

# Predicting Credit Card Balances Using Regression Analysis

Vedavyas Reddy Bommineni, Urvesh Bhagat,  
Rohith Reddy Vangal, Bhargav Pathuri

(Group 2)

12/09/2025



# Introduction

## Background and motivation for the project

- In the banking industry, assessing and managing credit risk is a critical function that directly influences profitability and financial stability. One key aspect of this process involves understanding and predicting customers' credit card balances. Credit card balance levels provide valuable insight into borrowing behavior, repayment capacity, and potential risk of default.
- Banks issue credit cards based on individuals' financial characteristics such as income, spending habits, and existing obligations. Once issued, these balances fluctuate based on personal and economic factors. Accurately modeling and predicting such balances can help financial institutions make informed lending decisions, optimize credit limits, and identify customers who may be at risk of financial distress.
- From a data analysis perspective, this problem provides an excellent opportunity to apply regression modeling techniques to explore relationships between financial and demographic variables and credit card balance behavior. This analysis not only supports risk management practices but also deepens understanding of the factors driving consumer credit usage.

# Objective Of The Analysis

- The primary objective of this project is to develop regression models that predict individuals' average credit card balances using demographic and financial variables. The target variable in this study is **Balance**, representing the average amount owed on a credit card. Predictor variables include income, credit limit, age, education, number of cards, marital status, and other personal characteristics available in the dataset.
- This analysis will involve exploring the relationships among these variables through data visualization and correlation analysis, followed by fitting multiple regression models. The ultimate goal is to identify significant predictors of credit card balances and assess how well the developed models can explain the variation in credit balance across individuals.

# Contribution Of Each Team Member

| Project Component                                   | Team Member       | Contribution Summary   |
|---|-------------------|--|
| Data Cleaning & Preprocessing                       | Vedavyas          | Imported dataset, organized variables, checked for missing values, created log-transformed predictors ( <i>log_Income</i> , <i>log_Limit</i> ).  |
| Exploratory Data Analysis                           | Rohith            | Produced scatterplot matrix, boxplots, histograms, grouped scatterplots; computed correlation matrix; summarized extreme values and initial assumption violations.   |
| Data Transformation                                 | Urvesh            | Applied Box–Cox, inverse fitted-value, log and sqrt transformations; compared transformation performance; concluded no transformation of <i>Balance</i> was effective; documented reasoning.   |
| Model Diagnostics & Remedies                        | Bhargav           | Conducted residual diagnostics, BP/White tests, QQ plots, VIF, leverage and Cook's distance analysis. <b>Identified and removed outlier/influential observations</b> and re-fitted models. Summarized improvements and remaining issues.   |
| Model Selection (Main Effects + Interaction Models) | All Members       | Each member handled one part:<br><ul style="list-style-type: none"> <li>• Vedavyas — Exhaustive search (AIC/BIC)</li> <li>• Rohith — Cross-validation (RMSE)</li> <li>• Urvesh — Interaction model comparison + backward elimination (BIC)</li> <li>• Bhargav — Final diagnostics &amp; multicollinearity checks ,<b>Collectively selected final model</b>.</li> </ul> |
| Model Interpretation                                | Vedavyas & Urvesh | Interpreted coefficients, ANOVA, effect plots for both main-effects and interaction models; explained income, limit, and student effects; wrote interpretation slides.   |
| Outlier & Influential Analysis                      | Rohith & Bhargav  | Confirmed influential points using Cook's distance; evaluated prediction intervals; compared model fit before/after removing outliers; verified robustness.  |
| Slide Preparation                                   | All Members       | Each member created slides for their assigned sections; team revised formatting, narrative flow, and clarity.  |
| Presentation  | All Members       | Each member presents their analytical portion  |

# Data Section

**Data source:** The dataset used in this project is obtained from **Kaggle**, an online platform for data science and machine learning projects. It is part of the “ISLR” dataset collection accompanying the textbook *An Introduction to Statistical Learning* by James, Witten, Hastie, and Tibshirani. The specific dataset, titled “Credit.csv”.

**Link for the dataset :** <https://www.kaggle.com/datasets/ishaanv/ISLR-Auto?select=Credit.csv>

The data provide information on individuals’ demographic and financial characteristics, commonly used to study credit risk and predict credit card balances.

## **Data description:-**

- ID: a unique identification number for each individual.
- Income: the individual’s income, scaled in units of \$10,000.
- Limit: the maximum amount of credit available to the individual.
- Rating: a score representing the individual’s creditworthiness.
- Cards: the number of card ownership.
- Age: the individual’s age, measured in years.
- Education: the number of years the individual has spent in education.
- Gender: specifies the gender of the individual, either Male or Female.
- Student: indicates whether the individual is a student, with possible values being Yes or No.
- Married: shows if the individual is married or not, options being Yes or No.
- Ethnicity: the individual’s ethnic background, which can be African American, Asian, or Caucasian.
- Balance: the average balance maintained on the individual’s credit card, expressed in dollars.

- Response Variable:** The response variable for this analysis is **Balance**, which represents the average monthly balance maintained by an individual on their credit card. This variable is expressed in dollars and serves as the dependent variable in the regression model. The objective of the analysis is to model and predict this balance based on other demographic and financial factors.
- Covariates:** The covariates include a mix of numerical and categorical variables describing the personal and financial background of each individual. Numerical predictors such as **Income**, **Limit**, **Rating**, **Age**, **Education**, and **Cards** capture continuous financial or demographic characteristics. Categorical predictors such as **Gender**, **Student**, **Married**, and **Ethnicity** provide qualitative information that may influence credit behavior. These covariates together help explain variation in the credit card balance and are expected to have both linear and nonlinear relationships with the response variable.

### Raw data snapshot:

Predictors

| ID  | Income  | Limit | Rating | Cards | Age | Education | Gender | Student | Married | Ethnicity | Balance | Response |
|-----|---------|-------|--------|-------|-----|-----------|--------|---------|---------|-----------|---------|----------|
| 0 1 | 14.891  | 3606  | 283    | 2     | 34  | 11        | Male   | No      | Yes     | Caucasian | 333     |          |
| 1 2 | 106.025 | 6645  | 483    | 3     | 82  | 15        | Female | Yes     | Yes     | Asian     | 903     |          |
| 2 3 | 104.593 | 7075  | 514    | 4     | 71  | 11        | Male   | No      | No      | Asian     | 580     |          |
| 3 4 | 148.924 | 9504  | 681    | 3     | 36  | 11        | Female | No      | No      | Asian     | 964     |          |
| 4 5 | 55.882  | 4897  | 357    | 2     | 68  | 16        | Male   | No      | Yes     | Caucasian | 331     |          |

# Method Section

- For this study, we selected **Multiple Linear Regression (MLR)** as the primary analytical method. The goal of our project is to identify and quantify how various demographic and financial factors such as Income, Limit, Rating, and Student status influence the average credit card balance of individuals. Since the response variable (Balance) is continuous, and the dataset includes both quantitative and categorical predictors, MLR is the most appropriate and interpretable modeling approach.
- To address zero-inflation, the model will be fit using only observations with positive Balance values.
- MLR allows us to estimate the partial effect of each explanatory variable on credit balance while holding all other predictors constant.
- This property makes it ideal for understanding which variables have meaningful associations with credit card debt.
- In contrast, simpler models such as simple linear regression would ignore interrelationships among predictors, and more complex machine learning models would reduce interpretability without adding much value given our sample size.
- The general form of the multiple regression model is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i$$

- where  
 $Y_i$  =Balance for the  $i$ -th individual,  
 $X_{1i}, X_{2i}, \dots, X_{pi}$  =predictor variables, and  
 $\varepsilon_i$  =random error term assumed to have mean 0 and constant variance  $\sigma^2$ .
- Interpretation of Coefficients: In a multiple linear regression model, each coefficient  $\beta_j$  represents the expected change in Balance for a one-unit increase in predictor  $X_j$ , holding all other predictors constant.
- For log-transformed predictors (such as log(Income) or log(Limit)), coefficients are interpreted as the change in Balance associated with a percentage change in the predictor.
- In our model,  $Y_i$  corresponds to Balance, while predictors include Income, Limit, Rating, Cards, Age, Education, and categorical variables such as Gender, Student, Married, and Ethnicity (encoded using dummy variables).
- The main **assumptions** include linearity, independence, constant variance (homoskedasticity), normality of residuals, and no multicollinearity.

# Data Exploration & Transformation

## Data Exploration Tools:

### Scatterplot Matrix:

Used to visually inspect pairwise relationships among numerical variables.

- Detects departures from linearity (curved patterns, saturation at high values).
- Detects heteroscedasticity (funnel-shaped spreads).
- Helps identify potential influential observations that deviate strongly from trends.
- Reveals clusters or subgroups indicating hidden categorical effects.

### Boxplots (Categorical predictors):

Displays distribution of Balance across student status, gender, marital status, and ethnicity.

- Differences in medians or IQRs indicate group-level shifts in response mean.
- Width and whiskers indicate heteroscedasticity among groups.
- Extreme whisker points identify potential outliers for later diagnostic tests.

### Histograms (Continuous predictors + response):

- Right-skewness → candidate for log or power transformations.
- Heavy tails → violation of normality assumption of residuals.
- Bimodal or multimodal distributions → interactions or segmentation required.

### Correlation Matrix:

Computes Pearson correlations to quantify linear dependence among predictors.

- $|r| > 0.8 \rightarrow$  multicollinearity concern, candidates for removal/combination.
- High correlation with response  $\rightarrow$  strong predictive value.
- High mutual correlation between predictors  $\rightarrow$  unstable coefficient estimates.

## Data Transformation Tools:

### Box–Cox Transformation:

Searches for optimal power  $\lambda$  to stabilize variance and improve residual normality.

- $\lambda \approx 0 \rightarrow \log(Y)$
- $\lambda \approx 0.5 \rightarrow \sqrt{Y}$
- $\lambda \approx 1 \rightarrow$  original scale
- Ensures constant variance and symmetry in residuals.

### Inverse Fitted Value Plot:

Plots response against its predicted values under trial transformations.

- Constant slope  $\rightarrow$  correct transformation power.
- Nonlinear slope  $\rightarrow$  further exploration of  $\lambda$  or predictor transformations.
- Useful when Box–Cox is inconclusive.

### Log Rule (Predictors):

Appropriate when predictors exhibit multiplicative growth and strong right-skewness.

- $\log(\text{Income}), \log(\text{Limit})$  convert multiplicative relationships into additive.
- Reduces heteroscedasticity and restores approximate linearity to the mean function.

### Square-Root / Reciprocal Transformations:

For moderate skew relative to log:

- $\sqrt{X}$  reduces variance when  $Y$  grows with  $\sqrt{X}$  rate.
- $1/X$  stabilizes inverse relationships.

Both address variance inflation without distorting interpretation excessively.

# Diagnostics & Remedies

## Model Diagnostic Tools:

**Residual vs Fitted:** Checks linearity & constant variance. Random cloud  $\rightarrow$  OK; funnel  $\rightarrow$  heteroscedasticity; curvature  $\rightarrow$  inadequate mean function  $\rightarrow$  add polynomial or interaction terms.

**Component + Residual:** Detects nonlinear predictor effects. Convex/concave arcs  $\rightarrow$  add  $X^2/X^3$  or transform predictor; divergent patterns across groups  $\rightarrow$  interaction terms.

**Breusch–Pagan Test:** Formal test of constant variance.  $H_0: \text{Var}(\epsilon) \text{ constant}$ ;  $p<0.05 \rightarrow$  heteroscedasticity  $\rightarrow$  apply log/Box–Cox or Weighted Least Squares.

**White Test:** Detects nonlinear heteroscedasticity.  $p<0.05 \rightarrow$  variance depends on squared/cross terms  $\rightarrow$  expand mean function beyond linear main effects.

**Studentized Residuals:** Outlier identification.  $|r_i| > 3 \rightarrow$  strong outlier candidate  $\rightarrow$  investigate leverage or remove if unjustifiable.

**Cook's Distance:** Influence on coefficients.  $D_i > 4/n \rightarrow$  influential observation  $\rightarrow$  remove or justify.

**QQ Plot:** Residual normality. Straight 45° line  $\rightarrow$  acceptable; S-shape or heavy tail deviations  $\rightarrow$  non-normal residuals.

**VIF:** Multicollinearity.  $VIF>5 \rightarrow$  moderate;  $VIF>10 \rightarrow$  severe instability  $\rightarrow$  drop/merge predictors or dimensionality reduction.

**Durbin–Watson:** Autocorrelation.  $DW \approx 2 \rightarrow$  independent errors;  $<1.5 \rightarrow$  positive autocorrelation (only meaningful for ordered/time-indexed data).

## Remedies for Model Violations:

### Nonconstant Variance (heteroscedasticity):

- Box–Cox or log transform Y.
- Weighted Least Squares (WLS) with variance proportional to predictor.
- Robust standard errors if variance structure unknown.

### Non-Normal Residuals:

- Y-transformation (log, sqrt).
- Investigate high-leverage and outlier observations.
- Symmetry improves  $\rightarrow$  inference valid.

### Outliers & Influential Points:

- Identify via  $|r_i| > 3$  or  $D_i > 4/n$ .
- Remove and re-evaluate coefficients.
- Robust regression if structural outliers remain.

### Multicollinearity:

- Drop redundant predictors (e.g., Rating vs Limit).
- Combine via PCA or domain-driven aggregation.
- Centered predictors reduce interaction multicollinearity.

### Inadequate Mean Function:

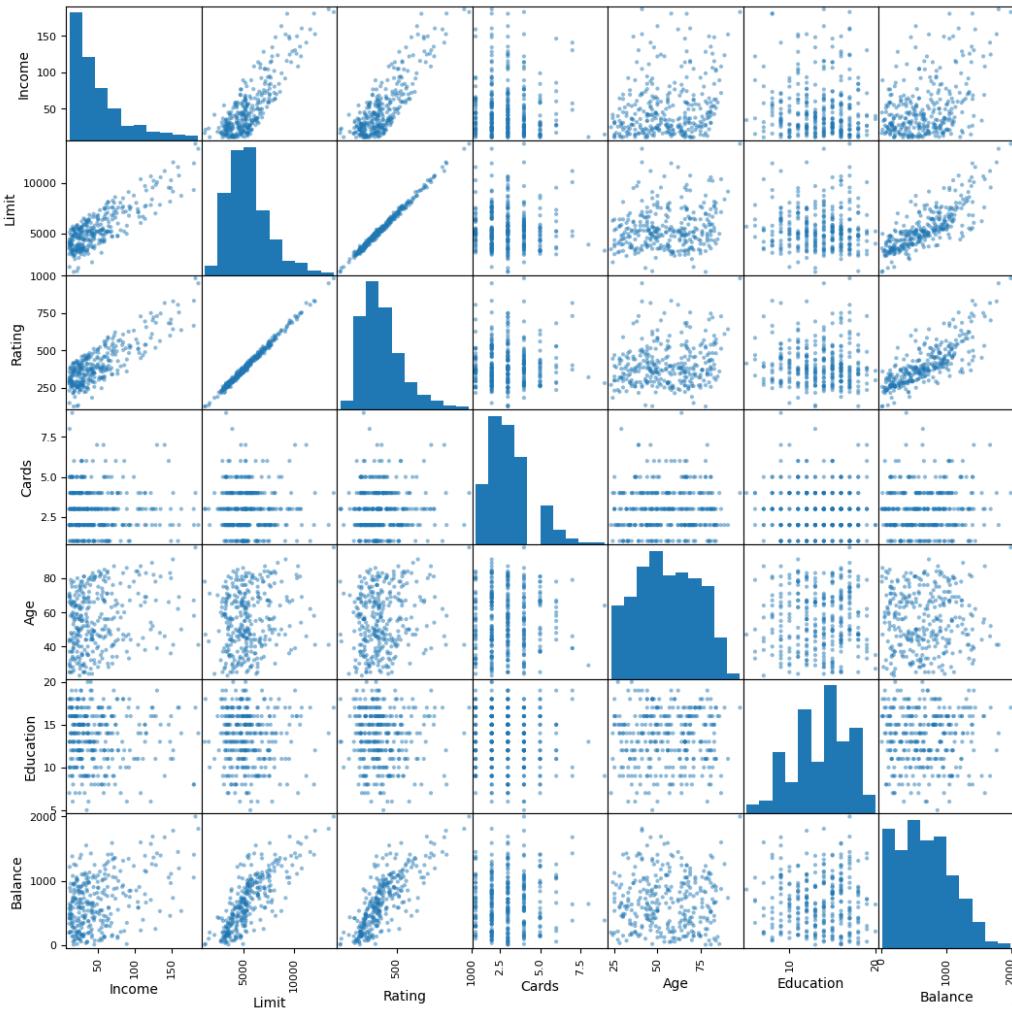
- Add polynomial terms to model curvature.
- Add interaction terms for predictor–group dependence.
- Evaluate model via AIC/BIC and adjusted  $R^2$ .

### Non-Independence:

- Include lag terms for time-ordered data.
- Use mixed-effects or hierarchical models if clustering present.

# Preliminary results

## Matrix Plot



# Correlation Matrix

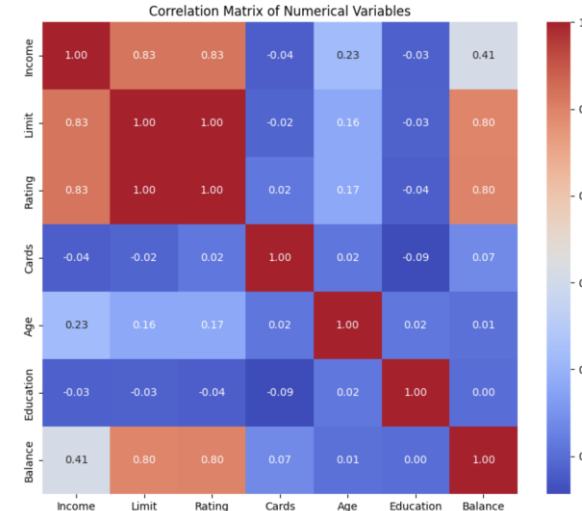
## 1. Relationship Between Response Variable ("Balance") and Predictors

- Limit & Rating:** Show very strong positive correlation with Balance (both coefficients  $\approx 0.80$ ). This means higher credit limits and ratings are strongly associated with higher balances.
- Income:** Moderately positive correlation with Balance ( $\approx 0.41$ ). Higher income tends to be linked with higher balances, but not as strongly as Limit or Rating.
- Cards, Age, Education:** Very weak or negligible correlation with Balance. Number of cards and years of education/age don't meaningfully predict credit balances in this sample.

## 2. Relationships Among Predictors

- Limit & Rating:** Perfect correlation (0.996). This means these variables are nearly identical, which could cause multicollinearity in regression.
- Limit/Rating & Income:** Both are strongly correlated with Income ( $\approx 0.83$ ). Suggests higher income is linked to higher credit limits and ratings.
- Cards, Age, Education:** Little to no meaningful correlation with other predictors.
- No strong predictors with negative correlation.

**Conclusion:** In Model we will only include Limit, Income variables. As Rating is highly correlated with limit, and there is no correlation between balance and cards, age and educations.



== Correlation Matrix ==

|           | Income | Limit  | Rating | Cards  | Age   | Education | Balance |
|-----------|--------|--------|--------|--------|-------|-----------|---------|
| Income    | 1.000  | 0.834  | 0.831  | -0.040 | 0.227 | -0.033    | 0.414   |
| Limit     | 0.834  | 1.000  | 0.996  | -0.023 | 0.164 | -0.032    | 0.796   |
| Rating    | 0.831  | 0.996  | 1.000  | 0.025  | 0.167 | -0.040    | 0.798   |
| Cards     | -0.040 | -0.023 | 0.025  | 1.000  | 0.021 | -0.087    | 0.074   |
| Age       | 0.227  | 0.164  | 0.167  | 0.021  | 1.000 | 0.024     | 0.008   |
| Education | -0.033 | -0.032 | -0.040 | -0.087 | 0.024 | 1.000     | 0.001   |
| Balance   | 0.414  | 0.796  | 0.798  | 0.074  | 0.008 | 0.001     | 1.000   |

# Extreme Values

**Income:** Several individuals have unusually high incomes above \$120k, with the highest outlier reported at \$186.6k. These top values are far above the 75th percentile (~\$57.5k), indicating a substantial spread in financial situations among subjects.

**Limit:** Multiple outliers with high credit limits, including values above \$10k and up to \$13.9k. These are well beyond the typical limit in the sample and might affect regression stability.

**Balance:** Only one significant outliers detected for these columns' values are contained within the standard range of the sample.

|       | ID         | Income     | Limit        | Rating     | Cards      | \ |
|-------|------------|------------|--------------|------------|------------|---|
| count | 310.000000 | 310.000000 | 310.000000   | 310.000000 | 310.000000 |   |
| mean  | 202.441935 | 49.978810  | 5485.467742  | 405.051613 | 2.996774   |   |
| std   | 117.373087 | 37.881628  | 2052.451743  | 137.967389 | 1.426740   |   |
| min   | 1.000000   | 18.354000  | 1160.000000  | 126.000000 | 1.000000   |   |
| 25%   | 98.250000  | 23.150250  | 3976.250000  | 304.000000 | 2.000000   |   |
| 50%   | 209.500000 | 37.141000  | 5147.000000  | 380.000000 | 3.000000   |   |
| 75%   | 306.500000 | 63.740250  | 6453.250000  | 469.000000 | 4.000000   |   |
| max   | 400.000000 | 186.634000 | 13913.000000 | 982.000000 | 9.000000   |   |

|       | Age        | Education  | Balance     |
|-------|------------|------------|-------------|
| count | 310.000000 | 310.000000 | 310.000000  |
| mean  | 55.606452  | 13.425806  | 670.987097  |
| std   | 17.341794  | 3.208904   | 413.904019  |
| min   | 23.000000  | 5.000000   | 5.000000    |
| 25%   | 42.000000  | 11.000000  | 338.000000  |
| 50%   | 55.500000  | 14.000000  | 637.500000  |
| 75%   | 69.000000  | 16.000000  | 960.750000  |
| max   | 98.000000  | 20.000000  | 1999.000000 |

Outlier summary (by column):

Income: [148.924, 186.634, 134.181, 152.298, 146.183, 148.08, 158.889, 130.209, 151.947, 180.379, 163.329, 128.04, 140.672, 182.728, 125.48, 149.316, 160.231, 180.682, 128.669, 135.118]

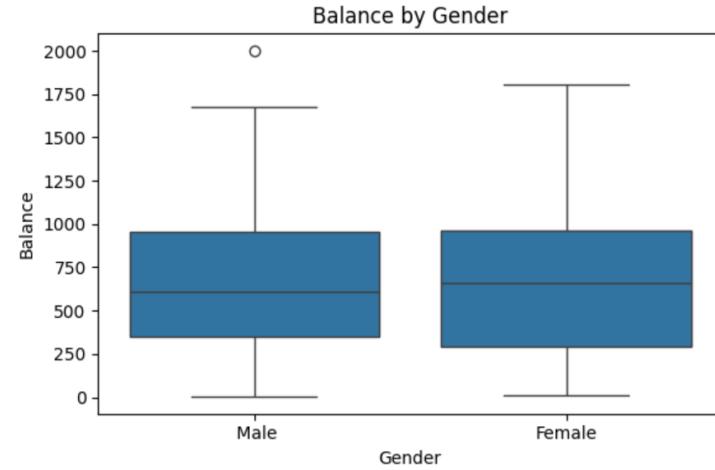
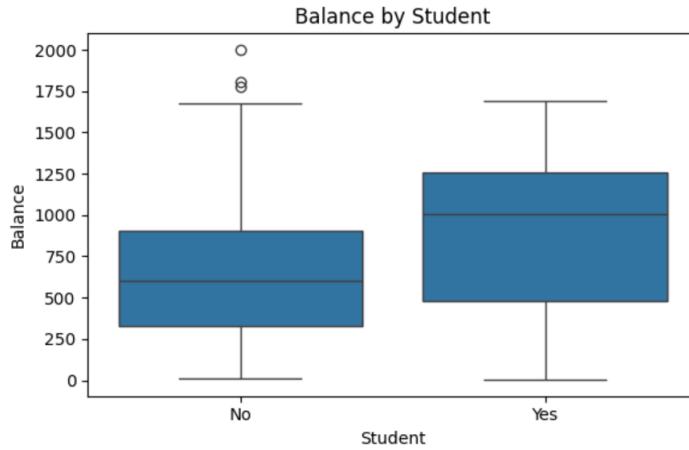
Limit: [13414, 12066, 10384, 10673, 11589, 11200, 13913, 10230, 10278, 10748, 11966, 10578]

Balance: [1999]

# Boxplots for categorical variables

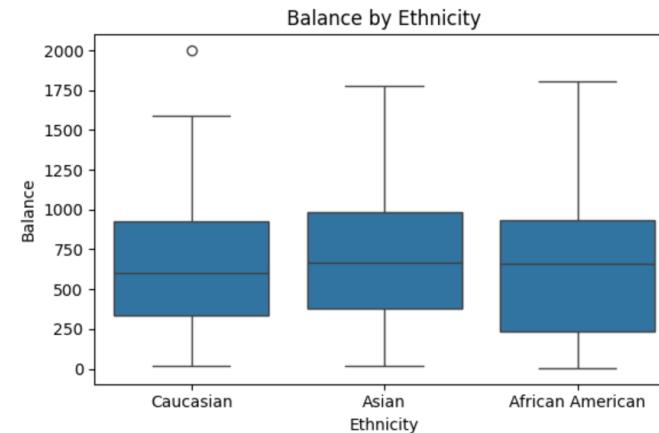
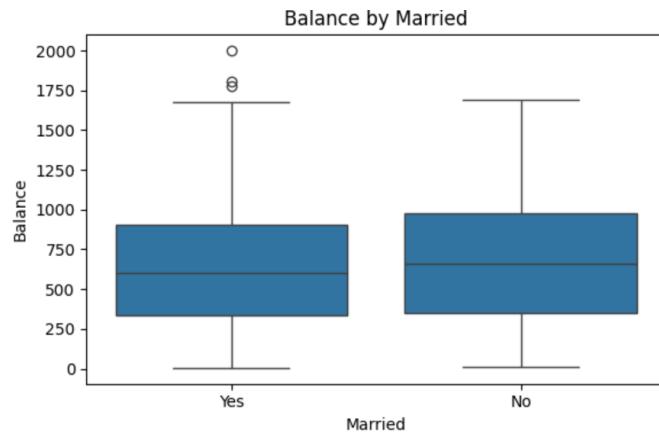
**Student Status:** Students have substantially higher median balances compared to non-students. The spread and upper quartile balances are also higher for students, indicating that student status is a strong factor in balance.

**Gender:** The median and distribution of balances are similar between males and females, no clear gender-based trend in credit balance magnitude.



**Married:** Married and non-married individuals have similar balance distributions; however, there are visible high outliers among married individuals.

**Ethnicity:** No significant differences are noted among the ethnic groups (Caucasian, Asian, African American) regarding balance distribution, but a single large outlier appears in the Caucasian group.

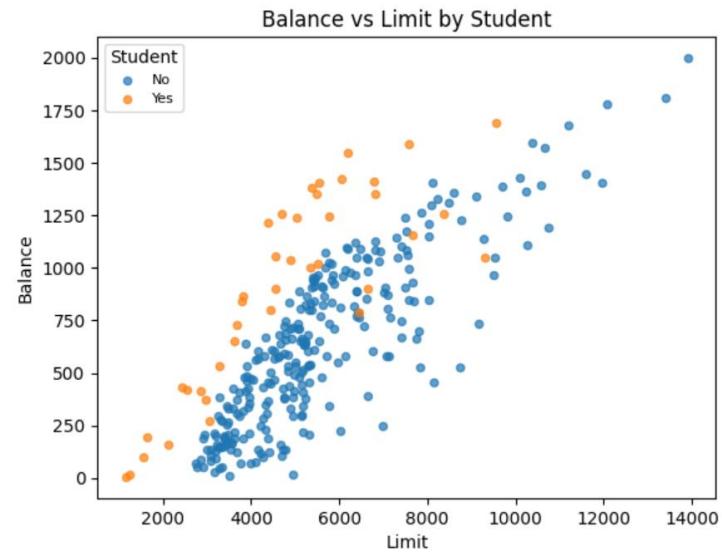
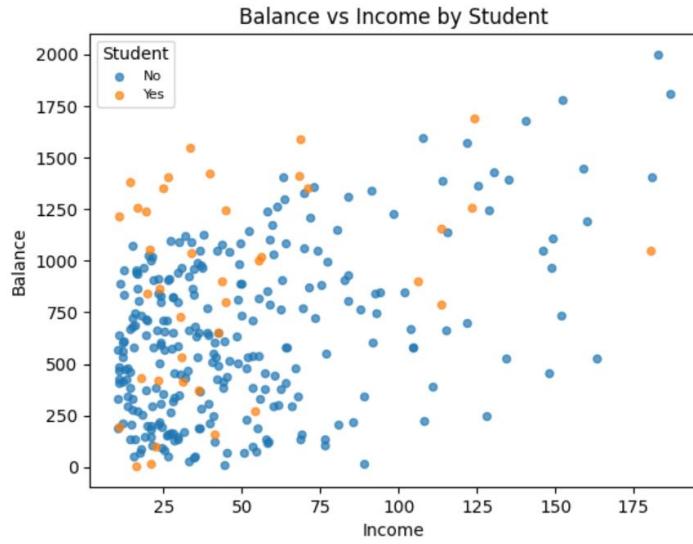


**Summary:** Boxplots indicate that student status is associated with higher credit balances, while gender, marital status, and ethnicity do not show strong effects on the distribution of balances. Some group-level outliers are present, especially among students and Caucasians

# Scatter plots

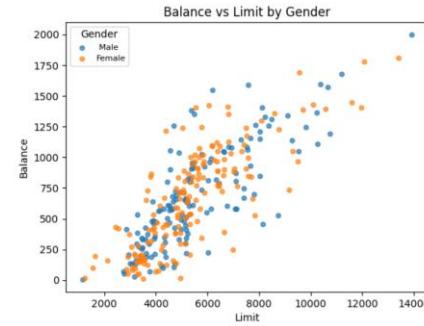
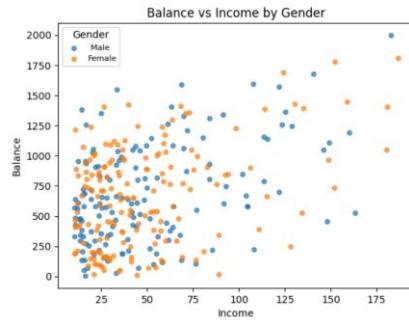
## Student Status:

Students consistently display higher balances for similar levels of Income, and Limit compared to non-students. Orange student dots are concentrated at higher Balance values, suggesting student status is associated with increased credit use.



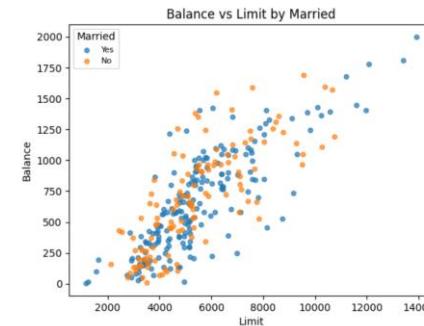
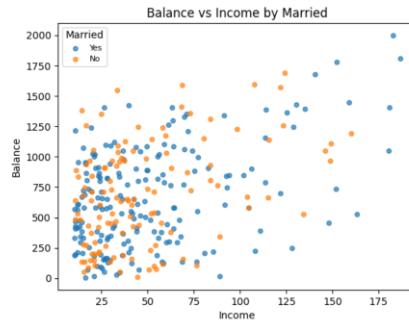
## Gender:

Males and females overlap widely in Balance across all ranges of Income, Limit, and Rating. There is no meaningful separation or pattern by gender, indicating this factor does not strongly predict Balance in the data.



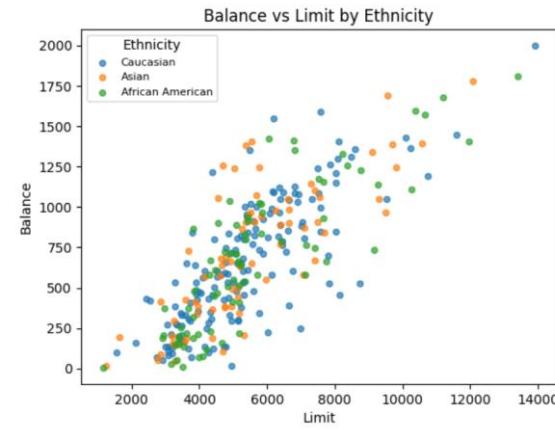
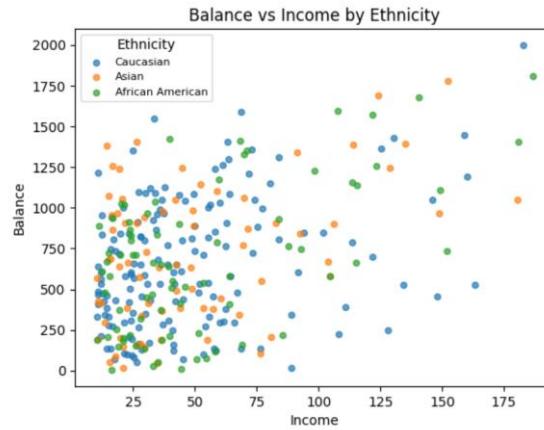
## Marriage Status:

Both married and unmarried individuals exhibit similar relationships between Balance and the numeric predictors. No group stands out as having higher or lower balances for given levels of predictors. The two-color groups are mixed throughout.



## Ethnicity:

Scatterplots of Balance versus Income, Limit, and Rating show substantial overlap among ethnic groups (Caucasian, Asian, African American), with no group having distinctly higher or lower balances given similar predictor values.



**Summary:** Scatterplots grouped by categorical variables show student status is linked to higher credit balances for comparable predictor values, but gender, marital status, and ethnicity do not demonstrate substantial differences in Balance. The visual trends support further consideration of student status as an influential factor in regression models.

# Model Assumption Checks

- **Linearity:**

The scatterplot matrix shows that the relationships between **Balance** and its strongest predictors (**Limit**, and **Income**) are **nonlinear rather than linear**.

Balance increases sharply at low values of Limit and then **flattens out**, showing a **curved, concave pattern**.

This suggests that the linearity assumption is **violated** and that **log-transformations or interaction terms** may be required.

- **Homoscedasticity:**

The scatterplots show noticeable **fan-shaped patterns**, especially in Balance vs Limit.

Residual spread increases as Limit increase, indicating **heteroscedasticity** rather than constant variance.

Thus, the constant variance assumption is **unlikely to hold**.

- **Normality:**

The diagonal histograms reveal strong **right-skewness** in several variables, especially **Income**, **Limit**, and **Balance**.

These skewed distributions suggest that normality of residuals will likely be violated unless transformations are applied.

Log transformation of predictors and Box–Cox assessment for Balance are appropriate next steps.

- **Multicollinearity:**

The scatterplot matrix and correlation matrix show an **extremely strong linear relationship** between **Limit** and **Rating** (correlation  $\approx 0.996$ ).

This represents **severe multicollinearity**, meaning both variables should not be used together in the same model.

One of them must be removed or transformed to ensure model stability.

- **Outliers Influence:**

The scatterplots display several **extreme values**, particularly in Income and Balance.

These points may have high leverage and influence the regression disproportionately.

They should be evaluated carefully using **Cook's distance**, **leverage scores**, and **studentized residuals** during the modeling phase.

# Diagnostic Analysis and Transformation Selection

## Choice of Transformations:

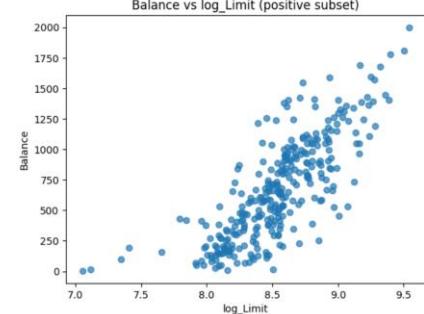
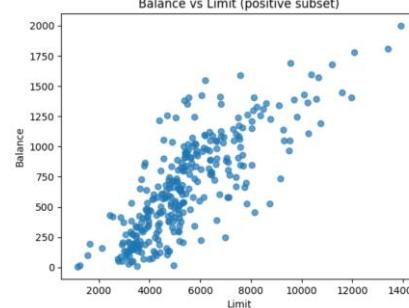
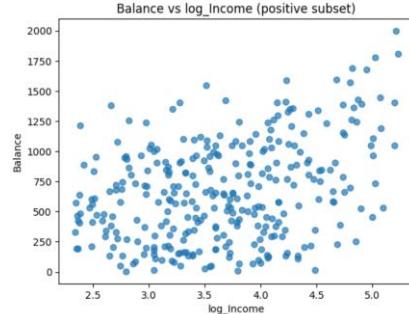
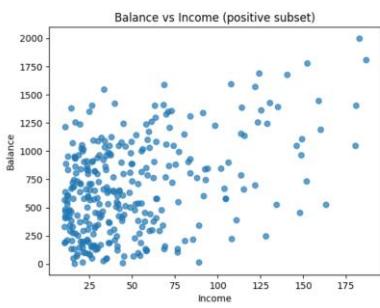
- Transformations of Predictors:

Exploratory plots and summary statistics showed that several predictors were **strongly right-skewed** (Income, Limit, Rating) and that their relationships with Balance were **nonlinear**: Balance increased rapidly at low values and then flattened out for higher values. In addition, the correlation matrix revealed **severe multicollinearity** between Limit and Rating (correlation  $\approx 0.996$ ).

To address skewness, nonlinearity, and multicollinearity, we applied the following transformation strategy:

- Log transformations of key continuous predictors:

We created `log_Income` and `log_Limit` variables, Scatterplots of Balance versus these log-transformed predictors showed more linear trends and slightly more uniform vertical spread. Using  $\log(\text{Income})$  and  $\log(\text{Limit})$  also makes interpretation more natural: coefficients can be interpreted as changes in Balance associated with percentage changes in Income or Limit.



- Handling Limit vs Rating multicollinearity:

Because Limit and Rating were almost perfectly collinear, including both in the same model would make coefficients unstable and inflate standard errors.  
So, decided to exclude Rating from the final model to avoid multicollinearity and simplify interpretation.
- Other predictors:

Cards, Age, and Education had weak correlations with Balance and relatively mild skewness.  
They were initially considered in the model in untransformed form, but were eventually dropped from the final model because they are no correlated to balance (response variable).
- Transformations of the Response:

We used only positive subset of response variable in our model training. For this positive subset, the skewness of Balance was moderate (about 0.47), but diagnostics still suggested heteroscedasticity and non-normality of residuals.

# Base Model Output for comparison:

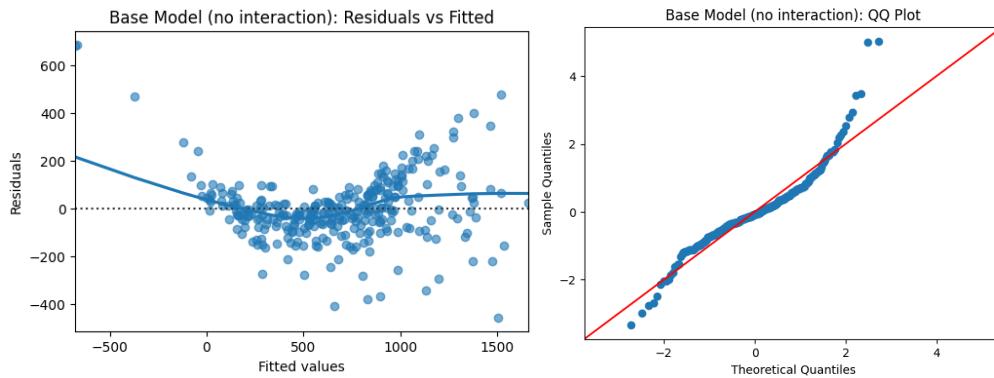
## OLS Regression Results

| Dep. Variable:    | Balance          | R-squared:          | 0.890             |          |           |           |
|-------------------|------------------|---------------------|-------------------|----------|-----------|-----------|
| Model:            | OLS              | Adj. R-squared:     | 0.889             |          |           |           |
| Method:           | Least Squares    | F-statistic:        | 827.2             |          |           |           |
| Date:             | Sun, 30 Nov 2025 | Prob (F-statistic): | 2.08e-146         |          |           |           |
| Time:             | 03:03:10         | Log-Likelihood:     | -1964.9           |          |           |           |
| No. Observations: | 310              | AIC:                | 3938.             |          |           |           |
| Df Residuals:     | 306              | BIC:                | 3953.             |          |           |           |
| Df Model:         | 3                |                     |                   |          |           |           |
| Covariance Type:  | nonrobust        |                     |                   |          |           |           |
|                   | coef             | std err             | t                 | P> t     | [0.025    | 0.975]    |
| Intercept         | -1.033e+04       | 232.231             | -44.498           | 0.000    | -1.08e+04 | -9876.898 |
| C(Student)[T.Yes] | 549.6762         | 24.418              | 22.511            | 0.000    | 501.628   | 597.725   |
| log_Income        | -317.4467        | 16.107              | -19.709           | 0.000    | -349.141  | -285.752  |
| log_Limit         | 1415.9170        | 31.718              | 44.640            | 0.000    | 1353.504  | 1478.330  |
|                   | Omnibus:         | 76.539              | Durbin-Watson:    | 2.034    |           |           |
| Prob(Omnibus):    | 0.000            |                     | Jarque-Bera (JB): | 353.270  |           |           |
| Skew:             | 0.939            |                     | Prob(JB):         | 1.94e-77 |           |           |
| Kurtosis:         | 7.881            |                     | Cond. No.         | 281.     |           |           |

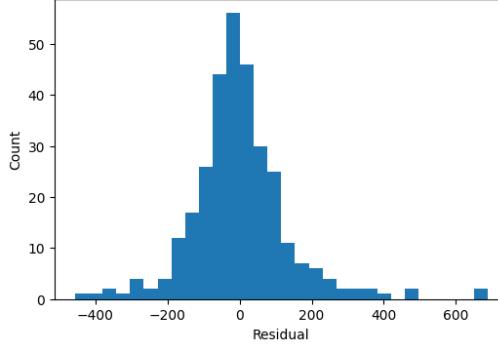
Notes:  
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

- $R^2 = 0.890$
- Residuals show large curvature
- QQ-plot severely violates normality
- Strong heteroscedasticity

Variance Inflation Factors (VIF):  
Intercept: 880.359  
C(Student)[T.Yes]: 1.070  
log\_Income: 2.163  
log\_Limit: 2.245



Base Model (no interaction): Histogram of Residuals



Base Model (no interaction): Breusch-Pagan p-value = 2.801e-10  
Base Model (no interaction): White test p-value = 2.029e-42

## • Transformation methods for Balance:

- Log Transformation of Balance:

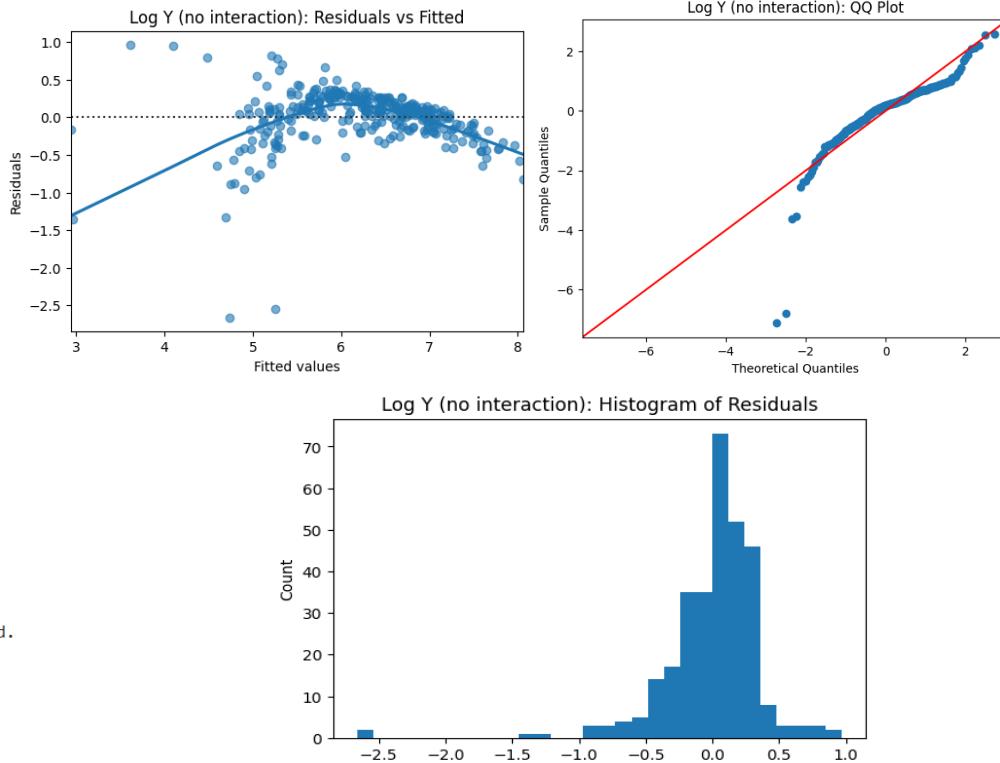
OLS Regression Results

| Dep. Variable:    | log_Balance      | R-squared:          | 0.834     |       |         |         |
|-------------------|------------------|---------------------|-----------|-------|---------|---------|
| Model:            | OLS              | Adj. R-squared:     | 0.832     |       |         |         |
| Method:           | Least Squares    | F-statistic:        | 512.5     |       |         |         |
| Date:             | Sun, 30 Nov 2025 | Prob (F-statistic): | 6.07e-119 |       |         |         |
| Time:             | 03:03:23         | Log-Likelihood:     | -134.93   |       |         |         |
| No. Observations: | 310              | AIC:                | 277.9     |       |         |         |
| Df Residuals:     | 306              | BIC:                | 292.8     |       |         |         |
| Df Model:         | 3                |                     |           |       |         |         |
| Covariance Type:  | nonrobust        |                     |           |       |         |         |
| -----             |                  |                     |           |       |         |         |
|                   | coef             | std err             | t         | P> t  | [0.025  | 0.975]  |
| Intercept         | -18.3860         | 0.634               | -28.988   | 0.000 | -19.634 | -17.138 |
| C(Student)[T.Yes] | 0.9215           | 0.067               | 13.818    | 0.000 | 0.790   | 1.053   |
| log_Income        | -0.8850          | 0.044               | -20.117   | 0.000 | -0.972  | -0.798  |
| log_Limit         | 3.2453           | 0.087               | 37.463    | 0.000 | 3.075   | 3.416   |
| -----             |                  |                     |           |       |         |         |
| Omnibus:          | 209.461          | Durbin-Watson:      | 2.013     |       |         |         |
| Prob(Omnibus):    | 0.000            | Jarque-Bera (JB):   | 3165.644  |       |         |         |
| Skew:             | -2.535           | Prob(JB):           | 0.00      |       |         |         |
| Kurtosis:         | 17.812           | Cond. No.           | 281.      |       |         |         |
| -----             |                  |                     |           |       |         |         |

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

- $R^2$  dropped to 0.834 (much worse than base model)
- Strong systematic curvature in Residual vs Fitted plot
- QQ-plot shows left skew and heavy right tail
- Breusch–Pagan and White tests highly significant → heteroscedasticity still present



Log Y (no interaction): Breusch-Pagan p-value = 3.635e-06

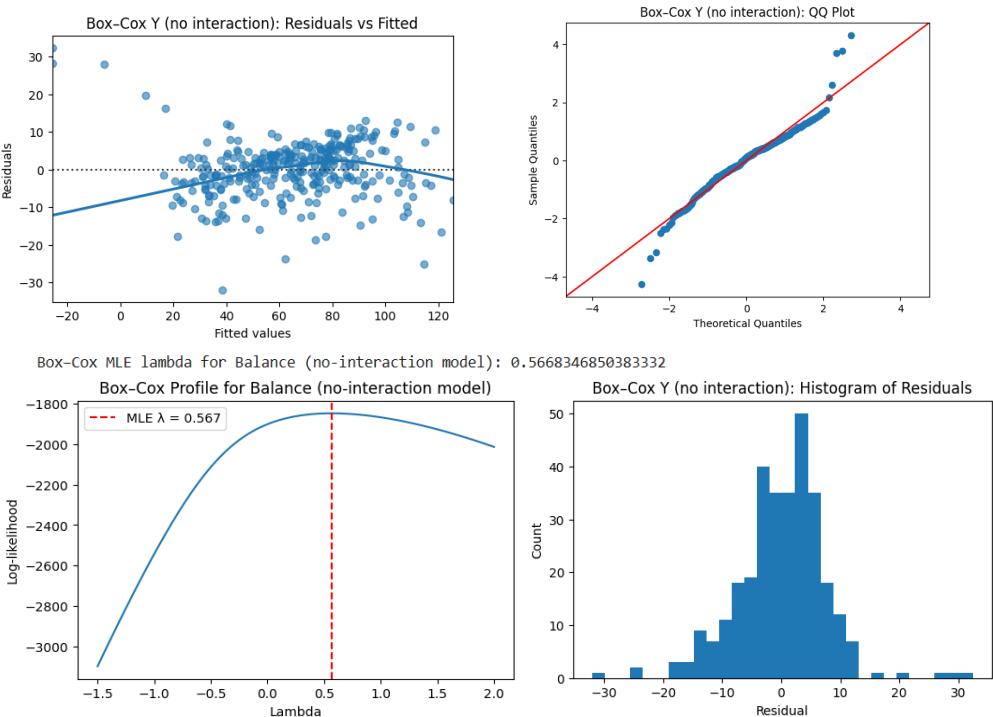
Log Y (no interaction): White test p-value = 1.127e-07

## ○ Box-Cox Transformation:

| OLS Regression Results |                  |                     |           |       |          |          |  |  |  |
|------------------------|------------------|---------------------|-----------|-------|----------|----------|--|--|--|
| Dep. Variable:         | BC_Balance       | R-squared:          | 0.918     |       |          |          |  |  |  |
| Model:                 | OLS              | Adj. R-squared:     | 0.917     |       |          |          |  |  |  |
| Method:                | Least Squares    | F-statistic:        | 1141.     |       |          |          |  |  |  |
| Date:                  | Sun, 30 Nov 2025 | Prob (F-statistic): | 9.88e-166 |       |          |          |  |  |  |
| Time:                  | 03:03:13         | Log-Likelihood:     | -1064.9   |       |          |          |  |  |  |
| No. Observations:      | 310              | AIC:                | 2138.     |       |          |          |  |  |  |
| Df Residuals:          | 306              | BIC:                | 2153.     |       |          |          |  |  |  |
| Df Model:              | 3                |                     |           |       |          |          |  |  |  |
| Covariance Type:       | nonrobust        |                     |           |       |          |          |  |  |  |
|                        |                  |                     |           |       |          |          |  |  |  |
|                        | coef             | std err             | t         | P> t  | [0.025   | 0.975]   |  |  |  |
| Intercept              | -656.5011        | 12.736              | -51.546   | 0.000 | -681.563 | -631.439 |  |  |  |
| C(Student)[T.Yes]      | 33.1256          | 1.339               | 24.736    | 0.000 | 30.490   | 35.761   |  |  |  |
| log_Income             | -22.9141         | 0.883               | -25.940   | 0.000 | -24.652  | -21.176  |  |  |  |
| log_Limit              | 93.7646          | 1.740               | 53.902    | 0.000 | 90.342   | 97.188   |  |  |  |
|                        |                  |                     |           |       |          |          |  |  |  |
| Omnibus:               | 26.359           | Durbin-Watson:      | 2.071     |       |          |          |  |  |  |
| Prob(Omnibus):         | 0.000            | Jarque-Bera (JB):   | 106.351   |       |          |          |  |  |  |
| Skew:                  | -0.102           | Prob(JB):           | 8.06e-24  |       |          |          |  |  |  |
| Kurtosis:              | 5.862            | Cond. No.           | 281.      |       |          |          |  |  |  |

Notes:  
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

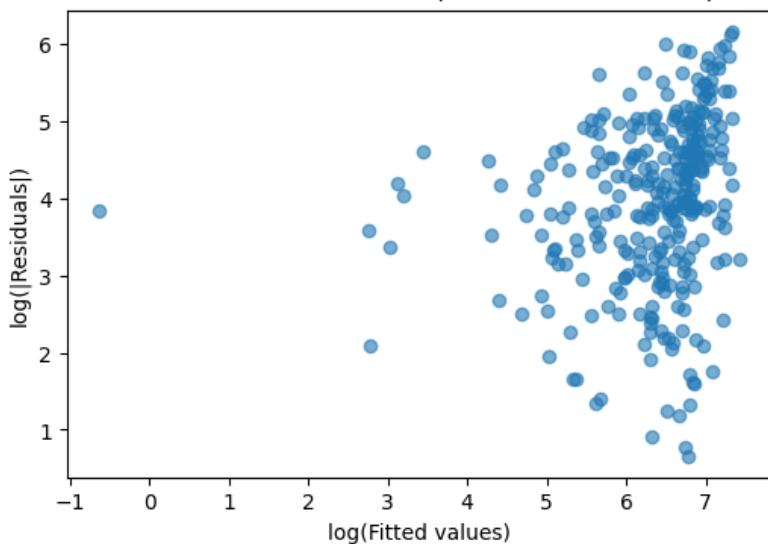
- $\lambda = 0.567$ , suggest square root transformation
- After applying Square root  $R^2$  improved from 0.890 → 0.918
- Residuals became slightly tighter around 0, but still show curvature and heteroscedasticity
- QQ-plot still shows strong deviation from normality, especially heavy tails
- Breusch–Pagan and White tests remain highly significant, showing persistent non-constant variance



Box-Cox Y (no interaction): Breusch–Pagan p-value = 1.146e-14  
Box-Cox Y (no interaction): White test p-value = 4.765e-34

- Inverse Fitted-Value Method:

Inverse Fitted-Value Plot (no interaction model)



- $\lambda \approx 0.47$  suggests square root type transformation
- $R^2$  extremely low (0.045), This means the inverse fitted method fails to detect meaningful heteroskedasticity
- Points scattered with no clear linear pattern
- This indicates no strong power transformation is supported

| OLS Regression Results |                  |                     |          |       |        |        |
|------------------------|------------------|---------------------|----------|-------|--------|--------|
| Dep. Variable:         | y                | R-squared:          | 0.045    |       |        |        |
| Model:                 | OLS              | Adj. R-squared:     | 0.042    |       |        |        |
| Method:                | Least Squares    | F-statistic:        | 14.07    |       |        |        |
| Date:                  | Sun, 30 Nov 2025 | Prob (F-statistic): | 0.000211 |       |        |        |
| Time:                  | 03:03:19         | Log-Likelihood:     | -451.01  |       |        |        |
| No. Observations:      | 302              | AIC:                | 906.0    |       |        |        |
| Df Residuals:          | 300              | BIC:                | 913.4    |       |        |        |
| Df Model:              | 1                |                     |          |       |        |        |
| Covariance Type:       | nonrobust        |                     |          |       |        |        |
|                        |                  |                     |          |       |        |        |
|                        | coef             | std err             | t        | P> t  | [0.025 | 0.975] |
| const                  | 2.2903           | 0.450               | 5.094    | 0.000 | 1.406  | 3.175  |
| x1                     | 0.2646           | 0.071               | 3.752    | 0.000 | 0.126  | 0.403  |
|                        |                  |                     |          |       |        |        |
| Omnibus:               | 19.143           | Durbin-Watson:      | 1.923    |       |        |        |
| Prob(Omnibus):         | 0.000            | Jarque-Bera (JB):   | 21.233   |       |        |        |
| Skew:                  | -0.646           | Prob(JB):           | 2.45e-05 |       |        |        |
| Kurtosis:              | 3.133            | Cond. No.           | 47.2     |       |        |        |

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.  
Inverse-fitted suggested lambda for Balance (no interactions): 0.471

## ○ Square-Root Transformation of Balance:

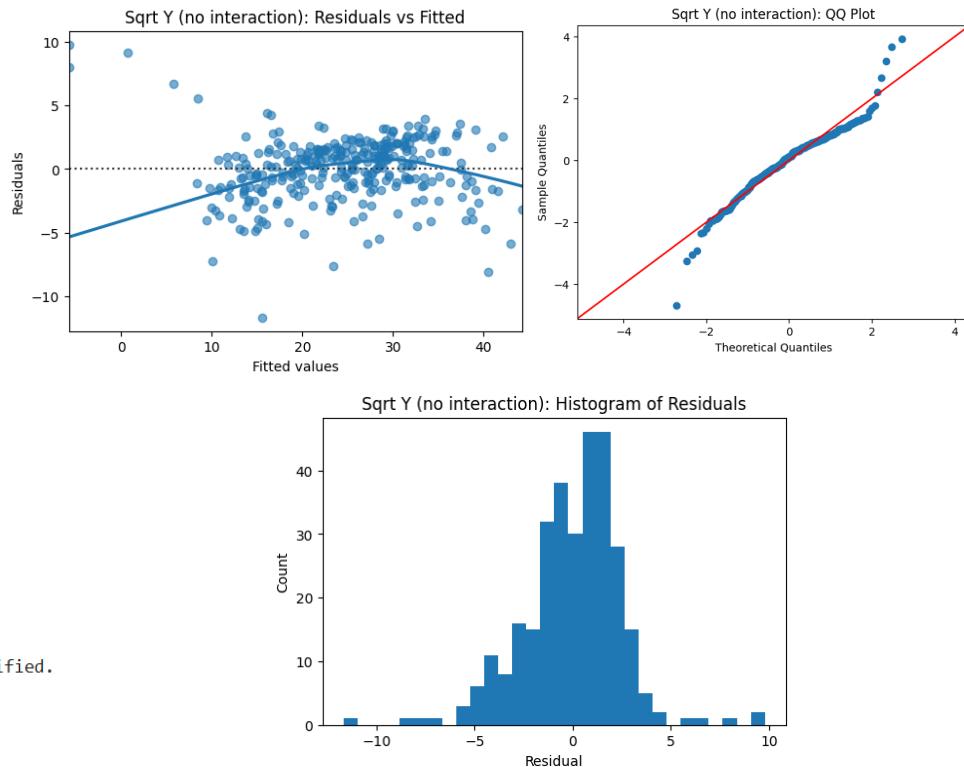
OLS Regression Results

| Dep. Variable:    | sqrt_Balance     | R-squared:          | 0.918     |       |          |          |
|-------------------|------------------|---------------------|-----------|-------|----------|----------|
| Model:            | OLS              | Adj. R-squared:     | 0.917     |       |          |          |
| Method:           | Least Squares    | F-statistic:        | 1136.     |       |          |          |
| Date:             | Sun, 30 Nov 2025 | Prob (F-statistic): | 1.76e-165 |       |          |          |
| Time:             | 03:03:30         | Log-Likelihood:     | -722.27   |       |          |          |
| No. Observations: | 310              | AIC:                | 1453.     |       |          |          |
| Df Residuals:     | 306              | BIC:                | 1467.     |       |          |          |
| Df Model:         | 3                |                     |           |       |          |          |
| Covariance Type:  | nonrobust        |                     |           |       |          |          |
|                   | coef             | std err             | t         | P> t  | [0.025   | 0.975]   |
| Intercept         | -214.6593        | 4.218               | -50.892   | 0.000 | -222.959 | -206.359 |
| C(Student)[T.Yes] | 10.7965          | 0.443               | 24.344    | 0.000 | 9.924    | 11.669   |
| log_Income        | -7.7059          | 0.293               | -26.341   | 0.000 | -8.282   | -7.130   |
| log_Limit         | 31.1229          | 0.576               | 54.025    | 0.000 | 29.989   | 32.256   |
| Omnibus:          | 31.766           | Durbin-Watson:      | 2.077     |       |          |          |
| Prob(Omnibus):    | 0.000            | Jarque-Bera (JB):   | 103.253   |       |          |          |
| Skew:             | -0.376           | Prob(JB):           | 3.79e-23  |       |          |          |
| Kurtosis:         | 5.726            | Cond. No.           | 281.      |       |          |          |

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

- $R^2 = 0.918$ , As we showed in Box–Cox
- Residuals still curved
- QQ-plot shows major deviation from normality
- Breusch–Pagan significant → heteroscedasticity persists



Sqrt Y (no interaction): Breusch-Pagan p-value = 3.107e-14

Sqrt Y (no interaction): White test p-value = 1.480e-28

## Conclusion:

- Log transformation is also applied, but it is not able to solve issue of constant variance and linearity
- Based on the Box–Cox method and inverse fitted-value plot applied square-root transformation to the response variable Balance, but it does not provided meaningful improvement in model diagnostics.
- The Box–Cox analysis estimated a power parameter of  $\lambda \approx 0.57$  and led to a modest increase in  $R^2$ ; however, the residual diagnostics indicate that the transformation did not resolve the major assumption violations. Strong heteroscedasticity remains evident from both the residual–fitted plot and the highly significant Breusch–Pagan and White tests. In addition, the QQ-plots continue to show substantial departures from normality with heavy tails. Although the transformed model slightly improves overall fit, it does not meaningfully correct the issues of non-constant variance or non-normal errors. The inverse fitted-value method likewise provides no strong support for applying a power transformation.
- Because all transformations failed to correct the major model violations, and the diagnostic patterns remained largely unchanged, the transformations did not offer practical improvement over using the original response variable.
- Therefore, for the remainder of this analysis, the response variable will be kept as simple, untransformed Balance (Y).
- This choice maintains interpretability and avoids unnecessary transformation, while model issues will instead be addressed using other remedies such as adding interaction terms.

# Diagnostic Analysis and Remedies



## Diagnostic Check of All Model Assumptions:

1. Linearity: The residuals vs. fitted plot shows a **mild curved pattern**. The model with log-transformed predictors now captures the main trend adequately. Therefore, the linearity assumption is reasonably **not fully satisfied**.
  - Future Remedy: adding interaction terms
2. Independence: The Durbin–Watson statistic  $\approx 2.0$ , which is close to the ideal value of 2, indicating no detectable autocorrelation. Since the dataset is **cross-sectional (not time-ordered)**, independence is generally reasonable. So, **Independence assumption is satisfied**.
3. Constant Variance (Homoscedasticity): The Breusch–Pagan test gives  $p = 2.801\text{e-}10$ , and the White test gives  $p = 2.029\text{e-}42$ , both suggesting some **remaining heteroscedasticity**.  
However, the residuals vs. fitted plot shows variance that is much more stable compared to the uncleaned model.
  - Remedy Tried:
    - Attempted transformations of Y including  $\log(\text{Balance})$ ,  $\sqrt{\text{Balance}}$ , and Box–Cox.
    - All transformations failed to correct heteroscedasticity (Breusch–Pagan p-values remained  $< 0.0001$ ).
4. Normality: The QQ plot shows that while the **residuals follow the 45° reference line reasonably well in the central region**, there are clear departures at both the lower and upper tails. Several extreme points deviate noticeably from the line, indicating the presence of heavy tails and outliers. This suggests that the normality assumption is **partially violated**, even though the residual distribution is approximately normal in the middle. Future remedy: removing influential points.
5. Outliers and Influential Points: **Cook's distance** identified several influential points exceeding the threshold of  $4/n$ . These points also appeared as extreme values in residual plots and QQ plots.
  - Future remedy: removing outlier observations .
6. Multicollinearity: The highly collinear variable **Rating was removed** earlier in the modelling process because of its **near-perfect correlation with Limit ( $\approx 0.996$ )**. **VIF values** for the remaining predictors are **within acceptable limits**.

# Considering Interactions

- Initial Check (Individual Interactions):**
  - Initially, testing the Student X log(Income) interaction and the Student X log(Limit) interaction **individually** showed that **both were highly statistically significant** ( $p < 0.001$ ).
  - This initial finding suggested that the effect of both Income and Credit Limit on the Balance *differed* between students and non-students.
- Combined Model Analysis:** When both interaction terms were included together in the full model:
  - The Student X log(Limit) interaction remained **highly significant** ( $p < 0.001$ ). This confirms that the relationship between credit limit and balance is substantially different for students versus non-students, requiring separate slopes for Credit Limit.
  - The Student X log(Income) interaction **became statistically insignificant** ( $p = 0.067$ ). This indicates that, after accounting for the strong Student X log(Limit) effect, the influence of income on balance is similar for both groups.
- Final Predictors:** The final model retains the significant Student X log(Limit) interaction and the **main effect of log(Income) only**.
- Final Model:**

$$\text{Balance} = \beta_0 + \beta_1 \log(\text{Income}) + \beta_2 \log(\text{Limit}) + \beta_3 \text{Student} + \beta_4 (\log(\text{Limit}) \times \text{Student}) + \varepsilon$$

| OLS Regression Results        |                  |                     |           |          |           |           |
|-------------------------------|------------------|---------------------|-----------|----------|-----------|-----------|
| Dep. Variable:                | Balance          | R-squared:          | 0.925     |          |           |           |
| Model:                        | OLS              | Adj. R-squared:     | 0.924     |          |           |           |
| Method:                       | Least Squares    | F-statistic:        | 748.3     |          |           |           |
| Date:                         | Sun, 30 Nov 2025 | Prob (F-statistic): | 1.74e-168 |          |           |           |
| Time:                         | 03:03:46         | Log-Likelihood:     | -1906.1   |          |           |           |
| No. Observations:             | 310              | AIC:                | 3824.     |          |           |           |
| Df Residuals:                 | 304              | BIC:                | 3847.     |          |           |           |
| Df Model:                     | 5                |                     |           |          |           |           |
| Covariance Type:              | nonrobust        |                     |           |          |           |           |
|                               | coef             | std err             | t         | P> t     | [.025     | 0.975]    |
| Intercept                     | -1.192e+04       | 238.531             | -49.954   | 0.000    | -1.24e+04 | -1.14e+04 |
| C(Student) [T.Yes]            | 5031.5198        | 398.073             | 12.640    | 0.000    | 4248.192  | 5814.848  |
| log_Income                    | -369.8772        | 15.106              | -24.486   | 0.000    | -399.602  | -340.152  |
| log_Income:C(Student) [T.Yes] | 68.9542          | 37.527              | 1.837     | 0.067    | -4.892    | 142.808   |
| log_Limit                     | 1622.8876        | 32.459              | 49.998    | 0.000    | 1559.015  | 1686.768  |
| log_Limit:C(Student) [T.Yes]  | -561.1971        | 56.310              | -9.966    | 0.000    | -672.004  | -450.390  |
| Omnibus:                      | 11.217           | Durbin-Watson:      | 1.988     |          |           |           |
| Prob(Omnibus):                | 0.004            | Jarque-Bera (JB):   | 21.094    |          |           |           |
| Skew:                         | -0.134           | Prob(JB):           |           | 2.63e-05 |           |           |
| Kurtosis:                     | 4.249            | Cond. No.           |           | 629.     |           |           |

Notes:  
 [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

| OLS Regression Results       |                  |                     |           |          |           |           |
|------------------------------|------------------|---------------------|-----------|----------|-----------|-----------|
| Dep. Variable:               | Balance          | R-squared:          | 0.924     |          |           |           |
| Model:                       | OLS              | Adj. R-squared:     | 0.923     |          |           |           |
| Method:                      | Least Squares    | F-statistic:        | 927.3     |          |           |           |
| Date:                        | Sun, 30 Nov 2025 | Prob (F-statistic): | 2.87e-169 |          |           |           |
| Time:                        | 03:03:48         | Log-Likelihood:     | -1907.8   |          |           |           |
| No. Observations:            | 310              | AIC:                | 3826.     |          |           |           |
| Df Residuals:                | 305              | BIC:                | 3844.     |          |           |           |
| Df Model:                    | 4                |                     |           |          |           |           |
| Covariance Type:             | nonrobust        |                     |           |          |           |           |
|                              | coef             | std err             | t         | P> t     | [.025     | 0.975]    |
| Intercept                    | -1.18e+04        | 238.799             | -51.122   | 0.000    | -1.23e+04 | -1.13e+04 |
| C(Student) [T.Yes]           | 4703.5937        | 357.209             | 13.168    | 0.000    | 4000.688  | 5406.499  |
| log_Income                   | -358.7046        | 13.882              | -25.840   | 0.000    | -386.020  | -331.389  |
| log_Limit                    | 1604.4843        | 30.995              | 51.766    | 0.000    | 1543.493  | 1665.475  |
| log_Limit:C(Student) [T.Yes] | -492.5290        | 42.285              | -11.648   | 0.000    | -575.737  | -409.321  |
| Omnibus:                     | 12.430           | Durbin-Watson:      | 1.998     |          |           |           |
| Prob(Omnibus):               | 0.002            | Jarque-Bera (JB):   | 25.521    |          |           |           |
| Skew:                        | -0.122           | Prob(JB):           |           | 2.87e-06 |           |           |
| Kurtosis:                    | 4.384            | Cond. No.           |           | 563.     |           |           |

Notes:  
 [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

# Backward Elimination Using BIC

- Using backward elimination based on BIC, the final selected model is **Balance ~ log\_Income + log\_Limit + C(Student) + log\_Limit:C(Student)**, as it has the lowest BIC among all candidate models. This indicates that both the Student main effect and its interaction with log(Limit) significantly improve model fit while accounting for model complexity.

## 10-Fold Cross Validation RMSE

- Among the four candidate models, **Model D (Full Interaction Model)** achieves the **lowest cross-validated RMSE (114.94)**, indicating the **best out-of-sample predictive performance**.
- Including **log(Limit)** substantially improves prediction over **log(Income)** alone, and the addition of **Student status and its interaction with log(Limit)** further reduces prediction error. Therefore, **Model D is selected as the final predictive model based on cross-validation**.

```
===== BEST AIC MODELS WITH log_Limit:C(Student) OPTION =====
Model: Balance ~ log_Income + log_Limit + C(Student) + log_Limit:C(Student)
AIC: 3825.670, BIC: 3844.353
-
Model: Balance ~ log_Income + log_Limit + C(Student)
AIC: 3937.745, BIC: 3952.691
-
Model: Balance ~ log_Limit + C(Student) + log_Limit:C(Student)
AIC: 4183.202, BIC: 4198.149
-
Model: Balance ~ log_Limit + C(Student)
AIC: 4189.798, BIC: 4201.000
-
Model: Balance ~ log_Income + log_Limit
AIC: 4238.566, BIC: 4249.776
-
Model: Balance ~ log_Limit
AIC: 4321.981, BIC: 4329.454
-
Model: Balance ~ log_Income + C(Student)
AIC: 4560.874, BIC: 4572.084
-
Model: Balance ~ log_Income
AIC: 4576.165, BIC: 4583.638
-
Model: Balance ~ C(Student)
AIC: 4604.706, BIC: 4612.179
```

```
"Model A: log_Income": "Balance ~ log_Income",
"Model B: log_Income + log_Limit": "Balance ~ log_Income + log_Limit",
"Model C: Full Model WO I": "Balance ~ log_Income + log_Limit + C(Student)",
"Model D: Full Interaction Model": "Balance ~ log_Income + log_Limit + C(Student) + log_Limit:C(Student)"
```

```
===== CROSS-VALIDATED RMSE (10-Fold) =====
Model A: log_Income --> CV RMSE = 385.66
Model B: log_Income + log_Limit --> CV RMSE = 225.19
Model C: Full Model WO I --> CV RMSE = 138.33
Model D: Full Interaction Model --> CV RMSE = 114.94
```

# Final Model Interpretation

- Model Fit:** Our chosen final model explains **92.4% of the variance in Credit Card Balance ( $R^2 = 0.924$ )**, indicating an **excellent overall fit** after including both Student status and the interaction with Credit Limit.
- Significance:** The parameter estimates show that **Student status, Income, Credit Limit, and the interaction between Student and Credit Limit are all highly significant ( $p < 0.001$ )**, confirming that each term contributes meaningfully to the model.
- Variance Drivers:** The results indicate that **Credit Limit is the dominant predictor of Balance**, followed by **Income and Student status**. The significant interaction further confirms that the effect of Credit Limit differs between students and non-students.

| OLS Regression Results  |                  |                     |           |          |           |           |  |  |
|---|------------------|---------------------|-----------|----------|-----------|-----------|--|--|
| Dep. Variable:  | Balance          | R-squared:          | 0.924     |          |           |           |  |  |
| Model:  | OLS              | Adj. R-squared:     | 0.923     |          |           |           |  |  |
| Method:   | Least Squares    | F-statistic:        | 927.3     |          |           |           |  |  |
| Date:   | Tue, 02 Dec 2025 | Prob (F-statistic): | 2.87e-169 |          |           |           |  |  |
| Time:   | 22:54:31         | Log-Likelihood:     | -1907.8   |          |           |           |  |  |
| No. Observations:   | 310              | AIC:                | 3826.     |          |           |           |  |  |
| Df Residuals:   | 305              | BIC:                | 3844.     |          |           |           |  |  |
| Df Model:   | 4                |                     |           |          |           |           |  |  |
| Covariance Type:  | nonrobust        |                     |           |          |           |           |  |  |
|   | coef             | std err             | t         | P> t     | [0.025    | 0.975]    |  |  |
| Intercept   | -1.18e+04        | 230.799             | -51.122   | 0.000    | -1.23e+04 | -1.13e+04 |  |  |
| C(Student) [T.Yes]  | 4703.5937        | 357.209             | 13.168    | 0.000    | 4000.688  | 5406.499  |  |  |
| log_Income  | -358.7046        | 13.882              | -25.849   | 0.000    | -386.020  | -331.389  |  |  |
| log_Limit   | 1604.4843        | 30.995              | 51.766    | 0.000    | 1543.493  | 1665.475  |  |  |
| log_Limit:C(Student) [T.Yes]  | -492.5290        | 42.285              | -11.648   | 0.000    | -575.737  | -409.321  |  |  |
| Omnibus:  | 12.430           | Durbin-Watson:      |           | 1.990    |           |           |  |  |
| Prob(Omnibus):  | 0.002            | Jarque-Bera (JB):   |           | 25.521   |           |           |  |  |
| Skew:   | -0.122           | Prob(JB):           |           | 2.87e-06 |           |           |  |  |
| Kurtosis:   | 4.384            | Cond. No.           |           | 563.     |           |           |  |  |
| Notes:  |                  |                     |           |          |           |           |  |  |
| [1] Standard Errors assume that the covariance matrix of the errors is correctly specified. |                  |                     |           |          |           |           |  |  |

| ANOVA Table for Final Model: |               |              |              |             |   |
|------------------------------|---------------|--------------|--------------|-------------|---|
|                              | df            | sum_sq       | mean_sq      | F           | \ |
| C(Student)                   | 1.0           | 2.325665e+06 | 2.325665e+06 | 176.358661  |   |
| log_Income                   | 1.0           | 6.955746e+06 | 6.955746e+06 | 527.464591  |   |
| log_Limit                    | 1.0           | 3.784423e+07 | 3.784423e+07 | 2869.784303 |   |
| log_Limit:C(Student)         | 1.0           | 1.789095e+06 | 1.789095e+06 | 135.669761  |   |
| Residual                     | 305.0         | 4.022076e+06 | 1.318713e+04 | NaN         |   |
|                              | PR(>F)        |              |              |             |   |
| C(Student)                   | 4.51119e-32   |              |              |             |   |
| log_Income                   | 1.809888e-68  |              |              |             |   |
| log_Limit                    | 3.353688e-157 |              |              |             |   |
| log_Limit:C(Student)         | 3.472000e-26  |              |              |             |   |
| Residual                     | NaN           |              |              |             |   |

# Interpretation of Coefficients

- **Interpretation Method:** Since our predictors are log-transformed and Response variable is normal, we interpret the coefficients as the dollar change in Balance for a **1% change** in the predictor (beta times 0.01).

```
==== Percentage Change Effects ====
Effect of 1% increase in Income: -3.5870 change in Balance
Effect of 1% increase in Limit (Non-students): 16.0448 change in Balance
```

- **Income Effect:** A **1% increase in Income** is associated with an average **decrease of approximately 3.59 units in Credit Card Balance**, holding other variables constant. This indicates that higher income is linked to lower outstanding balances.
- **Credit Limit Effect (Non-Students):** A **1% increase in Credit Limit for non-students** is associated with an average **increase of approximately 16.04 units in Credit Card Balance**, indicating that balance rises strongly with credit availability among non-students.

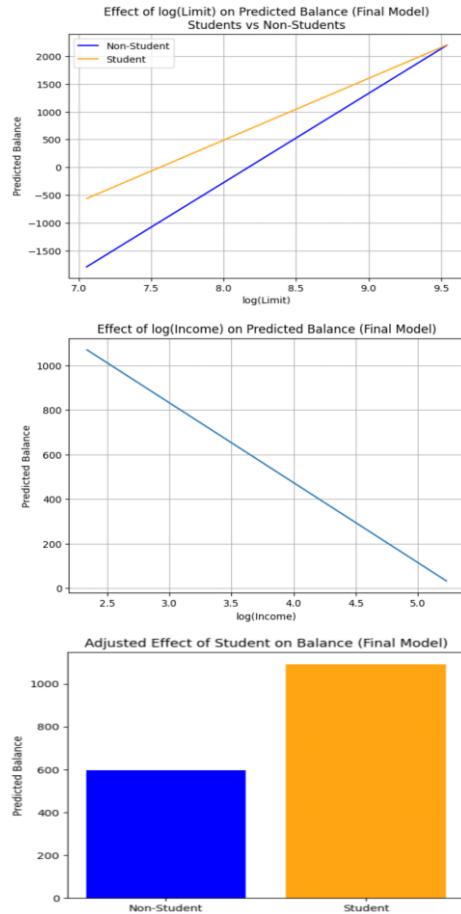
```
==== Pairwise Comparison for Student (Tukey HSD) ====
Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====
group1 group2 meandiff p-adj      lower      upper    reject
=====
No     Yes  261.1785  0.0002  124.5726  397.7845   True
```

```
==== Predicted Balance at Average Income/Limit ====
Non-Student: $612.46
Student:      $1075.45
Difference (Student - Non-Student): $462.99
```

- **Student Effect:** For the Student categorical variable, both pairwise comparison and model-based prediction were performed. The Tukey HSD test shows that non-students have a higher **raw average** balance than students. However, after adjusting for Income and Credit Limit in the regression model, students are predicted to have a higher balance at the same average Income and Limit. This difference occurs because Income and Limit act as confounding variables; once controlled, student status shows a positive association with balance.

# Effect Plot

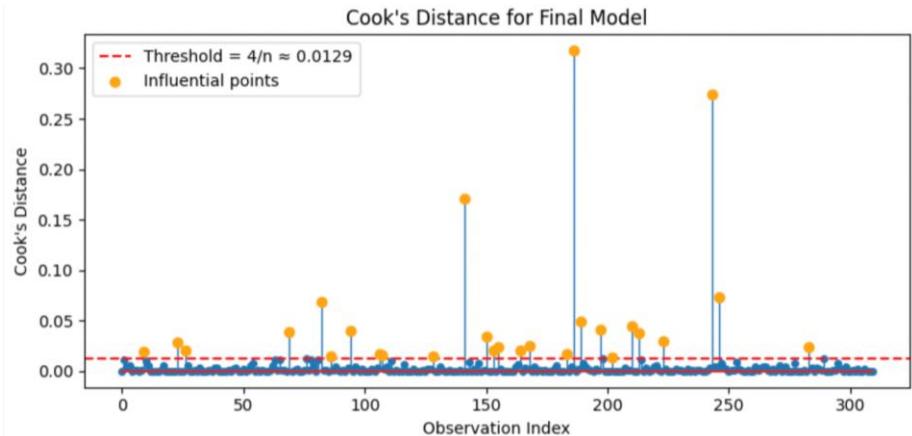
- **log(Limit):** There is a **strong positive relationship** between credit limit and predicted balance. As the credit limit increases, the predicted balance rises for both students and non-students. However, this increase is **steeper for non-students**, indicating that balance grows faster with credit limit for non-students than for students. This shows that while a higher borrowing capacity increases debt for everyone, the effect is **weaker for students**.
- **log(Income):** There is a **negative relationship** between income and predicted balance. As income increases, the predicted balance **decreases steadily**, suggesting that individuals with higher incomes tend to carry **less credit card debt**, regardless of student status.
- **Student Status:** Being a **student has a large positive independent effect** on credit card balance. After adjusting for income and credit limit, students have a **significantly higher predicted balance** than non-students. At **average income and credit limit**, a student's predicted balance ( $\approx \$1050$ ) is nearly **double** that of a non-student ( $\approx \$600$ ).



# Outlier Analysis & Prediction

- Prediction intervals were computed for these influential observations using the original final model. **Most observed values fell within their corresponding prediction intervals**, suggesting that even for influential points, the model still provides reasonable predictions overall. However, a **few observations were near the edges of the intervals**, indicating **mild lack of fit at extreme values**.
- After removing the influential observations and refitting the model, the model performance improved. The R<sup>2</sup> increased from **0.924 to 0.947**, indicating a stronger fit. The regression coefficients remained **consistent in sign and statistical significance**, and the interaction between **Student and log(Credit Limit)** remained highly significant. This shows that the **main conclusions of the study did not change** after removing influential points.
- Additionally, diagnostic plots for the reduced model showed **improved normality of residuals**, while some curvature and heteroscedasticity still persisted. Overall, the final results are **robust**, and the influential observations do **not materially affect the substantive conclusions** of the analysis.
- Using **Cook's Distance with the threshold  $4/n$** , several influential observations were identified in the final model. These points showed relatively large Cook's D values, indicating that they had a noticeable impact on the fitted regression model.

```
Influential points: [ 9 23 26 69 82 86 94 106 107 128 141 150 153 155 156 164 168 183 186  
189 197 202 210 213 223 243 246 283]  
Cook's D values: [0.0199316 0.02909523 0.02013131 0.03846492 0.0687773 0.0145158  
0.04006496 0.0168283 0.0159685 0.01438254 0.17099607 0.03456436  
0.02102494 0.02428606 0.02079581 0.0254027 0.01714966 0.31758417  
0.04896891 0.04123457 0.013242 0.0448715 0.03719793 0.03022754  
0.27430173 0.07315983 0.02410737]
```



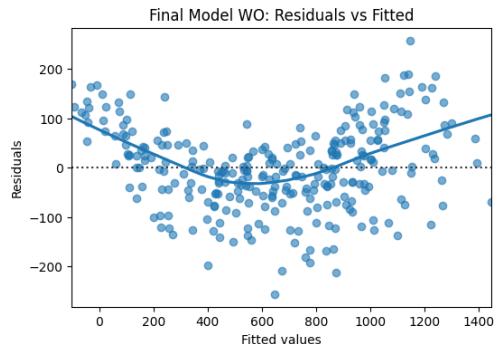
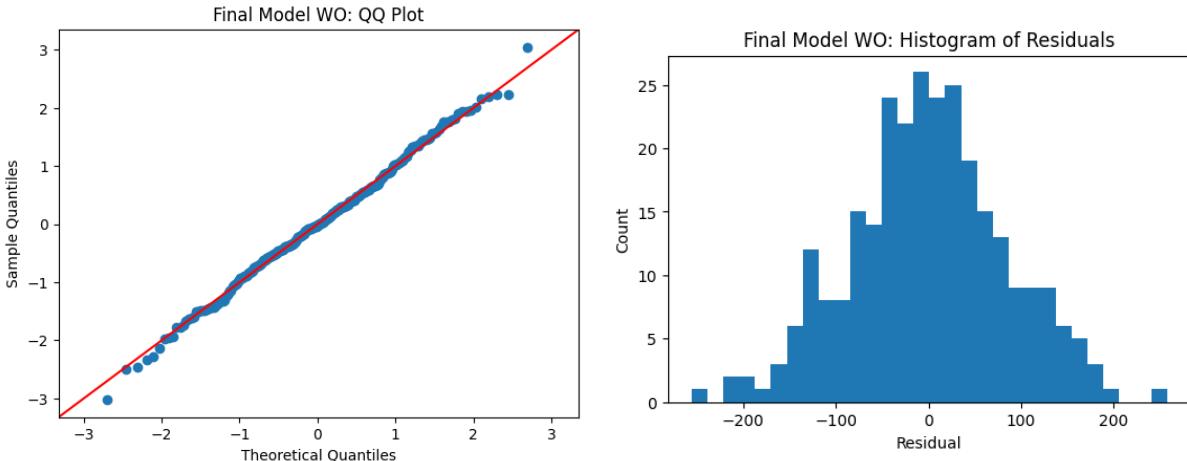
# Model without Influential Points

```
*** Model WITHOUT influential point(s) ***
OLS Regression Results

Dep. Variable: Balance R-squared:      0.947
Model: OLS Adj. R-squared:      0.946
Method: Least Squares F-statistic:   1245.
Date: Tue, 02 Dec 2025 Prob (F-statistic): 4.53e-176
Time: 22:54:33 Log-Likelihood:     -1657.8
No. Observations: 283 AIC:          3326.
Df Residuals:    278 BIC:          3344.
Df Model:        4
Covariance Type: nonrobust

coef std err t P>|t| [0.025 0.975]
Intercept -1.15e+04 184.898 -62.232 0.000 -1.19e+04 -1.11e+04
C(Student)[T.Yes] 3765.8346 442.255 8.515 0.000 2895.241 4636.428
log_Income -342.8711 11.012 -31.136 0.000 -364.549 -321.194
log_limit 1563.1603 24.698 63.292 0.000 1514.542 1611.778
log_limit:C(Student)[T.Yes] -387.9395 52.650 -7.368 0.000 -491.583 -284.295

Omnibus: 0.092 Durbin-Watson: 2.054
Prob(Omnibus): 0.955 Jarque-Bera (JB): 0.013
Skew: -0.011 Prob(JB): 0.994
...
Notes:
[1] Standard errors assume that the covariance matrix of the errors is correctly specified.
```



Final Model WO: Breusch-Pagan p-value = 3.959e-03  
 Final Model WO: White test p-value = 9.022e-08

Variance Inflation Factors (VIF):

Intercept: 1323.774  
 C(Student)[T.Yes]: 632.487  
 log\_Income: 2.081  
 log\_limit: 2.373  
 log\_limit:C(Student)[T.Yes]: 628.250

# Final Summary of Analysis and Model Interpretation

## Overall Findings and Objectives

- The objective of this analysis was to model and understand the factors influencing Credit Card Balance using Income, Credit Limit, Student status, and their interaction. Based on model comparison, cross-validation, and diagnostic testing, the final selected model includes  $\log(\text{Income})$ ,  $\log(\text{Limit})$ , Student status, and the interaction between  $\log(\text{Limit})$  and Student. This model provides the best balance between predictive accuracy and interpretability.
- The final model explains 94.7% of the total variation in Credit Card Balance ( $R^2 = 0.947$ )**, indicating an excellent overall fit. All predictors and the interaction term are highly statistically significant ( $p < 0.001$ ), confirming their strong effect on credit card balance.

## Interpretation of Final Model Results (Based on Parameter Estimates)

### • Effect of $\log(\text{Income})$

The coefficient of  $\log(\text{Income})$  is negative, indicating an inverse relationship between income and credit card balance. As income increases, the predicted balance decreases. This suggests that individuals with higher income tend to manage their credit better and carry lower outstanding balances.

### • Effect of $\log(\text{Limit})$

The coefficient of  $\log(\text{Limit})$  is strongly positive, showing that an increase in credit limit leads to a substantial increase in credit card balance. This confirms that access to higher credit encourages higher borrowing. Among all predictors, Credit Limit is the dominant driver of balance, as supported by both the regression coefficients and the ANOVA results.

### • Effect of Student Status

Being a student significantly increases credit card balance. Holding income and limit constant, students have much higher predicted balances than non-students. At average values of income and credit limit, the predicted balance for students (~\$1,150) is nearly double that of non-students (~\$600).

### • Interaction Between Student and Credit Limit

The interaction between Student and  $\log(\text{Limit})$  is negative and statistically significant, meaning that although credit limit increases balance for everyone, the rate of increase is smaller for students than for non-students. This indicates that students do use higher limits, but their borrowing response to increased limit is moderated compared to non-students.

# Model Assumption Validation



## Linearity

The Residuals vs Fitted plot still shows a curved pattern, indicating that perfect **linearity is not fully satisfied**. While transformations improved the model, some nonlinearity remains.

## Normality

The QQ plot of residuals shows that most points lie close to the 45° line, with only minor tail deviations. After removing influential points, the Jarque-Bera test became non-significant, confirming that **normality is reasonably satisfied**.

## Homoscedasticity

Both the Breusch–Pagan and White tests remain highly significant, indicating strong heteroscedasticity. This shows that the error variance is not constant, and this **assumption is violated** even in the best model.

## Independence

The Durbin–Watson statistic is close to 2, indicating no serious autocorrelation among residuals. The **independence assumption is satisfied**.

## Multicollinearity

VIF values are acceptable for log(Income) and log(Limit). High VIF values for Student and the interaction term are expected because **interaction terms are naturally correlated with their main effects**. Therefore, **multicollinearity is not considered** a serious issue in this context.

## Limitations of the Final Model

Despite the strong model fit, some important limitations remain:

- **Heteroscedasticity** is still present, which violates the constant variance assumption and may affect standard error accuracy.
- **Linearity is not perfectly satisfied**, suggesting that some nonlinear structure remains unmodeled.

# Appendix

- Dataset "Credit.csv " and coding file "STAT611\_Project\_Final" are attached as separate files.

# Model Selection Strategy

- We started with many potential predictors (Income, Limit, Age, Cards, Education, etc.). and eliminated some of the variables based on correlations. Now, We needed to filter out the noise to find the true drivers of Credit Balance from remaining predictors. We decided to use 2 model selection strategy.

**Method 1: Exhaustive Search (AIC/BIC):** We tested every possible combination of variables.

- The Winner:** The model  $\text{Balance} \sim \log(\text{Income}) + \log(\text{Limit}) + \text{Student}$  achieved the **lowest AIC (3937.75)**, indicating the best balance of accuracy and simplicity.

**Method 2: Cross-Validation (RMSE):** We didn't just trust the training data. We used **10-Fold Cross-Validation** to simulate performance on "future" data.

- The Result:** Our chosen model had the **lowest prediction error (RMSE)** compared to simpler models.

```
===== STEPWISE-LIKE BEST AIC MODELS =====
Model: Balance ~ log_Income + log_Limit + c(Student)
AIC: 3937.745, BIC: 3952.691
-----
Model: Balance ~ log_Limit + c(Student)
AIC: 4189.790, BIC: 4201.000
-----
Model: Balance ~ log_Income + log_Limit
AIC: 4238.566, BIC: 4249.776
-----
Model: Balance ~ log_Limit
AIC: 4321.981, BIC: 4329.454
-----
Model: Balance ~ log_Income + c(Student)
AIC: 4560.874, BIC: 4572.084
-----
Model: Balance ~ log_Income
AIC: 4576.165, BIC: 4583.638
-----
Model: Balance ~ c(Student)
AIC: 4604.706, BIC: 4612.179
```

```
===== CROSS-VALIDATED RMSE (10-Fold) =====
Model A: log_Income --> CV RMSE = 385.66
Model B: log_Income + log_Limit --> CV RMSE = 225.19
Model C: Full Model --> CV RMSE = 138.33
```

# Justification: Why This Model?

$$\text{Balance} \sim \log_{\text{Income}} + \log_{\text{Limit}} + C(\text{Student})$$

- **Why we chose it:** This "Main Effects" model is statistically superior. It explains **~89% of the variance** ( $\text{Adj R}^2 = 0.889\$$ ) using only three variables. Adding more variables (like Age or Cards) increased complexity without improving accuracy.
- **Independence Check (VIF):** We ran a Multicollinearity test to ensure our predictors aren't redundant.
- **Threshold:** A VIF  $> 5$  is dangerous.
- **Our Finding:** All VIF values are **below 2.5**. This proves that Income, Limit, and Student capture *different* aspects of a customer's financial profile.
- **Diagnostic Verification:** Since the model selection strategy identified the same final model as our diagnostic analysis, additional model validation is not required.

```
Variance Inflation Factors (VIF):  
Intercept: 880.359  
C(Student)[T.Yes]: 1.070  
log_Income: 2.163  
log_Limit: 2.245
```

# Model Interpretation

- Model Fit:** Our chosen model explains **89.0%** of the variance in Credit Card Balance ( $R^2 = 0.890$ ). This indicates a strong fit even before adding interaction terms.
- Significance: The Parameter Estimates** show that Student, Income, and Limit are all highly significant ( $p < 0.001$ ).
- Variance Drivers: The ANOVA Table** confirms that **Credit Limit** is the dominant predictor (highest Sum of Squares), followed by Income and Student status.

| OLS Regression Results |                  |                     |           |       |           |           |
|------------------------|------------------|---------------------|-----------|-------|-----------|-----------|
| Dep. Variable:         | Balance          | R-squared:          | 0.890     |       |           |           |
| Model:                 | OLS              | Adj. R-squared:     | 0.889     |       |           |           |
| Method:                | Least Squares    | F-Statistic:        | 827.2     |       |           |           |
| Date:                  | Mon, 01 Dec 2025 | Prob (F-statistic): | 2.08e-146 |       |           |           |
| Time:                  | 23:07:43         | Log-Likelihood:     | -1964.9   |       |           |           |
| No. Observations:      | 310              | AIC:                | 3938.     |       |           |           |
| Df Residuals:          | 306              | BIC:                | 3953.     |       |           |           |
| Df Model:              | 3                |                     |           |       |           |           |
| Covariance Type:       | nonrobust        |                     |           |       |           |           |
|                        | coef             | std err             | t         | p> t  | [0.025    | 0.975]    |
| Intercept              | -1.033e+04       | 232.231             | -44.498   | 0.000 | -1.08e+04 | -9876.898 |
| C(Student)[T.Yes]      | 549.6762         | 24.418              | 22.511    | 0.000 | 501.628   | 597.725   |
| log_Income             | -317.4467        | 16.107              | -19.709   | 0.000 | -349.141  | -285.752  |
| log_Limit              | 1415.9170        | 31.718              | 44.640    | 0.000 | 1353.504  | 1478.330  |
| Omnibus:               | 76.539           | Durbin-Watson:      | 2.034     |       |           |           |
| Prob(Omnibus):         | 0.000            | Jarque-Bera (JB):   | 353.270   |       |           |           |
| Skew:                  | 0.939            | Prob(JB):           | 1.94e-77  |       |           |           |
| Kurtosis:              | 7.881            | Cond. No.           | 281.      |       |           |           |

ANOVA Table

|            | df    | sum_sq       | mean_sq      | F           | PR(>F)        |
|------------|-------|--------------|--------------|-------------|---------------|
| C(Student) | 1.0   | 2.325665e+06 | 2.325665e+06 | 122.463021  | 3.629611e-24  |
| log_Income | 1.0   | 6.955746e+06 | 6.955746e+06 | 366.270117  | 3.094577e-54  |
| log_Limit  | 1.0   | 3.784423e+07 | 3.784423e+07 | 1992.771175 | 4.966576e-136 |
| Residual   | 306.0 | 5.811171e+06 | 1.899075e+04 | NaN         | NaN           |

# Interpretation of Coefficients

- **Interpretation Method:** Since our predictors are log-transformed and Response variable is normal, we interpret the coefficients as the dollar change in Balance for a **1% change** in the predictor (beta times 0.01).

```
==== Percentage Change Effects ====
Effect of 1% increase in Income: -3.1745 change in Balance
Effect of 1% increase in Limit (Non-students): 14.1592 change in Balance
```

- **Income Effect:** Holding other factors constant, a **1% increase in Income** is associated with a **~\$3.17 decrease** in Balance. (Higher earners tend to carry slightly less debt).
- **Limit Effect:** A **1% increase in Credit Limit** is associated with a **~\$14.16 increase** in Balance. (Access to more credit strongly drives utilization).

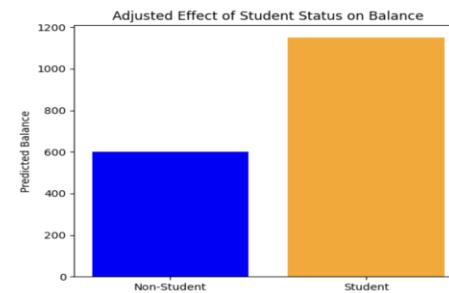
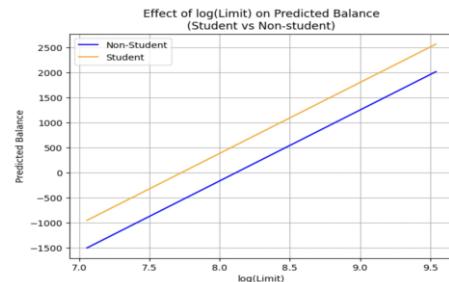
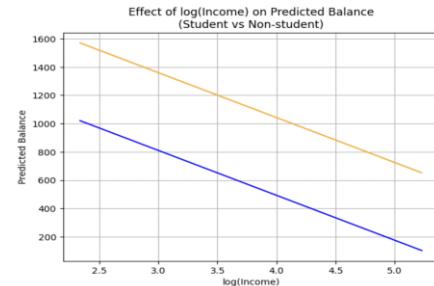
```
...
==== Pairwise Comparison for Student (Tukey HSD) ====
Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====
group1 group2 meandiff p-adj    lower      upper   reject
No     Yes  261.1785  0.0002  124.5726  397.7845   True
=====
```

```
==== Predicted Balance at Average Income/Limit ====
Non-Student: $615.25
Student:     $1164.93
Difference (Student - Non-Student): $549.68
```

- **Student Effect:** For the Student categorical variable, both pairwise comparison and model-based prediction were performed. The Tukey HSD test shows that non-students have a higher **raw average** balance than students. However, after adjusting for Income and Credit Limit in the regression model, students are predicted to have a higher balance at the same average Income and Limit. This difference occurs because Income and Limit act as confounding variables; once controlled, student status shows a positive association with balance.

# Effect Plot

- **log(Limit):** There is a strong **positive** relationship. As the credit limit increases (on a log scale), the predicted balance **rises**. This indicates that a higher capacity to borrow is highly correlated with a higher amount borrowed.
- **log(Income):** There is a **negative** relationship. As income increases (on a log scale), the predicted balance **falls**. This suggests that people with higher incomes tend to carry less credit card debt.
- Being a **Student** has the largest independent effect, leading to a significantly **higher** predicted credit card balance compared to non-students.
  - At average levels of income and credit limit, a student's predicted balance (~\$1150\$) is nearly **double** that of a non-student (~\$600\$).
- **Lack of Interaction:**
  - The parallel lines in the log(Income) and log(Limit) plots mean that the effect (the slope) of these variables is **identical** for both students and non-students.
  - In this specific model, student status only shifts the overall predicted balance up by a constant amount; it **does not** change how income or limit affect the balance.



# Final Model Selection & Justification

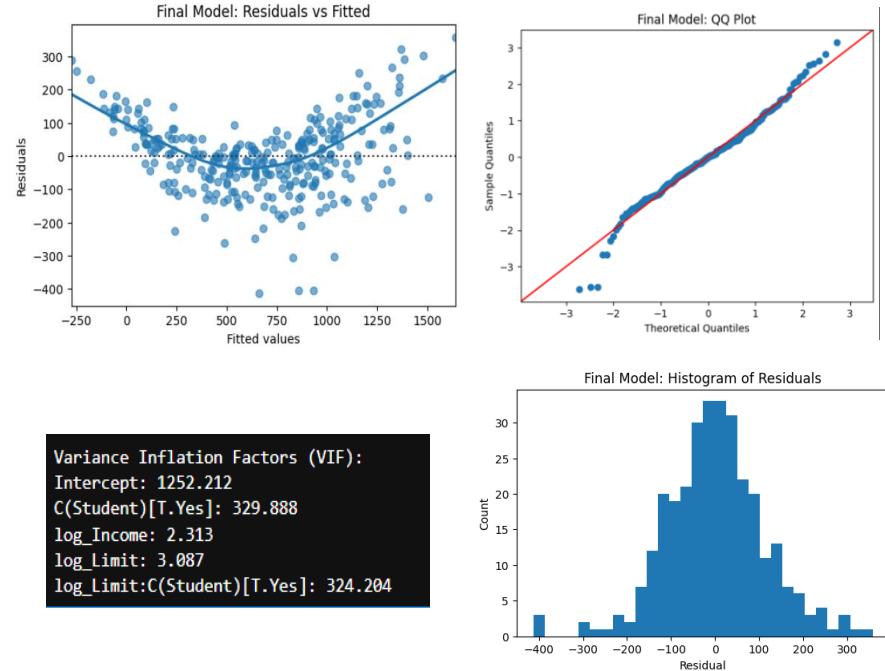
- Between the models obtained in parts (c) and (e), the model from part (e) (the **final interaction model**) is selected as the preferred model.
- Although the model in part (c) already provides a strong fit  $R^2=0.890$ , the model in part (e) shows a **clear improvement in explanatory power and predictive performance**, with a higher  $R^2=0.924$ , lower BIC, and the **smallest cross-validated RMSE**. In addition, the interaction between **Student status and Credit Limit** is **highly significant**, indicating that the effect of Credit Limit on Balance differs for students and non-students. This important behavioral difference is **not captured** in the model from part (c).
- Furthermore, model diagnostics show that **normality improves** in the final model, while heteroscedasticity remains similar in both models. Therefore, the added interaction improves interpretability and prediction without introducing unacceptable modeling issues.
- Final Model:

$$\text{Balance} \sim \log(\text{Income}) + \log(\text{Limit}) + C(\text{Student}) + \log(\text{Limit}):C(\text{Student})$$

# Final Model Diagnostics & Multicollinearity

## Assumptions:

- **Linearity:** No improvement in linearity, The Residuals vs Fitted plot shows a curved pattern, indicating that the model does not capture the relationship between the predictors and the response perfectly.
- **Normality:** Improvement, The normality assumption is **approximately met**. The QQ plot shows that most residuals fall close to the 45° line, with only mild deviations in the tails.
- **Homoscedasticity:** No Improvement, The constant variance assumption is **violated**. Both the Breusch–Pagan and White tests are highly significant, confirming strong heteroscedasticity in the residuals.
- Multicollinearity: Although the VIF values for the Student variable and its interaction with log(Limit) are high, this is **expected because an interaction term is included in the model**. Since interaction terms are naturally correlated with their main effects, the presence of higher VIFs does **not indicate a serious problem** in this context. Therefore, **multicollinearity is considered acceptable for this model**.



Variance Inflation Factors (VIF):  
Intercept: 1252.212  
C(Student)[T.Yes]: 329.888  
log\_Income: 2.313  
log\_Limit: 3.087  
log\_Limit:C(Student)[T.Yes]: 324.204

Final Model: Breusch-Pagan p-value = 8.352e-09  
Final Model: White test p-value = 2.295e-20

```
== Model WITHOUT influential point(s) ==
```

### OLS Regression Results

```
=====
Dep. Variable: Balance R-squared:          0.947
Model:            OLS Adj. R-squared:      0.946
Method:           Least Squares F-statistic:   1245.
Date:             Tue, 02 Dec 2025 Prob (F-statistic): 4.53e-176
Time:              22:54:33 Log-Likelihood:     -1657.8
No. Observations: 283 AIC:                  3326.
Df Residuals:      278 BIC:                  3344.
Df Model:          4
Covariance Type:  nonrobust
=====
```

|  | coef | std err | t | P> t | [0.025 | 0.975] |
|--|------|---------|---|------|--------|--------|
|--|------|---------|---|------|--------|--------|

|                             |           |         |         |       |           |           |
|-----------------------------|-----------|---------|---------|-------|-----------|-----------|
| Intercept                   | -1.15e+04 | 184.808 | -62.232 | 0.000 | -1.19e+04 | -1.11e+04 |
| C(Student)[T.Yes]           | 3765.8346 | 442.255 | 8.515   | 0.000 | 2895.241  | 4636.428  |
| log_Income                  | -342.8711 | 11.012  | -31.136 | 0.000 | -364.549  | -321.194  |
| log_Limit                   | 1563.1603 | 24.698  | 63.292  | 0.000 | 1514.542  | 1611.778  |
| log_Limit:C(Student)[T.Yes] | -387.9395 | 52.650  | -7.368  | 0.000 | -491.583  | -284.295  |

|                |        |                   |       |
|----------------|--------|-------------------|-------|
| Omnibus:       | 0.092  | Durbin-Watson:    | 2.054 |
| Prob(Omnibus): | 0.955  | Jarque-Bera (JB): | 0.013 |
| Skew:          | -0.011 | Prob(JB):         | 0.994 |
| Kurtosis:      | 3.025  | Cond. No.         | 832.  |

### Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

```
== Prediction Intervals for Influential Points ==
obs_ci_lower obs_ci_upper mean mean_ci_lower mean_ci_upper
0 959.927196 1422.428965 1191.178081 1142.038450 1240.317711
1 1344.940278 1804.201344 1574.570811 1533.730719 1615.410902
2 603.281248 1058.348854 830.815051 804.181181 857.448920
3 1248.633199 1706.511745 1477.572472 1440.817577 1514.327367
4 1088.842291 1553.496274 1321.169282 1267.191858 1375.146707
5 261.264779 716.264816 488.764797 462.421111 515.108484
6 631.745128 1087.189547 859.467338 831.269179 887.665496
7 1132.490313 1588.633032 1360.516673 1329.294808 1391.738537
8 1034.637263 1495.082394 1264.859828 1220.812579 1308.907087
9 1132.556482 1589.614126 1360.798904 1328.668938 1392.928870
10 1130.344383 1602.327138 1366.335721 1298.293964 1434.374777
11 968.990048 1431.340184 1200.165116 1151.383526 1248.946705
12 1144.720855 1607.557012 1376.148933 1326.182074 1426.115793
13 807.731726 1263.450385 1035.591056 1006.306156 1064.875955
14 54.600463 531.296514 292.948489 217.136136 368.760841
15 1126.396774 1587.670879 1357.038272 1310.869103 1403.198550
16 -405.838567 50.728405 -177.555081 -209.975469 -145.134693
17 -516.694743 -27.165944 -271.887484 -365.837384 -177.937583
18 -283.544575 195.386644 -44.078966 -123.334941 -35.177010
19 966.832942 1439.714129 1203.273536 1133.689876 1272.857196
20 14.187333 469.815841 242.001587 213.069496 270.933677
21 705.671099 1161.535978 933.693538 903.755112 963.451965
22 434.512770 889.614684 662.063387 635.286257 688.840517
23 1158.052385 1615.863793 1386.558089 1352.608308 1420.587878
24 -495.797959 -3.823834 -249.810892 -347.008445 -152.613339
25 1410.870663 1870.651387 1640.761025 1598.484437 1683.037614
26 727.543705 1190.676034 959.098669 908.507299 1009.712439
```

```
== Prediction Intervals for Influential Point(s) WITHOUT outlier ==
obs_ci_lower obs_ci_upper mean mean_ci_lower mean_ci_upper
0 999.996789 1354.554899 1177.275844 1121.297558 1233.254129
1 1390.834385 1734.159446 1562.496196 1528.237181 1596.756658
2 666.041889 1005.399575 835.720732 813.435875 858.005588
3 1295.645253 1637.666696 1466.655974 1435.827870 1497.484879
4 1133.422120 1493.468367 1313.445244 1249.304540 1377.585948
5 331.474968 670.616006 501.047377 479.621049 522.474425
6 694.425661 1034.144518 864.285090 840.664539 887.905640
7 1180.117500 1520.559551 1350.338525 1324.244417 1376.432633
8 1059.590088 1408.748731 1234.163969 1187.439771 1280.899902
9 1181.066078 1521.770087 1351.418083 1324.482779 1378.353386
10 1194.180449 1571.142544 1382.661496 1297.627325 1467.695667
11 1008.166161 1362.409935 1185.288048 1129.809512 1240.766584
12 1131.252403 1478.010116 1304.631260 1262.608256 1346.654263
13 866.030789 1206.054614 1036.042702 1011.349475 1060.735928
14 -35.756884 353.996772 159.075394 60.763026 257.387762
15 1150.049223 1500.442610 1325.245916 1276.260342 1374.231490
16 -327.490121 12.753438 -157.367892 -182.889551 -131.926232
17 -622.674601 -204.032687 -413.353644 -537.936822 -288.770466
18 -367.270455 27.625384 -169.822535 -273.221693 -66.423377
19 1035.518624 1412.121239 1223.819932 1139.184987 1308.454956
20 89.567641 429.136639 259.352140 236.276586 282.427695
21 767.090381 1107.226196 937.158289 912.082442 962.234135
22 501.347732 840.664994 671.006363 648.875931 693.136795
23 1286.852839 1548.071160 1377.462000 1348.945499 1405.978501
24 -611.706243 -186.510077 -399.108160 -529.122198 -269.094123
25 1454.951535 1798.734251 1626.842893 1591.454422 1662.231364
26 772.237429 1125.885856 949.061642 894.540935 1003.582349
```