Did Shareholders Affect the Credit Rating Decision? Evidence from a Semiparametric Ordered Model with Marginal Effects Analysis

Yixiao Jiang (Ethan)*

August 30, 2017

Abstract

This paper investigates whether firms with strong shared-ownership relation with the credit rating agency (CRA) benefit from higher credit ratings relative to firms with weaker relation with the CRA. We propose a semiparametric ordered response model in which the ratings are flexibly affected by firm/investment specific characteristics and a debt issuer’s shared-ownership relation with the rating agency. To capture the heterogeneous effect of ownership stake on ratings, this paper features estimation of the quantile-level average marginal effects and derive asymptotic results for this marginal effect estimator. We test hypotheses regarding whether the CRAs had compromised their impartiality due to possible conflicts of interest in the context of Moody’s. Using data on initial ratings from the Mergent’s Fixed Income Securities Database (FISD) from 2001-2007, we find ratings are improperly inflated on medium credit quality bonds, especially those issued by firms affiliated with large shareholders of the CRA. However, the findings of this paper suggest that when rating agencies are publicly held by only diffuse owner, the credibility of ratings remains intact.

Keywords: Semiparametric ordered response model, Marginal effects, Credit rating agencies, Conflict of interests.

JEL Classification: C25, G24

*Department of Economics, Rutgers University, 75 Hamilton St, New Brunswick, NJ 08854. I thank my advisor, Roger Klein, for his advising and comments. I’m also grateful to Hilary Sigman, Bruce Mizrach, Zhutong Gu, Shuyang Yang, Celso Brunetti, Yuliyan Mitkov, Tong Xu (discussant) and conference participants for helpful discussion. I gratefully acknowledge financial support provided by the graduate school and economics department of Rutgers University and the Chinese Economist Society. All errors remain mine.
1 Introduction

A number of provisions of the Dodd-Frank Act are directed at mitigating actual or perceived conflicts of interest faced by the Credit Rating Agencies (CRA), as Joseph Stiglitz labeling them one of the “key culprits” of the recent financial crisis. This paper investigates a prevalent, but sparsely studied form of conflict of interests, which arises when the raters and issuers engage in a shared-ownership relation. That is, when the rater and issuer are controlled by a common institutional investor.

Many CRAs today are owned by large financial institutions who simultaneously control firms that the CRAs rate, and this shared ownership relation has been assumed to exacerbate conflicts of interest. In the case of Moody’s, the leading rating agency in the U.S., there is a potential rating bias after it went public in 2001 because Moody’s rating decision might be affected by the economic interest of its shareholders. Surprisingly about 80% of bonds Moody’s rated in this period are issued by firms associated with Moody’s shareholders. Due to the increasing competition within the credit rating industry, the threat of losing business to their competitors has tilted the balance away from an independent risk arbiter towards a facilitator of risk transfer (Financial Crisis Inquiry Commission, 2011). The key question we want to examine is: How does a bond issuer’s relation with the rating agency affect this CRA favoritism? if any.

One empirical challenge in the bond rating literature is to formulate and estimate a model that does not overly restrict the manner in which the ratings depends on the explanatory variables, primarily because the rating procedure is not observed by the public. This paper proposes a new econometric framework to understand the rating process. Specifically, we define a latent variable $y^*$ as the dependent variable of theoretical interest (e.g., default risk), which is unobserved but presumably determined by a series of risk predictors satisfying an index restriction similar to Ichimura and Lee (1991) and Klein and Sherman (2002). The actual ordinal ratings are then assigned based on the relation between $y^*$ and a set of unknown threshold points between rating categories. For the models for which we found predictive results, our model outperforms them in terms of goodness-of-fit. Compared to extant parametric models, the semiparametric approach has two main advantages.

First, the semiparametric model imposes fewer functional and distributional restrictions on both explanatory variables and the error term, so it is more robust to misspecification errors. The main-
stream literature uses either ordered-probit/logit model (Blume et al., 1998, Kaplan and Urwitz, 1979, West, 1970), or the linear probability model (LPM) to model the rating process (Campbell and Taksler, 2003, Jiang et al., 2012, Kedia et al., 2016). These methods leverage strong assumptions on both the linearity of covariates as well as the additivity of the error term. The linearity assumption, in the context of bond ratings, is hard to justify in practice. As mentioned in a research manual by Moody’s analysts (Crosbie and Bohn, 2003), an investment’s default risk \( y_i^* \) can be modeled as the product of probability of default and loss given default\(^1\). This indicates factors that affect the default probability, such as a firm’s leverage ratio, and factors that affect loss given default, such as the issuing amount, may affect credit rating in a non-separable fashion. The semiparametric estimator proposed in this paper takes this potential non-separability into account by not specifying a priori the way ratings are affected by explanatory variables.

Second, and more importantly, the semiparametric model permits non-constant marginal effects. In this study, we construct a variable termed Moody-Firm-Ownership-Interaction (MFOI) to characterize the shared-ownership relation between a bond issuer and Moody’s, and we identify rating bias as the marginal effect of MFOI. By construction, a bond issuer associated with larger MFOI has a stronger connection with Moody’s through Moodys’ shareholders. In our dataset, most issuers are connected with Moody’s through minority shareholders, while others are subsidiaries of important shareholders such as Warran Buffett who owns 12% of Moody’s stock on average. To address the concern that bond issuers are heterogeneous in their shared-ownership relation with Moody’s, we estimate the average marginal effects of MOFI at different quantile level of MOFI. That is, if firms with high MFOI are rated higher than firms with low MFOI \( \text{ceteris paribus} \), we conjecture that Moody’s might have a upward bias towards firms that are related with Moody’s large shareholders. The semiparametric approach enables us to capture the variation of marginal effects across the distribution of MFOI, while parametric models in general are not capable of.

Using data on initial ratings from the Mergent’s Fixed Income Securities Database (FISD) from 2001-2007, we test hypotheses regarding whether the CRAs had compromised their impartiality due to possible conflicts of interest. To be more specific, we estimate a model where the ratings are

---

\(^1\)To be specific, default probability means the probability that the counterparty or borrower will fail to service obligations and loss given default stands for the extent of the loss incurred in the event the borrower or counterparty defaults. Crosbie and Bohn (2003) list other types of risk as well. For example, correlation of defaults also affects the default risk of structured products.
allowed to be affected by firm characteristics, bond characteristics and MFOI, the aforementioned proxy for shared-ownership relation. Our results indicates that it is 14% more likely for bonds issued by firms related with Moody’s large shareholders to receive higher ratings relative to other bonds with similar firm risk attributes and issue characteristics. The magnitude of this effect is close to what Kedia et al. (2016) found. However, we don’t find such preferential treatment on firms related to Moody’s small shareholders. In addition, we found high-yield bonds issued by any firms, regardless of their shared-ownership relation with Moody’s, do not receive higher ratings.

Our first finding is in accord with the literature on the role of large shareholders in corporate governance (Admati and Pfleiderer, 2009, Edmans, 2009). That is, large shareholders may utilize their governance power and/or threat of exit to extract private benefits. Consequently, only firms related with Moody’s large shareholders received favorable treatment. The second finding is also credible. Bolton et al. (2012) argues that when a CRA decides whether or not to assign a inflated rating, it compares the foreseen revenue generated by the inflated rating versus the “reputation loss” if the underlying debt turns out default. Since a low quality bond is more likely to default, implying a higher probability to trigger the reputation loss (which means that the CRA will “get caught” for assigning inflated ratings), the CRA should be less likely to assign inflated ratings on low quality bonds.

Following the recent financial crisis, the literature on conflicts of interest in credit rating agencies has focused on the “issuer-paid” model\(^2\). There is a vast empirical literature concerning with the effect of issuer-paid models on rating qualities (Jiang et al., 2012, Kraft, 2015, Mathis et al., 2009). Recently, several studies in this stream of literature also document a “premium” on ratings for firms that share a particular form of relation with the rating agency. For example, Baghai and Becker (2016) discovers a similar preferential treatment to issuers that have non-rating business relation with the CRA using a different dataset in India. The closest paper related to ours is Kedia et al. (2016) (KRZ henceforth), in which the authors find that Moody’s assigned higher ratings to firms that are associated with its long term large shareholders. Our study differs from KRZ in three respects:

\(^2\)See Bolton et al. (2012), Goel and Thakor (2015) and Camanho et al. (2009) for theoretical discussion on the issuer-paid model
(i) We construct a continuous variable (the aforementioned MFOI) to capture the shared-ownership relation between a bond issuer and the CRA using shareholding data. In contrast, KRZ defines a dummy variable which takes value one if a bond issuer is invested by either Berkshire Hathaway or Davis Selected Advisers, which are the two shareholders that consistently own at least 5% of Moody’s, and zero otherwise. In their model, bond issuers are partitioned into two groups: “Moody related” firms and others. However, in our sampling period, Moody’s has an average of 360 shareholders every quarter, each of whom has a different ownership stake and influential power in Moody’s. Considering the complex structure of this shareholder network, one should suspect there is still significant heterogeneity in terms of ownership status in the these two groups. Instead of categorizing the bond issuers into two groups, we characterize each bond issuer's connectivity with Moody’s in a continuum based on its affiliation with all Moody’s shareholders.

(ii) In absence of the actual default rate, the standard approach to identify the rating bias is to compare the CRA’s ratings with either the rating from another CRA or some indicator for firm’s default risk. KRZ uses a differences-in-differences approach to identify the rating bias: for each bond, they examine the difference in ratings from S&P’s and Moody’s and across different shared-ownership relation. When S&P’s ratings are used as the benchmark, a higher rating from Moody’s somehow reveals favorable treatment. However, S&P’s ratings could be more stringent than Moody’s in a systematic way so that a higher rating from Moody’s does not necessarily imply rating biases. To avoid this critique, this paper evaluates rating quality by directly examining the marginal effect of MFOI, and hence avoid using any external benchmark whose validity might subject to critique.

(iii) Methodologically our econometric model is more flexible. In contrast to the linear model in KRZ, the marginal effects in our specification are allowed to be affected by both a issuer’s shared-ownership status and quality of the underlying bond. By comparing the marginal effects of MFOI for issuers with high MFOI (those who are invested by large shareholders of Moody’s) and issuers with low MFOI (those who are invested by small shareholders of Moody’s), we

---

Footnote: One of this measure is the Expected-Default-Frequency (EDF), which is computed using firm’s financial data. See Duffie et al. (2007) and Crosbie and Bohn (2003)
reach a similar conclusion as KRZ. In addition, we find this differential effect reverse signs for low quality bonds. This indicates that the degree of rating bias is also affected by bond quality as well.

Our results are robust to several alternative specifications. To address the concern that the main variable of interest MFOI is constructed, we proxy the share-ownership relation between bond issuers and the CRA by several other variables such as the number of common shareholders. To check for omitted variable bias, we use the rating difference between Moody’s and S&P’s as the dependent variable. We do not find the results vary significantly in either case.

The rest of this paper is organized as follow: Section 2 introduces the data and develops hypotheses of interest. In Section 3 we describe the semiparametric rating model, motivates marginal effects, and discuss the estimation strategy. Section 4 compares the semiparametric model with various parametric models in terms of in-sample fitness. Section 5 presents the estimation results for the parameter of interests and marginal effects. Section 6 provides some robustness check and section 7 concludes. The Appendix will contain all the supplemental definition, assumption and asymptotic results, along with their proofs.

2 Data and Context

We examine the impact of a issuer’s shared-ownership relation with the CRA on ratings in the context of Moody’s, the leading credit rating agency in the U.S. After the collapse of AAA bonds during the recent financial crisis, credit rating agencies have come under increased public scrutiny. In the case of Moody’s, after it went public in 2001, there is a potential rating bias because Moody’s rating decision might be affected by the economic interest of its shareholders.

The literature on conflicts of interest in credit rating agencies is quite ample. Much of academic debate has focused on the conflict of interest inherent in the issuer-pay model. This model and its equilibrium structure are studied by Bolton et al. (2012) and Sangiorgi et al. (2009). The conflict of interests can be characterized as a “trade-off” between providing accurate (and hence unflattering) ratings versus exaggerating on the investment but risking a potential loss on their reputation. The reputation cost to the CRA is Since rating agencies care about their reputation, the CRA might
take more prudential actions when rating lower grade investments, as the potential reputation cost of overrating a bad investment is high (Bolton et al., 2012). In this regards, we conjecture that even a catering rating agency might be less biased when rating high-yield bonds, as the realized benefit of catering cannot outweigh the potential reputation cost.

**Hypothesis 1.** For a given issuer, would the sign and degree of rating bias differ for bonds with different quality?

The degree of bias might also be affected by a issuer’s shareholding relation with Moody’s. Kedia et al. (2016) point out that firms that are related to Moody’s large shareholders received significant higher ratings from Moody’s than S&P, and this rating difference cannot be explained by Moody’s private information. Given the vast literature on the role of large shareholders (Admati and Pfleiderer, 2009, Edmans, 2009, Shleifer and Vishny, 1986), Moody’s might exert more upwardly biased ratings to investments that are related to their large shareholders, while being absolutely impartial or even harsher to regular customers. Motivated by the above empirical concern, the first hypothesis we want to test is

**Hypothesis 2.** Ceteris paribus, is Moody’s appear be more biased towards issuers related with its large shareholder?

Our data derive from multiple sources. First, we obtain ratings on corporate bonds issued by firms covered in either CRSP or Compustat, along with a series of bond characteristics, from Mergent’s Fixed Income Securities Database (FISD). Since fixed effects in a nonlinear model cannot be easily “differenced out”, we obtain only initial ratings so that each bond appears in our dataset only once. Our sampling period goes from 2001, when Moody’s went to public, to 2007 to prevent any confounding effect of financial crisis and other regulation act thereafter. For each quarter of this period, we also obtain a series of issuer firm characteristics from CRSP-Compustat to match the rating data. In doing so, we obtain a pooled cross-section of data on initial ratings from Moody’s, consisting of 5913 bonds issued by 986 firms. Of the 5913 bonds, 3207 are also rated by S&P’s.

In the original rating scales, there are 21 rating categories from Aaa to C, descending in terms of credit qualities. Some categories are further partitioned into sub rating grades, e.g., Aa1, Aa2 and Aa3. To simplify exposition, we merge sub rating categories (Aa1-Aa3 into Aa). Bonds that are
rated at least Baa are classified as investment grade bonds, while others are classified as high-yield bonds. The distribution of initial ratings from Moody’s is shown in Figure 1.

Figure 1: Distribution of Ratings assigned by Moody’s

To identify each bond issuer’s shared-ownership relationship with Moody’s, we obtain shareholding data from Thomson-Reuters Institutional Holding (13F) Database. For each quarter in our sampling period, we first find the list of Moody’s shareholders and calculate their ownership stake in Moody’s (the percentage of Moody’s stock that they hold). We then access each of these shareholders’ investment portfolios to find out which bond issuers are also invested by the same shareholder. For each firm-shareholder pair, we calculate the issuer firm’s weight in the shareholder’s investment portfolio. To match the shareholding data to each bond issuers, we use the shareholder’s Manager Type Code (MGRNO) and firm’s CUSIP number.

Most importantly, we use shareholding data to construct measures of shared-ownership relation between each bond issuer and Moody’s. For the most part of this paper, we use a single variable, termed Moody-Firm-Ownership-Interaction (MFOI), to proxy the shared-ownership relation. Suppose Moody’s has \( j = 1, 2, \cdots, M \) shareholders at time \( t^4 \), and a issuing firm \( i \) can be invested by

\(^4\)Since all of the variable are time-specific, we drop the time \( t \) subscript for notational simplicity
any subset of those shareholders, we define MFOI in the following way:

\[ MFOI_i = \sum_{j=1}^{M} p_j \lambda_j \]  

(1)

with \( \lambda_j \) denotes shareholder j’s ownership take in the CRA, namely the percentage of the CRA’s owned by the shared owner j, and \( p_j \) denotes issuing firm i’s weight in shareholder j’s investment portfolio, namely the percentage of the shareholder’s portfolio that the issuing firm accounts for. We choose a product form because conflicts of interest would not be a concern if either piece is zero: that is, the bond issuer is not connected with Moody’s through shareholder j. It is important to note that by construction, a bond issuer associated with larger \( MFOI_i \) has a stronger connection with Moody’s through Moodys’ shareholders.

We plot the distribution of MFOI in Figure 2. Due to the fact that institutional investors hold very diverse portfolio, most \( p_j \) and \( \lambda_j \) take on very small values\(^{5}\), resulting the distribution of \( b_i \) extremely skewed to zero. As can be seen from the figure, about 20% of the bonds in our sample are issued by firms that are not affiliated with Moody’s at all. Most of the bonds are issued by firms related with Moody’s small shareholders. Only the top 5% of bonds are issued by firms with extremely large MFOI (those who are likely to related with Moody’s large shareholders).

\[ p = 0.25\%, \, \lambda = 0.07\% \] are the 75 percentile cutoffs respectively

Figure 2: Distribution of MFOI

---

\(^{5}\)
Since MFOI is constructed based on our economic knowledge, one shortcoming of using MFOI directly as an explanatory variable is the regression results may be subject to measurement error bias. To mitigate this concern, we use three other variables that are more observable to proxy the shared-ownership relation. Table 1 provides summary measures on the variables we use to characterize a bond issuer’s shared-ownership relation with Moody’s. First, we use the number of common shareholders of Moodys and the issuer firm who owns at least 5% of Moodys’ stock (num_largeSH). There are only 4 large shareholders of Moody’s appear in our sampling period (Berkshire Hathaway, Goldman Sachs, Davis Selected Advisers and Barclays Bank PLC), and the most popular bond issuer is simultaneously invested by three out of this four. Second, we use the number of common shareholders of Moodys and the issuer firm. (num_SH). Moody’s has an average of 360 shareholders in our sampling period (See Table 1 in Kedia et al. (2016)). An average bond issuer are invested by 40 shareholders of Moody’s. Third, we use the total percentage of Moody’s stock owned all common shareholders (overlapping share).

In choosing variables that affect credit ratings, we focus both on the characteristics of the issuer firm and characteristics of the underlying bond. The summary statistics of these measures are provided in Table 2. All financial ratios were computed using a 5-year arithmetic average of the annual ratios, as Kaplan and Urwitz (1979) points out that bond raters might look beyond a single year’s data to avoid temporary anomalies. The explanatory variables are: (1) Firm leverage, defined as the ratio of long-term debt to total assets (Leverage ratio). (2) Operating performance, defined as operating income before depreciation divided by sales (Profitability). (3) Issue size, defined as the par value of the bond issue (Issuing amount). (4) Issuer size, defined as the value of the firm’s total asset (Asset) and (5) Subordination status, a 0-1 dummy variable which is equal to one if the bond is

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>overlapping share</td>
<td>0.197</td>
<td>0.102</td>
<td>0</td>
<td>0.794</td>
</tr>
<tr>
<td>num_largeSH</td>
<td>0.606</td>
<td>0.706</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>num_SH</td>
<td>39.394</td>
<td>20.417</td>
<td>0</td>
<td>106</td>
</tr>
<tr>
<td>MFOI</td>
<td>0.005</td>
<td>0.004</td>
<td>0.000</td>
<td>0.037</td>
</tr>
</tbody>
</table>

We follow Kedia et al. (2016) and pick 5% as the threshold point.

MSDW, Sands Capital Management, and Harris Associates L.P. used to own more than 5% of Moody’s stock for 1 or 2 quarters in this 7 year period. Considering their relation with Moody’s is short-term, we do not classify them as large shareholders.
a senior bond (Seniority dummy). (6) Stability variable (Variance of asset), defined as the variance of the firm’s total asset in the last 16 quarters.

Table 2: Firm and Bond Characteristics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASSET</td>
<td>log(asset) of the issuer</td>
<td>9.643</td>
<td>2.280</td>
<td>4.360</td>
<td>14.324</td>
</tr>
<tr>
<td>STABILITY</td>
<td>Variance of asset</td>
<td>0.230</td>
<td>0.169</td>
<td>0.003</td>
<td>1.416</td>
</tr>
<tr>
<td>LEVERAGE</td>
<td>Firm leverage ratio</td>
<td>0.264</td>
<td>0.178</td>
<td>0.002</td>
<td>1.212</td>
</tr>
<tr>
<td>PROFIT</td>
<td>Operating performance</td>
<td>0.026</td>
<td>0.058</td>
<td>-0.739</td>
<td>0.436</td>
</tr>
<tr>
<td>AMT</td>
<td>log(issuing amount)</td>
<td>12.224</td>
<td>1.681</td>
<td>2.708</td>
<td>19.337</td>
</tr>
<tr>
<td>SENIORITY</td>
<td>a bond’s subordination status</td>
<td>0.809</td>
<td>0.393</td>
<td>0.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

3 Econometric Strategy

3.1 Model

We consider the following generic model for the rating process:

\[ y_i^* = g(F_i, B_i, Z_i, U_i) \] (2)

where \( y_i^* \) is a bond’s default risk, which is unobserved; \( F_i \) (\( B_i \)) denotes a vector of publicly observed issuer (bond) specific characteristics; \( Z_i \) represents a vector capturing the shared-ownership relation between the issuer and rating agency (henceforth, the “CRA-issuer relation”), which does not vary at bond level; and \( U_i \) denotes other unobserved factors that may potentially affect ratings. The unknown function \( g(\cdot, \cdot, \cdot, \cdot) \) represents the way rating agencies estimate a bond’s default risk based on public available as well as their private information.

The public observes an ordinal rating \( Y_i = 1, 2, 3 \ldots L \), which are assigned based on the magnitude of \( Y_i^* \) and a series of cut off points \( c_j \) between rating categories:

\[ Y_i = \sum_{j=1}^{L} j \mathbb{1}\{ c_{j-1} < y_i^* < c_j \}, \quad \mathbb{1}\{ \cdot \} \text{ is a indicator function} \] (3)

These cutoff points may be fixed points. Alternatively, these cutoff points may be random variables from different distributions that are independent of the explanatory variables. The estimator em-
ployed in this paper allows for either possibility.

Defining $X_i = (F_i, B_i, Z_i)$, we might write $Prob(Y_i \leq K|X_i) = F(X_i)$ in a general nonparametric formulation. This specification imposes few restrictions on the form of the joint distribution of the data, so there is little room for misspecification, and consistency of the estimator is established under much more general conditions than for parametric modeling. However, when the dimension of $X$ is large, it is well known that the resulting estimator end up having huge variance due to the “curse of dimensionality”. In order to estimate the above conditional probability well with a moderately sized sample, we make the following index assumption:

**Assumption 1. (Index Assumption)** There exists a firm aggregator (or index) $V_F \equiv F_i \beta_F^0$, a bond aggregator $V_B \equiv B_i \beta_B^0$, and a differentiable function $H(\cdot, \cdot, \cdot)$, such that for all category $K$:

$$Prob(Y_i \leq K|X_i) = Prob(Y_i \leq K|V_F, V_B, Z_i) = H_K(F_i \beta_F^0, B_i \beta_B^0, Z_i)$$  \hspace{1cm} (4)

Assumption 1 states the effects of the model’s explanatory variables $X_i$ on ratings operate through three channels: a firm index $V_F = F_i \beta_F^0$, a bond index $V_B = B_i \beta_B^0$, and the CRA-issuer relation $Z_i$. The function $H(\cdot, \cdot, \cdot)$ in turn denotes the nonparametric mapping of the three indices to the probability that a bond will be rated at least category $K$. Being nonparametric, the mapping $H(\cdot, \cdot, \cdot)$ allows for potential non-linearities, non-monotonicities as well as fully flexible interactive among indices in determining the rating assignments. From an ex ante perspective, all of these features may be desirable.

Since this function $H_K$ is not parametrically specified, identification of the index parameter $\beta_0$ is up to any multiplicative and additive constant, or the so-called identification up to location and scale. To be precise, from now on we rewrite $V_F = F_1 + F' \theta_F^0$ and $V_B = B_1 + B' \theta_B^0$ as function of the identified parameter $\theta_0 \equiv [\theta_F^0, \theta_B^0]$, where $F_1(B_1)$ is the firm (bond) characteristic that we choose to normalize on. With a slight abuse of notation, we may also use $F_1 \theta_F^0, B_1 \theta_B^0$ to stress the dependence of index on the identified parameter $\theta_0$.

This multiple-index specification arises naturally in many applications where a single index model cannot fully capture the underlying economic behaviors. Such example includes sample selection.

---

8Note that (1) the “index parameter” $\beta_0$ can be identified only up to location and scale. (2) we state the index in a linear form only for the purpose of exposition, but the index need not be linear.
(Heckman, 1977), extraneous variable (Stoker, 1986) and decision-making with multiple players (Lührmann and Maurer, 2008). In our context, there are evidence suggesting indices could enter the rating model in a non-additive way. As mentioned in a research manual by Moody’s analysts (Crosbie and Bohn, 2003), an investment’s default risk $Y_i^*$ is primarily driven by two types of risk: *probability of default* and *loss given default*\(^9\). The first type of risk captures the likelihood of default for an issuer, and thus is only affected by the issuer fundamentals $F_i$. The second risk is driven by the “type” of the bond, namely what $B_i$ represents\(^10\). Clearly these two risk seems affect the rating in a product form rather than linear, which motivates the potential non-separability of $H(\cdot, \cdot, \cdot)$.

We make additional assumptions to deal with the reality of data. Although our sample spans several years, the same bond is never observed twice because we use only initial ratings (e.g., we do not model rating upgrades/downgrades on the same bond). Therefore all bonds can be treated as independent statistical units that are observed at different points in time (Assumption 2). Another complication is in our sample is there are multiple bonds issued by the same firm both within or across different time periods. To deal with the with-in firm correlation for those bonds, we allow the error term $U$ to be clustered (Assumption 3).

**Assumption 2. (Pooled cross-section data)** $(Y_i, X_i)$ are obtained by collecting random samples from a large population of firms indexed by $g \in \{1, 2, \ldots N_F\}$ at different points in time. $(Y_i, X_i)$ are independent of each other across firms.

Assumption 2 states that bonds issued at different point in time are treated as distinct statistical units, even though they may be issued by the same firm. Therefore we suppress the time subscript $t$ and assume the rating $Y_i$ is solely driven by the contemporaneous components. To accommodate the fact that raters might look beyond a single year’s data, the covariates are computed using a 5-year arithmetic average. This assumption also implies that serial correlation of residuals is not an issue, when regression analysis is applied.

**Assumption 3. (Clustering)** For any issuer firm $g \in \{1, 2, \ldots N_F\}$, the error term $U_i$ does not have

\(^9\)To be specific, *default probability* means the probability that the counterparty or borrower will fail to service obligations and *loss given default* stands for the extent of the loss incurred in the event the borrower or counterparty defaults. Crosbie and Bohn (2003) list other types of risk as well. For example, correlation of defaults also affects the default risk of structured products.

\(^10\)For example, if a bond is classified as “senior debt”, then bondholders have priority to claim their money back before other regular debt holders once the issuer defaulted. For them, the *loss given default* is much lesser.
to be independent with $X_i$ as long as Assumption 1 holds. Moreover, for bonds issued by a firm are allowed to be correlated:

$$E(U_i U_j) = \begin{cases} 
\sigma_{ij} & \text{if } i, j \in g \\
0 & \text{otherwise}
\end{cases}$$

Assumption 3 makes our model more appealing and somehow robust to the data reality, because observations are very unlikely to be iid in practice. In this credit rating context, there could be firm level shocks that apply to all bonds issued by the same firm, resulting mechanical within-firm correlations. To accommodate common shocks, we assume to have a clustered sample with large number of firms and tolerate correlation within each firm. Strict exogeneity on explanatory variables is often a strong condition, especially so in our case that the CRA-issuer relation $Z_i$ and the ratings might be determined almost simultaneously. Our model does not require the error term to be independent from explanatory variables. Instead, identification of the model leverages purely on the index assumption 4, which is slightly weaker than assuming strict exogeneity.

**Remark 1.** (Endogeneity) Assumption 3 raises a question on the applicability of our model when explanatory variables are not strictly exogenous. Inarguably there are cases in which the endogeneity issue will indeed defy the index assumption. Although beyond the scope of this paper, one might deal with the potential endogeneity issue more directly. As illustrated in Blundell and Powell (2004), if a reduced-form model for $Z_i$ can be formulated with appropriate exclusive restrictions, then the estimated reduced form error term can be utilized as “control variables” for the endogeneity of $Z_i$. However, finding an instrument is difficult empirically and will exacerbate the curse of dimensionality problem\(^1\), so we resort to the index assumption to alleviate this issue.

### 3.2 Hypothesis Testing Based on Marginal Effects

Recall that the central goal of this paper is to study the impact of CRA-issuer relation on ratings. It is important to notice that under the index restriction (4), any conflict of interests faced by the CRA will affect the rating procedure through the marginal impact of $Z$. Therefore, one may test hypotheses regards rating bias by examining the marginal effects of $Z$. More formally, for any rating category $K$,

\(^{11}\)Suppose we have an instrument and use the methods proposed by Blundell and Powell (2004), the estimated reduced form error term will enter the model in (4) nonparametrically as another index. This is not impossible to handle, but would require some extra treatment when we estimating the object in (4).
the marginal effect ($ME$) of $Z_i$ from a base-level $z^b$ to $z^b + \Delta^{12}$ corresponds to the increments in (4):

$$ME_i(F_i, B_i, z_b; K, \theta_0) \equiv H_K(V_F, V_B, z^b + \Delta) - H_K(V_F, V_B, z^b),$$

(5)

In calculating the marginal effects as (5), we allow the responsiveness of rating from a strengthening CRA-issuer relation to depend on the firm/bond fundamentals. This extra degree of flexibility enables us to capture the heterogeneous marginal effects in a way that parametric models cannot.

To draw economic inferences and test our hypotheses of interest, we average the marginal effect, defined in (5), at a particular set rather than the entire sample. The set of interest is defined by the quantile levels of one or multiple $X_i$.

More formally, let $t(q)$ be a quantile trimming function indicating whether $Z_i$ belongs to a particular quantile $Z_q$. We define the “Quantile Marginal effects” (QME) as:

$$\mu^K(Z_q, \theta_0) \equiv E[ME_i(F_i, B_i, z_b; K, \theta_0)|Z \in Z_q]$$

$$= E[t_i(q)ME_i(F_i, B_i, z_b; K, \theta_0)] / E[t_i(q)]$$

(6) (7)

This notion of Quantile Marginal Effects is a natural generalization of Average Marginal Effects (AME) when one suspects the influence of explanatory variables are sufficiently heterogeneous. AME is often of primary interest in cases when policy-makers are interested in the general effectiveness of policy instead of at a particular point of region. That is, if the average marginal effect of $Z$ is small and insignificant, the credibility of the CRA as a independent and impartial risk arbiter seems confirmed. However, in cases when the impact of $Z$ is sufficiently nonlinear, solely examining AME might lead to incorrect inference. For example, suppose $Z_i$ has a negative (positive) effect on ratings when $Z_i$ is small (large), AME could conceivably be close to zero even though the real effect of $Z_i$ is clearly not. The limitation of AME in capturing heterogeneous effects leads to consideration of QME.

To test Hypothesis 1 - For a given issuer, would the sign and degree of rating bias differ for bonds with different quality? we are interested in the variation of marginal effects for bonds from different

---

12 The choice of $\Delta$ is completely arbitrary, usually $\Delta = 1$ for discrete regressor and one standard deviation for continuous regressor.
categories. Since the marginal effect can be calculated for any category K, it would be natural to hold
the quantile level fixed, and test whether the quantile marginal effects vary with the category level
K. To test Hypothesis 2 - *is Moody’s more biased towards issuers related with its large shareholder?*
we are only interested in the impact of $Z_i$ on ratings for issuers with a large $Z_i$. In this context, such
an impact can be assessed by averaging the marginal effects of $Z_i$ conditioning on a high quantile of
$Z_i$. For the purpose of testing Hypothesis 2, studying the size of $QME$ and its variation across $Z_q$
clearly leads to a more direct test.

It is important to note that these test strategies would not work in ordered-probit or linear
probability models. In linear probability models, the marginal effects of $Z_i$ is captured by the
corresponding regression coefficient, therefore there would be no variation across rating categories
or the level of $Z_i$. In ordered probit models, the variation of marginal effect of $Z_i$ is purely driven
by the cumulative distribution function of the error term. Consequently any misspecification of the
distribution will bias the testing results.

### 3.3 Estimation Strategy and Asymptotic Results

In this section, we briefly discuss the estimation strategy for quantile marginal effects and the asymp-
totic properties of the proposed estimator.

Evaluation of the $QME^K_q$ requires estimates of the index coefficients $\theta_0 \equiv [\theta^F_0, \theta^B_0]$, the quantile
trimming function $t_i(q)$ and most importantly, the structural equation $H_K(\cdot, \cdot, \cdot)$. Given the index
structure in equation (4), we employed the semiparametric maximum likelihood procedure proposed
in *Klein and Sherman (2002)*. Defining $P_i^k(\theta_0) \equiv \mathbb{E}(Y^k_i = 1 | X_i) = H_k(V_F, V_B, Z_i) - H_{k-1}(V_F, V_B, Z_i)$
for any rating notch k, the likelihood function can be written as:

$$Q(\theta) = \sum_{i=1}^N \tau(X_i) \{ \sum_{k=1}^L \{ Y_i = k \} \ln(P_i^k(\theta_0)) \}$$

(8)

where $\tau$ is a trimming function that removes observations with poorly estimated conditional prob-
abilities, which is defined formally in the appendix. Estimates of the index parameter $\theta_0$ will be
obtained by maximizing the above likelihood function:

\[
\hat{\theta} = \arg\max_{\theta} \hat{Q}(\theta) \equiv \sum_{i=1}^{N} \hat{\tau}(X_i) \left\{ \sum_{k=1}^{L} \{Y_i = k\} \ln(\hat{P}_i^k(\theta_0)) \right\} \tag{9}
\]

The object \(P_i^k(\theta_0)\), or the CRP differential, describes the conditional probability that a bond is assigned to category \(k\) given \(V_F, V_B\) and \(Z_i\). This object will be estimated nonparametrically. It is important to note that nearly every nonparametric estimator has bias, and its variances converges to zero slower than in the parametric case. While higher order kernels are often used in the literature as a bias control, they do not perform well in our case\(^{13}\). This is because the estimand \(P_i^k(\theta_0)\) is a (conditional) probability but higher order kernels, unlike regular kernels, can deliver estimated probability outside of \([0, 1]\). Therefore we estimate this object \(P_i^k(\theta_0)\) with a regular kernel first, and reduce the bias using a differences-in-differences approach proposed by Klein and Shen (2015). Details regarding the implementation of the estimator are provided in the appendix, with some standard regularity condition stated.

With the estimated index \(\hat{V}_{Fi} = F_i\hat{\theta}_F, \hat{V}_{Bi} = B_i\hat{\theta}_B\) and \(Z_i\), a “plug-in” estimator for \(QME^K_q\) is given as:

\[
\overline{QME}^K_q \equiv \frac{\sum_{i=1}^{N} \hat{t}_{qi} \overline{ME}_i(F_i, B_i, z_b; K, \theta_0)}{\sum_{i=1}^{N} \hat{t}_{qi}} \tag{10}
\]

where \(\overline{ME}_i(F_i, B_i, z_b; K, \theta_0) \equiv \overline{H}_K(\hat{V}_{Fi}, \hat{V}_{Bi}, Z_i + \Delta) - \overline{H}_K(\hat{V}_{Fi}, \hat{V}_{Bi}, Z_i)\) and the quantile trimming function \(\hat{t}_{qi} = \{Z_i \in Z_q\}\) ensures the average is taken over observations with \(Z\) in the quantile of interest. The asymptotic distribution of the estimator for \(\theta\) and \(QME^K_q\) is formally stated in Theorem B2 and B3 respectively in the Appendix.

\(^{13}\)We tried both smooth optimum kernels proposed by Muller (1984) and the twicing kernel by Newey et al. (2004) and found they do not perform well


4 Model Evaluation

Using the dataset on Moody’s initial ratings from 2001-2007, we first estimate the semiparametric model proposed in the previous section:

\[
Pr(Y_i = k|X) = E(R^k_i|V_F, V_B, MFOI) = P_k(V_F, V_B, MFOI) \tag{11}
\]

\[
V_F = ASSET + \theta_1^{F}STABILITY + \theta_2^{F}LEVERAGE + \theta_3^{F}PROFIT
\]

\[
V_B = AMT + \theta_1^{B}SENORITY
\]

with the following ordinal information on ratings:

\[
Y_i = \begin{cases} 
1 & \text{if the bond is rated as Aaa} \\
2 & \text{if the bond is rated as Aa} \\
\vdots \ & \text{...} \\
7 & \text{if the bond is rated as Caa}
\end{cases}
\]

Despite the generality with which this framework accounts for the influence of explanatory variables, the complexity of the estimation procedure raises the question of whether these features can be satisfactorily addressed by a simpler model.

In the empirical literature of credit ratings, the most prevailing approach is linear probability models (LPM, see Campbell and Taksler (2003), Jiang et al. (2012), Kedia et al. (2016) among other works), which assumes the observed ratings \( Y_i \) is a linear function on predictor variables \( X \) and error term \( U \), e.g., \( Y_i = X_i\beta_0 + U_i \). This approach ignores the discreteness in ratings and implicitly assumes, for example, Aa-grade bonds \( (Y_i = 2) \) are twice as likely to default as Aaa-grade bonds \( (Y_i = 1) \). Such an assumption is logically problematic, as rating scales convey only ordinal rather than quantitative relationship. Moreover, Hausman et al. (1992) argues that the linear probability model always deliver conditional distributions of dependent variable that is unimodal and have little weight in the tails. Therefore, in contexts where the focal interest is the conditional probabilities, as it is in our case, researchers switch to more advanced discrete-choice models. The second frequently employed approach is ordered probit/logit models (Blume et al., 1998, Kaplan and Urwitz, 1979,
Compared to our approach, this class of models assumes $y^* = X_i \beta_0 + U_i$. Hausman and his coauthors regarded ordered-probit as “a suitably extended version of LPMs” when the dependent variable is naturally discrete. However, ordered-probit model deliver consistent estimates only when the functional form on $y^*$ and the distributional assumption on $U$ are correctly specified. This concern leads to consideration of the semiparametric approach.

In this section, we run a horse-race between our semiparametric model and various parametric models in the literature as for in-sample fitness. In ordered models, usually fitness is measured by the percentage of being correctly predicted. The predicted rating is the category with highest conditional probability $Pr[Y^k_i = 1|X]$.

**Ordered-Probit vs Semiparametric** We also estimate the following ordered-probit model as a benchmark to compare with. The ordered-probit model replaces (11) with the following parametric structure:

$$
Pr(Y_i = k|X) = \begin{cases} 
\Phi(-ww_i) & \text{if } k = 1 \\
1 - \Phi(c_{L-1} - ww_i) & \text{if } k = 7 \\
\Phi(c_k^* - ww_i) - \Phi(c_{k-1}^* - ww_i) & \text{otherwise}
\end{cases}
$$

where

$$ww_i = \alpha_0 X_i + c_1^*, \quad \alpha_0 = \frac{\beta_0}{\sigma}$$

The $\Phi$ function is the cumulative distribution function for $U$ in (2), and the $\beta$s can be consistently estimated through the standard maximum likelihood estimation.

After estimating the semiparametric model and the benchmark ordered probit model, we report the fitness for the two in Table 3, which is the traditional format of displaying the results of bond-rating predictions. The upper table corresponds to the semiparametric model with three indices and the lower table corresponds to the linear parametric probit model. Collectively the semiparametric model correctly predicts 68 % of bonds, which is 10 % higher than the standard parametric model with the same explanatory variables. In addition, the semiparametric model performs a better predictive power in all the rating categories, especially the Aaa, A and Ba grades.
Table 3: Predictions of New-Issues

<table>
<thead>
<tr>
<th>Actual ratings</th>
<th>Predicted Ratings (Semiparametric)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Aaa</td>
</tr>
<tr>
<td>Aaa</td>
<td>13</td>
</tr>
<tr>
<td>Aa</td>
<td>7</td>
</tr>
<tr>
<td>A</td>
<td>9</td>
</tr>
<tr>
<td>Baa</td>
<td>7</td>
</tr>
<tr>
<td>Ba</td>
<td>6</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
</tr>
<tr>
<td>Caa</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Actual ratings</th>
<th>Predicted Ratings (Linear ordered probit)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Aaa</td>
</tr>
<tr>
<td>Aaa</td>
<td>0</td>
</tr>
<tr>
<td>Aa</td>
<td>7</td>
</tr>
<tr>
<td>A</td>
<td>3</td>
</tr>
<tr>
<td>Baa</td>
<td>4</td>
</tr>
<tr>
<td>Ba</td>
<td>7</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
</tr>
<tr>
<td>Caa</td>
<td>1</td>
</tr>
</tbody>
</table>

Note: The upper table is the prediction result for the 3-index model and the lower table is the prediction result for the probit model with specification below
- 3-index model: 4009/5913 = 67.72 % correct
- Probit model: 3419/5913 = 57.82 % correct

Semiparametric vs Other Models the Literature  We further compare the semiparametric model with previous models in the literature of bond ratings and reports the results in Table 4. It is important to note that, for each previous work, the statistics on the percentage of correctly predicted is directly imported from the corresponding paper. Therefore, we view the comparison as suggestive in the sense that the dataset and the explanatory variables being used are different. In terms of percentage of correctly predicted, the semiparametric model outperforms West (1970), Horrigan (1966) and Blume et al. (1998). The semiparametric model has roughly the same predictive power with Kaplan and Urwitz (1979). However, KU made poor prediction on Aa Ba and B bonds, while the semiparametric model shows more robust predictive power across all rating categories.
Table 4: Comparison with previous models

<table>
<thead>
<tr>
<th>Study</th>
<th>Aaa</th>
<th>Aa</th>
<th>A</th>
<th>Baa</th>
<th>Ba</th>
<th>B</th>
<th>Caa</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>West (1970)</td>
<td>0.00</td>
<td>0.65</td>
<td>0.76</td>
<td>0.45</td>
<td>0.57</td>
<td>0.67</td>
<td>0.6234</td>
<td></td>
</tr>
<tr>
<td>Horrigan (1966)</td>
<td>1.00</td>
<td>1.00</td>
<td>0.71</td>
<td>0.53</td>
<td>0.64</td>
<td>0.4</td>
<td>0.5857</td>
<td></td>
</tr>
<tr>
<td>BLM*(1998)</td>
<td>0.26</td>
<td>0.36</td>
<td>0.74</td>
<td>0.54</td>
<td></td>
<td></td>
<td>0.5721</td>
<td></td>
</tr>
<tr>
<td>PM**(1975)</td>
<td></td>
<td></td>
<td></td>
<td>0.71</td>
<td>0.83</td>
<td>0.48</td>
<td>0.89</td>
<td>0.74</td>
</tr>
<tr>
<td>KU (1979)</td>
<td>1.00</td>
<td>0.22</td>
<td>0.92</td>
<td>0.47</td>
<td>0.00</td>
<td>0.00</td>
<td>0.6875</td>
<td></td>
</tr>
<tr>
<td>3-index</td>
<td>0.22</td>
<td>0.82</td>
<td>0.68</td>
<td>0.83</td>
<td>0.27</td>
<td>0.69</td>
<td>0.31</td>
<td>0.6772</td>
</tr>
</tbody>
</table>

* - In Blume et al. (1998), the authors estimate only the investment grade bonds using S&P’s rating
** - In Pinches and Mingo (1975), the authors use Multiple Discriminant Analysis (MDA) instead of regular regression

5 Results

I discuss results in three sections. First, I present the estimates for parameters in the structural index, $\theta$. I then present the estimates for average marginal effects, which summarizes the “general impact” of a bond/firm characteristics or CRA-issuer relation measure on ratings. As motivated in the previous section, the impact of CRA-issuer relation $Z_i$ can be fairly heterogeneous. Last but not least, I present the estimates for quantile marginal effects ($QME^k_q$) to pick up the heterogeneous impact. In the first two sections, I compare the estimation results from the semiparametric model with its convenient parametric counterpart (ordered-probit).

5.1 Parameter Estimates

Table 5 shows the estimates of parameters for the parametric (ordered-probit) and semi-parametric models. The estimated standard errors, which are included in the parentheses, are based on formulas derived from asymptotic theory\(^\text{14}\). I estimate the semiparametric model in Column 1, noting that each index is only identified up to location and scale. For normalization, I set the coefficients of ASSET and AMT equal to one. In Column 2 and 3, I present the coefficient estimates from the (parametric) ordered-probit models. As the CRA may systematically changes rating standards year by year (Blume et al., 1998), I allow year fixed effect in Oprobit1 (Column 2). Most structural

\(^\text{14}\)The standard errors of the coefficients in the probit model is obtained with White (1982)’s specification-robust method and are asymptotically correct even if the model is misspecified
Table 5: New Issues rated by Moody’s 2001-2007

<table>
<thead>
<tr>
<th></th>
<th>Semiparametric Model</th>
<th>Oprobit 1</th>
<th>OProbit2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Firm and Bond Characteristics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lasset</td>
<td>-0.647***</td>
<td>-0.609***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.013)</td>
<td>(0.019)</td>
<td></td>
</tr>
<tr>
<td>CVTA</td>
<td>-0.548***</td>
<td>0.350***</td>
<td>-0.540</td>
</tr>
<tr>
<td></td>
<td>(0.094)</td>
<td>(0.090)</td>
<td>(0.098)</td>
</tr>
<tr>
<td>leverage</td>
<td>-1.360***</td>
<td>1.930***</td>
<td>-2.984</td>
</tr>
<tr>
<td></td>
<td>(0.137)</td>
<td>(0.108)</td>
<td>(0.129)</td>
</tr>
<tr>
<td>profit</td>
<td>11.010***</td>
<td>-9.399***</td>
<td>14.529</td>
</tr>
<tr>
<td></td>
<td>(0.447)</td>
<td>(0.303)</td>
<td>(0.819)</td>
</tr>
<tr>
<td>offering_amt</td>
<td>0.043</td>
<td>-0.067</td>
<td>0.027</td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td>(0.010)</td>
<td></td>
</tr>
<tr>
<td>seniority</td>
<td>0.398***</td>
<td>-0.488***</td>
<td>0.754</td>
</tr>
<tr>
<td></td>
<td>(0.043)</td>
<td>(0.040)</td>
<td>(0.044)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CRA-issuer Relation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MFOI</td>
<td>-22.234***</td>
<td>2.872</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(4.723)</td>
<td>(4.427)</td>
<td></td>
</tr>
<tr>
<td>Year FE</td>
<td>yes</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>5913</td>
<td>5913</td>
<td>5913</td>
</tr>
</tbody>
</table>

**Note:** - standard error in parentheses
Considering semi-parametric estimation can only identify ratio of parameters, we take the ratios of the probit parameters to facilitate comparison.

Coefficients have the predicted signs and are consistent across models: when a firm’s ASSET or PROFIT increases, the probability of getting a higher rating on the firm’s bond increases. When STABILITY goes up (meaning the variance of ASSET goes up) or the firm has a higher LEVERAGE ratio, the probability of getting a higher rating decreases. The issue amount (AMT) has an insignificant impact on ratings, but this finding is consistent with Kedia et al. (2016).

Unlike the consistent results in estimating the structural index, the two parametric models yield very different results in estimating the effect of the CRA-issuer relation. Such disagreement may highlight the usefulness of using a more flexible model. I find MFOI, the main variable of interest, has a significant impact on ratings in OProbit2 (Column 3), as the calculated t-statistics for the MFOI coefficient is 0.65. However, after adding year fixed effect in OProbit1 (Column 2), the

---

15 Note that higher rating means lower categories...
effect of MFOI turn to be significant in a sense that a stronger connection with Moody’s (which implies a larger MFOI) will increase the probability of receiving higher ratings. Since MFOI enters the semiparametric model nonparametrically, no parameter of interest directly capture the (average) effect of MOFI. We will calculate its marginal effects and present relevant results in the next section.

Since the ordered probit models are nonlinear by nature, it is hard to quantify the impact of those variables by simply examining the size of their coefficients. In the next two subsections, I estimate various marginal effects to develop more concrete arguments.

5.2 Average Marginal Effects

In the top panel of Table 6, I report the Average Marginal Effects from the semiparametric model for each category K; while in the middle and lower panel I present the same effects calculated from the two ordered-probit models (Oprobit1 and Oprobit2). All the average marginal effects are calculated by exogenously increasing the variable of interest by one if the variable is discrete or log-transformed, and by one standard deviation for continuous variables. All estimates in Table 6 should be interpreted as the change in probabilities.

The estimates of marginal effects for MFOI are very different for the two class of models (reported in the last row of each panel). In the semiparametric ordered model, MFOI has quite different effects for bonds in different notches: the A-notch bonds are 4.2% more likely to be rated higher when the cross-ownership relation MFOI increases by one standard deviation, but such effects become negative for lower notch bonds. One possible interpretation would be Moody’s is more conservative when rating bonds that seems to have high default risk. In contrast, we do not find a significant impact of MFOI on ratings from the two ordered-probit models.

Turning to the estimates of average marginal effects for other variables, the semiparametric model and the two ordered probit model almost give identical results in terms of economic magnitude. The two class of model only disagree on the impact of subordination status. Recall that subordination status is measured by a binary variable which is one when the bond is a senior bond\(^\text{16}\) and zero otherwise, the results from the semiparametric model implies when a bond goes from subordinated debt to senior debt, the probability that it will be rated higher can increases by as much as 9.2%.

\(^{16}\)“Senior debt”, in general, is borrowed money that a company must repay first if it goes out of business.
Table 6: Average Marginal Effects

<table>
<thead>
<tr>
<th></th>
<th>Aaa</th>
<th>Aa</th>
<th>A</th>
<th>Baa</th>
<th>Ba</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Semiparametric Model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>asset</td>
<td>N/A</td>
<td>0.022</td>
<td>0.193</td>
<td>0.239</td>
<td>0.182</td>
<td>0.118</td>
<td>0.015</td>
</tr>
<tr>
<td>CVTA</td>
<td>N/A</td>
<td>-0.001</td>
<td>-0.016</td>
<td>-0.021</td>
<td>-0.021</td>
<td>-0.015</td>
<td>-0.002</td>
</tr>
<tr>
<td>leverage</td>
<td>N/A</td>
<td>-0.001</td>
<td>-0.040</td>
<td>-0.056</td>
<td>-0.055</td>
<td>-0.042</td>
<td>-0.006</td>
</tr>
<tr>
<td>profit</td>
<td>N/A</td>
<td>0.008</td>
<td>0.108</td>
<td>0.136</td>
<td>0.116</td>
<td>0.079</td>
<td>0.010</td>
</tr>
<tr>
<td>amt</td>
<td>N/A</td>
<td>0.000</td>
<td>0.004</td>
<td>0.005</td>
<td>0.005</td>
<td>0.004</td>
<td>0.000</td>
</tr>
<tr>
<td>seniority</td>
<td>N/A</td>
<td>0.004</td>
<td>0.073</td>
<td>0.092</td>
<td>0.083</td>
<td>0.058</td>
<td>0.007</td>
</tr>
<tr>
<td>MFOI</td>
<td>N/A</td>
<td>-0.003</td>
<td>0.042</td>
<td>-0.011</td>
<td>-0.020</td>
<td>-0.014</td>
<td>0.004</td>
</tr>
<tr>
<td><strong>Oprobit 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>asset</td>
<td>N/A</td>
<td>0.029</td>
<td>0.186</td>
<td>0.249</td>
<td>0.244</td>
<td>0.185</td>
<td>0.043</td>
</tr>
<tr>
<td>CVTA</td>
<td>N/A</td>
<td>-0.001</td>
<td>-0.009</td>
<td>-0.012</td>
<td>-0.012</td>
<td>-0.009</td>
<td>-0.002</td>
</tr>
<tr>
<td>leverage</td>
<td>N/A</td>
<td>-0.007</td>
<td>-0.046</td>
<td>-0.062</td>
<td>-0.060</td>
<td>-0.046</td>
<td>-0.011</td>
</tr>
<tr>
<td>profit</td>
<td>N/A</td>
<td>0.011</td>
<td>0.072</td>
<td>0.096</td>
<td>0.095</td>
<td>0.072</td>
<td>0.017</td>
</tr>
<tr>
<td>offering.amt</td>
<td>N/A</td>
<td>0.000</td>
<td>-0.002</td>
<td>-0.002</td>
<td>-0.002</td>
<td>-0.002</td>
<td>0.000</td>
</tr>
<tr>
<td>seniority</td>
<td>N/A</td>
<td>0.004</td>
<td>0.025</td>
<td>0.034</td>
<td>0.033</td>
<td>0.025</td>
<td>0.006</td>
</tr>
<tr>
<td>MFOI</td>
<td>N/A</td>
<td>0.002</td>
<td>0.010</td>
<td>0.013</td>
<td>0.013</td>
<td>0.010</td>
<td>0.002</td>
</tr>
<tr>
<td><strong>Oprobit 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>asset</td>
<td>N/A</td>
<td>0.027</td>
<td>0.179</td>
<td>0.241</td>
<td>0.237</td>
<td>0.180</td>
<td>0.043</td>
</tr>
<tr>
<td>CVTA</td>
<td>N/A</td>
<td>-0.001</td>
<td>-0.007</td>
<td>-0.010</td>
<td>-0.009</td>
<td>-0.007</td>
<td>-0.002</td>
</tr>
<tr>
<td>leverage</td>
<td>N/A</td>
<td>-0.007</td>
<td>-0.048</td>
<td>-0.064</td>
<td>-0.063</td>
<td>-0.048</td>
<td>-0.011</td>
</tr>
<tr>
<td>profit</td>
<td>N/A</td>
<td>0.011</td>
<td>0.070</td>
<td>0.094</td>
<td>0.092</td>
<td>0.070</td>
<td>0.017</td>
</tr>
<tr>
<td>amt</td>
<td>N/A</td>
<td>0.000</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.000</td>
</tr>
<tr>
<td>seniority</td>
<td>N/A</td>
<td>0.004</td>
<td>0.026</td>
<td>0.035</td>
<td>0.034</td>
<td>0.026</td>
<td>0.006</td>
</tr>
<tr>
<td>MFOI</td>
<td>N/A</td>
<td>0.000</td>
<td>-0.002</td>
<td>-0.003</td>
<td>-0.003</td>
<td>-0.002</td>
<td>0.000</td>
</tr>
</tbody>
</table>

**Note:**
- This table reports the average marginal effect (AME) of covariates from the semiparametric model (top panel), ordered probit with year fixed effects (middle panel) and ordered probit without year fixed effects (lower panel).
- Each AME is computed by increasing the variable of interest (indicated in the first column in each panel) by one standard deviation if the variable is continuous, and by one if the variable is discrete.
- All numbers should be interpreted as the change in probability. For example, the number 0.022 means an Aa-rated bond is 2.2% more likely to be rated as Aaa when the firm’s log of asset increased by one standard deviation.
Table 7: Quantile Marginal effects of MFOI

<table>
<thead>
<tr>
<th></th>
<th>Aaa</th>
<th>Aa</th>
<th>A</th>
<th>Baa</th>
<th>Ba</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>N/A</td>
<td>-0.005*</td>
<td>-0.063***</td>
<td>-0.063***</td>
<td>-0.002</td>
<td>-0.004</td>
<td>-0.002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-1.647)</td>
<td>(-6.816)</td>
<td>(-4.648)</td>
<td>(-0.097)</td>
<td>(-0.197)</td>
<td>(-0.096)</td>
</tr>
<tr>
<td>Q2</td>
<td>N/A</td>
<td>0.001</td>
<td>0.018**</td>
<td>-0.013</td>
<td>-0.048***</td>
<td>-0.032**</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.358)</td>
<td>(1.955)</td>
<td>(-1.257)</td>
<td>(-3.772)</td>
<td>(-2.133)</td>
<td>(0.562)</td>
</tr>
<tr>
<td>Q3</td>
<td>N/A</td>
<td>-0.002</td>
<td>0.026***</td>
<td>-0.017</td>
<td>-0.043***</td>
<td>-0.027*</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-0.306)</td>
<td>(2.904)</td>
<td>(-1.485)</td>
<td>(-2.960)</td>
<td>(-1.667)</td>
<td>(0.520)</td>
</tr>
<tr>
<td>Q4</td>
<td>N/A</td>
<td>-0.004</td>
<td>0.089***</td>
<td>-0.006</td>
<td>-0.023</td>
<td>-0.012</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-0.887)</td>
<td>(8.670)</td>
<td>(-0.507)</td>
<td>(-1.582)</td>
<td>(-0.737)</td>
<td>(0.088)</td>
</tr>
<tr>
<td>Q5</td>
<td>N/A</td>
<td>-0.005*</td>
<td>0.140***</td>
<td>0.046***</td>
<td>0.014</td>
<td>0.006</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-1.704)</td>
<td>(15.891)</td>
<td>(3.262)</td>
<td>(0.784)</td>
<td>(0.288)</td>
<td>(0.078)</td>
</tr>
</tbody>
</table>

Note:
- This table reports the quantile marginal effect (QME) of MFOI from the semiparametric model, with the calculated t-statistics for testing $QME = 0$ reported in parentheses.
- Each different line refers to the quantile level of interest, with Q1 refers to the lowest quartile of MFOI (firms do not well connected with the Moody’s) and Q5 the highest (firms that are strongly connected).
- All numbers should be interpreted as the change in probability. For example, the first number -0.005 should be interpreted in the following way: the probability that an AA bond will be rated to AAA will decrease by 0.005 if the relation variable MFOI increase by 0.005 standard deviation. This impact is not statistically significant at 90% confidence level as the calculated t-statistics is -1.647.
- Compared to ordered-probit models, the semiparametric model captures more heterogeneity in marginal effects among different rating notches: taking MFOI as an example, the semiparametric estimates of average marginal effects vary from -2% to 4.2% as the rating notches changes; while their parametric analogue only vary from 0.2% to 1.3% in Oprobit 1 and -0.3% to 0% in Oprobit 2.
- The homogeneity of marginal effects in ordered-probit models is not surprising because they assume the probability of being rate into any notch is governed by a single error term whose distribution is known and fixed. This assumption, however, is hard to justify given the rating scheme can be very different for different rating categories.

5.3 Quantile marginal effects

To address the two hypotheses mentioned above, we compute the quantile marginal effect ($QME^K_q$) defined in (6). In Table 7 we report the impacts on ratings from a one standard deviation increment
of MFOI\textsuperscript{17}, where Q1 refers to the lowest and Q5 the highest quartile for MFOI. The horizontal dimension indexes the rating category that we choose to evaluate the marginal effects. Similarly as the average marginal effects, all the number should be interpreted as change in probabilities and the standard errors are calculated from the asymptotic theory derived in this paper.

From the quantile marginal effects analysis, we find Moody's assign favorable ratings only to firms that are related to its large shareholders. Recall that MFOI is a measure of “connectedness” between the CRA and the issuer firm. For example, when evaluating the quantile marginal effects conditional on the highest quartile (Q5) of MOFI, we focus on firms that are connected with Moody's through large shareholders. From Table 7, it is clear that as a firm’s shareholding relation with Moody’s strengthens, A-grade bonds are more likely to be rated as Aa or Aaa. This “upgrade probability” increases to 14% for firms that have the strongest connectedness with Moody (those with MFOI in the highest quartile). This implies that roughly one out of seven A-grade bonds issued by those firms might have received favorable treatment. A strengthening CRA-issuer relation also has a significant positive impact for Baa-grade bonds, but the economic magnitude is much smaller (4.6% vs 14%). Another interesting finding is the quantile marginal effect of MFOI is negative in its lowest quarter Q1, which implies firms related with Moody’s minority shareholders do not receive any favorable treatment.

In addition, we find bonds that are below investment grade are not benefit from the CRA-issuer relation regardless of the issuer’s relation with Moody’s. Even for firms with strong relation with Moody’s (those in Q5), the probability of being rated into a higher category is only 1.4% for Ba-grade bonds and 0.6% for B-grade bonds. Both effects are not statistically significant.

### 5.4 Discussion

Looking back to the two hypotheses of interest proposed in the previous section, we find evidence that Moody’s is indeed biased towards issuers related its large shareholders, and the degree of bias depends on the quality of the underlying bonds. Basically only investment grade bonds issued by firms related with Moody’s large shareholders received favorable treatment. High-yield bonds, regardless

\textsuperscript{17}We also examine the case when increasing MFOI by a smaller amount, the magnitude of marginal effects decrease as expected, but the pattern does not change
of the bond issuer’s shared-ownership relation with Moody’s never received preferential treatment.

Our results are robust to the number of quantiles. We calculate the quantile marginal effects of MFOI using deciles and plot the pattern in Figure 3. The red solid line indicates the quantile marginal effect corresponding to each decile level, and the green dotted lines are the upper and lower bounds of the 95% Confidence Interval. We choose A-grade and Ba-grade bonds only for the purpose of exposition, and the pattern for quantile marginal effects are similar for other investment-grade (high-yield) bonds.

For investment-grade bonds, QME becomes significantly positive after the seventh decile. However, QME for high-yield bonds remain statistically insignificant throughout. This differential effect between investment-grade and high-yield bonds is not documented in Kedia et al. (2016) or any empirical literature on rating bias. We interpret it in the following way: Bolton et al. (2012) argues that when a CRA decides whether or not to assign a inflated rating, it compares the predicted revenue generated by the inflated rating and the reputation loss if the underlying debt turns out default. Since a low quality bond is more likely to default, implying a higher probability to trigger the reputation loss, the CRA should be less likely to assign inflated ratings on low quality bonds.

6 Robustness Check

Alternative measures of the CRA-issuer relation  Previously we use a single variable, MFOI, to measure the CRA-issuer relation. In case this single measure did not fully capture the complicated relation, we perform the same analysis above replacing MFOI with a linear combination of three alternative measures. Recall that the three measures are (1) Number of Large Shareholder, defined as the number of common shareholders of Moodys and the issuer firm who owns at least 5% of Moodys’ stock (num_largeSH). (2) Number of Shareholder, defined as the number of common shareholders of Moodys and the issuer firm (num_SH). (3) Overlapping shares, defined as the total percentage of Moody’s stock owned by common shareholders (overlapping share).

Rather than assuming which measure is the appropriate one, we assume that the CRA-issuer

---

18We follow Kedia et al. (2016) and pick 5% as the threshold point.
relationship variable, $Z$, is an unknown function of all three measures above in that

$$Z = g(w_1 \ast \text{overlapping share} + w_2 \ast \text{num_largeSH} + w_3 \ast \text{num_SH})$$

(12)

The $w$-weights are parameters that we seek to estimate. It is important to note that the average marginal effect (AME) and quantile marginal effects (QME) of these three measures as well as the CRA-issuer relation $Z$ can still be identified even if $g(\cdot)$ is unknown. We report the quantile marginal effects of $\text{num_largeSH}$ in Table 8\textsuperscript{19}. That is, we examine the effect on ratings when the bond issuer

\textsuperscript{19}The weights $w$ are estimated first using the maximum likelihood procedure. Given the estimated $\hat{Z} = g(\hat{w}_1 \ast \text{overlapping share} + \hat{w}_2 \ast \text{num_largeSH} + \hat{w}_3 \ast \text{num_SH})$, we can estimate the marginal effects of $Z$ as well as $\text{overlapping share}$, $\text{num_largeSH}$ and $\text{num_SH}$. Additional results on the marginal effects for the three measures are available upon request.
is connected with one more large shareholders of Moody’s. The pattern of the quantile marginal effects, as compared to Table 7, remain unchanged.

Table 8: Quantile marginal effects of number of large shareholders

<table>
<thead>
<tr>
<th></th>
<th>Aaa</th>
<th>Aa</th>
<th>A</th>
<th>Baa</th>
<th>Ba</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>N/A</td>
<td>-0.001</td>
<td>-0.040***</td>
<td>-0.041**</td>
<td>-0.004</td>
<td>-0.013</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-0.172)</td>
<td>(-2.996)</td>
<td>(-1.789)</td>
<td>(-0.162)</td>
<td>(-0.433)</td>
<td>(0.030)</td>
</tr>
<tr>
<td>Q2</td>
<td>N/A</td>
<td>-0.003</td>
<td>0.007</td>
<td>-0.048***</td>
<td>-0.028</td>
<td>-0.039</td>
<td>-0.002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-0.917)</td>
<td>(1.096)</td>
<td>(-3.254)</td>
<td>(-1.091)</td>
<td>(-1.316)</td>
<td>(-0.056)</td>
</tr>
<tr>
<td>Q3</td>
<td>N/A</td>
<td>-0.002</td>
<td>0.038***</td>
<td>-0.010</td>
<td>0.026</td>
<td>0.044</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-0.431)</td>
<td>(2.847)</td>
<td>(-0.528)</td>
<td>(1.062)</td>
<td>(1.522)</td>
<td>(0.141)</td>
</tr>
<tr>
<td>Q4</td>
<td>N/A</td>
<td>0.000</td>
<td>0.053***</td>
<td>0.022</td>
<td>0.081**</td>
<td>0.081*</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.063)</td>
<td>(3.474)</td>
<td>(0.757)</td>
<td>(2.016)</td>
<td>(1.725)</td>
<td>(0.173)</td>
</tr>
<tr>
<td>Q5</td>
<td>N/A</td>
<td>-0.002</td>
<td>0.149***</td>
<td>0.052**</td>
<td>0.022</td>
<td>0.010</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-0.455)</td>
<td>(8.531)</td>
<td>(1.855)</td>
<td>(0.636)</td>
<td>(0.267)</td>
<td>(0.034)</td>
</tr>
</tbody>
</table>

Note:
- This table reports the quantile marginal effect (QME) of $Z = g(w_1 \ast overlapping\ share + w_2 \ast num\_largeSH + w_3 \ast num\_SH)$, with the calculated t-statistics for testing $QME = 0$ reported in parantheses.
- Each different line refer to the quantile level of interest, with Q1 refers to the lowest quartile of Z (firms do not well connected with the Moody’s) and Q5 the highest (firms that are strongly connected).
- All numbers should be interpreted as the change in probability in the same way as in Table 5.

Omitted Variable Bias  Because it is clear that rating agencies use more information than the explanatory variables that we specified, in this section we check whether the effect of CRA-issuer relation is driven by omitted variable bias. In our sample, 3207 out of 5913 issues are concurrently rated by S&P’s, the other major rating agency in the market. We calculate the difference\(^{20}\) in ratings for those bonds, and examine whether the relative rating difference can be explained by the issuer’s relation with Moody’s. This “difference-in-differences” approach was first proposed by Kedia et al. (2016) to study rating bias. In specific, we run a series of linear probability regressions controlling different combination of firm and bond characteristics, and the results are displayed in Table 9.

Without controlling for firm and bond characteristics, we find that having an additional Moody’s large shareholders involved (increasing the number of large shareholders by one) would increase the rating difference by 0.22 rating notch. Such a effect decreases to about 0.15 rating notch after controlling for these characteristics, but remains highly significant. This implies that on average, \(^{20}\)The rating difference is defined as Moody’s rating - S&P’s rating, therefore a negative rating difference indicates Moody’s rating is more favorable.
when assigning ratings to firms related with its large shareholders, Moody’s ratings are 0.15 notches higher than S&P’s. In terms of the magnitude, our finding is similar to the effect documented in Kedia et al. (2016) (Moody’s ratings are 0.213 notch higher than S&P’s on average).

Table 9: Rating Difference bewteen Moody’s and S&P

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Moody’s - S&amp;P’s ratings</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRA-Issuer relation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>overlapping share</td>
<td>1.820***</td>
<td>0.107</td>
<td>0.127</td>
<td>0.105</td>
</tr>
<tr>
<td></td>
<td>(7.58)</td>
<td>(0.43)</td>
<td>(0.51)</td>
<td>(0.38)</td>
</tr>
<tr>
<td>num_LargeSH</td>
<td>-0.217***</td>
<td>-0.155***</td>
<td>-0.145***</td>
<td>-0.142***</td>
</tr>
<tr>
<td></td>
<td>(-7.01)</td>
<td>(-5.22)</td>
<td>(-4.89)</td>
<td>(-4.70)</td>
</tr>
<tr>
<td>num_SH</td>
<td>-0.00360***</td>
<td>0.00124**</td>
<td>0.00117**</td>
<td>0.00129**</td>
</tr>
<tr>
<td></td>
<td>(-9.99)</td>
<td>(2.80)</td>
<td>(2.61)</td>
<td>(2.82)</td>
</tr>
<tr>
<td>MFOI</td>
<td></td>
<td></td>
<td></td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.29)</td>
</tr>
<tr>
<td>Firm characteristics</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Bond characteristics</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations</td>
<td>3207</td>
<td>3207</td>
<td>3207</td>
<td>3207</td>
</tr>
</tbody>
</table>

$t$ statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

7 Conclusion

This paper contributes to the literature by evaluating rating quality using a semiparametric ordered model. Compared to existing models in the bond rating literature, the semiparametric model proposed in this paper allows a richer set of interactions among covariates. Through the marginal effects analysis, we are able to test various hypotheses regarding the impartialness of the rating agency in presence of conflict of interests. In the empirical section, we study to what extent are Moody’s ratings affected by the economic interest of its large shareholders, which is very important for the regulation of credit rating agencies.

Given the complicated nature of the rating process, the standard ordered probit specification has a greater exposure of incorrectly specifying the model, which typically makes the estimator inconsistent. In contrast, the semiparametric specification does not subject to this critique because it allows a richer set of interaction among explanatory variables and leaves the error distribution unspecified.
In summary, we find that Moodys’ is likely to assign favorable ratings to firms that have a strong interaction with Moodys’ large shareholders. Based on marginal effect analysis, we found being a large investee firm of Moody’s large shareholder could increase the probability of receiving favorable treatment by as much as 14%, meaning that, on average, one out of seven bonds issued by those firms received favorable treatment. However, we found Moody’s does not assign favorable ratings to firms related to Moodys’ small shareholders. This “large shareholder bias” is in accord with the literature on the role of large shareholders in corporate governance. In addition, we found subprime bonds issued by any firms, regardless their ownership interaction with Moodys’, will unlikely to be treated with favor, which also seems credible because overrating a subprime bond would incur a greater expected reputation loss than overrating a safe bond. The policy relevance of the findings in this paper is that when credit rating agencies are publicly held by diffuse owners, their ratings are still highly trustable.
References


James J Heckman. Sample selection bias as a specification error (with an application to the estimation of labor supply functions), 1977.


Simi Kedia, Shivaram Rajgopal, and Xing Zhou. Does it matter who owns moody’s, 2016.


Appendix A  Assumptions and Definitions

We provide the assumptions and definitions that we employ to establish the asymptotic properties for the semiparametric estimators defined in Section 3.

A.1 The data The data $(Y^k_i, X_i, Z_i)$ are i.i.d. observations from the model in (1)-(3). The columns of $[X,Z]$ are linearly independent with probability 1. In addition, we require the support of $[X,Z]$ to be finite\(^{21}\).

A.2 The error term The error in the latent model (2), $U_i$, is independent over i and independent of $[X_i, Z_i]$.

A.3 Parameter space The vector of true parameters values $\theta_0 \equiv [\theta_0^F, \theta_0^B]$ for the model in (1)-(3) lies in the interior of a compact parameter space, $\Theta$.

A.4 Index assumptions Write the vector of indices $V(\theta_0) \equiv [X_F + X_F^T \theta_0^F, X_B + X_B^T \theta_0^B, Z]' = [V_F, V_B, Z]$, which depends on two vector, $X_F$ and $X_B$, and a continuous variable $Z$. We furthur assume that the first components of $X_F$ and $X_B$ are continuous and functionally independent, and the following index assumption holds:

$$E[Y^k_i = 1|X_i, Z_i] = E[Y^k_i = 1|V(\theta_0)]$$ (13)

In this paper we may refer this conditional probability as the category probability, as this is the probability that an observation will be placed into category $k$ given the true index.

A.5 Conditional densities With $V$ defined in A.4, denote $g(t|y, x, z)$ as its density conditioned on $Y = y$ and $X = x$, $Z = z$. Denote $\nabla^d g(t|y, x, z)$ as the partials or cross partials up to order d, with $\nabla^0 g(t|y, x, z) \equiv g(t|y, x, z)$. With $g$ defined on a compact support, we assume $g \geq 0$ and $\nabla^d g(t|y, x, z)$ uniformly bounded for $d = 0, 1, 2, 3$ on the interior of its support.

A.6 Residual property of $\nabla_\theta E[Y^k_i = 1|V(\theta)]$ (Result due to Newey) Assume A.4 holds, then the first derivative of the category probability function has mean zero conditioning on knowing

\(^{21}\)Since $X$ denotes variables like firm’s asset, leverage ratio...which seems to be naturally bounded from above
the true index

\[ E[\nabla_\theta E[Y^k_i = 1|V(\theta)]|V(\theta_0)] = 0 \]  

The residual property of this derivative is proved in Theorem 0 in Klein and Shen (2010)\textsuperscript{22}. It will become transparent that why this property can be used as a bias control when we show the index estimator is asymptotic normal.

\textbf{A.1-A.5} are standard in the literature. Namely we require each index has at least one continuous variable (A.4) and densities for continuous variables and the indices must be sufficiently smooth, as implied by (A.5). We defer the discussion on A.6 after laying out all the definitions. Additional window conditions will be required and stated directly in the theorems for which they are needed.

To formally define the estimators, we will also require the definitions that follow.

\textbf{D.1 The localized Model} With \( V(\theta_0) \equiv (V_F, V_B, Z) \) defined in A.4 and \( \epsilon_i \equiv Y^k_i - P^k(V_i(\theta_0)) \), the mean regression model in (???) can be written as \( Y^k_i = E[Y^k_i|V_i(\theta_0)] + \epsilon_i = P^k(V_i(\theta_0)) + \epsilon_i \).

Letting \( V_j \) be a point close to \( V_i(\theta_0) \) and \( \Delta_i(j) \equiv P^k(V_i(\theta_0)) - P^k(V_j) \), the localized model at \( V_j \) can be defined as:

\[ Y^k_i = P^k(V_j) + \Delta_i(j) + \epsilon_i \]  

Since this paper focuses on marginal effects, inevitably sometimes we will evaluate \( P^k \) conditioning on a “perturbed index”: \( V^\delta \equiv (V_F, V_B, Z + \delta) \). This perturbation results in the following \( \delta \)-localized model:

\[ Y^k_i = P^k(V^\delta_j) + \Delta^\delta_i(j) + \epsilon^\delta_i \]  

Where \( \epsilon^\delta_i \) and \( \Delta^\delta_i(j) \) are defined similarly as above but replace \( V(\theta_0) \) with \( V^\delta \). It is important to note that \( \epsilon \) and \( \epsilon^\delta \) shall follow different distribution.

\textbf{D.2 Trimming} With \( W_{ik} \) as the ith observation on a continuous variable, \( W_k, k = 1,\ldots,K \), let

\footnote{One of the authors acknowledges Whitney Newey for providing this result and proof in a private communication.}
\[ \hat{\tau}_{ik} \equiv \{ \hat{a}_k < W_{ik} < \hat{b}_k \} \text{ and } \hat{\tau}_i = \prod_k \hat{\tau}_{ik}, \text{ where } \hat{a}_k, \hat{b}_k \text{ are, respectively, lower and upper sample quantiles for } W_k. \]

- With \( X_k \) as an exogenous variable, when \( W_{ik} = X_{ik} \), we refer to \( \tau_{ix} \) as X trimming;
- With \( \hat{V} \) as the estimated index, when \( W_{ik} = \hat{V}_i \), we refer to \( \tau_{iv} \) as index trimming.
- In case when a smooth trimming function is needed, define

\[ \tau(z, \delta) \equiv (1 + \exp(-(\ln(N) \ast \ln(N) \ast (z - \delta))))^{-1} \quad (17) \]

as a smoothed approximation to an indicator on \( z \geq \delta \). A smoothed indicator on \( z \in [a, b] \) is then defined as \( \tau(z, a) \ast \tau(b, z) \).

**D.3 Estimated density** With the window size \( h = O(N^{-r}) \), the kernel function is termed regular if \( K(z) \leq 0, \int K(z) dz = 1, \) and \( K(z) = K(-z) \). Let \( V \equiv (V_B, V_F, Z) \) as the vector of indices, define:

\[
\hat{g}_{1i}(V_i | Y_{ik}^k = 1) \equiv \frac{1}{N_1h^3} \sum_{Y_{ik}^k=1} K_1\left(\frac{V_{iF} - V_{jF}}{h}\right) K_2\left(\frac{V_{iB} - V_{jB}}{h}\right) K_3\left(\frac{Z_i - Z_j}{h}\right) \quad (18)
\]

\[
\hat{g}_{0i}(V_i | Y_{ik}^k = 0) \equiv \frac{1}{N_0h^3} \sum_{Y_{ik}^k=0} K_1\left(\frac{V_{iF} - V_{jF}}{h}\right) K_2\left(\frac{V_{iB} - V_{jB}}{h}\right) K_3\left(\frac{Z_i - Z_j}{h}\right) \quad (19)
\]

where \( N_0, N_1 \) are the number of observations with \( Y_{ik}^k = 0 \) or \( 1 \). As noted in Klein and Shen (2010), some adjustments are needed on these densities when the indices are in a non-compact set.\(^{23}\)

**D.4 Estimated probability**\(^{24}\) From Bayes’ rule and based on the unadjusted density above, the estimated conditional expectation with window parameter \( r \) is denoted as \( \hat{P}_k \equiv \hat{P}_k(V_i; \theta) = \ldots \)

---

\(^{23}\)Refer the index trimming in D.1, define the adjustment \( \hat{\gamma}_i \) to be:

\[ \hat{\gamma}_i \equiv (1 - \hat{\tau}_i)h^{0.5}q. \quad (20) \]

where \( q = q_1q_2q_3 \) is the product of lower quantiles of the indices. For \( d = 0 \) or \( 1 \), let the adjusted density be: \( \hat{g}_{kid}(v | Y_{ik}^k = d) \equiv \hat{g}_{id}(v | Y_{ik}^k = d) + \hat{\gamma}_i \) and calculate the adjusted probability \( \hat{P}_{ikw} \) as (21) using these adjusted density.

\(^{24}\)In this paper, the author also refers expectation as “probability”, because in the ordered probit model, \( E[Y_{ik}^k = 1 | V] \) is the conditional probability of \( Y_{ik}^k = 1 \) given \( V \).
\[ \hat{E}[Y^* = 1|V = v_i] \text{ and is given by} \]

\[
\hat{p}_i^k \equiv \frac{\hat{P}(Y^*_i = 1)\hat{g}_i^k(V_i|Y^*_i = 1)}{\hat{g}_i^k(V_i)} \tag{21}
\]

\[ = \frac{1}{N-1} \sum_{j \neq i} Y^*_i K_1(\frac{V_{jE} - V_{jE}}{h} K_2(\frac{V_{jE} - V_{jE}}{h} K_3(\frac{Z_i - Z_i}{h}))}{\frac{1}{N-1} \sum_{j \neq i} K_{ij} = \hat{f}_i/\hat{g}_i}
\]

Where \( K_{ij} \) is a product of kernels \( K_1(\frac{V_{jE} - V_{jE}}{h}) K_2(\frac{V_{jE} - V_{jE}}{h} K_3(\frac{Z_i - Z_i}{h})) \)

\[ h^3 \text{ and } h = O(N^{-r}) \]

**D.5 Final estimator** Refer to the localized model and estimated probability in D.1 and D.4, the final estimator of \( E[Y^*_i = 1|V = v_i] \) is given by:

\[ \hat{p}_i^k = \frac{1}{N-1} \sum_{j \neq i} (Y^*_i - \hat{\Delta}_i(j)) K_{ij}/\frac{1}{N-1} \sum_{j \neq i} K_{ij} = \hat{f}_i/\hat{g}_i \tag{22} \]

Where \( \hat{\Delta}_i(j) \equiv \hat{p}_i^k(V_i(\theta_0)) - \hat{p}_i^k(V_j) \) is an estimate of the localization bias \( p_i^k(V_i(\theta_0)) - p_i^k(V_j) \).

**D.6 First- and Second-stage estimator** Based on the above definitions, we define:

\[ \hat{\theta}_1 = \arg\max_\theta Q_1(\theta), \quad Q_1(\theta) = N^{-1} \sum_{i=1}^N \sum_{k=1}^n Y^*_i \ln(\hat{p}_i^k) \tag{23} \]

\[ \hat{\theta}_2 = \arg\max_\theta Q_2(\theta), \quad Q_2(\theta) = N^{-1} \sum_{i=1}^N \sum_{k=1}^n Y^*_i \ln(\hat{p}_i^k) \tag{24} \]

**D.7 Quantile trimming** Let \( W \in [X, Z] \) be the variable whose quantile marginal effects we are interested in. Define \( t_{qj} \) to be the true quantile trimming function which takes value 1 only if \( W \) is in a given quantile: \( t_{qj} \equiv 1\{q(\lambda_1) < W < q(\lambda_2)\} \), where \( \lambda_1 \) and \( \lambda_2 \) are the upper and lower bound for that quantile and \( q(\lambda_1), q(\lambda_2) \) the corresponding population quantile for \( W \).

**D.8 Quantile Marginal effects estimators** Let \( V(\hat{\theta}) \equiv [X_{F1} + X_F\hat{\theta}_0^F, X_{B1} + X_B\hat{\theta}_0^B, Z] \) be the estimated index. For an observation with \( Y_i = K \), the true cumulative marginal effect \( CME_i(\theta_0; K) \)
defined in (5) can be estimated by:

\[
\hat{m}_j(\hat{\theta}) \equiv \sum_{k=1}^{L} \{Y_i \geq k\} [\hat{P}_{ka}(V_F(\hat{\theta}), V_B(\hat{\theta}), Z + \delta) - \hat{P}_{ka}(V_F(\hat{\theta}), V_B(\hat{\theta}), Z)]
\] (25)

Refer to D.7 for quantile trimming \(t_q\), we define the quantile marginal effect \(CME^K_q\) and its estimator by:

\[
QME^K_q \equiv \frac{E[t_q m_j(\theta_0)]}{E[t_q]}
\] (26)

\[
\hat{QME}^K_q \equiv \frac{\sum_{j=1}^{N} \hat{t}_{jq} \hat{m}_j(\hat{\theta})}{\sum_{j=1}^{N} \hat{t}_{jq}}
\] (27)

D.9 Bahadur Representation Refer to D.3 and D.6, with \(g_Z(\cdot)\) be the marginal density for \(Z\), let:

\[
B_j \equiv \begin{bmatrix}
(1\{Z \leq q(\lambda_2)\} - \lambda_2)/g_Z(\lambda_2) \\
(1\{Z \geq q(\lambda_1)\} - \lambda_1)/g_Z(\lambda_1)
\end{bmatrix}
\]

The Bahadur representation (Bahadur (1966), David (1981) and Klein and Shen (2015)) can now be defined as:

\[
\sqrt{N}[\hat{q}(\lambda) - q(\lambda)] = \sqrt{N\overline{B}} + o_p(1), \overline{B} \equiv \frac{1}{N} \sum_{j=1}^{N} B_j
\] (28)

Before discussing the asymptotic result, we need to discuss the two bias control strategies that we employ to obtain asymptotic normality. We briefly discuss them here and leave the formal argument in the actual proofs. Firstly we discuss why we need the residual property of \(\nabla \theta E[Y_i = 1|V(\theta)]\): the estimated gradient of the likelihood function in (8) can be reformulated as a sample mean of iid terms plus a sample mean of terms look like \(\tau_{iv} H(V_i) \nabla \theta E[Y_i = 1|V(\theta)]\), where \(H\) is a function of index with estimated parts evaluated at \(\theta_0\). On the assumption of A.6 and the second stage estimation with trimming on the index (D.6), taking iterated expectation conditional on the true index will make this extra piece vanish.

To use the residual property of \(\nabla \theta E[Y_i = 1|V(\theta)]\) as a bias control, we need to ensure that the
estimated counter part of $\nabla_\theta E[Y_i^k = 1|V(\theta)]$ in the gradient can be replaced by the truth. It turns out we do need another bias control to make the replacement permissible. Refer to the localized model in D.1, this localization enables us to treat the unknown function $P^k$ evaluated at $V_j$ as estimable parameter. While the kernel weighting (D.3) intends to reduce the localization bias $\Delta_i(j)$, the resulting bias in the estimator in D.4 is not small enough to establish asymptotic normality. For this reason, referring to the localized model in D.1, we define the bias corrected model has the form:

$$Y_i^k - \hat{\Delta}_i(j) = P^k(V_j) + [\Delta_i(j) - \hat{\Delta}_i(j)] + \epsilon_i$$  \hspace{1cm} (29)

$$\hat{\Delta}_i(j) \equiv \hat{P}^k(V_i(\theta_0)) - \hat{P}^k(V_j)$$  \hspace{1cm} (30)

As shown in Klein and Shen (2015), by removing an estimate of the localization bias from the outcome variable $Y_i^k$, we ensure the bias in $\hat{P}_j^k$, the final estimator defined in D.5, is small enough to permit the replacement.

**Appendix B  Asymptotic Results**

Recall that in order to use the bias correction due to Newey, we need to estimate the model twice, first trimming on the X’s and then trimming on the indices. The estimators form either stage is consistent.

**Theorem B.1 (Identification and consistency).** The model is identified under the assumption that there is at least one continuous explanatory variable in $V_F$ and $V_B$ are distinct and functionally independent. Denote $\hat{\theta}_1$ and $\hat{\theta}_2$ as the first- and second-stage estimators respectively and assume (A.1)-(A.5). Base the first(second)-stage estimator on a regular(adjusted) expectation (D4) with $1/12 < r_1 < 1/10$. Then:

$$|\hat{\theta}_1 - \theta_0| = o_p(1); \hspace{1cm} |\hat{\theta}_2 - \theta_0| = o_p(1)$$

**Proof.** A more general identification result of multiple-index models has been showed by Ichimura and Lee (1991), and our model is a special case. To show consistency, recall that $\theta$ maximize the estimated average log-likelihood function $\hat{Q}(\theta)$. From Lemma C.3, $\sup_\theta |\hat{Q}(\theta) - \tilde{Q}(\theta)| \xrightarrow{P} 0$, where $\tilde{Q}(\theta)$ is obtained from $\hat{Q}(\theta)$ by replacing all estimated functions with their probability limits. From
standard argument, \( \tilde{Q}(\theta) \) converges uniformly to its expectation \( E[\tilde{Q}(\theta)] \). Then, consistency would follow as long as \( E[\tilde{Q}(\theta)] \) is uniquely maximized at \( \theta_0 \), which is shown by Ichimura (1993).

**Theorem B.2 (Normality).** Assume A.1-A.5 and base the second stage estimator, \( \hat{\theta}_2 \), on an adjusted expectation (Referring to D.5) with the window size \( 1/12 < r < 1/10 \), with \( Q_2 \) the likelihood function defined in (8),

\[
\sqrt{N}(\hat{\theta}_2 - \theta_0) \overset{d}{\rightarrow} N(0, H_0^{-1})
\]

where \( H_0 \equiv E[\nabla_{\theta} Q_2(\theta_0)] \)

**Proof.** From a taylor expansion of the estimated gradient on \( \hat{\theta} \) and the fact that the estimated gradient is zero evaluated at \( \theta_0 \), we have

\[
\sqrt{N}(\hat{\theta} - \theta_0) = -\hat{H}(\hat{\theta}^+)^{-1}\sqrt{N}\hat{G}(\theta_0) \quad \theta^+ \in (\theta_0, \hat{\theta})
\]

where \( G \) refers to the gradient and \( H \) the Hessian of the likelihood function in (8):

\[
\hat{G}(\theta_0) = \nabla_{\theta} \hat{Q}_2(\theta_0) \quad (32)
\]

\[
\hat{H}(\theta_0) = \nabla_{\theta} \hat{Q}_2(\theta_0) \quad (33)
\]

Refer to the expression in (31), from Lemma C.4 the Hessian converges to some fixed matrix \( H_0 \equiv E[H(\theta_0)] \). Let \( P_i^k \equiv P_i^k(V_F, V_B, Z) \), then the normalized gradient \( \sqrt{N}\hat{G}(\theta_0) \), is given as:

\[
\sqrt{N}\hat{G}(\theta_0) \equiv \sqrt{N} \sum_i \hat{\tau}_{iv} \sum_k \frac{Y_i^k}{P_i^k} \frac{\partial \hat{P}_i^k}{\partial \theta}|_{\theta=\theta_0}
\]

\[
= \sqrt{N} \sum_i \hat{\tau}_{iv} \sum_k \frac{Y_i^k - \hat{P}_i^k}{\hat{P}_i^k} \frac{\partial \hat{P}_i^k}{\partial \theta}|_{\theta=\theta_0}
\]

\[
= \sqrt{N} \sum_i \hat{\tau}_{iv} \frac{Y_i^1 - \hat{P}_i^1}{\hat{P}_i^1} \frac{\partial \hat{P}_i^1}{\partial \theta}|_{\theta=\theta_0} + \frac{Y_i^2 - \hat{P}_i^2}{\hat{P}_i^2} \frac{\partial \hat{P}_i^2}{\partial \theta}|_{\theta=\theta_0} + \ldots
\]

The first equality above follows from the fact that

\[
\sum_k \frac{\partial \hat{P}_i^k}{\partial \theta} = \frac{\partial \sum_k \hat{P}_i^k}{\partial \theta} = \frac{\partial 1}{\partial \theta} = 0
\]

(37)
As the first term in (36) and all remaining terms have the same structure, it suffices to analyze the first term. From standard argument in Pakes and Pollard (1989) and Klein and Spady (1993), the estimated trimming function \( \hat{\tau}_{iv} \) can be replaced by the truth asymptotically. With \( \hat{w}_i^k \equiv \frac{\partial \hat{P}_i^k}{\partial \theta} |_{\theta=\theta_0} \), we may write this term as:

\[
\sqrt{N}(D_1 - D_2) = \sqrt{N} \sum_i \tau_{iv} [Y_i^k - \hat{P}_i^k] \hat{w}_i^k \\
D_1 \equiv \sum_i \tau_{iv} [Y_i^k - P_i^k] \hat{w}_i^k \\
D_2 \equiv \sum_i \tau_{iv} [\hat{P}_i^k - P_i^k] \hat{w}_i^k
\]

By making use of the fact that the “true” residual \( Y_i^k - P_i^k \) has zero conditional expectation, it can be shown that \( D_1 \) could be replaced by \( \sum_i \tau_{iv} [Y_i^k - P_i^k] \hat{w}_i^k \) through a mean-square convergence results (Klein and Spady (1993)). For \( D_2 \), with \( 1/12 < r < 1/6 \), from Lemma C.7, we have

\[
\sqrt{N}(D_2 - D_2^*) = \frac{1}{\sqrt{N}} \sum_i \tau_{iv} [\hat{P}_i^k - P_i^k] (w_i^k - \hat{w}_i^k) = o_p(1), \quad D_2^* \equiv \sum_i \tau_{iv} [\hat{P}_i^k - P_i^k] \hat{w}_i^k
\]

Next, with \( \hat{P}_i^k \) and \( K_{ij} \) defined in D.4 and D.5, letting \( \rho_{ij} \equiv [Y_j^k - \hat{\Delta}_j(i) - P_i] K_{ij} \tau_{vi} w_i^k / g_i \), Lemma C.7 gives

\[
\sqrt{N}(D_2^* - U_N) = o_p(1) \\
U_N \equiv \frac{1}{N} \sum_i (\hat{f}_i / \hat{g}_i - P_i) \tau_{vi} w_i^k / g_i = \frac{1}{N} \sum_i (\hat{f}_i - P_i \hat{g}_i) \tau_{vi} w_i^k / g_i = \frac{1}{N(N-1)} \sum_i \sum_{j \geq i} \rho_{ij}
\]

With \( \rho_{ij}^* = \frac{\rho_{ij} + \rho_{ji}}{2} \), \( U_N \) has the conventional second order U-statistic form. We could apply the standard approximation theory of U-statistics (Serfling (2009), Powell et al. (1989)) and obtain \( \sqrt{N}(U_N - \hat{U}_N) = o_p(1) \) where

\[
\sqrt{N}\hat{U}_N \equiv \frac{1}{\sqrt{N}} \sum_i (E[\rho_{ij}|Obs_i] + E[\rho_{ji}|Obs_i]) = T_1 + T_2
\]
It can be shown that for each term in $T_1$ (Need check):

$$E[\text{term}_i] \equiv E_i\{E[\rho_{ij}|\text{Obs}_i]\} = E[\rho_{ij}] = E_{V_i,V_j}\{E[\rho_{ij}|V_i,V_j]\} = 0 \quad (45)$$

$$\text{Var}[\text{term}_i] = O(1) \quad (46)$$

Therefore $T_1$ is $O(1)$ from standard sample mean property. $T_2 = 0$ from iterated expectation and the residual property of $w^k_i$ defined in A.6:

$$E[\rho_{ji}|\text{Obs}_i] = E_{V_j}\{E[\rho_{ji}|\text{Obs}_i,V_j]\} = 0 \quad (47)$$

Therefore, for the estimated gradient in (31), we have

$$\sqrt{N}G(\theta_0) = \sqrt{N}\frac{1}{N} \sum_i \tau_{iv} \frac{Y^k_i - P^k_i}{P^k_i} \frac{\partial P^k_i}{\partial \theta} + O_p(1) \quad (48)$$

Then, refer to the expression in (31), a central limited theorem shall apply to

$$\sqrt{N}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} H^{-1}_0 G_i + O_p(1) \quad (49)$$

$$G_i \equiv \sum_{k=1}^{L} \tau_{iv} \frac{Y^k_i - P^k}{P^k} \frac{\partial P^k}{\partial \theta} \quad (50)$$

Since the information equality $H_0 = E[G_i'G_i]$ holds here as in Klein and Sherman (2002), Theorem B.2 would follow then.

**Theorem B.3** (Consistency and Normality for Quantile Marginal effects). Under A.1-A.5, with the cumulative quantile marginal effect $CME^K_q$ and its estimator defined in D.8,

$$\sqrt{N}(GME^K_q - \hat{GME^K_q}) \sim N(0, \sum_{k=1}^{K-1} E[\psi^k_j\psi^k_j])$$

where $\psi^k_j \equiv \psi^k_{ij} + \psi^k_{2j} + \psi^k_{3j} + \psi^k_{4j}$. To define $\psi_{kj}$, recall the Bahadur representation $B_j$, and let $t_{qj}$ be the quantile trimming function for $Z$ in a particular quantile. Then, with $G_j$ as the true gradient
defined above and \( V_j = (V_F, V_B, Z) \) evaluated at \( \theta_0 \):

\[
\psi_{1j}^k \equiv \frac{E[t_{qj} \nabla_{\theta} P_j^k(V_F, V_B, Z + \delta; \theta_0)]] - E[t_{qj} \nabla_{\theta} P_j^k(V_F, V_B, Z; \theta_0)]]}{E[t_{qj}]} H_0^{-1} G_j
\]

(51)

\[
\psi_{2j}^k \equiv \{\nabla_q E[t_{qj} m_j(\theta_0)] - \nabla_q E[t_{qj} CME_q^k]\} \frac{B_j}{E[t_{qj}]}
\]

(52)

\[
\Psi_{3j}^k \equiv \frac{E[t_{qj} | V_F, V_B, Z + \delta; \epsilon_j] - E[t_{qj} | V_F, V_B, Z; \epsilon_j]}{E[t_{qj}]}
\]

(53)

\[
\psi_{4j}^k \equiv \frac{t_{qj} m_j(\theta_0) - E[t_{qj} CME_q^k]}{E[t_{qj}]} - \frac{t_{qj} - E[t_{qj}]}{E[t_{qj}]} GME_q^k
\]

(54)

Proof. For consistency, it can be shown that the estimated quantile trimming function \( \hat{t}_{qj} \) will converge in probability to the truth. Since \( GME_q^K \) is a continuous function of the estimated probability, and \( \sup_{\theta} | \hat{P}_k(\theta) - P_k(\theta) | \overset{p}{\rightarrow} 0 \) for all \( k \), consistency follows from continuous function theorem and Theorem B.1.

To show normality, note that the cumulative marginal effect \( \hat{CME}_q^K \) defined in (5) can be written as

\[
\hat{CME}_q^K \equiv \sum_{j=1}^{N} \hat{t}_{qj} \sum_{k=1}^{K-1} \frac{[\hat{P}_i^k(V_F, V_B, Z + \delta; \hat{\theta}) - \hat{P}_i^k(V_F, V_B, Z; \hat{\theta})]}{\sum_{j=1}^{N} \hat{t}_{qj}}
\]

(55)

\[
\Rightarrow \sqrt{N}(GME_q^K - \hat{GME}_q^K) = \sqrt{N} \sum_{k=1}^{K-1} (M^k_q - \hat{M}^k_q)
\]

(56)

Since all terms within the above summation have the same structure, it suffice to analyze just one term for any \( k \). To simplify notation, let

\[
\hat{m}^k_j \equiv \hat{P}_i^k(V_F, V_B, Z + \delta; \hat{\theta}) - \hat{P}_i^k(V_F, V_B, Z; \hat{\theta})
\]

(57)

\[
m^k_j \equiv P_i^k(V_F, V_B, Z + \delta; \theta_0) - P_i^k(V_F, V_B, Z; \theta_0)
\]

(58)

\[
\hat{N}_q \equiv \sum \hat{t}_{qj}
\]

(59)
and we proceed with the following decomposition:

\[
\sqrt{N}(\hat{M}_q^k - M_q^k) = \sqrt{N}\left(\frac{\sum_{j=1}^N \hat{t}_{qj} \hat{m}_j^k(\hat{\theta})}{\sum_{j=1}^N t_{qj}} - \frac{\sum_{j=1}^N t_{qj} m_j^k}{\sum_{j=1}^N t_{qj}}\right)
\]

\[
= \sqrt{N}(\Delta_1 + \Delta_2 + \Delta_3 + \Delta_4)
\]

(60)

\[
\Delta_1 = (N/\hat{N}_q) \frac{1}{N} \sum_j \hat{t}_{qj}(\hat{m}_j^k(\hat{\theta}) - m_j^k(\theta_0))
\]

\[
\Delta_2 = (N/\hat{N}_q) \frac{1}{N} \sum_j m_j^k(\theta_0)(\hat{t}_{qj} - t_{qj})
\]

\[
\Delta_3 = (N/\hat{N}_q)(\frac{1}{N} \sum_j t_{qj} m_j^k(\theta_0) - E[t_{qj}]CME_q^k)
\]

\[
\Delta_4 = (N/\hat{N}_q) \frac{1}{N} \sum_j (E[t_{qj}] - \hat{t}_{qj})CME_q^k
\]

For \(\Delta_1\), note that

\[
\sqrt{N}\Delta_1 \equiv \Delta_{11} + \Delta_{12}
\]

(61)

\[
\Delta_{11} \equiv (N/\hat{N}_q) \frac{\sqrt{N}}{N} \sum_j (\hat{t}_{qj} - t_{qj})(\hat{m}_j(\hat{\theta}) - m_j(\theta_0))
\]

(62)

\[
\Delta_{12} \equiv (N/\hat{N}_q) \frac{\sqrt{N}}{N} \sum_j t_{qj}(\hat{m}_j(\hat{\theta}) - m_j(\theta_0))
\]

(63)

\(\Delta_{11}\) is \(o_p(1)\) in a fashion identical to the term \(\Delta_{11}\) in Theorem 5 of Klein and Shen (2015). Turning to \(\Delta_{12}\), we further decompose \(\Delta_{12}\) into four pieces:

\[
\Delta_{12} = (N/\hat{N}_q)\sqrt{N}(T_1 + T_2 - T_3 - T_4)
\]

\[
T_1 = \frac{1}{N} \sum_{j=1}^N t_{qj}[\hat{P}_j^k(V_F, V_B, Z + \delta; \hat{\theta}) - \hat{P}_j^k(V_F, V_B, Z + \delta; \theta_0)]
\]

\[
T_2 = \frac{1}{N} \sum_{j=1}^N t_{qj}[\hat{P}_j^k(V_F, V_B, Z + \delta; \theta_0) - P_j^k(V_F, V_B, Z + \delta; \theta_0)]
\]

\[
T_3 = \frac{1}{N} \sum_{j=1}^N t_{qj}[\hat{P}_j^k(V_F, V_B, Z; \hat{\theta}) - \hat{P}_j^k(V_F, V_B, Z; \theta_0)]
\]

\[
T_4 = \frac{1}{N} \sum_{j=1}^N t_{qj}[\hat{P}_j^k(V_F, V_B, Z; \theta_0) - P_j^k(V_F, V_B, Z; \theta_0)]
\]

(64)
For $T_1$, from a Taylor expansion around $\theta_0$ and sup $|\nabla_\theta \hat{P}_i^k(\theta) - \nabla_\theta P_i^k(\theta)| \overset{p}{\to} 0$ (Lemma C.4), we have:

\[
\sqrt{N}T_1 = \frac{1}{N} \sum_{j=1}^{N} t_{qj} \nabla_\theta \hat{P}_j^k|_{\theta = \theta_0} \sqrt{N}(\hat{\theta} - \theta_0) + o_p(1) \tag{65}
\]

\[
= E[t_{qj} \nabla_\theta P_j^k(V_F, V_B, Z + \delta; \theta_0)] \sqrt{N}(\hat{\theta} - \theta_0) + o_p(1) \tag{66}
\]

Turning to $T_2$, in Lemma C.8 we show that $\sqrt{N}T_2$ can be written as the sum of two pieces $U_1$ and $U_2$. In Lemma C.9 we proceed to show $U_1$ is a degenerate third order U-stat which would vanish asymptotically. In Lemma C.10 we show that $U_2$ can be approximated by a sample mean of iid terms. Therefore we have:

\[
\sqrt{N}T_2 = \sqrt{N}(U_1 + U_2)
\]

\[
= \frac{1}{\sqrt{N}} \sum_{j=1}^{N} E[t_{qj}|V_F, V_B, Z + \delta]e_j^\delta + o_p(1) \tag{67}
\]

where $e_j^\delta \equiv Y_j^k - E[Y_j^k|V_F, V_B, Z + \delta]$. The proof strategy for $T_3$ and $T_4$ mimic the ones for $T_1$ and $T_2$:

\[
T_3 = E[t_{qj} \nabla_\theta P_j^k(V_F, V_B, Z; \theta_0)] \sqrt{N}(\hat{\theta} - \theta_0) + o_p(1) \tag{68}
\]

\[
T_4 = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} E[t_{qj}|V_F, V_B, Z]e_j + o_p(1) \tag{69}
\]

Combining (66)(67)(68)(69), we have

\[
\sqrt{N}\Delta_{12} = \sqrt{N}(T_1 - T_3 + T_2 - T_4)
\]

\[
= \frac{E[t_{qj} \nabla_\theta P_j^k(V_F, V_B, Z + \delta; \theta_0)] - E[t_{qj} \nabla_\theta P_j^k(V_F, V_B, Z; \theta_0)]}{E[t_{qj}]} \sqrt{N}(\hat{\theta} - \theta_0) \tag{70}
\]

\[
+ \frac{1}{E[t_{qj} \sqrt{N}]} \sum_{j=1}^{N} E[t_{qj}|V_F, V_B, Z + \delta]e_j^\delta - E[t_{qj}|V_F, V_B, Z]e_j + o_p(1) \tag{71}
\]

Both $\Delta_2$ and $\Delta_4$ in (60) are related to quantile estimation uncertainty. Turning to $\Delta_4$, we have,
up to $o_p(1)$:

$$\Delta_4 = (N/N_q) \frac{1}{N} \sum_j (\hat{t}_{qj} - E[t_{qj}]) CME_q^k$$

(72)

$$\Delta_4 = (N/N_q) \frac{1}{N} \sum_j (\hat{t}_{qj} - t_{qj}) CME_q^k + (N/N_q) \frac{1}{N} \sum_j (t_{qj} - E[t_{qj}]) CME_q^k$$

(73)

Both $\Delta_2$ and the first piece of $\Delta_4$ take the form $(N/N_q) \frac{1}{N} \sum W_j (\hat{t}_{qj} - t_{qj})$ where $W_j$ is either iid term over $j$ or constant. From Lemma 3 of Klein and Shen (2015), who exploits an important result from Pakes and Pollard (1989), the Bahadur representation, and the convergence of $N/N_q$ to $1/E[t_{qj}]$, we have:

$$\sqrt{N} \Delta_2 = \frac{1}{E[t_{qj}]} \nabla_q E[t_{qj} m_j(\theta_0)] \sqrt{N B} + o_p(1)$$

(74)

$$\sqrt{N} \Delta_4 = \frac{1}{E[t_{qj}]} \nabla_q E[t_{qj} GME_q] \sqrt{N B} + \frac{GME_q}{\sqrt{N E[t_{qj}]}} \sum_{j=1}^{N} (t_{qj} - E[t_{qj}]) + o_p(1)$$

(75)

Note that the normality of $B$, the sample average of $B_j$ has been derived by Bahadur (1966) and widely used in the econometric literature. The second piece of $\Delta_4$ is an average of iid terms and therefore could be handled by CLT.

$\Delta_3$ in (60) is exactly an sample average of iid terms minus its expectation. From Lindeberg-Levy CLT, we shall have

$$\sqrt{N} \Delta_3 = (N/N_q) \frac{1}{N} \sum_j t_{qj} m_j(\theta_0) - E[t_{qj}] GME_q$$

(76)

$$= \frac{1}{\sqrt{N E[t_{qj}]}} \sum_j (t_{qj} m_j(\theta_0) - E[t_{qj}] GME_q)$$

(77)

After incorporating the results of all four $\Delta$s and the asymptotic linear structure of $\sqrt{N}(\hat{\theta} - \theta_0)$
in (49), the vector of $\hat{M}_q^k - M_q^k$, where k=1,2,3...L, has an asymptotic linear form:

$$\sqrt{N}(M_q^k - \hat{M}_q^k) = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} [\Psi_{1j} + \Psi_{2j} + \Psi_{3j} + \Psi_{4j}] + o_p(1)$$  \hspace{2cm} (78)

$$\Psi_{1j}^k \equiv \frac{E[t_{qj} \nabla_0 P_j^k(V_F, V_B, Z + \delta; \theta_0)] - E[t_{qj} \nabla_0 P_j^k(V_F, V_B, Z; \theta_0)]}{E[t_{qj}]} H_0^{-1} G_j$$ \hspace{2cm} (79)

$$\Psi_{2j}^k \equiv \{\nabla_q E[t_{qj} m_j(\theta_0)] - \nabla_q E[t_{qj} CME_q^k]\} \frac{B_j}{E[t_{qj}]}$$ \hspace{2cm} (80)

$$\Psi_{3j}^k \equiv \frac{E[t_{qj} | V_F, V_B, Z + \delta] \epsilon_j^\delta - E[t_{qj} | V_F, V_B, Z] \epsilon_j}{E[t_{qj}]}$$ \hspace{2cm} (81)

$$\Psi_{4j}^k \equiv \frac{t_{qj} m_j(\theta_0) - E[t_{qj}] CME_q^k}{E[t_{qj}]} - \frac{t_{qj} - E[t_{qj}]}{E[t_{qj}]} CME_q^k$$ \hspace{2cm} (82)

and after applying the Central Limited Theorem to that vector, we have:

$$\sqrt{N} \begin{bmatrix} \hat{M}_1^1 - M_1^1 \\ \hat{M}_2^2 - M_2^2 \\ \vdots \\ \hat{M}_L^L - M_L^L \end{bmatrix} \xrightarrow{d} Z \sim N(0, \Omega)$$ \hspace{2cm} (83)

where $\Omega \equiv E[(\Psi_{1j}^k + \Psi_{2j}^k + \Psi_{3j}^k + \Psi_{4j}^k)'(\Psi_{1j}^k + \Psi_{2j}^k + \Psi_{3j}^k + \Psi_{4j}^k)]$ and each $\Psi_j^k$ is a L X 1 column vector calculated from the formula above. From (55), note that the object of interest $\sqrt{N}(\hat{G}ME_q^K - GME_q^K)$ can be obtained from the following linear transformation on the above vector. With $A = (1, 1, ..., 1, 0, 0, 0)$, an 1 X L dimension row vector with the first K-1 component equals one and the rest equals zero, we have

$$\sqrt{N}(\hat{G}ME_q^K - GME_q^K) = \sqrt{N} A \begin{bmatrix} \hat{M}_1^1 - M_1^1 \\ \hat{M}_2^2 - M_2^2 \\ \vdots \\ \hat{M}_L^L - M_L^L \end{bmatrix} \xrightarrow{d} AZ \sim N(0, A'\Omega A)$$ \hspace{2cm} (84)

Then the normality result follows. □
Appendix C  Intermediate lemmas

Lemma C.1 (Convergence of estimates to expectations). Let \( w \) be a \( K \) dimensional vector and assume that \( m(w) \) is a sample average of terms \( m(w; z_i) \), where \( z_i \) are i.i.d. Assume that with \( h \to 0 \) we have uniformly over \( N \):

\[
h^{r+1}|m(w; z_i)| < c, \quad r + 1 > 0 \quad \text{and} \quad h^s|\partial m(w; z_i)/\partial w| < c, \quad s > 0
\]

Let \( E[m(w)] \) be the expectation of \( m(w) \) taken over the distribution of \( z_i \). Then, with \( w \) in a compact set and for any \( \alpha \not= 0 \):

\[
N^{(1-\alpha)/2}h^{r+1}\sup| m(w) - E[m(w)] | \to 0 \quad \text{a.s.}
\]

Proof. See the proof of Lemma C.1 on pp 411 in Klein and Spady (1993). The proof utilizes important implication in Hoeffding (1963) and Bhattacharya (1967).

Lemma C.2 (Uniform Convergence for density and its derivatives). Let \( \hat{g} \) be a estimated density with 3 index and \( \nabla^r_{\theta} \hat{g} \) be the \( r \)-th density derivative with respect to \( \theta \), \( r = 0,1,2 \). With the kernal function \( K_{ij} \) defined in D.4. Let \( \hat{g}_i = \frac{1}{N-1}\sum_{j\not= i} K_{ij} \) as in D.4. If all \( x \)'s are bounded, then:

(C2.1) \( |\nabla^r_{\theta} \hat{g}(\theta) - E[\nabla^r_{\theta} \hat{g}(\theta)]| = O_p(N^{-\frac{1}{2}}h^{-r-3}) \)

(C2.2) \( |E[\nabla^r_{\theta} \hat{g}(\theta)] - \nabla^r_{\theta} g(\theta)| = O_p(h^2) \)

(C2.3) \( \sup| \nabla^r_{\theta} \hat{g}(\theta) - \nabla^r_{\theta} g(\theta) | = O(N^{-\frac{1}{2}}h^{-r-3}) \)

Proof. C2.1 directly from Lemma C.1, and C2.3 directly follows from C2.1 and C2.2 since

\[
\sup| \nabla^r_{\theta} \hat{g}(\theta) - \nabla^r_{\theta} g(\theta) | \leq \sup| \nabla^r_{\theta} \hat{g}(\theta) - E[\nabla^r_{\theta} \hat{g}(\theta)] | + \sup| E[\nabla^r_{\theta} \hat{g}(\theta)] - \nabla^r_{\theta} g(\theta) |
\]

The proof of C2.2 follows from the bias calculation of \( \nabla^r_{\theta} \hat{g}(\theta) \), which can be found in standard
literature such as Hansen (2009), so we just outline the proof for \( r = 0 \) here:

\[
E[\hat{g}(\theta)] = \int_{-\infty}^{\infty} K(u)g(\theta + hu)du = \int_{-\infty}^{\infty} K(u)(g(\theta) + hu)du \quad (86)
\]

Let \( u = \frac{V_i(\theta) - v(\theta)}{h} \)

After a second order Taylor expansion on \( h = 0 \)

\[
E[\hat{g}(\theta)] = \int_{-\infty}^{\infty} K(u)[g(v) + g'(v)hu + \frac{1}{2}g''(v)h^2u^2]du \quad (88)
\]

\[
= A_1 + A_2 + A_3 \quad (90)
\]

Given that \( K \) is the regular Gaussian kernel, \( A_1 = g(v(\theta)) \) because \( \int K(u)du = 1 \). \( A_2 = 0 \) since \( K \) is symmetric by D.4, so \( \int K(u)udu = 0 \). \( A_3 = O_p(h^2) \) because the Gaussian kernel has variance of 1 and \( g'' \) is bounded by A.5. Therefore, we have

\[
E[\hat{g}(\theta)] - g(\theta) = O_p(h^2) \quad (91)
\]

The proof for density derivatives are similar. \( \square \)

**Lemma C.3** (Convergence rate in the bias reduced model). *Refer to the bias correction model defined in (29), with \( \Delta_i(j) \) and its estimator defined in the bias corrected model in (29), let \( \hat{f}_i = \frac{1}{N-1} \sum_{j \neq i} (Y^k_i - \hat{\Delta}_i(j)) K_{ij} \). When all of the assumption and definition of Lemma C.2 holds, the following convergence rate holds for the bias reduced model:

\[
(C3.1) \quad \sup_{\theta} E\{\|\nabla_{\theta} f_i(\theta) - E[\nabla_{\theta} f_i(\theta)]\|^2\} = O_p(N^{-1}h^{-r-3})
\]

\[
(C3.2) \quad \sup_{\theta} |E[\nabla_{\theta}^r \hat{f}_i(\theta)] - P_i^k \nabla_{\theta}^r \hat{g}_i(\theta)| = O_p(h^4)
\]

**Proof.** Klein et al. (2015) shows he variance calculation in C3.1 is identical as the regular kernal estimator \( \hat{f}_i = \frac{1}{N-1} \sum_{j \neq i} Y^k_i K_{ij} \). From standard result, the order of bias in C3.2 is the same for
density and its derivatives. We illustrate for the case \( r=0 \):

\[
E[\hat{f}_i - P^k_i(\theta)\hat{g}] = E\left[\frac{1}{(N - 1)h} \sum_{j \neq i} [(Y^k_i - \hat{\Delta}_i(j))K_{ij} - P^k_i K_{ij}]\right]
\]

\[= E[(Y^k_i - \hat{\Delta}_i(j) - P^k_i)K_{ij}] \]  

\[= E[(\Delta_i(j) - \hat{\Delta}_i(j) + \epsilon_i)K_{ij}] \]  

\[= E[(\Delta_i(j) - \hat{\Delta}_i(j))K_{ij}] \]  

(92)

(93)

(94)

(95)

Recall from D.1 that \( \Delta_i(j) \equiv P^k_i(V_i(\theta_0) - P^k_j) \)

\[= E\left\{[P^k_i(V_i(\theta_0)) - P^k_j(V_j) - \hat{P}^k_i(V_i(\theta_0)) + \hat{P}^k_j(V_j)]K_{ij}\right\} \]  

\[= \int [P^k_i(V_i(\theta_0)) - P^k_j(V_j) - \hat{P}^k_i(V_i(\theta_0)) + \hat{P}^k_j(V_j)]K_{ij} \cdot g(V_j) dV \]  

(96)

(97)

(98)

Note that \( \hat{P}^k \) is a ratio of kernel estimators defined in D.4. From standard results, the bias of this kernel estimator is of order \( h^2 \). However, the bias could be reduced by the diff-in-diff structure in (98) when \( V_i \) are close to \( V_j \) (Klein et al. (2015)). Let \( t_1 \) be some point around \( V_j + uh \) and \( t_2 \) some point around \( V_j \). From the same change of variable technique, the Mean Value Theorem, and a standard Taylor expansion of \( g(V_j+uh) \) round \( h=0 \), the above expectation reduces to:

\[= \int (P^k(V_j + uh) - P^k(V_j) - \hat{P}^k(V_j + uh) + \hat{P}^k(V_j))K_{ij} \cdot g(V_j + uh) du \]  

\[= \int \left[ \frac{1}{2} g''(t_1)(uh)^2 - \frac{1}{2} g''(t_2)(uh)^2 + o(h^4) \right] K_{ij} \cdot g(V_j + uh) du \]  

\[\leq B \int (uh)^3 K(u)[g(V_j) + uh \cdot g'(uh)^2 \cdot g''(uh)] du \]  

\[= T_1 + T_2 + T_3 \]  

(99)

(100)

(101)

(102)

for some constant \( B \) which relate to the upper bound of \( g''' \). The first and third term in this expectation vanish because:

\[T_1 = B \int (uh)^3 K(u)g(w) du = 0 : \int u^3 K(u) du = 0 \]  

(103)

\[T_3 = B \int (uh)^5 K(u)g''(uh) du = 0 : \int u^5 K(u) du = 0 \]  

(104)
For $T_2$, since we assume $g'$ is uniformly bounded from above: $\sup_w |g'(w)| < C, C > 0$ and finite. Note that the central fourth moment of the Gaussian pdf is 3. It then follows that:

$$T_2 = B \int (uh)^4 K(u) g' du$$

$$\leq BCh^4 \int u^4 K(u) du$$

$$= 3BCh^4$$

Therefore, it follows that $T_2 = O_p(h^4)$, and hence $\sqrt{N}E[\hat{f}_{ki} - P_{ki}(\theta)\hat{g}] = O_p(N^{1/2}h^4)$. □

**Lemma C.4** (Convergence of estimated probability and their derivatives). Referring D.5 for the definition of the regular expectation $\hat{P}_k = \hat{f}/\hat{g}$, and the “adjusted expectation” with $\hat{f}_i = \frac{1}{N-1} \sum_{j \neq i}(Y_i - \Delta_i(j))K_{ij}$. Let $\nabla^t_{\theta} \hat{P}$ be the t-th derivative with respect to $\theta$, $t = 0,1,2$. Then, for $h = N^{-r}$ and under the conditions of Lemma C.1 and Lemma C.2,

(C4.1) Uniformly in $w$ and $\theta$: $\sup|\nabla^t_{\theta} \hat{P}_{ki}(\theta) - \nabla^t_{\theta} P_{ki}(\theta)| = O(N^{-\frac{1}{2}}h^{t-3})$

(C4.2) $\frac{1}{N} \sum_i (\hat{P}_k^i(\theta) - P_k^i(\theta))^2 = O_p(N^{-1}h^{-3}) + O_p(h^8)$

**Proof.** The proof for the single index version of C4.1 can be found in Klein and Spady (1993) (Lemma 4 pp.414), in which the author showed the convergence rate of the ratio $\hat{f}/\hat{g}$ shall follow the uniform convergence rate of $\hat{g}$ as calculated in Lemma C.3. It can be shown that $\hat{f}_i$ does not converge slower than $\hat{g}_i$, so the same rate will follow for the adjusted expectation.

To show C4.2, note that $\frac{1}{N} \sum_i (\hat{P}_k^i(\theta) - P_k^i(\theta)) = \frac{1}{N} \sum_i (\hat{f}_i - \hat{g}_i) = \sup_{\hat{f}}(\frac{1}{N}) \frac{1}{N} \sum_i (\hat{f}_i - P_k^i(\theta)\hat{g}_i)$. Since we trim $\hat{g}_i$ away from zero, $1/inf(\hat{g}_i)$ is clearly bounded from above. Therefore, for some constant $B$, from the convergence rate in Lemma C.3.

$$\frac{1}{N} \sum_i (\hat{P}_k^i(\theta) - P_k^i(\theta))^2 \leq B^2 \frac{1}{N} \sum_i (\hat{f}_i - P_k^i(\theta)\hat{g}_i)^2$$

$$= B^2 \frac{1}{N} \sum_i (\hat{f}_i - E[\hat{f}] + E[\hat{f}] - P_k^i(\theta)\hat{g}_i)^2$$

$$= B^2 \frac{1}{N} \sum_i (\hat{f}_i - E[\hat{f}])^2 + B^2 \frac{1}{N} \sum_i (E[\hat{f}] - P_k^i(\theta)\hat{g}_i)^2$$

$$= O_p^2(1)(O_p(N^{-1}h^{-3}) + O_p(h^8))$$
Lemma C.5 (Convergence of Hessian). Assume the window size $h = N^{-r}$ and $1/16 < r < 1/10$. Then, under the conditions of lemma C.1 and C.2 and with $\theta^+ \in [\hat{\theta}, \theta_0]$, 

$$\hat{H}(\theta^+)^{-1} \xrightarrow{p} H_0 = E[H(\theta_0)]$$

Proof. Given that the Hessian is a continuous function on $\theta$, the desired argument $\hat{H}(\theta^+) \xrightarrow{p} E[H(\theta_0)]$ would follow if we have the following two conditions holds: (a) $\theta^+ \xrightarrow{p} \theta_0$ (b) $\sup |\hat{H}(\theta) - E[H(\theta)]| \xrightarrow{p} 0$. Condition (b) implies that $\hat{H}(\theta_0) \xrightarrow{p} E[H(\theta_0)]$. If (a) holds, then by the continuous mapping theorem we have $\hat{H}(\theta^+) \xrightarrow{p} \hat{H}(\theta_0) \xrightarrow{p} E[H(\theta_0)]$.

Condition (a) directly follows from consistency as $\theta^+$ is some intermediate point between $\theta_0$ and $\hat{\theta}$. To show (b), note that:

$$\sup |\hat{H}(\theta) - E[H(\theta)]| \leq \sup |\hat{H}(\theta) - H(\theta)| + \sup |H(\theta) - E[H(\theta)]| \tag{112}$$

$$\leq T_1 + T_2 \tag{113}$$

$T_2 \xrightarrow{p} 0$ from Lemma C.1. Note that the hessian, by definition, is the second derivative of the quasi-log-likelihood function:

$$H(\theta_0) \equiv \frac{1}{N} \sum_i \left( \frac{Y_{i,k}^k}{P_{k,i}} \nabla_{\theta^0} P_{k,i} - \frac{Y_{i,k}^k}{P_{k,i}^2} \nabla_{\theta} P_{k,i} \right)$$

To make $T_1$ uniformly converge to 0, we need $\nabla_{\theta^0} \hat{P}_i^k$ uniformly converge to its associated estimand for $t = 0, 1, 2$. Lemma C3.2 ensures that given the window condition $1/12 < r < 1/10$, this is the case. Therefore we have $\hat{H}(\theta^+)^{-1} \xrightarrow{p} H_0 = E[H(\theta_0)]$. \hfill \square

Lemma C.6 (Double Convergence). Let $a_i, b_i$ be some iid quantity, and $\hat{a}_i, \hat{b}_i$ be their estimator respectively. If $\frac{1}{N} \sum_i (\hat{a}_i - a_i)^2 = O_p(N^{-\alpha_1})$, $\frac{1}{N} \sum_i (\hat{b}_i - b_i)^2 = O_p(N^{-\alpha_2})$, then $\frac{1}{N} \sum_i (\hat{a}_i - a_i)(\hat{b}_i - b_i) = O_p(N^{-\alpha_2 - \alpha_1})$
Proof. The proof follow directly from the Cauchy-Schwarz:

\[
\frac{1}{N} \sum_i (\hat{a}_i - a_i) (\hat{b}_i - b_i) \leq \sqrt{\frac{1}{N} \sum_i (\hat{a}_i - a_i)^2} \sqrt{\frac{1}{N} \sum_i (\hat{b}_i - b_i)^2}
\]

(114)

Lemma C.7 (Convergence rate for double sums). With \( h = O(h^{-r}) \), \( 1/12 < r < 1/10 \), then with \( \hat{P} \equiv \hat{f}_{ki}/\hat{g}_{ki} \) and \( \hat{M} \) as (i) \( \hat{g} \) or (ii) \( \hat{w} \equiv \frac{\partial \hat{P}_{ki}}{\partial \theta} / \hat{P} \):

\[
\Delta \equiv \sqrt{N} \frac{1}{N} \sum_i (P - \hat{P})(M - \hat{M}) = O_p(1)
\]

Proof. To show (i), it suffice to show that \( \sqrt{N} \frac{1}{N} \sum_i (\hat{f} - P\hat{g})(\hat{g} - \hat{g})/g\hat{g} = O_p(1) \). It is important to note that as long as \( \sqrt{N} \langle (f - P\hat{g})(g - \hat{g}) \rangle = O_p(1) \), the above triple product will be \( O_p(1) \) due to Slusky’s theorem since \( \hat{g} \) converges to \( g \) at some rate. Apply the double convergence result on the above quantity, from the convergence rates in Lemma C.2-C.4, we have:

\[
\sqrt{N} \frac{1}{N} \sum_i (\hat{f} - P\hat{g})(\hat{g} - \hat{g}) \leq \sqrt{N} \sqrt{\frac{1}{N} \sum_i (f - P\hat{g})^2} \frac{1}{N} \sum_i (g - \hat{g})^2 \leq \sqrt{N} \sqrt{(O_p(h^8) + O_p(1/Nh^3))(O_p(h^4) + O_p(1/Nh^3))} = O_p(1)
\]

given that \( 1/12 < r < 1/6 \)

To show (ii), apply (i) and note it suffice to show that \( \sqrt{N} \langle (f - P\hat{g})(w - \hat{w}) \rangle = O_p(1) \).

Lemma C.8. Refering to \( T_2 \) in (64), \( \sqrt{NT_2} = \sqrt{N}(U_1 + U_2) + o_p(1) \), where

\[
U_1 \equiv \frac{1}{\sqrt{N}(N-1)(N-2)} \sum_{j=1}^{N} \sum_{i \neq j} \sum_{s \neq i \neq j} \rho_{jirs}
\]

(119)

\[
U_2 \equiv \frac{1}{\sqrt{N}} \sum_{j=1}^{N} E[t_{ij}]|V_F, V_B, Z + \delta| \epsilon_j^\delta + o_p(1)
\]

(120)

\[
\rho_{jirs} = \frac{\delta_j}{g_j} K_{ij} \left[ \frac{1}{g_j} (Y_s K_{js} - P_j K_{js}) - \frac{1}{g_i} (Y_s K_{is} - P_j K_{is}) \right]
\]

(121)
Proof. To simplify notation, let $\delta_j = t_{gj}$ and $P_j = P^h_j(V_F, V_B, Z + \delta; \theta_0)$.

\[
\sqrt{NT_2} = \frac{\sqrt{N}}{N} \sum_{j=1}^N [\hat{P}_j - P_j] \delta_j = \frac{\sqrt{N}}{N} \sum_{j=1}^N \left[ \frac{f_j}{g_j} - \frac{\hat{f}_j}{\hat{g}_j} \right] \delta_j
\]

\[
= \frac{\sqrt{N}}{N} \sum_{j=1}^N \left[ \frac{\hat{f}_j - P_j \hat{g}_j}{\hat{g}_j} \right] \delta_j + o_p(1)
\]

\[
= \frac{\sqrt{N}}{N(N-1)} \sum_{j=1}^N \sum_{i \neq j} \left[ Y_i - \hat{\Delta}_i(j) - P_j K_{ij} \right] \delta_j
\]

\[
= \frac{\sqrt{N}}{N(N-1)} \sum_{j=1}^N \sum_{i \neq j} \left[ Y_i - \hat{P}_i + \hat{P}_j - P_j \right] K_{ij} \delta_j
\]

\[
= \frac{\sqrt{N}}{N(N-1)} \sum_{j=1}^N \sum_{i \neq j} \left[ P_i - \hat{P}_i + \epsilon_i + \hat{P}_j - P_j \right] K_{ij} \delta_j
\]

\[
= \frac{\sqrt{N}}{N(N-1)} \sum_{j=1}^N \sum_{i \neq j} \left[ \Delta_i(j) - \hat{\Delta}_i(j) + \epsilon_i \right] K_{ij} \delta_j
\]

The equality of (122) and (123) follows from double convergence. For the transformation from (123)-(126), recall that we have a $\delta$-localized model: (The same localized model but conditioning on $Z+\delta$... Notation abbreviation: $V_i^\delta \equiv (V_{Fi}, V_{Bi}, Z_i + \delta)$)

\[
Y_i = P(V_i^\delta) + \epsilon_i^\delta \quad \text{(Note: } \epsilon_i^\delta = Y_i - E[Y_i|V_i^\delta])
\]

\[
= P(V_j^\delta) + P(V_i^\delta) - P(V_j^\delta) + \epsilon_i^\delta
\]

\[
\equiv P(V_j^\delta) + \Delta_i(j) + \epsilon_i
\]

and a localized bias reducing estimator:

\[
\hat{P}(V_j^\delta) = \frac{\hat{f}_j}{\hat{g}_j} = \frac{1}{(N-1)h} \sum_{i \neq j} (Y_i - \hat{\Delta}_i(j)) K_{ij}
\]

(\text{second stage estimator})

\[
K_{ij} \equiv K \left( \frac{V_j^\delta - V_i^\delta}{h} \right)
\]

\[
\hat{\Delta}_i(j) \equiv \hat{P}_i - \hat{P}_j \quad \text{(difference of first stage estimator)}
\]

Proceed from (127), note that both $\hat{P}_i, \hat{P}_j$ could be estimated by the first stage kernel estimator, so
the difference of $\Delta s$ in (127) could be replaced by

$$
\Delta_i(j) - \hat{\Delta}_i(j) = \left( \frac{f_j}{g_j} - \frac{\hat{f}_j}{g_j} \right) - \left( \frac{f_i}{g_i} - \frac{\hat{f}_i}{g_i} \right)
$$

(134)

$$
= \frac{1}{g_j}(\hat{f}_j - P_j\hat{g}_j) - \frac{1}{g_i}(\hat{f}_i - P_i\hat{g}_i)
$$

(135)

$$
= \frac{1}{(N-2)g_j} \sum_{s \neq i \neq j} (Y_sK_{js} - P_jK_{js}) - \frac{1}{(N-2)g_i} \sum_{s \neq i \neq j} (Y_sK_{is} - P_iK_{is})
$$

(136)

$$
= \frac{1}{(N-2)g_j} \sum_{s \neq i \neq j} C_{sj} - \frac{1}{(N-2)g_i} \sum_{s \neq i \neq j} C_{si}
$$

(137)

So substitute the expression in (137) into (127) yields the following:

$$
\sqrt{N}U = \sqrt{N}(U_1 + U_2)
$$

(138)

$$
\sqrt{N}U_1 = \frac{\sqrt{N}}{N(N-1)(N-2)} \sum_{j=1}^{N} \delta_j \sum_{i \neq j} K_{ij} \left[ \frac{1}{g_j} \sum_{s \neq i \neq j} C_{sj} - \frac{1}{g_i} \sum_{s \neq i \neq j} C_{si} \right]
$$

(139)

$$
\sqrt{N}U_2 = \frac{\sqrt{N}}{N(N-1)} \sum_{j=1}^{N} \delta_j \sum_{i \neq j} \epsilon_i K_{ij}
$$

(140)

**Lemma C.9.** Referring to the proof in Lemma C.8, $\sqrt{N}U_1 = o_p(1)$

**Proof.** With the definition of $U_1$ in (139), and

$$
\rho_{jis} = \frac{\delta_j}{g_j} K_{ij} \left[ \frac{1}{g_j} (Y_sK_{js} - P_jK_{js}) - \frac{1}{g_i} (Y_sK_{is} - P_jK_{is}) \right]
$$

(141)

we have:

$$
\sqrt{N}U_1 \equiv \frac{1}{\sqrt{N}(N-1)(N-2)} \sum_{j=1}^{N} \sum_{i \neq j} \sum_{s \neq i \neq j} \rho_{jis}
$$

(142)

letting

$$
\rho_{jis}^* = [\rho_{jis} + \rho_{ijs} + \rho_{isj} + \rho_{sis} + \rho_{sji}] / 6
$$

(143)
we can write (142) as a centered U-statistic:

\[
\sqrt{N}U_1 = \sqrt{N} \left( \frac{N}{3} \right)^{-1} \sum_{j=1}^{N} \sum_{i \neq j} \sum_{s \neq i \neq j} \rho_{jis}^*
\] (144)

From the approximation theory of Serfling (2009) and Powell et al. (1989),

\[
\sqrt{N}(U_1 - \hat{U}_1) = o_p(1) \quad (145)
\]

\[
\hat{U}_1 = \frac{1}{N} \sum_{j=1}^{N} E[\rho_{jis}^*|\text{obs}_j]
\] (146)

we show \(\sqrt{NU}_1\) is \(o_p(1)\) through the following steps:

(a) \(E[\rho_{jis}|\text{obs}_j] = o(N^{-1/2})\)

(b) \(E[\rho_{jis}^*] = o(N^{-1/2})\)

(c) \(\sqrt{N}\hat{U}_1\) is \(o_p(1)\)

For (a), we may write \(E[\rho_{jis}|\text{obs}_j]\) as \(E_1 - E_2\) where

\[
E_1 \equiv E\left\{ \frac{\delta_j}{g_j} K_{ij} \frac{1}{g_j} (Y_s K_{js} - P_j K_{js})|\text{obs}_j \right\}
\] (147)

\[
E_2 \equiv E\left\{ \frac{\delta_j}{g_j} K_{ij} \frac{1}{g_i} E[(Y_s K_{is} - P_i K_{is})|V_i]|\text{obs}_j \right\}
\] (148)

For the interior expectation in \(E_2\) term, we have:

\[
E[(Y_s K_{is} - P_i K_{is})|V_i] = E_{V_i} E[(Y_s K_{is} - P_i K_{is})|V_i, V_s]
\] (149)

\[
= E_{V_i} E(E[Y_s|V_s] K_{is} - P_i K_{is})|V_s]
\] (150)

\[
= \int (P_s - P_i) K \left( \frac{V_i - V_s}{h} \right) g(V_s) d(V_s)
\] (151)

\[
= \int (P(V_i + zh) - P(V_i)) K(z) g(V_i + zh) d(z)
\] (152)

\[
= \frac{1}{2} \left[ g(V_i) \sup_w (\nabla^2 P(w)) + \sup_w (\nabla P(w) \nabla g(w)) \right] h^2 + O(h^4)
\] (153)

where \(\nabla\) and \(\nabla^2\) denote a first(second) derivative taken with respect to \(V_i\). Substituting this expres-
sion into (148) yields:

\[
E_2 = E\left\{ \frac{\delta_j}{g_j} K_{ij} \frac{1}{g_i} \left[ \frac{g(V_i) \sup_w (\nabla^2 P(w))}{2} + \sup_w (\nabla P(w) \nabla g(w)) h^2 + O(h^4) \right] \right\}_{\text{obs}_j} \tag{154}
\]

\[
= \frac{\delta_j}{g_j} E\left\{ K_{ij} \frac{1}{g_i} \left[ \frac{g(V_i) \sup_w (\nabla^2 P(w))}{2} + \sup_w (\nabla P(w) \nabla g(w)) h^2 + O(h^4) \right] \right\}_{\text{obs}_j} \tag{155}
\]

\[
= h^2 \frac{\delta_j}{g_j} \frac{g(V_j) \sup_w (\nabla^2 P(w))}{2} + \sup_w (\nabla P(w) \nabla g(w)) + O(h^4) \tag{156}
\]

Turning to the \( E_1 \) term, through a similar derivation, we shall have:

\[
E_1 = E\left\{ \frac{\delta_j}{g_j} K_{ij} \frac{1}{g_j} \left[ \text{obs}_j \right] \right\} E\left\{ Y_s K_{is} - P_j K_{is} \right\}_{\text{obs}_j} \tag{157}
\]

\[
= \frac{\delta_j}{g_j} \left[ \frac{g(j)}{2} \sup_w (g''(w)) + O(h^4) \right] \left[ \frac{g(V_j) \sup_w (\nabla^2 P(w))}{2} + \sup_w (\nabla P(w) \nabla g(w)) h^2 + O(h^4) \right] \tag{158}
\]

\[
= h^2 \frac{\delta_j}{g_j} \frac{g(V_j) \sup_w (\nabla^2 P(w))}{2} + \sup_w (\nabla P(w) \nabla g(w)) + O(h^4) \tag{159}
\]

Therefore we have \( E_1 - E_2 = O(h^4) = o(N^{-1/2}) \) with \( h = O(N^{-1/7.99}) \).

For (b), the proof follows immediately from (a) and the fact that all terms in \( \rho^*_ijk \) have the same unconditional expectations. For (c), since (b) would imply that \( E[\sqrt{N}U_1] = o_p(1) \) and it can be shown that \( \text{Var}(\sqrt{N}U_1) = o_p(1) \), the result \( \sqrt{N}U_1 = o_p(1) \) now follows. \( \square \)

**Lemma C.10.** the \( U_2 \) piece : \( \sqrt{N}U_2 = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} E[t_q | V_j^\delta] e_j^\delta + o_p(1) \)
Proof. Refer to $U_2$ in (140), defining $\rho_{ij} \equiv \frac{\delta_i}{g_i} \epsilon_i K_{ij}$, we have:

\[
\sqrt{N}U_2 = \frac{1}{\sqrt{N(N-1)}} \sum_{j=1}^{N} \sum_{i \neq j} \frac{\delta_j}{g_j} \epsilon_i K_{ij} 
\]

\[
= \sqrt{N} \left( \begin{array}{c} N \end{array} \right)^{-1} \sum_{j=1}^{N} \sum_{i \geq j} [\rho_{ji} + \rho_{ij}]/2 \tag{160}
\]

\[
= 2 \sqrt{N} \sum_{j=1}^{N} E[\rho_{ji} + \rho_{ij}|Y_j, X_j]/2 + o_p(1) \tag{161}
\]

\[
= \frac{1}{\sqrt{N}} \sum_{j=1}^{N} E[\rho_{ij}|Y_j, X_j] + o_p(1) \tag{162}
\]

\[
= \frac{1}{\sqrt{N}} \sum_{j=1}^{N} E\left[\frac{\delta_i}{g_i} K_{ji}|Y_j, X_j\right] \epsilon_j + o_p(1) \tag{163}
\]

\[
= \frac{1}{\sqrt{N}} \sum_{j=1}^{N} E\left[E\left[\delta_i|V_i\right] K_{ji}|Y_j, X_j\right] \epsilon_j + o_p(1) \tag{164}
\]

\[
= \frac{1}{\sqrt{N}} \sum_{j=1}^{N} E[\delta_j|V_j^\delta] \epsilon_j + o_p(1) \tag{165}
\]

The third equality follows from standard U-statistics projection theorem. (165) follows from an iterated expectation conditioning on $V_i$. The last equality follows from a Taylor expansion on $E[\delta_i|V_i]$. When defining the localized model, recall that we let $\epsilon_j^\delta = Y_j - E[Y_j|V_j^\delta]$, therefore:

\[
\sqrt{N}U_2 = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} E[\delta_j|V_j^\delta][Y_j - E[Y_j|V_j^\delta]] + o_p(1) \tag{166}
\]