Regression Discontinuity Designs with an Endogenous Forcing Variable and an Application to Contracting in Health Care

Patrick Bajari, University of Minnesota and NBER

Han Hong, Stanford University

Minjung Park, University of Minnesota

Robert Town, University of Minnesota and NBER

April 24, 2010

Abstract

Regression discontinuity designs (RDDs) are a popular method to consistently estimate linear econometric models when ordinary least squares would be biased. However, researchers have observed that the continuity condition of RDDs may fail if the forcing variable can be manipulated by the agent. In this paper, we propose sufficient conditions for consistent estimation of an RDD-style model with this feature. In our paper, we first write down a canonical model from economic theory in which a maximizing agent faces a non-linear constraint set. We show that if there is no "bunching" of types at the discontinuity in the optimal solution to this problem, a modified RDD estimation strategy is possible. A novel aspect of our paper is that we ground our RDD-style estimator in economic theory and clarify primitive economic assumptions that guarantee consistent estimation when

the forcing variable is endogenous. As an application of our method, we study contracts between a large health maintenance organization and leading hospitals for the provision of organ transplants. The contracts used to reimburse the hospitals have "donut holes" and therefore there are sudden, discontinuous changes in hospitals' reimbursement rate. We use these discontinuous changes to estimate how the total claims filed by the hospitals depend (locally) on the generosity of the reimbursement structure. Our results clearly demonstrate that hospitals will submit significantly larger bills if they are not faced with incentives to economize. Our results thus suggest that informational asymmetries in the relationship between HMOs and hospitals are significant in this market.

1 Introduction

Regression discontinuity design (RDD) is a commonly used method to consistently estimate linear econometric models. In empirical work, researchers often encounter econometric models where ordinary least squares estimation will generate biased estimates. This happens when the classic exogeneity condition fails because one or more of the regressors is correlated with unobservables or determined simultaneously with the dependent variable.

In an RDD, the researcher searches for a forcing variable that exogenously shifts the regressor of interest. If such a variable can be found, an RDD generates a consistent estimate so long as the continuity condition holds (see Hahn, Todd, and Van der Klaauw, 2001). A number of researchers have pointed out that this condition may fail if the forcing variable can be manipulated by an agent (see McCrary, 2008; Lee and Lemieux, 2010; Imbens and Lemieux, 2008). Lee and Lemieux (2010) argue that many published papers using RDDs may suffer from this problem and that the standard arguments guaranteeing the consistency of the estimator may fail as a result. Urquiola and Verhoogen (2009) demonstrate that economic theories of sorting may predict failure of the continuity condition. They point out that previously proposed RDD identification strategies in hedonic regressions and public finance may not be consistent as a result.

In this paper, we propose an alternative strategy that can be applied when the forcing variable is determined simultaneously with the dependent variable. We begin by examining a canonical model from microeconomics of an agent's choice when she faces a non-linear constraint set. In this framework, the agent can choose the dependent variable of interest as well as influence the forcing variable of the RDD. The constraint set depends on this forcing variable in a non-linear fashion.

As labor economists have known for decades, models with non-linear budget sets may generate "bunching." That is, choice models may predict that a positive mass of agents will make

precisely the same choice near a kink point in a budget set. Labor economists have also observed that there may be a "gap" near a kink point. That is, models may predict that no choices should be observed in some neighborhood of the kink point in the constraint set. When there is either bunching or a gap, standard RDDs would not work. In case of bunching, the continuity condition is likely to fail. The presence of a gap also violates the continuity condition of RDDs since the dependent variable will not have full support in this region. In summary, it can be problematic to apply standard RDDs to important choice models since the forcing variable can be manipulated by the agent.

Some researchers have argued that it may be possible to resolve these issues if there is "optimization error" or some other factor which limits the ability of agents to control the forcing variable near the kink point (see Lee and Lemieux, 2010). We demonstrate that this solution to the problem of an endogenous forcing variable relies on assumptions which may be inconsistent with canonical choice models from microeconomic theory.

In our paper, we propose a modified RDD estimator which can yield consistent estimates when the forcing variable can be manipulated by the agent. For many choice models, the optimal solution would imply a strictly monotonic relationship between the type of the agent and the agent's choice, except at bunching. The key insight in our paper is that we can exploit this monotonicity to recast the problem such that the type of the agent is seen as the forcing variable. We then apply an RDD-style estimation to this reformulated problem. We provide a set of conditions under which our estimator is consistent. Since our estimator relies on strict monotonicity between the type and the dependent variable, the estimator can be used where economic theory predicts the possibility of a gap, but not where there is bunching.

We apply our estimator to understand a fundamental question in health economics—provider agency. Hospitals, physicians and other health care providers possess more information about the appropriateness and necessity of care than the patient or, importantly, their insurer. This fact combined with the likelihood that health care providers are concerned with their own financial

well-being implies that first-best contracts may be difficult to write and implement. Understanding the magnitude of this agency problem is a requisite step to both assessing the welfare consequences of provider agency and designing the optimal contracts in health care settings. Physicians and hospitals control most of the flow of resources in the health care system and medical care expenditures are a large component of most industrialized countries' GDP—in the U.S., health care expenditures are currently over 16% of GDP with a real growth rate of approximately 4% (Congressional Budget Office, 2008). Thus, the welfare gain from better aligning incentives in these contracts with societal objectives is potentially very large. Despite the importance of this issue and the existence of a large theoretical literature (McGuire, 2000), little quality empirical work exists studying the role of financial incentives in affecting provider behavior.[1]

We have collected a unique data set on contracts between hospitals and one of the largest U.S. health insurers for organ transplants. Organ transplants are an extremely expensive but rare procedure. In 2007, 27,578 organs were transplanted in the U.S. and the average total billed charges for kidney transplantation in our data, the least expensive and most commonly transplanted organ, exceed $140,000. Between 2005 and 2008, the cost of organ transplant rose at an annual rate of 14%—a rate that is higher than general health care cost inflation. Our analysis will allow us to determine if generous reimbursement rates can partly explain this rate of increase or whether it is attributable to some other sources such as underlying hospital costs. To the best of our knowledge, no other study in the literature has gathered a panel data set of reimbursement contracts between a major insurer and hospitals of this form and detail.

The form of the contracts in our data is fairly simple. As a hospital treats patients, it uses its information system to keep track of all reimbursable expenses, which include, but are not limited to, drugs, nights in the hospital and care from health providers. Our hospitals have

[1]The discussion in the Centers for Medicare and Medicaid Services report on implementing pay-for-performance in Medicare well summarizes the state of knowledge. No definitive body of research exists and indicates the optimal payment policy parameters for achieving the goals of the Value-Based Purchasing program (Centers for Medicare and Medicaid Services, p. 8).

standard "list prices" for each of these reimbursable expenses which are maintained in their "chargemaster." The sum of all of these list prices times the reimbursable items is referred to as charges. The contract specifies what fraction of the charges submitted by the hospital for each patient will be reimbursed by the insurer.

A key feature of the reimbursement schedules is that the total reimbursement amount for each patient follows a piecewise linear schedule: the marginal reimbursement rate changes discontinuously when certain levels of expenditure are reached. This generates discontinuities in the marginal price received by the hospital for its provision of health care. Even a visual inspection of the data suggests that these incentives are important since the frequency of expenditures changes substantially from the left hand side to the right hand side of these discontinuity points.

In this problem, health care expenses would be the forcing variable (as well as the outcome variable) and different reimbursement rates would be treatments in the terminology of standard RDDs. Clearly, the forcing variable is a choice variable of the hospitals. Using a model of hospitals' optimal health care provision, we verify that in the model the key conditions required for our estimator are satisfied at one of the discontinuity points in the reimbursement schedule. We then apply our estimator to that discontinuity point to estimate how the total claims filed by the hospital depend on the reimbursement structure. Our results clearly demonstrate that hospitals will submit significantly larger bills if they are not faced with incentives to economize. When the marginal reimbursement rate changes from 0% to 50%, a magnitude of change typically found in our contracts, the marginal increase in hospitals' expenditures for a given increase in patients' illness severity becomes 2 to 14 times larger. These results suggest that there are significant informational asymmetries in the relationship between HMOs and hospitals in this market.

Our paper makes the following contribution to the literature. Researchers have argued that RDDs may not work if the forcing variable is a choice variable of the agent. This makes a

straightforward application of standard RDDs to estimating demand, supply or other behavioral responses difficult. We show that the problems, under certain conditions, can be recast so that the agent's type is viewed as the forcing variable. Our proposed estimator thus allows researchers to study a wider set of problems within an RDD-style framework.

The rest of this paper proceeds as follows. In Section 2, we present a model of hospitals' health care choice. In Section 3, we propose our estimator and discuss its asymptotic properties. Section 4 provides literature review on agency problems in health care markets. Section 5 describes our data and Section 6 presents model estimates. Section 7 concludes the paper.

2 Model

2.1 The Agency Problem

Consider a firm that designs compensation contracts (the principal) for the provider of a medical service (the agent). A patient is identified who has a health condition with severity $\theta \geq 0$, and θ is a random variable with a pdf $f(\theta)$ and cdf $F(\theta)$. The health shock captures heterogeneity in the demand for health care. We assume that patients' heterogeneity is one-dimensional, fully captured by θ. The provider then chooses a level of treatment $q \geq 0$. The value of the health outcome to the patient is given by $v(q, \theta)$, which is twice continuously differentiable. The cost of providing treatment at level q is given by $c(q)$. Given these, the socially optimal level of treatment solves

$$\max_{q \geq 0} \; v(q, \theta) - c(q).$$

The optimal level of health care treatment considers both the benefits to the patient and the costs of treatment.

The patient, who is a passive player in this framework, arrives at the agent's facilities with

her realization of severity θ. The agent observes θ and chooses the level of health care q. The principal, however, cannot observe θ but can observe the choice q. Hence, the principal cannot directly contract on the optimal level of q, and instead must rely on a compensation scheme to the agent of the general form $r(q)$ in order to implement the desired q.

To continue, one needs to specify the payoff functions of the principal and the agent. Naturally, the cost of treatment is born by the agent, and $r(q)$ is paid to the agent by the principal. We assume that the net monetary benefits of the principal are $k - r(q)$, where k is some fixed payment that he receives from the patient (insurance premium). We assume that the agent's net monetary benefits are just $r(q) - c(q)$. Furthermore, we assume that each party receives a non-pecuniary benefit that is proportional to the patient's payoff. This captures the idea that both the principal and the agent benefit from successful health outcomes.[2] We also assume quasi-linear utility functions so that there are no income effects. We can write the payoffs of the principal and the agent as

$$\begin{aligned} u^p &= \gamma^p v(q,\theta) + k - r(q), \\ u^a &= \gamma^a v(q,\theta) - c(q) + r(q). \end{aligned}$$

In this paper, our main interest lies in understanding hospitals' behavioral responses to the reimbursement structure, not in understanding what the optimal reimbursement scheme should look like. Therefore, we abstract away from the optimal contract design problem faced by the principal, and just focus on the agent's optimal choice of health care spending given a reimbursement scheme. We note that in reality we might observe an incentive scheme that departs from the optimal one for various reasons, such as institutional constraints, lack of information or complexity in implementing the optimal contract.

The agent maximizes $\gamma^a v(q,\theta) - c(q) + r(q)$ and the FOC is (for now, ignoring potential

[2]For example, reputational concerns for attracting future patients or for deflecting scrutiny by regulators.

discontinuities in $r(q)$),

$$\gamma^a \frac{\partial v(q,\theta)}{\partial q} = c'(q) - r'(q). \tag{1}$$

The equality in (1) has a simple economic interpretation: the left hand side is the agent's *marginal benefit* from treatment while the right hand side is her *net marginal cost* (total marginal costs less marginal reimbursement).

2.2 Assumptions

We shall assume that the payoffs obey the following conditions:

$$\begin{aligned}
\frac{\partial v(q,\theta)}{\partial q} &> 0 && (2)\\
\frac{\partial^2 v(q,\theta)}{\partial^2 q} &< 0 && (3)\\
\frac{\partial v(q,\theta)}{\partial \theta} &< 0 && (4)\\
\frac{\partial^2 v(q,\theta)}{\partial \theta \partial q} &> 0 && (5)\\
\frac{\partial c(q)}{\partial q} &> 0 && (6)\\
\frac{\partial^2 c(q)}{\partial^2 q} &\geq 0 && (7)
\end{aligned}$$

Assumptions (2) and (3) state that the value of the health outcome to the patient is increasing and strictly concave in q. Assumption (4) implies that health shocks adversely affect utility. Assumption (5) implies that the value of the health outcome to the patient exhibits strict increasing differences in (q, θ): the marginal utility of health care increases as agents receive more adverse health shocks. According to assumptions (6) and (7), the cost of providing treatment is an increasing and (weakly) convex function in q.

This structure captures the intuitive idea that (i) health expenditures have an interior optimal level after which more health care decreases social welfare; (ii) a more severe condition has a higher marginal benefit of extra treatments; and (iii) providing more treatment costs more money, and marginal treatments are (weakly) more expensive. As a result, when a patient's condition is more severe she should be offered more treatment.

When the agent consumes q dollars of health care to treat a patient, the agent is reimbursed $r(q)$ by the principal. As we discussed in the introduction, we are interested in situations where the constraint set faced by the agent displays kinks or jumps. Reflecting the reimbursement schedules used by the health insurer in our data, we shall assume that $r(h)$ satisfies:

$$r(0) = 0 \tag{8}$$

$$r'(q) = \delta_1 \text{ for } 0 < q < q_1 \tag{9}$$

$$r'(q) = 0 \text{ for } q_1 \leq q \leq q_2 \tag{10}$$

$$r'(q) = \delta_2 \text{ for } q > q_2. \tag{11}$$

This assumption implies that the amount of reimbursement for each patient is piecewise linear. For expenditures between 0 and q_1, the hospital is reimbursed δ_1 for every dollar spent to treat the patient. Once expenditures exceed q_1, the hospital hits what is called the donut hole and is forced to bear all of its health care expenses at the margin. Finally, for expenditures above q_2, the hospital is reimbursed δ_2 for every dollar spent. Figure 1 illustrates a reimbursement scheme implied by assumptions (8)–(11).

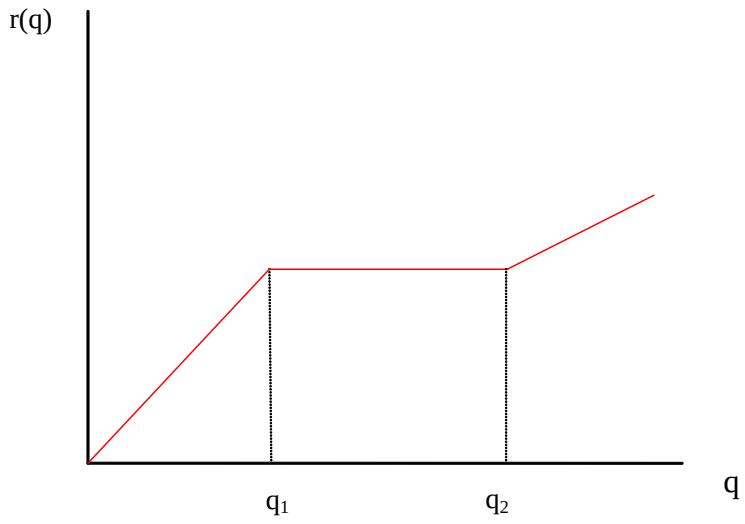

Figure 1: A Typical Reimbursement Scheme

2.3 Optimal Decision Rule

Under the assumptions written above, the optimal decision rule of an agent who treats a pool of patients exhibits the following features:

1. There will be bunching at q_1.
2. There may be a gap near q_2 and the size of the gap depends on the shape of $u^a(q, \theta)$.
3. The optimal choice of q is monotonically increasing in θ. In fact, the optimal choice q is strictly increasing in θ except for bunching at q_1.

Figure 2 illustrates these observations. In drawing the figure, we assume that $0 < \delta_2 < \delta_1 < 1$, which is what we typically observe in the data. The marginal benefit curve for a given level of θ is decreasing in q, and is given by $\gamma^a \frac{\partial v(q,\theta)}{\partial q}$. The lower is γ^a, the flatter are the marginal benefit curves. A higher θ is associated with a marginal benefit curve that is more to

the right (crosses 0 at a higher q). The net marginal cost curve is just $c'(q) - r'(q)$. For this figure, we assume that $c'(q)$ is constant, which is not crucial for any of our results but simplifies the graphical analysis.

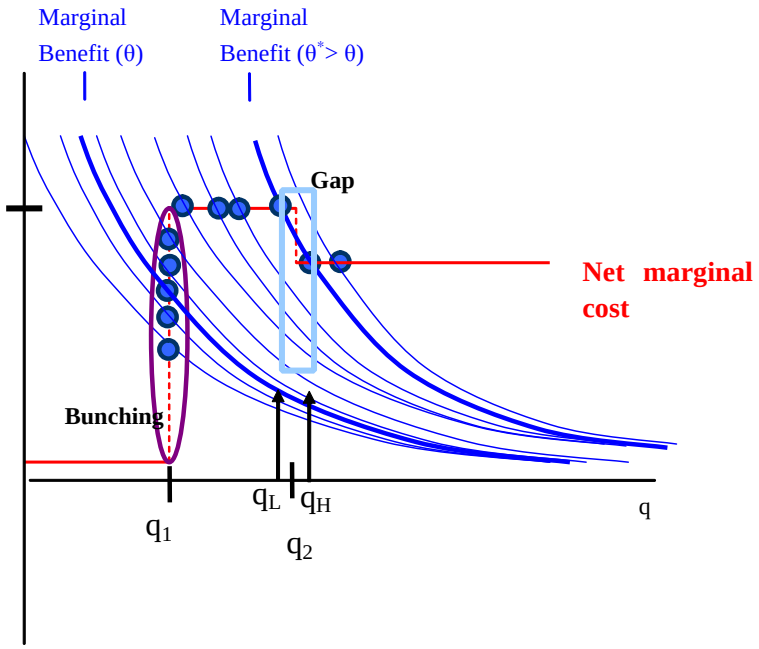

Figure 2: Optimal Decision Rule

Imagine a level of θ that corresponds to an optimal choice below q_1. As θ increases, the optimal choice will also increase until some level θ_1 at which it will be exactly q_1. Given the kink in the incentive scheme, there is an upward jump in the net marginal cost curve, causing bunching at q_1 for levels higher than θ_1. At some point, however, high enough levels of θ above θ_1 will cause the marginal benefit curve to shift enough so that optimal choices will exceed q_1 and be on the part of the net marginal cost curve that is $c'(q)$ (i.e., $r'(q) = 0$). The choice of q then continues to rise monotonically with θ until we hit a gap in choices just around q_2, where the net marginal cost drops. To see why we have a gap, consider the level θ^* that is depicted in

Figure 2. For this level of severity the agent is indifferent between choosing two levels of health care, one strictly below q_2 (say q_L) and another strictly above (say q_H). By the monotonicity of $q(\theta)$ which follows from the assumption of increasing differences in (q, θ), there will not be any choices of treatment that correspond to expenditures within the interval (q_L, q_H). We see that the flatter the marginal benefit curves are (smaller γ^a), the larger the gap (q_L, q_H) will be. Finally, for all $\theta > \theta^*$, $q(\theta)$ is strictly increasing. The graphical analysis of Figure 2 offers a complete treatment of what the agent's behavior would be in face of a kinked incentive scheme as described in Figure 1.

Throughout our discussion, we have assumed that the agent cannot "cheat" and fraudulently announce costs that were not incurred. If this can happen, then we might observe patterns that are not implied by the optimal decision rule. Although such a fraudulent reporting is not impossible, we think it is uncommon among the hospitals in our data because they are large, established hospitals that are subject to regular audits.

3 Estimation

In this section we propose our estimator that will yield consistent estimates of the agent's behavioral responses when the forcing variable is endogenously chosen by the agent. We first discuss the key intuition behind our approach and then outline our estimation procedures.

3.1 Using Discontinuous Changes for Identification

At the two discontinuity points q_1 and q_2, the marginal reimbursement rate faced by the hospital changes discontinuously. These discontinuities seem to present a natural setting for an RDD. Our problem, however, differs from typical RDD settings because q is both the forcing variable (the level of q determines the marginal reimbursement rate) and the dependent variable (our

goal is to estimate how the level of q responds to the marginal reimbursement rate). In this canonical choice model, the forcing variable is clearly endogenous.

One might think that an RDD estimator might be still consistent if there is "optimization error" or some other factor that prevents agents from precisely controlling the forcing variable (see Lee and Lemieux, 2010). This solution to the problem of an endogenous forcing variable does not work for our problem due to the fact that the forcing variable is the same as the dependent variable. When we add optimization error to the forcing variable, we are adding the same optimization error to the dependent variable. Thus, patients who are on the left hand side of a discontinuity and patients who are on the right hand side of the discontinuity will have systematically different optimization errors added to their outcomes, which will lead to inconsistent estimates under standard RDD estimation.

In our paper, we propose an alternative solution to the problem. A key step in our approach is to transform the problem so that we make the type of the patient θ a forcing variable. From the earlier discussion, and more generally the monotone comparative statics literature of Topkis (1978) and Milgrom and Shannon (1994), we know that the assumption of strict increasing differences in (θ, q) implies that the optimal health care provision q is a strictly increasing function of patient type θ, with the exception of where there is bunching at q_1. As a result, the percentiles of q will identify θ. That is, if we see a patient with the 5^{th} percentile of health expenditure within a hospital, that patient will have the 5^{th} percentile of health shock within that hospital. This means that for all practical purposes, the health shocks are observable to the econometrician. Since q is only weakly increasing in θ around the first discontinuity point due to the presence of bunching, the econometrician cannot infer θ from the cdf of q in the region. Hence, our estimation procedure can be applied to the second kink, but not the first one.

Once we reformulate the problem so that the patient type θ is a forcing variable (which is exogenously endowed and cannot be manipulated), a shift in the patient type θ determines whether the hospital's choice of q for that patient will be on the left hand side or right hand side of the

second discontinuity point. This then generates an exogenous change in the marginal price faced by the hospital, allowing for identification of the hospital's response to incentives. Our approach boils down to estimating a variant of regression discontinuity models in the empirical quantile function of the hospital's choice q.

Figure 3 helps illustrate the idea behind our approach. Among patients who come to the hospital with a realization of health shock θ, there will be a value of θ at which the hospital is indifferent between choosing q_L $(< q_2)$ and q_H $(> q_2)$. Let θ^* denote the level of severity which leads to such an indifference. Then for all patients whose θ is greater than θ^*, the hospital will choose q larger than q_H and will face a marginal reimbursement rate of δ_2. For patients whose θ is smaller than θ^*, the hospital will choose q smaller than q_L and will face a marginal reimbursement rate of 0. Thus, the hospital's supply of health care services will be more responsive to an increase in θ on the right hand side of θ^* than on the left hand side of θ^*. Therefore, by comparing how quickly q rises with an increase in θ for values of θ just below and just above θ^*, we can infer how the total claims filed by the hospital depend on the reimbursement structure. We let ϕ_4 denote change in the slope of the quantile function at θ^*.

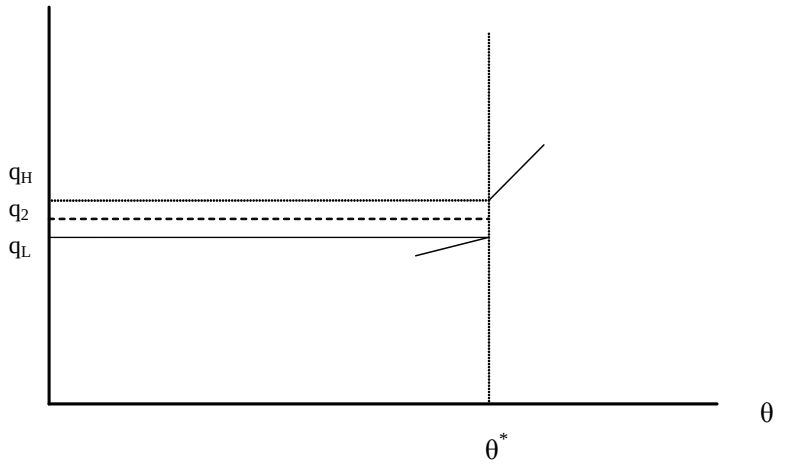

Figure 3: The slope of quantile function $q(\theta)$ changes at θ^*

The figure also shows a possibility of gap $\phi_2 = q_H - q_L$ at θ^*. As discussed earlier, for a high value of γ^a the gap could be very small, while for a low value of γ^a the gap might be large. Hence, it is an empirical question whether only the slope of $q(\theta)$ changes (i.e., $\phi_4 \neq 0$), or both the slope and intercept of $q(\theta)$ change at the discontinuity point θ^* (i.e., $\phi_4 \neq 0$ and $\phi_2 \neq 0$).

3.2 Estimation

We consider several different estimation approaches that deal with different levels of hospital heterogeneities. The first method applies to individual hospital data. The second method makes use of a global parametric assumption to pool information from data across all hospitals.

3.2.1 Individual Hospital Estimates

Suppose that there are $i = 1, ..., n$ individuals treated in the hospital under consideration. Let q_i denote the health expenditure of individual i. Let $\hat{F}(\cdot)$ denote the empirical distribution of the observed q's for the hospital. We propose estimators for ϕ_2 and ϕ_4 at the upper regression discontinuity point of q_2 and derive the asymptotic distribution of the estimators.

The incentive scheme is such that for θ approaching a cutoff value θ^*, $q(\theta)$ approaches a limit value q_L. As soon as θ moves to the right of θ^*, $q(\theta)$ takes a discrete jump at the point of θ^* by an amount $\phi_2 > 0$ to q_H.

By normalization, θ is estimated as the empirical CDF of the observed q. Hence θ^* is estimated by

$$\hat{\theta}^* = \hat{F}(q_2) = \frac{1}{n} \sum_{i=1}^{n} 1(q_i \leq q_2),$$

where $\hat{F}(\cdot)$ is the empirical distribution of the observed q's. Given that we define $\theta^* = F(q_2)$

where $F(\cdot)$ is the true distribution function of q, the asymptotic distribution of $\hat{\theta}^*$ is immediate:

$$\sqrt{n}(\hat{\theta}^* - \theta^*) \xrightarrow{d} N\left(0, F(q_2)(1 - F(q_2))\right).$$

We are interested in estimating the magnitude of the discontinuity ϕ_2. This is estimated by

$$\hat{\phi}_2 = \hat{q}_H - \hat{q}_L = \min\{q_i : q_i > q_2\} - \max\{q_i : q_i \leq q_2\}.$$

The goal is to derive the asymptotic distribution of $\hat{\phi}_2 - \phi_2$. It suffices to show that the joint distribution of $n(\hat{q}_L - q_L)$ and $n(\hat{q}_H - q_H)$ are independent exponential distributions. To see this, note that

$$\begin{aligned} P(n(\hat{q}_L - q_L) &\leq -x, n(\hat{q}_H - q_H) \geq y) \\ &= P(q_i \leq q_L - x/n, q_i \geq q_H + y/n, \forall i) \\ &= (1 - P(q_L - x/n \leq q_i \leq q_H + y/n))^n \\ &= \left(1 - f^- x/n - f^+ y/n + o(1)\right)^n \xrightarrow{n\to\infty} e^{f^- x + f^+ y}. \end{aligned}$$

In other words, $n(\hat{q}_L - q_L)$ and $n(\hat{q}_H - q_H)$ converge to two independent (negative and positive) exponential random variables with hazard rates $f^- = f(q_L)$ and $f^+ = f(q_H)$, where we have used f^- and f^+ to denote the (left and right) densities at q_L and q_H. The limiting distribution of $n(\hat{\phi}_2 - \phi_2)$ is therefore the sum of these two independent exponential random variables. When $f^- = f^+$,the limit distribution is a standard Gamma random variable with two degrees of freedom.

Next we turn to the estimation of difference between the slopes of $q(\theta)$ at q_H and q_L, defined as $\phi_4 = \lim_{\theta\to\theta^*_+} q'(\theta) - \lim_{\theta\to\theta^*_-} q'(\theta)$. Note that $\phi_4 = \frac{1}{f^+} - \frac{1}{f^-}$. Hence it suffices to obtain

consistent nonparametric estimators for f^+ and f^-. This can be done using standard one sided kernel smoothing methods.

Define

$$\hat{f}^- = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}k(\frac{\hat{q}_L - q_i}{h})1(q_i \leq \hat{q}_L),$$

and

$$\hat{f}^+ = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}k(\frac{q_i - \hat{q}_H}{h})1(q_i \geq \hat{q}_H).$$

In the above, $k(\cdot)$ is a one-sided density function supported on $(0, \infty)$, and h is a sequence of bandwidth parameters used in typical kernel smoothing. It is straightforward to show that as long as $nh \longrightarrow \infty$ and $nh^3 \longrightarrow 0$,

$$\sqrt{nh}\left(\hat{f}^- - f^-\right) \xrightarrow{d} N\left(0, f^- \int K(u)^2 du\right),$$

and

$$\sqrt{nh}\left(\hat{f}^+ - f^+\right) \xrightarrow{d} N\left(0, f^+ \int K(u)^2 du\right)$$

and that they are asymptotically independent. Therefore

$$\sqrt{nh}\left(\hat{\phi}_4 - \phi_4\right) \xrightarrow{d} N\left(0, \frac{1}{f^{-3}} \int K(u)^2 du + \frac{1}{f^{+3}} \int K(u)^2 du\right).$$

3.2.2 A Parametric Model Using Multiple Hospital Data

Now we consider how to extend the previous method to allow for pooling information from data across multiple hospitals. Consider first $\phi_2 = q_H - q_L$. We define $y_i = q_i 1(q_i \leq q_2)$ and $z_i = q_i 1(q_i > q_2) + M1(q_i \leq q_2)$. In the homogeneous case, we have defined $\hat{q}_L = \max\{y_i\}$ and $\hat{q}_H = \min\{z_i\}$, where M is number that is larger than any of the data points. This definition of the estimators can be extended to incorporate heterogeneous data from all hospitals.

With cross-hospital data, the observed threshold value q_2 can be hospital dependent, which we will denote as $q_2(t)$, where we have used t to index hospitals. Suppose hospital heterogeneity is captured by covariates x_t, where x_t can be $q_2(t)$ itself. Let I_t be the number of patient observations for each hospital. We specify the following parametric assumption that

$$q_L(t) \equiv q_L(x_t) = g_L(x_t, \beta_L) \quad \text{and} \quad q_H(t) \equiv q_H(x_t) = g_H(x_t, \beta_H).$$

In the above, we can use a flexible series expansion functional form of $g_L(x_t, \beta_L)$ and $g_H(x_t, \beta_H)$ so that they are linear in the parameters β_L and β_H. The structure of this problem fits into the boundary parameter estimation method studied in the literature. Possible estimators include the linear programming approach and nonstandard likelihood estimator (c.f. Donald and Paarsch, 1996; Chernozhukov and Hong, 2004) and the extreme quantile regression approach of Chernozhukov (2005). We describe these alternatives in the following.

The linear (or quadratic, etc.) programming approach estimates the parameters by

$$\begin{aligned} \hat{\beta}_L \quad &= \quad \arg\min_{\beta_L} \sum_{t=1}^{T} I_t^L g_L(x_t, \beta_L), \quad \text{where} \quad I_t^L = \sum_{i=1}^{I_t} 1(y_i > 0), \\ \text{such that} \quad & y_i \leq g_L(x_t, \beta_L), \forall i = 1, ..., I_t^L, t = 1, ..., T, \end{aligned}$$

and

$$\begin{aligned} \hat{\beta}_H \quad &= \quad \arg\max_{\beta_H} \sum_{t=1}^{T} I_t^H g_H(x_t, \beta_H), \quad \text{where} \quad I_t^H = \sum_{i=1}^{I_t} 1(z_i < M), \\ \text{such that} \quad & z_i \geq g_H(x_t, \beta_H), \forall i = 1, ..., I_t^H, t = 1, ..., T. \end{aligned}$$

The objective functions $\sum_{t=1}^{T} I_t^L g_L(x_t, \beta_L)$ and $\sum_{t=1}^{T} I_t^H g_H(x_t, \beta_H)$ can be replaced by

$$\sum_{t=1}^{T} \sum_{i=1}^{I_t} 1(y_i > 0)(y_i - g_L(x_t, \beta_L))^2 \text{ and } \sum_{t=1}^{T} \sum_{i=1}^{I_t} 1(z_i < M)(z_i - g_H(x_t, \beta_H))^2$$

or other types of penalization functions. The linear programming approach however seems to be the easiest to implement.

Alternatively, β_L and β_H can be estimated by the extreme quantile regression method of Chernozhukov (2005):

$$\hat{\beta}_L = \arg\min_{\beta_L} \sum_{t=1}^{T} \sum_{i=1}^{I_t} 1(y_i > 0)\rho_{\tau_L}(y_i - g_L(x_t, \beta_L)),$$

where $\rho_\tau(u) = (\tau - 1(u \leq 0))u$ is the *check function* of Koenker and Bassett (1978), such that

$$\tau_L \to 1 \quad \text{as} \quad n_L = \sum_t I_t^L \to \infty.$$

Similarly, $\hat{\beta}_H = \arg\min_{\beta_H} \sum_{t=1}^{T} \sum_{i=1}^{I_t} 1(z_i < M)\rho_{\tau_H}(z_i - g_H(x_t, \beta_H))$, where

$$\tau_H \to 0 \quad \text{as} \quad n_H = \sum_t I_t^H \to \infty.$$

The quantile regression approach has the advantage of being robust against a certain fraction of outliers in the data. On the other hand, the programming estimators always satisfy the constraints of the relation between y_i, z_i and $g_L(x_t, \beta_L)$ and $g_H(x_t, \beta_H)$.

By adopting a parametric functional form on $q_L(x_t)$ and $q_H(x_t)$ we are maintaining a strong specification assumption which can potentially be tested by the data. An implicit assumption of the parametric functional form is that $g_L(x_t, \beta_L^0) \leq q_2(t) \leq g_H(x_t, \beta_H^0)$ for all t at the true parameters β_L^0 and β_H^0. Of course their estimates introduce sampling noise, but we still expect that it should be largely true for most t:

$$g_L(x_t, \hat{\beta}_L) \leq q_2(t) \leq g_H(x_t, \hat{\beta}_H).$$

The approximate validity of this condition can be used as the basis of a model specification test.

Then $\phi_2(x_t)$ will be estimated consistently by

$$\hat{\phi}_2(x_t) = g_H(x_t, \hat{\beta}_H) - g_L(x_t, \hat{\beta}_L).$$

Conducting statistical inference on $\hat{\phi}_2(x_t)$ requires the limiting *joint* distribution of $\hat{\beta}_L$ and $\hat{\beta}_H$. They converge to a nonstandard distribution at a fast $1/n$ rate for $n = \sum_t I_t$. The limiting distribution can be obtained by simulation which we will describe below in the context of the parametric likelihood approach.

In fact we can also adopt a maximum likelihood approach. This will be useful in case we are interested in the shape of the distribution of q_{it} in order to conduct counter-factual welfare calculations. To this end, assume that

$$\epsilon_{it}^L = g_L(x_t, \beta_L) - y_{it} \sim f_L(\epsilon_{it}^L, x_t, \alpha_L) \text{ for } y_{it} \leq g_L(x_t, \beta_L),$$

and

$$\epsilon_{it}^H = z_{it} - g_H(x_t, \beta_H) \sim f_H(\epsilon_{it}^H, x_t, \alpha_H) \text{ for } z_{it} \geq g_H(x_t, \beta_H).$$

The maximum likelihood estimator for α_L, α_H and β_L, β_H can then be written as

$$\begin{aligned}(\hat{\alpha}_L, \hat{\beta}_L) &= \arg\max_{\alpha_L, \beta_L} \sum_{t=1}^{T} \sum_{i=1}^{I_t} 1(y_{it} > 0) \log f_L(g_L(x_t, \beta_L) - y_{it}, x_t, \alpha_L) \\ &\quad \text{such that} \quad y_{it} \leq g_L(x_t, \beta_L), \forall i = 1, ..., I_t, t = 1, ..., T,\end{aligned}$$

and

$$\begin{aligned}(\hat{\alpha}_H, \hat{\beta}_H) &= \arg\max_{\alpha_H, \beta_H} \sum_{t=1}^{T} \sum_{i=1}^{I_t} 1(z_{it} < M) \log f_H(z_{it} - g_H(x_t, \beta_H), x_t, \alpha_H) \\ &\quad \text{such that} \quad z_{it} \geq g_H(x_t, \beta_H), \forall i = 1, ..., I_t, t = 1, ..., T.\end{aligned}$$

In fact the linear programming estimator is a special case of the above maximum likelihood estimator when the densities $f_L(\epsilon_{it}^L, x_t, \alpha_L)$ and $f_H(\epsilon_{it}^H, x_t, \alpha_H)$ are exponential distribution with a homogeneous hazard rate parameter: $f(\epsilon) = \lambda e^{-\lambda\epsilon}$. In this case, in addition to obtaining $\hat{\beta}_L$ and $\hat{\beta}_H$ from the linear programming estimators, we also estimate the hazard parameters by

$$1/\hat{\lambda}_L = \frac{\sum_{t=1}^{T} \sum_{i=1}^{I_t} 1(y_{it} > 0) \left(g_L(x_t, \hat{\beta}_L) - y_{it} \right)}{\sum_{t=1}^{T} \sum_{i=1}^{I_t} 1(y_{it} > 0)}$$

and

$$1/\hat{\lambda}_H = \frac{\sum_{t=1}^{T} \sum_{i=1}^{I_t} 1(z_{it} < M) \left(z_{it} - g_H(x_t, \hat{\beta}_H) \right)}{\sum_{t=1}^{T} \sum_{i=1}^{I_t} 1(z_{it} < M)}$$

Even though $\hat{\beta}_L$ and $\hat{\beta}_H$ converge at $1/n$ rate to a nonstandard limit distribution, $\hat{\alpha}_L$ and $\hat{\alpha}_H$ are still root n consistent and asymptotically normal, as long as there is no functional relations between α and β.

To estimate $\phi_4(x_t)$, we can use

$$\hat{\phi}_4(x_t) = \frac{1}{f_H(0, x_t, \hat{\alpha}_H)} - \frac{1}{f_L(0, x_t, \hat{\alpha}_L)}.$$

Since $\hat{\phi}_4(x_t)$ is root n consistent and asymptotically normal, its limiting distribution can be obtained by the standard sandwich formula, or by simulation or bootstrap, in which $\hat{\beta}_L$ and $\hat{\beta}_H$ can be held fixed because they do not affect the asymptotic distribution.

The joint asymptotic distribution for $\hat{\beta}_L$ and $\hat{\beta}_H$ can be obtained by parametric simulations. Given the assumption that the parametric model is correctly specified, it is possible to simu-

late from the model using the estimated parameters $\hat{\beta}_L$, $\hat{\beta}_H$, $\hat{\alpha}_L$ and $\hat{\alpha}_H$. The approximate distribution can be obtained from repeated simulations. Instead of recomputing the maximum likelihood estimator at each simulation, it suffices to recompute weighted programming estimators of β_L and β_H at each simulation:

$$\begin{aligned}\tilde{\beta}_L &= \arg\min_{\beta_L} \sum_{t=1}^{T}\sum_{i=1}^{I_t} f_L(0, x_t, \hat{\alpha}_L)\frac{\partial g_L(x_t, \hat{\beta}_L)'}{\partial \beta_L}\beta_L \\ &\quad \text{such that} \quad y_i \leq g_L(x_t, \beta_L) \ \ \forall i, t,\end{aligned}$$

and

$$\begin{aligned}\tilde{\beta}_H &= \arg\max_{\beta_H} \sum_{t=1}^{T}\sum_{i=1}^{I_t} f_H(0, x_t, \hat{\alpha}_H)\frac{\partial g_H(x_t, \hat{\beta}_H)'}{\partial \beta_H}\beta_H \\ &\quad \text{such that} \quad z_i \geq g_H(x_t, \beta_H) \ \ \forall i, t.\end{aligned}$$

We can also consider the possibility that $g_L(x_t, \beta_L)$ and $g_H(x_t, \beta_H)$ are correctly specified but $f_L(\epsilon_{it}^L, x_t, \alpha_L)$ and $f_H(\epsilon_{it}^H, x_t, \alpha_H)$ are misspecified. In this case, each of the above methods (linear and quadratic programmings, extreme quantile regression, (pseudo) maximum likelihood estimation) will still deliver consistent estimates of β_L and β_H and hence ϕ_2. But the estimates for α_L, α_H and hence ϕ_4 are clearly inconsistent.

In this case, if we are willing to impose parametric assumptions on ϕ_2 through $g_L(x_t, \beta_L)$ and $g_H(x_t, \beta_H)$, but are not willing to make parametric assumption on ϕ_4, we can still estimate ϕ_4 using nonparametric density estimators. We can also use nonparametric density estimators to perform semiparametric simulations for consistent inference about $\hat{\phi}_2$. Suppose x_t is continuously distributed with dimension d. Let

$$\hat{f}^-(x) = \frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{I_t}\frac{1}{h}w(x_t, x)k(\frac{g_L(x_t, \hat{\beta}_L) - q_{it}}{h})1(q_{it} \leq g_L(x_t, \hat{\beta}_L)),$$

and

$$\hat{f}^{+}(x) = \frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{I_t}\frac{1}{h}w(x_t,x)k(\frac{q_{it}-g_H(x_t,\hat{\beta}_H)}{h})1(q_{it}\geq g_H(x_t,\hat{\beta}_H)),$$

where

$$w(x_t,x) = k^d(\frac{x_t-x}{h})/\sum_{t=1}^{T}k^d(\frac{x_t-x}{h}).$$

Then we can form the estimate $\hat{\phi}_4(x_t) = 1/f^{+}(x_t) - 1/f^{-}(x_t)$.

The limiting distribution of $\hat{\beta}_L$ and $\hat{\beta}_H$ can be obtained by recomputing the following weighted programmming estimators:

$$\begin{aligned}\tilde{\beta}_L &= \arg\min_{\beta_L}\sum_{t=1}^{T}\sum_{i=1}^{I_t}f^{-}(x_t)\frac{\partial g_L(x_t,\hat{\beta}_L)'}{\partial\beta_L}\beta_L \\ &\quad \text{such that} \quad y_i \leq g_L(x_t,\beta_L)\ \ \forall i,t,\end{aligned}$$

and

$$\begin{aligned}\tilde{\beta}_H &= \arg\max_{\beta_H}\sum_{t=1}^{T}\sum_{i=1}^{I_t}f^{+}(x_t)\frac{\partial g_H(x_t,\hat{\beta}_H)'}{\partial\beta_H}\beta_H \\ &\quad \text{such that} \quad z_i \geq g_H(x_t,\beta_H)\ \ \forall i,t.\end{aligned}$$

As before, the simulated distributions of $n(\tilde{\beta}_L - \hat{\beta}_L)$ and $n(\tilde{\beta}_H - \hat{\beta}_H)$ should approximate the unknown limit distributions of $n(\hat{\beta}_L - \beta_L^0)$ and $n(\hat{\beta}_H - \beta_H^0)$.

4 Literature on Provider Agency in Health Care Markets

Health care markets are rife with informational asymmetries which can be leveraged by health care providers to increase incomes relative to full information, first-best equilibrium. Arrow (1963) noted that a first-best insurance contract would specify a state-dependent payment. However, these contracts are not generally negotiated because health states are not readily ob-

served. Since the work of Arrow (1963), a large theoretical literature has arisen characterizing the optimal payment contract under different informational, preference and market structure scenarios.[3] However, most current provider-insurer contracts do not correspond to the structures generally prescribed by theory.[4] The failure of insurers to negotiate first-best contracts suggests that there is meaningful scope for provider agency. Given the size of the health care sector (approximately 16% of GDP), the potential welfare consequences of provider agency are extremely large.

Empirical analyses of the magnitude of the agency problem date to the work on physician induced demand of Fuchs (1978). The most important evidence on the presence of provider agency is provided by the *Dartmouth Atlas Project*. They find that there are large geographic variations in utilization by Medicare enrollees that are unrelated to health status. The *Dartmouth Atlas Project* suggests (but only provides limited econometric evidence) that the geographic differences are driven by geographic variation across providers in demand inducement. Many papers have attempted to estimate physician agency, but the identification strategies employed in these papers are generally suspect.[5] There are important exceptions, however. Gruber and Owings (1996) find that within state declines in fertility are associated with increases in cesarean sections. Yip (1997) shows that cardiac surgeons responded to payment reductions by increasing the number of procedures they performed. During the 1980s, Medicare changed its hospital reimbursement system from retrospective to prospective using Diagnostic Related Groups (DRG) as the basis of the payment. The incentive under prospective payment is to reduce the length of stay of Medicare beneficiaries and the policy appeared to have had the expected impact (Hodgkin and McGuire, 1994) without dramatically impacting the quality of care (Cutler, 1995). Even within the DRG system, hospitals appear to leverage their superior information into more generous payments. Dafny (2005) finds that when Medicare changed the

[3] McGuire (2000) provides an excellent review of this literature.

[4] For example, private insurers generally pay physicians on a fee-for-service or percentages of billed charges basis, while Medicare pays physicians on a fee-for-service basis and hospitals by groupings of diagnoses (Diagnosis Related Groups).

[5] See Dranove and Wehner (1994) for a discussion of the limitation of the attempts to estimate physician agency.

structure of the DRG payment generosity, hospitals responded by upcoding patients into more generous payment groups.

There has been very little detailed analysis of the response of providers to the specific incentives embedded in reimbursement and remuneration contracts. Gaynor and Gertler (1995) find that physicians reduce their effort when faced with lower powered incentives. Gaynor, Rebitzer and Taylor (2004) analyze a model of physician behavior under group incentives and test the predictions using detailed contract and financial data from a network HMO. They rely on variation in the size of the panel over which the group incentive is implemented to identify the parameters. They find that the HMO's incentive contract provides a typical physician with an increase, at the margin, of $0.10 in income for each $1.00 reduction in medical utilization expenditures. The presence of these high powered incentives reduced medical expenditures by 5%.

5 Data

Our estimation uses two data sets. The first data set comes from one of the largest private health insurers in the U.S. and has information on its contracts with 127 hospitals which specify reimbursement schedules for organ transplant surgeries. The other data set, also from the same insurer, has information on the set of patients who received organ transplants in each of the 127 hospitals. We merge the two, and the resulting data set has (i) claim-level information, such as the admission and discharge dates of the patient, the type of organ transplant received by the patient, the size of the bill submitted by the hospital to the insurer and the reimbursement amount paid by the insurer, as well as (ii) hospital-level information, such as the name and location of the hospital and the reimbursement schedule the hospital faces for each type of organ transplant surgery it performs. The data run from 2004 through 2006.

The insurer, a fortune 50 company, uses this network of hospitals for its own enrollees and

also sells access to this network to other health insurers and self-insured employers. This insurer is a major player in the organ transplant market, with its 80% market share among private vendors.

There are various types of organ and tissue transplants covered by the contracts, major ones being bone marrow transplant (BMT), kidney transplant, liver transplant, heart transplant and lung transplant. Organ transplants are a rare but exceedingly expensive procedure. In 2007, 27,578 organs were transplanted in the U.S. and the average total billed charges for kidney transplantation in our data, the least expensive and most commonly transplanted organ, exceed $140,000. Between 2005 and 2008, the cost of organ transplant rose at an annual rate of 14%—a rate that is larger than general health care cost inflation. An organ transplant is an extremely challenging and complex procedure taking anywhere from 3 (kidney) to 14 hours (liver). Organ transplants usually require significant post-operative care (up to 3 weeks of inpatient care) and careful medical management to prevent rejection. The infrequency of the procedures, the complexity of the treatments and the large variation across patients in their response to transplantation make it difficult for insurers to determine the appropriateness of the care for a given episode. That, in turn, implies that hospitals are in a position to engage in agency in response to the incentives embodied in their contracts.

The insurer negotiates a separate contract with each individual hospital, instead of having one common contract applied to all participating hospitals. As a result, the reimbursement schedule differs across hospitals. Typically, the reimbursement schedule takes a form as shown in Figure 1, but the exact locations of the first kink (q_1, also called inlier threshold) and the second kink (q_2, also called outlier threshold), the marginal reimbursement rate for each of the segments (δ_1 and δ_2) and the height of the donut hole ($\delta_1 q_1$) all vary across hospitals.[6] These differences likely reflect variation in bargaining power as well as heterogeneity in the patient pool across hospitals.

[6]There are some hospitals whose contracts do not have non-linearities and rather specify reimbursements as a fixed proportion of the bills. We exclude these hospitals from the estimation sample.

One practical issue we encounter is that the number of patients who receive a certain type of organ transplant within a hospital is typically very small. The average number of patients per hospital for a given type of organ transplant is 9.08 in year 2004, 8.74 in year 2005 and 8.05 in year 2006. The small number of patients could pose a serious problem for the performance of our estimator, particularly when estimation is done separately for each hospital (Section 3.2.1), since it leads to an imprecise estimation of the true quantile function of q. Since a hospital is likely to differ in its price sensitivity across different types of organ transplants, we cannot pool observations across organ types. However, we can pool observations across years for a given hospital and organ type since it seems plausible to expect that a given hospital's price sensitivity does not change during the sample period. To further reduce the potential bias arising from the small number of patients, we restrict our attention to (hospital, organ) pairs that have more than 30 patients over the years.[7] Table 1 presents summary statistics for the final sample.

[7] For individual hospital estimation, we use cutoff of 50.

Table 1: Summary Statistics

	BMT	Kidney	Liver
Total # Patients	791	388	396
Total # Hospitals	10	5	8
Avg. Reimbursement per Patient (in $1000)	96.65 (51.04)	72.15 (38.39)	187.7 (114.88)
Avg. # Patients per Hospital	79.1 (35.23)	77.6 (44.07)	49.5 (18.91)
Avg. q_1 across Hospitals (in $1000)	117.59 (18.75)	80.84 (22.89)	188.97 (16.22)
Avg. q_2 across Hospitals (in $1000)	167.86 (31.63)	136.05 (41.17)	265.69 (24.79)
Avg. δ_1 across Hospitals	0.75 (0.07)	0.79 (0.09)	0.79 (0.07)
Avg. δ_2 across Hospitals	0.53 (0.06)	0.48 (0.08)	0.56 (0.06)
Avg. % Patients with $q < q_1$	39.17 (20.2)	10.17 (5.26)	18.68 (13.39)
Avg. % Patients with $q_1 \leq q \leq q_2$	28.95 (11.64)	45.72 (11.49)	37.16 (9.01)
Avg. % Patients with $q > q_2$	31.88 (19.51)	44.11 (10.51)	44.16 (14.07)

Inside the parentheses are standard deviations.

In Figure 4 we show the distribution of q's of our sample, pooled across hospitals, organs and years. In order for us to be able to pool observations in the presence of different reimbursement schemes, we need to normalize q's. We show the distribution of q's that are normalized against the outlier threshold, $q_{OUT} = \frac{q - q_2}{\delta_1 q_1}$.

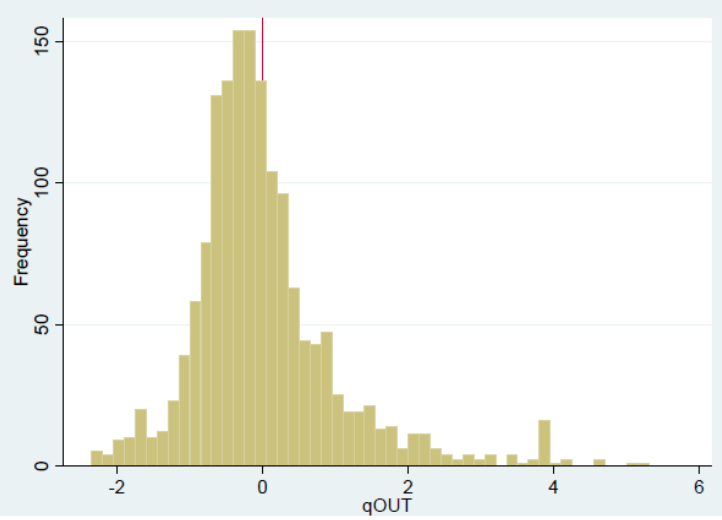

Figure 4: Distribution of q_{OUT}

The figure shows that the frequency of expenditures changes substantially from the left hand side to the right hand side of the discontinuity point q_2 (where q_{OUT} = 0), suggesting that the changes in the marginal price received by the hospitals are important for their choice of health care provision.

6 Results

We apply our proposed estimators to the second discontinuity point q_2 in order to estimate how the amount of health care provision depends on the reimbursement structure. In the first set of results, we apply maximum likelihood estimation to data pooled across multiple hospitals. The maximum likelihood estimation will yield $\hat{\alpha}_L, \hat{\alpha}_H, \hat{\beta}_L$ and $\hat{\beta}_H$, and these allow us to obtain the size of gap at the discontinuity point, $\hat{\phi}_2(x_t) = g_H(x_t, \hat{\beta}_H) - g_L(x_t, \hat{\beta}_L)$, and the change in the slope of the quantile function, $\hat{\phi}_4(x_t) = \frac{1}{f_H(0, x_t, \hat{\alpha}_H)} - \frac{1}{f_L(0, x_t, \hat{\alpha}_L)}$, for each hospital characterized by x_t. We use exponential distribution for densities f_L and f_H, with hazard rate parameter $\lambda_L(x_t, \alpha_L)$ and $\lambda_H(x_t, \alpha_H)$, respectively. All contract variables that potentially differ across

hospitals, such as the locations of the first and second kinks (q_1 and q_2) and the marginal reimbursement rates (δ_1 and δ_2), are included in x_t. We also include higher-order polynomials of these variables in x_t to flexibly capture the distribution of q for multiple hospitals. To compute standard errors, we use parametric bootstrap using 500 simulations. We apply MLE to each organ type separately.

In Tables 2-4, we report maximum likelihood estimates of ϕ_2 and ϕ_4 for each hospital in the data, along with hospital characteristics. Table 2 reports estimates for BMT, Table 3 for kidney transplants, and Table 4 for liver transplants.

Table 2: Maximum Likelihood Estimates, Bone Marrow Transplant (q measured in \$1,000)

	q_1	q_2	δ_1	δ_2	$\hat{\phi}_2$	Slope L	Slope R	$\hat{\phi}_4$
H1	135.7	158.3	0.7	0.6	15.75 (8.9)	16.67 (2.0)	107.9 (8.8)	91.18 (9.5)
H2	135.7	190.0	0.7	0.5	8.61 (18.4)	24.99 (3.6)	63.59 (7.7)	38.59 (8.7)
H3	135.7	172.7	0.7	0.55	9.93 (19.3)	19.78 (1.8)	82.15 (6.7)	62.37 (7.5)
H4	94.8	146.0	0.77	0.5	2.62 (3.7)	9.52 (2.6)	49.69 (7.0)	40.17 (7.4)
H5	133.3	181.8	0.75	0.55	5.13 (5.7)	24.15 (3.4)	86.44 (9.8)	62.29 (10.9)
H6	107.3	123.9	0.75	0.65	4.70 (6.8)	9.04 (2.0)	126.5 (15.8)	117.5 (15.9)
H7	120.0	150.0	0.75	0.6	7.10 (2.6)	13.89 (1.6)	102.9 (7.1)	89.04 (7.6)
H8	117.3	195.6	0.75	0.45	3.05 (14.3)	24.42 (4.9)	46.09 (6.7)	21.67 (8.1)
H9	130.8	170.0	0.65	0.5	3.44 (3.3)	16.12 (2.3)	56.85 (5.1)	40.73 (5.6)
H10	100.0	153.3	0.92	0.6	10.27 (22.1)	14.94 (1.7)	104.9 (7.6)	89.93 (8.2)

Inside the parentheses are bootstrapped standard errors.

Table 3: Maximum Likelihood Estimates, Kidney Transplant (q measured in \$1,000)

	q_1	q_2	δ_1	δ_2	$\hat{\phi}_2$	Slope L	Slope R	$\hat{\phi}_4$
H1	97.1	123.6	0.7	0.55	0.78 (8.34)	11.30 (1.30)	51.54 (5.16)	40.23 (5.50)
H2	74.7	124.4	0.75	0.45	6.69 (9.81)	14.25 (2.72)	74.07 (18.66)	59.82 (18.97)
H3	75.3	130.6	0.85	0.49	4.67 (3.98)	14.66 (1.48)	57.15 (6.04)	42.49 (6.60)
H4	79.4	137.8	0.85	0.49	3.80 (4.25)	16.75 (2.03)	50.41 (5.98)	33.65 (6.53)
H5	83.4	144.6	0.85	0.49	1.95 (4.51)	19.05 (3.26)	44.66 (8.04)	25.60 (8.76)
H6	70.7	108.3	0.92	0.6	1.44 (2.50)	7.62 (1.53)	55.86 (8.65)	48.24 (8.83)

Inside the parentheses are bootstrapped standard errors.

Table 4: Maximum Likelihood Estimates, Liver Transplant (q measured in \$1,000)

	q_1	q_2	δ_1	δ_2	$\hat{\phi}_2$	Slope L	Slope R	$\hat{\phi}_4$
H1	202.9	258.2	0.7	0.55	51.76 (31.42)	28.47 (2.92)	149.6 (12.02)	121.1 (13.52)
H2	120.0	163.6	0.75	0.55	9.94 (9.24)	13.04 (3.94)	149.6 (12.02)	136.5 (13.81)
H3	178.5	206.0	0.75	0.65	2.39 (91.49)	16.00 (2.93)	182.0 (17.12)	166.0 (17.34)
H4	166.5	288.8	0.85	0.49	13.99 (16.09)	39.99 (7.82)	132.9 (18.06)	92.9 (20.50)
H5	160.0	218.2	0.75	0.55	10.01 (22.74)	20.46 (3.31)	149.6 (12.02)	129.1 (13.43)
H6	200.0	254.6	0.7	0.55	7.77 (12.99)	27.63 (2.84)	149.6 (12.02)	121.9 (13.46)
H7	198.8	271.1	0.75	0.55	26.02 (14.20)	31.67 (3.61)	149.6 (12.02)	117.9 (13.88)
H8	198.7	250.0	0.78	0.62	6.54 (5.42)	24.04 (2.96)	171.6 (11.95)	147.5 (12.87)

Inside the parentheses are bootstrapped standard errors.

From the results in Tables 2-4, we see that $\hat{\phi}_4$ is positive and statistically significant for all hospitals across all organ types. This suggests that for a given increase in the severity of patient

health shock, hospitals tend to increase their health care spending by a larger amount when they face a positive marginal reimbursement rate than when the marginal reimbursement rate is zero. To interpret the magnitude of the coefficients, take the results for hospital 1 in Table 2. The hospital increases its bone marrow transplant spending by $166.7 for one percentile increase in patient illness severity when it is on the LHS of the kink (marginal reimbursement rate = 0%), while it increase its spending by $1079 for one percentile increase in illness severity when it is on the RHS of the kink (marginal reimbursement rate = 60%). This amounts to approximately six and a half times larger sensitivity of hospitals' health care spending to BMT patients' health condition due to the hike in the reimbursement rate. Similarly, take the results for hospital 1 in Table 3. The hospital increases its kidney transplant spending by $113 for one percentile increase in illness severity when it is on the LHS of the kink (marginal reimbursement rate = 0%), while it increase its spending by $515.4 for one percentile increase in illness severity when it is on the RHS of the kink (marginal reimbursement rate = 55%). This amounts to approximately four and a half times larger sensitivity of hospitals' health care spending to kidney transplant patients' health condition due to the increase in the reimbursement rate. Similar results hold for liver transplants as well.

Overall, the sensitivity of health care spending to patient illness is 2 to 14 times larger on the RHS than on the LHS for bone marrow transplants, 2 to 7 times larger on the RHS than on the LHS for kidney transplants, and 3 to 11 times larger on the RHS than on the LHS for liver transplants. What is also interesting is that $\hat{\phi}_4$ tends to be larger when δ_2 is larger, which again suggests that hospitals are sensitive to reimbursement rates in their health care provision decision. For instance, In Table 2, $\hat{\phi}_4$ is largest for Hospital 6, which has largest δ_2, and $\hat{\phi}_4$ is smallest for Hospital 8, which has smallest δ_2. Similar patterns hold for kidney transplants (Table 3) and liver transplants (Table 4).

Another pattern we observe in Tables 2-4 is that $\hat{\phi}_2$ is always positive, although almost always insignificant. The fact that $\hat{\phi}_2$ is always positive alleviates concerns about possible model

misspecification. As we discussed in Section 3.2.2, an implicit assumption of the parametric functional form is that $g_L(x_t, \beta_L^0) \leq q_2(t) \leq g_H(x_t, \beta_H^0)$ for all t at the true parameters β_L^0 and β_H^0. Since we find that $g_L(x_t, \hat{\beta}_L) < g_H(x_t, \hat{\beta}_H)$ holds for all hospitals in the data, there is no evidence of model misspecification. The fact that $\hat{\phi}_2$ is statistically insignificantly different from zero suggests no clear gap at q_2. This is consistent with a visual inspection of the data, where there isn't any clear evidence of gap around the second discontinuity point.

Since the results suggest no evidence of gap, we re-estimate the model, this time not allowing the possibility of gap. This is equivalent to assuming $g_L(x_t, \beta_L) = g_H(x_t, \beta_H) = q_2$. In this model, we only need to estimate α_L and α_H. The results are reported in Tables 5-7.

Table 5: Maximum Likelihood Estimates, Bone Marrow Transplant (q measured in \$1,000)

	q_1	q_2	δ_1	δ_2	Slope L	Slope R	$\hat{\phi}_4$
Hospital 1	135.71	158.33	0.7	0.6	21.02 (2.68)	114.89 (9.63)	93.87 (10.14)
Hospital 2	135.71	190.00	0.7	0.5	27.29 (3.86)	68.20 (8.41)	40.91 (9.24)
Hospital 3	135.71	172.73	0.7	0.55	23.23 (2.22)	87.66 (7.34)	64.43 (7.81)
Hospital 4	94.81	146.00	0.77	0.5	10.72 (3.06)	50.55 (7.02)	39.82 (7.83)
Hospital 5	133.33	181.82	0.75	0.55	28.17 (3.98)	93.26 (10.94)	65.08 (11.92)
Hospital 6	107.33	123.85	0.75	0.65	12.41 (2.84)	131.32 (16.63)	118.91 (16.96)
Hospital 7	120.00	150.00	0.75	0.6	17.61 (2.15)	108.55 (7.78)	90.94 (8.20)
Hospital 8	117.33	195.56	0.75	0.45	25.00 (4.96)	49.00 (7.10)	24.00 (8.32)
Hospital 9	130.77	170.00	0.65	0.5	17.85 (2.57)	59.52 (5.13)	41.67 (5.69)
Hospital 10	100.00	153.33	0.92	0.6	18.90 (2.28)	111.04 (8.27)	92.14 (8.70)

Inside the parentheses are bootstrapped standard errors.

Table 6: Maximum Likelihood Estimates, Kidney Transplant (q measured in \$1,000)

	q_1	q_2	δ_1	δ_2	Slope L	Slope R	$\hat{\phi}_4$
Hospital 1	97.14	123.64	0.7	0.55	11.94 (1.42)	52.41 (5.31)	40.47 (5.55)
Hospital 2	74.67	124.44	0.75	0.45	17.43 (3.35)	78.05 (20.70)	60.62 (21.07)
Hospital 3	75.29	130.61	0.85	0.49	16.62 (1.68)	58.82 (6.39)	42.20 (6.74)
Hospital 4	79.41	137.76	0.85	0.49	18.63 (2.28)	51.35 (6.29)	32.72 (6.81)
Hospital 5	83.38	144.64	0.85	0.49	20.80 (3.58)	45.04 (8.29)	24.24 (9.10)
Hospital 6	70.65	108.33	0.92	0.6	7.79 (1.63)	57.03 (8.91)	49.24 (9.08)

Inside the parentheses are bootstrapped standard errors.

Table 7: Maximum Likelihood Estimates, Liver Transplant (q measured in \$1,000)

	q_1	q_2	δ_1	δ_2	Slope L	Slope R	$\hat{\phi}_4$
Hospital 1	202.86	258.18	0.7	0.55	38.61 (4.02)	159.80 (14.85)	121.18 (15.49)
Hospital 2	120.00	163.64	0.75	0.55	17.81 (5.97)	142.31 (22.91)	124.50 (23.67)
Hospital 3	178.53	206.00	0.75	0.65	20.06 (3.76)	181.56 (17.29)	161.50 (17.84)
Hospital 4	166.47	288.78	0.85	0.49	56.86 (11.28)	147.88 (24.11)	91.02 (26.76)
Hospital 5	160.00	218.18	0.75	0.55	27.83 (4.62)	152.15 (13.28)	124.32 (14.08)
Hospital 6	200.00	254.55	0.7	0.55	37.48 (3.90)	159.09 (14.26)	121.60 (14.88)
Hospital 7	198.80	271.09	0.75	0.55	42.92 (4.98)	162.35 (17.48)	119.43 (18.33)
Hospital 8	198.72	250.00	0.78	0.62	30.79 (3.83)	180.92 (16.70)	150.12 (17.34)

Inside the parentheses are bootstrapped standard errors.

The results in Tables 5-7 are very similar to those in Tables 2-4, which is not surprising given that gap around q_2 did not seem important in earlier results.

In the second set of results, we perform estimation for each hospital separately. We apply kernel estimator as discussed in Section 3.2.1 to estimate ϕ_4. We do not estimate ϕ_2, taking our earlier results into account. For individual hospital estimation, we use hospitals that have more than 50 patients in order to ensure that we have enough observations for each estimation. In our estimates, half-normal kernels with various choices of bandwidth were used to construct the weights. The bandwidths listed in Table 8 correspond to 0.6 to 1.7 times the sample standard deviation for LHS estimation, and 0.05 to 0.26 times the sample standard deviation for RHS estimation (sample standard deviation is larger for RHS data points since they are more dispersed). Table 8 reports kernel estimates of ϕ_4 for each hospital and each organ type.

Table 8: Kernel Estimates (q measured in \$1,000)

	Bandwidth	Slope L	Slope R	$\hat{\phi}_4$
H5 (BMT)	15	38.04 (0.69)	48.40 (1.71)	10.36 (2.4)
H7 (BMT)	15	27.67 (0.50)	236.27 (51.1)	208.60 (51.64)
H3 (Kidney)	15	43.03 (1.87)	44.21 (1.20)	1.18 (3.08)
H5 (Kidney)	15	35.73 (0.78)	75.45 (9.50)	39.73 (10.28)
H6 (Kidney)	15	22.31 (0.21)	63.53 (2.61)	41.23 (2.82)
H4 (Liver)	15	133.45 (55.87)	70.57 (8.81)	-62.88 (64.69)
H8 (Liver)	15	54.78 (1.99)	108.71 (9.48)	53.93 (11.47)
H5 (BMT)	18	40.02 (0.67)	51.23 (1.6)	11.21 (2.36)
H7 (BMT)	18	30.27 (0.54)	221.10 (34.9)	190.83 (35.47)
H3 (Kidney)	18	42.68 (1.52)	48.11 (1.29)	5.42 (2.82)
H5 (Kidney)	18	38.06 (0.79)	73.34 (7.27)	35.29 (8.06)
H6 (Kidney)	18	25.48 (0.26)	67.30 (2.58)	41.81 (2.84)
H4 (Liver)	18	121.77 (35.38)	74.41 (8.61)	-47.37 (43.98)
H8 (Liver)	18	54.75 (1.66)	106.62 (7.45)	51.87 (9.11)
H5 (BMT)	21	42.14 (0.67)	54.13 (1.70)	11.99 (2.37)
H7 (BMT)	21	33.08 (0.61)	211.20 (26.0)	178.12 (26.70)
H3 (Kidney)	21	43.50 (1.38)	51.96 (1.4)	8.46 (2.78)
H5 (Kidney)	21	40.43 (0.81)	72.53 (6.03)	32.10 (6.84)
H6 (Kidney)	21	28.83 (0.32)	71.01 (2.6)	42.18 (2.92)
H4 (Liver)	21	116.48 (26.53)	78.30 (8.6)	-38.18 (35.13)
H8 (Liver)	21	54.92 (1.44)	106.71 (6.40)	51.78 (7.84)

Inside the parentheses are standard errors.

The results in Table 8 are similar to our earlier results, although the magnitudes differ. For one of the hospitals, ϕ_4 is estimated to be negative but is not significant. The fact that our global estimator (MLE) and local estimator (kernel) lead to similar conclusions is reassuring.

To sum up, an overall picture that consistently appears in all results is that hospitals tend to submit much larger bills when marginal reimbursement rates are higher. The exact magnitude differs across organ types, hospitals, and specifications, but when the marginal reimbursement rate jumps from 0% to approximately 50%, the marginal increase in hospitals' expenditures for a given increase in patients' illness severity becomes 2 to 14 times larger.

7 Conclusion

In this paper, we propose a modified RDD estimator that will be consistent when the forcing variable can be manipulated by agents. We ground our RDD-style estimator in economic theory and provide primitive economic assumptions that guarantee the consistency of our estimator. Our proposed estimator can be applied to many interesting settings that have been considered to be outside of the RDD framework. For instance, it is not possible to use standard RDDs to recover consumers' price sensitivity in the presence of non-linear budget constraints or workers' labor supply elasticity in the presence of higher marginal tax rates for higher tax brackets, because agents optimally choose their forcing variable. Our paper shows that a modification to the standard RDDs allows us to consider these types of problems within an RDD-style framework.

The assumptions required for our estimator are unlikely to hold for all settings, and thus it is important for researchers to examine whether the assumptions hold for their problems of interest. A key assumption is the strict monotonicity between the type and the dependent variable. This is likely to be violated if the type is multi-dimensional or if there is optimization error. In future work, we plan to investigate the performance of our estimator under these more general conditions and improve our estimator to make it robust against these complications.

References

[1] **Arrow, Kenneth**. 1963. "Uncertainty and the Welfare Economics of Medical Care." *American Economic Review,* 53: 941-973.

[2] **Chernozhukov, Victor**. 2005. "Extremal Quantile Regression." *The Annals of Statistics*, 33(2): 806-839.

[3] **Chernozhukov, Victor and Han Hong**. 2004. "Likelihood Inference for a Class of Nonregular Econometric Models." *Econometrica*, 72(5): 1445-1480.

[4] **Cutler, David**. 1995. "The Incidence of Adverse Medical Outcomes Under Prospective Payment." *Econometrica*, 63(1): 29-50.

[5] **Dafny, Leemore**. 2005. "How Do Hospitals Respond to Price Changes?" *American Economic Review*, 95(5): 1525-1547.

[6] **Donald, Stephen and Harry Paarsch**. 1996. "Identification, Estimation, and Testing in Parametric Empirical Models of Auctions within Independent Private Values Paradigm." *Econometric Theory*, 12: 517-567.

[7] **Dranove, David and Paul Wehner**. 1994. "Physician-Induced Demand for Childbirths." *Journal of Health Economics,* 13(1): 61-73.

[8] **Fuchs, Victor**. 1978. "The Supply of Surgeons and the Demand for Operations." *Journal of Human Resources*, 13: 121-133.

[9] **Gaynor, Martin and Paul Gertler**. 1995. "Moral Hazard and Risk Spreading in Partnerships." *RAND Journal of Economics*, 26(4): 591-613.

[10] **Gaynor, Martin, James Rebitzer and Lowell Taylor**. 2004. "Physician Incentives in Health Maintenance Organizations." *Journal of Political Economy*, 112(4): 915-931.

[11] **Gruber, Jonathan and Maria Owings**. 1996. "Physician Financial Incentives and Cesarean Section Delivery." *Rand Journal of Economics*, 27(1): 99-123.

[12] **Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw**. 2001. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica*, 69(1): 201-209.

[13] **Hodgkin, Dominic and Thomas McGuire**. 1994 "Payment Levels and Hospital Response to Prospective Payment." *Journal of Health Economics*, 13: 1-29.

[14] **Imbens, Guido and Thomas Lemieux**. 2008. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics*, 142(2): 615-635.

[15] **Koenker, Roger and Gilbert Bassett**. 1978. "Regression Quantiles." *Econometrica*, 46: 33-50.

[16] **Lee, David and Thomas Lemieux**. 2010. "Regression Discontinuity Designs in Economics." forthcoming in *Journal of Economic Literature*.

[17] **McCrary, Justin**. 2008. "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test." *Journal of Econometrics*, 142(2): 698-714.

[18] **McGuire, Thomas**. 2000. "Physician Agency." in *Handbook of Health Economics,* eds. A. Cuyler and J. Newhouse, North-Holland, 467-536.

[19] **Milgrom, Paul and Chris Shannon**. 1994. "Monotone Comparative Statics." *Econometrica*, 62(1): 157-180.

[20] **Topkis, Donald**. 1978. "Minimizing a Submodular Function on a Lattice." *Operations Research*, 26: 305-321.

[21] **Urquiola, Miguel and Eric Verhoogen**. 2009. "Class-Size Caps, Sorting, and the Regression-Discontinuity Design." *American Economic Review*, 99(1): 179-215.