# MULTIPLE REGRESSION MODELLING of U.S. UNIVERSITY ADMISSION FACTORS

TEAM 4 FINAL PROJECT

WEN HSIN KO & ANUSHKA MONDAL

MAY 12, 2024

DS 853: APPLIED MULTIVARIATE ANALYSIS

PROFESSOR ROBERT SALTZMAN

LAM FAMILY COLLEGE OF BUSINESS

SAN FRANCISCO STATE UNIVERSITY

**Introduction**

This project aims to explore the important variables affecting applications for Master's programs in the United States and their impact on admission outcomes. Key factors in the application process include GRE, TOEFL, university rating, research experience, undergraduate GPA, statement of purpose, and strength of letters of recommendation. By collecting and analyzing these critical parameters, we seek to develop a model to predict the likelihood of an applicant being admitted to a specific university. The dataset used is hypothetical and designed to simulate real-world scenarios of applying for Master's programs. Through this analysis, we will investigate the relationships among these variables and their influence on final admission decisions. The objective of this project is to provide applicants with a reasonable prediction that helps them understand their potential competitiveness when applying to specific universities and provides valuable information and insights to applicants to aid them in making informed application decisions.

**Dataset Description and Expectations about the Predictors**

The data were collected from "graduate admission" and were inspired by the UCLA Graduate dataset, which is hypothetically created Source. The response variable, the chance of admission, is expressed from 0-1. There are seven explanatory variables: GRE score, TOEFL score, Research, Undergraduate GPA, University Rating, Statement of Purpose, and Letter of Recommendation. The selection of these variables is based on the admissions committee's use of this information to evaluate and make decisions regarding prospective candidates for admission.

The GRE score is often used as a standardized measure of a candidate's academic aptitude and readiness for graduate-level studies. A higher GRE score can indicate strong analytical and quantitative abilities, which are valued in many programs. For international students whose native language is not English, the TOEFL score is crucial. A higher TOEFL score demonstrates the ability to communicate effectively in English, which is necessary for coursework, research, and participation in academic discussions. A well-written statement of purpose provides insights into the applicant's academic and professional goals, motivations, and reasons for pursuing the specific program. It allows the admissions committee to assess the applicant's fit with the program and their potential contributions to the academic community. Strong letters of recommendation from professors, employers, or professionals who can attest to the applicant's academic abilities, character, and potential for success in graduate studies are highly influential. Thus, there will be a positive relationship from the above four points.

While a highly ranked university may initially seem advantageous, it's important to consider that admissions committees assess applicants based on their individual achievements, academic performance, and potential for success in the specific graduate program. Therefore, the university rating itself might not directly determine admission outcomes. Admissions decisions are more likely to be influenced by the applicant's qualifications, experiences, and fit for the program, rather than solely by the ranking of their undergraduate institution.

$$\widehat{Chance\ of\ Admit} = b_0 + b_1 X_{GRE\ Score} + b_2 X_{TOEFL\ Score} + b_3 X_{Univerity\ Rating} + b_4 X_{SOP} + b_5 X_{LOR} + b_6 X_{CGPA} + b_7 X_{Research\ Experience}$$ ------- **Equation [1]**

Equation [1] above is the generic form of the equation for this project, prior to doing any modeling, where *b*0 is the intercept and b1 – b7 are the coefficients of the respective variables. The table below summarizes our expectations about each predictor's possible impact on the chance of admission.

| **Predictor Variable** | **Expected Direction of Relation to Chance of admit** |
|---|---|
| GRE score | Positive relationship |
| TOEFL score | Positive relationship |
| University Rating | Negative relationship |
| Statement of purpose | Positive relationship |
| Letter of recommendation | Positive relationship |
| Undergraduate CGPA | Positive relationship |
| Research Experience | Positive relationship |

**Stepwise Regression Modeling**

We used stepwise regression to develop the best multiple regression model for predicting the chance of admission to graduate school. In our initial step, we created a correlation matrix for all variables. Among these, CGPA showed the highest correlation with the chance of admission, so we chose CGPA as the first predictor. We then conducted a simple linear regression of Chance of Admit against CGPA, resulting in a highly significant F value, confirming the statistical significance of the model at $\alpha = 5\%$.

In step 2, we proceeded with multiple regressions, pairing CGPA with each of the other six predictors to observe the impact of adding a second predictor. Among these, the GRE Score exhibited the lowest coefficient p-value and highest adjusted R-squared value, leading us to select it as our second predictor (X2). Moving to step 3, we simultaneously included CGPA and GRE Scores in the model, introducing a third predictor variable. Here, the Letter of Recommendation (LOR) showed the lowest coefficient p-value and highest adjusted R-squared, prompting its selection as our third predictor (X3).

For step 4, we expanded the model to include CGPA, GRE Score, LOR, and each of the remaining four predictors one at a time. Research experience yielded the lowest coefficient p-value and highest adjusted R-squared among these, becoming our fourth predictor (X4). \ In step 5, with CGPA, GRE Score, LOR, and

Research already in the model, we introduced the remaining predictors—TOEFL Score, University Rating, and Statement of Purpose (SOP). The TOEFL Score demonstrated the most significant coefficient p-value and highest adjusted R-squared, leading to its inclusion as the fifth predictor (X5).

By step 6, with CGPA, GRE Score, LOR, Research, and TOEFL Score included, the remaining two predictors did not meet the significance threshold, indicating their non-significance. Therefore, we concluded our stepwise regression analysis at this point. Our final regression equation was the following:

$$\widehat{Chance\ of\ Admit} = -1.34 + 0.002X_{GRE\ Score} + 0.03X_{TOEFL\ Score} + 0.02X_{LOR} + 0.12X_{CGPA} + 0.03X_{Research\ Experience}$$

**Results from Stepwise Regression**

| | |
|---|---|
| CGPA | Contrary to our initial thoughts, CGPA has a positive effect on a chance of admit. This means that for each increase of one unit in CGPA, the predicted chance of admission will increase by 0.12 units, *i.e.,* by 12 percentage points. The reason that admissions committees value CGPA might be that it can reflect a candidate's academic ability, motivation, and work ethic. |
| GRE score | GRE score aligns with our initial thoughts of positively impacting the chance of admission. For every one-point increase in GRE score, the chance of admit increases by 0.002 units (0.2%), assuming that the other variables are held constant. |
| LOR | LOR aligns with our initial thoughts of positively impacting the chance of admission. For every one-unit increase in LOR score, the chance of admit increases by 0.02 units (2%), assuming that the other variables are held constant. |
| Research | Contrary to our initial thoughts, Research has a positive effect on a chance of admission. This means that if the candidates have the experience of ~~R~~esearch, the predicted chance of admission will increase by 0.03 units (3%). The reason that admissions committees value Research might be that this experience enhances candidates' skills, attributes, and potential contributions to the academic community, making them more competitive and desirable applicants for admission to academic programs. |
| TOEFL score | TOEFL score aligns with our initial thoughts of positively impacting the chance of admission. For every one-point increase in TOEFL score, the chance of admit increases by 0.003 units (0.3%), assuming that the other variables are held constant. |

**Cross Validation**

After getting the 5 best predictors for the model through stepwise regression, cross-validation was used to validate the model i.e. to test its performance on the independent data. Step 1 consisted of splitting the available data modified through stepwise regression into parts:

1. Training (or model-building) set (n = 400).

2. Validation (or holdout, testing, prediction) set (n = 100).

We used a single 80:20 split. 400 rows of the data were assigned to the training set and 100 rows of the data were assigned to the validation set.

In step 2, we first focused on the training set. Chance of admit was the response variable and CGPA, GRE score, LOR, Research, and TOEFL score were the predictors In step 3, we focused on the validation set. With the same response and predictor variables, we performed the regression analysis for the validation set. In step 4, we calculated the Predicted Chance of Admit using the following formula:

***predicted chance of admit = -1.480 + (0.002 * GRE score) + (0.003 * TOEFL score) + (0.010 * LOR) + (0.126 * CGPA) + (0.025 * Research)*** --------- **equation [2]**

In step 5, we calculated the model error using the following formula: ***Model Error = (Chance of Admit) – (Predicted Chance of Admit)***

In step 6, we calculated the SSPE, MSPE, and RSPE using the Excel formulae.

In step 7, we finally created a comparison table to compare both the training and validation regression models. Our focus was to compare the validation model's MSPE with the training model's MSE. MSPE is based on the training model's predictions of all the validation data cases. We figured that MSE had a much larger value than MSPE. Our predictions on the validation data are much more accurate than they are for the training data, so we don't think we overfit the training data.

## Appendix

### A. Stepwise Regression Method Details

- We first calculated the correlations between the predictor variables and the response variable chance of admit. We found CGPA to be the predictor with the highest correlation (0.882) with a chance of admit.

| | Chance of Admi | GRE Score | TOEFL Score | niversity Ratin | SOP | LOR | CGPA | Research |
|---|---|---|---|---|---|---|---|---|
| Chance of Adı | 1 | | | | | | | |
| GRE Score | 0.810 | 1 | | | | | | |
| TOEFL Score | 0.792 | 0.827 | 1 | | | | | |
| University Rat | 0.690 | 0.635 | 0.650 | 1 | | | | |
| SOP | 0.684 | 0.613 | 0.644 | 0.728 | 1 | | | |
| LOR | 0.645 | 0.525 | 0.542 | 0.609 | 0.664 | 1 | | |
| CGPA | 0.882 | 0.826 | 0.811 | 0.705 | 0.712 | 0.637 | 1 | |
| Research | 0.546 | 0.563 | 0.467 | 0.427 | 0.408 | 0.373 | 0.501 | 1 |

- In Step 1, we ran a regression between chance of admit and CGPA and got almost zero for the coefficient's p-value (statistically significant at α = 5%) and 0.778 for the Adjusted R Square, so it seems to be a good predictor for chance of admit. Then CGPA became the first predictor of our model.

| SUMMARY OUTPUT | | Chance of admit vs CGPA | | | | |
|---|---|---|---|---|---|---|
| *Regression Statistics* | | | | | | |
| Multiple R | 0.882 | | | | | |
| R Square | 0.779 | | | | | |
| Adjusted R Sq | 0.778 | | | | | |
| Standard Erro | 0.066 | | | | | |
| Observations | 500 | | | | | |
| ANOVA | | | | | | |
| | *df* | *SS* | *MS* | *F* | *Significance F* | |
| Regression | 1 | 7.74010112 | 7.74010112 | 1751.850429 | 3.3965E-165 | |
| Residual | 498 | 2.20028508 | 0.00441824 | | | |
| Total | 499 | 9.9403862 | | | | |
| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| Intercept | -1.044 | 0.042 | -24.689 | 1.76142E-88 | -1.127 | -0.961 |
| CGPA | 0.206 | 0.005 | 41.855 | 3.3965E-165 | 0.196 | 0.216 |

- In Step 2, we included each of the six remaining predictors one by one in the model along with including and CGPA. Among these predictors, we found that the GRE Score had the lowest p-value (nearly zero) and was statistically significant at $\alpha = 5\%$. Also, its Adjusted R2 is higher than others, so we decided to include the GRE score as our second predictor.

| $\alpha$ = 0.05 | | | | | |
|---|---|---|---|---|---|
| X2 | T | P value | R-sq | Adj R2 | |
| GRE Score | 7.206 | 0.000000000002 | 0.800 | 0.799 | significant, select as X2 |
| TOEFL Score | 6.486 | 0.000000000213 | 0.796 | 0.795 | significant |
| University Rating | 4.629 | 0.000004701566 | 0.788 | 0.787 | significant |
| SOP | 3.816 | 0.000152826380 | 0.785 | 0.784 | significant |
| LOR | 5.234 | 0.000000244988 | 0.790 | 0.789 | significant |
| Research | 5.861 | 0.000000008385 | 0.793 | 0.792 | significant |

- In step 3, we repeated the previous steps and compared the GRE score, CGPA, and the other remaining five variables one by one. As a result of the picture below, we got letter of recommendation (LOR) as our third predictor.

| X3 | T | P value | R-sq | Adj R2 | |
|---|---|---|---|---|---|
| TOEFL Score | 3.573 | 0.00039 | 0.805 | 0.803 | significant |
| University Rating | 3.897 | 0.00011 | 0.806 | 0.804 | significant |
| SOP | 3.539 | 0.00044 | 0.805 | 0.803 | significant |
| LOR | 5.544 | 0.00000 | 0.811 | 0.810 | significant, select as X3 |
| Research | 3.999 | 0.00007 | 0.806 | 0.805 | significant |

- Repeated the previous steps. We continued to use stepwise regression in step 4. Currently, our model already has LOR, CGPA, GRE score. We found that Research would be our fourth predictor.

| X4 | T | P value | R-sq | Adj R2 | |
|---|---|---|---|---|---|
| TOEFL Score | 3.315 | 0.00098 | 0.8154 | 0.8139 | significant |
| University Rating | 2.444 | 0.01487 | 0.8135 | 0.8120 | significant |
| SOP | 1.572 | 0.11658 | 0.8135 | 0.8120 | nonsignificant |
| Research | 3.644 | 0.00030 | 0.8162 | 0.8147 | significant, select as X4 |

- So far, our model already has LOR, CGPA, GRE score, research. At step 5, we got a TOEFL score as our fifth predictor.

| X5 | T | P value | R-sq | Adj R2 | |
|---|---|---|---|---|---|
| TOEFL Score | 3.501 | 0.00051 | 0.821 | 0.819 | significant, select as X5 |
| University Rating | 2.249 | 0.02493 | 0.818 | 0.816 | significant |
| SOP | 1.453 | 0.14687 | 0.817 | 0.815 | nonsignificant |

- At step 6, our model already has LOR, CGPA, GRE score, research, TOEFL score, and we added the last two variables to the model separately. However, both of them are non-significant at $\alpha = 5\%$. Hence, step 5 would be our last meaningful step of the method.

| X6 | T | P value | R-sq | Adj R2 | |
|---|---|---|---|---|---|
| University Rating | 1.820 | 0.069 | 0.822 | 0.820 | nonsignificant |
| SOP | 0.991 | 0.322 | 0.821 | 0.819 | nonsignificant |
| | | | | | |
| | | | | | stop here |

- Below is the output from our final regression model:

SUMMARY OUTPUT - Chance of Admit vs CGPA,GRE Score,LOR ,Research & TOEFL Score

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.9059 |
| R Square | 0.8207 |
| Adjusted R Sq | 0.8188 |
| Standard Error | 0.0601 |
| Observations | 500 |

**y(hat) = -1.34+0.12x1+0.002x2+0.02x3+0.03x4+0.003x5**
y(hat) = chance of admit
x1=CGPA
x2 = GRE Score
x3 = LOR
x4 = Research
x5 = TOEFL score

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 5 | 8.15767788 | 1.631535576 | 452.1090555 | 9.9711E-182 |
| Residual | 494 | 1.78270832 | 0.003608721 | | |
| Total | 499 | 9.9403862 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
|---|---|---|---|---|---|---|---|---|
| Intercept | -1.336 | 0.0991 | -13.4817 | 0.0000 | -1.5304 | -1.1410 | -1.5304 | -1.1410 |
| CGPA | 0.123 | 0.0093 | 13.2210 | 0.0000 | 0.1047 | 0.1413 | 0.1047 | 0.1413 |
| GRE Score | 0.002 | 0.0005 | 3.7604 | 0.0002 | 0.0009 | 0.0029 | 0.0009 | 0.0029 |
| LOR | 0.019 | 0.0038 | 5.0924 | 0.0000 | 0.0119 | 0.0268 | 0.0119 | 0.0268 |
| Research | 0.025 | 0.0066 | 3.8135 | 0.0002 | 0.0122 | 0.0381 | 0.0122 | 0.0381 |
| TOEFL Score | 0.003 | 0.0009 | 3.5009 | 0.0005 | 0.0013 | 0.0047 | 0.0013 | 0.0047 |

**B. Cross-Validation Method details**

- We first started by splitting the data at 80:20. We had a total of 500 rows in our dataset. As per the split, 400 rows were assigned to the training set.

- 100 rows were assigned to the validation set as per the split.

- In step 2, we performed the linear regression analysis on the training set and got the following output. The adjusted r square value is 0.80 showing a moderately strong linear relationship. The significance F (p-value) is 2.0427E-136 which is extremely small and less than 0.05 making the training model statistically significant. An important value to be noted is the MSE which is 0.004063.

SUMMARY OUTPUT

*Regression Statistics*

| | |
|---|---|
| Multiple R | 0.895948 |
| R Square | 0.802722 |
| Adjusted F | 0.800219 |
| Standard I | 0.063742 |
| Observatic | 400 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regressior | 5 | 6.513795 | 1.302759 | 320.6369 | 2.0427E-136 |
| Residual | 394 | 1.600836 | 0.004063 | | |
| Total | 399 | 8.114631 | | | |

| | *Coefficients* | *andard Err* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *ower 95.0%* | *pper 95.0%* |
|---|---|---|---|---|---|---|---|---|
| Intercept | -1.29846 | 0.11729 | -11.0705 | 5.51E-25 | -1.52905714 | -1.06787 | -1.52906 | -1.06787 |
| GRE Score | 0.001782 | 0.000596 | 2.99227 | 0.002943 | 0.000611181 | 0.002953 | 0.000611 | 0.002953 |
| TOEFL Sco | 0.003032 | 0.001065 | 2.846564 | 0.00465 | 0.000937911 | 0.005126 | 0.000938 | 0.005126 |
| LOR | 0.022776 | 0.004804 | 4.741156 | 2.97E-06 | 0.013331644 | 0.032221 | 0.013332 | 0.032221 |
| CGPA | 0.121004 | 0.011735 | 10.3115 | 3.13E-22 | 0.097933404 | 0.144075 | 0.097933 | 0.144075 |
| Research | 0.024577 | 0.00792 | 3.103032 | 0.002054 | 0.00900559 | 0.040148 | 0.009006 | 0.040148 |

- In step 3, we performed linear regression analysis on the validation model. The adjusted r square has increased to 0.90 depicting a very strong linear relationship. The significance F (p-value) is also very small (1.72538E-46) making the validation model statistically significant.

SUMMARY OUTPUT

*Regression Statistics*

| | |
|---|---|
| Multiple R | 0.951 |
| R Square | 0.905 |
| Adjusted F | 0.900 |
| Standard I | 0.043 |
| Observatic | 100 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regressior | 5 | 1.640 | 0.328 | 179.598 | 1.73528E-46 |
| Residual | 94 | 0.172 | 0.002 | | |
| Total | 99 | 1.812 | | | |

| | *Coefficients* | *andard Err* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *ower 95.0%* | *pper 95.0%* |
|---|---|---|---|---|---|---|---|---|
| Intercept | -1.480 | 0.165 | -8.952 | 0.000 | -1.809 | -1.152 | -1.809 | -1.152 |
| GRE Score | 0.002 | 0.001 | 2.831 | 0.006 | 0.001 | 0.004 | 0.001 | 0.004 |
| TOEFL Sco | 0.003 | 0.001 | 2.636 | 0.010 | 0.001 | 0.006 | 0.001 | 0.006 |
| LOR | 0.010 | 0.005 | 1.918 | 0.058 | 0.000 | 0.020 | 0.000 | 0.020 |
| CGPA | 0.126 | 0.013 | 9.980 | 0.000 | 0.101 | 0.152 | 0.101 | 0.152 |
| Research | 0.025 | 0.010 | 2.476 | 0.015 | 0.005 | 0.046 | 0.005 | 0.046 |

- Then we got the equation for the predicted chance of admit as follows:

***predicted chance of admit = -1.480 + (0.002 * GRE score) + (0.003 * TOEFL score) + (0.010 * LOR) + (0.126 * CGPA) + (0.025 * Research)*** --------- **equation [2]**

- In step 4, we created the predicted chance of admit column for all the 100 rows in the validation set using the equation [2].

| Chance of Admit | GRE Score | TOEFL Score | LOR | CGPA | Research | Pred Chance of Admit |
|---|---|---|---|---|---|---|
| 0.63 | 304 | 100 | 3 | 8.22 | 0 | 0.609446502 |
| 0.66 | 315 | 105 | 3 | 8.34 | 0 | 0.658728912 |
| 0.78 | 324 | 109 | 3 | 8.94 | 1 | 0.784074251 |
| 0.91 | 330 | 116 | 3.5 | 9.23 | 1 | 0.862469332 |
| 0.62 | 311 | 101 | 2.5 | 7.64 | 1 | 0.567958891 |
| 0.52 | 302 | 99 | 3 | 7.45 | 0 | 0.509677285 |
| 0.61 | 322 | 103 | 2.5 | 8.02 | 1 | 0.639606523 |
| 0.58 | 298 | 100 | 4 | 7.95 | 1 | 0.61343638 |
| 0.57 | 297 | 101 | 4 | 7.67 | 1 | 0.580805149 |
| 0.61 | 300 | 98 | 2.5 | 8.02 | 0 | 0.560665609 |
| 0.54 | 301 | 96 | 4 | 7.56 | 0 | 0.534886053 |
| 0.56 | 313 | 94 | 1.5 | 8.13 | 0 | 0.562238193 |
| 0.59 | 314 | 102 | 2 | 7.88 | 1 | 0.593989796 |
| 0.49 | 317 | 101 | 2 | 7.94 | 1 | 0.603564123 |
| 0.72 | 321 | 110 | 4 | 8.35 | 1 | 0.733143884 |
| 0.76 | 327 | 106 | 4.5 | 8.75 | 1 | 0.791497895 |
| 0.65 | 315 | 104 | 2.5 | 8.1 | 0 | 0.615267851 |
| 0.52 | 316 | 103 | 2 | 7.68 | 0 | 0.551808045 |
| 0.6 | 309 | 111 | 4 | 8.03 | 0 | 0.651493477 |
| 0.58 | 308 | 102 | 3.5 | 7.98 | 1 | 0.629562434 |
| 0.42 | 299 | 100 | 3 | 7.42 | 0 | 0.503733084 |
| 0.77 | 321 | 112 | 4.5 | 8.95 | 1 | 0.823198416 |
| 0.73 | 322 | 112 | 2.5 | 9.02 | 1 | 0.787898344 |
| 0.94 | 334 | 119 | 5 | 9.54 | 1 | 0.950368833 |

data | training | validation | comparison | +

- In step 5, we calculated the model error using the following formula:

***Prediction Error = (Chance of Admit) – (Predicted Chance of Admit)***

ere we will use last 100 rows as the validation set

| Chance of Admit | GRE Score | TOEFL Score | LOR | CGPA | Research | Pred Chance of Admit | Model error |
|---|---|---|---|---|---|---|---|
| 0.63 | 304 | 100 | 3 | 8.22 | 0 | 0.609446502 | 0.020553498 |
| 0.66 | 315 | 105 | 3 | 8.34 | 0 | 0.658728912 | 0.001271088 |
| 0.78 | 324 | 109 | 3 | 8.94 | 1 | 0.784074251 | -0.00407425 |
| 0.91 | 330 | 116 | 3.5 | 9.23 | 1 | 0.862469332 | 0.047530668 |
| 0.62 | 311 | 101 | 2.5 | 7.64 | 1 | 0.567958891 | 0.052041109 |
| 0.52 | 302 | 99 | 3 | 7.45 | 0 | 0.509677285 | 0.010322715 |
| 0.61 | 322 | 103 | 2.5 | 8.02 | 1 | 0.639606523 | -0.02960652 |
| 0.58 | 298 | 100 | 4 | 7.95 | 1 | 0.61343638 | -0.03343638 |
| 0.57 | 297 | 101 | 4 | 7.67 | 1 | 0.580805149 | -0.01080515 |
| 0.61 | 300 | 98 | 2.5 | 8.02 | 0 | 0.560665609 | 0.049334391 |
| 0.54 | 301 | 96 | 4 | 7.56 | 0 | 0.534886053 | 0.005113947 |
| 0.56 | 313 | 94 | 1.5 | 8.13 | 0 | 0.562238193 | -0.00223819 |
| 0.59 | 314 | 102 | 2 | 7.88 | 1 | 0.593989796 | -0.0039898 |
| 0.49 | 317 | 101 | 2 | 7.94 | 1 | 0.603564123 | -0.11356412 |
| 0.72 | 321 | 110 | 4 | 8.35 | 1 | 0.733143884 | -0.01314388 |
| 0.76 | 327 | 106 | 4.5 | 8.75 | 1 | 0.791497895 | -0.03149789 |
| 0.65 | 315 | 104 | 2.5 | 8.1 | 0 | 0.615267851 | 0.034732149 |
| 0.52 | 316 | 103 | 2 | 7.68 | 0 | 0.551808045 | -0.03180804 |
| 0.6 | 309 | 111 | 4 | 8.03 | 0 | 0.651493477 | -0.05149348 |
| 0.58 | 308 | 102 | 3.5 | 7.98 | 1 | 0.629562434 | -0.04956243 |
| 0.42 | 299 | 100 | 3 | 7.42 | 0 | 0.503733084 | -0.08373308 |
| 0.77 | 321 | 112 | 4.5 | 8.95 | 1 | 0.823198416 | -0.05319842 |
| 0.73 | 322 | 112 | 2.5 | 9.02 | 1 | 0.787898344 | -0.05789834 |
| 0.94 | 334 | 119 | 5 | 9.54 | 1 | 0.950368833 | -0.01036883 |

data | training | validation | comparison

- In step 6, we calculated the 3 important values SSPE, MSPE, and RMSPE using the Excel formulae and got the following output: SSPE = 0.185253, MSPE = 0.001853, RMSPE = 0.043041

- In step 7, we moved on to create a comparison table for the training model vs validation model as follows:

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Statistic | ($n$ = 400) | ($n$ = 100) | Difference | Difference | |
| 2 | $b_0$ | -1.2985 | -1.4803 | 0.1818 | -14.0% | |
| 3 | GRE Score | 0.0018 | 0.0023 | -0.0005 | -29.4% | |
| 4 | TOEFL Score | 0.0030 | 0.0032 | -0.0002 | -5.7% | |
| 5 | LOR | 0.0228 | 0.0099 | 0.0129 | 56.5% | |
| 6 | CGPA | 0.1210 | 0.1264 | -0.0054 | -4.5% | |
| 7 | Research | 0.0246 | 0.0254 | -0.0008 | -3.5% | |
| 8 | s[$b_0$] | 0.1173 | 0.1654 | -0.0481 | -41.0% | |
| 9 | s[GRE Score] | 0.0006 | 0.0008 | -0.0002 | -36.8% | |
| 10 | s[TOEFL Score] | 0.0011 | 0.0012 | -0.0002 | -14.2% | |
| 11 | s[LOR] | 0.0048 | 0.0052 | -0.0004 | -7.7% | |
| 12 | s[CGPA] | 0.0117 | 0.0127 | -0.0009 | -7.9% | |
| 13 | s[Research] | 0.0079 | 0.0103 | -0.0023 | -29.6% | |
| 14 | r-squared | 80.27% | 90.52% | -0.10 | -12.8% | |
| 15 | SSE | 1.6008 | 0.1717 | 1.43 | 89.3% | |
| 16 | MSE | 0.0041 | 0.0018 | 0.00 | 55.0% | |
| 17 | RMSE | 0.06 | 0.04 | 0.02 | 32.9% | |
| 18 | MSPE | 0.001853 | Mean Squared Prediction Error | | | |
| 19 | RMSPE | 0.043 | Root Mean Squared Prediction Error | | | |
| 20 | | | | | | |

- On carefully comparing the MSE from the training model with MSPE from the validation model, we figured that MSE is greater than MSPE (0.0041 > 0.001853). With this, we conclude the model does even better on the validation data (smaller error).

## C. Checking the Four Major Assumptions of Regression

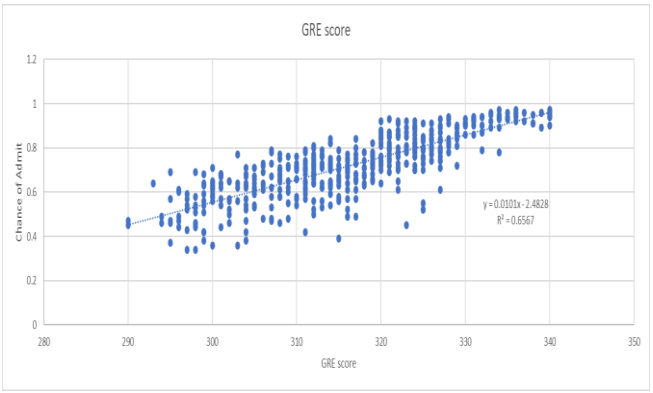
1. The model is linear.

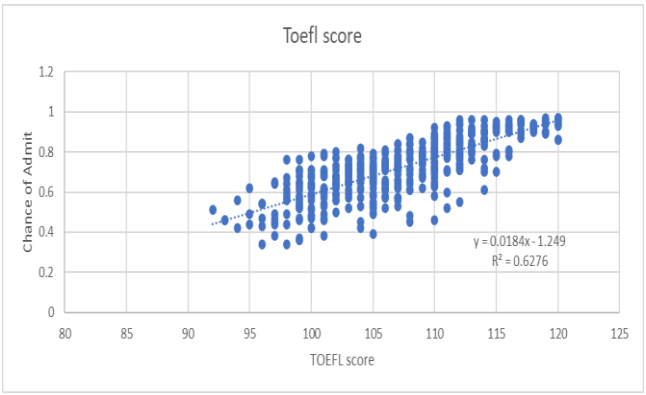
2. The error terms have constant variance.

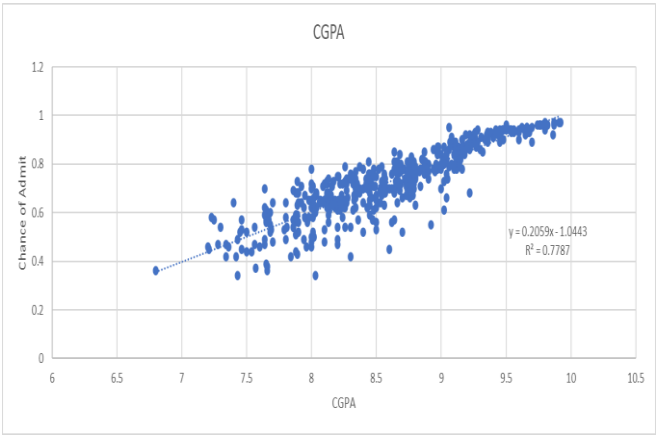
3. The error terms are independent.

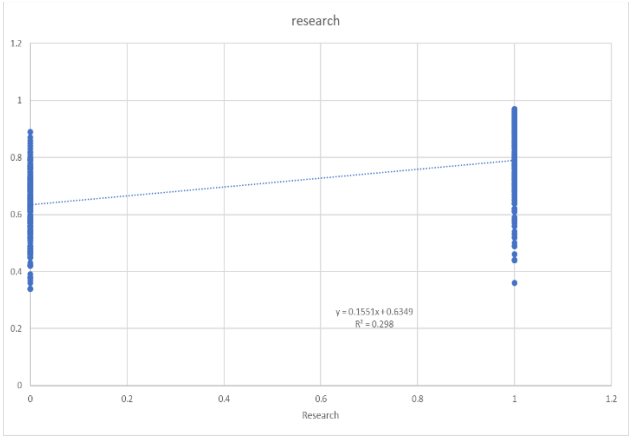
4. The error terms are normally distributed.

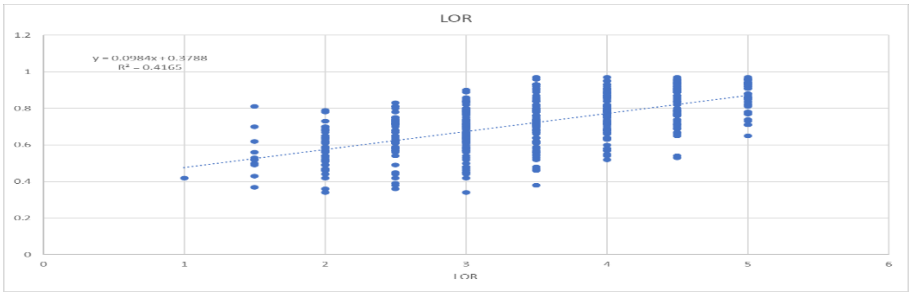
## Scatter plots



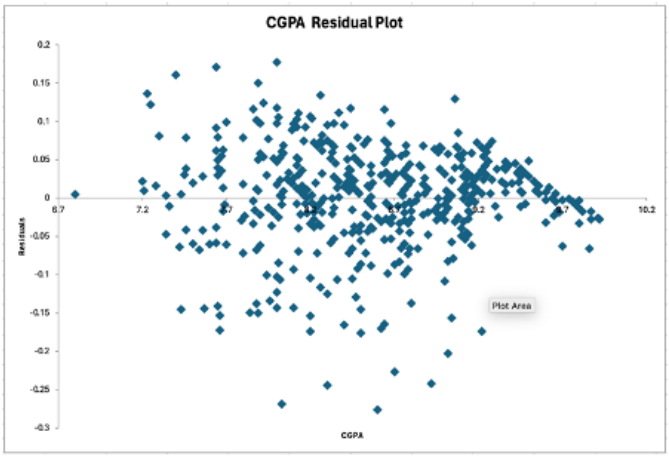



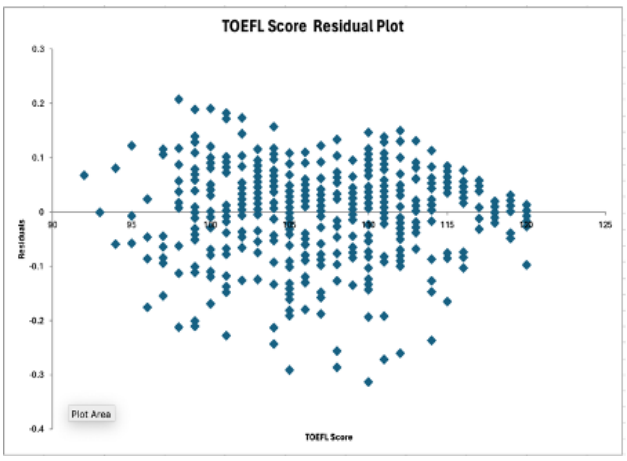



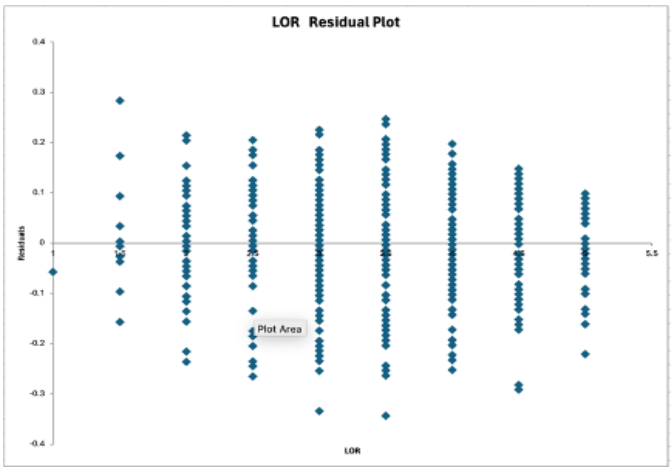



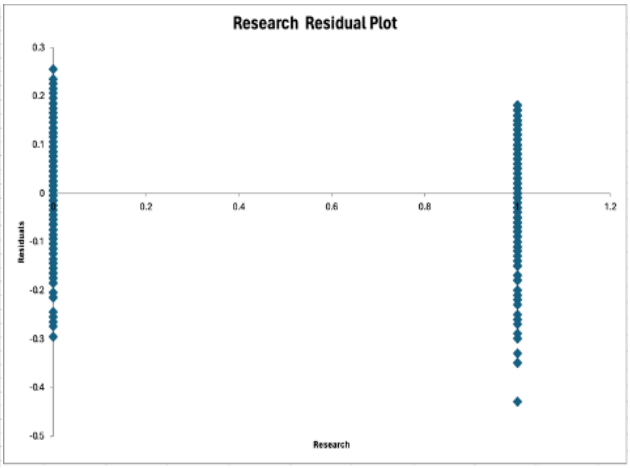



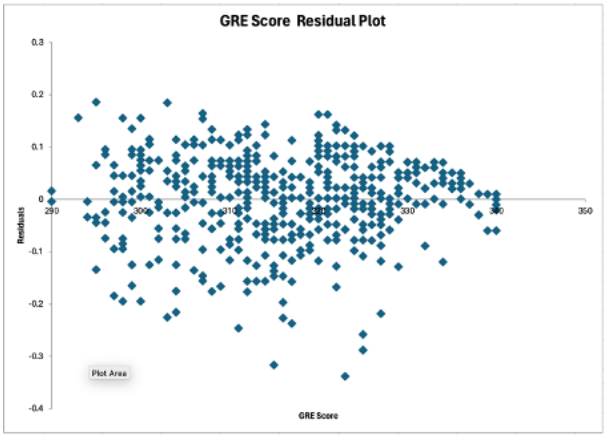

## Residual plots:






**1. Assumption One (Linearity) is Satisfied:**

- There is no apparent pattern, trend, curve, or shape observed in the scatter plots.
- This indicates that the assumption of linearity is likely satisfied, as there are no indications of non-linear relationships between the predictor (X) and the response (Y).

2. **Assumption Two is Satisfied**:

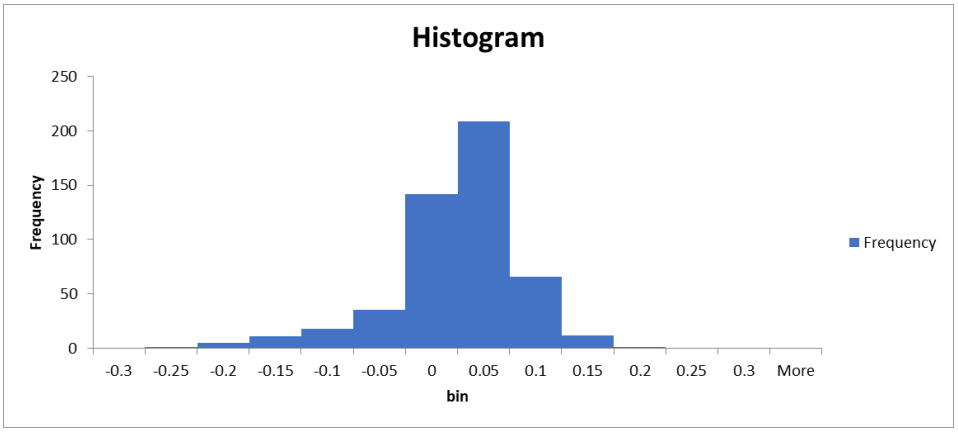
- From the residual error graphs, we also see a roughly constant variance of the errors, *i.e.*, the errors do not seem to be either expanding or contracting as the X variable changes,

**3. Assumption Three (Error terms independence) is Satisfied:**

- We don't see any clear trend or pattern in the residual error plots which is why we state that the assumption that error terms are independent is satisfied.

**4. Assumption Four (normal distribution) is Satisfied:**

- We created a histogram of the residuals (see below). Since the errors seem almost normally distributed, Assumption 4 seems to hold approximately true.




**Team 4:** 43/50

Main Body: 17/20. Working with hypothetical data is not ideal. You never stated why you think the relationship between Y and Research, and between Y and Undergraduate GPA should be negative. Your response variable is a probability, so you should express changes to it as a percent, *e.g.,* "by 12%" or (better) "12 percentage points" rather than "by 12 units." Stepwise regression should have terminated after Step 3 because the Adjusted $R^2$ improved a neglible amount after that.

I appreciate your validation effort. However, (1) you should have split the data first, before doing stepwise regression modeling; the results of that process gives you the model that you use on the validation data. (2) There seems to be some text missing in "step 4" and "step 5." (3) Your model's predictions on the validation data seem *more* accurate than they are for the training data, so I don't think you overfit the training data.

Appendix: 19/20. It's not necessary to include screen shots of how the data were split. Good job with the four regression assumptions.

Report Quality: 7/10. There are a lot of grammatical and capitalization errors, as well as some unclear/missing statements. Use the past tense consistently when you are talking about what you did.