STA302 Proposal - Project Group 83

11/29/2024

Bryan Chen, Jack Lewandowski, Pio Sleiman
Contributions: Introduction and Ethic Discussion - Jack
Methods, Conclusion, and Limitation - Bryan
Results - Pio

Introduction

Online peer-to-peer (P2P) lending platforms like Lending Club connect borrowers seeking loans directly with individual investors, bypassing traditional financial institutions. Borrowers must share detailed personal and financial information, such as loan purpose, debt-to-income ratio, and credit score, to help investors assess risk. This market is particularly relevant given that Americans held $16.9 trillion in debt in 2022 (Fay, 2023), with an average of $90,460 per individual across various obligations like credit cards, student loans, and mortgages (DeMatteo, 2023).

This study examines how geographic location influences the amount investors fund for loans, controlling for key variables like FICO (credit) score and debt-to-income ratio. Previous research identifies important predictors of loan performance; Serrano-Cinca et al. (2015) find that high interest rates and low loan grades are associated with defaults, while Emekter et al. (2014) highlight the importance of debt-to-income ratio and credit score. Additionally, Iyer et al. (2009) show that investors rely on "soft information" to assess credit worthiness, such as personal descriptions or images of the borrower, especially when assessing low-grade loans. These findings suggest that geographic location, an underexplored form of soft information, could also affect investor funding behavior.

We use a linear regression model to analyze this relationship, focusing on interpretability to assess the direct impact of location while controlling for variables previously documented in literature. Linear regression is appropriate as it allows us to isolate the effect of location and draw insights about its role in loan funding decisions within the P2P market.

Methods

To develop a linear regression model for predicting the funded amount in P2P lending, we implemented a rigorous approach using R. This process involved data preparation, model

building, assumption assessment, and diagnostic analysis, with each step grounded in statistical methodologies.

The initial phase focused on curating a subset of variables informed by relevant literature. We performed stratified sampling with respect to region on the Lending Club dataset to include predictors prevalent in the literature, and then cleaned the data by removing observations with missing values to ensure completeness and accuracy in subsequent analyses. We then subsetted the data for a pool of variables relevant in the literature for fitting our model. This filtering and cleaning ensured that our pool of predictors was relevant to both the research question and the literature base.

We employed forward automated model selection to systematically build a model starting with a simple intercept-only structure. This approach sequentially added predictors to identify an optimal model while minimizing overfitting. The forward selection process was based on Akaike Information Criterion (AIC) to balance model complexity and goodness-of-fit. Once the forward selection process identified a candidate model, we fit the linear regression model manually to ensure it aligned with expectations, specifically our assumptions. Key plots were used to preliminarily assess the validity of key assumptions for multiple linear regression, including the conditional mean predictor and mean response, linearity, independence, homoscedasticity, and normality. In terms of linearity and independence, residuals vs fitted plots confirmed an absence of significant patterns, indicating compliance. For homoscedasticity, residual vs predictor and residual vs fitted plots helped to reveal heteroscedasticity. To address any heteroscedasticity, we experimented with a natural logarithmic transformation of the response variable, funded amount, as a variance-stabilizing transformation. However, this transformation had limited effect on the fanning patterns. Thus, a Box-Cox transformation was applied using the PowerTransform function in R to identify optimal transformations for predictors such as dti, annual income, and total high credit limit.

Once we confirmed that our model was void of any assumption violations, comprehensive diagnostic evaluations were conducted. This included an ANOVA F-test of overall significance, assessing if at least one predictor in the model is statistically different from zero. To evaluate the significance of our research question's main predictor region, we performed partial F-tests to compare the full model with a reduced model excluding the predictor. Adjusted R^2 was also examined to quantify the proportion of variance in the response variable explained by the model.

Finally, multicollinearity was assessed using the Variance Inflation Factor, which we computed for all predictors.

In regards to our dataset, we identified potential influential observations by setting thresholds for leverage, Cook's distance, and standardized residuals. Problematic points were flagged and assessed for their impact on parameter estimates.

By integrating these methods systematically, our final model was robust and theoretically sound, providing a reliable framework to answer our research question and predict the funded amount in P2P lending.

Results

We initiated the modeling process with an automated variable selection procedure to identify the most relevant predictors for our analysis. This approach suggested a total of 10 predictors to be regressed on by our response variable, the funded amount of the loan. These 10 variables include the Debt-to-Income Ratio (DTI), region, employment length, average FICO score, home ownership status, annual income, number of publicly recorded bankruptcies, and the total high credit limit of the borrower, in addition to the grade (ranging from A-G), and application type (Individual or Joint application) of the loan. Importantly, at this point in our analysis, these variables are in their original form and have not been transformed in any way, aside from the data-cleaning process. Taking these 10 explanatory variables, we created our first model, referred to as Model 1, shown below:

$$\text{Model 1:} \quad \textit{Funded Amount}_i = \alpha_0 DTI_i + \alpha_1 Region_i + \alpha_2 Employment\ Length_i + \alpha_3 Average\ FICO_i + \alpha_4 Home\ \ Ownership_i + \alpha_5 Annual\ Income_i + \alpha_6 Publicly\ Recorded\ Bankruptcies_i + \alpha_7 Grade_i + \alpha_8 High\ Credit\ Limit_i + \alpha_9 Application\ Type_i$$

Using Model 1 as a starting point, we conducted statistical plot analyses to validate the core assumptions made with multiple linear regression models before interpreting our results. Through these methods, we found violations in both the Constant Variance and Linearity assumptions, implying that the model was not fit to describe the relationship between these variables. Figure 1 below shows a few of the problematic plots that led to the development of our model. Note that due to the amount of observations in our data, the points are very clustered.

Figure 1: Plots Showing Signs of Assumption Violations

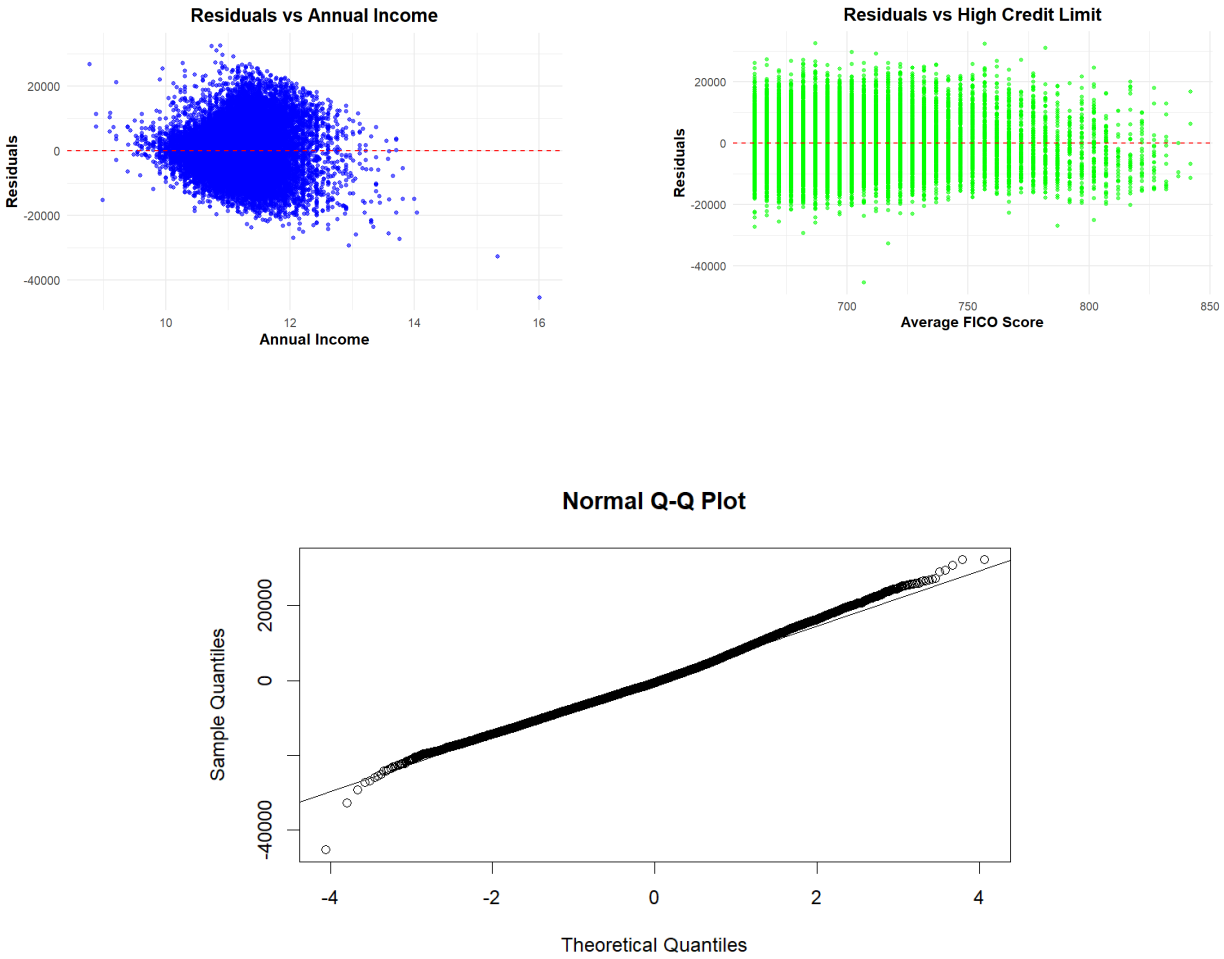
The figures above show the plots of Residuals vs Annual Income, Residuals vs High Credit Limit, and Fitted vs Residuals for Model 1. Note that the plots were manipulated for the readability of the reader

At this point, we decided to apply Box-Cox Transformations to the problematic numerical variables to rectify both of the violations in our model. We then applied three of the recommended transformations to our models, resulting in Model 2:

Model 2: $$Funded\ Amount_i = \beta_0(DTI_i)^{0.45} + \beta_1 Region_i + \beta_2 Employment\ Length_i + \beta_3 Average\ FICO_i + \beta_4 Home\ \ Ownership_i + \beta_5 ln(Annual\ Income_i) + \beta_6 Publicly\ Recorded\ Bankruptcies_i + \beta_7 Grade_i + \beta_8 (High\ Credit\ Limit_i)^{0.33} + \beta_9 Application\ Type_i$$

After once again applying various statistical methods to test the validity of the multiple linear regression assumptions, our Model 2 passed all the necessary thresholds to determine that it does not clearly violate any assumptions.

Figure 2: Plots Showing Model 2's Validity



The figures above show the plots of Residuals vs Annual Income, Residuals vs High Credit Limit, and Fitted vs Residuals for Model 1. These are only a few of the many plots analyzed before determining that Model 2 does not violate any assumptions.

After fitting Model 2 and determining its validity, we began analyzing its overall significance. This analysis started with the ANOVA test, shown towards the bottom of Table 1; we found that the F-Statistic for our model is very significant, passing the 1% significance level with ease.

Taking a look at both the R-squared and the Adjusted R-squared allowed us to determine that although at least one of our variables is statistically different from zero, the model overall does not explain a lot of the variation that is seen in the Funded Amount.

In regards to the overall fit of the model, we can see that the significant F-statistic and low R-squared point to the fact that the model cannot explain a large amount of the variety we see in the Funded Amount, but does have some predictive power.

At this point, we decided to analyze any problematic observations in our data that could influence the variation in the funded amount that our model fails to capture. Through the set thresholds, we identified that over 10,000 of our ~20,000, observations were deemed "problematic". Upon further analysis of the problematic observations, we found that the statistically problematic observations would not be deemed problematic following intuition gained from previous research, meaning that we could not reasonably justify removing them. With this in mind, we concluded that Model 2 was the most straightforward and well-validated model among our options, making it the basis for our final analysis.

Table 1: Model 2 Regression Outputs

	Dependent variable:
	Funded Amount
DTI	2,449.855***
	(88.848)
Average FICO Score	43.280***
	(1.943)
Home Ownership (Own)	-692.511***
	(186.830)
Home Ownership (Rent)	-802.955***
	(124.646)
Region (North East)	-49.518
	(170.881)
Region (South East)	174.350
	(166.754)
Region (South West)	-8.531
	(197.972)
Region (West)	61.225
	(172.632)
Annual Income	10,444.150***
	(135.443)
Constant	-138,583.000***
	(2,171.283)
Observations	20,196
R^2	0.315
Adjusted R^2	0.314
Residual Std. Error	7,674.608 (df = 20167)
F Statistic	331.325*** (df = 28; 20167)
Note:	*p<0.1; **p<0.05; ***p<0.01

	Dependent variable:
	Funded Amount
Grade B	1,910.117***
	(170.601)
Grade C	3,348.094***
	(178.812)
Grade D	4,608.423***
	(213.650)
Grade E	6,312.471***
	(269.183)
Grade F	7,192.207***
	(414.852)
Grade G	8,255.633***
	(659.253)
Emp Length (1 Year)	-275.265
	(273.107)
Emp Length (10+ Years)	-928.154***
	(206.604)
Emp Length (2 Years)	-386.595
	(252.904)
Constant	-138,583.000***
	(2,171.283)
Observations	20,196
R^2	0.315
Adjusted R^2	0.314
Residual Std. Error	7,674.608 (df = 20167)
F Statistic	331.325*** (df = 28; 20167)
Note:	*p<0.1; **p<0.05; ***p<0.01

	Dependent variable:
	Funded Amount
Pub Rec Bankruptcies	-390.116***
	(149.646)
Emp Length (3 Years)	-523.768**
	(260.167)
Emp Length (4 Years)	-642.089**
	(283.096)
Emp Length (5 Years)	-560.749**
	(276.788)
Emp Length (6 Years)	-577.347*
	(304.890)
Emp Length (7 Years)	-961.301***
	(314.115)
Emp Length (8 Years)	-188.764
	(316.686)
Emp Length (9 Years)	-456.999
	(329.364)
High Credit Limit	-118.713***
	(6.194)
Application Type (Joint App)	4,886.483***
	(250.539)
Constant	-138,583.000***
	(2,171.283)
Observations	20,196
R^2	0.315
Adjusted R^2	0.314
Residual Std. Error	7,674.608 (df = 20167)
F Statistic	331.325*** (df = 28; 20167)
Note:	*p<0.1; **p<0.05; ***p<0.01

Displayed above is the formatted regression output for our final model, Model 2. Note that the bottom of the graph on all three sides contains information regarding the overall fit of the model, such as the ANOVA/F-statistic, R-squared, and Adjusted R-squared. Furthermore, variable names are not displayed according to their transformations, when applicable, for visualization purposes.

Conclusion and Limitations

Our final model demonstrates that geographic location is not a significant predictor of the funded amount in peer-to-peer (P2P) lending. Partial F-tests comparing the full model to a reduced model excluding region yielded a high p-value, indicating that the contribution of region to the funded amount is statistically negligible. This finding addresses our research question directly, suggesting that funding decisions in P2P lending platforms are influenced more by borrower-specific financial metrics than by geographic location. This result aligns with the literature emphasizing the importance of creditworthiness and borrower financials over regional factors in P2P lending contexts.

One key coefficient of interest is the Box-Cox transformed debt-to-income ratio, which showed a statistically significant positive relationship with the funded amount. Specifically, for every one-unit increase in the square root of debt-to-income ratio, the funded amount is expected to increase by approximately $2449, holding all other predictors constant. This aligns with prior studies highlighting debt-to-income ratio – or more generally, borrower financials – as a key determinant in loan approval and funding amounts, reinforcing the validity of our model.

Despite the model's utility, several limitations must be acknowledged. First, while the residual vs. predictor plots indicated minimal violations of the constant variance assumption, slight fanning patterns persisted even after applying Box-Cox transformations to key predictors. This suggests underlying heteroscedasticity that could lead to biased standard errors and potentially affect hypothesis testing.

Second, our analysis of problematic observations revealed that approximately half of the sample data points exhibited high leverage, significant influence, or large residuals based on diagnostic measures such as Cook’s distance and standardized residuals. These problematic observations may reflect irregularities or outliers inherent in the dataset, potentially introducing noise or bias into our model.

Lastly, the residual vs. fitted plots also raised concerns about the adequacy of the model fit. While our transformations improved the diagnostic metrics slightly, lingering issues suggest that

unobserved variables or non-linear relationships may still exist, limiting the model's generalizability.

Future research should explore alternative models, such as generalized linear models (GLMs) or machine learning approaches, to capture potential non-linear effects and better handle influential observations. Additionally, expanding the dataset to include more diverse borrower attributes may help address these limitations and further refine our understanding of the funded amount determinants in P2P lending.

Ethics Discussion: Automated vs. Manual Selection Techniques

To develop our final model, our group opted to use a mix of manual and automated selection techniques. However, we prioritized manual selection for most of our analysis due to concerns about assumption violations and the importance of including key variables. Many variables required transformations to prevent assumption violations, which automated methods would not have appropriately addressed. Additionally, we suspected that our primary variable of interest, geographic location, might be excluded by automated selection due to its low statistical significance in the model. We used automated selection to help us identify the key variables for our final model, using the AIC results as a baseline. From this baseline model, we used manual selection techniques to reach our final model. The manual selection allowed us to check violations and ensure additional theoretically relevant variables were included to maintain model accuracy and relevance.

Reflecting on the virtues discussed in the ethics module, particularly trustworthiness and wisdom, we recognized that using automated methods alone would be reckless, leading to biased or misleading outcomes. Automated methods, while easy and efficient, skip essential steps that are required to create robust models that can provide answers to our research question. While the AIC model would have served as a decent predictor of funded amount, we had the wisdom to identify its shortcomings and fix them with manual selection techniques. We wanted to create a model with new patterns that had not been identified in past literature, but we prioritize statistical rigor to avoid creating false and misleading results.

Bibliography

Bachmann, A., Becker, A., Buerckner, D., Hilker, M., Kock, F., Lehmann, M., & Tiburtius, P. (2011). Online Peer-to-Peer Lending – A Literature Review. *Journal of Internet Banking and Commerce*, *16*(2), 2–18.

DeMatteo, M. (2023, November 14). *The average American has $90,460 in debt-here's how much debt Americans have at every age*. CNBC. https://www.cnbc.com/select/average-american-debt-by-age/

Emekter, R., Tu, Y., Jirasakuldech, B., & Lu, M. (2014). Evaluating credit risk and loan performance in online peer-to-peer (P2P) lending. *Applied Economics*, *47*(1), 54–70. https://doi.org/10.1080/00036846.2014.962222

Fay, B. (2023, December 4). *Debt in america: Statistics and demographics*. Debt.org. https://www.debt.org/faqs/americans-in-debt/demographics/

Iyer, R., Khwaja, A. I., Luttmer, E. F., & Shue, K. (2009). Screening in new credit markets: Can individual lenders infer borrower creditworthiness in peer-to-peer lending? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1570115

Nathan George (2018), *All Lending Club loan data*. Kaggle. https://www.kaggle.com/datasets/wordsforthewise/lending-club

Serrano-Cinca, C., Gutiérrez-Nieto, B., & López-Palacios, L. (2015). Determinants of default in P2P lending. *PLOS ONE*, *10*(10). https://doi.org/10.1371/journal.pone.0139427