consumer has paid the deductible and a fixed amount in coinsurance, the consumer reaches the "stoploss," and the insurer pays all expenses. For consumers that have met the stoploss, the marginal price is zero. Figure 1-1 depicts how the deductible, coinsurance rate, and stoploss induce a nonlinear relationship between the total amount paid by the consumer and the total amount paid by the consumer plus the insurer. The consumer faces three distinct marginal prices, depicted as the slope of each segment. The intercepts on each axis are exact for a consumer insured as an individual with no family members, but they can move toward the origin for a consumer insured as

part of a family.

If a consumer is insured as a member of a family, the general cost sharing structure is the same, but an additional family-level deductible and stoploss enable one family member's spending to affect another family member's marginal price. As a concrete example, suppose that a plan has an individual deductible of $500, and it also has a family deductible that is three times the individual deductible ($1,500). Each family member must meet the individual deductible unless total family spending toward individual deductibles exceeds the family deductible. Since the family deductible is three times the individual deductible, if a family has fewer than four members, all family members must meet the individual deductible. In a family of four, when the first, second, and third family members go to the doctor, they each face the individual deductible of $500, and then they pay according to the coinsurance rate, as if they were insured as individuals. However, when the fourth family member goes to the doctor, if the family deductible of $1,500 has been met through the fulfillment of three individual $500 deductibles, he makes his first payment at the coinsurance rate. In families with more than four members, the family deductible is fixed at $1,500, and it can be met by any combination of payments toward individual $500 deductibles. A similar interaction occurs at the level of the stoploss. Given the family-level cost sharing parameters, some individuals will face lower marginal prices than their own medical spending would dictate.

The marginal price variation induced by the family cost sharing parameters suggests a simple way to study price responsiveness by comparing expenditures of individuals whose families have and have not met the family deductible. The flaw with this type of identification strategy is that individuals in families that have met the family deductible may be more likely to consume medical care for reasons unrelated to its price, such as contagious illnesses or hereditary diseases. For this reason, instead of comparing individuals according to whether or not their family members have met the family deductible, I compare individuals according to an instrumental variable.

1.3 Identification Strategy

To identify the effect of marginal price on an individual's medical care expenditure, I use an instrumental variable – whether or not a family member has an injury. The first stage effect of a family member's injury on the individual's marginal price is possible in families of four or more because of the family deductible and family stoploss described above. When one family member receives treatment for an injury, the family is more likely to meet the family deductible than it otherwise would have been, and any individual in the family is more likely to face a lower marginal price than his own spending would dictate. Empirically, I find that one family member's injury does indeed affect another family member's marginal price.

Given the first stage, the key to the identification strategy is an exclusion restriction: one family member's injury cannot affect another family member's medical spending outside of its effect on his marginal price. Strictly speaking, direct violations of the exclusion restriction are not possible. Since the outcome that I study is the medical spending of an individual in a family, and not the medical spending of the entire family, expenditure for the treatment of one family member's injury is not included in the outcome variable. Furthermore, since one family member's injury does have a direct effect on his own medical expenditure, and the injury itself likely influences his decision to consume follow-up medical care and care for secondary illnesses, I use injured family members only to construct the instrument, and I do not include them in the estimation sample. If two or more family members are injured, all injured family members are excluded from the estimation sample.

Other potential violations of the exclusion restriction involve indirect effects of one family member's injury on another family member's medical spending that occur through a mechanism other than the marginal price. I include only specific injury categories in the determination of the instrument to preclude any mechanisms that involve physical contagion. The complete set of injury categories included in the determination of the instrument are intracranial injuries, superficial injuries, crushing injuries, foreign body injuries, burns, and complications of trauma and injuries to the

nerves and spinal cord. These injury categories should be severe and unexpected enough that treatment for an injury in these categories should not be related to an underlying family-level propensity to seek treatment, which could lead to a violation of the exclusion restriction. An indirect test for violations of the exclusion restriction based on family-level propensities to seek treatment, presented in Section 7.1, lends support to my identification strategy.

To further avoid violations of the exclusion restriction, and also to avoid measurement error, I determine the instrument only on the basis of whether an individual was treated for an injury, and not on the basis of the spending associated with the treatment. If the instrument included a measure of injury spending, the instrument could be related to another family member's medical spending through a family-level propensity to go to expensive doctors, thus violating the exclusion restriction. Since my instrument is only based on the treatment margin, a family-level propensity to go to expensive doctors will not violate the exclusion restriction. However, such a propensity could raise concerns if the price elasticity of expenditure on medical care is not homogenous in the population.

In any instrumental variables setting, if the treatment effect of interest is not homogenous in the population, the estimated effect is a "local average treatment effect," which is intuitively the average effect on "compliers" who would not have received the treatment absent the intervention of the instrument. In this setting, compliers are people who have a family injury which causes them to face a lower price than they would have absent the injury. Although it is not possible to identify compliers because doing so would involve the observation of a counterfactual state in which a family member did not get injured, Angrist, Imbens, and Rubin (1996) propose a formal methodology to examine the average characteristics of compliers in a setting with a binary treatment and a binary instrument. The multivalued treatment in my application precludes the use of the Angrist et al. (1996) methodology, but I can still informally describe the compliers as the population for which the first stage is likely to be the strongest. For example, the first stage will likely be strongest among people who go to more expensive doctors, because the higher the expense, the

higher the likelihood of meeting the family deductible. In addition, the first stage will likely be strongest among accident-prone families, because having an injury in the family is a necessary prerequisite to being a complier. Lastly, the first stage is likely to be strongest among large families because large families have more people to contribute to the fixed family deductible.

1.4 Data

1.4.1 Data Description

I use recent proprietary data from a US firm with over 500,000 insured employees. The data for my analysis are merged together from several databases compiled and distributed by Medstat. In my merged dataset, in addition to observing inpatient, outpatient, and prescription drug claims, I also observe characteristics of the offered plans and associated enrollment characteristics. The Medstat claims data are particularly well-suited to my analysis because the medical claims data identify the beneficiary and insurer contributions on each claim. Because beneficiaries must submit claims to receive reimbursement, and because the firms that pay the claims collect the data, incentives are aligned to ensure the accuracy and completeness of the claims data.

A major advantage of the Medstat data over standalone claims data is that if beneficiaries do not file any claims or discontinue enrollment, I can still verify their coverage and observe their demographic characteristics in the enrollment database. Since the data are longitudinal, I can track individuals and their covered family members over time as long as the subscriber remains at the same firm. One limitation of the Medstat data is that I do not observe employees or family members who are not covered, and I do not observe health insurance options available outside the firm. However, according to the 2006 Kaiser Annual Survey of Employer Health Benefits, 82% of eligible workers enroll in plans offered by their employers, so I should observe a large majority of workers at the firm that I study.

I focus on data from one firm to isolate marginal price variation from other factors that could vary by firm and plan. The main advantage of the firm I that I study is that the four plans that it offered in 2003 and 2004 varied only in the deductible and stoploss. I rely on within-plan price variation for identification, but plan-related local average treatment effects are possible, and I investigate them in Section 6.2. Table 1.1 presents a comparison of the cost sharing parameters across plans. The individual deductibles vary from $350 to $1,000, and the family deductible is always three times the individual deductible, as in the example described above. Net of deductibles, the family stoplosses are always twice as large as the individual stoplosses.

Overall, the simple cost sharing parameters introduced above provide a very accurate description of the marginal prices that consumers face at this firm. Almost all covered medical spending counts toward the deductible and stoploss, except for spending on prescription drugs, which I analyze separately because it is covered separately. Unlike in many medical plans, there is no fixed per-visit payment.

The only complication in the cost sharing structure at the firm that I study is that the plans offer incentives for beneficiaries to go to providers that are part of a network. All four plans are a common type of health insurance plans called preferred provider organization (PPO) plans. According to the Kaiser 2006 Annual Survey of Employer Health Benefits, 60% of workers with employer-sponsored health insurance are covered by PPO plans. PPO plans do not require a primary care physician or a referral for services, and there are no capitated physician reimbursements. However, there is an incentive to visit providers in the network because there is a higher coinsurance rate for expenses outside of the network. In the firm that I study, the general coinsurance rate is 20%, and the out-of-network coinsurance rate is 40%. The network itself does not vary across plans. In the data, there are no identifiers for out-of-network expenses, but, as demonstrated by Figure 1-2, which plots beneficiary expenses on total expenses, beneficiary expenses follow the in-network schedule with a high degree of accuracy, indicating that out-of-network expenses are very rare. Accordingly, in my analysis, I assume that everyone who has met the deductible faces

the in-network marginal price for care. My main results do not change when I exclude the small number of beneficiaries whose out-of-pocket payments deviate from the in-network schedule.

1.4.2 Sample Selection

Although selection into the firm that I study could be a cause for concern, the firm has employees in every region of the United States, and it is large enough that idiosyncratic medical usage should not be a problem. With over 800,000 people covered by the plans offered by this firm, this firm is large, even among other large firms in the Medstat data. Furthermore, all of the component Medstat databases are available for this firm for 2003 and 2004, so I can check for internal consistency by comparing results across both years. Beginning in the 2003 data, the Medstat data include fields that make the determination of marginal price and continuous enrollment very accurate. Since these data are so recent, they should provide an accurate description of current health insurance offerings and usage. Because the covered population consists of active, non-union employees in the retail trade industry, my findings should have widespread external validity.

Within the firm, the main selection criterion that I apply is a continuous enrollment restriction. Since my outcome of interest is year-end expenditure, and family members play a role in the determination of the instrument, I only include individuals in my sample if their entire families, with the exception of newborns, are enrolled for the entire plan year. I retain families with newborns on the grounds that child birth is an important medical expense. Care before death is also an important medical expense, but I cannot make an exception for individuals who die because I only observe in-hospital deaths, and there are none recorded in the unselected sample. In my main results, which use the 2004 and 2003 data as separate cross-sections, I only require that the family is enrolled from January 1 to December 31 of the given year. Selection due to the continuous enrollment restriction eliminates over 30% of the original sample in each year. Analysis of other firms in the Medstat data suggests

that the rate of turnover at this firm is comparable to the rate of turnover at other large firms.

Through selection based on the detailed fields in the Medstat data, I can be confident that my selected sample consists of accurate records. Since families are important to my analysis, I perform all selection steps at the family level. I eliminate families that switch plans, families that have changes in observable covariates over the course of the year, and families that have demographic information that is inconsistent between enrollment and claims information. I also eliminate families that have unresolved payment adjustments. Statistics on each step of the sample selection are available on request. Taken together, these steps eliminate less than seven percent of individuals from the continuously enrolled sample.

In this clean sample, just over 25% of employees with other insured family members are insured in families of four or more. In my main specifications, I restrict the estimation sample to people in families of four or more to ensure that intra-family interactions in cost sharing parameters are possible. The 2004 main estimation sample includes 127,119 individuals from 29,010 families of four or more. Although the stoploss induces some intra-family interactions in marginal price in families of three, I restrict the estimation sample to families of four or more so that deductible interactions are also possible. In a robustness check, I examine employee-spouse couples precisely because price interactions are not possible.

To better control for unobservables, in some specifications, I limit my estimation sample to the employee in each family, and I use other family members only in the determination of the instrument. In some specifications, I also include individuals identified as spouses in the estimation sample. Restricting the sample to employees or employees and spouses sacrifices power because it does not take the price responsiveness of all family members into account, but it arguably provides the best control for unobservables on the grounds that employees at the same firm have some common characteristics that they do not necessarily share with the spouses and children of their co-workers. Moreover, restricting the sample to employees eliminates the need to address possible correlations in price responsiveness among family members.

1.4.3 Summary Statistics

In the 2004 sample, mean year-end medical expenditure by the beneficiary and the insurer is $1,484.74 in the sample of employees and $1,134.83 in the sample that also includes spouses and dependents. However, the mean is not a very informative summary statistic for medical expenditures because many people consume zero care, and the distribution of medical spending among those who do consume care traditionally has a long right tail. In my full sample, almost 40% of people consume zero care in the entire year, and people in the top 25% of the expenditure distribution are responsible for 94.5% of expenditures. Given this skewness, I analyze the logarithm of expenditure instead of the level. The first panel of Table 1.2 summarizes the expenditure distribution across bins that follow a logarithmic scale. Excluding individuals with zero expenditure, the distribution of positive expenditure follows an approximately lognormal distribution, with 31.1% of individuals in the expenditure range between $100 and $1,000, and smaller percentages of individuals in the bins above and below this range. The distribution of expenditures in the full sample, summarized in the second column, is similar. Table 1.3 presents analogous summary statistics for the 2003 samples.

In Table 1.4, I compare the expenditure distribution in my sample to the expenditure distribution from a nationally representative sample, the 2004 Medical Expenditure Panel Survey (MEPS). I restrict the MEPS sample to include only non-elderly individuals, and I exclude expenditures on prescription drugs. As shown in Table 1.4, on a percentage basis, the skewness in my sample is relatively comparable to the skewness in the MEPS, but my sample has a slightly more concentrated right tail. On a levels basis, spending in the MEPS is higher than spending in my sample. For example, the .95 quantile of the expenditure distribution in my sample is $5,457 as compared to $8,282 in the MEPS. Part of this discrepancy seems linked to the lower tail of the distribution. In the MEPS sample, only 19.40% of the sample consumes zero care, roughly half of the percentage that consumes zero care in my sample. Although it is possible that my sample selection criteria are more likely to eliminate

individuals with nonzero expenditure, I do not eliminate enough individuals to explain the difference between my sample and the MEPS. It is perhaps more plausible that my sample is representative of people with employer-sponsored coverage, but people with other types of coverage in the MEPS consume more care. Expenditure within employer-sponsored plans is interesting in its own right because roughly three out of five nonelderly Americans have access to them (Kaiser 2006). It is also plausible that the actual claims data are more accurate than the MEPS survey data on the grounds that people are more likely to respond to a survey on medical expenditure if they have nonzero expenditure.

The second panel in Table 1.2 depicts the distribution of the endogenous variable, the marginal price for the next dollar of care at the end of the year. I calculate the marginal price to reflect the spending of the individual and his family members. If the individual has not consumed any care and the family deductible has not been met, the marginal price takes on a value of one because the individual still needs to meet the deductible. In the employee sample, 57.3% of beneficiaries face a marginal price of one, 38.3% of employees face the coinsurance rate of .2, and 3.9% of employees have met the stoploss and face a marginal price of zero. This price variation should be large enough to be meaningful.

The distribution of the instrument, "family injury," shows that 13.4% of employees have at least one family member who is injured in the course of the year. Since injured employees are excluded from the sample, all of the injuries included in the determination of the instrument in the employee sample are to spouses and other dependents. In the full sample, injuries to employees are included in the determination of the instrument, and the same injury can be reflected as a "family injury" for more than one person. Overall, 12.6% of individuals in the full sample have an injury in the family.

Even though injured people are excluded from all estimation samples, I report statistics on the injured people in Table 1.5. If a person has any claim for an injury with an ICD-9 code in one of the listed categories, he is included in the count in the first column. Complications of trauma and injuries to the nerves and spinal

cord are the most prominent. The distribution of injuries across 2003 and 2004 is remarkably stable, which could indicate that the firm is large enough that injuries are not idiosyncratic. In the second column, I report the mean year-end total expenditures for the injured people to demonstrate that their spending should be large enough to have a meaningful effect on the price that their family members face. The last three columns of Table 1.5 show the number of affected family members in each estimation sample by injury category.

The remaining panels of Table 1.2 summarize the distribution of covariates. Family size varies from four to eleven, with 60.2% of people in families of four. The composition of the full sample reflects the composition of a family of four, with almost as many spouses as employees and twice as many dependents. The full sample is gender balanced, but 57.4 percent of employees are male. All employees are between the ages of 20 and 65 in 2004. The distribution of "year of birth" in the full sample shows a bimodal age distribution – the 7% of people in the sample aged between 21 and 30 are in a valley in the age distribution between the parents and the children.

Given the variation in the covariates, I control as flexibly as possible for family size and family composition. If some factor related to family structure causes one family member to get injured and another member to spend more on medical care, there could be a violation of the exclusion restriction. Conditional on flexible controls for family structure, the exclusion restriction should be valid.

The panel that depicts the distribution of "employee class," shows that 70.1% of the employees are salaried, and the remaining employees are hourly. One of the limitations of the Medstat data is that it does not include any income measures, but subscriber class could serve as a crude proxy. At this firm, hourly and salaried employees have health insurance, but their medical expenditure patterns could differ.

The distribution of the sample by Census region in the penultimate panel demonstrates that the firm has a very national reach. The largest concentration of employees is in the West South Central Census region, where 28.3% of the sample resides. There are also high concentrations of employees in other central regions.

The final panel depicts the distribution of employees and families across the

four plans. Each plan has a unique individual deductible, which I use as the plan identifier. A comparison of the plan distribution between the employee sample and the full sample shows that larger families do not select differentially into plans. Almost 60% of employees and families are enrolled in the most generous plan, which has a $350 deductible. Since this plan is the most popular, and since the low deductible makes the people in this plan the most likely to experience a price change for a fixed amount of spending, it is likely that the behavior of the people in this plan has a substantial influence on my results.

1.5 Results

1.5.1 Graphical Results

The raw variation in the data that drives my instrumental variables approach is so pronounced that it can be discerned graphically, without the assistance of complex estimators. In instrumental variables parlance, the effect of family injury on expenditure is the "reduced form," and the effect of family injury on the year-end price is the "first stage." The simple instrumental variables estimate is the ratio of the reduced form to the first stage. To show the variation that drives the instrumental variables strategy, I present graphical depictions of the reduced form and the first stage in the 2004 sample of employees.

To demonstrate the reduced form, in the top panel of Figure 1-3, I present the cumulative distribution (cdf) of expenditure conditional on family injury. The cdf of expenditure for employees with no family injury is represented by a solid line, and the cdf of expenditure for employees with a family injury is represented by a dashed line. In this depiction, each quantile on the y axis is associated with a value of the logarithm of expenditure on the x axis. Since the lines never cross, it is clear from the figure that employees with family injuries have higher expenditures at all quantiles. This should reassure us that regression results will not be driven by a few large spenders with family injuries. The y intercepts of each line indicate that

family injuries affect the extensive margin decision of whether or not to consume any care; only 30% of people with family injuries consume zero care, as opposed to 37% of people with no family injuries. The figure also suggests that family injuries affect the intensive margin decision of how much care to consume conditional on consuming any care; median expenditure is $120 among employees with no family injuries and $203 among employees with family injuries. To examine whether the difference between the lines at all quantiles is driven by effects on the extensive margin, I create a similar figure, not shown here, that depicts cumulative distributions conditional on positive expenditure. The lines of the new figure do not cross, indicating that even among employees with positive expenditure, employees with family injuries have higher expenditure at each quantile. Columns 3-4 and 5-6 in the first panel of Table 1.2 demonstrate the same finding with conditional probability density functions.

To demonstrate the first stage effect of family injury on the year-end price, in the bottom panel of Figure 1-3, I present the cumulative distribution of year-end price conditional on family injury. Since the year-end price takes on only three values, the cdf is a step function, but I connect the points of the step function with straight lines to aid in the visual interpretation. The lines in this figure do not cross, indicating that employees with family injuries are more likely to face lower prices than their counterparts without family injuries. Labels on the y axis show that 56% of employees with family injuries spend more than the deductible, while only 41% of employees without family injuries spend more than the deductible. Similarly, 6.8% of employees with family injuries spend more than the stoploss, while only 3.5% of employees without family injuries spend more than the stoploss.

The depiction in the bottom panel also allows us to assess which price change, the change from 1 to .2 or the change from .2 to 0, yields the most identification. Following Angrist and Imbens (1995), the vertical difference between the cdf's at the new price is proportional to the weight in an instrumental variables estimate formed from a weighted combination of separate Wald estimates for each price change. Since the difference in the cdf's is largest at the price of .2, the figure indicates that most identification comes from the price change between 1 and .2, and some identification

comes from the price change between .2 and 0.

As a more formal alternative to the bottom panel of Figure 1-3, a simple ordinary least squares (OLS) regression of year-end price on family injury and a set of covariates discussed below indicates that having an injury in the family decreases the year-end price by 11 percentage points, with a standard error of .7 percentage points. The R-squared of this first stage regression with the covariates partialled out is .0096, implying a concentration parameter (defined as $NR^2/(1-R^2)$) of 281. Based on this evidence, "weak instruments bias" is unlikely to be a problem in this application.

If the instrumental variables strategy mimics a randomized experiment, the inclusion of control variables should not have a substantial impact on the estimate, but should merely make it more precise. One way to assess the importance of control variables to the instrumental variable strategy is to examine the distribution of each variable conditional on the values of the instrument. Ideally, in this setting, individuals who have any injured family member would be similar in all observable ways to those who do not have an injured family member.

Columns 3-4 and 5-6 of Table 1.2, starting with the panel on family size, give the distribution of covariates conditional on family injury. The distribution of family size shows that individuals in larger families are slightly more likely to have injuries in their families, as is to be expected if the incidence of injures is distributed evenly across individuals. Given this discrepancy, I include flexible controls for family structure in my formal estimates. Specifically, I include a dummy for the presence of a spouse on the policy, the year of birth of the oldest and youngest dependent, and the count of family members born in each of the year ranges in the table, with the 1999-2004 range saturated by year. In the remaining panels of Table 1.2, the distribution of the other control variables appears much less sensitive to the instrument. However, complex interactions between these variables would not be visible in the table. To test for complex interactions between variables, I regress family injury on the flexible controls for family structure and saturated controls for all of the other covariate rows in Table 1.2. The F test of the null hypothesis that the coefficients on all variables

are zero is rejected with a test statistic of 13.35 and degrees of freedom (32, 28997), suggesting the need to control for covariates. In my formal estimates, I include the controls from the aforementioned regression specification to account for complex interactions and to improve the precision of the estimates.

1.5.2 Results using Traditional Estimators

The large frequency of zeros in the distribution of year-end expenditure complicates the formalization of the above graphical results. Even though the zeros arise from a decision to consume zero medical care, econometrically, they can be treated in the same manner as zeros that arise from a traditional censoring mechanism, such as a report of a zero when desired medical expenditure is negative. It is intuitive to understand the zeros in the context of a traditional censoring mechanism when analyzing the logarithm of expenditure, which is advisable here given the skewness in the expenditure distribution. It is always possible to represent censoring as a monotonic transformation of a variable, and here, the monotonic transformation of expenditure takes the following form:

$$\ln E_i = \max((\ln E_i)^*, \varepsilon) = T((\ln E_i)^*) \tag{1.1}$$

where $T(x) \equiv \max(x, \varepsilon)$, E represents total year-end medical expenditures by the beneficiary and the insurer, and $(\ln E_i)^*$ is the hypothetical uncensored value of $\ln E_i$. When estimators account for censoring at ε, the specific value of ε is immaterial as long as ε is more extreme than all observed uncensored values, so that no information is lost. In practice, I set ε equal to -.7 in my data, so that all observations with zero expenditure have a value that is smaller than the logarithm of 50 cents, the smallest observed nonzero expenditure.

If estimators that do not allow for censoring are used on data censored at ε, bias can arise. Intuitively, when ε is observed in the place of a value that should be much smaller, a line that fits the observed values will be biased toward zero. Formally, in OLS and traditional instrumental variables models, censoring in the

dependent variable induces a correlation between the error term and the independent variable that leads to bias.

In the econometrics literature, there are several techniques to deal with censoring in the dependent variable. Three of the most popular censored models are the truncated model, the two-part model, and the Tobit model, and I present estimates using each model in turn. I also present estimates that incorporate instrumental variables into these models. Estimation of these models with my data and instrumental variables strategy allows me to compare my results to those in the existing literature on the price elasticity of expenditure on medical care.

Truncated Model

The truncated model deals with censored observations in the simplest way, by dropping them. Because of data limitations, Eichner's sample was truncated in the sense that he did not observe a large fraction of the sample that consumed zero care. To facilitate a simple comparison of my results to Eichner's results, I estimate maximum likelihood regressions that assume a truncated normal distribution for the error term. In the first column of Table 1.6, I present results from the truncated regression of the logarithm of expenditure on year-end price P and a set of controls X as given by the following specification:

$$\ln E = \alpha P + X'\beta + u, \qquad E > 0, \quad u \backsim TN(0, \sigma^2). \tag{1.2}$$

The coefficient on year-end price and the associated lower and upper bounds of the 95% confidence interval are presented in the table. The coefficient would be in the form of an elasticity if I estimated a specification using the logarithm of year-end price, but I cannot do so because year-end price can take on values of zero. Instead, I transform the coefficient using the following arc elasticity formula:

$$\eta_{arc} = \frac{\ln(\frac{y_a}{y_b})}{\ln(\frac{a}{b})}. \tag{1.3}$$

I use an arc elasticity instead of a point elasticity because, as discussed above, identification comes mainly from the large price drop from 1 to .2. Specifically, as a function of the estimated coefficient $\widehat{\alpha}$, and the prices of interest, the transformation that I use is as follows:

$$\widehat{\eta} = \frac{(\ln \widehat{E|P = .2}) - (\ln \widehat{E|P = 1})}{\ln(\frac{.2}{1})} = \frac{\widehat{\alpha}(.2 - 1)}{\ln(\frac{.2}{1})} \approx .50\widehat{\alpha}. \tag{1.4}$$

This formula yields the "price elasticity of expenditure." For a homogenous good with a linear price, the "price elasticity of expenditure" η, is related to the "price elasticity of demand" η_{demand}, by the following equation:

$$\eta = \frac{\partial \ln E}{\partial \ln P} = \frac{\partial \ln(Q \cdot P)}{\partial \ln P} = \frac{\partial \ln Q + \ln P}{\partial \ln P} = \frac{\partial \ln Q}{\partial \ln P} + 1 = \eta_{demand} + 1 \tag{1.5}$$

where Q measures units of medical care. By subtracting one from the expenditure elasticity, I could arrive at the price elasticity of demand for medical care under the simplifying assumption that medical care is a homogenous good with a linear price that is borne entirely by the consumer. However, since the literature generally reports expenditure elasticities, I report expenditure elasticities in brackets under each coefficient.

The estimated expenditure elasticity from the truncated regression is -1.4, indicating that a one percent increase in price is associated with a 1.4% decrease in expenditure. Truncation should bias this estimate toward zero, but endogeneity should bias this estimate away from zero. In the second column, I present results from the truncated IV regression based on the following model:

$$lnE = \alpha P + X'\beta + u, \qquad E > 0, \quad u \backsim TN(0, \sigma^2) \qquad \text{(structural)} \tag{1.6}$$

$$P = Z'\gamma + X'\psi + v, \qquad E > 0 \qquad \text{(first stage)} \tag{1.7}$$

where all variables are defined as above, Z is a dummy variable that indicates

family injury, and α is the coefficient of interest. In the truncated instrumental variables regression, the elasticity computed according to the arc elasticity formula is -.8. Given that endogeneity should bias the estimate OLS away from zero, it is intuitive that this estimate is smaller than the OLS estimate. Even with deliberate truncation, both of these estimates are substantially larger than Eichner's estimate of -.33, suggesting that estimates obtained with other estimators on the full set of data are likely to be large.

Two-part Model

Many of the results from the RAND experiment come from estimators that model censored outcomes with a two-part model (2PM), which is very common in the health literature (see Duan et al. (1983)). The advantage of the two-part model is that it allows the decision to consume any care to be determined in a process separate from the decision of how much care to consume. The probability density function for year-end expenditure, E, in the two-part model is as follows:

$$f(E|X) = \left\{[\Pr(d=0|X)^{1-d} \Pr(d=1|X)^{d}] f(E|d=1,X)^{d}\right\} \tag{1.8}$$

where I define $d = 1$ if expenditure is positive and $d = 0$ if expenditure is zero, and other notation is the same as above. Because the probability density function, and hence the likelihood function, is separable between the term in brackets and the last term, this model can be estimated in two parts. As emphasized by Duan et al. (1983), this separability does not depend on an independence assumption between the two parts. The first part is generally estimated by a probit model:

$$\Pr(d=1|X) = \Phi(X_1'\beta_1) \tag{1.9}$$

where $\Phi(\cdot)$ is the standard normal cumulative density function. The second part is generally estimated by a truncated log-normal OLS regression as in (1.2), which imposes the following distributional assumption:

$$\ln E|d=1,X \quad \backsim \quad N(X_2'\beta_2, \sigma_2^2). \tag{1.10}$$

Although the logarithmic transformation deals with skewness as in the truncated model, the primary purpose of the logarithmic transformation in the two-part model is actually to allow for a tractable linear conditional mean function. Although expenditure can take on only positive values, the logarithm of expenditure can take on both positive and negative values, so the assumption of a linear conditional mean is more justifiable. Expected medical expenditure is given by the expression:

$$\mathrm{E}[E|X] = \Phi(X_1'\beta_1)\exp[X_2'\beta_2 + \sigma^2/2]. \tag{1.11}$$

The two-part nature of this model makes the calculation of marginal effects and elasticities complex. See Mullahy (1998) for a discussion. In particular, if there is heteroskedasticity, so that σ_2 is a function of X_2, even though β_1 and β_2 can be estimated consistently, the marginal effects can be inaccurate. Several techniques have been developed to deal with potential heteroskedasticity in these models, but I will follow Duan et al (1983) here in making the simplifying assumption of homoskedasticity for computational convenience. I calculate the marginal effect of year-end price P (included in the vectors $X_1 = X_2$) as follows:

$$\frac{\partial \ln \mathrm{E}[E|X]}{\partial P} = \frac{\ln(\Phi(X_1'\beta_1)) + X_2'\beta_2 + \sigma_2^2/2}{\partial P} = \beta_{1P}\frac{\phi(X_1'\beta_1)}{\Phi(X_1'\beta_1)} + \beta_{2P} \tag{1.12}$$

where β_{1P} and β_{2P} are the coefficients on P. To estimate this expression, I estimate β_2 with the truncated OLS and IV models, I estimate β_1 with the probit model, and I estimate the inverse Mills ratio, $\phi(X_1'\beta_1)/\Phi(X_1'\beta_1)$, by predicting it for every observation in the data and taking the average. I transform this marginal effect into an elasticity using the arc elasticity formula as in (1.4). In the fourth column labeled "OLS 2PM," I present the marginal effects and the elasticity for the Two-Part model including a probit first step and an OLS second step. As expected because

this estimate accounts for the zeros, the implied elasticity of -1.9 is larger than the elasticity implied by the truncated OLS regression.

Because of random assignment to plans in the RAND experiment, it was not necessary to account for endogeneity in the RAND two-part model. However, in the fifth column, I attempt to account for endogeneity here by replacing the truncated OLS second part of the model with a truncated IV second part of the model. I do not run probit IV in the first step because the probit IV coefficients may not be identified given the first stage homoskedasticity assumption and the discreteness of the endogenous variable. (See Chesher (2005).) With this attempt to account for endogeneity, the estimated elasticity is -1.6, which is slightly smaller than the two-part model estimate that does not account for endogeneity, but it is still much larger than estimates from the RAND experiment.

Tobit Model

Unlike the two-part model, the Tobit model, developed by Tobin (1958), models the intensive and extensive margins simultaneously. In addition, the Tobit model can be readily extended to deal with endogeneity with a two-step estimator developed by Newey (1987). Tobit estimates are obtained by maximizing a censored log likelihood function that is derived by assuming that the error term is normally distributed. In the presence of non-normality or heteroskedasticity, the estimates are inconsistent. The instrumental variable version of Tobit is particularly restrictive because it imposes additional distributional assumptions: a homoskedasticity assumption on the first stage error term and a joint normality distributional assumption on the structural and first stage error terms. Furthermore, it is unlikely that the Tobit IV assumption of homoskedasticity in the structural equation holds given the discreteness of year-end price, the endogenous variable. However, I present results from the Tobit model here because it is arguably the most popular way to deal with censoring, and because Eichner's results are based on a variation of the Tobit model.

In column 6 of Table 1.6, I present results from a Tobit regression. The

estimated elasticity of -4.1 is quite large, but we might expect it to be biased away from zero because of endogeneity. Column 7 presents results obtained with Tobit IV. After dealing with endogeneity, the estimated elasticity shrinks to -3.2.

The elasticity presented in column 6 is based on the marginal effect of year-end price on the conditional mean of the latent variable, which can be thought of as desired medical spending. However, the marginal effect of year end-price on *actual* medical spending may also be of interest. It can be calculated according to the following equation based on the censored conditional mean:

$$\frac{\partial \mathrm{E}[\ln E|X, E>0]}{\partial P} = \beta_P \Phi(X'\beta/\sigma)$$

where all variables are defined as in the two-part model. I estimate this marginal effect by predicting the scaling factor $\Phi(X'\widehat{\beta}/\widehat{\sigma})$ for every observation. I then multiply the sample mean of this expression by $\widehat{\beta_P}$, and I calculate the arc elasticity according to (1.4). As shown in column 8, labeled "Tobit IV mfx," the estimated elasticity of actual spending is still quite large at -1.9.

1.5.3 Introduction to CQIV

The advantage of CQIV relative to the traditional estimators discussed above is that it does not require distributional assumptions to handle censoring, and it allows for estimation at several conditional quantiles of the dependent variable, instead of just at the conditional mean. In this application, given that the distribution of medical expenditure is skewed, effects on the upper end of the expenditure distribution are generally of interest in discussions of cost containment. CQIV allows me to estimate the price sensitivity of the people with the highest expenditure. It also allows me to examine how price sensitivity varies across the expenditure distribution. Since CQIV is a quantile estimator, it is robust to extreme values.

Even absent outliers, which are likely in this application given the skewness in medical expenditures, quantile estimators and mean estimators are not likely to yield the same point estimates because they do not estimate the same quantities.

Quantile estimates and mean estimates will be similar if the underlying treatment effect is linear and the error distribution is symmetric and homoskedastic. To the extent that quantile and mean estimators are linear approximations to underlying nonlinear functions, it is likely that they will yield different estimates.

CQIV estimates are also likely to be different from estimates obtained with traditional estimators because CQIV is a semiparametric estimator, and the traditional estimators discussed above depend on distributional assumptions on the error term. If the distributional assumptions are correct, the other estimators are more efficient. If the distributional assumptions are incorrect, CQIV is more robust.

The functional form of the CQIV model that I estimate is very flexible, in that it allows for random coefficients on year-end price and the control variables that vary with the quantiles of the expenditure distribution. Specifically, I estimate the following model:

$$\begin{aligned}
\ln E &= T(\alpha(U)P + X'\beta(U)) \equiv T(\widetilde{X}_i'\theta(U)) && \text{(1.13a)}\\
P &= \phi(X, Z, V) && \text{(1.13b)}\\
& V \text{ statistically dependent on } U && \text{(1.13c)}\\
& U|X, Z \backsim \text{Uniform(0,1)} && \text{(1.13d)}\\
\tau &\mapsto \alpha(\tau)P + X'\beta(\tau) \text{ strictly increasing in } \tau && \text{(1.13e)}
\end{aligned}$$

where the variables are as defined above, the censoring function $T(x)$ is defined as above, and $\alpha(\tau)$, the coefficient of interest, varies with the quantile, τ. The expression given by (1.13d) is completely general because the *quantiles* of any distribution always follow a uniform distribution. The expression given by (1.13e) is a standard "rank invariance" condition. A sufficient condition for rank invariance is that the ordering of observations does not change with treatment: people who begin at the median remain at the median when the entire distribution experiences a price change. A necessary condition for rank invariance is that deviations from the original ordering are not systematic after the price change. To understand the theoretical and computational

underpinnings of how the CQIV estimator applies to this model, it is intuitive to consider censoring and endogeneity separately.

The CQIV estimator handles the theoretical issues associated with censoring in the spirit of Powell's censored quantile regression (CQR) model (Powell (1986)). Censoring induces attenuation bias in traditional quantile regression much in the same way it induces bias in mean regression. Quantile regression is based on the assumption that the conditional quantiles of lnE depend linearly on P. Since quantile regression uses information from the entire sample, if some observations on lnE are censored, the quantile regression lines can be biased toward zero at *all* quantiles. However, because quantiles are invariant to monotonic transformations, and censoring is a monotonic transformation, quantile estimators can be extended to handle censoring without any distributional assumptions. Intuitively, Powell's estimator eliminates attenuation bias by using the entire sample to determine which observations are least likely to be censored and estimating the coefficients based on those observations. Technically, Powell's model incorporates a censoring mechanism directly into the estimator. Powell's censored regression model, as applied to Equation (1.13a), abstracting away from endogeneity, is as follows:

$$\widehat{\theta}(\tau) \text{ minimizes } \sum_{i=1}^{n} \rho_\tau(\ln E_i - T(\widetilde{X}_i'\theta)). \tag{1.14}$$

where $\rho_\tau(u) = \{(1-\tau)1(u<0) + \tau 1(u>0)\}|u|$. Despite its intuitive appeal, this model is rarely used because the function $T(x)$ induces nonconvexities in the objective function that present computational difficulties.

Chernozhukov and Hong (2002) devised a tractable computational algorithm for Powell's model based on the idea that Powell's censored regression model estimates the coefficients using observations that are not likely to be censored. Accordingly, the algorithm is a three-step procedure that predicts which observations are least likely to be censored and estimates the coefficients based on those observations. The first step involves a parametric prediction of the probability of censoring based on a probit or logit model. A set fraction of observations that are unlikely to be censored

are retained for estimation via quantile regression in the second step. After the second step, a larger set of observations is retained based on the predicted values of the dependent variable. This sample gets asymptotically close to the ideal sample of non-censored observations, and consistent estimates are obtained through a third step of quantile regression on this sample. In the computation of the CQIV estimator, I use an analog of the Chernozhukov and Hong (2002) algorithm to handle censoring.

In its theoretical handling of endogeneity, the CQIV estimator is based on the instrumental variable quantile regression estimator of Chernozhukov and Hansen (2008). Following Chernozhukov and Hansen (2008), the "structural quantile function"

$$S_{\ln E}(\tau|p,x) = T(\alpha(\tau)p + x'\beta(\tau)) \tag{1.15}$$

describes the quantile function of the latent outcome variable $\ln E_p$, which could be observed for each potential price p and potential set of covariates x if these variables could be manipulated as in a randomized experiment. Given the conditional uniform distribution of U, and the rank invariance condition, the event $\{\ln E \leq S_{\ln E}(\tau|P,X)\}$ is equivalent to the event $\{U \leq \tau\}$, yielding the following conditional moment restriction:

$$\Pr[\ln E \leq S_{\ln E}(\tau|P,X)|Z,X] = \tau. \tag{1.16}$$

Intuitively, this conditional moment restriction requires that the instrument is independent of the residual at each quantile τ, which can be interpreted as a generalization of a standard instrumental variables assumption to a quantile framework. Based on this conditional moment restriction, the true coefficients will satisfy the structural quantile function which satisfies

$$\arg\min_{f \in F} E\rho_\tau[(\ln E - S_{\ln E}(\tau|P,X) - f(Z,X)] \tag{1.17}$$

where F is the class of measurable functions of (Z,X). CQIV estimates are based on a finite sample analog of (1.17). In Chernozhukov and Kowalski (2008),

Victor Chernozhukov and I propose an algorithm to compute estimates according to a finite sample analog of the above equation using a control function approach. In the current paper, which was completed prior to Chernozhukov and Kowalski (2008), I use a slightly different algorithm, which is a more direct combination of the Chernozhukov and Hong (2002) and Chernozhukov and Hansen (2008) algorithms. The main CQIV results of this paper are the same across both algorithms up to the reported number of significant digits. Formally, in this paper, I estimate the CQIV coefficients for each quantile τ as follows:

$$(\widehat{\gamma}(\alpha,\tau),\widehat{\beta}(\alpha,\tau)) = \arg\min_{\beta,\gamma} Q_n(\tau,\alpha,\beta,\gamma) \tag{1.18}$$

$$Q_n(\tau,\alpha,\beta,\gamma) = \frac{1}{n}\sum_{i=1}^{n}\rho_\tau(\ln E_i - T(\alpha P_i + X_i'\beta) - Z_i'\gamma). \tag{1.19}$$

To execute the algorithm, for each τ, I run the $\tau-$censored quantile regression of $\ln E_i - \alpha P_i$ on X and Z, over a grid of $\alpha \in A$, where A is chosen to be large enough to incorporate all plausible values of the coefficient of interest. I choose $\widehat{\alpha}(\tau)$ to be the value for which the coefficient on the instrument, $\widehat{\gamma}(\tau)$, divided by its variance, is closest to zero. Formally, $\widehat{\alpha}(\tau)$ is the value in the grid that yields the smallest value of the objective function $W(\alpha)$ as follows:

$$\widehat{\alpha}(\tau) = \arg\inf_{\alpha\in A}[W(\alpha)], \quad W(\alpha) := [\widehat{\gamma}(\alpha,\tau)']\widehat{A}(\alpha)[\widehat{\gamma}(\alpha,\tau)]. \tag{1.20}$$

In practice, I set $A(\alpha)$ equal to the inverse of the asymptotic covariance matrix of $\sqrt{n}(\widehat{\gamma}(\alpha,\tau) - \gamma(\alpha,\tau))$ so that $W(\alpha)$ is the Wald statistic for testing $\gamma(\alpha,\tau) = 0$. All values of α that produce a value of $W(\alpha)$ lower than the .95 critical value of the chi-squared distribution fall within a 95% confidence interval on $\widehat{\alpha}(\tau)$. In practice, the highest and lowest such values of α are reported as upper and lower bounds on the estimated $\widehat{\alpha}(\tau)$. Although this procedure does not yield a standard error and associated symmetric confidence interval, the upper and lower bounds serve the same purpose. Following a proof by Chernozhukov and Hansen (2008) these upper and

lower bounds are robust to weak identification.

I transform the estimated $\widehat{\alpha}(\tau)$ into an expenditure elasticity via a direct analog of (1.4). For point estimates of the coefficients on the other covariates, I report the values of those coefficients at $\widehat{\alpha}(\tau)$. The upper and lower bounds of these coefficients are computed as the highest and lowest values of these coefficients for which $\widehat{\alpha}(\tau)$ is within its 95% confidence interval.

1.5.4 Main Results

Although it is intuitive to compare mean elasticities to median elasticities, it is not possible to obtain reliable CQIV results at the median because of the heavy censoring of expenditure in this application. Intuitively, at conditional quantiles where zero expenditure is likely, the marginal price can have an effect on two margins - the decision to spend anything at all, and the decision to change spending conditional on spending a positive amount. If changes in price and other factors are not sufficient to induce people to visit the doctor at all, it is not possible to estimate the effect of small changes in price. With 40% censoring, it seems reasonable that CQIV coefficients are not reliable at the median. In CQR, results cannot be obtained at the quantile corresponding to the percent of censored observations because there is noise in the prediction of which observations are least likely to be censored. In CQIV, the prediction of which observations are least likely to be censored depends on the instrument, which adds noise to the prediction. The lowest quantile for which results can be obtained in this paper is the .55 quantile, but estimates at the lowest estimable quantiles are not very precise.

To demonstrate the relative precision of the CQIV estimates across quantiles from .55 to .95, Figure 1-4 presents separate graphs of the CQIV objective function, $W(\alpha)$, at each quantile, as estimated in the sample of 2004 employees. The horizontal axis of each graph is the estimation grid over the year-end price coefficient α, which extends from -10 to -2 in increments of .1. The vertical axis of each graph is the value of $W(\alpha)$ evaluated at each α. The horizontal line, drawn at the same value in

the graph for each quantile, is the critical value for a 95% confidence interval on α. In each graph, all values of α for which the value of $W(\alpha)$ is below the horizontal line are within the 95% confidence interval on the coefficient on year-end price at that quantile.

As shown in the three graphs in the top row of Figure 1-4, identification is very limited in the .55 to .65 quantiles, where a wide range of values of α, including the lower limit of the grid, are included in the 95% confidence interval. However, at higher quantiles, the shape of $W(\alpha)$ becomes more convex, and the range of values within the 95% confidence interval narrows. At all quantiles, the value of the objective function is relatively large at the upper limit of the grid. To facilitate comparison across quantiles, a few extreme values of $W(\alpha)$ greater than 30 are omitted from the graphs. Values of the objective function at an α of zero (not shown) are well above those depicted, indicating that zero is well outside the 95% confidence interval at all estimated quantiles. Even though the grid is over the year-end price coefficient α, and not the elasticity, it is worth noting that zero is never in the 95% confidence interval on α, indicating that the transformed elasticity will not be zero.

In each graph, the point estimate on the year-end price coefficient is the value of α at which the objective function $W(\alpha)$ is minimized. The first row of Table 1.7 presents the corresponding point estimates and their associated upper and lower bounds in tabular form. In the table, I only report estimates from the .65 quantile and above. As in the previous table, expenditure elasticities are shown in brackets. In all of the CQIV results in this paper, since the grid on α is in increments of .1, the expenditure elasticity is estimated in increments of .05, and I report its value to one decimal place.

In all of the estimated quantiles, the CQIV expenditure elasticities are an order of magnitude larger than those in the literature. For example, at the .85 quantile of the expenditure distribution, the implied expenditure elasticity is -2.3, which indicates that a one percent increase in price would decrease spending at the .85 quantile of the expenditure distribution by 2.3 percent. This elasticity estimate is fairly stable across the quantiles from .65 to .95, indicating that price responsiveness,

though strong, does not tend to vary among people in the highest quantiles of the expenditure distribution.

The next two sets of rows in Table 1.7 present coefficients estimated on less-restrictive samples that include spouses and other dependents in each family. The patterns in the estimates across the quantiles are very similar to those in the employee sample, but the estimates are slightly more precise given the larger sample sizes. The elasticities estimated on the 2003 sample, presented in the bottom panel of Table 1.7, show remarkably similar patterns. The similarity of the estimates between 2003 and 2004 provides some evidence of robustness, and it suggests that price responsiveness did not change between 2003 and 2004. Even though price responsiveness does not tend to vary across the estimated quantiles, the coefficients on the covariates, not reported here, vary dramatically, indicating that models such as Tobit IV, which impose constant treatment effects for all coefficients, sacrifice flexibility.

The last column of Table 1.7 presents Tobit IV coefficients for comparison to the CQIV coefficients. I present the untransformed Tobit IV coefficients, which give the effect on desired spending, because they should be the most comparable to the CQIV coefficients at high quantiles. Since Tobit IV imposes a constant treatment effect across all quantiles, the single Tobit IV coefficient can be compared directly to the CQIV coefficient at each quantile. In all specifications, the estimated Tobit IV coefficient is more negative than all of the quantile coefficients above the .65 quantile. If the underlying price responsiveness is constant across the expenditure distribution, a large difference between the Tobit IV estimate and the CQIV estimates could indicate that the distributional assumptions made by Tobit IV are restrictive. Formally, a Hausman (1978) test statistic can be constructed to compare the Tobit IV estimate to the CQIV estimates at each quantile. This is a joint test of the Tobit IV normality and homoskedasticity assumptions. If these conditions hold, Tobit IV should be consistent and efficient, and CQIV should be consistent. It is possible to construct the test statistic through a bootstrapping procedure, but the discrepancy between the estimates is so large that the test null hypothesis is rejected though informal comparison.

As another method of comparison, it is also informative to construct an informal bound on the mean estimate by transforming the distribution CQIV of estimates. Assume, based on the main CQIV estimates, that the expenditure elasticity is constant at -2.3 from the .65 quantile to the top of the expenditure distribution. Since we cannot measure price responsiveness at other quantiles of the distribution, make the conservative assumption that the expenditure elasticity is zero at these quantiles. If the true price responsiveness at these quantiles is zero, a lower bound on the true mean elasticity, assumed constant over all quantiles is $(1 - .65) \times -2.3 = -.805$. This lower bound is very similar to the truncated instrumental variables elasticity, the lowest elasticity attained in the above comparison of traditional estimators.

1.5.5 Closer Examination of Endogeneity

As discussed above, a simple OLS regression in this application should be biased toward zero because of censoring, and it should be biased away from zero because of the mechanical link between expenditure and marginal price. In contrast, the CQIV estimator should not be biased because it accounts for censoring and endogeneity. To get a sense of the magnitude of the endogeneity, it seems instructive to compare the main CQIV coefficients to similar coefficients obtained through CQR, just as one would compare IV coefficients to OLS coefficients; however, censoring makes the comparison less straightforward.

Table 1.8 presents CQIV coefficients and CQR coefficients. The CQR coefficients in the second row are estimated according to the algorithm of Chernozhukov and Hong (2002), and those in the third row are estimated directly using Powell's censored quantile regression objective function. In both sets of censored quantile coefficients, I report an extra significant digit to demonstrate that although the estimates are asymptotically the same, they differ slightly in finite samples. It is easy to report an extra significant digit for the CQR estimates because, unlike the CQIV estimates, they are not estimated on a grid. At the lowest estimated quantiles, both sets of CQR estimates are larger in magnitude than the CQIV estimates, as expected. However, this relationship reverses above the .70 quantile, where the CQIV coefficients

are estimated more precisely.

To understand why the CQIV coefficients can sometimes be larger in magnitude than the CQR coefficients, it is helpful to recall that both estimators use information to select the sample based on the observations that are least likely to be censored. However, the CQIV estimator uses more information than the CQR estimator because it selects the sample based on the instrument as well as the covariates. Thus, in the presence of endogeneity, the CQR estimator still suffers from censoring bias toward zero, even though the CQIV estimator does not.

1.6 Specification Tests

1.6.1 Timing of Family Injury

Regardless of the timing of the family injury, I aggregate medical spending over the entire plan year to ensure that I capture all expenditure responses to injuries. Unlike in other analyses, where the year is a useful construct that allows for a discrete representation of variables that move continuously, the plan year has intrinsic meaning because the cost sharing parameters reset at the end of the plan year. An event-study design based on the timing of family injuries would be difficult to implement here because there is no reason that family members should react immediately as long as they react before the end of the plan year. Similarly, a regression discontinuity design based on the timing of the price change would not have much power. Immediately after a family injury, consumers can see price changes coming even if the actual meeting of the family deductible does not occur until much later, so price responses need not coincide with the timing of the price change. Furthermore, since such a small fraction of the sample receives care on any given date, intertemporal patterns in medical care usage are difficult to detect, and they are easily confounded with seasonal patterns. Even though my estimates will not generally capture shifting of expenditures from month to month, they can capture shifting of expenses from year to year. To the extent that many insurance contracts only cover beneficiaries for a

single year, year-end expenditures are a policy-relevant outcome.

In my main specifications, by examining year-end expenditures, I make the implicit assumption that individuals have enough time to react to all family injuries before the end of the plan year. As the timing of the family injury approaches the end of the year, this assumption becomes less plausible, but I do not automatically omit family injuries that occur near the end of the year for fear of imposing a seasonal bias. Table 1.9 shows the distribution of the first family injury claim by month. I measure the timing of the first claim on the grounds that the first claim is the first possible opportunity for a family response. Measurement of the first claim shifts observed injury incidence toward the start of the year, but a seasonal pattern is still visible in the table.

As a specification test, I drop the 810 employees that have their first family injury in October or later, and I re-estimate my main specification. If these employees do not have sufficient time to react to family injuries, or if their expenditures are driving my main results, the results in this specification test should differ from the main results. The second panel in Table 1.10 shows the results from this family injury timing specification test. At all quantiles, the point estimates are almost exactly the same as the point estimates from the main specification, shown in the first panel. For comparative purposes, the next specification in Table 1.10 shows results obtained by dropping all 3,076 employees with family injuries before October. As expected given that this specification eliminates almost 80% of injuries, the confidence intervals on these estimates are large and often include the upper and lower bounds of the estimation grid. A finding of larger price responsiveness in this specification would be cause for concern, because, if anything, we expect that estimated price responsiveness should be lower if some people with family injuries late in the year will not have adequate time to respond. However, the point estimates suggest the same or slightly less price responsiveness than that in the main specification.

In the next two specifications of Table 1.10, I report results from similar specification tests that include injuries from only the first and second half of the year, respectively. Elasticities based on injuries from the first half of the year are slightly

higher than the main estimates, at approximately -2.5, and elasticities based on injuries from the second half of the year are slightly lower than the main estimates. These specification tests suggest that the effect of the timing of the injury on the main elasticity estimates is small relative to their overall magnitude. Thus, I include injuries that occur throughout the year in my main specification to avoid potential seasonal bias within my year-end expenditure model. An extension of my year-end expenditure model to a full dynamic model could be an interesting area for future research.

1.6.2 Income Effects

In my analysis, there is potential cause for concern if family injuries affect family income and family income affects expenditure. In my analysis, I cannot control for income directly because I do not observe it. As mentioned above, I try to proxy for income with covariates. In addition, I can rely on Newhouse (1977), which shows that income effects on medical expenditure tend to be large across countries and small within countries, to argue that income effects should be small in response to an injury. However, it is perhaps more convincing to test directly for income effects directly in my data.

Since all of the individuals in my data are insured, the income loss from an injury itself is likely to be small, but the income loss associated with an injury can be larger if it prevents the injured party from working. The idea behind my specification test is that if there are large income effects due to the injury of a wage earner, we might expect an employee's response to a spouse's injury to be different than an employee's response to a child's injury. Accordingly, I re-estimate the main specification two times: one time keeping just the employees with child injuries or no family injuries, and another time keeping just the employees with spouse injuries or no family injuries. As shown in the third panel of Table 1.10, the specification with just child injuries gives almost the exact same point estimates as the main specification with spouse AND child injuries, which is not surprising given that 4/5 of the injuries in my sample are to children. The specification with just spouse injuries, which is not as

well identified, also yields point estimates that are the similar or, if anything, smaller in magnitude. All in all, this specification test suggests that income effects are not large relative to my main elasticity estimates.

1.6.3 Plan Variation

As discussed above, employees can select into four different plans, and most employees are included in the $350 deductible plan. The main specification includes a saturated set of plan controls, but it does not allow price responsiveness to vary by plan. If we assume that injury occurrence is independent of factors leading to plan selection, then we can test the viability of restricting price responsiveness to be the same across plans. Though the main specification relies on within-plan identification, this test uses across-plan variation. In plans with lower deductibles, injuries should have larger first stage effects on price than they would in plans with higher deductibles because a given injury has a larger chance of causing the deductible to be met when the deductible is lower. If the first stage does indeed vary by plan as expected, and price responsiveness does not vary by plan, then the reduced form effect of price on expenditure should also vary by plan, leading to instrumental variables estimates that are the same across plans. However, if price responsiveness is not the same across plans, instrumental variables estimates will yield a local average treatment effect that gives the most weight to the price responsiveness in the plan with the largest first stage.

In Table 1.11, I present results from OLS first stage regressions by plan. Although the CQIV algorithm used in the main specification does not explicitly use an OLS first stage, OLS estimates should be informative because the econometric issues that motivate the use of CQIV do not arise in the first stage. As shown in the first column, in the sample that includes all plans, a family injury reduces year-end price by .11, relative to a mean price of .65. As expected, columns 2 through 4 show that the first stage coefficient decreases as the deductible increases. Furthermore, the magnitude of the decrease in the point estimates corresponds roughly to the magnitude of the

decrease in the deductible. Continuing the test, in the fourth panel of Table 1.10, I present results from a full CQIV specification estimated only on the employees in the $350 deductible plan. In separate CQIV regressions for the other plans that are not shown in the table, the point estimates have similar magnitudes, but the 95% confidence intervals are very wide. In all specifications, price responsiveness is broadly the same by plan as it is in the main specification, suggesting that there must be a differential reduced form effect by plan to compensate for the differential first stage. Since the lowest deductible plan is the most popular, and it has the largest first stage, most identification comes from the lowest deductible plan, but this test lends support to the restriction that price responsiveness is the same across plans.

1.6.4 Outpatient Spending vs. Inpatient Spending

Since the potential for cross-substitution is so vast among the medical services covered by the plans that I study, I do not examine expenditure responses by therapeutic category. However, it could make theoretical sense to separate inpatient expenditure from outpatient expenditure because it is conceivable that they are not close substitutes and that price responsiveness varies across these two types of expenditure. The intuition behind estimating separate inpatient and outpatient expenditure specifications comes from a concern that inpatient expenditures could be driving the results in my main specification. These specifications have precedent because the RAND study examined both types of spending separately, and they are feasible because the Medstat data clearly differentiates inpatient spending from outpatient spending.

With few exceptions, individuals with any medical expenditure have outpatient expenditure. Approximately 64% of the sample has some outpatient expenditure, and only 4% of the sample has some inpatient expenditure. On average, individuals with any outpatient expenditure spend $1,585.90, and individuals with any inpatient expenditure spend $9,068.30. If I omit inpatient expenditures from the dependent variable and get similar estimates, I can be more confident that large inpatient expenditures, which could be less elastic *a priori*, are not driving the results.

In the fifth panel of Table 1.10, I present the results from a modification of the

main specification in which the dependent variable includes only outpatient spending. The results in the table suggest that the elasticity of outpatient expenditure with respect to marginal price is approximately -2.0 across the .65 to .95 quantiles, which is slightly smaller than the elasticity of total expenditure. However, the estimates at each quantile of the outpatient specification are not directly comparable to the estimates in the main specification. In the main specification, the quantile assumption is that the log of the conditional quantiles of *inpatient plus outpatient expenditures* are linear in the price, and in the outpatient specification, the quantile assumption is that the log of the conditional quantiles of *outpatient expenditures only* are linear in the price. Given these differing assumptions, the coefficients at specific quantiles cannot be compared without further restrictive assumptions. For example, it is likely that people with inpatient expenditures are likely to be above the .95 quantile of the main specification, but it would be restrictive to assume that they are also above the .95 quantile in the outpatient specification. This example highlights a general phenomenon: although it is natural to use mean estimators to compare models with different dependent variables, similar comparisons are less natural with quantile estimators. Quantile estimates from a regression that includes only inpatient spending in the dependent variable provide an even starker example of the difficulty of comparing quantile models with different dependent variables. Since inpatient expenditures are zero at the .95 quantile, the quantile coefficients are not even identified at or below any of the quantiles reported in the table, making comparison to the main and outpatient specifications very difficult.

However, the general intuition that suggests comparing a specification with inpatient expenditures only to the main specification comes from experience with mean estimators which can be biased because of extreme values. Such tests have less merit in applications that use quantile estimators because quantile estimators are not as sensitive to outliers. Rather, if people with inpatient expenditures are at the highest quantiles of the main specification, and they have different underlying price responsiveness, this will be reflected in coefficients estimated at the highest quantiles. In the last panel of Table 1.10, I present results estimated on the baseline specification

at the .975 through the .995 quantiles in increments of .005. The point estimates suggest that price responsiveness is larger among consumers with the very highest expenditures. However, the confidence intervals are very large and often include the endpoints of the estimation grid. The large confidence intervals are likely a result of poor identification in the region of the very highest expenditures; employees at the highest quantiles are unlikely to change expenditure in response to a family injury. Especially since inpatient expenditures are so infrequent in the sample, and since individuals with inpatient expenditures are generally in the highest quantiles of the total expenditure distribution, it is likely that identification in the main specification comes mostly from outpatient spending. Furthermore, on the whole, the estimated outpatient elasticities are similar to the elasticities in the main specification.

1.7 Robustness Tests

1.7.1 Couples Data

For the instrument to satisfy the exclusion restriction in the main specification, it must be true that one family member's injury does not affect another family member's spending outside of its effect on his marginal price. At the firm that I study, in families of two, there is no mechanical effect of one family member's spending on another family member's marginal price. Therefore, any effects of one family member's injury on another family member's spending presumably operate through another channel. Although the exclusion restriction is not an econometrically testable restriction in the main sample of families of four or more, evidence that there is no effect of one family member's injury on another family member's spending in a family of two supports the validity of the exclusion restriction in the main specification.

To formalize this test, I use the following model, which I estimate with censored quantile regression:

$$\begin{aligned} \ln E &= Z'\delta(U) + X'\beta(U) \\ & \quad U|X,Z \backsim \text{Uniform}(0,1) \\ \tau &\mapsto Z'\delta(\tau) + X'\beta(\tau) \text{ is strictly increasing in } \tau. \end{aligned} \tag{1.21}$$

This specification differs from the main specification only in that, in instrumental variables terminology, it examines the "reduced form" effect of the family injury on lnE directly. A traditional instrumental variables specification would not be informative here because the first stage cannot exist in families of two.

I estimate this specification on the "couples" sample of 2004 employees in employee-spouse families of two. For comparison, I also estimate this specification on the sample of employees in families of four or more. Because price interactions are possible through the stoploss for people in families of three, I do not include people in families of three in this test. To ensure that the results from the family sample are as comparable as possible to the results from the couples sample, I estimate an additional family specification that is only identified off of injures to spouses.

Column 7 of Table 1.2 presents summary statistics on the couples sample. Comparison with Column 1 shows that employees in couples tend to be much older than employees in families of four or more, suggesting that the couples population consists mostly of older "empty nesters" and young couples without children. Furthermore, employees in couples have much higher average expenditures on medical care than their counterparts in families of four or more. Only 24% of employees in couples consume zero care, as opposed to 36% in families of four or more. Given that employees in the couples sample consume more medical care, we should be more likely to observe spurious effects of other family injuries on spending in the couples sample than in the family sample. Since the couples sample is much larger than the family sample, to remove effects of sample size from the comparison, I conduct the estimation in 100 random subsets of the couples sample of the same size as the family sample.

The results in the first row of Table 1.12, which were estimated on one couples sample, show that the effect of the instrument on expenditure in couples is not statistically different from zero. In the 100 random couples samples taken together, the median point estimate at each quantile is generally not statistically different from zero. In contrast, the coefficients in the family specification in the second row suggest that employees with an injured spouse or dependent spend .27 to .45 percent more on their own medical care. In many quantiles, the entire confidence interval for the families exceeds the entire 95% confidence interval for the couples. In the family specification, the 95% confidence interval never includes zero. In contrast, the 95% confidence interval includes zero at almost all quantiles in the couples specification. In the couples point estimates shown, even though the pointwise confidence intervals at the .65 and .75 quantiles do not include zero, a conservative calculation of a uniform confidence interval over all quantiles would include zero, given that the lower bounds at these quantiles are already so close to zero. Tobit coefficients, shown in the last row for comparison, do not include zero in the confidence interval, but they are substantially smaller in the couples specification than they are in the family specification. Overall, this comparison lends strong support to the validity of the exclusion restriction.

One concern with the couple/family comparison is that identification in the couples specification comes from family injures to spouses, and identification in the family specification comes from family injures to spouses as well as other dependents. To the extent that injures to spouses are fundamentally different than injuries to dependents, this comparison becomes less informative. Furthermore, it is possible that injuries to children are less likely to violate the exclusion restriction than injuries to spouses. For example, if an employee is sick, he might spend less time watching his child, and his child may be more likely to get injured. In this scenario, there will be a violation of the exclusion restriction because the employee will have medical expenses for his own illness, and his child will be injured, and these two phenomena are not related through the marginal price of medical care. However, a similar story is less plausible in relation to the injury of a spouse. Thus, a family specification identified

off of injuries to spouses provides a better comparison for the couples specification, and it could be interesting in its own right. A CQIV version of this specification was presented above in the discussion of income effects, but I re-estimate a reduced form version of this specification for comparison to the couples specification here.

The third row of Table 1.12 presents results from a family specification that is identified off of injuries to spouses. In this specification, 760 employees have injured spouses and 25,124 employees have no family injuries. Relative to the main family specification, 3,126 of the 3,886 families with non-spouse injuries are eliminated from the sample. Even though the instrument should have less power in the modified family specification, the confidence intervals do not include zero at any quantile, further reinforcing the validity of the exclusion restriction when compared to the couples specification. Moreover, the point estimates are stable across the two family specifications, suggesting that the identification strategy is robust to the source of the family injuries included in the instrument.

1.7.2 Longitudinal Data

Identification in my main specification relies on price shocks faced by employees when dependents have an injury. Given the price responsiveness suggested by the main specification, if injuries are true shocks to the price, then employees in families with injuries should spend more in the year of the injury than they did in the previous year. To formalize this test, I limit my sample to employees in families of four or more in which every family member is continuously enrolled in 2003 and 2004. This requirement severely limits the sample size. I also exclude employees who have injuries themselves in either year. On the resulting estimation sample of 18,743 individuals, I estimate the following specification with ordinary least squares:

$$E_{2004} - E_{2003} = b_1 Z_{2004only} + b_2 Z_{2003only} + b_3 Z_{2004\&2003} + X'\beta + v \tag{1.22}$$

where the dependent variable, $(E_{2004} - E_{2003})$, is the change in expenditure from

2003 to 2004. The first three independent variables flexibly represent three of the four possible changes in family injury status from 2003 to 2004: $Z_{2004only}$ is a dummy variable that indicates that the individual had a family injury in 2004 but not in 2003, $Z_{2003only}$ is a dummy variable that indicates that an individual had a family injury in 2003 but not in 2004, and $Z_{2004\&2003}$ is a dummy variable that indicates that the individual had a family injury in both years. The vector X, which contains the 2003 values of the standard set of controls, is omitted from initial estimates.

The coefficient b_1 is of interest because it gives the change in expenditure between 2003 and 2004 for individuals that only have a family injury in 2004 relative to individuals who have no family injury in either year. If a family injury represents a true shock to the price, and family members respond in the year of the injury, this coefficient should be positive. Similarly, the coefficient b_2 is of interest because it gives the change in expenditure between 2003 and 2004 for individuals that only have a family injury in 2003 relative to individuals who have no family injuries. This coefficient should be negative.

Conditional on positive b_1 and negative b_2, comparison of the magnitudes of b_1 and b_2 allows for a one-sided test of the null hypothesis that consumers increase contemporaneous expenditures through a mechanism other than inter-year expenditure shifting. Assuming that the contemporaneous expenditure increase in response to a family injury is the same in 2003 and 2004, and consumers achieve this contemporaneous increase by shifting forward 2004 expenditures to 2003, b_2 will be negative and larger in magnitude than b_1, and the null hypothesis will be rejected in favor of inter-year expenditure shifting. If, instead, b_2 is smaller in magnitude than b_1, we cannot reject the null hypothesis because it is possible that employees begin treatment in 2003 in response to an injury but then have some residual expenditure that extends into 2004. The sign of the coefficient on $Z_{2004\&2003}$ is theoretically ambiguous because the first stage price response need not be the same in both years, but it is included as a control variable.

The specification estimated here differs in several ways from other specifications estimated in this paper. Unlike in the previous specifications, the dependent

variable is specified in levels so that the dummy variables allow for a fixed level change in expenditure from 2003 to 2004 instead of a fixed percentage change. Because I am fundamentally interested in the effect of a difference in family injury status on the overall sample, and not on the people who experience the largest changes in expenditures, I do not estimate a quantile model. Furthermore, because the values of the dependent variable can be negative or positive, there is no need for a censored estimator. There is a mass point in the dependent variable at zero because 21.5% of individuals have no change in expenditures from 2003 to 2004, but the overall distribution of the dependent variable is approximately symmetric, so OLS should be an appropriate estimator.

In the sample, only 1.6% of employees have a family injury in 2003 and 2004, suggesting that there is limited persistence in family injury status, so it is plausible that family injuries are indeed shocks to most families that experience them. A further 16.3% of the sample has a family injury in one year but not in the other: 5.5% of employees have a family injury in 2004 only, and 10.8% of employees have a family injury in 2003 only. The remaining employees have no family injury in either year.

The estimates in Table 1.13 support the conclusion that family injuries induce shocks to spending in the year that they are experienced. The first coefficient in the first column, the estimate of b_1, indicates that individuals with a family injury in 2004 and no family injury in 2003 have an expenditure difference between 2004 and 2003 that is \$482.58 larger than the analogous expenditure difference for individuals who never have a family injury. This coefficient is statistically significant at the 5% level, and it remains statistically significant and of a similar magnitude in the specification in the second column, which includes demographic control variables. Although the estimated b_2 is not statistically significant in either column, the sign is negative as expected. The point estimate indicates that individuals with a family injury in 2003 and no family injury in 2004 spend \$82.12 more in 2003 relative to 2004 as compared to individuals who do not have a family injury in either year. Though the standard errors on this point estimate are large, they do not include negative amounts larger in

magnitude than the estimate of b_1, so the null hypothesis that increased expenditure in 2003 is not the result of direct shifting of expenditure from 2004 is maintained.

To further investigate potential inter-year expenditure shifting by decomposing the 2004 effect out of the difference specification, I estimate an analog of (1.13) with 2004 expenditure as the dependent variable. If the negative estimated b_2 comes primarily from shifting of expenses from 2004 to 2003 in response to an injury, the new coefficient on $Z_{2003only}$ should be negative. Alternatively, if the negative estimated b_2 comes from follow up to care initiated in response to an injury in 2003, the new coefficient on $Z_{2003only}$ should be positive. Unlike in the previous specification, OLS is no longer as appropriate because of censoring and skewness, but I use it for comparison to the previous specification. The new estimated coefficient on $Z_{2003only}$ in the specification with covariates indicates that employees with family injuries in 2003 spend \$244.23 more in 2004 than employees with no family injuries in 2003 or 2004. The 95% confidence interval includes only a limited negative range, providing little evidence in favor of expenditure shifting from 2004 as the mechanism for increased spending in 2003.

Overall, the results using longitudinal data indicate that employees with a family injury in 2004 but not 2003 spend more in the year of the injury than they did in the previous year, relative to people with no family injuries in either year. Examination of the panel data in the other direction indicates that employees with a family injury in 2003 but not 2004 spend more in the year of the injury than they do in the subsequent year, relative to people with no family injuries in either year. Further examination indicates that the mechanism for the increased expenditure in 2003 is not likely to be full shifting of expenditures from 2004, but these results are not conclusive. In another context, Oyer (1998) finds evidence of inter- and intra-year shifting of key business variables by business executives with annual contracts. Given the large expenditure elasticities estimated in this paper, identifying the mechanisms through which consumers manipulate medical expenditure is an important topic for future research.

1.8 Extension: Prescription Drug Cross-Price Elasticity

In the plans that I study, prescription drugs are not included in the main cost sharing provisions, so the marginal price of prescription drugs differs from the marginal price of other medical services. Thus, it is possible to examine the cross-price elasticity of prescription drug expenditure with respect to the marginal price of other medical services. To do so, I estimate the main CQIV specification with prescription drug expenditure in the place of other medical expenditure as the outcome variable.

In the drug specification, the first stage is the same as it is in the main specification, but the exclusion restriction changes. If the exclusion restriction is valid in the main specification, it is plausibly also valid in the drug specification. A violation of the exclusion restriction in the drug specification would require that one family member's injury is related to another family member's drug expenditure through a mechanism outside of the marginal price of other medical services.

The cross-price elasticity can be obtained from the coefficient on the marginal price of other medical services in the drug specification. If services and drugs are perfect complements, the cross-price elasticity of prescription drugs will be the same as the own-price elasticity of services. If they are complements, but not perfect complements, the cross-price elasticity will be smaller in magnitude but still negative. If they are not related, the cross-price elasticity will be zero, and if they are substitutes, the cross-price elasticity will be positive.

In the main estimation sample of 2004 employees, there is slightly more censoring in drug expenditures than in overall expenditures: 52% of people in the main estimation sample consume no drugs, and 38% of people in the main 2004 estimation sample consume no services. Given the higher degree of censoring, I am not able to estimate the cross-price elasticity at the .65 quantile, but I can estimate it at all of the other quantiles estimated in the main specification. However, because of reasons discussed with regard to inpatient and outpatient spending, it is not necessarily appropriate to compare estimates from the drug specification to estimates from the main

specification on a quantile-by-quantile basis, though comparisons of overall patterns can be informative.

Table 1.14 presents results from the main specification and the drug specification. Overall, the estimated cross-price elasticity varies non-systematically across the quantiles from -1.3 to -2.3, suggesting a strong complementarity between prescription drug expenditures and expenditures on other medical services. It is important to note that this complementarity is on a percentage basis, and not on a dollar-for-dollar basis. In the sample, mean prescription drug spending is $277.94, and mean spending on other services is $1,484.75.

The estimated strong complementarity is surprising given recent results by Li et al. (2005), who examine changes in prescription drug copayments over time and find evidence of substitution between prescription drugs and other medical services. Mathematically, the cross-price elasticity of other medical services with respect to the marginal price of prescription drugs should be the same as the cross-price elasticity of prescription drugs with respect to the price of other medical services. Empirically, it seems likely that the Li et al. results differ from mine precisely because their variation comes from drug prices instead of the price of other medical services. Furthermore, they examine dynamic variation in drug prices across years, and I examine within-year variation in the price of medical services. Findings from the RAND experiment, which also examined within-year price variation, suggest that complementarity is possible because prescription drug expenditures have a strong relationship to the number of visits to the doctor, which varies with plan cost sharing parameters. However, the RAND experiment did not allow for direct estimation of the cross-price elasticity because prescription drugs were covered under the same cost sharing provisions as other medical services.

In the final rows of Table 1.14, I present results from another CQIV specification, in which the dependent variable is the logarithm of the sum of prescription drug expenditures and expenditures on other medical services. Given the evidence of strong complementarity, this specification arguably estimates a more policy-relevant parameter because it captures the effect of marginal prices on a wider range of medi-

cal expenditure. The estimated coefficients suggest that the elasticity of spending on prescription drugs and other medical services with respect to the price of other medical services is approximately -1.9 across the .65 to .90 quantiles. At the .95 quantile, this elasticity is even larger.

1.9 Comparison to RAND

1.9.1 Scope of Comparison

The estimates that I present here are an order of magnitude larger than those commonly cited from the RAND experiment. There could be a multitude of reasons for this discrepancy, including a possible change in the underlying expenditure elasticity over the decades between the RAND study and my study and a difference in behavior between people in experimental plans and people in actual plans. While some potential explanations of this disagreement are difficult to assess, it is possible and instructive to examine differences in methodology behind the RAND estimates and my estimates.

Below I discuss the calculation of the RAND estimates of the price elasticity of expenditure on medical care. I emphasize that the RAND methodology assumes a myopic response to contemporaneous marginal price, and my methodology assumes a forward-looking response to year-end marginal price. Next, I present evidence of forward-looking behavior among the individuals in my data. Lastly, I conduct a simulation in my data under conditions intended to mimic the plans and assumptions of the RAND experiment. The simulation shows that by assuming myopia when some individuals are forward-looking, it is possible to estimate an elasticity that is an order of magnitude smaller than the true elasticity.

1.9.2 Review of RAND estimates

To induce subjects to participate in the RAND experiment, researchers had to guarantee that participants would be subject to very low out-of-pocket costs, so all plans

in the experiment had a yearly stoploss of $1,000 or less in 1974-1982 dollars. Furthermore, each year, all families were given lump sum payments that equaled or exceeded their out-of-pocket payments. The experimenters randomized families into plans with initial marginal prices of 0%, 25%, 50%, 95%, but after family spending reached the stoploss, marginal price was zero for the rest of the year, regardless of plan. In practice, the stoploss was binding for a large fraction (roughly 20%) of participants. Approximately 35% of individuals in the least generous plan exceeded the stoploss, as did approximately 70% of individuals with any inpatient care. To put these rates in a broader context, less than 4% of individuals met the stoploss in my non-experimental data.

RAND researchers recognized that the stoploss affected their ability to calculate the price elasticity of expenditure on medical care based on the experimentally randomized prices:

> "In order to compare our results with those in the literature, however, we must extrapolate to another part of the response surface, namely, the response to coinsurance variation when there is no maximum dollar expenditure. Although any such extrapolation is hazardous (and of little practical relevance given the considerable departure from optimality of such an insurance policy), we have undertaken such an extrapolation rather than forego entirely any comparison with the literature." (Manning et al. (1987), page 267)

Manning et al. (1987) cited three sources of estimates of the price elasticity of expenditure on medical care in the RAND data, the most prominent of which was based on a simulation by Keeler and Rolph (1988) and not on the Manning et al. (1987) four-part model. Keeler and Rolph (1988) recognized that a comparison of year-end expenditures based on the experimentally induced coinsurance rates across plans could be misleading because behavior was influenced by stoplosses. They therefore used the experimental data to simulate year-end-expenditures in hypothetical plans without stoplosses, and they based their elasticity estimates on this simulated be-

havior. To conduct the simulation, they assumed myopic responses to marginal price and examined the frequency of visits for all participants in the period for which their families still had over $400 remaining before meeting the stoploss. Notably, they included people in families that far exceeded the stoploss in the simulation. Based on calibrated parametric assumptions on the frequency of visits and the cost per visit, they forecasted year-end expenditures, and they compared forecasted expenditures across coinsurance plans relative to the free plan to attain their elasticity estimates using the following midpoint arc elasticity formula:

$$\eta_{midpoint} = \frac{(e_1 - e_2)/(e_1 + e_2)}{(p_1 - p_2)/(p_1 + p_2)} \tag{1.23}$$

where p denotes the coinsurance rate and e denotes simulated expenditures relative to the free care plan. The often-cited RAND elasticity estimate of -.22 comes from a comparison of predicted expenditures across plans with 95% and 25% coinsurance rates as follows:

$$\eta_{RAND} = \frac{(71 - 55)/(71 + 55)}{(25 - 95)/(25 + 95)} \approx -.22 \tag{1.24}$$

The magnitude of this arc elasticity should be roughly comparable the arc elasticities that I calculate, which are based on a price change from 100% before the deductible to the 20% coinsurance rate. One key methodological difference, however, is that I use within-plan price variation instead of across-plan price variation. Given the current policy environment, which focuses on the effect of high deductibles on medical spending, the ideal experiment for today's policy environment would arguably focus on price variation within plans. Another key difference between the RAND methodology and my methodology comes from the underlying treatment of myopia vs. foresight.

1.9.3 Evidence of Foresight

In the simple model of medical care expenditure on which I base my analysis, the most important parameter is the year-end marginal price. According to the model, if an individual expects to meet the stoploss by the end of the year, he will consume medical care all year as if his marginal price is zero, and expenditures paid at the randomized marginal rate will induce only an income effect. In contrast, by forecasting expenditures based on expenditure patterns before the stoploss is met, the Keeler and Rolph (1988) analysis assumes a strong form of myopia.

To address the assumption of strong myopia, I present simple suggestive evidence of forward-looking behavior in my data. The test that produces this evidence is that if individuals are forward-looking, individuals who expect to meet the deductible should not change the intra-year pattern of expenditures when a family injury occurs, but individuals who do not expect to meet the deductible should. To examine people who plausibly expected to meet the deductible in 2004 absent family injuries, I identify individuals whose 2003 own spending exceeded the 2003 individual deductible as "High 2003" spenders. I identify all other individuals as "Low 2003" spenders. Within these two 2003 spending categories, I compare average monthly expenditures before and after the month of the first family injury. As in the main estimation sample, individuals with own injuries are excluded from the sample. I also omit individuals whose first family injuries occur in January or December so that it is always possible to observe spending before and after the family injury.

The top panel of Figure 1-5 presents the results from the sample of 2,265 employees with 2004 family injuries and complete 2003 expenditure data. A comparison of the two bars on the left to the two bars on the right shows that individuals with high 2003 spending spend more on average in 2004, regardless of the timing of the family injury. Within each set of bars, the comparisons provide evidence of forward-looking behavior. As expected, the left set of bars shows that employees with low 2003 spending spend more on average after the family injury than they did before the family injury. Also as expected, the right set of bars shows that employees with high

2003 spending do not appear to alter their spending patterns in response to the timing of a family injury.

Formally, the t statistic for the paired t test of the difference in mean spending before and after the injury is 1.17 for low 2003 spenders and .1083 for high spenders, so neither difference is statistically significant. However, in the bottom panel, when instead of restricting the sample to employees, I use the entire sample, low spenders also spend more on average after the injury, and the difference in means is statistically significant for low 2003 spenders (t=-2.74) and is not statistically significant for high 2003 spenders (t=.4748). Given that this test is only conducted on the universe of people with family injuries from February-December, it is plausible that the employee sample size of 2,265 is not large enough to detect statistically significant effects, but the full sample size of 9,075 is.

This test has several limitations, notably that it relies on averages even though medical expenditures are censored and skewed, and it has imperfect controls for seasonality of medical expenditures. The question of whether consumers are myopic or forward-looking is complicated and interesting in its own right, and should be investigated more completely. However, this test provides suggestive evidence against the Keeler and Rolph (1988) assumption of myopia.

If consumers are forward-looking, it is problematic to assume that the initial statutory marginal price ever governs behavior of participants who expect to meet the stoploss, even in the period before the stoploss is met. Including these participants in the simulation should bias estimates of price responsiveness downward because variation across plans will be less pronounced among participants who expect to meet the stoploss and thus do not respond to at all to the statutory marginal price. Furthermore, participants with the highest coinsurance rates are more likely than participants with the lowest coinsurance rates to meet the stoploss, and thus they are more likely to behave as if care is free, which further attenuates elasticity estimates toward zero. More broadly, the lack of experimental price variation among the highest spenders is unfortunate because, given the skewness in the distribution of medical expenditure, the price responsiveness of the highest spenders is a very policy-relevant

parameter.

1.9.4 Simulation Exercise

To calculate expenditure elasticities, Keeler and Rolph (1988) simulated the expenditure response to plans with a higher stoploss than the true stoploss in their data. To illustrate potential bias in the Keeler and Rolph (1988) methodology, I conduct a theoretical reverse of the RAND exercise, in which I simulate the response to plans with a lower stoploss than the true stoploss in my data. One advantage of my simulation over the RAND simulation is that it leads to within-sample predictions, whereas the RAND simulation led to out-of-sample predictions.

Since the RAND simulation included people who faced a zero effective year-end marginal price but attributed their behavior to a nonzero statutory marginal price, the RAND estimates should be biased toward zero. In my simulation, I simulate behavior governed by a zero effective marginal price, but I attribute this behavior to a nonzero statutory marginal price in the estimates, and I demonstrate the magnitude of the resulting bias toward zero. Under assumptions intended to mimic the conditions of the RAND experiment in my simulation, I estimate a simulated elasticity that is an order of magnitude smaller than the true elasticity.

The simulation steps are as follows:

1. Estimate the following specification using my data and my methodology:

$$\ln E = \alpha P + X'\beta + u \tag{1.25}$$

   where all variables are defined as above. Retain estimates for subsequent steps. In practice, I estimate my model in my data using Tobit IV, and I estimate a price elasticity of -3.2. I do not use CQIV for this simulation because I am interested in a mean estimate for comparison to RAND.

2. Predict log expenditure for all individuals using the estimated coefficients and

the empirical values of P and X:

$$\widehat{\ln E} = \widehat{\alpha} P + X'\widehat{\beta} \tag{1.26}$$

3. To mimic the spending response to a new, lower stoploss than that in the actual plans, choose a group of individuals for whom the new stoploss will be low enough that they will reasonably expect to meet it. Calibrate the size of this group according to the percentage of individuals who met the stoploss in the RAND study. For this group, compute a simulated predicted expenditure, which assumes an effective marginal price of zero, even though the nominal year-end marginal price for these individuals in the actual plans is often non-zero:

$$\widetilde{\ln E} = \widehat{\alpha} * 0 + X'\widehat{\beta} \tag{1.27}$$

   Since $\widehat{\alpha} < 0$ and $P \geq 0$, it follows that $\widetilde{\ln E} > \widehat{\ln E}$. This makes intuitive sense because, given downward sloping demand, people who face a price of zero will spend more on medical care than they would if they faced a nonzero marginal price. For example, in the data, there is an individual who faces a year-end nominal marginal price of .2, and has total year-end spending of \$927.00. Based on his nominal marginal price and the values of his values of X, his predicted log spending is 5.7244, which by exponentiation, translates into \$306.25. In the simulation, when I predict his log spending based on a year-end effective price of zero, the new predicted value is 6.9970, which by exponentiation, translates into \$1,093.25.

4. Re-estimate the price elasticity using my methodology on the dataset of predicted expenditures and nominal marginal prices, and compare it to the "true" elasticity as computed by the price coefficient $\widehat{\alpha}$, estimated in the first step.

To determine whose expenditures to alter in the third step, I examine expenditures on the family level because the RAND stoplosses were on the family level. Since approximately 20% of subjects met the stoploss in the RAND study, I place

approximately 20% of my sample into in hypothetical plans in which the effective marginal price is zero. Specifically, this subset includes 6,015 people with no family injuries whose total family spending exceeds $5,500 (20.7% of the entire sample, and 23.9% of the sample with no family injuries).

It is plausible that families without injuries whose expenditures exceed $5,500 would have met the $1,000 stoploss in the RAND plans, even accounting for overall and medical inflation. In the least generous plan in my data, when family total beneficiary plus insurer spending is $5,500, beneficiary spending is $3,000+($5,500-$3,000)*.2=$3,500. Similarly, in the most generous plan in my data, when family total beneficiary plus insurer spending is $5,500, beneficiary spending is $1,050+($5,500-$1,050)*.2= $1,940. In my data, since the stoplosses are so much higher than they were in the RAND experiment, very small numbers of individuals meet the stoploss. Among the individuals whose expenditures I alter, the average statutory marginal price is .4 (29.4% at 1, 52.6% at .2, and 14.6% at 0).

When I re-estimate the model in the fourth step using predicted expenditures and nominal marginal prices, I estimate a price elasticity of -.34, which is an order of magnitude smaller than the original estimate of -3.2. It is possible to alter the expenditures of other plausibly-sized subsets of individuals to yield similar results. For example, when I alter the spending of a random 15% of individuals with no family injuries, I estimate a price elasticity of -.33. In addition, when I alter the spending of a random 50% of individuals with family spending that exceeds $2,000 and no family injuries, I estimate a price elasticity of -.28. Overall, the results of these simulation exercises suggest that if plausibly-sized groups of individuals are forward-looking, but they are assumed to be myopic, estimates of the price elasticity of expenditure on medical care could reflect a substantial bias toward zero.

1.10 Conclusion

This paper makes several contributions. Using recent, detailed data and a careful identification strategy, I use the current econometric standard to estimate mean elas-

ticities and compare my results to those in the literature. I find elasticities that are an order of magnitude larger than those in the literature. I then go beyond the current econometric standard to relax the distributional assumptions traditionally used to deal with censoring by using a new censored quantile instrumental variables estimator. My CQIV estimates vary by quantile, which is advantageous because the distribution of medical spending is so skewed. I find that the price elasticity of expenditure on medical care is very large among people who spend the most. Specifically, across the .65 to .95 quantiles of the expenditure distribution, the price elasticity of expenditure is approximately -2.3. This finding is stable across a variety of specification tests, and the results from other analyses support the robustness of the identification strategy. In an extension of my main results, I find evidence of strong complementarity between prescription drugs expenditures and expenditures on other types of medical care.

The task for my future research is to incorporate the findings from this paper into a broader model of medical care consumption. The results in this paper are inherently "reduced form" in the sense that they do not impose restrictions from economic theory to estimate the parameters. In a future paper, I intend to compare these results to results from a "structural model" that incorporates restrictions based on consumer utility maximization subject to a nonlinear budget set in the spirit of Hausman (1985). A rigorous comparison of results from both strategies could lead to a broader methodological contribution as well as a substantive one.

Bibliography

[1] Angrist, Joshua, and Imbens, Guido. "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity." Journal of the American Statistical Association. 1995. 90(430), pp. 431-442.

[2] Angrist, Joshua, Imbens, Guido, and Rubin, Donald. "Journal of the American Statistical Association." 1996, 91(434), pp. 444-455.

[3] Chandra, Amitabh, Gruber, Jonathan, and McKnight, Robin. "Medical Price Sensitivity and Optimal Health Insurance for the Elderly." 2006. mimeo.

[4] Chernozhukov, Victor, and Hansen, Christian. "Instrumental variable quantile regression: A robust inference approach." Journal of Econometrics. January 2008. 142(1), pp.379-398.

[5] Chernozhukov, Victor, and Hong, Han. "Three-Step Quantile Regression and Extramarital Affairs." Journal of The American Statistical Association. September 2002, 97(459). pp. 872-882.

[6] Chernozhukov, Victor, and Kowalski, Amanda. "Censored Quantile Instrumental Variables Regression via Control Functions." 2008. mimeo.

[7] Chesher, Andrew. "Nonparametric Identification under Discrete Variation." Econometrica, 2005, 73(5), pp.1525-1550.

[8] Duan, Naihua, Manning, Willard G Jr., Morris, Carl N., Newhouse, Joseph P. "A Comparison of Alternative Models for the Demand for Medical Care." Journal of Business & Economic Statistics, 1983, 1(2), pp. 115-126.

[9] Eichner, Matthew J. "Medical Expenditures and Major Risk Health Insurance,"Massachusetts Institute of Technology, 1997, 1-66.

[10] Eichner, Matthew J. "The Demand for Medical Care: What People Pay Does Matter." The American Economic Review. Papers and Proceedings of the Hundred and Tenth Annual Meeting of the American Economic Association. May 1998. 88(2) pp. 117-121.

[11] Eichner, Matthew J.; McClellan, Mark B.; and Wise, David A. NBER Tax Policy and the Economy. Cambridge, M.A.: MIT Press, 1997, 92-128.

[12] Finkelstein, Amy. "The Aggregate Effects of Health Insurance: Evidence from the Introduction of Medicare." Quarterly Journal of Economics, 2007, 122(3), pp. 1-37.

[13] Hausman, Jerry A. "Specification Tests in Econometrics." Econometrica, 1978, 46(6), pp. 1251-71.

[14] Hausman, Jerry A. "The Econometrics of Nonlinear Budget Sets." Econometrica, 1985, 53(6), pp. 1255-82.

[15] Hsu, John, Price, Mary, Huang, Jie, Brand, Richard, Fung, Vicki, Hui, Rita, Fireman, Bruce, Newhouse, Joseph P, and Selby, Joseph V. "Unintended Consequences of Caps on Medicare Drug Benefits." The New England Journal of Medicine, 2006, 354(22), pp. 2349-2359.

[16] Kaiser Family Foundation and Health Research Educational Trust. "Employer Health Benefits Annual Survey 2006."

[17] Keeler, Emmett. "Effects of Cost Sharing on Use of Medical Services and Health." Journal of Medical Practice Management, 1992, 8(Summer), pp. 317-21.

[18] Keeler, Emmett and Rolph, John E. "The Demand for Episodes of Treatment in the Health Insurance Experiment." Journal of Health Economics, 1988, 7, pp. 337-367.

[19] Li, Jian, Gaynor, Martin, and Vogt, William B. "Is Drug Coverage a Free Lunch? Cross-Price Elasticities and the Design of Prescription Drug Benefits." Carnegie Mellon University mimeo. December 2005.

[20] Manning, Willard G., Newhouse, Joseph P., Duan, Naihua, Keeler, Emmett B., and Leibowitz, Arleen. "Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment." The American Economic Review, Jun 1987, 77(3), pp. 251-277.

[21] "Medical Expenditure Panel Survey." Agency for Healthcare Research and Quality. 2004.

[22] "MarketScan Database," Ann Arbor,MI: The MEDSTAT Group Inc., 2003.

[23] Mullahy, John. "Much ado about two: reconsidering retransformation and the two-part model in health econometrics." Journal of Health Economics, 1998, 17, pp. 247-281.

[24] Newey, Whitney K. "Efficient Estimation of Limited Dependent Variable Models with Endogenous Explanatory Variables." Journal of Econometrics, 1987, 36, pp. 231-250.

[25] Newhouse, Joseph P. "Medical-Care Expenditure: A Cross-National Survey." The Journal of Human Resources, 1977, 12(1), pp. 115-125.

[26] Newhouse, Joseph P and the Insurance Experiment Group. Free for All? Lessons from the RAND Health Insurance Experiment. Harvard University Press. Cambridge: 1993.

[27] Oyer, Paul. "Fiscal Year Ends and Nonlinear Incentive Contracts: The Effect on Business Seasonality." The Quarterly Journal of Economics, 1998, 113(1), pp. 149-185.

[28] Powell, James L. "Censored Regression Quantiles." Journal of Econometrics, 1986, 23, pp.143-155.

[29] Tobin, J. "Estimation of Relationships for Limited Dependent Variables." Econometrica, 1958, pp. 24-36.

Figure 1-1: Cost Sharing for Individuals

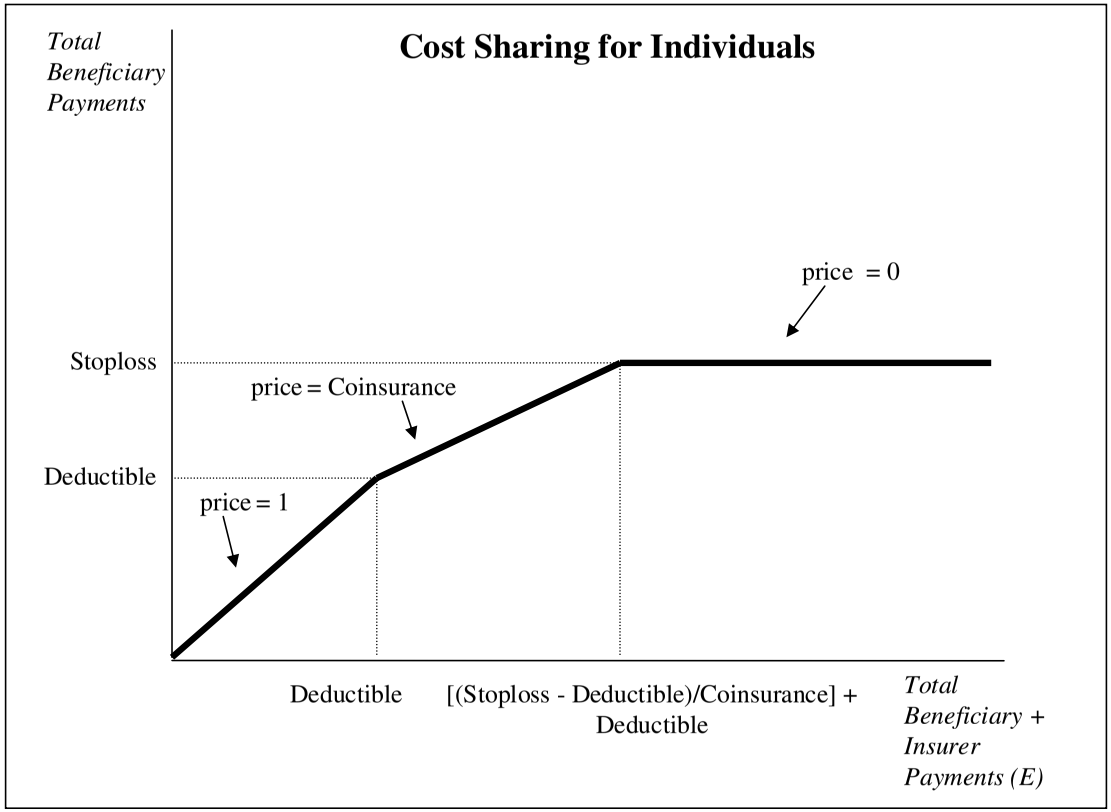


Figure 1-2: Empirical Cost Sharing for Individuals

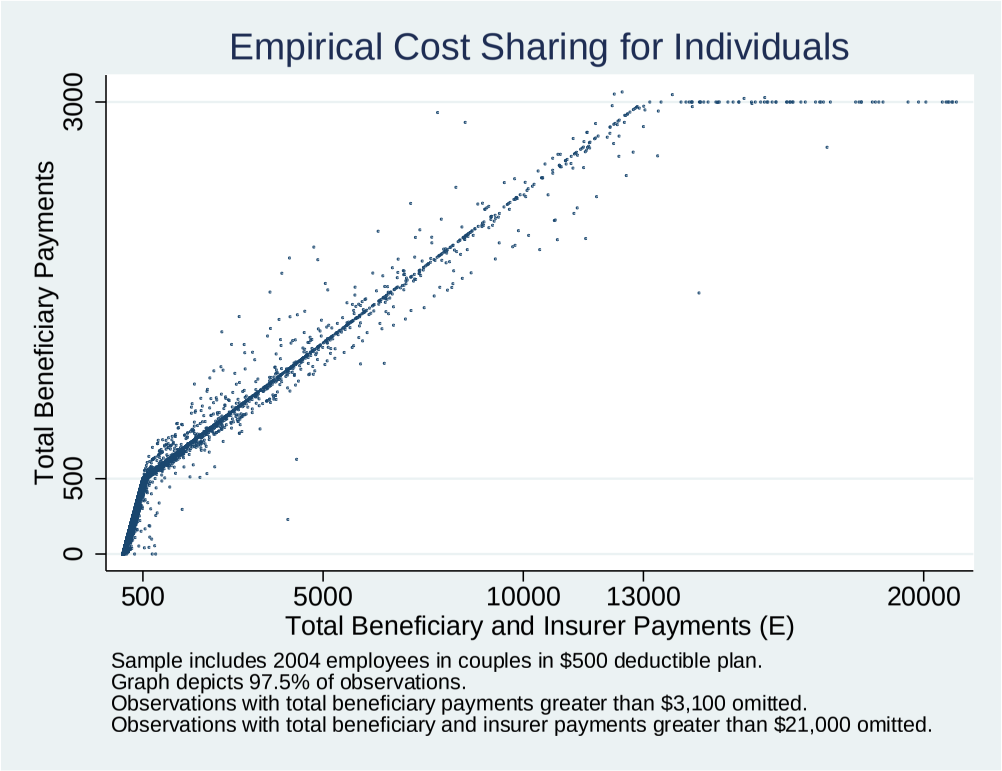
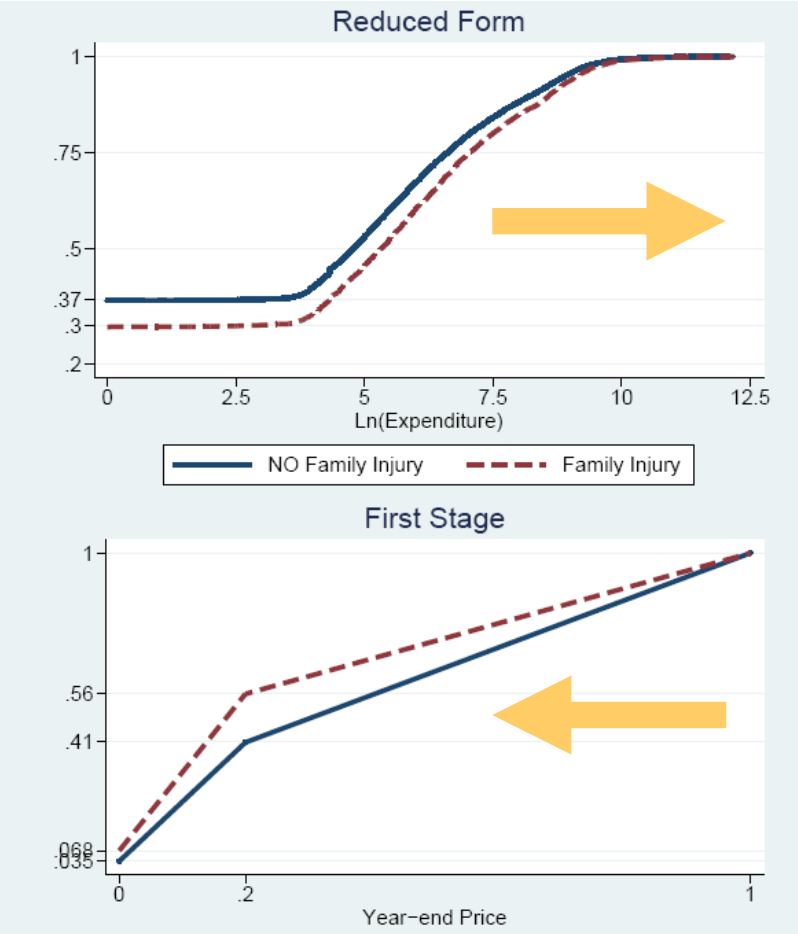


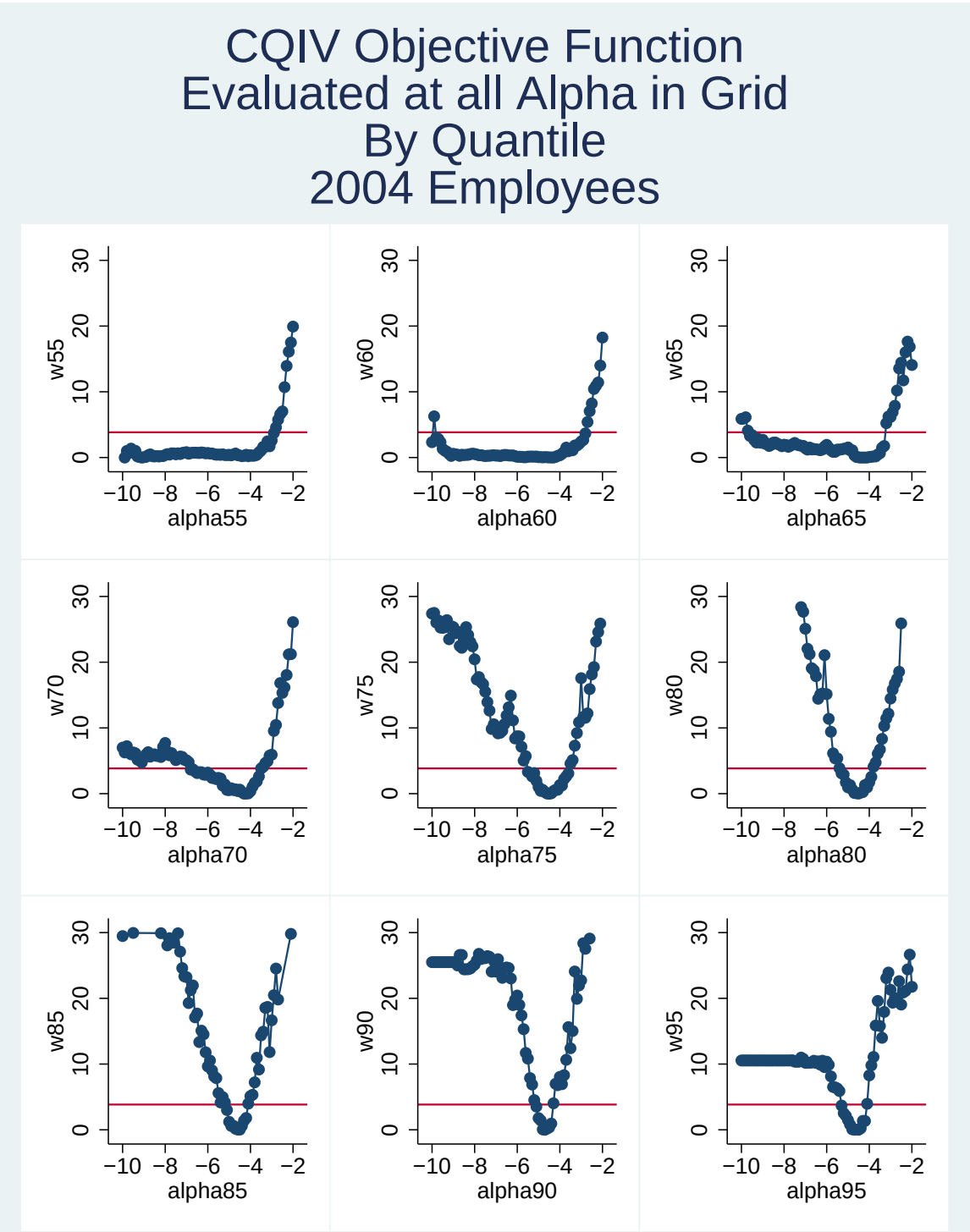
Figure 1-3: Reduced Form and First Stage



Higher Expenditure

Lower Year-end Price

Figure 1-4: CQIV Objective Function



Values of objective function above 30 omitted from graphs.
Alpha grid over −10 to −2 in increments of 0.1.
Horizontal line depicts the critical value for a 95% confidence interval on alpha.

Figure 1-5: Expenditure Before and After Month of First Family Injury

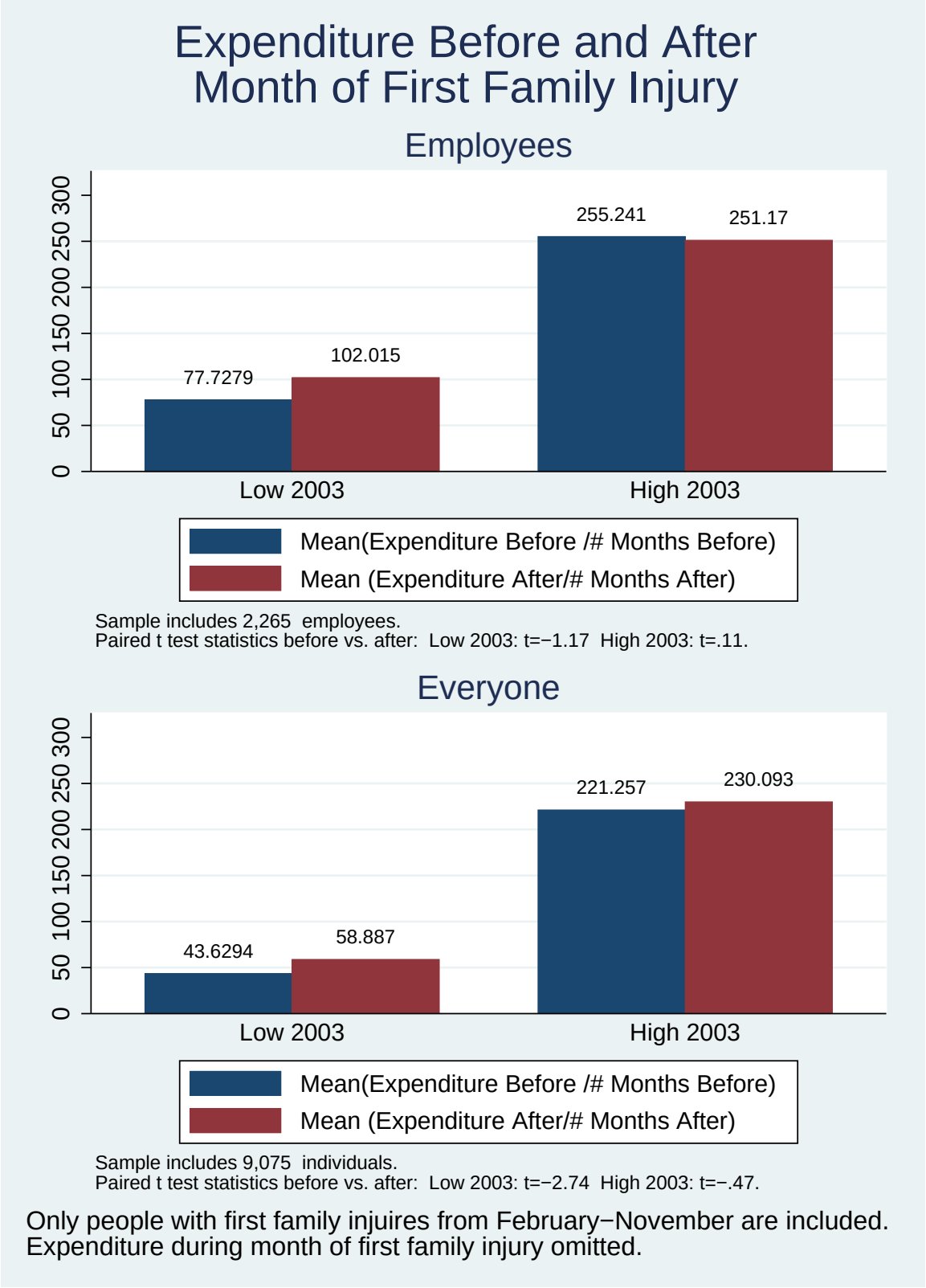


Table 1.1: Plan Comparison

Cost Sharing Comparison		Plan A	Plan B	Plan C	Plan D
Deductible	Individual	$350	$500	$750	$1,000
	Family	$1,050	$1,500	$2,250	$3,000
Stoploss (Includes Deductible)	Individual	$2,100	$3,000	$4,500	$6,000
	Family	$4,550	$6,500	$9,750	$10,000 $13,000 in 2004
Coinsurance (Beneficiary)	In-Network	20%	20%	20%	20%
	Out-of-Network	40%	40%	40%	40%

Table 1.2: 2004 Summary Statistics

2004 Summary Statistics

Cells report column % by variable

		Families of Four or More						Couples
		Employees	Everyone	Employees		Everyone		Employees
		All	All	NO Family Injury	Family Injury	NO Family Injury	Family Injury	All
Variable		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Year-end Expenditure ($)								
	0	35.7	39.9	36.6	29.8	40.9	32.3	24.2
	.01 to 100.00	11.0	12.2	11.0	10.9	12.3	11.4	7.9
	100.01 to 1,000	31.1	31.4	30.8	32.8	30.9	35.0	33.8
	1,000.01 to 10,000	19.0	14.4	18.5	22.1	13.8	18.2	27.6
	10,000.01 to 100,000	3.2	2.1	3.0	4.5	2.0	3.0	6.4
	100,000.01 and up	0.0	0.0	0.0	0.0	0.0	0.0	0.1
Year-end Price								
	0	3.9	3.1	3.5	6.8	2.7	6.1	6.7
	0.2	38.8	32.8	37.2	49.1	30.9	46.0	47.2
	1	57.3	64.1	59.3	44.1	66.4	48.0	46.1
Family Injury								
	0 (NO Family Injury)	86.6	87.4	100.0	0.0	100.0	0.0	96.1
	1 (Family Injury)	13.4	12.6	0.0	100.0	0.0	100.0	3.9
Family Size								
	2	0.0	0.0	0.0	0.0	0.0	0.0	100.0
	3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	4	66.9	60.2	68.2	58.2	61.7	49.6	0.0
	5	24.4	27.5	23.8	28.5	26.9	31.6	0.0
	6	6.6	8.8	6.1	9.6	8.3	12.5	0.0
	7	1.6	2.5	1.4	2.8	2.3	4.3	0.0
	8 to 11	0.5	1.0	0.5	0.9	0.9	1.9	0.0
Relation to Employee								
	Employee	100.0	22.8	100.0	100.0	22.6	24.3	100.0
	Spouse	0.0	19.0	0.0	0.0	18.9	19.8	0.0
	Child/Other	0.0	58.2	0.0	0.0	58.5	55.9	0.0
Male								
	0 (Female)	42.6	49.9	42.7	41.9	49.9	50.2	60.2
	1 (Male)	57.4	50.1	57.3	58.1	50.1	49.8	39.8
Year of Birth								
	1934 to 1943	0.1	0.1	0.1	0.2	0.1	0.1	10.9
	1944 to 1953	4.0	1.8	4.1	3.2	1.8	1.5	44.3
	1954 to 1963	30.9	12.9	31.1	29.7	12.9	12.8	26.5
	1964 to 1973	51.8	20.8	51.5	53.7	20.5	22.7	10.6
	1974 to 1983	13.2	7.0	13.2	13.2	6.9	7.6	7.6
	1984 to 1993	0.0	27.9	0.0	0.1	28.0	27.1	0.1
	1994 to 1998	0.0	16.0	0.0	0.0	16.1	15.4	0.0
	1999 to 2004	0.0	13.5	0.0	0.0	13.6	12.8	0.0
Employee Class								
	Salary Non-union	29.9	30.2	29.9	30.4	30.2	30.0	10.3
	Hourly Non-union	70.1	69.8	70.1	69.6	69.8	70.0	89.7
US Census Region								
	New England	1.4	1.4	1.4	1.5	1.4	1.6	1.6
	Middle Atlantic	1.6	1.6	1.6	1.3	1.6	1.2	1.7
	East North Central	15.6	15.7	15.8	14.5	15.8	15.1	14.2
	West North Central	11.9	12.0	11.8	12.2	12.0	12.0	11.1
	South Atlantic	19.0	18.9	19.3	16.9	19.2	17.2	23.7
	East South Central	11.6	11.3	11.2	14.4	11.0	13.7	13.9
	West South Central	28.3	28.3	28.4	27.4	28.5	27.3	24.5
	Mountain	7.5	7.6	7.3	8.4	7.5	8.3	6.3
	Pacific	3.1	3.2	3.1	3.4	3.1	3.5	2.9
Plan by Individual Deductible								
	350	59.8	59.9	58.7	67.2	58.7	67.8	67.1
	500	17.0	16.9	17.3	15.6	17.2	15.2	15.4
	750	6.3	6.3	6.6	4.8	6.5	4.7	5.3
	1000	16.8	16.9	17.5	12.4	17.6	12.3	12.2
Sample Size		*29,010*	*127,119*	*25,124*	*3,886*	*111,124*	*15,995*	*37,490*

Table 1.3: 2003 Summary Statistics

2003 Summary Statistics

Cells report column % by variable

		Families of Four or More					
	Employees	Everyone	Employees		Everyone	
	All	All	NO Family Injury	Family Injury	NO Family Injury	Family Injury
Variable	(1)	(2)	(3)	(4)	(5)	(6)
Year-end Expenditure ($)						
0	36.8	40.7	38.0	28.7	41.9	31.6
.01 to 100.00	11.4	12.5	11.4	11.4	12.5	12.2
100.01 to 1,000	31.1	31.7	30.7	34.4	31.1	36.2
1,000.01 to 10,000	17.8	13.3	17.2	22.3	12.7	17.5
10,000.01 to 100,000	2.8	1.8	2.7	3.3	1.7	2.4
100,000.01 and up	0.0	0.1	0.0	0.1	0.0	0.1
Year-end Price						
0	3.4	2.5	3.1	5.3	2.2	4.8
0.2	37.5	31.6	35.8	49.4	29.8	44.9
1	59.1	65.9	61.1	45.3	68.0	50.3
Family Injury						
0 (NO Family Injury)	87.7	88.3	100.0	0.0	100.0	0.0
1 (Family Injury)	12.3	11.7	0.0	100.0	0.0	100.0
Family Size						
2	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0
4	66.1	59.2	67.2	58.1	60.5	49.3
5	24.9	27.9	24.3	29.0	27.4	32.1
6	6.8	9.1	6.4	9.5	8.6	12.8
7	1.7	2.6	1.5	2.4	2.4	3.9
8 to 12	0.6	1.2	0.6	0.9	1.1	1.9
Relation to Employee						
Employee	100.0	22.7	100.0	100.0	22.5	23.9
Spouse	0.0	18.8	0.0	0.0	18.7	19.9
Child/Other	0.0	58.5	0.0	0.0	58.8	56.2
Male						
0 (Female)	43.4	50.1	43.4	43.1	50.1	50.2
1 (Male)	56.6	49.9	56.6	56.9	49.9	49.8
Year of Birth						
1934 to 1943	0.2	0.1	0.2	0.2	0.1	0.1
1944 to 1953	4.7	2.1	4.8	4.4	2.1	2.0
1954 to 1963	33.9	14.1	34.0	33.3	14.0	14.6
1964 to 1973	50.2	20.0	50.0	51.5	19.8	21.6
1974 to 1983	10.9	6.7	11.0	10.6	6.7	6.8
1984 to 1993	0.0	30.4	0.0	0.0	30.5	29.8
1994 to 1998	0.0	15.5	0.0	0.0	15.6	14.8
1999 to 2003	0.0	11.0	0.0	0.0	11.4	10.2
Employee Class						
Salary Non-union	29.4	29.8	29.4	29.5	29.9	29.0
Hourly Non-union	70.6	70.2	70.6	70.5	70.1	71.0
US Census Region						
New England	1.4	1.4	1.4	1.6	1.4	1.7
Middle Atlantic	1.7	1.7	1.7	1.7	1.7	1.5
East North Central	15.5	15.6	15.4	16.7	15.5	16.9
West North Central	12.3	12.2	11.8	15.4	11.9	15.1
South Atlantic	18.5	18.4	19.2	13.5	19.1	13.2
East South Central	10.8	10.7	10.7	11.3	10.6	11.7
West South Central	29.1	29.1	29.2	28.7	29.2	28.5
Mountain	7.8	8.0	7.7	8.4	7.9	8.7
Pacific	2.9	2.9	2.9	2.8	2.9	2.8
Plan by Individual Deductible						
350	63.4	63.4	62.7	68.8	62.7	69.0
500	17.1	17.0	17.4	15.6	17.2	15.4
750	5.7	5.6	5.8	4.5	5.8	4.2
1000	13.8	13.9	14.1	11.0	14.2	11.4
Sample Size	*29,886*	*131,815*	*26,201*	*3,685*	*116,393*	*15,422*

Table 1.4: Comparison of Skewness

Comparison of Skewness

	2004 Full Sample		2004 MEPS	
Expenditure Rank	$ Expenditure	% Expenditure	$ Expenditure	% Expenditure
top 1%	16,074	34.00%	26,881	27.68%
top 5%	5,457	64.90%	8,282	54.89%
top 10%	2,267	80.69%	4,362	70.10%
top 15%	1,173	87.91%	2,754	78.83%
top 20%	717	91.96%	1,848	84.40%
top 25%	473	94.53%	1,305	88.15%
top 50%	84	99.58%	360	97.02%
mean	1,135		2,019	
N	127,119	127,119	28,990	28,990

2004 MEPS expenditures based on the sample of individuals under age 65.

2004 MEPS expenditures exclude prescription drug expenditures.

Table 1.5: Summary Statistics on Individuals with Injuries and Their Families

Summary Statistics on Individuals with Injuries and Their Families

	Injured Individuals (Excluded from Estimation Sample)		Non-Injured Individuals in Family (Estimation Sample)		
2004 Sample	Count	Mean Expenditure	Count of Everyone	Count of Employees and Spouses	Count of Employees
Intracranial Injuries	331	$9,873.39	1,049	480	272
Superficial Injuries	1,276	$2,447.52	4,172	1,846	1,014
Crushing Injuries	59	$2,296.21	196	83	46
Foreign Body Injuries	536	$2,591.30	1,764	805	443
Burns	238	$3,146.49	819	336	189
Complications of Trauma and Injuries to the Nerves and Spinal Cord	3,241	$4,639.26	10,069	4,451	2,462
All Injuries	5,249	$3,871.19	15,995	7,052	3,886
No Injury	127,119	$1,134.83	111,124	46,133	25,124
Everyone	132,368	$1,243.34	127,119	53,185	29,010
2003 Sample					
Intracranial Injuries	293	$11,134.06	1,004	465	249
Superficial Injuries	1,178	$2,291.38	3,857	1,702	927
Crushing Injuries	62	$5,937.69	197	92	50
Foreign Body Injuries	462	$2,516.10	1,541	685	390
Burns	250	$8,873.55	868	354	205
Complications of Trauma and Injuries to the Nerves and Spinal Cord	3,168	$4,125.15	9,809	4,300	2,328
All Injuries	5,031	$3,789.94	15,422	6,761	3,685
No Injury	131,815	$1,038.19	116,393	47,922	26,201
Everyone	136,846	$1,139.36	131,815	54,683	29,886

Note: Categories of selected injuries need not be mutually exclusive.
Statistics on non-injured people in family exclude people with ANY type of selected injury.

Table 1.6: Comparison of Traditional Estimators

Comparison of Traditional Estimators

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Estimator	Trunc MLE	Trunc MLE IV	Probit	OLS 2PM	IV 2PM	Tobit	Tobit IV	Tobit IV mfx
Dependent Variable	ln(Expend)	ln(Expend)	1(Expend>0)	NA	NA	ln(Expend)*	ln(Expend)*	ln(Expend)*
Year-end price	-2.7	-1.7	-2.4	-3.8	-3.2	-8.2	-6.4	-3.9
lower bound	-2.8	-2.2	-2.5	-3.9	-3.8	-8.3	-7.4	-4.5
upper bound	-2.7	-1.1	-2.3	-3.7	-2.7	-8.1	-5.3	-3.2
[Elasticity]	-[1.4]	-[0.8]	NA	-[1.9]	-[1.6]	-[4.1]	-[3.2]	-[1.9]
Population	Expend>0	Expend>0	all	all	all	all	all	all
N	18,646	18,646	29,010	29,010	29,010	29,010	29,010	29,010

All regressions estimated on the sample of 2004 employees.
* In logarithmic specifications that use all observations, the left-hand side is set to -.7 when Expend=0.
Probit average estimated inverse mills ratio: .450.
2PM coefficients obtained as the sum of OLS or IV coefficient plus the problit average estimated inverse mills ratio multiplied by the probit coefficient.
Tobit IV mfx average scaling factor: .610.
Controls include: male dummy, plan (saturated), census region (saturated), salary dummy (vs. hourly), spouse on policy dummy, YOB of oldest dependent, YOB of youngest dependent, family size (saturated with 8-11 as one group), count family born 1944 to 1953, count family born 1954 to 1963, count family born 1974 to 1983, count family born 1984 to 1993, count family born 1994 to 1998, count family born 1999, count family born 2000, count family born 2001, count family born 2002, count family born 2003, count family born 2004 (when applicable).

Table 1.7: 2004 and 2003 CQIV Year-End Price Coefficients for Various Samples

2004 and 2003 CQIV Year-End Price Coefficients for Various Samples

Dependent variable: Ln(Expenditure)

		Censored Quantile IV							
2004 Sample		65	70	75	80	85	90	95	Tobit IV
Employee									
N= 29,010	Year-end price	-4.4	-4.3	-4.5	-4.5	-4.5	-4.7	-4.5	-6.4
	lower bound	-9.6	-6.8	-5.5	-5.3	-5.1	-5.1	-5.3	-7.4
	upper bound	-3.3	-3.5	-3.6	-3.9	-4.2	-4.4	-4.2	-5.3
	[Elasticity]	-[2.2]	-[2.2]	-[2.3]	-[2.3]	-[2.3]	-[2.3]	-[2.3]	-[3.2]
Employee and Spouse									
N= 53,185	Year-end price	-4.9	-4.8	-4.7	-4.6	-4.6	-4.7	-4.8	-6.6
	lower bound	-8.8	-5.4	-5.3	-5.2	-5.1	-5	-5.2	-7.3
	upper bound	-3.9	-4.2	-4.1	-4.2	-4	-4.3	-4.2	-5.9
	[Elasticity]	-[2.5]	-[2.4]	-[2.3]	-[2.3]	-[2.3]	-[2.3]	-[2.4]	-[3.3]
Everyone									
N= 127,119	Year-end price	-4.1	-4.1	-4.1	-4.1	-4.2	-4.3	-4.1	-6.8
	lower bound	-4.6	-4.7	-4.6	-4.3	-4.5	-4.7	-4.6	-7.3
	upper bound	-3.9	-3.6	-3.7	-3.7	-3.7	-3.9	-3.9	-6.3
	[Elasticity]	-[2.0]	-[2.0]	-[2.0]	-[2.0]	-[2.1]	-[2.2]	-[2.0]	-[3.4]
2003 Sample									
Employee									
N= 29,886	Year-end price	-9.1	-4.9	-4.6	-4.3	-4.5	-4.5	-4.5	-7.5
	lower bound	-10	-9.7	-5.8	-5.2	-5.1	-5	-10	-8.6
	upper bound	-3.7	-4.1	-3.9	-3.8	-3.6	-3.8	-4	-6.5
	[Elasticity]	-[4.6]	-[2.5]	-[2.3]	-[2.2]	-[2.3]	-[2.3]	-[2.3]	-[3.8]
Employee and Spouse									
N= 54,683	Year-end price	-9.7	-5.4	-5.1	-4.6	-4.6	-4.6	-4.6	-7.8
	lower bound	-10	-9.5	-5.8	-5.5	-5.2	-5	-5.4	-8.6
	upper bound	-4.9	-4.6	-4.3	-4	-4	-4	-4.2	-7.1
	[Elasticity]	-[4.8]	-[2.7]	-[2.5]	-[2.3]	-[2.3]	-[2.3]	-[2.3]	-[3.9]
Everyone									
N= 131,815	Year-end price	-5.4	-4.5	-4.5	-4.3	-4.2	-4.2	-4.2	-7.7
	lower bound	-7.4	-5.1	-4.9	-4.9	-4.6	-4.7	-4.9	-8.3
	upper bound	-4.4	-4.1	-4	-4	-4	-3.9	-4	-7.2
	[Elasticity]	-[2.7]	-[2.3]	-[2.3]	-[2.2]	-[2.1]	-[2.1]	-[2.1]	-[3.9]

Censored quantile IV results from a grid search over -10 to -2 in increments of .1.

Controls include: employee dummy (when applicable), spouse dummy (when applicable), male dummy, plan (saturated), census region (saturated), salary dummy (vs. hourly), spouse on policy dummy, YOB of oldest dependent, YOB of youngest dependent, family size (saturated with 8-11 as one group), count family born 1944 to 1953, count family born 1954 to 1963, count family born 1974 to 1983, count family born 1984 to 1993, count family born 1994 to 1998, count family born 1999, count family born 2000, count family born 2001, count family born 2002, count family born 2003, count family born 2004 (when applicable).

Table 1.8: Closer Examination of Endogeneity

Closer Examination of Endogeneity

Dependent Variable: Ln(Expenditure)

	Censored Quantile (IV)							
	65	70	75	80	85	90	95	Tobit (IV)
CQIV								
Year-end price	-4.4	-4.3	-4.5	-4.5	-4.5	-4.7	-4.5	-6.4
lower bound	-9.6	-6.8	-5.5	-5.3	-5.1	-5.1	-5.3	-7.4
upper bound	-3.3	-3.5	-3.6	-3.9	-4.2	-4.4	-4.2	-5.3
[Elasticity]	-[2.2]	-[2.2]	-[2.3]	-[2.3]	-[2.3]	-[2.3]	-[2.3]	-[3.2]
CQR - Chernozhukov and Hong								
Year-end price	-4.63	-4.51	-4.44	-4.37	-4.25	-4.19	-4.15	-8.21
lower bound	-4.72	-4.59	-4.51	-4.44	-4.31	-4.24	-4.19	-8.33
upper bound	-4.54	-4.44	-4.38	-4.30	-4.19	-4.14	-4.11	-8.09
[Elasticity]	-[2.32]	-[2.26]	-[2.22]	-[2.19]	-[2.13]	-[2.10]	-[2.08]	-[4.11]
CQR - Powell								
Year-end price	-4.67	-4.47	-4.44	-4.33	-4.27	-4.20	-4.15	-8.21
lower bound	-4.75	-4.54	-4.51	-4.39	-4.33	-4.25	-4.19	-8.33
upper bound	-4.60	-4.40	-4.38	-4.28	-4.21	-4.15	-4.12	-8.09
[Elasticity]	-[2.34]	-[2.24]	-[2.22]	-[2.17]	-[2.13]	-[2.10]	-[2.08]	-[4.11]

Estimated on the 2004 sample of 29,010 employees in families.

Controls include: male dummy, plan (saturated), census region (saturated), salary dummy (vs. hourly), spouse on policy dummy, YOB of oldest dependent, YOB of youngest dependent, family size (saturated with 8-11 as one group), count family born 1944 to 1953, count family born 1954 to 1963, count family born 1974 to 1983, count family born 1984 to 1993, count family born 1994 to 1998, count family born 1999, count family born 2000, count family born 2001, count family born 2002, count family born 2003, count family born 2004.

Table 1.9: Month of Family Injury

Month of First Family Injury

	2004 Estimation Sample			
	Everyone		Employees	
	Count	%	Count	%
Jan	1,019	6.4%	255	6.6%
Feb	1,050	6.6%	253	6.5%
Mar	1,329	8.3%	326	8.4%
Apr	1,512	9.5%	382	9.8%
May	1,566	9.8%	377	9.7%
Jun	1,536	9.6%	387	10.0%
Jul	1,445	9.0%	353	9.1%
Aug	1,542	9.6%	367	9.4%
Sep	1,554	9.7%	376	9.7%
Oct	1,488	9.3%	357	9.2%
Nov	1,039	6.5%	250	6.4%
Dec	915	5.7%	203	5.2%
Total	15,995	100%	3,886	100%

Samples limited to individuals in families with injuries.

Table 1.10: CQIV Specification Tests

CQIV Specification Tests

Dependent variable: Ln(Expenditure) or Ln(Outpatient Expenditure)

		Censored Quantile IV							
2004 Employee Sample		65	70	75	80	85	90	95	Tobit IV
Baseline									
N= 29,010	Year-end price	-4.4	-4.3	-4.5	-4.5	-4.5	-4.7	-4.5	-6.4
	lower bound	-9.6	-6.8	-5.5	-5.3	-5.1	-5.1	-5.3	-7.4
	upper bound	-3.3	-3.5	-3.6	-3.9	-4.2	-4.4	-4.2	-5.3
	[Elasticity]	-[2.2]	-[2.2]	-[2.3]	-[2.3]	-[2.3]	-[2.3]	-[2.3]	-[3.2]
First Family Injury Jan-Sep (Q1-Q3)									
N= 28,176	Year-end price	-4.2	-4.3	-4.5	-4.5	-4.6	-4.6	-4.6	-6.7
	lower bound	-9.6	-6.9	-5.6	-5.5	-5.2	-5.4	-5.4	-7.7
	upper bound	-3.3	-3.6	-3.7	-3.9	-4.2	-4.2	-4.1	-5.6
	[Elasticity]	-[2.1]	-[2.2]	-[2.3]	-[2.3]	-[2.3]	-[2.3]	-[2.3]	-[3.3]
First Family Injury Oct-Dec (Q4)									
N= 25,934	Year-end price	-3.2	-3.1	-4.0	-4.1	-4.1	-4.8	-4.6	-4.6
	lower bound	-10.0	-10.0	-7.2	-5.9	-5.7	-5.5	-10.0	-7.8
	upper bound	-2.0	-2.0	-2.0	-2.5	-2.7	-2.0	-2.0	-1.3
	[Elasticity]	-[1.6]	-[1.5]	-[2.0]	-[2.0]	-[2.0]	-[2.4]	-[2.3]	-[2.3]
First Family Injury Jan-Jun (Q1-Q2)									
N= 27,104	Year-end price	-5.7	-5.4	-5.0	-5.2	-4.8	-4.9	-4.9	-7.3
	lower bound	-10.0	-10.0	-7.3	-6.6	-7.0	-5.8	-5.9	-8.5
	upper bound	-3.8	-4.0	-4.3	-4.0	-4.2	-4.2	-4.2	-6.0
	[Elasticity]	-[2.8]	-[2.7]	-[2.5]	-[2.6]	-[2.4]	-[2.5]	-[2.5]	-[3.6]
First Family Injury Jul-Dec (Q3-Q4)									
N= 27,030	Year-end price	-2.7	-3.0	-3.6	-3.9	-4.4	-4.5	-4.2	-5.0
	lower bound	-9.7	-6.7	-5.3	-4.9	-5.0	-4.9	-5.5	-6.8
	upper bound	-2.0	-2.0	-2.3	-2.8	-3.0	-3.6	-3.9	-3.3
	[Elasticity]	-[1.4]	-[1.5]	-[1.8]	-[2.0]	-[2.2]	-[2.3]	-[2.1]	-[2.5]
Injuries to Children Only									
N= 25,386	Year-end price	-4.0	-4.2	-4.6	-4.6	-4.6	-4.7	-4.5	-6.3
	lower bound	-9.7	-6.8	-5.5	-5.4	-5.4	-5.1	-5.1	-7.4
	upper bound	-3.3	-3.3	-3.6	-3.9	-4.1	-4.3	-4.1	-5.2
	[Elasticity]	-[2.0]	-[2.1]	-[2.3]	-[2.3]	-[2.3]	-[2.3]	-[2.3]	-[3.1]
Injuries to Spouses Only									
N= 25,884	Year-end price	-4.7	-4.3	-4.3	-4	-4.2	-4.5	-4.6	-6.8
	lower bound	-10	-10	-9.6	-6.6	-6.9	-5.5	-5.5	-9.0
	upper bound	-2.5	-2.8	-2.5	-2.7	-2.7	-2.7	-3.5	-4.6
	[Elasticity]	-[2.3]	-[2.2]	-[2.2]	-[2.0]	-[2.1]	-[2.3]	-[2.3]	-[3.4]
$350 Deductible Plan Only									
N= 17,353	Year-end price	-3.9	-4.2	-4.4	-4.6	-4.5	-4.5	-4.7	-5.8
	lower bound	-9.5	-6.1	-5.4	-4.9	-4.9	-4.9	-5.6	-6.8
	upper bound	-3.3	-3.3	-3.6	-3.7	-4.1	-4.2	-4.1	-4.8
	[Elasticity]	-[2.0]	-[2.1]	-[2.2]	-[2.3]	-[2.3]	-[2.3]	-[2.3]	-[2.9]
Ln(Outpatient Expenditure)									
N= 29,010	Year-end price	-4.0	-4.0	-3.8	-4.0	-4.0	-4.0	-3.7	-6.1
	lower bound	-5.9	-4.9	-4.8	-4.5	-4.5	-4.5	-5.1	-7.0
	upper bound	-2.9	-3.0	-3.2	-3.4	-3.5	-3.6	-3.2	-5.2
	[Elasticity]	-[2.0]	-[2.0]	-[1.9]	-[2.0]	-[2.0]	-[2.0]	-[1.9]	-[3.1]

		Censored Quantile IV				
		97.5	98	98.5	99	99.5
Baseline (Higher Estimated Quantiles)						
N= 29,010	Year-end price	-4.7	-5.0	-5.2	-5.6	-4.3
	lower bound	-10.0	-10.0	-10.0	-10.0	-10.0
	upper bound	-4.2	-4.2	-4.1	-3.5	-2.0
	[Elasticity]	-[2.4]	-[2.5]	-[2.6]	-[2.8]	-[2.2]

Censored quantile IV results from a grid search over -10 to -2 in increments of .1.

Controls include: employee dummy (when applicable), spouse dummy (when applicable), male dummy, plan (saturated), census region (saturated), salary dummy (vs. hourly), spouse on policy dummy, YOB of oldest dependent, YOB of youngest dependent, family size (saturated with 8-11 as one group), count family born 1944 to 1953, count family born 1954 to 1963, count family born 1974 to 1983, count family born 1984 to 1993, count family born 1994 to 1998, count family born 1999, count family born 2000, count family born 2001, count family born 2002, count family born 2003, count family born 2004 (when applicable).

Table 1.11: OLS First Stage By Plan

OLS First Stage By Plan

Dependent Variable: Year-end Price

		By Plan (Individual Deductible)			
	All	350	500	750	1000
2004 Employee Sample	(1)	(2)	(3)	(4)	(5)
Family Injury	-0.111	-0.132	-0.083	-0.071	-0.044
lower bound	-0.124	-0.149	-0.115	-0.125	-0.075
upper bound	-0.098	-0.116	-0.050	-0.017	-0.014
Controls	yes	yes	yes	yes	yes
R-squared	0.123	0.073	0.089	0.110	0.099
N	29,010	17,353	4,945	1,834	4,878

Controls include: plan (saturated, when applicable), male dummy, census region (saturated), salary dummy (vs. hourly), spouse on policy dummy, YOB of oldest dependent, YOB of youngest dependent, family size (saturated with 8-11 as one group), count family born 1944 to 1953, count family born 1954 to 1963, count family born 1974 to 1983, count family born 1984 to 1993, count family born 1994 to 1998, count family born 1999, count family born 2000, count family born 2001, count family born 2002, count family born 2003, count family born 2004.

Table 1.12: Robustness Test: Effect of Family Injury on Couples and Families

Robustness Test: Effect of Family Injury on Couples and Families

Dependent variable: Ln(Expenditure)

			Censored Quantile Regression							
Employees in Couples			65	70	75	80	85	90	95	Tobit
N=	29010[+]	Family Injury	0.18	0.11	0.20	0.13	0.03	-0.03	-0.08	0.43
Mean Expenditure:	$2,882.57	lower bound	0.01	-0.06	0.01	-0.04	-0.15	-0.22	-0.30	0.17
		upper bound	0.35	0.29	0.38	0.31	0.21	0.17	0.15	0.69
		Includes zero:	no	yes	no	yes	yes	yes	yes	no
Employees in Families of Four or More										
N=	29,010	Family Injury	0.45	0.43	0.42	0.43	0.39	0.34	0.27	0.84
Mean Expenditure:	$1,484.74	lower bound	0.33	0.32	0.32	0.31	0.27	0.23	0.16	0.65
		upper bound	0.58	0.53	0.53	0.55	0.52	0.45	0.38	1.02
		Includes zero:	no	no	no	no	no	no	no	no
Employees in Families of Four or with Injury to Spouse or No Family Injury										
N=	25,884	Family Injury	0.50	0.44	0.48	0.46	0.43	0.34	0.31	0.89
Mean Expenditure:	$1,442.12	lower bound	0.25	0.22	0.26	0.20	0.15	0.10	0.07	0.50
		upper bound	0.76	0.67	0.70	0.73	0.70	0.58	0.55	1.28
		Includes zero:	no	no	no	no	no	no	no	no

[+]Statistics shown are for a random sample of 29,010 drawn from the full sample of 37,490 employees in couples.

Couple controls: employee dummy (when spouses included), male dummy, plan (saturated), census region (saturated), salary dummy (vs. hourly), count family born 1944 to 1953, count family born 1954 to 1963, count family born 1964 to 1973, count family born 1974 to 1983, count family born 1984 to 1993.

Family controls: couple controls, spouse on policy dummy, YOB of oldest dependent, YOB of youngest dependent, family size (saturated with 8-11 as one group), count family born 1994 to 1998, count family born 1999, count family born 2000, count family born 2001, count family born 2002, count family born 2003, count family born 2004.

Table 1.13: Robustness Tests Using Longitudinal Data

Robustness Tests Using Longitudinal Data

Continuously Enrolled 2003-2004 Employee Sample

Dependent Variable:	(Expenditure in 2004 - Expenditure in 2003)		(Expenditure in 2004)	
N= 18,743	OLS	OLS	OLS	OLS
2004 Family Injury Only (n= 1,037)	482.58	478.80	671.66	668.60
lower bound	73.00	69.82	312.87	309.75
upper bound	892.16	887.78	1030.44	1027.44
2003 Family Injury Only (n= 2,024)	-82.12	-121.01	279.74	244.23
lower bound	-383.97	-423.56	15.32	-21.23
upper bound	219.73	181.54	544.15	509.69
2004 & 2003 Family Injury (n= 295)	525.01	442.52	879.19	824.85
lower bound	-225.36	-306.51	221.88	167.64
upper bound	1275.38	1191.55	1536.51	1482.05
Controls	no	yes	no	yes

Mean dependent variable: $139.66

Continuously enrolled 2003-2003 employee sample includes all employees for whom the entire family meets the selection criteria for 2003 and 2004.

People with selected injuries in 2003 or 2004 are dropped in both years.

Controls include (2003 values): employee dummy (when applicable), spouse dummy (when applicable), male dummy, plan (saturated), census region (saturated), salary dummy (vs. hourly), spouse on policy dummy, YOB of oldest dependent, YOB of youngest dependent, family size (saturated with 8-11 as one group), count family born 1944 to 1953, count family born 1954 to 1963, count family born 1974 to 1983, count family born 1984 to 1993, count family born 1994 to 1998, count family born 1999, count family born 2000, count family born 2001, count family born 2002, count family born 2003.

Table 1.14: Extension: Prescription Drug Expenditure

Extension: Prescription Drug Expenditure

		Censored Quantile IV							
2004 Employee Sample		65	70	75	80	85	90	95	Tobit IV
Dependent Variable: Ln(Expenditure)									
N= 29,010	Year-end price	-4.4	-4.3	-4.5	-4.5	-4.5	-4.7	-4.5	-6.4
	lower bound	-9.6	-6.8	-5.5	-5.3	-5.1	-5.1	-5.3	-7.4
	upper bound	-3.3	-3.5	-3.6	-3.9	-4.2	-4.4	-4.2	-5.3
	[Elasticity]	-[2.2]	-[2.2]	-[2.3]	-[2.3]	-[2.3]	-[2.3]	-[2.3]	-[3.2]
Dependent Variable: Ln(Drug Expenditure)									
N= 29,010	Year-end price	*	-4.5	-3.9	-4.1	-3.6	-2.7	-2.6	-5.9
	lower bound	*	-6.3	-5.5	-5.0	-4.3	-3.8	-10.0	-7.0
	upper bound	*	-3.2	-3.2	-2.8	-2.6	-2.4	-2.0	-4.7
	[Elasticity]	NA	-[2.3]	-[2.0]	-[2.0]	-[1.8]	-[1.4]	-[1.3]	-[2.9]
Dependent Variable: Ln(Expenditure + Drug Expenditure)									
N= 29,010	Year-end price	-3.9	-4.1	-4.0	-3.9	-3.8	-3.7	-3.8	-5.6
	lower bound	-5.2	-4.8	-4.4	-4.5	-4.4	-4.1	-10.0	-6.5
	upper bound	-3.2	-3.1	-3.3	-3.4	-3.4	-3.2	-3.2	-4.7
	[Elasticity]	-[2.0]	-[2.0]	-[2.0]	-[2.0]	-[1.9]	-[1.9]	-[1.9]	-[2.8]

* Estimator did not converge.

Censored quantile IV results from a grid search over -10 to -2 in increments of .1.

Controls include: employee dummy (when applicable), spouse dummy (when applicable), male dummy, plan (saturated), census region (saturated), salary dummy (vs. hourly), spouse on policy dummy, YOB of oldest dependent, YOB of youngest dependent, family size (saturated with 8-11 as one group), count family born 1944 to 1953, count family born 1954 to 1963, count family born 1974 to 1983, count family born 1984 to 1993, count family born 1994 to 1998, count family born 1999, count family born 2000, count family born 2001, count family born 2002, count family born 2003, count family born 2004 (when applicable).

Chapter 2

Censored Quantile Instrumental Variables Regression via Control Functions (by Victor Chernozhukov and Amanda Ellen Kowalski)

2.1 Introduction

Censoring in the dependent variable can introduce bias and inconsistency in traditional mean and quantile estimators because it induces a correlation between independent variables and the error term. Several mean estimators such as Tobit IV have been developed to produce consistent estimates in models with censored dependent variables, but they often require strong parametric assumptions. Through a generalization of a traditional quantile estimator, Powell (1986) developed a semi-parametric way to achieve consistent quantile estimates on censored data. However, the Powell

estimator has proven computationally difficult to execute, and it does not incorporate endogeneity. In this paper, we develop a new censored quantile instrumental variables (CQIV) estimator that handles censoring nonparametrically in the tradition of Powell (1986) and generalizes standard censored quantile regression (CQR) methods to incorporate endogeneity. Furthermore, we set forth a CQIV computational algorithm that is simple to execute using standard statistical software. The results of a Monte-Carlo simulation exercise demonstrate that the performance of CQIV is comparable to that of Tobit IV in data generated to satisfy the Tobit IV assumptions.

The CQIV computational algorithm that we develop here uses a control term approach to control for endogeneity in the structural equation. Newey, Powell, and Vella (1999) describe the use of the control function approach in triangular simultaneous equations models with constant coefficients. Lee (2007) sets forth an estimation strategy using a control function approach in a model with quantile structural and first stage equations. Our model differs from his in that our model has a censored dependent variable, and our first stage equation does not need to be additive. An application of our CQIV method to the estimation of the price elasticity of expenditure on medical care appears in Kowalski (2008).

In Section 2, we present the CQIV model and estimation methods. In Section 3, we describe the associated computational algorithm and present results from a Monte-Carlo simulation exercise. In Section 4, we provide conclusions and discuss potential empirical applications of CQIV.

2.2 Censored Quantile Instrumental Variables Regression

2.2.1 The Model

The general stochastic model we consider is the following "triangular" system of quantile equations:

$$Y = \max(Y^*, C) \tag{2.1}$$

$$Y^* = Q_{Y^*}(U|D, W, V) \tag{2.2}$$

$$D = Q_D(V|W, Z). \tag{2.3}$$

In this system, Y^* is the latent response variable, Y is obtained by censoring Y^* above at the censoring variable C, D is the endogenous variable, W is a vector of regressors, possibly containing C, V is a latent unobserved regressor, and Z is a vector of instruments. Further, $Q_{Y^*}(\cdot|D, W, V)$ is the conditional quantile function of Y^* given the endogenous variable D, regressors W, and unobserved regressor V; and $Q_D(\cdot|W, Z)$ is the conditional quantile function of the endogenous variable D given regressors W and instruments Z. Here, U is a Skorohod disturbance that satisfies the independence assumption

$$U \sim U(0, 1)|D, W, C, V, \tag{2.4}$$

and V is a Skorohod disturbance such that

$$V \sim U(0, 1)|W, C, Z. \tag{2.5}$$

In the last two equations, we make the assumption that the censoring variable C is independent of the disturbances U and V. This variable can, in principle, be related

to W. Indeed, our notation allows us to capture possible dependence of W and C by simply treating C as a component of W.

In the model above, to recover the structural function of interest, $Q_{Y^*}(\cdot|D, W, V)$, it is important to condition on an omitted regressor V called the "control function." The instrumental equation allows us to recover this omitted regressor as a residual that explains movements in the variable D, conditional on the set of instruments and other regressors. Nonparametric triangular models for uncensored data are developed in Imbens and Newey (2002) and Chesher (2003); parametric nonlinear variants of these models are also discussed in Wooldridge (2002); linear variants of these models appear in the analysis of Hausman (1978). The model treated in this paper differs from these earlier models by explicitly treating the case of a censored response variable.

From the system of equations above, we have that

$$Y = Q_Y(U|D, W, V, C) = \max(Q_{Y^*}(U|D, W, V), C). \tag{2.6}$$

Thus, the conditional quantile function of the observed response variable Y is equal to the conditional quantile function of the latent variable Y^*, transformed by the censoring transformation function $max(\cdot, C)$.

2.2.2 Estimation

To make estimation both practical and realistic, we make a flexible semi-parametric restriction on the functional form of the structural quantile function. In particular, we assume that

$$Q_{Y*}(u|D, W, V) = \tilde{X}'\beta(u), \quad \tilde{X} = T(D, W, V) = (X, \dot{V}), \tag{2.7}$$

where $T(D, W, V)$ is a collection of continuously differentiable transformations of initial regressors D, W, V. The transformations could be, for example, polynomial, trigonometric, B-spline or other basis functions that have good approximating ability for economic problems. In this notation, we also need to distinguish the part of the vector $T(D, W, V)$ that only depends on V; we denote this part $\dot{V}$. An important property of this functional form is linearity parameters, which will lead us to a construction of a computationally efficient estimator. The resulting functional form for the conditional quantile function of the censored random variable is given by

$$Q_Y(u|D, W, V, C) = \max(\tilde{X}'\beta(u), C). \tag{2.8}$$

This is the standard functional form first derived by Powell (1984) in the exogenous context.

We then form the estimator for parameters of this function as

$$\widehat{\beta}(u) = \arg\min_{\beta \in R^k} \frac{1}{n} \sum_{i=1}^{n} [1((\dot{X}_i, \widehat{\dot{V}_i})'\hat{\gamma} > c)\rho_u(Y_i - (\tilde{X}_i, \widehat{\dot{V}_i})'\beta)], \tag{2.9}$$

where $\rho_u(x) = (u - 1(x < 0))x$ is the asymmetric absolute loss function of Koenker and Bassett (1978), and $\dot{X}$ is a vector of transformations of vector $(\tilde{X}, C)$. This estimator adapts the estimator developed in Chernozhukov and Hong (2002) to deal with endogeneity. We call the multiplier $1((\dot{X}_i, \widehat{\dot{V}_i})'\hat{\gamma} > c)$ the selector, as its purpose is to predict the subset of regressors where the probability of censoring is sufficiently low to permit using a linear – in place of a censored linear – functional form for the conditional quantile. We formally state the conditions on the selector in the next subsection. This notational formulation allows for this estimator to be computed through several steps all taking the form above. We provide necessary practical details in the next section. This estimator may also be seen as a computationally

attractive approximation to the Powell estimator applied to our case:

$$\widehat{\beta}_p(u) = \arg\min_{\beta \in R^k} \frac{1}{n} \sum_{i=1}^{n} [\rho_u(Y_i - \max((\tilde{X}_i, \widehat{V}_i)'\beta, C_i))]. \tag{2.10}$$

The control function V can be estimated in several ways. We can see that

$$V = V(D, W, Z) \equiv Q_D^{-1}(D|W, Z) = \int_0^1 1\{Q_D(v|W, Z) \leq D\} dv. \tag{2.11}$$

Take any estimator for $Q_D(v|W, Z)$ or for $Q_D^{-1}(D|W, Z)$, based on any parametric or semi-parametric functional form. Denote the resulting estimator for the control function as

$$\widehat{V} = \widehat{V}(D, W, Z) \equiv \widehat{Q}_D^{-1}(D|W, Z) = \int_0^1 1\{\widehat{Q}_D(v|W, Z) \leq D\} dv. \tag{2.12}$$

There are several examples: in the classical additive example, we have that

$$Q_D(v|W, Z) = \tilde{Z}'\delta + Q(v), \tag{2.13}$$

where Q is a quantile function, and $\tilde{Z}$ is a vector collecting transformations of W and Z, so that

$$V = Q^{-1}(D - \tilde{Z}'\delta); \tag{2.14}$$

in a non-additive example, we have that

$$Q_D(v|W, Z) = \tilde{Z}'\delta(v), \tag{2.15}$$

and

$$V = \int_0^1 1\{\tilde{Z}'\delta(v) \leq D\} dv. \tag{2.16}$$

The estimators then take the form

$$\widehat{V} = \int_0^1 1\{\tilde{Z}'\widehat{\delta}(v) \leq D\}dv. \tag{2.17}$$

Their asymptotic theory has been developed in Chernozhukov, Fernandez-Val, and Galichon (2006).

2.2.3 Regularity Conditions for Estimation

In order to estimate and make inference on $\beta(u)$ where u is the probability index of interest in $(0, 1)$, we make the following assumptions:

Condition 1 (Sampling) *We have a sample of size n of identically and independently distributed vectors (Y_i, D_i, W_i, Z_i). The distribution function of (Y_i, D_i, W_i, Z_i) has a compact support and satisfies conditions stated below.*

Condition 2 (Conditions on the Estimator of the Control Function) *We have that*

$$\widehat{\dot{V}} = \widehat{\dot{V}}(D, Z, W), \text{ where } \widehat{\dot{V}} \in \mathcal{V}, \tag{2.18}$$

where $\mathcal{V}$ is class of functions that are sufficiently smooth, in the sense that the class satisfies Pollard's entropy condition, and

$$\sqrt{n}(\widehat{\dot{V}} - \dot{V}) = B(D, Z, W)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_i + o_p(1), \tag{2.19}$$

where $S_1, ..., S_n$ are i.i.d. random vectors with finite second moments, and $B(D, Z, W)$ also has finite second moments.

Condition 3 (Conditions on the Selector) *The selection rule is equivalent to the form*

$$1(\widehat{\dot{X}}'\hat{\gamma} > c), \text{ where } \hat{\gamma} \to_p \gamma. \tag{2.20}$$

where for some $b > 0$

$$1(\dot{X}'\gamma > c) \leq 1(Pr[Y = C|X, Z, V] < u + b). \tag{2.21}$$

The selector must also be nontrivial in the sense that

$$1(\dot{X}'\gamma > c) = 1(Pr[Y = C|X, Z, V] < u + b) \tag{2.22}$$

with positive probability.

Condition 4 (Smoothness Conditions) *(a) The conditional density* $f_Y(y|X = x)$ *is differentiable in the argument* y*, with a derivative that is uniformly bounded in* y *and* x *varying over the support of* (Y, X)*. (b) the mapping* $(\alpha, V') \mapsto P((\tilde{X}_i, V')'\alpha > v)$ *is Lipschitz in* α *and in* V'*, for* α *in an open neighborhood of* γ_0 *and* V' *in* $\mathcal{V}'$*.*

Condition 5 (Design Conditions) *The matrices* $\dot{J} \equiv Ef_Y(X_i'\beta(u)|X_i)X_iX_i'1[\dot{X}_i'\gamma > c]$ *and* $\dot{\Lambda} \equiv Var[\{(u-1(Y_i < X_i'\beta(u)))X_i + E[f_Y(X_i'\beta(u)|X_i)X_iB(X_i)]S_i\} \cdot 1(\dot{X}_i'\gamma > c)]$ *are of full rank.*

Assumption 1 imposes standard independence conditions as well as compactness of support of the data variables. We can relax the compactness at the cost of more complicated notation and proofs. Assumption 2 imposes a high-level condition on the estimator of the control function. This condition is plausible, and it holds for the parametric estimators of the control function in the additive set-up, and also for semi-parametric estimators of the control function in the non-additive set-up using quantile regression (see Chernozhukov, Fernandez-Val, and Galichon, 2006). Assumption 3 imposes a high-level condition on the estimator of the selector function. This condition is plausible, and it holds for a variety of selectors based on the initial estimates of the censoring probability and estimates of the conditional quantile functions. Assumption 4 imposes some smoothness assumptions on the distribution of Y and on

the distribution of the linear index entering the selector function. This assumption is more or less standard, and it also appears to be plausible. Assumption 4 imposes a design condition that allows us to identify the parameters of interest and also estimate them at the standard $\sqrt{n}$ rate.

2.2.4 Main Theorem

We obtain the following result that states that the CQIV estimator is consistent, converges to the true parameter at $\sqrt{n}$ rate, and is normally distributed in large samples.

Theorem 6 *Under the stated assumptions*

$$\sqrt{n}(\hat{\beta}(u) - \beta(u)) \xrightarrow{d} N(0, J^{-1}(u)\Lambda_0(u)J^{-1}(u)) \tag{2.23}$$

See Appendix A for a proof. We can estimate the variance-covariance matrix using standard methods and carry out analytical inference based on the normal distribution. In practice, we find it more practical to use bootstrap and subsampling to perform inference.

2.3 Implementation Details and Monte-Carlo Illustrations

We begin our CQIV computational algorithm with Step 0 to facilitate comparison with the Chernozhukov and Hong (2002) 3-Step CQR algorithm, which we follow closely. For each desired quantile u,

0. Obtain a prediction of the control term, $\widehat{\dot{V}}$ (and its transformations). A simple additive strategy is to obtain $\widehat{\dot{V}}$ by predicting the OLS residuals from the first

stage regression of D on W and Z. If desired, higher order functions of the predicted residuals can be included in $\widetilde{X}$. We mentioned non-additive strategies in the previous section.

1. Select a subset of observations, J_o, which are not likely to be censored using a parametric probability model:

$$1(Y_i > C_i) = p(\widetilde{X}_i'\gamma) + \varepsilon_i \tag{2.24}$$

where $1(Y_i > C_i)$ takes on a value of 1 if the observation is not censored and takes on a value of zero otherwise. Note that the control term, $\widehat{\dot{V}}$, is included in $\widetilde{X}$. In practice, a probit, logit, or any other model that fits the data well can be used. Select the sample J_0 according to the following criterion:

$$J_0 = \{i : p(\widetilde{X}_i'\widehat{\gamma}) > 1 - u + c\}. \tag{2.25}$$

In practice, it is advisable to choose c such that a constant fraction of observations satisfying $p(\widetilde{X}_i'\widehat{\gamma}) > 1 - u$ are excluded from J_0 for each quantile. To do so, set c so that $(1 - u - c)$ is the $q0$th quantile of $p(\widetilde{X}_i'\widehat{\gamma})$ such that $p(\widetilde{X}_i'\widehat{\gamma}) > 1 - u$, where $q0$ is a percentage (10 worked well in our simulation). The empirical value of c and the percentage of observations retained in J_0 can be computed as simple robustness diagnostic test at each quantile.

2. Estimate the standard quantile regression on the sample J_o:

$$\widehat{\beta}(u) \text{ minimizes } \sum_{J_0} \rho_u(Y_i - \widetilde{X}_i'\beta(u)). \tag{2.26}$$

and, using the predicted values, select another subset of observations, J_1, from the full sample according to the following criterion:

$$J_1 = \{i : \widetilde{X}_i' \widehat{\beta(u)} > C_i + \delta_n\}. \tag{2.27}$$

In practice, it is advisable to choose δ_n such that a constant fraction of observations satisfying $\widetilde{X}_i' \widehat{\beta(u)} > C_i$ are excluded from J_1 for each quantile. To do so, set $(C_i + \delta_n)$ to be the $q1$th quantile of $\widetilde{X}_i' \widehat{\beta(u)}$ such that $\widetilde{X}_i' \widehat{\beta(u)} > C_i$, where $q1$ is a percentage less than $q0$ (3 worked well in our simulation). In practice, it should be true that $J_0 \subset J_1$. If this is not the case, it is advisable to alter $q0$, $q1$, or the regression models. At each quantile, the empirical value of δ_n, the percentage of observations from the full sample retained in J_1, the percentage of observations from J_0 retained in J_1, and the number of observations in J_1 but not in J_0 can be computed as simple robustness diagnostic tests. Coefficient estimates $\widehat{\beta(u)}$ obtained in this step are consistent but will be inefficient relative to estimates obtained in the subsequent step.

3. Estimate the standard quantile regression on the sample J_1. Formally, replace J_0 with J_1 in (2.26). The new estimates, $\widehat{\beta(u)}$, are the 3-Step CQIV coefficient estimates.

4. (Optional) With results from the previous step, select a new sample J_2. Repeat this and the previous step as many times as desired.

Beginning with Step 2, each successive step of the algorithm should yield estimates that come closer to minimizing the Powell objective function. As a simple robustness diagnostic test, we recommend computing the value of the Powell objective function using the full sample and the estimated coefficients after each step, starting with Step 2. This diagnostic test is computationally straightforward because computing the value of the objective function for a given set of values is much simpler than maximizing it. In practice, this diagnostic test can be used to determine when to stop the CQIV algorithm for each quantile. If the value of the Powell objective

function increases from Step s to Step $s+1$ for $s \geq 2$, estimates from step s can be retained as the coefficient estimates.

We recommend obtaining confidence intervals through a bootstrapping procedure, though analytical formulas can also be used. If the estimation runs quickly on the desired sample, it is straightforward to draw $R \geq 100$ bootstrap samples with replacement and run each bootstrapped sample through all steps of the algorithm. A confidence interval for each coefficient estimate can be formed from the .05 and .95 quantiles of the vector of point estimates obtained for each coefficient.

2.3.1 Monte-Carlo

The goal of the following Monte-Carlo simulation is to quantify the empirical performance of CQIV relative to Tobit IV. For our simulation, we generate data according to a model that satisfies the Tobit IV assumptions. When the Tobit IV assumptions are satisfied, Tobit IV is consistent and efficient, and CQIV at each quantile is consistent but inefficient. Thus, estimates from both models satisfy the criteria for a Hausman (1978) specification test, in which the null hypothesis is that the Tobit IV assumptions are satisfied. Since the Tobit IV assumptions are satisfied in our simulated data, a comparison of Tobit IV coefficients to CQIV coefficients at each quantile quantifies the relative efficiency of CQIV in a model where Tobit IV can be expected to perform as well as possible.

A model with constant coefficients facilitates comparison between the conditional mean estimate of Tobit IV and the conditional quantile estimates of CQIV. Specifically, for each of R Monte-Carlo repetitions, we generate N observations according to the following model:

$$D_i = \alpha_0 + \alpha_1 Z_i + \alpha_2 W_i + \Phi^{-1}(V_i), \quad V_i \backsim U(0,1) \tag{2.28}$$

$$Y_i^* = \gamma_0 + \gamma_1 D_i + \gamma_2 W_i + \Phi^{-1}(U_i'), \quad U_i' \backsim U(0,1) \tag{2.29}$$

where $(\Phi^{-1}(V_i), \Phi^{-1}(U_i'))$ is distributed multivariate normal with mean zero and covariance matrix

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}. \tag{2.30}$$

Though we can observe Y_i^* in the simulated data, in the censored data, we instead observe

$$\max(y_i, C_i) = \gamma_0 + \gamma_1 D_i + \gamma_2 W_i + \Phi^{-1}(U_i'). \tag{2.31}$$

From properties of the multivariate normal distribution, we know $\Phi^{-1}(U_i') = \rho\Phi^{-1}(V_i) + \Phi^{-1}(U_i)$, where $\Phi^{-1}(U_i)$ is distributed $N(0, 1-\rho^2)$. Using this expression, we can combine (2.28) and (2.31) for an alternative formulation of the censored model in which the control term, V, is included in the structural equation:

$$\max(Y_i, C_i) = \gamma_0 + \gamma_1 D_i + \gamma_2 W_i + \rho\Phi^{-1}(V_i) + \Phi^{-1}(U_i) \tag{2.32}$$

This formulation is useful because it indicates that when we include the control term in the structural equation, its true coefficient is ρ.

In our simulated data, we create extreme endogeneity by setting $\rho = .9$. For simplicity, we set $\alpha_0 = \gamma_0 = 0$, and $\gamma_0 = \gamma_1 = 1$. To generate the data, we draw the disturbances $\Phi^{-1}(V_i)$ and $\Phi^{-1}(U_i')$ from a multivariate normal distribution with mean zero and covariance matrix (2.30). We draw Z_i from a standard normal distribution, and we generate W_i to be a log-normal random variable that is censored from the right

at its 95th percentile, r. Formally, we draw $\widetilde{W_i}$ from a standard normal distribution. We then calculate $r = Q_W(.95)$, which differs across replication samples. Next, we set $W_i = \min(e^{\widetilde{W_i}}, r)$. For comparative purposes, we set the amount of censoring in the dependent variable to be comparable to that in Kowalski (2008). Specifically, we set $C_i = C = Q_Y(.38)$ in each replication sample. In results not reported here, we set $N = 30,000$ for comparison to Kowalski (2008). Here, we report results with $N = 1,000$ to demonstrate CQIV performance in a more conventional sample size.

For each of $R = 100$ replications, on the uncensored data, we compute traditional IV estimates and IV estimates using the control function approach. On the censored data, we compute traditional Tobit IV estimates and Tobit estimates using the control function approach. Also on the uncensored data, we compute CQIV estimates at the .05 to .95 quantiles in increments of .10.

In Table 2.1, we report the median bias and interquartile range (IQR) of the IV and Tobit estimates. Bias on the coefficients on D and W is computed as $(1 - estimate)$. Bias on the estimated control term, $\widehat{V}$, the predicted residual from the first stage regression of D on Z, W, and a constant, is computed as $(.9 - estimate)$. The IV and Control IV results in the first two rows of each section are numerically identical, given the equivalence of the traditional approach and the control function approach in a linear model. The median bias and IQR of these estimates provide a bound on median bias and IQR absent censoring. In the censored data, the next set of results demonstrates that Control Tobit IV represents a substantial improvement over Tobit IV in terms of median bias and IQR. This comparison illustrates the value of the control function approach in a nonlinear model.

In Table 2.2, we present the CQIV robustness diagnostic tests suggested above. In our estimates, we used a probit model in the first step, and we set $q1 = 10$ and $q2 = 3$. In empirical practice, we do not necessarily recommend reporting the diagnostics in Table 2.2, but we have included them here for expositional purposes. In the top

section of the table, we present diagnostics computed after CQIV Step 1. At the 0.05 quantile, observations are retained in J_0 if their predicted probability of being uncensored exceeds $1 - u + c = 1 - .05 + .0445 = .9945$. Empirically, this leaves 47.0% of the total sample in J_0 in the median replication sample. In all statistics, the variation across replication samples appears small. However, as intended by the algorithm, there is meaningful variation across the estimated quantiles. As the estimated quantile increases, the percentage of observations retained in J_0 increases. From these diagnostics, the CQIV estimator appears well-behaved in the sense that the percentage of observations retained in J_0 is never very close to 0 or 100.

In the second section of Table 2.2, we present robustness test diagnostics computed after CQIV Step 2. Observations are retained in J_1 if the predicted Y_i exceeds $C_i + \delta_n$, where the median value of C_i, as shown in the table, is 1.575, and the median value of δ_n at the .05 quantile is 1.694. As desired, at each quantile, the percentage of observations retained in J_1 is smaller than the percentage of observations with predicted values above C_i but larger than the percentage of observations retained in J_0. As shown in sections of the table labeled "Percent J_0 in J_1" and "Count J_1 not in J_0" J_0 is almost a proper subset of J_1.

In the last section of Table 2.2, we report the value of the Powell objective function obtained after CQIV Step 2 and CQIV Step 3. As shown at the far right of the last section, on average, across the estimated quantiles, the final CQIV step represents an improvement in the objective function in 36-51% of replication samples. In our CQIV simulation results, we report the results from the second and third steps separately for comparative purposes. In empirical practice, we recommend selecting results from the second or third step based on the value of the objective function.

In Table 2.3, we report the median bias and IQR of the CQIV estimates from Step 3 and Step 2. Roughly, the absolute value of the median bias and IQR are highest at the extreme quantiles. It is notable that even with 38% censoring, we are able

to attain estimates at low quantiles. As shown in the penultimate set of columns, CQIV Step 3 estimates of the coefficients on D and W have smaller absolute median bias than comparable CQIV Step 2 estimates in 30% of the estimated quantiles, with no clear pattern across the quantiles. In results with $N = 30,000$, the gains to CQIV Step 3 relative to CQIV Step 2 are larger. In terms of interquartile range, CQIV Step 2 almost always out-performs CQIV Step 3, illustrating the potential disadvantage of increasing the number of steps in the CQIV algorithm.

The last two columns of Table 2.3 present the most important results of the Monte-Carlo Simulation, the comparison between Control Tobit IV estimates from Table 2.1 and the CQIV estimates in Table 2.3. In terms of median bias, CQIV Step 3 estimates of the coefficient on D out-perform Tobit IV estimates in 90% of estimated quantiles. In terms of IQR, Tobit IV estimates almost always out-perform CQIV estimates, but comparison of the actual IQR values in Table 2.1 and Table 2.3 shows that the CQIV IQR has the same order of magnitude as the Tobit IV IQR. Given that the simulated data satisfy the Tobit IV assumptions, the results of this simulation should give a lower bound of CQIV performance relative to Tobit IV. Since Tobit IV requires several parametric assumptions, the advantages of CQIV are likely to be large relative to Tobit IV in applied work.

2.4 Conclusion

In this paper, we develop a new censored quantile instrumental variables estimator, and we demonstrate its computation and finite sample performance using a Monte-Carlo simulation. Censoring and endogeneity abound in empirical work, so CQIV should be readily applicable. Kowalski (2008) uses CQIV to estimate the price elasticity of expenditure on medical care across the quantiles of the expenditure distribution, where censoring arises because of the decision to consume zero care, and

endogeneity arises because marginal prices explicitly depend on expenditure. In another joint paper, we use CQIV in a duration model context. Specifically, we re-estimate the McClellan, McNeil, and Newhouse (1994) model of the effect of cardiac catheterization on elderly mortality, which uses differential distance to a cardiac catheterization facility as an instrument. In addition to allowing for the censored mortality of patients that are still alive, the CQIV estimator allows us to examine the effect of cardiac catheterization across the quantiles of the mortality distribution. Since CQIV can be implemented using standard statistical software, it should prove useful to applied researchers.

Bibliography

[1] Chernozhukov, Victor, Fernandez-Val, Ivan, and Galichon, Alfred. "Quantile and Probability Curves without Crossing." 2006. MIT Working Paper.

[2] Chernozhukov, Victor, and Hansen, Christian. "Instrumental variable quantile regression: A robust inference approach." Journal of Econometrics. January 2008. 142(1) pp.379-398.

[3] Chernozhukov, Victor, and Jong, Hong. "Three-Step Quantile Regression and Extramarital Affairs." Journal of The American Statistical Association. September 2002, 97(459). pp. 872-882.

[4] Chesher, A. "Identification in Nonseparable Models." Econometrica, 2003, 71(5), pp. 1405-1441.

[5] Hausman, Jerry A.."Specification Tests in Econometrics." Econometrica, 1978, 46(6), pp. 1251-71.

[6] Imbens, Guido W., and Newey, Whitney K.. "Identification and Estimation of Triangular Simultaneous Equations Models without Additivity." NBER Technical Working Paper 285.

[7] Kowalski, Amanda E. "Censored Quantile Instrumental Variable Estimates of the Price Elasticity of Expenditure on Medical Care." Mimeo. 2008.

[8] Koenker, Roger, and Bassett, Gilbert Jr. "Regression Quantiles." Econometrica, 1978, 46(1), pp. 33-50.

[9] Lee, Sokbae. "Endogeneity in quantile regression models: A control function approach." Journal of Econometrics. 2007. 141, pp. 1131-1158.

[10] McClellan, M., McNeil, B.J., and Newhouse, J.P. "Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables." Journal of the American Medical Association. 1994. 272(11). pp. 859-866.

[11] Newey, Whitney K., Powell, James L., Vella, Francis. "Nonparametric Estimation of Triangular Simultaneous Equations Models." Econometrica. 1999. 67(3), 565-603.

[12] Powell, James L. "Censored Regression Quantiles." Journal of Econometrics, 1986. 23. pp-143-155.

[13] Powell, James L. "Least absolute deviations estimation for the censored regression model." Journal of Econometrics, 1984, 25(3), pp. 303-325.

[14] Wooldridge, Jeffrey M. Econometric Analysis of Cross Section and Panel Data. MIT Press. Cambridge, MA. 2002.

Table 2.1: Median Bias and IQR of IV and Tobit IV Estimators

Median Bias and IQR of IV and Tobit IV Estimators

	Estimator	Censored Y	Median Bias	IQR
Enodgenous Variable (p)	IV	no	0.0043203	0.0454494
	Control IV	no	0.0043203	0.0454494
	Tobit IV	yes	0.0332017	0.3733709
	Control Tobit IV	yes	0.0081629	0.0533828
Covariate (x)	IV	no	-0.0009606	0.0558725
	Control IV	no	-0.0009606	0.0558725
	Tobit IV	yes	-0.0160931	0.0656968
	Control Tobit IV	yes	-0.0007842	0.0527166
Control Term (vhat)	IV	no	NA	NA
	Control IV	no	-0.0021902	0.0444503
	Tobit IV	yes	NA	NA
	Control Tobit IV	yes	-0.0021902	0.0584879

N=1,000
Replications=100

Table 2.2: CQIV Optimization Statistics Across Monte Carlo Replications

CQIV Optimization Statistics Across Monte Carlo Replications

CQIV Step 1

	c			Percent J0		
Quantile	Median	Min	Max	Median	Min	Max
0.05	0.0445	0.0396	0.0478	47.0	45.1	50.3
0.15	0.1266	0.1024	0.1411	50.2	48.7	52.3
0.25	0.2038	0.1640	0.2336	52.2	50.9	53.4
0.35	0.2675	0.2353	0.3226	53.8	52.9	54.4
0.45	0.3248	0.2765	0.4082	55.1	54.3	56.1
0.55	0.3769	0.2760	0.4962	56.5	55.7	57.3
0.65	0.4145	0.3105	0.5828	57.8	56.3	58.6
0.75	0.4320	0.2575	0.6026	59.4	58.1	61.2
0.85	0.3990	0.2803	0.5566	61.3	59.7	63.0
0.95	0.3012	0.2071	0.4588	64.3	62.3	66.4

CQIV Step 2

	Deltan			Percent J1			Percent Predicted Above C		
Quantile	Median	Min	Max	Median	Min	Max	Median	Min	Max
0.05	1.6942	1.4888	1.9101	50.9	47.3	53.8	52.5	48.8	55.5
0.15	1.7066	1.4541	1.9218	54.2	52.2	56.3	55.9	53.9	58.1
0.25	1.6927	1.5111	1.9317	56.0	54.6	58.2	57.8	56.3	60.0
0.35	1.6940	1.5115	1.9402	57.9	56.1	59.8	59.7	57.9	61.7
0.45	1.7061	1.5343	1.9061	59.3	57.6	61.2	61.2	59.4	63.1
0.55	1.7178	1.5086	1.9763	61.1	58.6	62.3	63.0	60.5	64.3
0.65	1.7162	1.4424	1.9716	62.3	59.9	63.7	64.3	61.8	65.7
0.75	1.6988	1.4070	2.0034	64.0	61.3	66.1	66.0	63.2	68.2
0.85	1.7106	1.4774	1.9687	66.0	62.9	68.4	68.1	64.9	70.6
0.95	1.7457	1.4790	2.0917	69.4	66.7	72.4	71.6	68.8	74.7

	C			Percent J0 in J1			Count in J1 not in J0		
Quantile	Median	Min	Max	Median	Min	Max	Median	Min	Max
0.05	1.5753	1.3161	1.9234	100	98.7	100	36	2	77
0.15	1.5753	1.3161	1.9234	100	99.8	100	39	21	63
0.25	1.5753	1.3161	1.9234	100	100.0	100	39.5	24	59
0.35	1.5753	1.3161	1.9234	100	99.8	100	42	25	60
0.45	1.5753	1.3161	1.9234	100	100.0	100	42	23	60
0.55	1.5753	1.3161	1.9234	100	99.6	100	44	23	62
0.65	1.5753	1.3161	1.9234	100	99.8	100	45	25	60
0.75	1.5753	1.3161	1.9234	100	100.0	100	46	28	67
0.85	1.5753	1.3161	1.9234	100	99.8	100	47	19	70
0.95	1.5753	1.3161	1.9234	100	100.0	100	50	21	79

Comparison of Objective Functions

	Objective Step 3			Objective Step 2			Objective Step 3<Objective Step 2	
Quantile	Median	Min	Max	Median	Min	Max	Median	Mean
0.05	4999.2	4542.6	5676.9	5058.9	4544.1	5676.9	0	0.44
0.15	12149.9	10771.4	13387.1	12153.1	11074.9	13555.3	0	0.45
0.25	17174.0	15224.2	20048.9	17302.6	15603.5	20039.8	1	0.51
0.35	20517.6	17544.4	24204.7	20639.9	18449.6	24270.9	0	0.39
0.45	22304.0	19531.0	26483.2	22266.7	19668.4	26629.4	0	0.49
0.55	22399.1	19605.3	27112.7	22493.3	19327.3	27431.3	0	0.44
0.65	20336.7	16101.3	26245.7	20487.1	15402.0	26502.1	0	0.46
0.75	15684.5	10436.3	23983.3	15711.9	10984.3	23995.9	0	0.41
0.85	6822.4	-1123.4	15435.6	6744.1	165.9	15435.6	0.5	0.50
0.95	-14979.2	-23792.3	-6720.3	-15015.9	-23192.6	-6720.3	0	0.36

N=1,000
Replications=100

Table 2.3: Median Bias and IQR of CQIV Estimator

Median Bias and IQR of CQIV Estimator

		CQIV Step 3		CQIV Step 2		|CQIV 3|<|CQIV 2|		|CQIV 3|<|Control Tobit IV|	
	Quantile	Median Bias	IQR	Median Bias	IQR	Median Bias	IQR	Median Bias	IQR
Enodgenous Variable (p)	0.05	0.0069862	0.0752635	0.0130589	0.0721436	FALSE	TRUE	TRUE	FALSE
	0.15	-0.0018179	0.0663269	0.0027844	0.0610748	FALSE	TRUE	TRUE	FALSE
	0.25	0.0002347	0.0509720	0.0059425	0.0574622	FALSE	FALSE	TRUE	TRUE
	0.35	0.0001040	0.0498580	0.0000295	0.0491355	TRUE	TRUE	TRUE	TRUE
	0.45	0.0013962	0.0555833	0.0025396	0.0578727	FALSE	FALSE	TRUE	FALSE
	0.55	-0.0021000	0.0556921	0.0049081	0.0565318	FALSE	FALSE	TRUE	FALSE
	0.65	0.0034332	0.0509731	0.0021315	0.0536874	TRUE	FALSE	TRUE	TRUE
	0.75	0.0029098	0.0612026	0.0078800	0.0634939	FALSE	FALSE	TRUE	FALSE
	0.85	0.0140356	0.0627072	0.0114948	0.0654458	TRUE	FALSE	FALSE	FALSE
	0.95	0.0044383	0.0662451	0.0065193	0.0607038	FALSE	TRUE	TRUE	FALSE
Covariate (x)	0.05	-0.0046611	0.0818740	-0.0056756	0.0812227	FALSE	TRUE	FALSE	FALSE
	0.15	0.0015344	0.0625166	-0.0055931	0.0645390	FALSE	FALSE	FALSE	FALSE
	0.25	-0.0036421	0.0501522	-0.0041385	0.0573029	FALSE	FALSE	FALSE	TRUE
	0.35	-0.0019922	0.0550888	0.0013665	0.0565156	TRUE	FALSE	FALSE	FALSE
	0.45	0.0006542	0.0588022	-0.0016363	0.0545364	FALSE	TRUE	TRUE	FALSE
	0.55	-0.0029429	0.0605592	-0.0035919	0.0638523	FALSE	FALSE	FALSE	FALSE
	0.65	0.0081584	0.0685686	0.0102680	0.0667563	FALSE	TRUE	FALSE	FALSE
	0.75	0.0068385	0.0748311	0.0064830	0.0772778	TRUE	FALSE	FALSE	FALSE
	0.85	0.0033740	0.0723600	0.0009685	0.0732648	TRUE	FALSE	FALSE	FALSE
	0.95	-0.0080097	0.0868069	-0.0087110	0.0876752	FALSE	FALSE	FALSE	FALSE
Control Term (vhat)	0.05	-0.0000815	0.0850520	-0.0083427	0.0763623	FALSE	TRUE	TRUE	FALSE
	0.15	0.0045069	0.0741163	-0.0033683	0.0736479	TRUE	TRUE	FALSE	FALSE
	0.25	0.0037275	0.0646972	-0.0042917	0.0684041	FALSE	FALSE	FALSE	FALSE
	0.35	0.0016332	0.0618684	-0.0049022	0.0605122	FALSE	TRUE	TRUE	FALSE
	0.45	-0.0072254	0.0607281	-0.0059266	0.0578150	TRUE	TRUE	FALSE	FALSE
	0.55	-0.0023095	0.0617227	-0.0033601	0.0627773	FALSE	FALSE	FALSE	FALSE
	0.65	0.0004744	0.0691199	-0.0027954	0.0693297	FALSE	FALSE	TRUE	FALSE
	0.75	-0.0037466	0.0650522	-0.0035111	0.0663501	TRUE	FALSE	FALSE	FALSE
	0.85	-0.0035439	0.0685245	-0.0063274	0.0668902	FALSE	TRUE	FALSE	FALSE
	0.95	-0.0029650	0.0796339	0.0000105	0.0749753	TRUE	TRUE	FALSE	FALSE

N=1,000
Replications=100

A Proof of Theorem 1.

Below, *const* and K are generic positive constants. C_i denotes the censoring point.

Step 1. The rescaled statistic $Z_n = \sqrt{n}(\hat{\beta}(u) - \beta(u))$ minimizes

$$Q_n(z, \hat{\gamma}, \widehat{V}) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_{in}(z) 1[(\tilde{X}_i, \widehat{\dot{V}}_i)'\hat{\gamma} > c], \quad \text{where} \tag{2.33}$$

$V_{in}(z, \widehat{V}) \equiv \sqrt{n}[\rho_u(\epsilon_i - (\tilde{X}_i, \widehat{\dot{V}}_i)'z/\sqrt{n}) - \rho_u(\epsilon_i)]$ and $\epsilon_i \equiv Y_i - X_i'\beta(u)$. The claim is that for any finite collection of points $z_j, j \leq l$

$$\left(Q_n(z_j, \hat{\gamma}, \widehat{V}), \quad j \leq l \right) \xrightarrow{d} \left(Q_\infty(z_j), \quad j \leq l \right), \tag{2.34}$$

where

$$\begin{aligned} & Q_\infty(z) \equiv \dot{W}'z + \tfrac{1}{2} z' \dot{J} z \\ & W \stackrel{d}{=} N(0, \dot{\Lambda}) \\ & \dot{J} \equiv E f_Y(X_i'\beta(u)|X_i) X_i X_i' 1[\dot{X}_i'\gamma > c], \\ \equiv \; & Var[\{(u - 1(Y_i < X_i'\beta(u)))X_i + E[f_Y(X_i'\beta(u)|X_i)X_i B(X_i)]S_i\} \cdot 1(\dot{X}_i'\gamma > c)] \end{aligned}$$

This claim above follows immediately from the standard CLT and LLN and some standard calculations applied to the first order approximation

$$Q_n(z, \hat{\gamma}, \hat{V}) = Q_n(z, \gamma, V) + z' E[f_Y(X_i'\beta(u)|X_i)X_i B(X_i)] \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_i + o_p(1), \tag{2.35}$$

which is obtained in the Step 2 below.

Matrix $\dot{J}$ is invertible by assumption. Since functions Q_n and Q_∞ are convex, finite, and continuous in z, and since function Q_∞ is uniquely minimized at random

vector $-\dot{J}^{-1}\dot{W} = O_p(1)$, (2.34) implies

$$Z_n \xrightarrow{d} -\dot{J}^{-1}\dot{W} \tag{2.36}$$

by the convexity theorem (e.g. Pollard, 1989).

Step 2. For any fixed z, the empirical process

$$\{Q_n(z, \gamma', V') - EQ_n(z, \gamma', V'), \gamma' \in \mathcal{G}, V' \in \mathcal{V}\} \tag{2.37}$$

is stochastically equicontinuous in γ, where $\mathcal{G} \equiv \{\gamma : |\gamma - \gamma_0| \leq \delta\}$ and $\delta > 0$ is small. Indeed, let

$$\mathcal{F} = \left\{(W, Z, D) \mapsto 1[(\tilde{X}', [V'(W, Z, D)])\gamma > c], \quad \gamma \in \mathcal{G}, V' \in \mathcal{V}\right\} \tag{2.38}$$

and

$$\mathcal{G}_n = \{(W, Z, D, \epsilon) \mapsto \sqrt{n}[\rho_u(\epsilon - (\tilde{X}', V'(W, Z, D))z/\sqrt{n}) - \rho_u(\epsilon)], V' \in \mathcal{V}]. \tag{2.39}$$

and, finally,

$$\mathcal{H}_n = \mathcal{F} \times \mathcal{G}_n. \tag{2.40}$$

By the boundedness assumptions, $\mathcal{H}_n$ has a constant envelope that is bounded. The class of functions $\mathcal{G}_n$ is a uniformly Liphitz transformation of $\mathcal{V}$. Using this fact it is not difficult to show that the bracketing integral for $\mathcal{H}_n$ satisfies

$$J_{[]}(\delta_n, \mathcal{H}_n, L_2(P)) \searrow 0, \text{ as } \delta_n \searrow 0. \tag{2.41}$$

Indeed, the $L_2(P)$ pseudo-metric on $\mathcal{H}_n$ is equivalent to the following pseudo-metric on $\mathcal{G} \times \mathcal{F}$. Let $h_1 \in H_n$ be defined by pair γ_1, V_1 and $h_2 \in \mathcal{H}_n$ be defined by pair

γ_2, V_2, then we define the pseudo-metric on $\mathcal{G} \times \mathcal{F}$ as

$$\begin{aligned}
&\rho(\gamma_1, \gamma_2, V_1, V_2) \equiv \sup_{n \geq 1} \quad \sqrt{E|h_1 - h_2|^2} \\
\lesssim \;& \sqrt{\|\gamma_2 - \gamma_1\|_2 + \sqrt{E|\dot{V}_1(X, D, Z) - \dot{V}_2(X, D, Z)|^2}} + \sqrt{E|\dot{V}_1(X, D, Z) - \dot{V}_2(X, D, Z)|^2} \\
\lesssim \;& \sqrt{\|\gamma_2 - \gamma_1\|_2 + \sqrt{E|V_1(X, D, Z) - V_2(X, D, Z)|^2}} + \sqrt{E|V_1(X, D, Z) - V_2(X, D, Z)|^2} \\
\lesssim \;& (\|\gamma_2 - \gamma_1\|_2)^{1/2} + (E|V_1(X, D, Z) - V_2(X, D, Z)|^2)^{1/4} + [E|V_1(X, D, Z) - V_2(X, D, Z)|^2]^{1/2}
\end{aligned}$$

where the first inequality follows by triangular inequality and some simple direct calculations, and the second from $\dot{V}$ being a uniform Lipschitz transform of V, and the last inequality is elementary. Using this inequality we can conclude that

$$J_{[]}(\delta_n, \mathcal{H}_n, L_2(P)) \lesssim J_{[]}(\delta_n, \mathcal{V}, L_2(P)) + J_{[]}(\delta_n, \mathcal{G}, L_2(P)), \tag{2.42}$$

where

$$J_{[]}(\delta_n, \mathcal{V}, L_2(P)) + J_{[]}(\delta_n, \mathcal{G}, L_2(P)) \searrow 0 \text{ as } \delta_n \searrow 0. \tag{2.43}$$

where the first terms goes to zero by assumption on the class $\mathcal{V}$; and the second term converges to zero trivially.

The stochastic equicontinuity condition implies that

$$Q_n(z, \hat{\gamma}, \hat{V}) - Q_n(z, \gamma, V) - EQ_n(z, \hat{\gamma}, \hat{V}) + EQ_n(z, \gamma, V) = o_p(1). \tag{2.44}$$

Thus, to complete the proof, it remains to examine the behavior of

$$EQ_n(z, \hat{\gamma}, \hat{V}) - EQ_n(z, \gamma, V) = EQ_n(z, \hat{\gamma}, V) - EQ_n(z, \gamma, V) + EQ_n(z, \hat{\gamma}, \hat{V}) - EQ_n(z, \hat{\gamma}, V) \tag{2.45}$$

We first show that $EQ_n(z, \hat{\gamma}, V) - EQ_n(z, \gamma, V) = o_p(1)$. We can suppress V in the

analysis. We will show that for $s_i(\gamma,\gamma_0) \equiv 1[X_i'\gamma > c] - 1[X_i'\gamma_0 > c]$:

$$EQ_n(z,\gamma) - EQ_n(z,\gamma_0)|_{\gamma=\hat{\gamma}} \equiv \sqrt{n}EV_{in}(z)s_i(\gamma,\gamma_0)|_{\gamma=\hat{\gamma}} = O_p(\hat{\gamma}-\gamma_0), \tag{2.46}$$

Write $\sqrt{n}V_{in}(z) \equiv -\sqrt{n}\big[\{u-1[\epsilon_i \le 0]\}X_i'z\big]+\sqrt{n}\big[-\eta_i(z)\{X_i'z-\epsilon_i\sqrt{n}\}\big] \equiv \sqrt{n}V_{in}'(z) + \sqrt{n}V_{in}''(z)$, where $\eta_i(z) \equiv \big[1(\epsilon_i \le 0) - 1(\epsilon_i \le X_i'z/\sqrt{n})\big]$. For γ close enough to γ_0, $X_i'\gamma > c$ implies $X_i'\beta(u) < C_i - v$, a.s. for $v > 0$ small, for all i, so that

$$E\big[\sqrt{n}V_{in}'(z)s_i(\gamma,\gamma_0)|X_i,C_i\big] = 0 \text{ uniformly in } i, \tag{2.47}$$

since $P\big[\epsilon_i \le 0|X_i,C_i,X_i\beta(u) < C_i - v\big] = u$ [if $X_i\beta(u) < C_i$, ϵ_i has u-th conditional quantile at 0] . Also $E\big[\sqrt{n}V_{in}''(z)s_i(\gamma,\gamma_0)|X_i,C_i\big] = O\big[f_u(0|X_i)z'X_iX_i'z1(X_i'\beta(u) < C_i - v)\big] \times s_i(\gamma,\gamma_0)$, uniformly in i . Therefore,

$$EE\big[\sqrt{n}V_{in}''(z)s_i(\gamma,\gamma_0)|X_i,C_i\big] = O\big(E\big[s_i(\gamma,\gamma_0)\big]\big) = O(\gamma-\gamma_0). \tag{2.48}$$

Next we consider the second term, and we can see by a direct calculation that

$$EQ_n(z,\hat{\gamma},\hat{V}) - EQ_n(z,\hat{\gamma},V) = z'E[f_Y(X_i'\beta(u)|X_i)X_iB(X_i)]\frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_i + o_p(1)$$

□

Chapter 3

Nonlinear Budget Sets and Medical Care

3.1 Introduction

Questions related to marginal prices for medical care are important for policy. First, how does medical expenditure respond to the marginal prices that consumers face? Second, what is the extent of preference heterogeneity across consumers? Third, what are the welfare consequences of price changes? In this paper, I develop a structural model and associated estimation strategy to answer these questions using medical claims data.

Estimates of the price elasticity of expenditure on medical care must address a fundamental endogeneity problem: in traditional insurance plans, marginal price is often a function of the quantity of care consumed. The existing literature has addressed the endogeneity problem with two main techniques. One way that the literature has addressed endogeneity is through randomization. The RAND health insurance experiment, which began in the 1970's, randomized consumers into plans with varying marginal prices. Using data from this experiment, Manning et al. (1987)

estimated the price elasticity of expenditure on medical care to be -0.2. However, some element of endogeneity remained because participants knew that they would face a zero marginal price once their total family expenditures reached an amount that was specified to them at the start of the experiment.

Another way that the literature has addressed endogeneity is through instrumental variables techniques. Eichner (1997, 1998) uses a clever instrumental variables strategy that relies on family cost sharing provisions to estimate the price elasticity of expenditure on medical care to be approximately -.3. In the first chapter of this dissertation, I capture the spirit of Eichner's instrument to estimate the price elasticity of expenditure at several quantiles of the expenditure distribution, and I attain estimates that are an order of magnitude larger than those in the literature. One disadvantage of instrumental variables procedures is that they can produce local estimates. Specifically, instrumental variables techniques in this setting do not allow expenditure responses to prices manipulated by the instrument to lead to even lower prices.

In this paper, I develop a third way to address the endogeneity problem in the estimation of the price elasticity of expenditure on medical care. Specifically, I develop a model based on utility theory in which expenditure and marginal price are jointly determined. As in the other two methods, I must specify a functional form for the demand function. Unlike the other two methods, my model allows me to estimate the extent of preference heterogeneity, and it allows for welfare calculations. Furthermore, it incorporates the decision to consume zero medical care directly into the model instead of treating zeros as a "censored" outcome. The estimation strategy that accompanies my model shares a tight link with the theory.

My model builds on the literature developed to estimate the elasticity of labor supply using nonlinearities in the budget set induced by progressive taxes, following Hausman (1985). Two other papers, Keeler, Newhouse, and Phelps (1977) and

Eichner (1997), discuss nonlinearities in the budget set for medical care, but my model incorporates additional elements of utility theory. In addition, I extend the nonlinear budget set literature to allow for estimation when the budget set contains more than one nonconvex kink.

In the following section, I describe the agent's general problem of utility maximization subject to a nonlinear budget set, and I describe the sources of the nonlinearities in the budget set in my setting. In Section 3, I compare my application to other applications in the nonlinear budget set literature. In Section 4, I derive the equations for the formal model and discuss regularity conditions. In Section 5, I develop estimation strategies for empirical applications of the model. In the final section, I conclude and suggest several directions for future research.

3.2 The Agent's Problem

Consider a partial equilibrium setting in which agents consume dollars of medical care, Q, and dollars of all other goods, A. For simplicity, utility is defined over Q, and A, but the model could be extended in the spirit of Phelps and Newhouse (1974) so that agents derive utility from health instead of medical care. I measure medical care in terms of dollars of expenditure on all types of medical care rather than in terms of specific services under the assumption that in most health insurance policies, the marginal price that the consumer pays for a dollar of medical care does not vary with the type of care consumed.

In traditional demand theory, expenditure is equal to quantity of units demanded multiplied by the per-unit price. In my model, I make some slight modifications to the standard notation from demand theory to incorporate expenditure on behalf of the consumer by another party, the insurer. To do so, I measure the quantity of units demanded, Q, in dollars of medical care, and I measure the per-unit price, p,

in terms of the marginal price that the *consumer* pays for a dollar of medical care. The marginal price that the *insurer* pays for a dollar of medical care is given by $(1-p)$. Since the marginal price paid by the consumer and the insurer always sums to unity, the number of units of medical care demanded by the consumer, Q, is equal to total expenditure on behalf of the consumer, $Q \times 1 = Q$. Thus, unlike in standard demand models, Q measures demand as well as total expenditure. To fit this model into traditional demand theory, I model Q as a function of p. In this framework, I define the "price elasticity of expenditure on medical care" as

$$\eta = \frac{d \ln Q}{d \ln p} \tag{3.1}$$

The traditional expression for expenditure, $Q \times p$, defines expenditure by the consumer, and it appears in the budget set. Expenditure by the consumer and the insurer, Q, is an argument of the utility function.

Formally, an agent maximizes utility subject to a budget set following the general constrained optimization problem:

$$v(y,p) = \max_{Q} U(Q, A : pQ \leq y) \tag{3.2}$$

where U is direct utility, v is indirect utility, y is virtual income as defined below, and p is the marginal price of medical care, which can be a function of Q. The Q that achieves the maximum can be expressed in terms of the demand function $Q(y,p)$. From standard utility theory, Roy's Identity relates indirect utility to demand:

$$-\frac{\partial v(y_{is}, p_s)/\partial p_s}{\partial v(y_{is}, p_s)/\partial y_{is}} = Q(y_s, p_s). \tag{3.3}$$

Therefore, given the budget set and conditions for integrability discussed below, this model requires one and only one functional form for direct utility, indirect utility,

or demand. Even though all three approaches are equivalent in terms of the model, I proceed by specifying a functional form for demand. On the grounds that reduced form work also requires a demand specification, this approach might be more palatable and transparent than the approach of specifying a utility function. This model incorporates reduced form identification as well as identification from restrictions from utility theory, which are discussed in detail below. One advantage of this approach relative to reduced form work is that the estimates can be interpreted in the context of utility theory, allowing for welfare comparisons. Once I have specified the functional form of the demand function, I move from the agent's problem to the population problem by specifying sources of unobserved heterogeneity. Before discussing the specific functional forms for demand and utility, however, I describe the budget set.

3.2.1 Nonlinear Budget Set for Medical Care

A traditional health insurance plan has three basic components: a deductible, a coinsurance rate, and a stoploss. The "deductible," is defined as the yearly amount that the beneficiary must pay before the plan covers any expenses. The percentage of expenses that the beneficiary pays after the deductible is met is known as the "coinsurance rate". The insurer pays the remaining fraction of expenses until the beneficiary meets the "stoploss," (also known as the "maximum out-of-pocket"), and the insurer pays all expenses for the rest of the year.

Figure 3-1 illustrates how these three parameters generate nonlinearities in the consumer budget set. This traditional partial equilibrium diagram relates medical care expenditure in dollars by the beneficiary and insurer, Q, to expenditure on all other goods, A. In this diagram, D denotes the deductible, C denotes the coinsurance rate, and S denotes the stoploss. The budget set has three linear segments, denoted by a, b, and c. The marginal price associated with each segment s is p_s. Specifically,

$$p_a = 1 \tag{3.4}$$

$$p_b = C \tag{3.5}$$

$$p_c = 0 \tag{3.6}$$

A central issue in nonlinear budget set models is that it is difficult to control for income because nonlinearities in the budget set create a disparity between marginal income and actual income. One approach to deal with this difficulty is to control for what Burtless and Hausman (1978) call "virtual income." Virtual income is the income that the consumer would have if each segment of the budget set were extended to the vertical axis. In the figure, actual income is denoted by Y, and virtual income on each segment is denoted by y_s. In terms of income and plan characteristics, virtual income on each segment can be expressed as follows:

$$y_a = Y \tag{3.7}$$

$$y_b = Y - (1 - C)D \tag{3.8}$$

$$y_c = Y - S \tag{3.9}$$

In practice, there are many other possible health insurance plan provisions. For example, some plans restrict care to a certain provider network, require a per-visit "copayment," and impose lifetime limits on plan payments. Furthermore, in the non-group market, premiums can vary with characteristics of the beneficiary unless prohibited by community rating laws. However, for many policies, the three parameters discussed above provide a relatively complete description of plan attributes.

3.3 Comparison to Nonlinear Budget Set Literature

The original nonlinear budget set literature estimated the labor supply elasticity using nonlinear budget sets induced by progressive taxes. Hausman (1985) provides a survey of the early literature. Some early estimates of the labor supply elasticity using nonlinear budget set models include those of Hurd (1976) Rosen (1979), and Burtless and Moffit (1985). Other applications of the nonlinear budget set model include the demand for air conditioners in Hausman (1979), the disability insurance program in Halpern and Hausman (1986), the Social Security earnings test in Friedberg (2000), and 401(k) saving in Englehardt and Kumar (2006). However, the labor supply elasticity remains the most prevalent application of the nonlinear budget set model.

To facilitate comparison of the nonlinear budget set in my application to the nonlinear budget set in the labor supply application, Figure 3-2 depicts a nonlinear budget set induced by a simple progressive tax. The after-tax wage, w, that a worker faces varies with the tax rate, t. Comparison with Figure 3-1 is slightly difficult because hours are a "bad," but both figures are drawn so that the hypothetical arrow of increasing preference points to the upper right. The labor supply application examines the effect of the after-tax wage (the slope) on hours (the horizontal axis) controlling for income (the vertical axis). Similarly, I examine the effect of the marginal price (the slope) on quantity of medical care consumed in dollars (the horizontal axis) controlling for income (the vertical axis).

Some difficulties that are present in the labor supply application are not present in my application. For example, in the labor supply application, one important issue is that several individuals work zero hours, and the potential wage for these individuals is unknown. The medical care application does not suffer from this difficulty, however. Although several individuals do not consume any medical care, the price

that they would face is observable because it is determined by the insurance policy. This transparency is possible because, unlike the wage, the price does not vary at the individual level.

One advantage of the transparency of the price schedule in the budget set for medical care is that the agent and the econometrician are likely to be aware of the agent's precise location on the budget set. Liebman and Zeckhauser (2004) hypothesize that individuals respond suboptimally to complex schedules - a phenomena that they call "schmeduling." While "schmeduling" may be very likely with respect to the complex tax rules addressed by the labor supply elasticity estimates, it is arguably less likely with respect to medical care because the price schedule is so simple. In the labor supply application, since the slope of each segment varies with the underlying marginal wage, the exact segment is often unknown to econometrician and possibly the agent.

The transparency of the price schedule in the medical care application comes at the cost of reduced underlying variation for identification. Blomquist and Newey (2004) have developed nonparametric techniques to estimate nonlinear budget set models which have been applied by Kumar (2004) and others. These nonparametric techniques would likely have less power in this application because the slopes of the segments of the budget set do not vary across individuals. Furthermore, the Blomquist and Newey (2004) approach requires that the budget set be convex.

As is apparent from the comparison of Figure 3-1 to Figure 3-2, the budget set induced by health insurance is inherently nonconvex, but the budget set induced by progressive taxes is inherently convex. Nonconvexities make utility maximization more complex because it is possible to have multiple tangencies between an indifference curve and a nonconvex budget set. While convex budget sets imply "bunching" at the kinks, nonconvex budget sets imply dispersion at the kinks. However, techniques to examine "bunching" developed by Saez (2004) and Liebman and Saez (2006)

can arguably be applied to study dispersion in this setting. Furthermore, although progressive taxes generally lead to convex budget sets, more complex budget sets, especially those that result from public assistance programs, can be nonconvex. Several papers, including Burtless and Hausman (1978), Hausman (1980), and Hausman (1981) estimate models that incorporate nonconvex segments. However, I am not aware of any other papers that incorporate two or more nonconvex segments as I do in my model.

3.4 Model Specification

Given a functional form for a demand function and a budget set, provided that regularity conditions hold, an agent's utility maximization problem is fully specified.

Proposition:

Given the following linear specification of the demand function:

$$Q(y_{is}, p_s) = \alpha + \beta y_{is} + \gamma p_s + X_i \delta \tag{3.10}$$

and the budget set:

$$A_i = y_{is} - p_s Q_i \tag{3.11}$$

where Q_i is the total amount spent on medical care on behalf of individual i, y_{is} denotes virtual income associated with segment s for individual i, p_s denotes the price per dollar of medical care associated with segment s, and X_i is a vector of covariates, and A_i is expenditure on all goods other than medical care, if the Slutsky condition $\gamma_i + \beta Q_i \leq 0$ is satisfied and $\beta \neq 0$, it follows that:

1. Indirect utility is given by:

$$v(y_{is}, p_s) = e^{-\beta p_s}[\frac{\alpha}{\beta} + y_{is} + \frac{\gamma}{\beta}p_s + \frac{X_i\delta}{\beta} + \frac{\gamma}{\beta^2}] \tag{3.12}$$

2. Utility is given by:

$$U(Q_i, A_i) = \frac{(\beta Q_i + \gamma_i)}{\beta^2}\exp[\frac{\beta(\beta A_i - Q_i + \alpha)}{\beta Q_i + \gamma_i}] \tag{3.13}$$

3. The agent has a convex indifference curve for any fixed utility level $\overline{U}$ given by:

$$A_i(Q_i, \overline{U}) = \frac{1}{\beta^2}(\beta Q_i - \alpha\beta + \beta Q_i \log(\frac{\overline{U}\beta^2}{\beta Q_i + \gamma_i})) + \frac{\gamma_i}{\beta^2}\log(\frac{\overline{U}\beta^2}{\beta Q_i + \gamma_i}) \tag{3.14}$$

I discuss the Slutsky condition in more detail in a later section.

Proof:

1. To derive indirect utility, recall Roy's Identity (3.3), which relates indirect utility, $v(y_{is}, p_s)$, to demand. The general solution of this partial differential equation for $v(y_{is}, p_s)$ is as follows, where F is a general function:

$$v(y_{is}, p_s) = F\left(e^{-\beta p_s}[\frac{\alpha}{\beta} + y_{is} + \frac{\gamma}{\beta}p_s + \frac{X_i\delta}{\beta} + \frac{\gamma}{\beta^2}]\right) \tag{3.15}$$

Since utility only has meaning up to a monotonic transformation in this model, set F equal to the function that returns its argument:

$$v(y_{is}, p_s) = e^{-\beta p_s}[\frac{\alpha}{\beta} + y_{is} + \frac{\gamma}{\beta}p_s + \frac{X_i\delta}{\beta} + \frac{\gamma}{\beta^2}] \tag{3.16}$$

This functional form satisfies Roy's Identity:

$$
\begin{aligned}
-\frac{\partial v(y_{is},p_s)/\partial p_s}{\partial v(y_{is},p_s)/\partial y_{is}} &= -\frac{(-\beta e^{-\beta p_s}[\frac{\alpha}{\beta}+y_{is}+\frac{\gamma}{\beta}p_s+\frac{X_i\delta}{\beta}+\frac{\gamma}{\beta^2}]+e^{-\beta p_s}[\frac{\gamma}{\beta}])}{e^{-\beta p_s}} \quad (3.17)\\
&= \beta[\frac{\alpha}{\beta}+y_{is}+\frac{\gamma}{\beta}p_s+\frac{X_i\delta}{\beta}+\frac{\gamma}{\beta^2}]-[\frac{\gamma}{\beta}] \quad (3.18)\\
&= \alpha+\beta y_{is}+\gamma p_s+X_i\delta = Q(y_{is},p_s) \quad (3.19)
\end{aligned}
$$

Note that when $\beta = 0$, indirect utility is undefined.

2. To derive direct utility, using the demand function and the budget set, solve for p_s and y_{is} in terms of Q_i and A_i:

$$
\begin{aligned}
p_s &= -\frac{(\beta A_i - Q_i + \alpha)}{\beta Q_i + \gamma_i} \quad (3.20)\\
y_{is} &= \frac{-Q_i^2 + Q_i\alpha - A_i\gamma_i}{\beta Q_i+\gamma_i} = A_i + \frac{Q_i(\beta A_i - Q_i+\alpha)}{\beta Q_i+\gamma_i}
\end{aligned}
$$

Substitute these expressions into the indirect utility function.

Although many general utility functions imply infinite utility and demand when the price of one good is zero, this utility function implies finite demand when the price of medical care is zero. From (3.10), we can see that when the price is zero, demand is determined entirely by virtual income and covariates. This property makes the model tractable without requiring any ad hoc assumptions about factors that make the price of medical care nonzero after the stoploss is met.

3. To derive the indifference curve, fix utility and solve for A_i :

$$
\begin{aligned}
\overline{U} &= \frac{(\beta Q_i+\gamma_i)}{\beta^2}\exp[\frac{\beta(\beta C_i - Q_i+\alpha)}{\beta Q_i+\gamma_i}] \quad (3.21)\\
A_i &= \frac{1}{\beta^2}(\beta Q_i - \alpha\beta + \beta Q_i \log(\frac{\overline{U}\beta^2}{\beta Q_i+\gamma_i})) + \frac{\gamma_i}{\beta^2}\log(\frac{\overline{U}\beta^2}{\beta Q_i+\gamma_i}) \quad (3.22)
\end{aligned}
$$

If and only if indifference curves are convex, the second derivative with respect to Q_i will be positive:

$$\frac{\partial^2 A_i}{\partial Q_i^2} = \frac{-1}{(\beta Q_i + \gamma_i)} \geq 0 \tag{3.23}$$

It follows immediately that this condition will be satisfied when the Slutksy condition holds.

3.4.1 Discussion of Conditions for Integrability

Symmetry and negativity of the Slutsky matrix is necessary to recover preferences from demand. (See Mas-Collel et al. (1995).) In a partial equilibrium model, the Slutsky matrix is necessarily symmetric. From the Slutsky equation, the Slutsky matrix S is defined as.

$$S = \frac{\partial Q(y_{is}, p_s)}{\partial p_s} + \frac{\partial Q(y_{is}, p_s)}{\partial y_{is}} Q(y_{is}, p_s) \tag{3.24}$$

It follows directly from this equation that for linear demand to satisfy negative semidefiniteness of the Slutsky matrix, the following condition must hold:

$$\gamma_i + \beta Q_i \leq 0 \tag{3.25}$$

Since it must be true that $Q_i \geq 0$, and since we expect $\gamma_i \leq 0$, $\beta \geq 0$, this condition is unlikely to be satisfied globally. Rather than making the unrealistic assumption that $\beta \leq 0$, so that the condition would be satisfied globally, or imposing the condition directly, I impose $\gamma_i \leq 0$ and check the estimated coefficients to see if this condition holds. In practice, when the Slutsky condition is not met, indirect utility will not be single-valued, and some consumers will locate along extensions of budget segments that are in the interior of the budget set. Thus, violations of the

Slutsky condition are readily apparent in data simulated according to the model.

In the nonlinear budget set literature, Slutsky conditions have received a great deal of attention. In the labor supply literature, the Slutsky condition can be satisfied globally if the labor supply elasticity is positive and the income elasticity is negative, but it is not automatically satisfied. MaCurdy et al. (1990) and MaCurdy (1992) brought attention to the role of Slutsky condition in the labor supply literature and proposed an alternative local linearization method to smooth around the kinks in the budget set and relax the Slutsky condition. However, Blomquist (1995) shows that even under local linearization, the Slutsky condition must be satisfied for the estimated parameters to be interpreted as labor supply parameters. He also shows that neither method automatically produces parameter estimates that satisfy the Slutsky condition. More recently, Heim and Meyer (2003) emphasize that though the MaCurdy work is valuable because it demonstrates where the Slutsky condition matters, it does not provide an alternative method. Following the literature, I proceed subject to the caveat that the Slutsky condition must be satisfied.

3.5 Estimation

For purposes of exposition, I first describe an estimation strategy for a simple empirical application with one nonconvex kink and one source of unobserved preference heterogeneity. In turn, I then incorporate a second kink, a corner solution outcome, and another source of unobserved preference heterogeneity. Finally, I discuss other practical considerations for estimation.

3.5.1 Simple Case: One Nonconvex Kink

Assume that there is no stoploss. After meeting the deductible, agents purchase all further medical care at the coinsurance rate. Formally, agents face a nonconvex

budget set with only two segments, $s = \{a, b\}$. For now, assume that all agents consume a positive amount of medical care. Given convex preferences, for each individual, there is a region around each nonconvex kink point that will not be chosen. We can allow the size of this region to vary by allowing γ, the coefficient on the price term in the demand function, to vary across individuals with cdf F_γ. The generalized demand specification is as follows:

$$Q(y_{is}, p_s) = \alpha + \beta y_{is} + \gamma_i p_s + X_i\delta \tag{3.26}$$

Following the above derivation, the indirect utility function corresponding to this demand function is as follows:

$$v(y_{is}, p_s) = e^{-\beta p_s}[\frac{\alpha}{\beta} + y_{is} + \frac{\gamma_i}{\beta}p_s + \frac{X_i\delta}{\beta} + \frac{\gamma_i}{\beta^2}] \tag{3.27}$$

Using this indirect utility function, it is possible to calculate a critical γ_i^{ab} for which an individual is indifferent between a point on either segment a or b:

$$v(y_{ia}, p_a) = v(y_{ib}, p_b) \tag{3.28}$$

$$e^{-\beta p_a}[\frac{\alpha}{\beta} + y_{ia} + \frac{\gamma_i^{ab}}{\beta}p_a + \frac{X_i\delta}{\beta} + \frac{\gamma_i^{ab}}{\beta^2}] = e^{-\beta p_b}[\frac{\alpha}{\beta} + y_{ib} + \frac{\gamma_i^{ab}}{\beta}p_b + \frac{X_i\delta}{\beta} + \frac{\gamma_i^{ab}}{\beta^2}] \tag{3.29}$$

Rearranging, we can express γ_i^{ab} in closed form:

$$\gamma_i^{ab} = \frac{-\beta\{e^{-\beta p_a}[\alpha + \beta y_{ia} + X_i\delta] - e^{-\beta p_b}[\alpha + \beta y_{ib} + X_i\delta]\}}{e^{-\beta p_a}[\beta p_a + 1] - e^{-\beta p_b}[\beta p_b + 1]} \tag{3.30}$$

If medical expenditure is a normal good, we expect γ_i to be negative for all i (as the price increases, quantity demanded decreases). For very negative values of γ_i, the quantity of medical care consumed will be small. As γ_i increases, there will be a critical γ_i^{ab} for which the agent will be indifferent between consumption of a bundle

on the first segment and another bundle on the second segment. Formally, we expect $-\infty < \gamma_i^{ab} < 0 \; \forall \; i$.

To develop a better understanding of the model, we can derive an expression for the region of the budget set that will not be chosen for each individual. To do so, express demand as a function of the budget segment and γ_i^{ab}:

$$Q_i = \left\{ \begin{array}{l} Q_{ia} = \alpha + \beta y_{ia} + \gamma_i p_a + X_i\delta \;\; \text{if } \gamma_i < \gamma_i^{ab} \\ Q_{ib} = \alpha + \beta y_{ib} + \gamma_i p_b + X_i\delta \;\; \text{if } \gamma_i > \gamma_i^{ab} \; . \end{array} \right\} \tag{3.31}$$

Technically, the case where $\gamma_i = \gamma_i^{ab}$ occurs with zero probability. However, this case is important theoretically because it determines the size of the region that will not be chosen. We can express the gap in quantity for each individual, g_i, as follows:

$$\begin{aligned} g_i &= (Q_{ib} - Q_{ia} | \gamma_i = \gamma_i^{ab}) \\ &= \beta(y_{ib} - y_{ia}) + \gamma_i^{ab}(p_b - p_a) \end{aligned}$$

To further simplify this expression, we can express virtual income on each segment in terms of actual income, Y_i, the coinsurance rate C, and the deductible D by generalizing (3.7)-(3.9) for individual heterogeneity

$$y_{ia} = Y_i \tag{3.32}$$

$$y_{ib} = Y_i - (1 - C)D \tag{3.33}$$

$$y_{ic} = Y_i - S \tag{3.34}$$

and substituting $p_a = 1$ and $p_b = C$:

$$
\begin{aligned}
g_i &= \beta(Y_i - (1-C)D - Y_i) - (1-C)\gamma_i^{ab} \\
&= -(1-C)(\beta D + \gamma_i^{ab})
\end{aligned}
$$

The size of this region varies across individuals only through γ_i^{ab}, which, as apparent from (3.30), varies only through the covariates X_i, and income Y_i. It follows directly that if and only if income and the covariate do not vary across individuals, the gap will not vary across individuals. We can determine how income affects the size of the region by taking the following partial derivative:

$$
\begin{aligned}
\frac{\partial g_i}{\partial Y_i} &= \frac{\partial(-(1-C)\gamma_i^{ab})}{\partial Y_i} && (3.35) \\
&= \frac{\beta^2(e^{-\beta(1-C)} - 1)}{\beta + 1 - (1+C)e^{-\beta(1-C)}} && (3.36)
\end{aligned}
$$

This expression depends only on the income coefficient β and the coinsurance rate C. In general, we expect $\beta \in (0,1)$. Therefore, $e^{-\beta(1-C)} \leq 1$ when $\beta \in (0,1)$. When $C = 1/5$, as is common in many PPO plans, it follows that an increase in income increases the size of the gap around each kink point. Intuitively, when income increases, γ_i^{ab} increases. It follows directly from (3.23) that an increase in γ_i^{ab} makes the indifference curve that intersects segments a and b less convex. Therefore, g_i increases. Similarly, assuming one covariate for simplicity,

$$
\begin{aligned}
\frac{\partial g_i}{\partial X_i} &= \frac{\partial(-(1-C)\gamma_i^{ab})}{\partial X_i} && (3.37) \\
&= (1-C)\delta\beta\left(\frac{(1 - e^{-(1-C)\beta})}{\beta + 1 - (C\beta + 1)e^{-(1-C)\beta}}\right)
\end{aligned}
$$

Since $(C\beta+1)e^{-(1-C)\beta} < (\beta+1)e^{-(1-C)\beta} < \beta+1$, the denominator is positive, so the whole expression shares the size of δ. Intuitively, if δ is positive, an increase in X_i makes the indifference curve that intersects segments a and b less convex, and g_i increases. The opposite argument holds for negative δ.

Likelihood Function for one nonconvex kink

In this simple model, γ_i^{ab} allows for a simple formulation of the likelihood. In general terms, the likelihood of observing an individual i on a given segment of the budget set is as follows:

$$L_i = \Pr(p_a) + \Pr(p_b) \tag{3.38}$$

where

$$\begin{aligned} \Pr(p_a) &= \Pr(v_{ia} > v_{ib}) \\ \Pr(p_b) &= \Pr(v_{ib} > v_{ia}) \end{aligned} \tag{3.39}$$

using the expression derived for γ_i^{ab} and the distribution F_γ,

$$\begin{aligned} \Pr(p_a) &= F_\gamma(\gamma_i^{ab}) \\ \Pr(p_b) &= 1 - F_\gamma(\gamma_i^{ab}) \end{aligned} \tag{3.40}$$

Although this likelihood appears similar to that of a probit or logit model, the functional form is different because heterogeneity enters multiplicatively instead of additively.

Summing the log of the likelihood over all individuals, we can use MLE to solve:

$$\max_{\alpha,\beta,\delta,\theta_\gamma} \sum_i \log L_i \tag{3.41}$$

where θ_γ are the unknown parameters of F_γ. As an alternative to traditional maximum likelihood techniques, which can be difficult to implement when the likelihood is complicated, this model can be combined with diffuse and hierarchical priors so that it can be estimated in a Bayesian setting. Since Bayesian techniques require integration of a posterior density instead of maximization of a function, they can be more computationally tractable.

3.5.2 General Case: Two Nonconvex Kinks

In practice, estimating a model that assumes away the existence of the second nonconvex kink would arguably involve selection based on the dependent variable, leading to bias. Therefore, I extend the model to incorporate a second nonconvex kink. The presence of a second nonconvex kink leads to two more indifference conditions: one for indifference between segments b and c and another for indifference between segments c and a.

Following the above derivation, and substituting $p_c = 0$ and $p_a = 1$, we can express γ^{bc} and γ^{ac} as follows:

$$\gamma_i^{bc} = \frac{-\beta\{e^{-\beta p_b}[\alpha + \beta y_{ib} + X_i\delta] - 1\}}{e^{-\beta p_b}[\beta p_b + 1] - 1} \tag{3.42}$$

$$\gamma_i^{ac} = \frac{-\beta\{e^{-\beta}[\alpha + \beta y_{ia} + X_i\delta] - 1\}}{e^{-\beta}[\beta + 1] - 1} \tag{3.43}$$

The indifference condition between segments a and c makes it difficult to calculate expressions for the region around each nonconvex kink that will not be chosen. Furthermore, it makes the functional form of the likelihood much more complicated.

The likelihood becomes a general multinomial model instead of an ordered model.

Likelihood Function for Two Nonconvex Kinks

In a traditional ordered model, like the ordered probit model, the outcomes are ordered, and a monotonic change in heterogeneity produces a monotonic change in the ordered outcome. In contrast, in this model, the outcomes are ordered, but a monotonic change in heterogeneity need not produce a monotonic change in the ordered outcome, and the direction of the change can vary across agents. Thus, a more general multinomial model is required. Specifically, the probability of a particular segment is equal to the probability that virtual income is higher on that segment than it is on the other two segments:

$$\Pr(p_s) = \Pr(v_{is} > v_{it} \ \forall t \neq s)$$

Rather than express these probabilities as a complicated function of the three indifference conditions defined by (3.30),(3.42), and (3.43), we can formulate the multinomial model more succinctly in terms of indirect utility using a simple trick. First, add an error term, ε_{is}, to indirect utility for each segment, and assume that ε_{is} is distributed iid extreme value :

$$\widetilde{v}(y_{is}, p_s) = e^{-\beta p_s}[\frac{\alpha}{\beta} + y_{is} + \frac{\gamma}{\beta}p_s + \frac{X_i\delta}{\beta} + \frac{\gamma}{\beta^2}] + \varepsilon_{is} \tag{3.44}$$

Now, if we integrate out the ε_{is}, we are left with a simple closed form expression (See Train (2003), page 78 for a proof):

$$\Pr(p_s|\gamma_i) = \frac{e^{v_{is}}}{\Sigma e^{v_{is}}} \tag{3.45}$$

where, as defined above, v_{is} is a function of γ_i. Assuming ε_{is} and γ_i are independent, we can integrate this closed form over the distribution of γ_i to express the following probability of each segment:

$$\Pr(p_s) = \int \frac{e^{v_{is}}}{\Sigma e^{v_{is}}} f(\gamma_i) dF(\gamma_i) \tag{3.46}$$

The new likelihood for each observation is as follows:

$$L_i' = \Pr(p_a) + \Pr(p_b) + \Pr(p_c) \tag{3.47}$$

We can estimate the coefficients by substituting L_i' for L_i in (3.41).This likelihood is now a random coefficients logit model. A variant of this model, in which (indirect) utility is linear in the parameters, is used extensively in the industrial organization literature, and it has been shown to be globally concave. However, the nonlinearity of the parameters in this model makes estimation more difficult.

3.5.3 Extension to Include Zero Care

Until this point, we have not modeled the probability of choosing zero medical care. Unlike in the labor supply application, in which the probability of zero working hours by prime-aged males is low, in this application, the probability of choosing zero medical care is high, with over 30% of agents choosing zero care in each year. Following the industrial organization literature, one way to model the probability of choosing zero medical care would be to model it as a fourth "outside good" with its own associated indirect utility. Alternatively, we could attempt to handle censoring that arises from the decision to consume zero care nonparametrically. However, it is possible to gain more identification and bring the model closer to the theory by explicitly modeling the choice of zero care as a corner solution. The new demand function is as follows:

$$\max(Q(y_{is}, p_s), 0) = \alpha + \beta y_{is} + \gamma p_s + X_i\delta \tag{3.48}$$

In utility theory, indirect utility is direct utility evaluated at the chosen point on the interior of the budget set. However, if the agent chooses a corner solution, his choice is governed by utility, not indirect utility. The utility associated with consuming zero care, obtained by plugging $Q = 0$ into (3.13) is as follows:

$$U(0, A_i) = \frac{\gamma_i}{\beta^2} \exp[\frac{\beta(\beta y_{ia} + \alpha)}{\gamma_i}] \tag{3.49}$$

The agent will choose to consume zero care if the utility associated with consuming zero care is larger than indirect utility on all segments:

$$U(0, A_i) > v(y_{is}, p_s) \quad \forall s \tag{3.50}$$

As above, it is possible to express this condition in terms of a critical γ_i^{0s} for each segment s.

$$\frac{\gamma_i^{0s}}{\beta^2} \exp[\frac{\beta(\beta y_{is} + \alpha)}{\gamma_i^{0s}}] > e^{-\beta p_a}[\frac{\alpha}{\beta} + y_{is} + \frac{\gamma_i^{0s}}{\beta} p_s + \frac{X_i\delta}{\beta} + \frac{\gamma_i^{0s}}{\beta^2}] \quad \forall s \tag{3.51}$$

However, because the ordering of these critical values need not be the same across agents, it is simpler to formulate the model in terms of utility and indirect utility instead of in terms of the critical values. As above, we can add an iid extreme value error term ε_{is} to the expression for utility of zero care:

$$\widetilde{U}(0, A_i) = \frac{\gamma_i}{\beta^2} \exp[\frac{\beta(\beta y_{ia} + \alpha)}{\gamma_i}] + \varepsilon_{iz} \tag{3.52}$$

Now, we can add zero care as a fourth alternative, z, in the likelihood above. Because the unobserved component of heterogeneity is additive and distributed iid

with the ε_{is} of the three segments, a variant of the closed form expression (3.45) applies. However, there is an additional complication. The Slutsky condition ensures that maximum indirect utility on each segment will occur on the outer envelope of the budget set. However, it does not ensure that maximum indirect utility will occur on the region of the budget set associated with positive values of Q. In fact, because unconstrained optimization always achieves weakly higher utility than constrained optimization, and utility of zero represents a constrained alternative, the new segment z will never be chosen. Thus, instead of adding utility of zero as a fourth alternative, we can condition on Q_i in the likelihood through the use of indicator functions:

$$\Pr(p_z) = \int f_\gamma(\gamma_i) dF_\gamma(\gamma_i) 1(Q_i = 0) \tag{3.53}$$

$$\Pr(p_s) = \int \frac{e^{v_{is}}}{\Sigma e^{v_{is}}} f(\gamma_i) dF(\gamma_i) 1(Q_i > 0) \tag{3.54}$$

The new likelihood for an individual observation is as follows:

$$L_i' = \Pr(p_z) + \Pr(p_a) + \Pr(p_b) + \Pr(p_c) \tag{3.55}$$

Unlike in the labor supply application, we know the chosen segment with certainty, so we can condition on it in the likelihood. If an agent chooses zero care, we know that utility governs his decision, and if an agent chooses a positive amount of care, we know that indirect utility governs his decision. [Unfortunately, the likelihood as written does not incorporate the functional form of the utility function. In Hausman (1980) the utility function is incorporated into the model through the indifference condition defined by (3.52). Is there another way to incorporate utility at zero into the model?]

3.5.4 Accounting for Additional Heterogeneity

In all likelihoods presented above, once the segment is known, the value of Q_i is completely determined by the demand equation, or, at the corner solution, the value of Q_i is zero. As a consequence, all of the likelihoods presented above can be estimated without the actual expenditure data as long as the segment is known. However, in applications where precise expenditure data are available, it is desirable to incorporate these data into the likelihood. Furthermore, the extent to which predicted expenditure differs from actual expenditure might be due to optimization error, which we can model. In this application, we can think of optimization error as a manifestation of a health shock or an inability to control the exact amount spent on care. Burtless and Hausman (1978) allow an optimization error term to allow individuals to locate in regions around the nonconvex kinks.

Intuitively, one way to incorporate the expenditure data is to maximize the likelihood subject to a moment condition that requires that the difference between predicted expenditure and actual expenditure is as small as possible. Within the maximum likelihood framework, we can achieve a similar objective by adding an additional error term to the demand equation:

$$\max(Q^*(y_{is}, p_s), 0) = \alpha + \beta y_{is} + \gamma_i p_s + X_i\delta + \omega_i \tag{3.56}$$

$$Q_i^* = Q_i + \omega_i \tag{3.57}$$

Note that we observe Q_i^*, while Q_i is predicted by the model. Assume ω_i is distributed $N(0, \sigma_\varepsilon^2)$. Given the additional assumption that γ_i, ε_{is}, and ω_i are inde-

pendent, the likelihood of observing an individual Q_i is as follows:

$$\Pr(p_s) = \int f(\gamma_i) d\gamma_i (Q_i^* = 0) + \int \frac{1}{\sigma_\varepsilon} \phi(\frac{Q_i - \alpha - \beta y_{ia} - \gamma_i p_a - X_i \delta}{\sigma_\varepsilon}) \frac{e^{v_{is}}}{\Sigma e^{v_{is}}} f(\gamma_i) d\gamma_i (Q_i^* > 0) \tag{3.58}$$

This additional optimization error only affects the likelihood for those observations with positive expenditure. As discussed in Hausman (1980), it is unlikely that there is a divergence between actual and preferred consumption for people who consume zero care.

3.5.5 Practical Considerations for Estimation

Extension to Unobserved Income

In many sources of medical claims data, information on income is not available, so the model must be extended to deal with this data limitation. In practice, lack of data on the insurance premium has the same implications as lack of income because, like income, the premium shifts the entire budget set downward. At first blush, the simplest way to deal with the lack of data is to assume that the coefficient on income in my demand function is zero. However, as discussed above, indirect utility is not defined when $\beta = 0$. To get around this difficulty, recall that virtual income varies even when income does not. When we do not observe income, we can still substitute for virtual income correction terms as follows:

$$y_{ia} = \overline{Y} \tag{3.59}$$

$$y_{ib} = \overline{Y} - (1 - C)D \tag{3.60}$$

$$y_{ic} = \overline{Y} - S \tag{3.61}$$

where $\overline{Y}$ is a measure of underlying income that is common to individuals. As shown in (3.35), given this assumption, the convexity of indifference curves will no longer vary across individuals with respect to income. However, as shown in (3.37), the convexity of indifference curves will still vary with respect to covariates. Simulation methods can be used to examine the robustness of estimates of γ_i when income is not observed.

Other Extensions

The model that I have presented here can be extended to gain more identification and take other features of health insurance policies into account. For example, the model can be readily extended to allow for varying deductibles, coinsurance rates, and stoplosses across different plans. Theoretically, the model can be extended to incorporate time series variation. To facilitate comparison with the instrumental variables literature, the model can also be extended to account for family interactions in the cost sharing provisions.

3.6 Conclusion

In this paper, I have developed a model to estimate the price elasticity of expenditure on medical care using medical claims data. This model allows me to estimate heterogeneity in preferences, and it allows for welfare calculations following Hausman (1981). Furthermore, the model incorporates censoring due to the decision to consume zero care with an approach that is firmly grounded in utility theory. By generalizing the model to incorporate more than one nonconvex kink, I contribute to the nonlinear budget set literature.

In future research, I intend to estimate this model with Marketscan medical claims data used in the first chapter of this dissertation to produce reduced form instrumen-

tal variables estimates. I can then compare and combine the structural and reduced form techniques. Specifically, I can use estimates from the reduced form paper to simulate the theoretical distribution of consumers that should be observed in the regions around the kinks and then compare it to the empirical distribution following Liebman and Saez (2006). More formally, I can incorporate the reduced form instrument into the structural model, and I can re-estimate the reduced form model controlling for a virtual income correction. By comparing and combining reduced form and structural approaches, I aim to make a methodological contribution as well as a substantive one.

Bibliography

[1] Blomquist, Soren. "Restriction in labor supply estimation: Is the MaCurdy critique correct?" Economics Letters. 47(1995), pp. 229-235.

[2] Blomquist, Soren and Newey, Whitney. "Nonparametric Estimation with Nonlinear Budget Sets." Econometrica, 2002, 70(6), pp. 2455-80.

[3] Burtless, Gary and Hausman, Jerry A. "The Effect of Taxation on Labor Supply: Evaluating the Gary Negative Income Tax Experiment." The Journal of Political Economy, 1978, 86(6), pp. 1103-30

[4] Burtless, Gary, and Moffit, Robert A. "The Joint Choice of Retirement Age and Postretirement Hours of Work." Journal of Labor Economics April 1985. 3(2), pp. 209-236.

[5] Eichner, Matthew J. "Medical Expenditures and Major Risk Health Insurance," Massachusetts Institute of Technology, 1997, 1-66.

[6] Eichner, Matthew J. "The Demand for Medical Care: What People Pay Does Matter." The American Economic Review. Papers and Proceedings of the Hundred and Tenth Annual Meeting of the American Economic Association. May 1998. 88(2) pp. 117-121.

[7] Englehardt, Gary V. and Kumar, Anil. "Employer matching and 401(k) Saving: Evidence from the Health and Retirement Study." Mimeo. 2006.

[8] Friedberg, Leora. "The Labor Supply Effects of the Social Security Earnings Test." The Review of Economics and Statistics. February 2000. 82(1), pp. 48-63.

[9] Halpern, Janice, and Hausman, Jerry. "Choice Under Uncertainty: A Model of Applications for the Social Security Disability Insurance Program." Journal of Public Economics 31(1986), pp. 131-161.

[10] Hausman, Jerry A. "Individual Discount Rates and the Purchase and Utilization of Energy-Using Durables." The Bell Journal of Economics. 1979. 10(1), pp. 33-54.

[11] ____. "The Econometrics of Nonlinear Budget Sets." Econometrica, 1985, 53(6), pp. 1255-82.

[12] ____. "The Effect of Taxes on Labor Supply," H. Aaron and J. Pechman, How Taxes Affect Economics Behavior. Washington, D.C.: Brookings, 1981, 27-84.

[13] ____. "The Effect of Wages, Taxes, and Fixed Costs on Women's Labor Force Participation." Journal of Public Economics, 1980, 14, pp. 161-94.

[14] Heim, Bradley T. Meyer, Bruce D. "Structural Labor Supply Models when Constraints are Nonlinear." Mimeo. June 10, 2003.

[15] Hurd, Michael. "Estimation of Nonlinear Labor Supply Functions iwth Taxes form a Truncated Sample." Stanford Research Institute. Research Memorandum 36. November 1976.

[16] Keeler, E.B. Newhouse, J.P., Phelps, C.E. "Deductibles and the Demand for Medical Care Services: The Theory of a Consumer Facing a Variable Price Schedule under Uncertainty." Econometrica. April 1977. 45(3), pp. 641-656.

[17] Kowalski, Amanda E. "Censored Quantile Instrumental Variable Estimates of the Price Elasticity of Expenditure on Medical Care." Mimeo. 2008.

[18] Kumar, Anil. "Nonparametric Estimation of the Impact of Taxes on Female Labor Supply." Federal Researve Bank of Dallas. July 2004.

[19] Liebman, Jeffrey and Saez, Emmanuel. "Earnings Responses to Increases in Payroll Taxes." September 2006.

[20] Liebman, Jeffrey, and Zeckhauser, Richard. "Schmeduling." Mimeo. October 2004.

[21] MaCurdy, Thomas. "Work Disincentive Effects of Taxes: A Reexamination of Some Evidence." American Economic Review, 1992, 82(2), pp. 243-49.

[22] MaCurdy, Thomas, Green, David, and Paarsh, Harry. "Empirical Approaches for Analyzing Taxes and Labor Supply." The Journal of Human Resources. Special Issue on Taxation and Labor Supply in Industrial Countries. 1990. 25(3) pp.4415-490.

[23] Manning, Willard G, Newhouse, Joseph P., Duan, Naihua, Keller, Emmett B., and Leibowitz, Arleen. "Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment." The American Economic Review, Jun 1987, 77(3), pp. 251-277.

[24] "MarketScan Database," Ann Arbor,MI: The MEDSTAT Group Inc., 2005.

[25] Mas-Collell, Andreu, Whinston, Michael D., and Green, Jerry R. "Microeconomic Theory." Oxford University Press. 1995.

[26] Newhouse, Joseph P. and the Insurance Experiment Group. Free for All? Lessons from the RAND Health Insurance Experiment. Harvard University Press. Cambridge: 1993.

[27] Phelps, Charles E. and Newhouse, Joseph P. "Coinsurance, the Price of Time, and the Demand for Medical Services." The Review of Economics and Statistics. August 1974, 56(3), pp. 334-342.

[28] Rosen, Harvey. "Taxes in a Labor Supply Model with Joint Wage-Hours Determination." Journal of Public Economics, 1979, 11, pp. 1-23.

[29] Saez, Emmanuel. "Do Taxpayers Bunch at Kink Points?" NBER Working Paper 7366, 1999.

[30] Train, Kenneth E. "Discrete Choice Methods with Simulation." Cambridge University Press 2003.

Figure 3-1: Nonlinear Budget Set for Medical Care

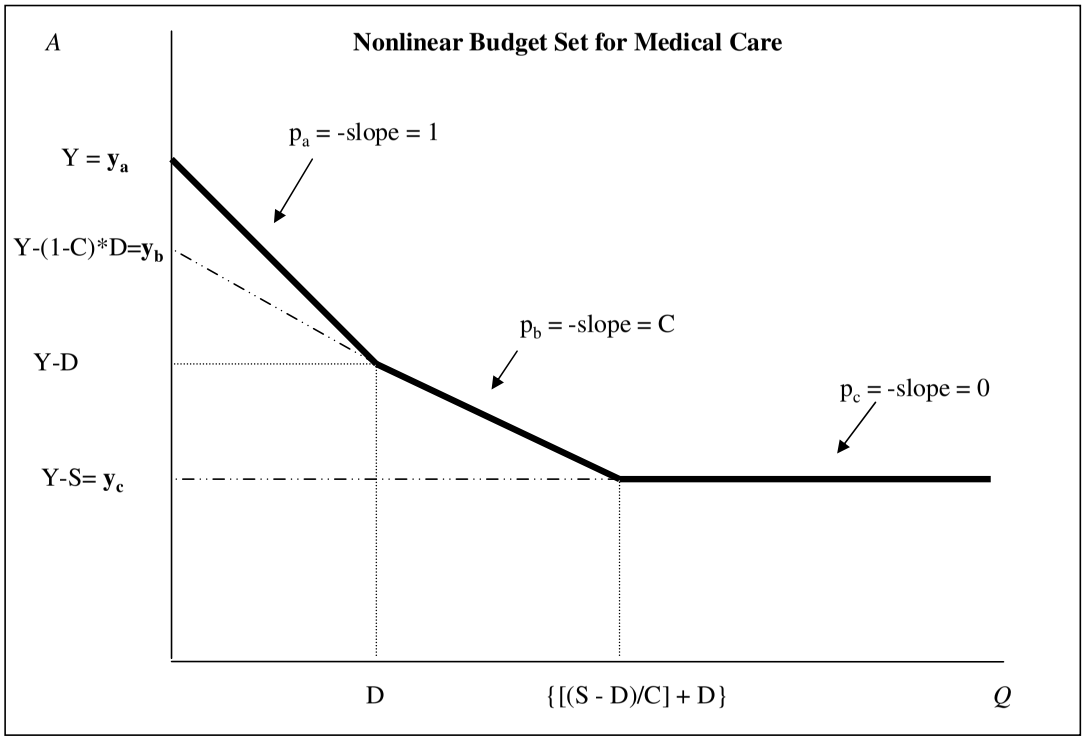


Figure 3-2: Reference Case: Nonlinear Budget Set Under Simple Progressive Tax