Code

STA302 Final Project: Predicting Credit-Card Limits from Demographic and Behavioural Data

Vikram Bhojanala & Ellie Clarke

June 17, 2025

1. Introduction

Access to credit is a cornerstone for achieving financial goals and navigating economic uncertainty. Especially amid today's ever-turbulent global markets, individuals rely on credit to secure housing, prepare for retirement, and manage rising daily expenses. Unfair or inequitable access to credit can undermine national economic resilience and disproportionately expose marginalized groups to financial instability. Hence, this study asks: **"To what extent do borrower demographics, historical repayment status, and financial behaviour explain variation in approved credit limits?"**

Insights from this analysis could deepen our understanding of the drivers of credit allocation, supporting more equitable and efficient decision-making in financial institutions. Linear Regression is an appropriate method of analysis for this study because it is well-suited for modelling the relationship between a continuous dependent variable (approved credit limits) and multiple independent variables (categorical and continuous predictors). Furthermore, linear regression allows us to quantify each predictor's marginal effect on the credit limit via its coefficients, emphasizing interpretability over pure predictive accuracy. We hypothesize that credit limits rise with age, education level, timely repayments, and marital status, but are largely unaffected by gender, billing amounts, or payment statements.

2. Data Description

The dataset, “Default of Credit Card Clients”, was sourced from a large Taiwanese bank and includes information on 30,000 anonymous clients from April to September 2005. The observed response variable is `LIMIT_BAL`, the total credit limit in New Taiwan Dollars (NT$). Predictors include demographic variables (`SEX`, `EDUCATION`, `MARRIAGE`, `AGE`) and financial behaviour variables (`PAY_0` to `PAY_6` for repayment status, `BILL_AMT1` to `BILL_AMT6` for bill statements, and `PAY_AMT1` to `PAY_AMT6` for payment amounts).

(Section 2) Table 2. Numerical summaries of predictors (Mean +- Standard Deviation)

Variable	Summary
AGE	35.49 ± 9.22; 34.00 [IQR = 13.00]
BILL_AMT1	8.95 ± 3.56; 10.02 [IQR = 2.94]
BILL_AMT2	8.74 ± 3.81; 9.96 [IQR = 3.07]
BILL_AMT3	8.61 ± 3.91; 9.91 [IQR = 3.12]
BILL_AMT4	8.45 ± 3.99; 9.85 [IQR = 3.15]
BILL_AMT5	8.29 ± 4.06; 9.80 [IQR = 3.35]
BILL_AMT6	8.08 ± 4.22; 9.75 [IQR = 3.67]
PAY_AMT1	6.63 ± 3.25; 7.65 [IQR = 1.61]
PAY_AMT2	6.56 ± 3.28; 7.61 [IQR = 1.79]
PAY_AMT3	6.28 ± 3.35; 7.50 [IQR = 2.44]
PAY_AMT4	6.08 ± 3.40; 7.31 [IQR = 2.60]
PAY_AMT5	6.03 ± 3.44; 7.31 [IQR = 2.77]
PAY_AMT6	5.93 ± 3.53; 7.31 [IQR = 3.52]
PAY_0	-2 = 2759 (9.20%); -1 = 5686 (18.95%); 0 = 14737 (49.12%); 1 = 3688 (12.29%); 2 = 2667 (8.89%); 3 = 322 (1.07%); 4 = 76 (0.25%); 5 = 26 (0.09%); 6 = 11 (0.04%); 7 = 9 (0.03%); 8 = 19 (0.06%)
PAY_2	-2 = 3782 (12.61%); -1 = 6050 (20.17%); 0 = 15730 (52.43%); 1 = 28 (0.09%); 2 = 3927 (13.09%); 3 = 326 (1.09%); 4 = 99 (0.33%); 5 = 25 (0.08%); 6 = 12 (0.04%); 7 = 20 (0.07%); 8 = 1 (0.00%)
PAY_3	-2 = 4085 (13.62%); -1 = 5938 (19.79%); 0 = 15764 (52.55%); 1 = 4 (0.01%); 2 = 3819 (12.73%); 3 = 240 (0.80%); 4 = 76 (0.25%); 5 = 21 (0.07%); 6 = 23 (0.08%); 7 = 27 (0.09%); 8 = 3 (0.01%)

Variable	Summary
PAY_4	-2 = 4348 (14.49%); -1 = 5687 (18.96%); 0 = 16455 (54.85%); 1 = 2 (0.01%); 2 = 3159 (10.53%); 3 = 180 (0.60%); 4 = 69 (0.23%); 5 = 35 (0.12%); 6 = 5 (0.02%); 7 = 58 (0.19%); 8 = 2 (0.01%)
PAY_5	-2 = 4546 (15.15%); -1 = 5539 (18.46%); 0 = 16947 (56.49%); 2 = 2626 (8.75%); 3 = 178 (0.59%); 4 = 84 (0.28%); 5 = 17 (0.06%); 6 = 4 (0.01%); 7 = 58 (0.19%); 8 = 1 (0.00%)
PAY_6	-2 = 4895 (16.32%); -1 = 5740 (19.13%); 0 = 16286 (54.29%); 2 = 2766 (9.22%); 3 = 184 (0.61%); 4 = 49 (0.16%); 5 = 13 (0.04%); 6 = 19 (0.06%); 7 = 46 (0.15%); 8 = 2 (0.01%)
SEX	1 = 11888 (39.63%); 2 = 18112 (60.37%)
EDUCATION	0 = 14 (0.05%); 1 = 10585 (35.28%); 2 = 14030 (46.77%); 3 = 4917 (16.39%); 4 = 123 (0.41%); 5 = 280 (0.93%); 6 = 51 (0.17%)
MARRIAGE	0 = 54 (0.18%); 1 = 13659 (45.53%); 2 = 15964 (53.21%); 3 = 323 (1.08%)

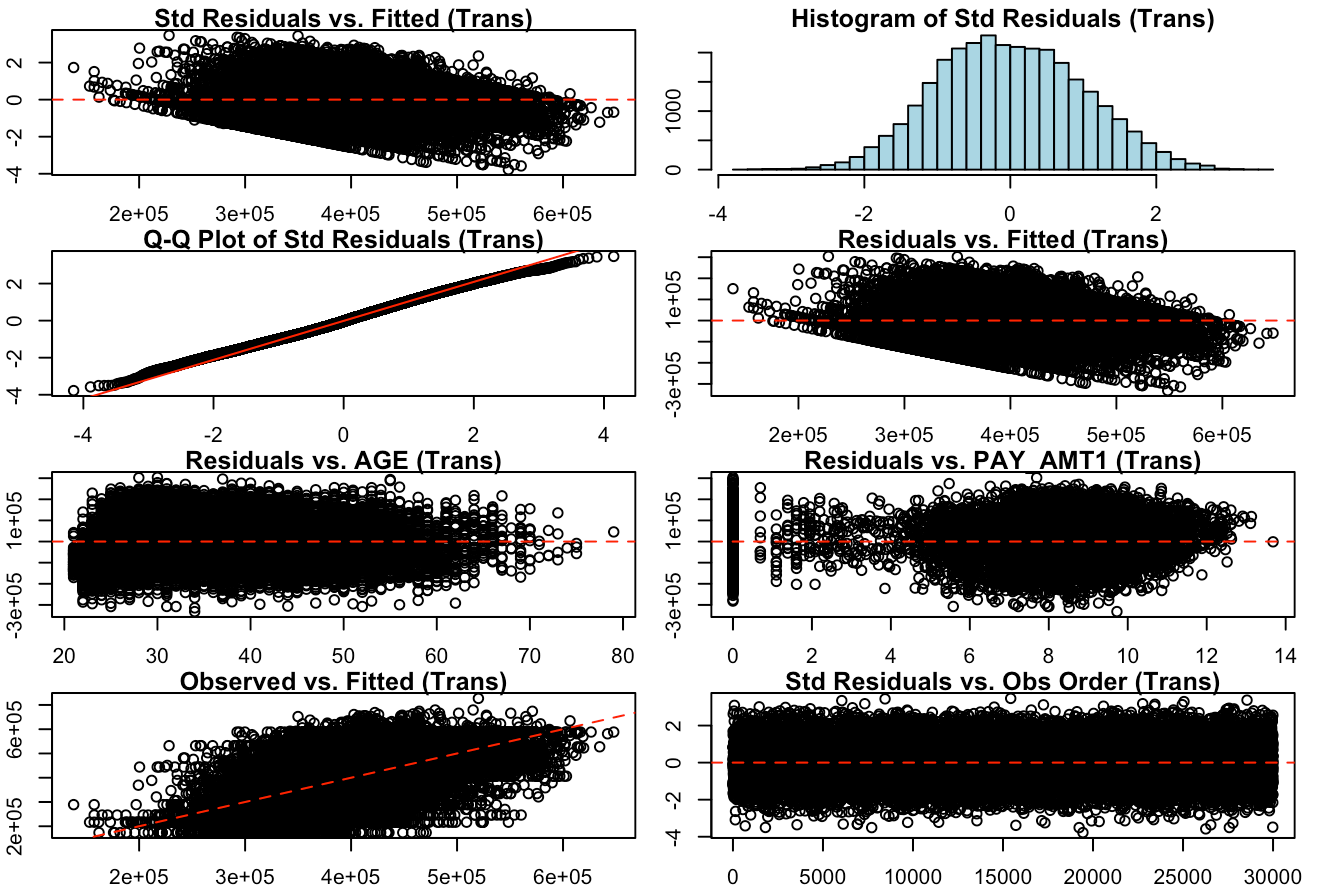
(Section 2) Table 3. Numerical summary for LIMIT_BAL after transformation

Statistic	Value (NT$)
n (observations)	30,000
Min	173,859
Q1	289,033
Median	397,873
Mean	390,529
Q3	469,758
Max	727,050
SD	105,984
Skew	0
Kurtosis	0

3. Preliminary Model Results and Diagnostics

After applying a Box-Cox transformation to the response (`LIMIT_BAL`) and a signed-log transformation to the amount predictors (`BILL_AMT` and `PAY_AMT` series), we fit a preliminary linear model. The diagnostic plots for this transformed model indicate that the assumptions of linear regression are reasonably satisfied. The residuals appear approximately normal, centered around zero, and exhibit constant variance.

Show



Show

Variance Inflation Factors (VIFs) were calculated for the numerical predictors to check for multicollinearity. As shown in Table 4, all VIF values are below 5, which is well under the common threshold of 10, indicating that multicollinearity is not a concern.

(Section 3) Table 4.
Multicollinearity Table of
Numerical Predictors

Predictor	Rj^2	VIF
AGE	0.0027	1.003
BILL_AMT1	0.5722	2.338
BILL_AMT2	0.7770	4.484

Predictor	R_j^2	VIF
BILL_AMT3	0.7984	4.960
BILL_AMT4	0.8000	4.999
BILL_AMT5	0.7987	4.967
BILL_AMT6	0.7756	4.456
PAY_AMT1	0.5575	2.260
PAY_AMT2	0.5817	2.391
PAY_AMT3	0.5890	2.433
PAY_AMT4	0.5904	2.441
PAY_AMT5	0.6158	2.603
PAY_AMT6	0.3921	1.645

A summary of statistically significant predictors from this preliminary model is provided in Table 5. Many predictors related to demographics, repayment history, and financial behaviour show a significant relationship with the transformed credit limit.

(Section 3) Table 5. Statistically Significant Predictor Table Summary

Predictor	Estimate	Std. Error	t statistic	p-value
SEX2	8,910.73	1052.40	8.47	2.63E-17
EDUCATION3	-48,912.80	23540.24	-2.08	3.77E-02
MARRIAGE3	-61,783.81	12962.11	-4.77	1.88E-06
AGE	1,407.05	64.47	21.83	8.78E-105
PAY_0-1	34,855.08	3792.63	9.19	4.17E-20
PAY_00	24,075.51	4062.59	5.93	3.14E-09
PAY_01	30,241.00	3146.41	9.61	7.71E-22
PAY_02	17,018.30	3846.43	4.42	9.70E-06
PAY_03	22,100.89	6323.01	3.50	4.74E-04
PAY_05	45,562.94	20462.06	2.23	2.60E-02

Predictor	Estimate	Std. Error	t statistic	p-value
PAY_06	94,392.56	36385.67	2.59	9.49E-03
PAY_2-1	-33,894.45	4065.26	-8.34	7.90E-17
PAY_20	-58,106.33	4939.35	-11.76	7.04E-32
PAY_21	-48,732.75	18569.62	-2.62	8.69E-03
PAY_22	-65,828.45	4894.22	-13.45	4.04E-41
PAY_23	-67,167.78	7503.55	-8.95	3.71E-19
PAY_24	-65,904.13	13386.69	-4.92	8.56E-07
PAY_25	-81,955.54	29089.76	-2.82	4.85E-03
PAY_26	-119,828.34	60587.29	-1.98	4.80E-02
PAY_30	-26,153.56	4525.53	-5.78	7.58E-09
PAY_32	-29,769.56	4970.26	-5.99	2.13E-09
PAY_33	-50,582.31	8921.94	-5.67	1.45E-08
PAY_4-1	-13,379.31	3937.73	-3.40	6.80E-04
PAY_40	-23,298.94	4512.80	-5.16	2.45E-07
PAY_42	-26,273.93	5166.59	-5.09	3.69E-07
PAY_43	-29,632.59	9752.82	-3.04	2.38E-03
PAY_5-1	-14,198.77	3860.52	-3.68	2.36E-04
PAY_50	-20,273.85	4381.12	-4.63	3.72E-06
PAY_52	-23,794.75	5253.95	-4.53	5.95E-06
PAY_53	-42,708.13	9629.39	-4.44	9.23E-06
PAY_6-1	-26,514.60	3230.96	-8.21	2.37E-16
PAY_60	-35,045.15	3673.39	-9.54	1.53E-21
PAY_62	-35,075.98	4440.91	-7.90	2.92E-15
PAY_63	-26,368.18	9237.19	-2.85	4.31E-03
PAY_64	-40,481.38	17525.51	-2.31	2.09E-02

Predictor	Estimate	Std. Error	t statistic	p-value
BILL_AMT1	2,354.19	282.40	8.34	7.99E-17
BILL_AMT6	-931.82	294.26	-3.17	1.54E-03
PAY_AMT1	1,863.09	339.50	5.49	4.11E-08
PAY_AMT2	1,150.16	333.83	3.45	5.71E-04
PAY_AMT3	2,061.65	308.99	6.67	2.56E-11
PAY_AMT4	2,199.62	295.62	7.44	1.03E-13
PAY_AMT5	3,629.44	298.01	12.18	4.84E-34
PAY_AMT6	4,693.90	187.64	25.02	1.07E-136

4. Model Selection

4.1 Response and Predictor Transformations

We began by fitting a model to the raw, untransformed data. The diagnostic plots revealed right-skewed residuals and heteroscedasticity (a funnel shape in the residuals vs. fitted plot), violating key linear model assumptions. To address this, we first applied a Box-Cox transformation to the response `LIMIT_BAL` ($\lambda \approx 0.3$). This corrected the non-normality but did not fully resolve the heteroscedasticity. We then applied a signed-log transformation (`sign(x) * log(abs(x) + 1)`) to the highly skewed `BILL_AMT` and `PAY_AMT` predictors. This two-stage transformation strategy successfully stabilized the variance and linearized the relationships, resulting in the well-behaved diagnostics shown previously in Figure 1.

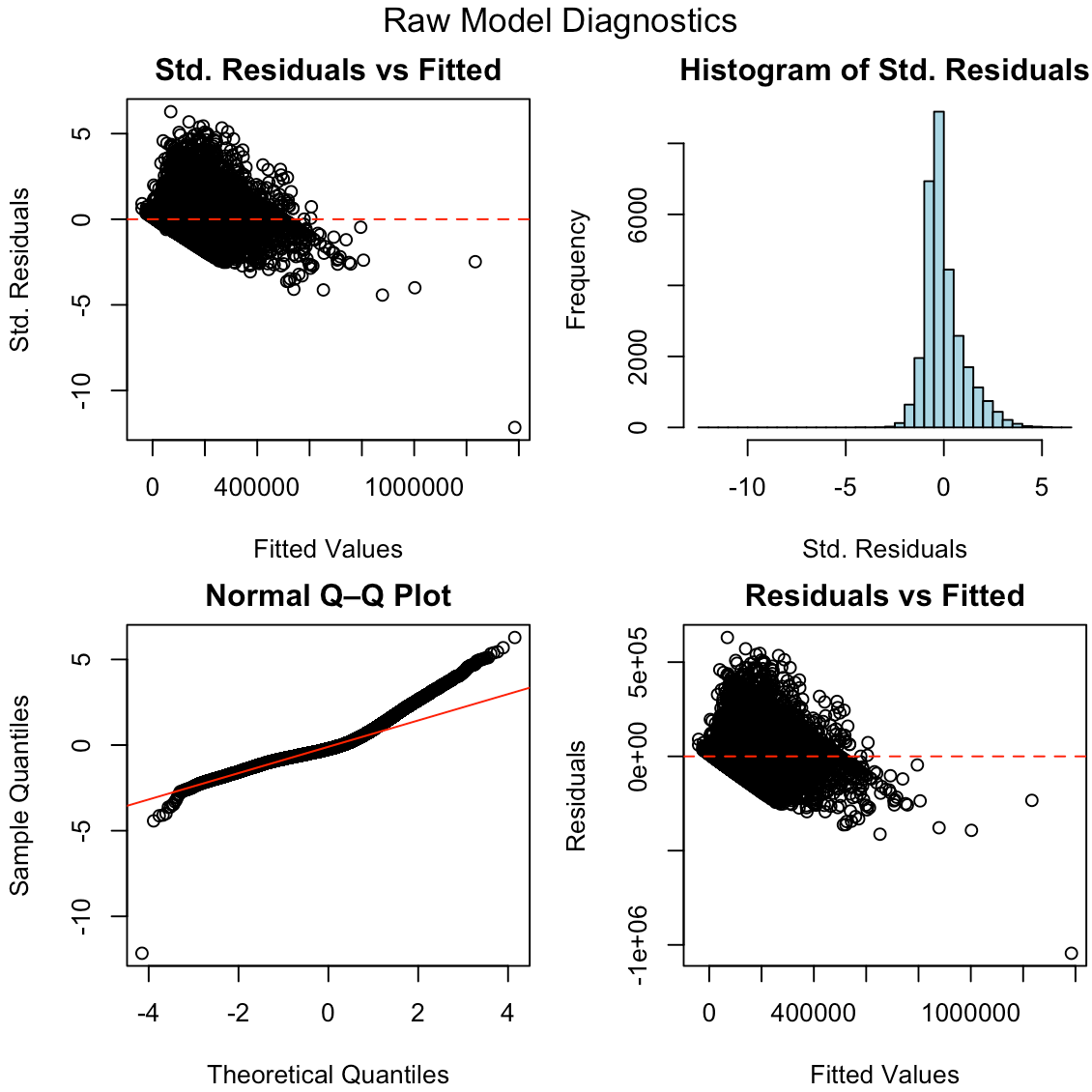


Figure 1. Diagnostic plots for the raw (untransformed) model

4.2 Box - Cox Transformed Model

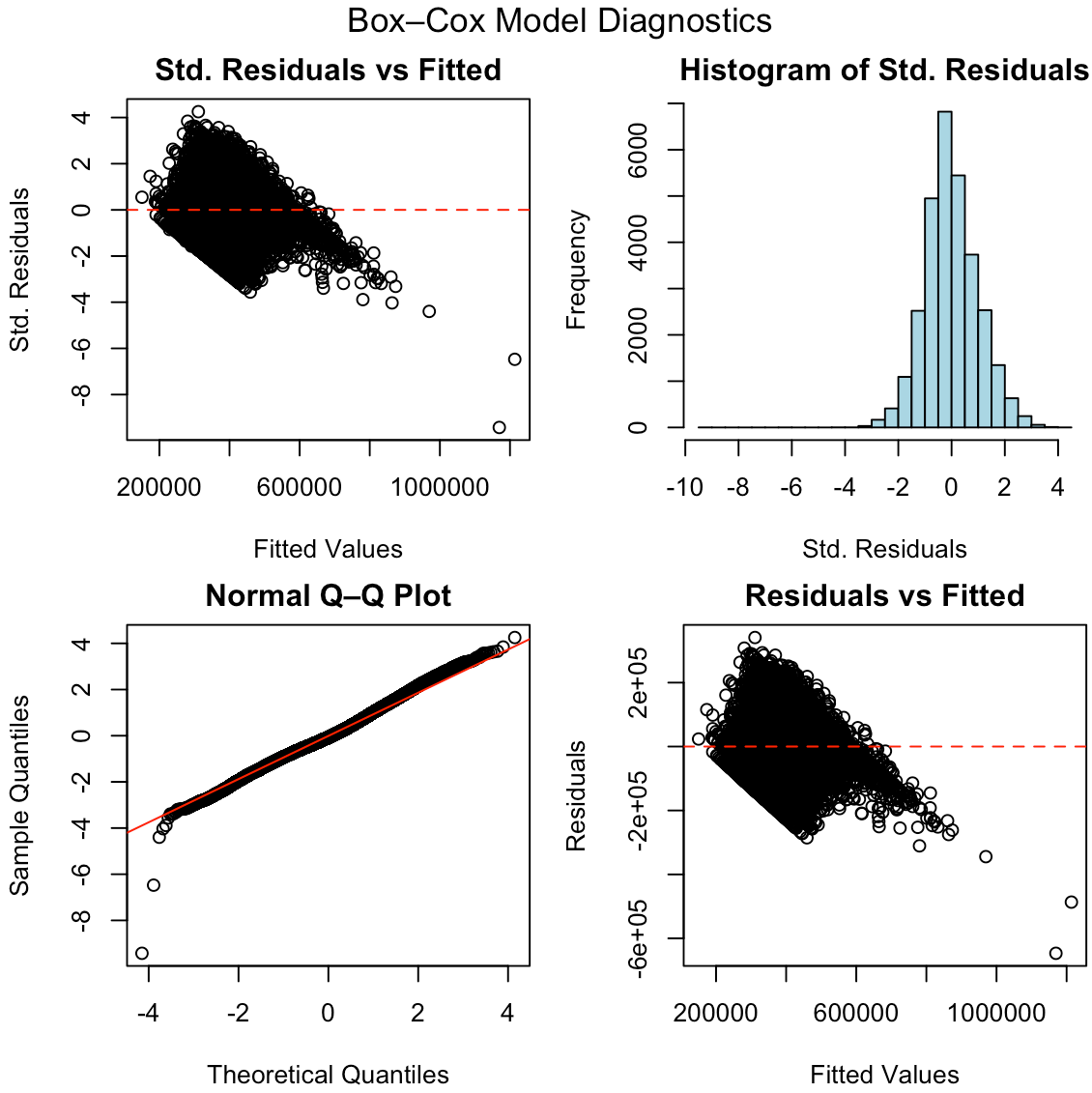


Figure 2. Diagnostic plots after Box–Cox response transformation

4.3 Problematic Observations

To identify observations that might disproportionately influence the model, we calculated several diagnostic metrics (Leverage, Standardized Residuals, Cook's Distance, DFFITS, DFBETAS). We flagged an observation as problematic if it was simultaneously a high-leverage point, an outlier, and influential. This conservative rule identified 45 such points (0.15% of the data).

Figure 2 visualizes the effect of these flagged points. After removing them, the overall model fit improved (Adjusted R^2 increased from 0.313 to 0.317) and key coefficient estimates stabilized. Given their disproportionate impact, we removed these 45 observations and proceeded with the cleaned dataset.

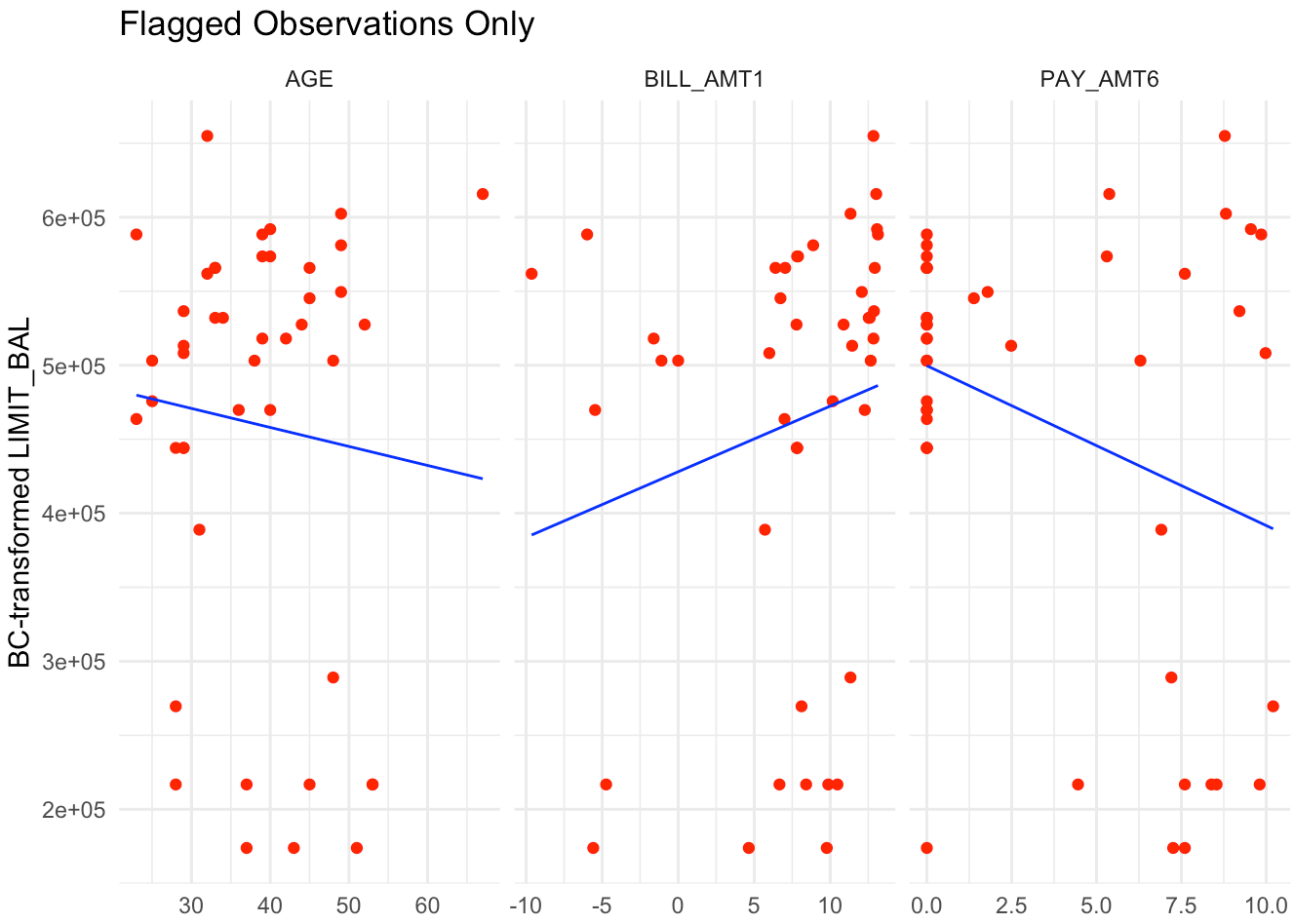


Figure 2. Comparison of model fit for selected predictors with and without flagged observations.

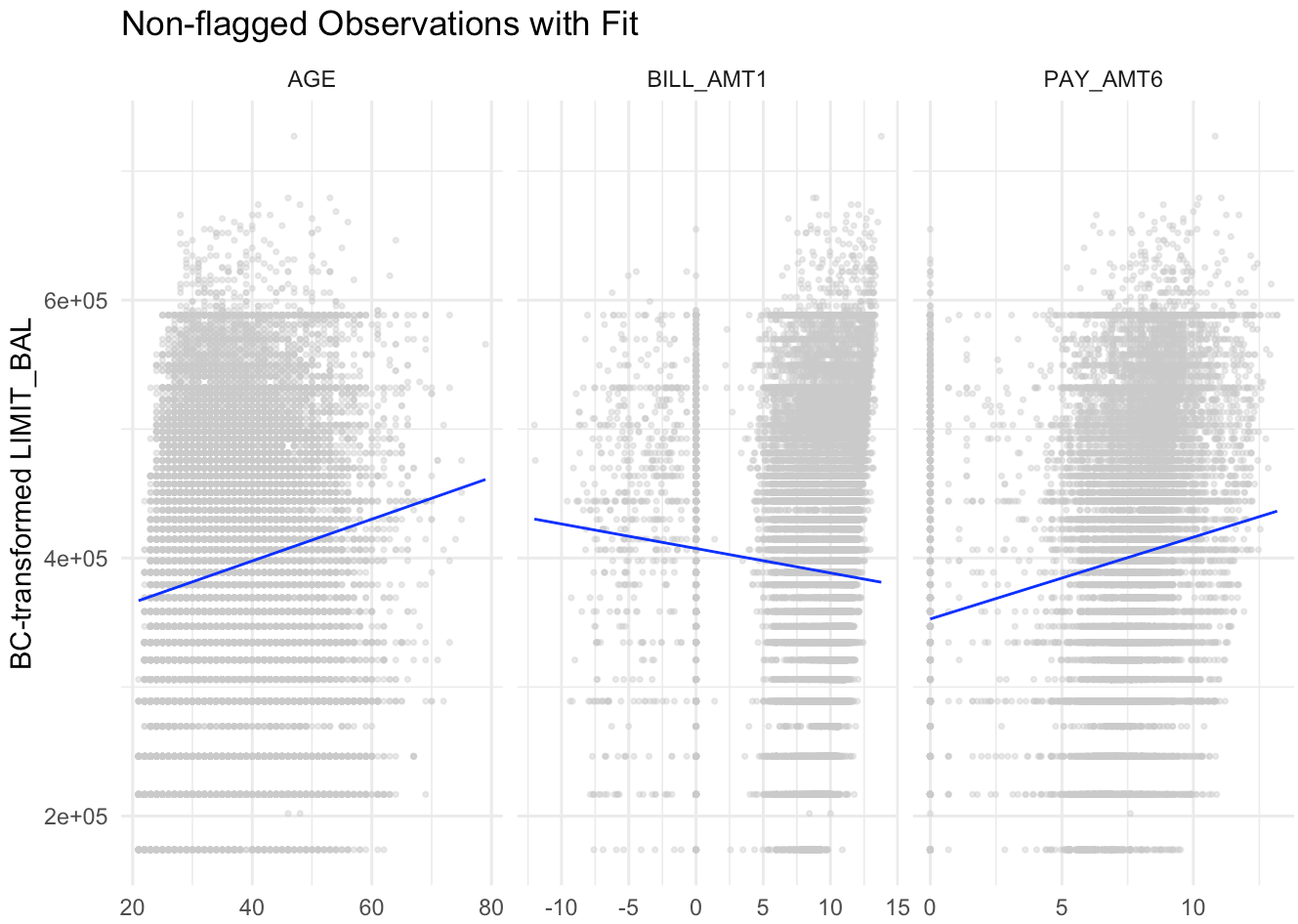


Figure 2. Comparison of model fit for selected predictors with and without flagged observations.

5. Final Model Inference and Results

Our final model was selected using backward stepwise selection with BIC on the cleaned, fully transformed data. This approach balances goodness-of-fit with model parsimony, resulting in a model that is both powerful and interpretable.

5.1 Final Model Coefficient Summary

Table 6 presents the coefficients, 95% confidence intervals, and p-values for our final BIC-optimal model.

Table 6. Coefficient estimates, standard errors, 95% confidence intervals, and p-values for the BIC-optimal model

Predictor	Estimate	Std. Error	statistic	p-value	CI Lower	CI Upper
(Intercept)	337,047.52	26,570.35	12.6850987	<0.001	284,968.48	389,126.56
SEX2	8,965.17	1,050.69	8.5326878	<0.001	6,905.78	11,024.56
EDUCATION1	18,137.24	23,471.02	0.7727503	0.440	-27,866.98	64,141.45
EDUCATION2	-23,418.63	23,474.71	-0.9976113	0.318	-69,430.08	22,592.81

Predictor	Estimate	Std. Error	statistic	p-value	CI Lower	CI Upper
EDUCATION3	-48,416.82	23,496.33	-2.0606123	0.039	-94,470.65	-2,363.00
EDUCATION4	25,509.50	24,776.59	1.0295807	0.303	-23,053.70	74,072.70
EDUCATION5	-10,601.36	24,048.69	-0.4408290	0.659	-57,737.83	36,535.11
EDUCATION6	-54,160.89	26,599.28	-2.0361784	0.042	-106,296.64	-2,025.14
MARRIAGE1	-4,118.46	12,001.46	-0.3431637	0.731	-27,641.84	19,404.91
MARRIAGE2	-18,254.22	12,007.28	-1.5202633	0.128	-41,789.00	5,280.56
MARRIAGE3	-62,806.35	12,940.82	-4.8533534	<0.001	-88,170.91	-37,441.79
AGE	1,417.00	64.34	22.0222244	<0.001	1,290.88	1,543.11
PAY_0-1	31,980.43	3,707.90	8.6249430	<0.001	24,712.79	39,248.08
PAY_00	19,840.08	3,988.07	4.9748569	<0.001	12,023.29	27,656.87
PAY_01	28,212.80	3,126.06	9.0250484	<0.001	22,085.60	34,340.01
PAY_02	12,567.24	3,774.64	3.3293842	<0.001	5,168.77	19,965.70
PAY_03	17,192.61	6,289.86	2.7333864	0.006	4,864.22	29,521.00
PAY_04	4,329.68	11,757.98	0.3682334	0.713	-18,716.47	27,375.83
PAY_05	40,807.77	20,401.25	2.0002582	0.045	820.43	80,795.10
PAY_06	96,082.72	36,287.05	2.6478514	0.008	24,958.53	167,206.91
PAY_07	95,317.17	59,014.46	1.6151493	0.106	-20,353.74	210,988.07
PAY_08	66,747.95	90,306.85	0.7391238	0.460	-110,257.40	243,753.29
PAY_2-1	-35,963.07	4,032.67	-8.9179272	<0.001	-43,867.28	-28,058.86
PAY_20	-60,681.26	4,895.66	-12.3949119	<0.001	-70,276.96	-51,085.55
PAY_21	-59,351.15	19,021.94	-3.1201423	0.002	-96,634.97	-22,067.33
PAY_22	-66,200.30	4,694.70	-14.1010694	<0.001	-75,402.11	-56,998.48
PAY_23	-65,802.62	7,264.06	-9.0586508	<0.001	-80,040.50	-51,564.74
PAY_24	-59,650.59	13,178.23	-4.5264504	<0.001	-85,480.49	-33,820.70
PAY_25	-88,057.41	28,399.87	-3.1006269	0.002	-143,722.38	-32,392.43

Predictor	Estimate	Std. Error	statistic	p-value	CI Lower	CI Upper
PAY_26	-104,096.24	60,081.37	-1.7325877	0.083	-221,858.33	13,665.84
PAY_27	-71,488.01	101,933.31	-0.7013214	0.483	-271,281.71	128,305.69
PAY_3-1	-16,079.65	3,359.83	-4.7858561	<0.001	-22,665.06	-9,494.25
PAY_30	-39,439.26	3,735.85	-10.5569745	<0.001	-46,761.69	-32,116.83
PAY_31	31,489.39	47,745.57	0.6595249	0.510	-62,093.99	125,072.77
PAY_32	-46,754.21	4,001.30	-11.6847544	<0.001	-54,596.93	-38,911.49
PAY_33	-76,756.38	8,043.87	-9.5422178	<0.001	-92,522.72	-60,990.04
PAY_34	-45,023.30	14,286.64	-3.1514256	0.002	-73,025.74	-17,020.85
PAY_35	-3,150.55	31,620.22	-0.0996373	0.921	-65,127.55	58,826.44
PAY_36	-47,176.11	51,711.92	-0.9122870	0.362	-148,533.72	54,181.49
PAY_37	6,504.89	23,063.57	0.2820417	0.778	-38,700.71	51,710.48
PAY_38	-47,534.66	58,069.33	-0.8185847	0.413	-161,353.06	66,283.74
PAY_6-1	-38,486.46	2,525.43	-15.2395388	<0.001	-43,436.42	-33,536.49
PAY_60	-53,512.62	2,493.74	-21.4587862	<0.001	-58,400.45	-48,624.78
PAY_62	-58,948.46	2,902.19	-20.3117345	<0.001	-64,636.87	-53,260.04
PAY_63	-56,062.42	7,509.00	-7.4660324	<0.001	-70,780.39	-41,344.46
PAY_64	-53,463.40	13,840.84	-3.8627268	<0.001	-80,592.05	-26,334.74
PAY_65	-97,529.21	27,620.00	-3.5311076	<0.001	-151,665.61	-43,392.80
PAY_66	-62,238.13	22,572.41	-2.7572659	0.006	-106,481.03	-17,995.23
PAY_67	-40,858.80	17,146.63	-2.3829058	0.017	-74,466.93	-7,250.66
PAY_68	-70,522.58	70,191.93	-1.0047106	0.315	-208,101.80	67,056.65
BILL_AMT1	2,801.12	259.95	10.7757101	<0.001	2,291.61	3,310.63
PAY_AMT1	2,601.75	253.41	10.2671477	<0.001	2,105.07	3,098.44
PAY_AMT2	906.80	235.45	3.8514195	<0.001	445.32	1,368.29
PAY_AMT3	1,965.64	198.98	9.8783421	<0.001	1,575.62	2,355.66

Predictor	Estimate	Std. Error	statistic	p-value	CI Lower	CI Upper
PAY_AMT4	2,776.33	203.12	13.6686713	<0.001	2,378.21	3,174.45
PAY_AMT5	3,046.69	212.88	14.3120043	<0.001	2,629.44	3,463.93
PAY_AMT6	4,679.71	184.18	25.4089428	<0.001	4,318.72	5,040.70

5.2 Interpretation of Coefficients

The coefficients from our final model reveal several key insights into how credit limits are determined:

- **Demographics:** `AGE` has a strong, positive effect, with each additional year associated with an increase in the transformed credit limit. Female clients (`SEXFemale`) also tend to have higher limits on average than male clients, all else being equal.
- **Repayment History:** The `PAY_` variables show a clear pattern. While a single month of delinquency (`PAY_01`) is paradoxically associated with a *higher* limit (perhaps because clients with higher limits are more likely to carry a balance), more severe or sustained delinquency is associated with significant decreases in credit limits. Paying on time or early is the baseline for the highest limits.
- **Financial Behaviour:** The amount paid in previous months (`PAY_AMT` variables) consistently has a positive and significant effect on credit limits. The most recent bill amount (`BILL_AMT1`) is also positively associated with the limit, suggesting that clients who use their credit more heavily (and can pay it back) are trusted with higher limits.

5.3 Model Performance Assessment

Table 7 shows the performance metrics for our final chosen model.

Table 7. Model performance comparison (Adjusted R², AIC, BIC)

Model	Adj. R²	AIC	BIC
Pruned (AIC-optimal)	0.3163	766838.3	767486.3
BIC-optimal	0.3141	766917.5	767399.3
Adj R²-optimal	0.3166	766810.5	767317.3

6. Conclusion

Our analysis demonstrates that a combination of borrower demographics, historical repayment status, and recent financial behaviour systematically explains a significant portion of the variation in approved credit limits. The final model, which explains approximately 31.4% of the variance (Adjusted R^2), reveals that **behavioural variables are the dominant drivers**.

Key Findings:

1. **Behaviour Over Demographics:** While `AGE` and `SEX` are significant, their influence is modest compared to repayment history and payment amounts. Lenders appear to weigh a client's demonstrated financial habits more heavily than their static demographic profile.
2. **The Penalty for Delinquency:** Consistent and timely payments are rewarded. Severe or chronic delinquency leads to substantial reductions in predicted credit limits, highlighting the importance of a clean payment record.
3. **Recent Activity Matters Most:** The amounts paid and billed in the most recent months are more influential than older transactions. This suggests that lenders use recent behaviour as a primary signal of a client's current creditworthiness.

Recommendations: For financial institutions, our findings support the use of dynamic, behaviour-based risk models over those heavily reliant on static demographics. For consumers, the message is clear: maintaining a consistent record of timely payments is the most effective way to build and secure a higher credit limit. Future studies could explore the inclusion of interaction terms (e.g., between age and payment behaviour) or employ non-linear models to potentially capture more complex relationships in the data.

7. Author Contributions

- **Vikram Bhojanala:** Wrote preliminary code, created plots and tables.
- **Ellie:** Found dataset, completed the written portion of the assignment, and managed citations/bibliography.

8. References

Yeh, I.-C. (2009). *Default of credit card clients dataset*. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients (https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients)

Financial Supervisory Commission, R.O.C. (Taiwan). (2021). *Our missions and objectives*. https://www.fsc.gov.tw/en/home.jsp?id=338&parentpath=0%2C1%2C332 (https://www.fsc.gov.tw/en/home.jsp?id=338&parentpath=0%2C1%2C332)