On the Heterogeneous Effects of Non-credit-related Information in Online P2P Lending: A Quantile Regression Analysis

Sirong Luo, Dengpan Liu, Yinmin Ye

Outline

- Introduction and Motivation
- Literature Review
- Quantile Regression Model
  - Standard Quantile
  - Binary Quantile
- Data Description
- Parameter Estimation and Heterogeneity
  - Different Heterogeneity Effects
  - Market Performance: 1) Funded Probability Model, 2) Borrow Interest Rate Model
  - Loan Performance: 1) Default Probability Model, 2) Loss Model
- Model Performance and Investment Optimization
  - Prediction Performance Comparison
  - Investment Decision Optimization

Introduction and Motivation

- Industrial Practice:
  - Peer to Peer (P2P) Lending has proven to be a very viable business model. As of 2014, Prosper and Lending Club, the two largest P2P lending platforms founded in 2005 and 2006, respectively, have originated over $6 billion in loans
  - Information asymmetry between borrowers and lenders may contribute to such relatively high default rates
  - Non-credit-related information could be used as a supplement to existing credit information (e.g., credit rating) to better infer the credibility of a borrower.

- Motivation:
  - The heterogeneous effect of non-credit-related information has not been addressed in the literature
  - A better understanding of the different relationships between the underlying borrower's information and credit market outcomes can help investors to make better investment decisions

The Literature Review

- Online P2P Lending
  - Herding Behaviors: Herzenstein et al. (2011) and Zhang and Liu (2013)
  - Reputation and Friendship Network: Ghasemkhani et al. (2013), and Lin et al. (2013) , Collier and Hampshire (2010)
  - Disintermediation of financial markets: Hume and Wright (2006)
- Heterogeneity
  - Heterogeneity is a critical issue: 1) Marketing: Allenby and Rossi (1999); 2) Economics: Laporte et al. (2010) ;
  - Individual Heterogeneities: Chintagunta (1991), Gonul and Srinivasan (1993), Hyslop (1999);
- Quantile regression and Heterogeneous effects
  - Quantile regression: Koenker and Bassett (1978), Koenker (2005), Abreveya (2001), Kordas (2006)
  - Heterogeneous effect: Buchinsky (2009) , Arias et al. (2001) , Engle and Manganelli (2004)

Quantile Regression Model

- Standard Quantile Regression Model:
  - A linear quantile regression estimates the τ-th conditional quantile Q_τ for a given x_i, i.e., $Q_\tau(Y_i|x_i)$, with a linear predictor $x_i^T\beta(\tau)$ for a given different x_i , where $\beta(\tau)$ is a regression coefficient vector for x_i, and x_i^T is the transpose of x_i. Letting $z = Y_i - x_i^T\beta(\tau)$ denote the residuals of estimation, for the τ-th conditional quantile Q_τ, $\beta(\tau)$ can be estimated by solving the minimization problem below:

$$\mathrm{Min}_{\beta(\tau)} \sum_{i=1}^{n} \rho_\tau\left(Y_i - x_i^T\beta(\tau)\right)$$

    where the loss function $\rho_\tau(z) = z\big(\tau - I(z < 0)\big)$ measures the estimation errors of $\beta(\tau)$, and $I(\cdot)$ is an indicator function, which is 1 if $z < 0$, and 0 otherwise.

  - $Q_\tau(Y|x) = \beta_0(\tau) + \beta_1(\tau)\ \mathrm{Listing} + \beta_2(\tau)\ \mathrm{Member} + \beta_3(\tau)\ \mathrm{Friendship} + \beta_4(\tau)\ \mathrm{Group}$
  - The model is used to estimate borrow rate, loss and profit

Quantile Regression Model

- Binary Quantile Model:
  - General Model
    - Manski (1975)

$$\begin{cases} Y_i^* = x_i^T \beta(\tau) + \varepsilon_i \\ Y_i = 1 \text{ if } Y_i^* \geq 0 \text{ and } Y_i = 0, \text{otherwise} \end{cases}$$

    - Probability Estimation: Kordas (2006)
  - Specification:
    - $Q_\tau(Y^*|x) = \beta_0(\tau) + \beta_1(\tau) \text{ Listing} + \beta_2(\tau) \text{ Member} + \beta_3(\tau) \text{ Friendship} + \beta_4(\tau)$
    - $\text{Probability}(Y = 1|x) = 1 - \tau, \text{ where } \tau = \underset{\theta}{\operatorname{argmin}} Q_\theta(Y^*|x) > 0$,
    - Bayesian Based Approach: Benoit and Van den Poel (2012)
    - The model is used to estimate the probability of funding and default

Data

- A unique sample collected in Oct, 2012 from Prosper.Com:
  - 5 datasets: Bid, Listing, Loan, Member and Group.
  - Control variables: Listing review requirement, Listing Option, category group
  - 47 variables selected from 18 original variables and 76 derived variables
  - Collinearity is controlled using VIF
  - Non-Credit-Related variables:
    - Listing variables: bid counts, group leader rewards, bid maximum rate,
    - Member variables: Homeownership, social roles, group association,
    - Friendship variables: the number of friends, the number of friends with roles, and the number of friends as borrowers;
    - Behavior variables about the current and historical behaviors of the member's friends and group, e.g., the number of friends who have paid off loans, the total bids submitted by the member's friends, the size of the member's group, and the total amounts of loans borrowed by the member's friends up to the listing month.

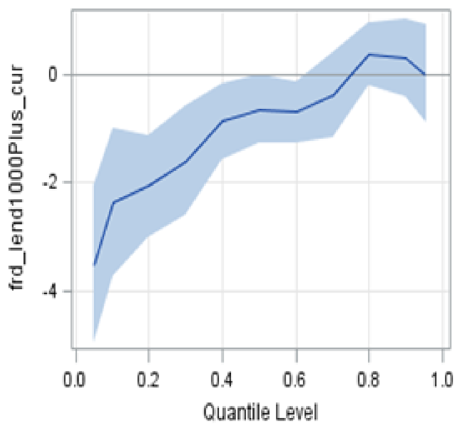
Parameter Estimation and Heterogeneity

- Different Heterogeneity effect Patterns are identified:
  - The *Monotone* pattern: the effect of non-credit-related information monotonically increases or decreases with quantiles
  - The *Cross* pattern: the effect exhibited in lower quantile listings is completely opposite to that at higher quantiles
  - The *Side* pattern: the effect is only significant at either low or high quantiles
  - The *Tail* pattern, where the effect is significant (or more pronounced) only for the lowest and highest quantiles.
  - The *Central* pattern, where the effect is only significant at the middle quantiles.
  - The relative importance of a particular covariate compared to other covariates may vary with the quantiles.

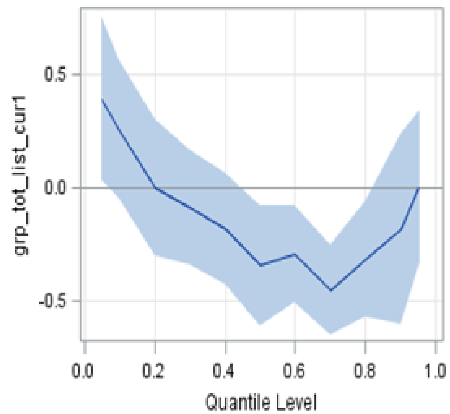
Parameter Estimation and Heterogeneity

- Different Heterogeneity effect Pattern are identified:

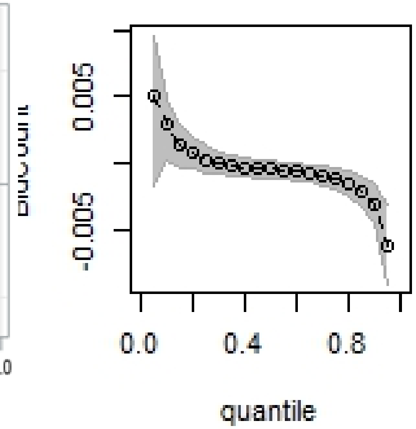
Monotone effect in Rate model



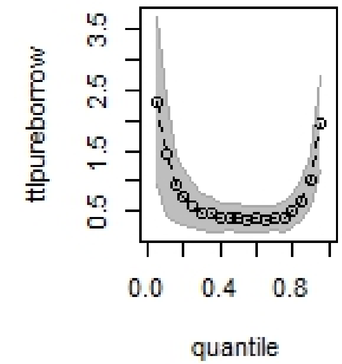
Central effect in Rate model



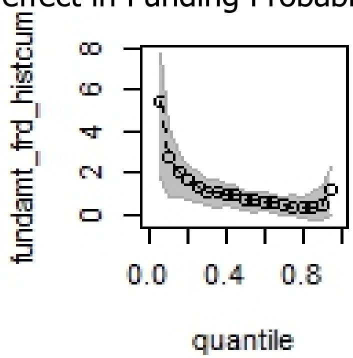
Cross Effect in Risk model



Tail Effect in Risk model



Side effect in Funding Probability Model



Parameter Estimation and Heterogeneity

- New insight about the non-credit-related information are found:
  - Market Performance Model--**Probability of Funding**
    - The effect of friendship depends more on the number of quality friends (e.g., friends with roles) than the total number of friends.
    - If a member has more friends with borrower roles or if her friends have a larger total funded amount, her listing has a higher probability of being funded. But this effect is only significant for the quantiles below or equal to 0.5
    - The effects of the *GroupleaderRewardRate* are only significant at the middle quantile levels
    - At the low quantile, the effect of *the total number of friends with borrower roles* is significant, while that of *the total number of friends having lent more than $1,000* is not; however, at the high quantile, the latter, but not the former, is significant. Thus, the borrowing friends are more related to listings with lower funding chance. However, the lendinding friends are more related to listings with higher funding chance.

Parameter Estimation and Heterogeneity

- New insight about the non-credit-related information are found:
  - Market Performance Model— **Borrow Interest Rate**
    - If the member has more friends with roles, then it implies that her friendship is more engaged and more responsible in the online microloan platform, and thus, the interest rate is lower.
    - If the member of the funded listing has more friends who have default loans, or her friends have borrowed many money before, the interest rate of the funded listing is higher.
    - If the member is a group leader, then the borrowing rate for her funded listing is lower, and such an effect is also more evident at lower quantiles (Monotone)
    - the effects of *GroupleaderRewardRate* are positive and significant at the lower and higher quantiles, whereas the effects are insignificant at the middle quantiles (Tail).

Parameter Estimation and Heterogeneity

- New insight about the non-credit-related information are found:
  - Loan Performance Model— **Probability of Default**
    - If a member has more friends who are pure borrowers or leaders, have default loans, or currently have listings, then the default risk of her loan is also higher.
    - If the member's group currently has more listings, the default risk is also higher.
    - If the member has more friends who have roles and funded listings, or if their loans are paid off, or if her friends have currently lent more money, the default risk of her loan is lower.
    - The variables *Ttlpureborrow* and *frd_default_histcum* have a positive impact on default risk, which is larger in the extreme low or high quantiles than in the middle quantiles. This implies that only modeling the conditional mean could underestimate their effects on default risk.
    - The variables *Ttlrole, frd_borrowfundlist_cur,* and *frd_payoff_histcum lendamt_frd_cur* have a negative impact on default risk, which is also larger in the extreme low or high quantiles than in the middle quantiles. If we use the traditional logistic regression to model the average risk of loans, we might overestimate their negative effects

Parameter Estimation and Heterogeneity

- New insight about the non-credit-related information are found:
  - Loan Performance Model— **Loss of Default Loan**
    - The effect of the number of friends depends on whether the friends have roles, and whether the members have many friends without roles. If many friends have no roles, the information regarding the number of friends is misleading, and thus the incurred loss is larger.
    - If the member has more friends who have defaulted loans or have listings requesting more than $10,000, or if the member's group currently has more listings, then the net loss of the default loan is larger.
    - The variables *Ttlleader, frd_lend1000Plus_hiscum, fundamt_frd_histcum* and *borrowamt_frd_cur* also have obvious *"Side-pattern"* effects. Specifically, the variable *Ttlleader* has significant negative effects on loans with a higher loss. However, other variables have a large impact only on loans with a smaller loss.

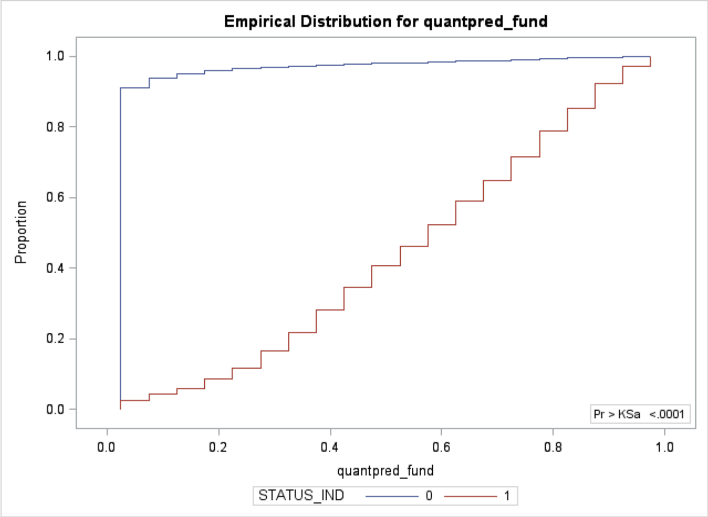
Model Performance

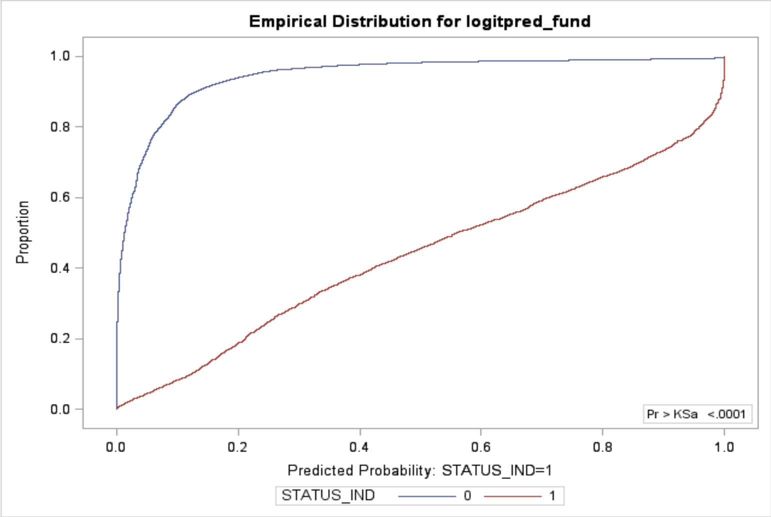
- Prediction Accuracy Comparison:
  - We compare the prediction accuracy of quantile regression model with traditional parameter models: Logistic and GLM model
  - Model is estimated in training dataset and prediction performance is validated in out of time sample
  - 20 representative quantiles are estimated for binary quantile model and 11 representative quantiles are estimated for standard quantile regression model
  - We use the popular nonparametric Kolmogorov-Smirnov statistic to compare the prediction performance of the two probability models. And we use the popular Mean Squared Error (MSE) to measure the prediction accuracy of other three models.

Model Performance

- Binary quantile model can outperform the logistic regression model:
  - Compared to the traditional logistic regression, which only focuses on “average” effects, binary quantile regression can capture the heterogeneous relationship between the covariates and the response variable. In particular, it can capture the important effects in those extreme quantiles of the response distribution

Kolmogorov-Smirnov	Probability of Funding		
Test	Binary Quantile	Logistic Regression	Improved Accuracy
K-S Statistic	0.265	0.235	12.77%





Model Performance

- Quantile model can outperform the traditional model:
  - the semi-parameter quantile regression method, which does not assume any distribution for the dependent variable, can improve the prediction accuracy when the distribution of the dependent variable is irregular and highly skewed.

	Loss			Profit		
	Quantile Regression	GLM Model	Improved Accuracy	Quantile Regression	GLM Model	Improved Accuracy
MSE	10,006,585	10,265,724	2.52%	883,667	897,610	1.55%

Investment Optimization

- The quantile regression model can be used for better investment decisions:
  - We formulate the Investment Decision Optimization Problem (IDOP) as a mixed integer linear programming (MILP) model.
  - Different constraints are incorporated, i.e., risk control, minimum investment amount
  - The MILP model:
    - $\text{Max} \quad \sum_{i=1}^{\mathrm{N}} \mathrm{x_i}\, \mathrm{P}_i/\mathrm{B}_i$
    - S.t. $\quad \sum_{i=1}^{\mathrm{N}} \mathrm{Y}_i \geq \mathrm{MinN};$
    - $\mathrm{x}_i \leq \mathrm{MaxAmt} \quad , \forall i,$
    - $\mathrm{x}_i \leq \mathrm{M} * \mathrm{Y}_i \quad , \forall i,$
    - $\mathrm{x}_i \geq \mathrm{Y}_i * \mathrm{MinAmt} \ , \forall i,$
    - $\sum_{i=1}^{N} \mathrm{x}_i\, \mathrm{R}_i \geq \overline{\mathrm{R}} \sum_{i=1}^{\mathrm{N}} \mathrm{x}_i$
    - $\sum_{i=1}^{N} \mathrm{x}_i \leq \mathrm{C}$

Investment Optimization

- The quantile regression model can be used for better investment decisions:
  - Three different approaches are compared:
    - Heuristic approach: invest listings with lowest risk and highest rate
    - Traditional model based approach: using traditional statistical models' prediction
    - Quantile model based approach: using quantile models' prediction
  - 612 active listings in 2007-05-15 are selected, three different approaches are used
  - Results:

	Heuristic	**Traditional Model-**	**Quantile Model-**	**Improvement**
	Based Approach	Based Approach	Based Approach	Quantile v.s. Traditional
Profit	($1,825.79)	$1,457.52	$1,679.26	15.21%

Future Research

- Two future research directions:
  - Quantile survival model for default risk:
    - High censor rate
    - Computation demanding
  - Dynamic Investment Optimization:
    - More factors need consider: dynamically changed market
    - Bid price optimization need be incorporated
    - Complex dynamic programming model need quick solution