An Investigation of Debt Delinquency in the United States

Stats 140XP Final Project: Report

Mason Kellerman, Sophia Yi, Diandian Shi, Aly Tan, Calvin Windmiller

December 12, 2025

1 Abstract

Debt delinquency, a term used to describe the failure to make a scheduled payment on a loan or credit account, is a metric often applied to evaluate the overall financial health of households and individuals nationwide. Debt delinquency can have drastic effects on the well being of an individual, leading to financial and personal consequences like credit score reduction, foreclosure, repossession, account closure, lawsuits, physical health detriments, and even higher crime rates. Because these outcomes intersect with broader structures of inequality, understanding the social determinants of debt delinquency is critical for identifying the underlying mechanisms that perpetuate the social and economic disparity in the United States.

In this study, we focus on how debt delinquency rates vary within socioeconomic and demographic characteristics, with particular attention to childhood family income, race and education levels. Notably, we consider the following research questions:

1. Do children from lower income families have different delinquency rates compared to higher income families?
2. How do childhood socioeconomic environments predict adult debt delinquency?

After data analysis of demographic and financial measures for cohorts in the United States, we observe clear disparities in adult delinquency rates across cohorts grouped by childhood parental income percentile, indicating that early-life economic environments are strongly linked to later financial outcomes. Moreover, the three components of socioeconomic context (child's race, parental income percentile, and college education rates) are all strong predictors on the future debt delinquency of a child, accounting for a significant portion of the variation in debt delinquency rates. The most important factor in prediction is the child's race, followed by parental income percentile and lastly the college education rates.

The analysis in this report is intended to provide insights that may inform policymakers, financial institutions and community organizations about the relationships between contextual factors and fiscal health outcomes. Ultimately, we hope that these findings will support the development of more equitable credit practices and interventions that enhance both the financial stability and overall well-being of all individuals.

2 Introduction

2.1 Background

Opportunity Insights is a non profit organization led by a group of economics professors from Harvard University, with the goal of identifying economic barriers in the United States and developing solutions to help those who need it. In 2020 (during the COVID-19 pandemic) they published their Opportunity Insights Economic Tracker[1], which is a large database built on private sector data that contains nationwide financial data at both the level of county and commuting zone. These figures include social data, social capital data, migration patterns, employment levels,

[1] "Opportunity Insights Economic Tracker", 2020.

consumer spending, etc. Relevantly, it has extensive data on credit access in the United States; this county-level credit access data will be the focus of our analysis.

Credit access, or the ability to take on loans/debt, is one of the most important abilities one can have. It opens up a significant amount of high-value economic activities, such as buying a car, buying a house, attending higher education, or starting a business. Generally though, credit access can be limited by factors such as a poor credit score or a high income to debt ratio. These factors are indicators of poor financial health that can come with their own devastating consequences, and they may lead a bank/lender to believe a debt will not be repaid.

This is why we are examining debt delinquency in particular, since it refers retrospectively to previously unpaid debts. This avoids the questions concerning credit score accuracy in favor of real outcomes, an important step since we address that relationship directly. The general perception understands unpaid loans as a consequence of personal decisions alone; however, there are of course more factors involved in the real world. To understand these contextual factors, we consider three components of one's socioeconomic environment: race, parental income, and prevailing level of education.

2.2 Literature Review

Many studies have investigated the socioeconomic and demographic factors, including race or ethnicity, neighborhood characteristics and parent income in explaining debt burdens and delinquency. Anders et al. (2023) conclude in their studies that the socioeconomic perspective of young people's family background could significantly influence their financial capabilities, mindsets or behaviors. Using survey data from 3,745 UK families, they found that young people from disadvantaged households have less parental inputs regarding financial education, and thus face an disadvantage in financial capabilities, including money confidence, money management, and financial connections. It implies potential intergenerational cycles of debt, poverty and inequality. Although the study does not follow participants' delinquency rate, it highlights the impact of people's socioeconomic familial background on financial behaviors. In our study, we will investigate whether the inter-generational impact exist across the United States.

In addition, the research of Hoeve et al. (2014) supports the association of debt stress, low financial control, and risk behaviors such as crime. Their systematic review found that nearly half of adolescents and young adults reported some debt and that financial problems were significantly correlated with criminal behavior. This supports the idea that economic stress and low perceived control over finances may contribute to delinquent outcomes. These results can be used within our research, in which individuals living in counties facing many social disparities can contribute to an individual's understanding (or lack of understanding) of financial literacy, and thus debt delinquency. Furthermore, this paper touches on the social risk factors that emerge from familial and inter-generational relationships, which can be used in our research on the effects of debt delinquency among families living in historically marginalized counties.

Furthermore, Blumenberg et al. (2024) examine automobile debt burdens and delinquency across neighborhoods in California, finding that structural racial and socioeconomic inequalities shape patterns of automobile debt. In particular, they find that Latino neighborhoods carry disproportionately high auto-loan debt burdens, while borrowers in Black neighborhoods have the highest delinquency rates. The authors emphasize that these disparities reflect systemic differences in economic opportunity, credit access, and neighborhood conditions. Building on their results, our study extends the geographic scope from California neighborhoods to a broader set of U.S. counties, allowing us to examine whether similar structural patterns emerge across different regions and more general types of debt.

From the health sciences discipline, Chan et al. (2014) investigate how different county-level characteristics are associated with mortality, then separating by cause-specific mortality like cancer or respiratory diseases. Stemming from the growing idea that social determinants of health, such as socioeconomic status or demographics, play a significant role in population health outcomes, this study adds onto previous research that has linked poverty, education, racial composition, and other community-level factors with health disparities. In particular, this study compares the relative strength of different types of community characteristics, such has sociodemographic vs. environmental, in predicting mortality. These findings present a compelling case that community-level sociodemographic and economic characteristics are among the most powerful predictors of county-level mortality in the U.S., lending support to the conclusion that public health interventions and policy changes addressing socioeconomic conditions could have sizable impacts on reducing mortality. Our study aims to use these relationships between community-based areas (such as counties) and its effects on debt delinquency and financial literacy among generations.

2.3 Dataset and Data Cleaning

Our analysis uses publicly available data from the Opportunity Insights[2] research group, which compiles large-scale administrative records to study economic mobility in the United States. Each observation in this set of datasets represents a **cohort** group defined by county, race, and parental income percentile, rather than individual-level records. This structure allows us to examine broad socioeconomic patterns while preserving privacy.

In particular, we utilized a collection of measures related to credit access. Each variable was provided in a separate file, which were combined using the shared demographic/cohort information.

The data is aggregated from children born between 1978 and 1985, when measured as adults in 2020. The adult financial outcomes come from 2020 credit bureau data, provided in aggregated form by cohort, including average credit score, average student loan, mortgage, auto loan, and credit card balances, the fraction of individuals with a 90+ day delinquency between 2016–2020, as well as delinquency adjusted for income.

While county names were provided, state was indicated only by a state FIPS code. As such, we used publicly available data from the U.S. Census[3] to add state abbreviations, which significantly aids in interpretability. County names for Puerto Rican municipalities were also added from U.S. Census data by joining on state and county FIPS code.

Table 1 gives a code book for the final set of variables used in this analysis. There are a total of six categorical variables defining demographic/cohort information, and seven relevant numerical variables containing credit access data.

For the second research question we required additional information about the socioeconomic context of the county during childhood. As such, we add an additional variable from the U.S. Census Bureau's Decennial Censuses, hosted by the USDA Economic Research Service[4], which reports the percentage of adults in each county with at least four years of college education. We use data from 1980, which is within the range for which individuals included in the Opportunity Insights data were born; as such, it helps to provide a picture of the local culture and environment growing up.

A short code book for this additional data is provided in Table 2.

2.4 Exploratory Data Analysis

We include selected summary plots of important variables in the analysis.

First, Figure 1 shows summary bar plots for the distribution of race and parental income percentile groups in the data.

[2] The data used is available from their portal under *Credit Access* https://opportunityinsights.org/data/.

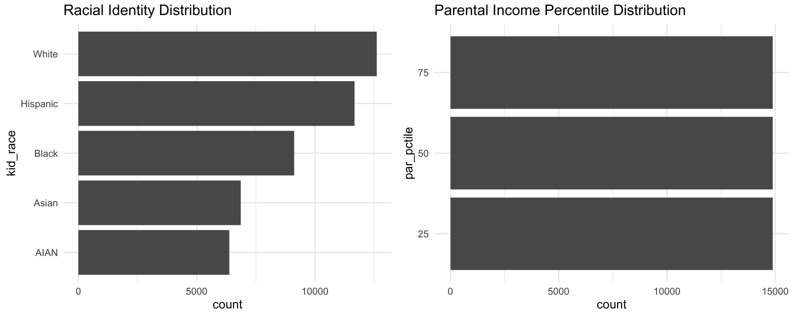
[3] https://www.census.gov/library/reference/code-lists/ansi.html

[4] https://www.ers.usda.gov/data-products/county-level-data-sets/county-level-data-sets-download-data

Table 1: Codebook for Opportunity Insights Credit Access Data

Variable	Variable Name	Data Type	Description
Parent's State (FIPS code)	`par_state`	categorical	The FIPS code for the Parent's State (or State-Like Entity)
Parent's State (Abbreviation)	`par_state_abbr`	categorical	The USPS Abbreviation for the Parent's State (or State-Like Entity)
Parent's County (FIPS code)	`par_county`	categorical	The FIPS code for the Parent's County
Parent's County Name	`par_county_name`	categorical	The Name of the Parent's County
Parent's Income Percentile	`par_pctile`	categorical	The National Income Percentile of the Parent. (`25`, `50`, `75`). Average is coded as (`-9`)
Child's Race	`kid_race`	categorical	The race of the child. (`AIAN` (American Indian and Alaska Native), `Asian`, `Black`, `Hispanic`, `White`)
Credit Score	`credit_score`	numeric	The average credit score in 2020, measured by Vantage 4.0
Student Loan Balance	`student_loan_balance`	numeric	The average balance of student loans held in 2020
Mortgage Balance	`mortgage_balance`	numeric	The average balance of mortgages held in 2020
Auto Loan Balance	`auto_loan_balance`	numeric	The average balance of auto loans held in 2020
Credit Card Balance	`credit_card_balance`	numeric	The average credit card balance in 2020
Debt Delinquency Rate	`debt_delinquency`	numeric	The rate of individuals with a 90+ day delinquency between 2016–2020, as a percentage
Debt Delinquency Controlling for Income	`debt_delinquency_ income_controlled`	numeric	The average residual from a regression of `debt_delinquency` on 2016 household income rank

Figure 1: Summary bar plots for race and parental income percentile



We see that we have the most observations for White and Hispanic cohorts, and the lease observations for AIAN (American Indian and Alaska Native) and Asian cohorts. Still, there is a substantial amount of observations for all races.

We note that there are exactly the same number of observations for each parental income percentile group. Recall that this data is at the level of cohorts. As such, this distribution reflects

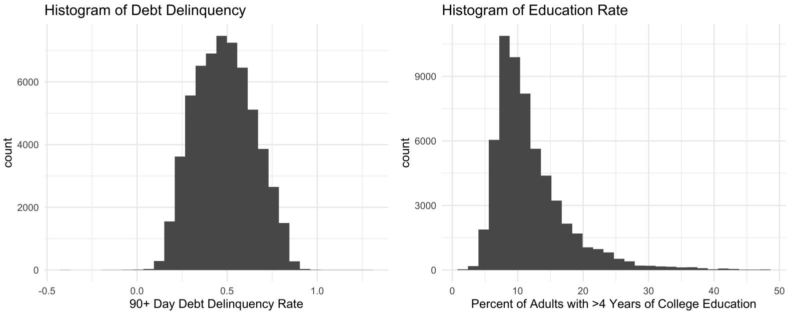
Table 2: Codebook for USDA Educational Attainment Data

Variable	Variable Name	Data Type	Description
Educational Attainment (4 years of College)	`college_4y`	numeric	The percentage of adults, by county, with at least four years of college education in 1980. (No data for Puerto Rico)

the fact that there are no county and race combinations that did not have an observation for all parental income percentile groups.

We also produce histograms for the distribution of the important numeric variables: debt delinquency rates, and the external data for four-year college education rates. These are shown in Figure 2.

Figure 2: Summary histograms for debt delinquency and education rates



We see that debt delinquency rates appear to be mostly normally distributed, while there is a substantial right-skew to the distribution of county-level college education rates.

3 Research Questions

We consider the following research questions, formulated as testable hypothesis:

1. *Do children from different family income percentile have different delinquency rates in adulthood?*
   - $\mathbf{H_0}$: Children from all levels of family income have *equal* delinquency rates in adulthood.
   - $\mathbf{H_1}$: Children from at least one parental income percentile group has *different* delinquency rates in adulthood compared to other groups.
2. *How do childhood socioeconomic environments predict adult debt delinquency?*
   - $\mathbf{H_0}$: Childhood socioeconomic environment *does not* predict variance in adult debt delinquency scores.
   - $\mathbf{H_1}$: Childhood socioeconomic environment *significantly* predicts variance in adult debt delinquency scores.

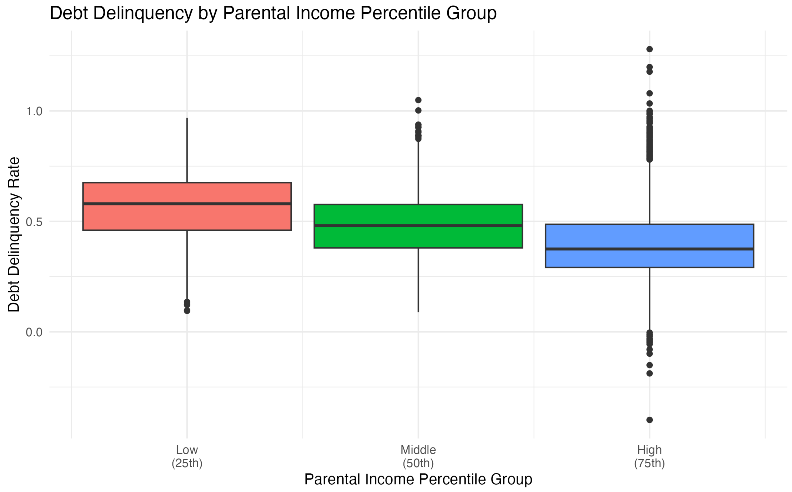
4 Methodology and Analysis

4.1 Research Question 1

We will test whether children from lower-income families have different adult delinquency rates than those from higher-income families. Our objective is to examine whether adult 90+ day delinquency rates differ across children from low (25th pctile), middle (50th pctile), and high (75th pctile) parental income groups. Delinquency is measured as the fraction of individuals with a 90+ day delinquency between 2016–2020 (debt_delinquency), and parental income percentile data comes from "par pctile" variable of the dataset. The goal is to determine whether childhood economic background is associated with different rates of adult financial distress.

4.1.1 Data Exploration

Figure 3: Debt delinquency rates across parental income groups



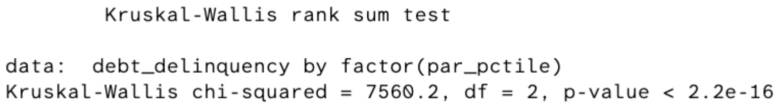
The boxplot in Figure 3 shows clear differences in adult debt delinquency rates across the three childhood income groups. The lowest income group (25th percentile) exhibits the highest delinquency rate, around 0.55 to 0.60. The middle-income group (50th percentile) has a lower median delinquency rate around 0.48 and 0.50 alongside a narrower distribution. The highest-income group (75th percentile) shows the lowest delinquency rates, with a median closer to 0.35 or 0.4, but with more extreme values. (just curious but why are there values above 1?)

Overall, the descriptive patterns suggest a negative relationship between parental income percentile and adult delinquency, demonstrating that people from wealthier childhood backgrounds tend to have lower adult delinquency rates.

4.1.2 Statistical Testing

To formally test if the differences are statistically significant, we conducted a Kruskal–Wallis test which is non-parametric, suggesting that it does not assume normality. We chose the Kruskal–Wallis test instead of ANOVA also because the delinquency variable is a rate bounded between 0 and 1 and displays notable skew and outliers.

Figure 4: Outcome of Kruskal–Wallis test



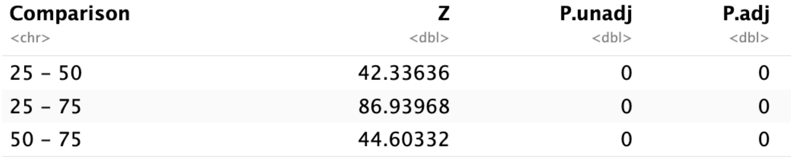
```
	Kruskal-Wallis rank sum test

data:  debt_delinquency by factor(par_pctile)
Kruskal-Wallis chi-squared = 7560.2, df = 2, p-value < 2.2e-16
```

As shown in Figure 4, the Kruskal-Wallis test revealed significant differences across income groups (p < 2.2e-16), indicating that the distribution of adult delinquency rates varies systematically by childhood income percentile.

Furthermore, we conducted a Post-hoc Dunn tests (Bonferroni-adjusted) to confirm how the delinquency rates differ by parental income groups. Specifically, as shown in Figure 5, the positive Z values on the first two rows suggest that children in the low parental income group (25th percentile) have higher adult debt delinquency rate than the other two groups with medium and high parental income percentiles, and the p-value of 0 suggests that the result is extremely significant. Additionally, the positive value of $Z = 44.60332$ when comparing 50 - 75th percentiles suggests that children with parents in the 50th percentile for income have lower adult debt delinquency than those in the 75th percentile parental income group. Again, the p-value of 0 indicates that this relationship is extremely significant.

Figure 5: Outcome of Post-hoc Dunn Test



Comparison	Z	P.unadj	P.adj
<chr>	<dbl>	<dbl>	<dbl>
25 - 50	42.33636	0	0
25 - 75	86.93968	0	0
50 - 75	44.60332	0	0

In general, the results suggest a strong negative relationship between childhood family income and adult financial distress measured by adult delinquency rate. Individuals raised in lower-income households are consistently at higher risk of debt delinquency in adulthood, which is consistent with the findings of prior research about early economic disadvantage contributing to long-term financial vulnerability.

4.2 Research Question 2

4.2.1 Additional Data Set

To incorporate a measure of childhood socioeconomic environment beyond parental income and race, we merged the Opportunity Insights dataset with an additional measure of the percentage of adults per county with at least four years of college education. This variable was sourced from data collected in the U.S. Census Bureau's Decennial Censuses hosted by the USDA Economic Research Service, as noted in Section 2.3.

We use education rates as a proxy for local human capital and economic opportunity. Furthermore, using data for the year 1980 best reflects the early childhood years of our sample, which are children born between 1978 and 1985. Therefore, this additional data set provides additional context for the socioeconomic environment during our selected populations' childhood and formative years

Also noteworthy is that we include race as a factor in childhood socioeconomic status (SES). Although race is a social or cultural identity and is not directly a component of SES which more focuses on education and income, it systematically affects the financial resources, economic opportunities, and upward mobility for different communities. Therefore, we believe that incorporating race can offer meaningful insights for investigating social inequality issues such as our second research question.

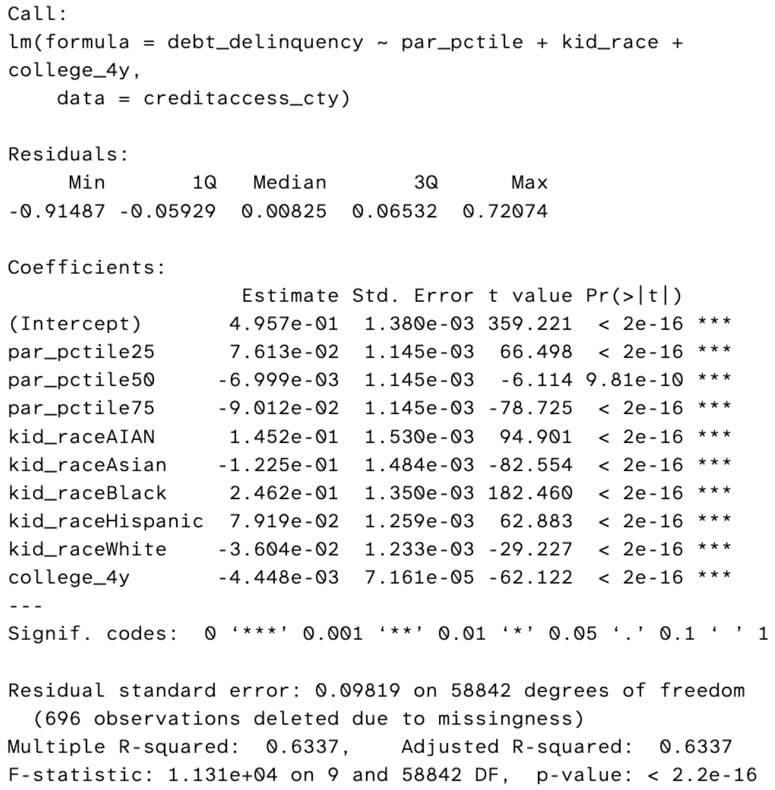
4.2.2 Model 1: Multivariate Linear Regression

We first fit a multivariate linear regression model, regressing debt delinquency (`debt_deliquency`) on parental income percentile group (`par_pctile`), race (`kid_race`), and rate of college education (`college_4y`), the formula for which is given in Equation 1.

$$\begin{aligned}\texttt{debt\_deliquency} =& \beta_0 + \beta_1(\texttt{par\_pctile}_{\text{25th}}) + \beta_2(\texttt{par\_pctile}_{\text{50th}}) + \beta_3(\texttt{par\_pctile}_{\text{75th}}) + \\ & \beta_4(\texttt{kid\_race}_{\text{AIAN}}) + \beta_5(\texttt{kid\_race}_{\text{Asian}}) + \beta_6(\texttt{kid\_race}_{\text{Black}}) + \\ & \beta_7(\texttt{kid\_race}_{\text{Hispanic}}) + \beta_8(\texttt{kid\_race}_{\text{White}}) + \beta_9(\texttt{college\_4y}) + \epsilon\end{aligned} \tag{1}$$

The results of this regression are displayed in Figure 6. Note that the default level for parental income percentile group and race is the overall/pooled average. We find that *all* of the model coefficients are statistically significant at $\alpha = 0.05$.

Figure 6: Multivariate linear regression model output for Research Question 2



```
Call:
lm(formula = debt_delinquency ~ par_pctile + kid_race +
college_4y,
    data = creditaccess_cty)

Residuals:
     Min       1Q   Median       3Q      Max
-0.91487 -0.05929  0.00825  0.06532  0.72074

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.957e-01  1.380e-03 359.221  < 2e-16 ***
par_pctile25     7.613e-02  1.145e-03  66.498  < 2e-16 ***
par_pctile50    -6.999e-03  1.145e-03  -6.114 9.81e-10 ***
par_pctile75    -9.012e-02  1.145e-03 -78.725  < 2e-16 ***
kid_raceAIAN     1.452e-01  1.530e-03  94.901  < 2e-16 ***
kid_raceAsian   -1.225e-01  1.484e-03 -82.554  < 2e-16 ***
kid_raceBlack    2.462e-01  1.350e-03 182.460  < 2e-16 ***
kid_raceHispanic  7.919e-02  1.259e-03  62.883  < 2e-16 ***
kid_raceWhite   -3.604e-02  1.233e-03 -29.227  < 2e-16 ***
college_4y      -4.448e-03  7.161e-05 -62.122  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09819 on 58842 degrees of freedom
  (696 observations deleted due to missingness)
Multiple R-squared:  0.6337,    Adjusted R-squared:  0.6337
F-statistic: 1.131e+04 on 9 and 58842 DF,  p-value: < 2.2e-16
```

The model has an adjusted R^2 value of 0.6337, meaning that the model explains approximately 63.37 percent of the variance in adult debt delinquency. The estimated intercept coefficient of $\hat{\beta}_0 = 0.4957$ indicates that the predicted debt delinquency rate across cohorts of all parental income percentile groups and racial identities, with 0% of adults having at least four years of college education, is about 49.57%.

For parental income percentile, the estimated coefficient for 25th percentile parental income is $\hat{\beta}_1 = 0.07613$. This positive coefficient value indicates that we expect cohorts with lower levels of parental income have debt delinquency rates about 7.7 percentage points higher in adulthood. On the other hand, the estimated coefficients for the 50th and 75th percentile parental income groups are $\hat{\beta}_2 = -0.006999$ and $\hat{\beta}_3 = -0.09012$, respectively. These negative coefficients indicate that we expect cohorts with moderate and higher levels of parental income have *lower* debt delinquency rates in adulthood, by about 0.70 percentage points for the 50th percentile, and 9.01 percentage points for the 75th percentile.

These results suggest that consequences for increased debt delinquency rates are mostly felt by those from lower-income families. Medium- and high-income families in this model do not face a similar penalty.

The estimated model coefficients for each racial category are shown in detail in Figure 6. Importantly, they have varying magnitudes and directions; for example, the estimated coefficient for Hispanic cohorts is $\hat{\beta}_7 = 0.07919$, while the estimated coefficient for White cohorts is $\hat{\beta}_8 = -0.03604$. This indicates that we expect debt delinquency rates for Hispanic cohorts to be about 6.92 percentage points *higher*, while expected debt delinquency rates for White cohorts are about 3.60 percentage points *lower*. This model, then, indicates that there are significant differences in debt delinquency rates along racial lines.

Lastly, the estimated model coefficient for college education rate is $\hat{\beta}_9 = -0.004448$. This means that across cohorts of all parental income percentile groups and racial identities, we expect a one percentage point increase in the percentage of adults with at least four years of college education to result in a 0.445 percentage point decrease in cohort debt delinquency rate. This constitutes a moderate relationship between education rates and debt delinquency; a one standard deviation increase in education rates (a change of 5.7 percentage points) would correspond with a lowering of debt delinquency rates by about 2.54 percentage points.

Overall, this multivariate linear regression model clearly shows that there are significant relationships between the socioeconomic environment of a person when growing up and their financial health in adulthood. In particular, higher levels of parental income and college education are associated with lower levels of debt delinquency, while undeserved racial identities face higher debt delinquency rates in adulthood.

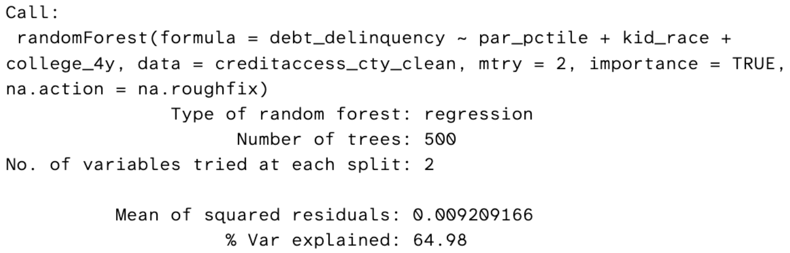
4.2.3 Model 2: Random Forest

We also built a random forest model, because it can be more suitable for capturing complex, nonlinear relationships in socioeconomic and demographic data across many variables. The method also works better with mixed variable types (both numerical and categorical). Also, random forest models provide measures of variable importance, allowing us to assess the relative contribution of each component of socioeconomic environment in predicting adult delinquency.

Since a random forest cannot handle missing data, we cleaned in two steps. First, we simply removed cohort observations for which the predicted variable of debt delinquency was missing. This removed 420 observations in total, constituting 0.705% of all observations. This is less than a single percentage point, so the removal is justified in this case. In order to handle missing values in the predictor variables, we use the `R` function `randomForest::na.roughfix()`, which imputes missing data using either the variable median or mode, depending on its data type. While more advanced methods of imputation exist, this strategy is both resource efficient and effective for limited levels of missing data.

When computing the random forest, we use the `R` function `randomForest` (from the package of the same name). The results from the model are given in Figure 7.

Figure 7: Random Forest result



```
Call:
 randomForest(formula = debt_delinquency ~ par_pctile + kid_race +
college_4y, data = creditaccess_cty_clean, mtry = 2, importance = TRUE,
na.action = na.roughfix)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 2

          Mean of squared residuals: 0.009209166
                    % Var explained: 64.98
```

The random forest model explains about 64.98% of the variance in debt delinquency, which means these three childhood and community factors together explain over half the variation in adult debt delinquency. This suggests that early-life socioeconomic context plays a major role in shaping long-term financial outcomes. The model's mean squared error ($MSE = 0.00921$) was low relative to the overall variance in delinquency rates, demonstrating strong predictive accuracy even with only three predictors.

Figure 8: Random Forest Variable Importance Plot

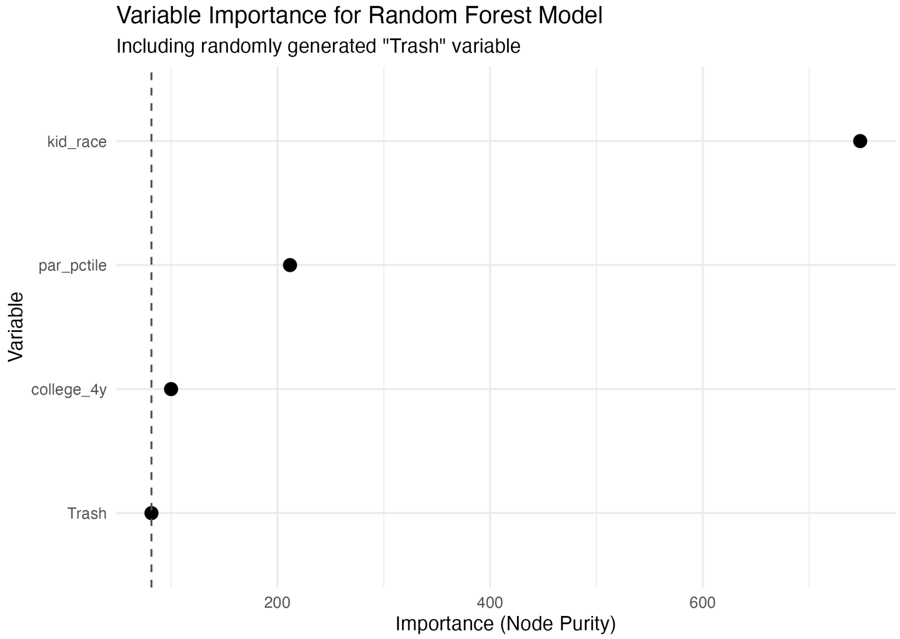


Figure 8 shows the importance plots for the random forest, after including a truly random "Trash" variable for comparison. While no definite conclusions can be made, a variable with similar importance to random noise is likely not important in the model.

We see that race appears to be by far the most important variable in predicting debt delinquency rates, likely reflecting systemic racial differences in the financial space. Parental income percentile group also appears to be important, representing the differential financial health outcomes for children of households with varying wealth. College education rates are the least important of the three predictors, perhaps signaling a reduced role in determining future debt delinquency rates.

Overall, the random forest model mostly supports the findings from the linear regression model in Section 4.2.2 in that all three axis of socioeconomic environment appear to play a role in determining future debt delinquency rates. Additionally, the specific demographic information of race and parental income play an increased role in this relationship than the county-level college education rates. One could imagine that cohorts aggregated on parental education would be more important, as parents play an outsized role on child development and opportunities.

5 Conclusions

Based these analyses, it appears that childhood socioeconomic factors do indeed have a significant effect on future debt delinquency rates.

In Research Question 1, through a Kruskal-Wallis test and a Post-hoc Dunn test, we found that children from all three parental income groups (25th, 50th, and 75th percentile) have significantly different debt delinquency rates, where a consistent and significant negative relationship between parental income percentile group and future debt delinquency exists.

In Research Question 2, a multivariate linear regression model was fit using race, parental income percentile group, and rate of four-year college education. We found that all three of these factors (and across all levels) were significant predictors of a child's future debt delinquency, explaining 63.37 percent of all the variance in adult debt delinquency. From the estimated regression coefficients, we make the following conclusions about their individual relationships:

- 25th income percentile = ~7.7 point increase from intercept in debt delinquency
- 50th income percentile = ~0.7 point decrease from intercept in debt delinquency

- 75th income percentile = ∼9.0 point decrease from intercept in debt delinquency
- AIAN = ∼14.5 point increase from intercept in debt delinquency
- Asian = ∼12.3 point decrease from intercept in debt delinquency
- Black = ∼24.6 point increase from intercept in debt delinquency
- Hispanic = ∼7.9 point increase from intercept in debt delinquency
- White = ∼3.6 point decrease from intercept in debt delinquency
- Every one point increase in college education rate = ˜0.445 point decrease from intercept

We also fit a more complex random forest model, predicting debt delinquency rate using the same factors. The random forest explained a very similar amount of the variation in debt delinquency (64.98 percent) of the variance in debt delinquency. By considering variable importance, we find that the child's race appears to be by far the most significant predictor of debt delinquency among the predictors, with parental income percentile second and college education rate third.

Overall, we find that the three socioeconomic factors: child's race, parental income percentile, and college education rates have a significant effect on the future debt delinquency of a child, accounting for a considerable 64.98 percent of the variation in adult debt delinquency rates. We find that the most significant factor is a child's race, with AIAN, Black, and Hispanic children having higher future debt delinquency rates and White and Asian children having lower rates. Parental income percentile is the second most significant factor, with a higher income percentile resulting in lower debt delinquency rates. Lastly, local college education rates (a proxy for local human capital and economic opportunity) is also found to be significant, with higher college education rates leading to lower debt delinquency rates. We can conclusively state that there appears to be a consistent and significant relationship between high debt delinquency rates and the the childhood socioeconomic factors in which people were raised.

6 Potential Impact

In terms of potential policy implications, this report finds evidence which suggests that mapping delinquency patterns by income, county, or other demographic factors could allow organizations and governments to allocate support and resources resources more efficiently. This avoids a "one-size-fits-all" approach, and ensures that support reaches those who are statistically most at risk. If delinquency is strongly associated with an individual's upbringing or parental income, then credit repair or lending policies may need to account for a structural disadvantage. This can be reformed through policies enacted by state or federal government to create a more equitable credit system.

Perhaps more important is the critical realization that the external circumstances of a person's youth can have a significant and substantial effect on their eventual financial health. Based on these findings, we suggest that policy makers and credit-lending institutions consider a person's financial status not merely as a direct consequence of personal actions, but as a complex product of myriad social and cultural factors outside of any individual's control. Financial education programs are sometimes seen as a way to help those struggling to make ends meet. However, our findings suggest that these individually-oriented programs should be augmented with larger-scale social policy directives to increase the opportunities available for all people, especially those groups revealed here to be at a systemic disadvantage in the credit market.

Moving forward, it is this larger social context which has the most untapped potential. By developing more complex, larger-scale models that include more measures of this socioeconomic context, future research can build towards a more detailed picture of how these disadvantages are distributed throughout the population and are felt by individuals.

7 References

Anders, J., Jerrim, J., & Macmillan, L. (2023). Socio-Economic Inequality in Young People's Financial Capabilities. *British Journal of Educational Studies*, *71*(6), 609–635. https://doi.org/10.1080/00071005.2023.2195478

Blumenberg, E., Speroni, S., Siddiq, F., & Wasserman, J. L. (2024). Putting Automobile Debt on the Map: Race and the Geography of Automobile Debt in California. *Transportation Research Part A: Policy and Practice*, *190*, 104230. https://doi.org/10.1016/j.tra.2024.104230

Chan, K. S., Roberts, E., McCleary, R., Buttorff, C., & Gaskin, D. J. (2014). Community Characteristics and Mortality: The Relative Strength of Association of Different Community Characteristics. *American Journal of Public Health*, *104*(9), 1751–1758. https://doi.org/10.2105/AJPH.2014.301944

Hoeve, M., Stams, G. J. J. M., Van Der Zouwen, M., Vergeer, M., Jurrius, K., & Asscher, J. J. (2014). A Systematic Review of Financial Debt in Adolescents and Young Adults: Prevalence, Correlates and Associations with Crime (Z. Wang, Ed.). *PLoS ONE*, *9*(8), e104909. https://doi.org/10.1371/journal.pone.0104909

Opportunity Insights Economic Tracker. (2020). Retrieved December 12, 2025, from https://opportunityinsights.org/tracker-resources/

8 Appendix

8.1 Reproducible R Code

All R code used to generate all figures and analyses used in the report is included here. In addition, an R markdown file ("project_code.Rmd") is submitted as well for easy replication. The tidyverse family of packages is used throughout, with specific packages loaded where needed.

8.1.1 Data Acquisition and Cleaning

We first download the .csv files from Opportunity Insights Recall that each credit access measure is delivered in a separate file, sharing cohort information. As such, we join on that demographic information to create a single data frame.

```r
### file urls by county
file_urls <- c(
  "https://opportunityinsights.org/wp-content/uploads/2025/07/avg_credit_score_2020_cty.csv",
  "https://opportunityinsights.org/wp-content/uploads/2025/07/avg_student_loan_balance_2020_cty.csv",
  "https://opportunityinsights.org/wp-content/uploads/2025/07/avg_mortgage_balance_2020_cty.csv",
  "https://opportunityinsights.org/wp-content/uploads/2025/07/avg_auto_loan_balance_2020_cty.csv",
  "https://opportunityinsights.org/wp-content/uploads/2025/07/avg_credit_card_balance_2020_cty.csv",
  "https://opportunityinsights.org/wp-content/uploads/2025/07/avg_delinq_rate_2020_cty.csv",
  "https://opportunityinsights.org/wp-content/uploads/2025/07/avg_delinq_rate_2020_cont_income_cty.csv"
)

### join dataframes by demographic info (cols 1, 2, 3, 4, 6)
creditaccess_cty <- apply(
  # load all data for the different variables
  as.matrix(file_urls), 1, function(file) read.csv(file)
) |>
  # combine by merging on demographic info
  reduce(full_join, by = names(read.csv(file_urls[1]))[c(1:4, 6)]) |>
  # reorder...
  select(1, 2, 6, 4, 3, everything()) |>
  # rename variables
  rename(
    par_county_name = county_name,
    credit_score = shrunk_xkid_vscore2020,
    student_loan_balance = shrunk_xkid_stubalance2020,
    mortgage_balance = shrunk_xkid_mtabalance2020,
    auto_loan_balance = shrunk_xkid_auabalance2020,
    credit_card_balance = shrunk_xkid_brcbalance2020,
    debt_delinquency = shrunk_xkid_delinq90_02020,
    debt_delinquency_income_controlled = shrunk_income_resid_2020
  )
```

Since the only indicator for par_state is the FIPS code value, use data from the U.S. Census to replace with state abbreviations (from USPS code). Also, use the Census' list of FIPS county-level area names to get Puerto Rican municipality names, which were missing.

```r
## replace state FIPS indicator with state abbreviations, also from census bureau
# get 2010 State FIPS codes
state_FIPS_table <- read.table(
  "https://www2.census.gov/geo/docs/reference/state.txt",
  sep = "|", header = TRUE,
  col.names = c("par_state", "state_abbr", "state_name", "state_GNISID")
)
# update state FIPS code ‘par_state‘ in data; reorder
creditaccess_cty <- creditaccess_cty |>
  left_join(state_FIPS_table[, 1:2], by = "par_state") |>
  rename(par_state_abbr = state_abbr) |>
  select(1, 13, everything())

## update Puerto Rico to include municipality names
# get data from U.S. Census Bureau
```

```r
PR_cty_key <- read.csv(
  "https://www2.census.gov/geo/docs/reference/codes/files/national_county.txt",
  header = FALSE,
  col.names = c("state_abbr", "par_state", "par_county", "county_name", "FIPS_
    classcode")
)
PR_cty_key <- PR_cty_key[PR_cty_key$state_abbr == "PR", 2:4]
# add PR municipality names
creditaccess_cty <- creditaccess_cty |>
  left_join(PR_cty_key, by = c("par_state", "par_county")) |>
  mutate(
    par_county_name = ifelse(
      par_state_abbr == "PR",
      county_name, par_county_name
    )
  ) |>
  select(-14)
```

Add data about education rate (4 years of college or more, 1980) hosted by USDA Economic Research Service:

```r
# read data from https://www.ers.usda.gov/data-products/county-level-data-sets/
#                county-level-data-sets-download-data
edurates <- read.csv("https://ers.usda.gov/sites/default/files/_laserfiche/
    DataFiles/48747/Education2023.csv?v=58365")
# get state FIPS code
edurates <- edurates |>
  mutate(
    par_state = as.numeric(str_sub(FIPS.Code, 1, -4)),
    par_county = as.numeric(str_sub(FIPS.Code, -3)),
    year = str_sub(Attribute, -4),
    # get county name, too
    par_county_name = str_sub(Area.name, 1, -8)
  ) |>
  # isolate 1980 values
  filter(
    year == "1980",
    str_sub(Attribute, 30, -7) == "four years of college or higher",
  ) |>
  # select relevant columns; exclude US level data
  select(5, 6, 7)
edurates <- edurates[-1, ]

# combine data by state and county name
creditaccess_cty <- creditaccess_cty |>
  left_join(
    edurates,
    by = c("par_state", "par_county")
  ) |>
  rename(college_4y = Value)
```

Finally, code to save out the updated data set as a .csv file for sharing and use.

```r
### save as csv (don't forget to use setwd())
creditaccess_cty |>
  write.csv("creditaccess_cty.csv", row.names = FALSE)
```

8.1.2 Figure Generation

Bar chart of frequencies for kid_race, and for par_pctile. Use the gridExtra package in R to combine plots to export:

```r
# race bar chart
# excluding pooled average
raceplot <- creditaccess_cty[creditaccess_cty$kid_race != "Pooled", ] |>
  ggplot(aes(y = kid_race)) +
  geom_bar() +
  ggtitle("Racial Identity Distribution") +
  labs(ylab = "Race") +
  theme_minimal()
raceplot
# par_percentile bar chart
# excluding overall average
```

```r
pctileplot <- creditaccess_cty[creditaccess_cty$par_pctile != -9, ] |>
  ggplot(aes(y = par_pctile)) +
  geom_bar() +
  ggtitle("Parental Income Percentile Distribution") +
  labs(ylab = "Parental Income Percentile Group") +
  theme_minimal()
pctileplot
# combine plots & save out
library(gridExtra)
png(
  "figures/EDAplots_1.png",
  width = 10, height = 4, units = "in", res = 500
)
grid.arrange(raceplot, pctileplot, ncol = 2)
dev.off()
```

Histograms for `debt_delinquency`, and for college_4y:

```r
# debt delinquency histogram
debtdelplot <- creditaccess_cty |>
  ggplot(aes(debt_delinquency)) +
  geom_histogram() +
  ggtitle("Histogram of Debt Delinquency") +
  xlab("90+ Day Debt Delinquency Rate") +
  theme_minimal()
debtdelplot
# education rate histogram
educplot <- creditaccess_cty |>
  ggplot(aes(college_4y)) +
  geom_histogram() +
  ggtitle("Histogram of Education Rate") +
  xlab("Percent of Adults with >4 Years of College Education") +
  theme_minimal()
educplot
# combine plots & save out
png(
  "figures/EDAplots_2.png",
  width = 10, height = 4, units = "in", res = 500
)
grid.arrange(debtdelplot, educplot, ncol = 2)
dev.off()
```

Bar chart of debt delinquency rates (`debt_delinquency`) grouped by parental income percentile (`par_pctile`)

```r
# "debtdel_by_parpctile.png"
# only 25th, 50th, 75th percentiles
debtdel_by_parpctile <- creditaccess_cty[creditaccess_cty$par_pctile != -9, ] |>
  ggplot(aes(y = debt_delinquency, group = factor(par_pctile), fill = factor(par_
    pctile))) +
  geom_boxplot() +
  ggtitle("Debt Delinquency by Parental Income Percentile Group") +
  xlab("Parental Income Percentile Group") + ylab("Debt Delinquency Rate") +
  scale_x_continuous(
    breaks = c(-0.25, 0, 0.25),
    labels = c("Low\n(25th)", "Middle\n(50th)", "High\n(75th)")
  ) +
  theme_minimal() +
  theme(
    legend.position = "none"
  )
debtdel_by_parpctile
# save plot out
ggsave(
  "figures/debtdel_by_parpctile.png",
  debtdel_by_parpctile,
  width = 8, height = 5, units = "in"
)
```

8.1.3 Statistical Analyses

RQ 1: Difference in debt delinquency by childhood parental income

Kruskal-Wallis Test if significantly different debt delinquency rates by parental income percentile group:

```r
# again, only focus on 25th, 50th, and 75th percentile (ignore average group data)
kruskal.test(
  debt_delinquency ~ factor(par_pctile),
  data = creditaccess_cty[creditaccess_cty$par_pctile != -9, ]
)
```

And, post-hoc Dunn's Test with Bonferroni corrections. Use the `FSA` package.

```r
library(FSA)
dunnTest(
  debt_delinquency ~ factor(par_pctile),
  data = creditaccess_cty[creditaccess_cty$par_pctile != -9, ],
  method = "bonferroni"
)
```

RQ 2: Using socioeconomic context as kid to predict debt delinquency as adult:

Model 1: Multivariate linear regression. We fit a multivariate linear regression model, regressing debt delinquency (`debt_deliquency`) on parental income percentile group (`par_pctile`), race (`kid_race`), and rate of college education (`college_4y`).

Note: no education level data for Puerto Rico, so this analysis does not include the territory. Since the data is missing, this is handled automatically by `lm()`:

```r
# treat categorical variables as factor; adjust level order to control default levels
creditaccess_cty$par_pctile <- factor(creditaccess_cty$par_pctile)
creditaccess_cty$kid_race <- factor(
  creditaccess_cty$kid_race,
  levels = c("Pooled", "AIAN", "Asian", "Black", "Hispanic", "White")
)
# fit linear model
rq2_1 <- lm(
  debt_delinquency ~ par_pctile + kid_race + college_4y,
  data = creditaccess_cty
)
summary(rq2_1)
```

Find standard deviation of education rate for aid in interpretation:

```r
# sd of education rate
sd(creditaccess_cty$college_4y, na.rm = TRUE)
# percentage point change in debt delinquency of
# average cohort for 1 sd increase in education rate
rq2_1$coefficients[10] * sd(creditaccess_cty$college_4y, na.rm = TRUE) * 100
```

Model 2: Random Forest. Before we can fit the random forest, we need to clean the data; in particular, we need to remove missing values in `debt_delinquency` outcome variable. Simply remove all cohort observations for which this value is missing, taking note of how many were removed in total:

```r
# number of missing ‘debt_deliquency‘ values:
sum(is.na(creditaccess_cty$debt_delinquency))
# as a percentage of all observations
sum(is.na(creditaccess_cty$debt_delinquency)) / nrow(creditaccess_cty) * 100
creditaccess_cty_clean <- creditaccess_cty[!is.na(creditaccess_cty$debt_delinquency), ]
```

Next, fit the random forest model using `randomForest::randomForest()`, predicting debt delinquency from parental income percentile group, race, and education rate. Note that this requires the `randomForest` package. Use the parameter `na.action = na.roughfix` to impute missing values in the predictors using median or mode, depending on type:

```r
library(randomForest)
set.seed(140)
# fit random forest model
rq2_2 <- randomForest(
  debt_delinquency ~ par_pctile + kid_race + college_4y,
  data = creditaccess_cty_clean,
  mtry = 2,
  importance = TRUE,
  na.action = na.roughfix # impute missing values in predictors with median or mode
)
rq2_2
```

Compare the RMSE with the standard deviation of `debt_delinquency`:

```r
# compare RSME and sd(debt_delinquency)
sqrt(rq2_2$mse[length(rq2_2$mse)])
sd(creditaccess_cty_clean$debt_delinquency)
```

Generate the variable importance plot from the random forest model object. We add a truly random "Trash" variable for comparison:

```r
# add "Trash" variable; rerun random forest:
creditaccess_cty_clean_trash <- data.frame(
  creditaccess_cty_clean,
  "Trash" = rnorm(nrow(creditaccess_cty_clean))
)
rq2_2_trash <- randomForest(
  debt_delinquency ~ par_pctile + kid_race + college_4y + Trash,
  data = creditaccess_cty_clean_trash,
  mtry = 2,
  importance = TRUE,
  na.action = na.roughfix # impute missing values in predictors with median or mode
)
# Plot variable importance
ggplot(
  data.frame(
    variable = rownames(importance(rq2_2_trash)),
    importance = importance(rq2_2_trash)[, "IncNodePurity"]
  ),
  aes(reorder(variable, importance), y = importance)
) +
  geom_point(size = 3) +
  geom_hline(
    yintercept = importance(rq2_2_trash)[4, "IncNodePurity"],
    linetype = "dashed", color = "grey30"
  ) +
  coord_flip() +
  ggtitle("Variable Importance for Random Forest Model") +
  labs(subtitle = 'Including randomly generated "Trash" variable') +
  ylab("Importance (Node Purity)") +
  xlab("Variable") +
  theme_minimal()
# save plot
ggsave(
  "figures/rq2_rfimportanceplot.png",
  width = 7, height = 5, units = "in"
)
```