Mrinal Kanti Bose

grade_buddy

Coding Theory

Document Details

Submission ID
trn:oid:::3618:128199093

Submission Date
Feb 13, 2026, 8:03 PM GMT+5:30

Download Date
Feb 13, 2026, 8:05 PM GMT+5:30

File Name
grade_buddy.docx

File Size
569.0 KB

29 Pages

3,380 Words

18,292 Characters

49% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Detection Groups

59 AI-generated only 49%
Likely AI-generated text from a large-language model.

0 AI-generated text that was AI-paraphrased 0%
Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

February 12, 2026

Contents

Introduction

This analysis examines factors affecting college graduation rates using multiple regression analysis. The dataset contains 777 observations with 16 variables including the dependent variable Grad_Rate (graduation rate) and 14 independent variables representing various college characteristics.

Variables:

- **Dependent Variable (Y):** Grad_Rate (Graduation Rate)
- **Independent Variables:** Private, Accept_pct, Elite10, F_Undergrad, P_Undergrad, Outstate, Room_Board, Books, Personal, PhD, Terminal, S_F_Ratio, perc_alumni, Expend

Question (a): Distribution of Grad_Rate

SAS Output - Histogram

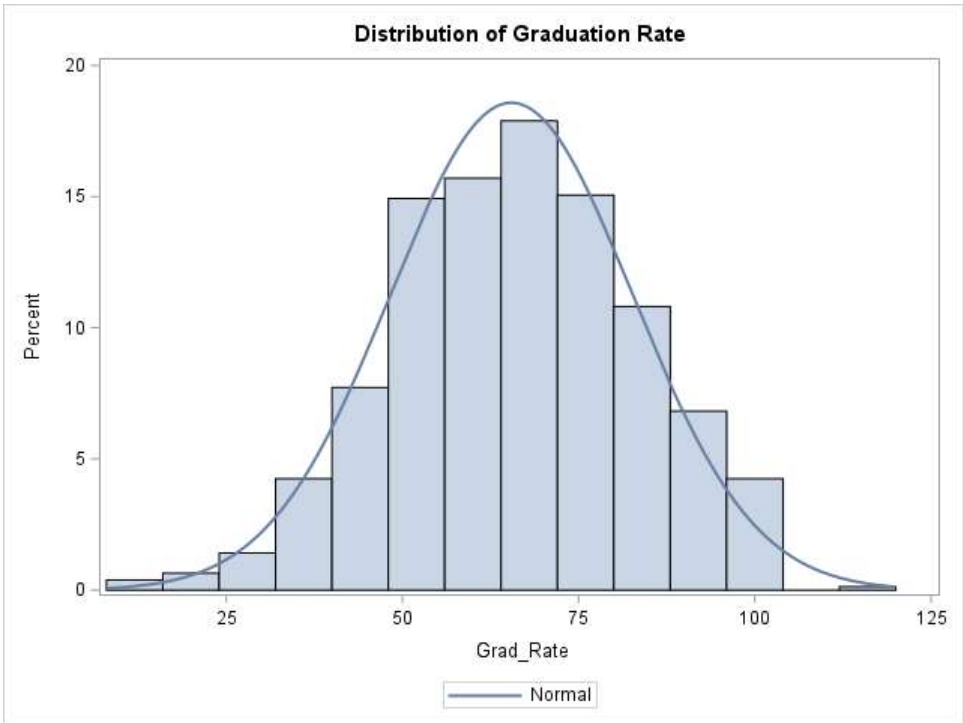

Figure 1: Distribution of Graduation Rate

SAS Output - Descriptive Statistics

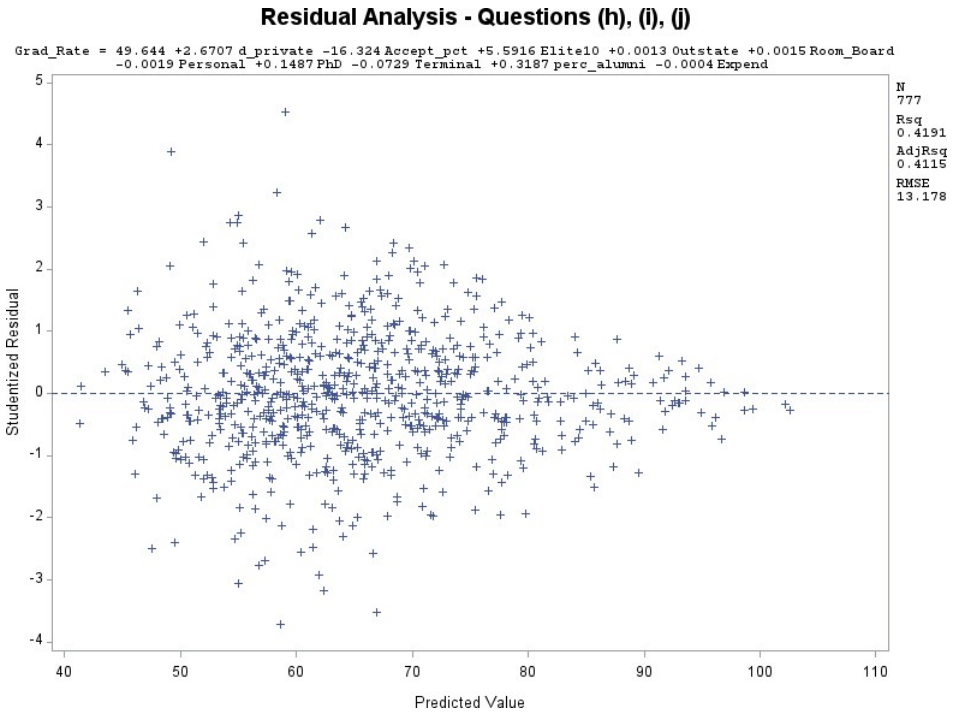

Figure 2: Descriptive Statistics for Grad_Rate

Statistical Summary

Table 1: Descriptive Statistics for Grad_Rate

Statistic	Value
N	777
Mean	65.46
Median	65.00
Std Dev	17.18
Minimum	10
Maximum	118
Q1	53
Q3	78
Skewness	-0.114
Kurtosis	-0.205

Interpretation

The distribution of Grad_Rate is **approximately symmetric**. The mean (65.46) is nearly equal to the median (65.00), indicating no substantial skewness. The skewness value of -0.114 is very close to zero (within ± 1), confirming the distribution is approximately symmetric. The kurtosis value of -0.205 indicates a slightly platykurtic (flatter) distribution compared to normal.

Conclusion: Since the distribution is approximately symmetric and shows no severe departures from normality, **no transformation of the dependent variable is needed** for regression analysis.

SAS Code for Question (a)

```
proc means data=college n mean median std min max q1 q3 skewness
   kurtosis;
    var Grad_Rate;
    title "Descriptive Statistics for Grad_Rate";
run;

proc sgplot data=college;
    histogram Grad_Rate;
    density Grad_Rate;
    title "Distribution of Graduation Rate";
run;
```

Question (b): Scatterplots

SAS Output - Scatterplot Matrix

The scatterplot matrix displays bivariate relationships between Grad_Rate and all continuous predictors.

Interpretation

Positive Relationships with Grad_Rate:

- **Outstate:** Strong positive relationship — higher out-of-state tuition is associated with higher graduation rates
- **perc_alumni:** Moderate positive relationship — higher alumni giving percentage correlates with higher graduation rates
- **Room_Board:** Moderate positive relationship
- **PhD and Terminal:** Moderate positive relationships — more faculty credentials associated with higher graduation rates
- **Expend:** Positive relationship — higher expenditure per student correlates with higher graduation rates

Negative Relationships with Grad_Rate:

- **Accept_pct:** Negative relationship — schools with higher acceptance rates (less selective) tend to have lower graduation rates
- **S_F_Ratio:** Weak negative relationship — higher student-faculty ratios slightly associated with lower graduation rates

- **P_Undergrad:** Weak negative relationship — more part-time undergraduates associated with somewhat lower graduation rates

Weak/No Clear Relationship:

- **Books:** No clear linear relationship with graduation rate
- **Personal:** Weak or no clear linear relationship
- **F_Undergrad:** No strong linear pattern

Conclusion: The scatterplots suggest that **Outstate**, **perc_alumni**, and **Accept_pct** are likely to be strong predictors in the regression model.

SAS Code for Question (b)

```
proc sgscatter data=college;
    title "Scatterplot Matrix for College Data";
    matrix Grad_Rate Accept_pct F_Undergrad P_Undergrad Outstate
           Room_Board Books Personal PhD Terminal S_F_Ratio
           perc_alumni Expend;
run;
```

Question (c): Boxplots for categorical variables

SAS Output - Private vs Public Boxplot

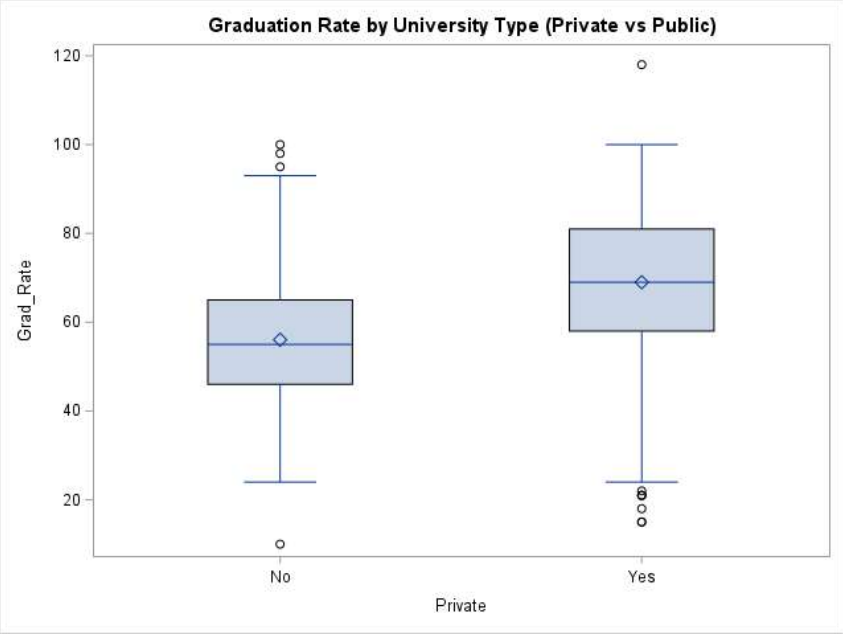

Figure 3: Graduation Rate by University Type (Private vs Public)

Interpretation - Private vs Public

Private universities (Private = “Yes”) have a **higher median graduation rate** compared to public universities (Private = “No”). The interquartile range (box) for private

schools is positioned higher on the y-axis, indicating that private institutions generally achieve better graduation rates.

SAS Output - Elite vs Non-Elite Boxplot

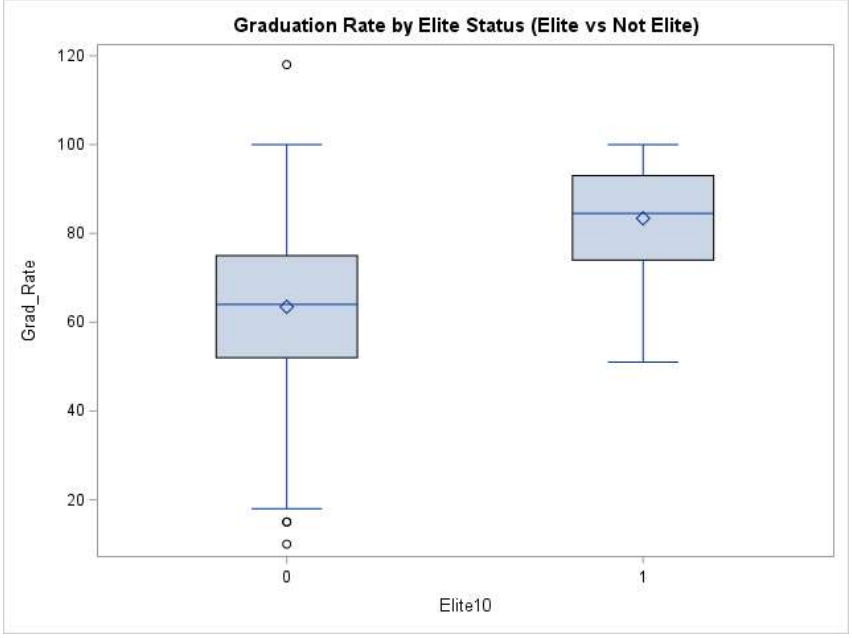

Figure 4: Graduation Rate by Elite Status (Elite10 = 1 vs Elite10 = 0)

Interpretation - Elite vs Non-Elite

Elite universities (Elite10 = 1) show **substantially higher graduation rates** compared to non-elite universities (Elite10 = 0). The median graduation rate for elite schools is noticeably higher, and the entire distribution is shifted upward. This suggests that **elite status is positively associated with graduation rates**.

SAS Code for Question (c)

```
proc sgplot data=college;
    vbox Grad_Rate / category=Private;
    title "Graduation Rate by University Type (Private vs Public)";
run;

proc sgplot data=college;
    vbox Grad_Rate / category=Elite10;
    title "Graduation Rate by Elite Status";
run;
```

Question (d): Full Model Results

SAS Output - Full Model

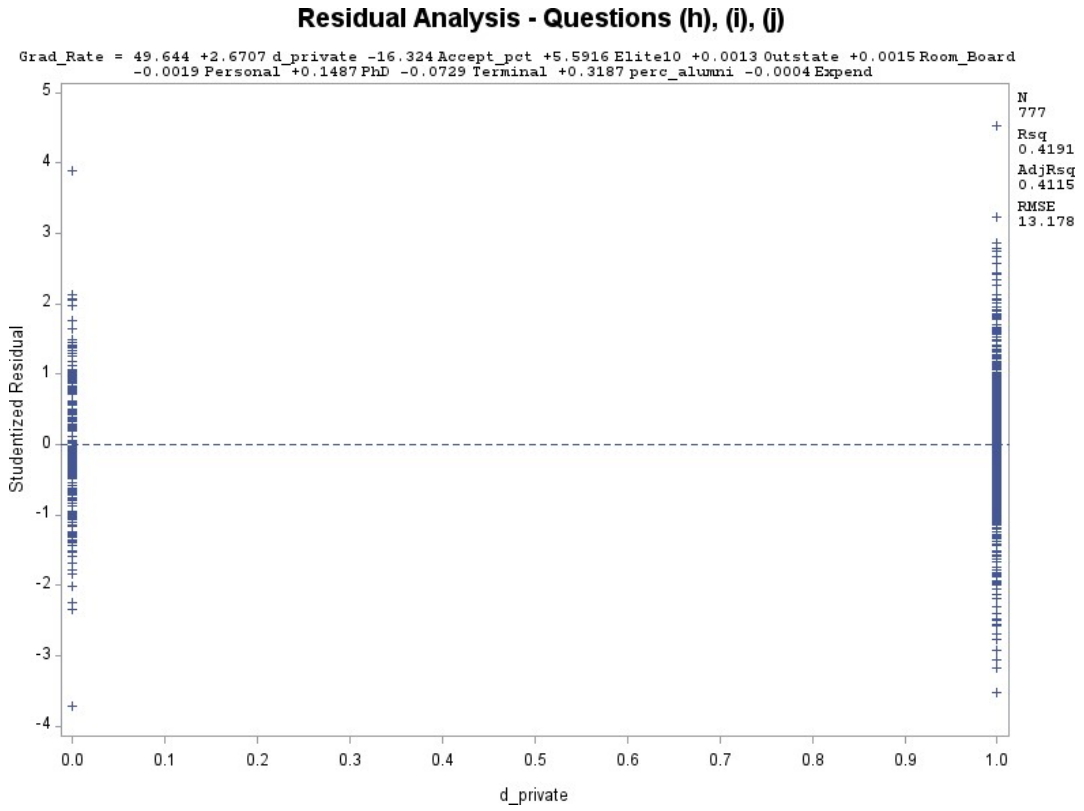

Figure 5: ANOVA Table for Full Model

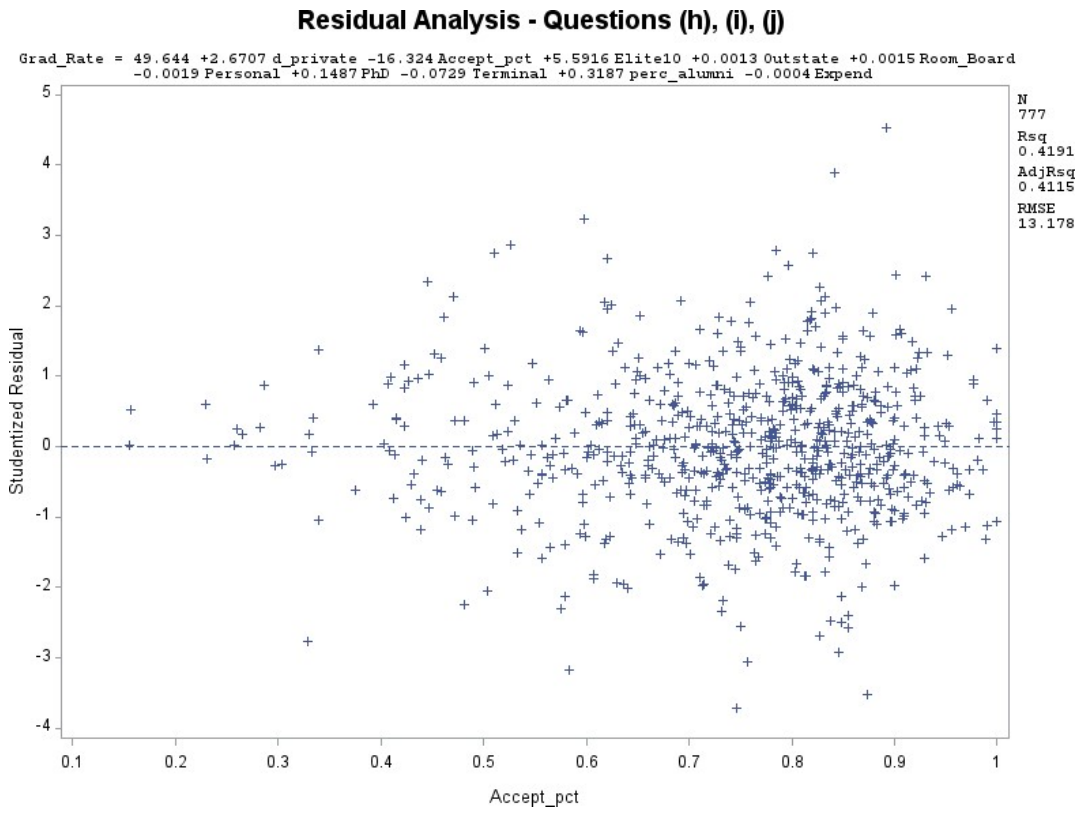

Figure 6: Parameter Estimates for Full Model

Model Statistics

Table 2: Full Model Summary

Statistic	Value
R-Square	0.4448
Adjusted R-Square	0.4346
F-value	43.61
p-value	< .0001

Significant Variables ($p < 0.05$)

Table 3: Significant Predictor Variables

Variable	Estimate	t-value	p-value
d_private	4.620	2.68	0.0075
Accept_pct	-18.109	-4.71	< .0001
Elite10	4.017	2.01	0.0453
F_Undergrad	0.00068	4.77	< .0001
P_Undergrad	-0.00196	-5.01	< .0001
Outstate	0.00123	5.40	< .0001
Room_Board	0.00167	2.80	0.0052
Personal	-0.00172	-2.21	0.0275
PhD	0.13064	2.32	0.0204
perc_alumni	0.30920	6.39	< .0001
Expend	-0.00044	-2.88	0.0041

Non-Significant Variables ($p > 0.05$)

Table 4: Non-Significant Predictor Variables

Variable	p-value
Books	0.3951
Terminal	0.2447
S_F_Ratio	0.9951

Interpretation

The full model is **statistically significant** ($F = 43.61$, $p < .0001$), indicating that at least one predictor variable has a significant relationship with Grad Rate.

Model Fit: The $R^2 = 0.4448$ indicates that **44.48% of the variation** in graduation rates is explained by the model. The Adjusted $R^2 = 0.4346$ accounts for the number of predictors.

Key Findings:

- **Positive predictors:** Private status (+4.62 points), Elite status (+4.02 points), perc_alumni (+0.31 per 1% increase)
- **Negative predictors:** Accept pct (−18.11 per unit increase) — more selective schools have higher graduation rates
- **Non-significant:** Books, Terminal, and S F_Ratio do not significantly predict graduation rate

SAS Code for Question (d)

```
proc reg data=college2;
    model Grad_Rate = d_private Accept_pct Elite10 F_Undergrad
       P_Undergrad
                      Outstate Room_Board Books Personal PhD Terminal
                      S_F_Ratio perc_alumni Expend;
    title "FULL MODEL";
run;
quit;
```

Question (e): Multicollinearity Analysis (VIF)

SAS Output - VIF Table

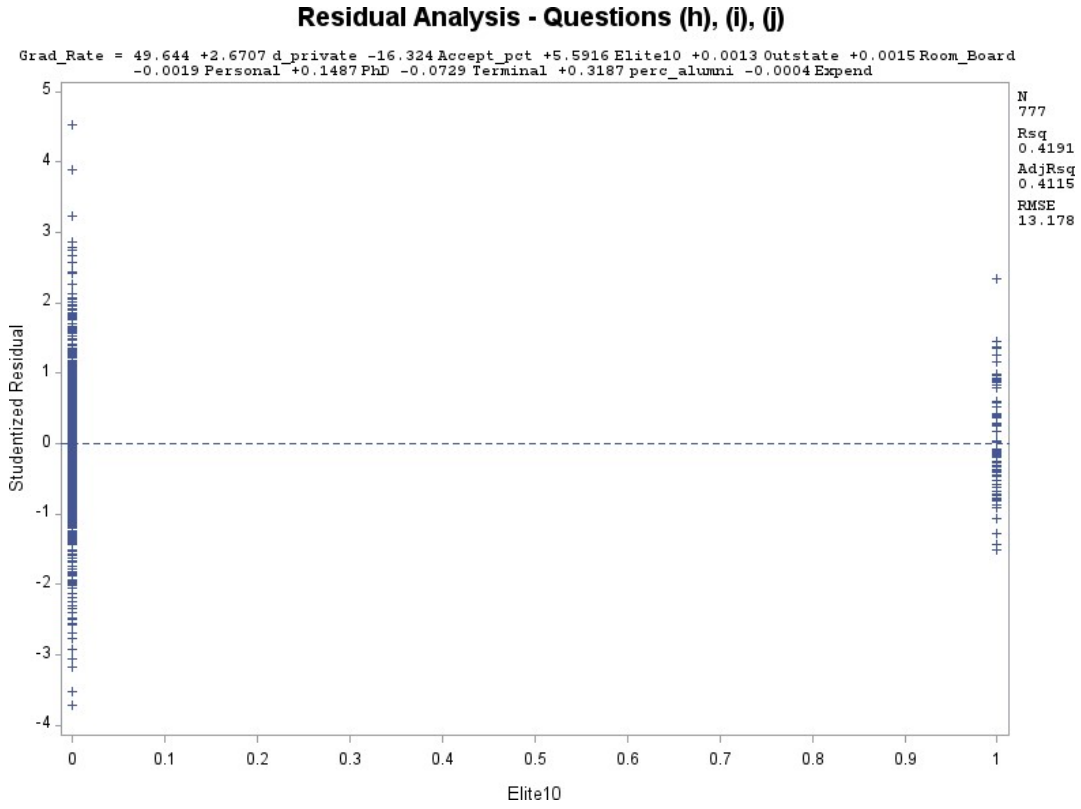

Figure 7: VIF and Tolerance Statistics

VIF and Tolerance Results

Table 5: Variance Inflation Factors

Variable	Tolerance	VIF
d_private	0.365	2.740
Accept_pct	0.673	1.487
Elite10	0.592	1.688
F_Undergrad	0.448	2.233
P_Undergrad	0.609	1.643
Outstate	0.254	3.935
Room_Board	0.506	1.977
Books	0.896	1.116
Personal	0.775	1.291
PhD	0.255	3.918
Terminal	0.253	3.947
S_F_Ratio	0.524	1.910
perc_alumni	0.598	1.672
Expend	0.342	2.923

Interpretation

There is **NO multicollinearity problem** in the full model.

Rule of Thumb:

- VIF $\geq$ 10 indicates severe multicollinearity
- Tolerance $\leq$ 0.10 indicates severe multicollinearity

Results:

- All VIF values are well below 10 (the largest VIF is 3.947 for Terminal)
- All tolerance values are above 0.10

Conclusion: Since no variable exceeds these thresholds, we can proceed with the analysis without concern for multicollinearity.

SAS Code for Question (e)

```
proc reg data=college2;
    model Grad_Rate = d_private Accept_pct Elite10 F_Undergrad
       P_Undergrad
                      Outstate Room_Board Books Personal PhD Terminal
                      S_F_Ratio perc_alumni Expend / vif tol;
    title "FULL MODEL with VIF";
run;
quit;
```

Question (f): Model Selection Methods

SAS Output - Stepwise Selection

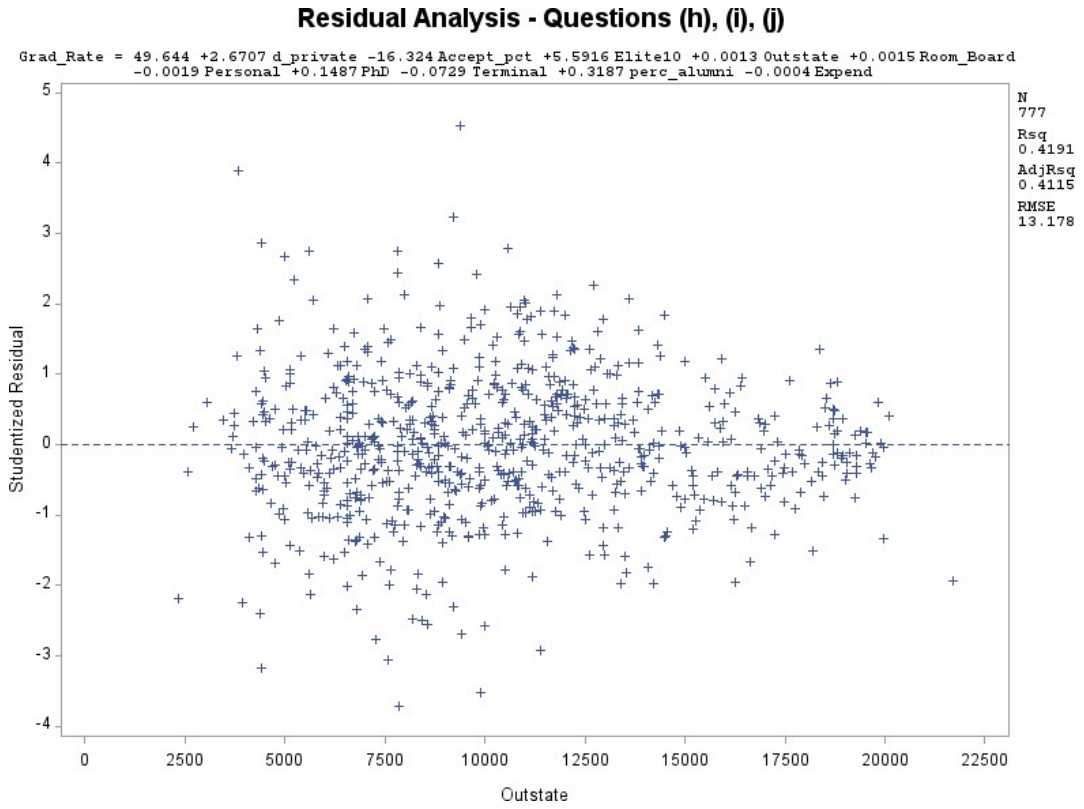

Figure 8: Stepwise Selection Summary

Stepwise Selection Summary

Table 6: Stepwise Selection - Variables Entered

Step	Variable Added	Model R^2
1	Outstate	0.3264
2	perc_alumni	0.3676
3	Accept_pct	0.3916
4	P_Undergrad	0.4036
5	F_Undergrad	0.4164
6	Room_Board	0.4230
7	Expend	0.4287
8	Personal	0.4326
9	d_private	0.4360
10	PhD	0.4401

Final stepwise model includes **10 variables** with $R^2 = 0.4401$.

SAS Output - Backward Elimination

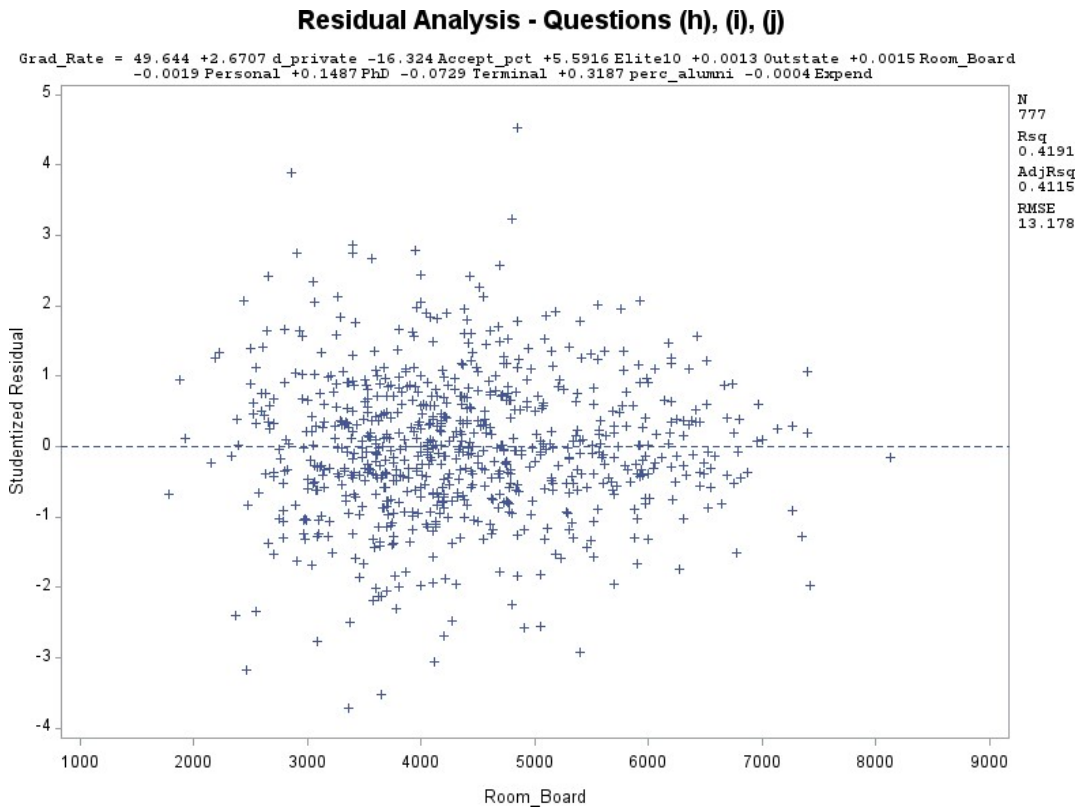

Figure 9: Backward Elimination Summary

Backward Elimination Summary

Table 7: Backward Elimination - Variables Removed

Step	Variable Removed	Model R^2
0	(All 14 variables)	0.4448
1	S_F_Ratio	0.4448
2	Books	0.4443
3	Terminal	0.4435
4	Elite10	0.4412

Interpretation

Both methods converged to similar models, with minor differences in which variables were retained. Both identified that **S_ F_ Ratio** and **Books** are not significant predictors.

SAS Code for Question (f)

```
proc reg data=college2;
    model Grad_Rate = d_private Accept_pct Elite10 F_Undergrad
       P_Undergrad
```

```
                        Outstate Room_Board Books Personal PhD Terminal
                        S_F_Ratio perc_alumni Expend / selection=stepwise
                            ;
    title "STEPWISE Selection";
run;
quit;

proc reg data=college2;
    model Grad_Rate = d_private Accept_pct Elite10 F_Undergrad
       P_Undergrad
                        Outstate Room_Board Books Personal PhD Terminal
                        S_F_Ratio perc_alumni Expend / selection=backward
                            ;
    title "BACKWARD Elimination";
run;
quit;
```

Question (g): Final Model M1

SAS Output - Final Model

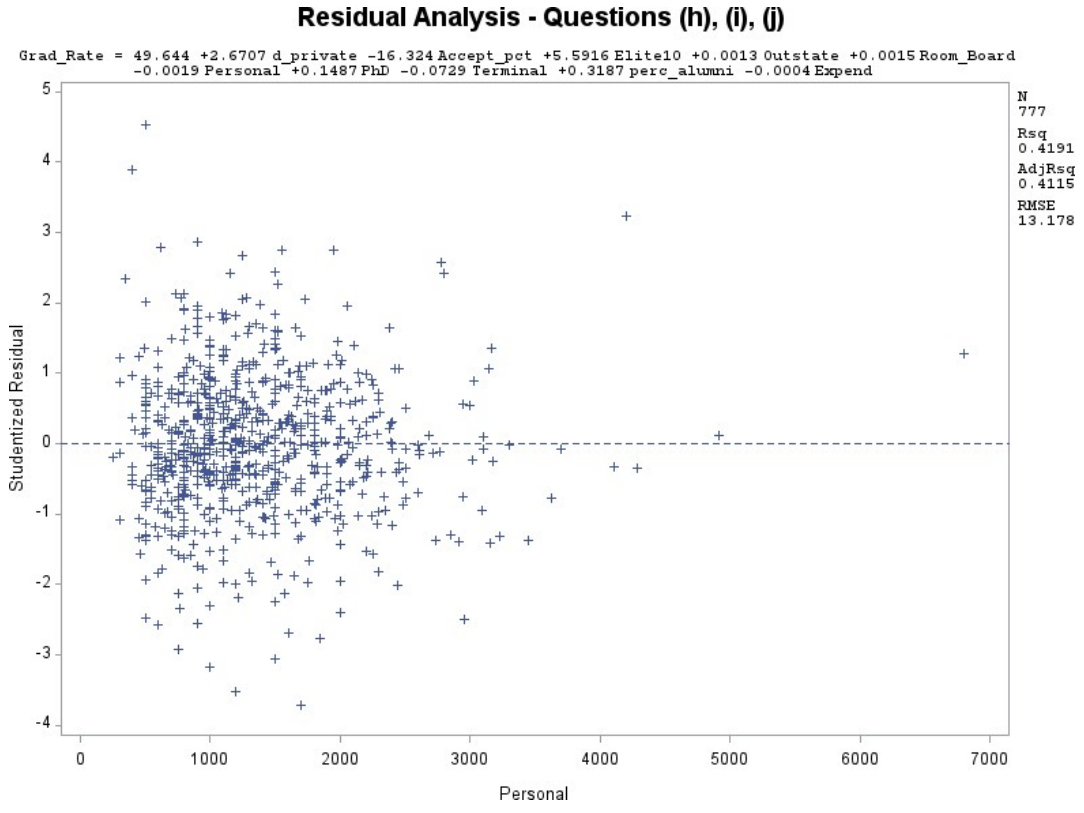

Figure 10: Final Model M1 Parameter Estimates

Final Model Statistics

Based on the selection methods, the final model includes:

d_private, Accept_pct, Elite10, Outstate, Room_Board, Personal, PhD, Terminal, perc_alumni, Expend

Table 8: Final Model Summary

Statistic	Value
R-Square	0.4191
Adjusted R-Square	0.4115
F-value	55.26
p-value	< .0001

Regression Equation

$$\widehat{\text{Grad\_Rate}} = 49.64 + 2.67(\text{d\_private}) - 16.32(\text{Accept\_pct}) + 5.59(\text{Elite10}) + 0.0013(\text{Outstate}) + 0.0015(\text{Room\_Board}) - 0.0019(\text{Personal}) + 0.15(\text{PhD}) - 0.073(\text{Terminal}) + 0.32(\text{perc\_alumni}) - 0.00042(\text{Expend})$$

SAS Code for Question (g)

```
proc reg data=college2;
    model Grad_Rate = d_private Accept_pct Elite10 Outstate Room_Board
                      Personal PhD Terminal perc_alumni Expend;
    title "FINAL MODEL M1";
run;
quit;
```

Question (h): Residuals vs Predicted Values Plot

SAS Output - Residual Plot

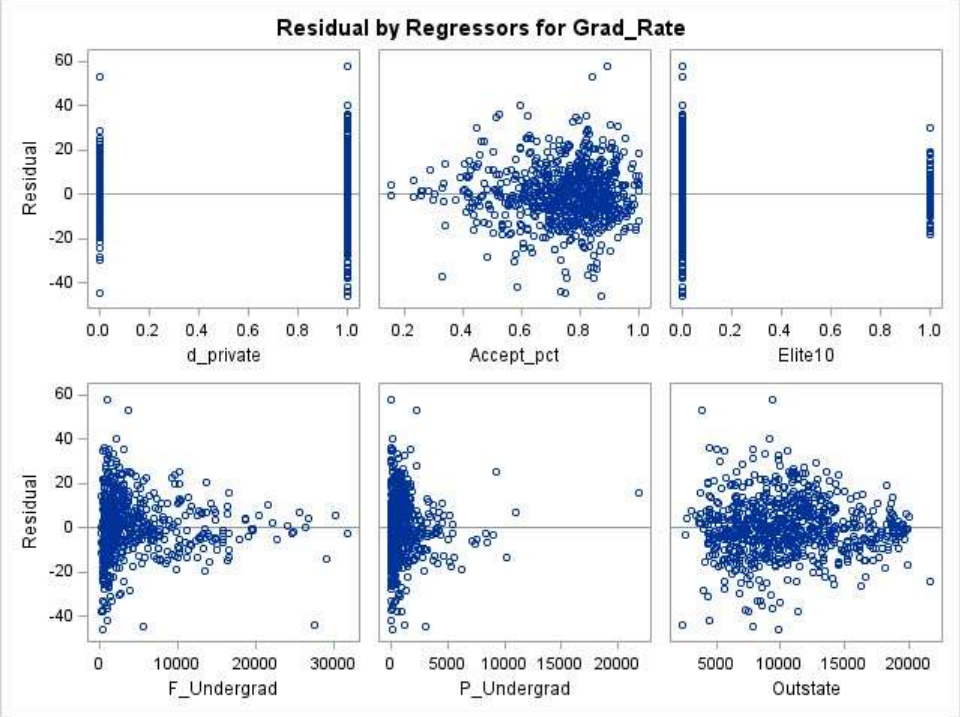

Figure 11: Studentized Residuals vs Predicted Values

Interpretation

The plot of studentized residuals versus predicted values shows a **random scatter of points around zero** with no apparent pattern. This indicates:

1. **Linearity assumption is satisfied** — no curved pattern observed
2. **Constant variance (homoscedasticity) assumption appears to be met** — the spread of residuals is roughly constant across predicted values
3. **No systematic pattern** suggesting model misspecification

The residuals are randomly distributed between approximately −3 and +3, with most falling within ±2 standard deviations.

SAS Code for Question (h)

```
proc reg data=college2;
    model Grad_Rate = d_private Accept_pct Elite10 Outstate Room_Board
                      Personal PhD Terminal perc_alumni Expend /
                         influence r;
    plot student.*predicted.;
    title "Residuals vs Predicted Values";
run;
quit;
```

Question (i): Normal Probability Plot

SAS Output - Normal Probability Plot

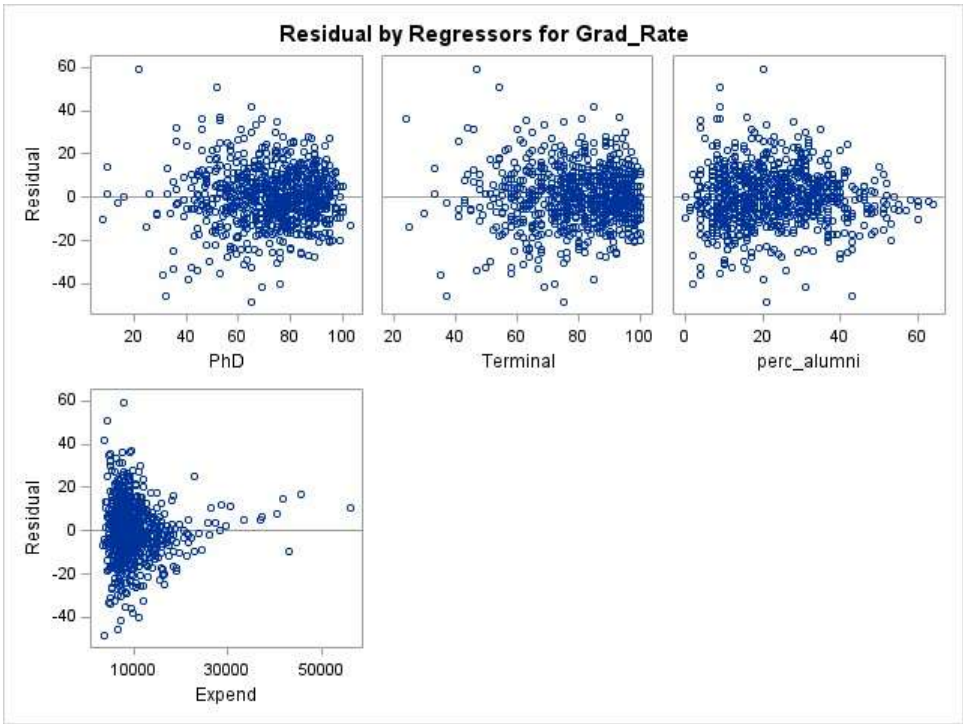

Figure 12: Normal Probability Plot of Residuals

Interpretation

The normal probability plot shows the residuals **following the diagonal reference line reasonably well**, with minor deviations at the tails. This indicates that the **normality assumption for residuals is approximately satisfied**.

Some slight departures from normality are observed at the extreme ends, but these are not severe enough to invalidate the regression analysis. The central portion of the distribution closely follows the expected normal pattern.

SAS Code for Question (i)

```
proc reg data=college2;
    model Grad_Rate = d_private Accept_pct Elite10 Outstate Room_Board
                      Personal PhD Terminal perc_alumni Expend /
                        influence r;
    plot npp.*student.;
    title "Normal Probability Plot of Residuals";
run;
quit;
```

Question (j): Outliers and Influential Points

SAS Output - Influence Diagnostics

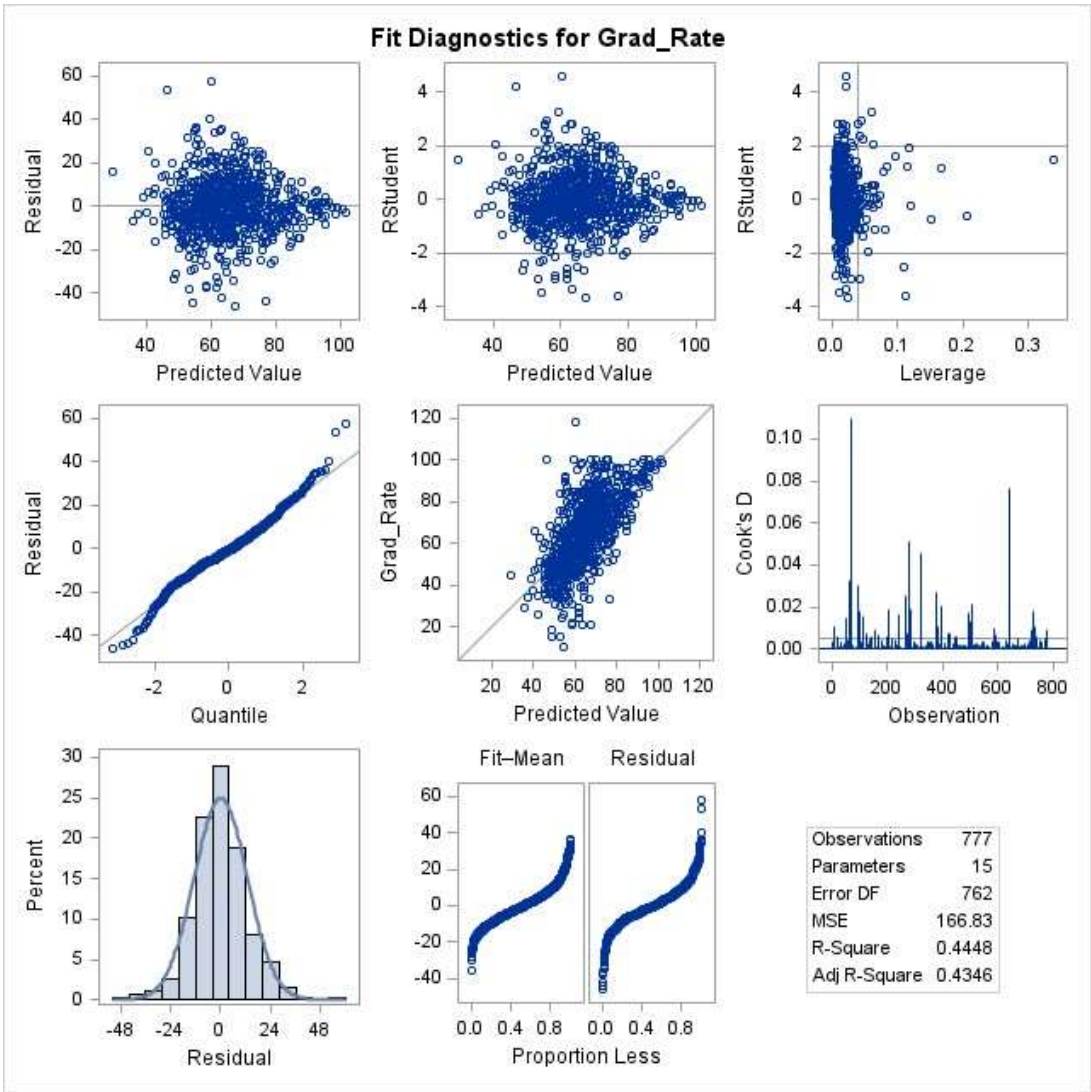

Figure 13: Influence Diagnostics Output

Criteria for Identification

To identify outliers and influential observations, we examine:

Table 9: Diagnostic Thresholds

Diagnostic	Threshold
Outliers	$\lvert\text{Studentized Residual}\rvert > 2$
Influential Points (Cook's D)	$D > \frac{4}{n} = \frac{4}{777} \approx 0.005$
High Leverage	$h > \frac{2(p+1)}{n}$

Interpretation

- **Outliers:** Observations with studentized residuals exceeding ±2 are potential outliers with unusually large residuals given the model
- **Influential Points:** Points with Cook's D exceeding 0.005 may have substantial influence on the regression coefficients
- **High Leverage Points:** Points with unusual combinations of predictor values

[Note: Examine the influence output table to identify specific observations that exceed these thresholds]

SAS Code for Question (j)

```
proc reg data=college2;
    model Grad_Rate = d_private Accept_pct Elite10 Outstate Room_Board
                      Personal PhD Terminal perc_alumni Expend /
                         influence r;
    title "Outliers and Influential Points Analysis";
run;
quit;
```

Question (k): Standardized Coefficients - Predictor Rankings

SAS Output - Standardized Estimates

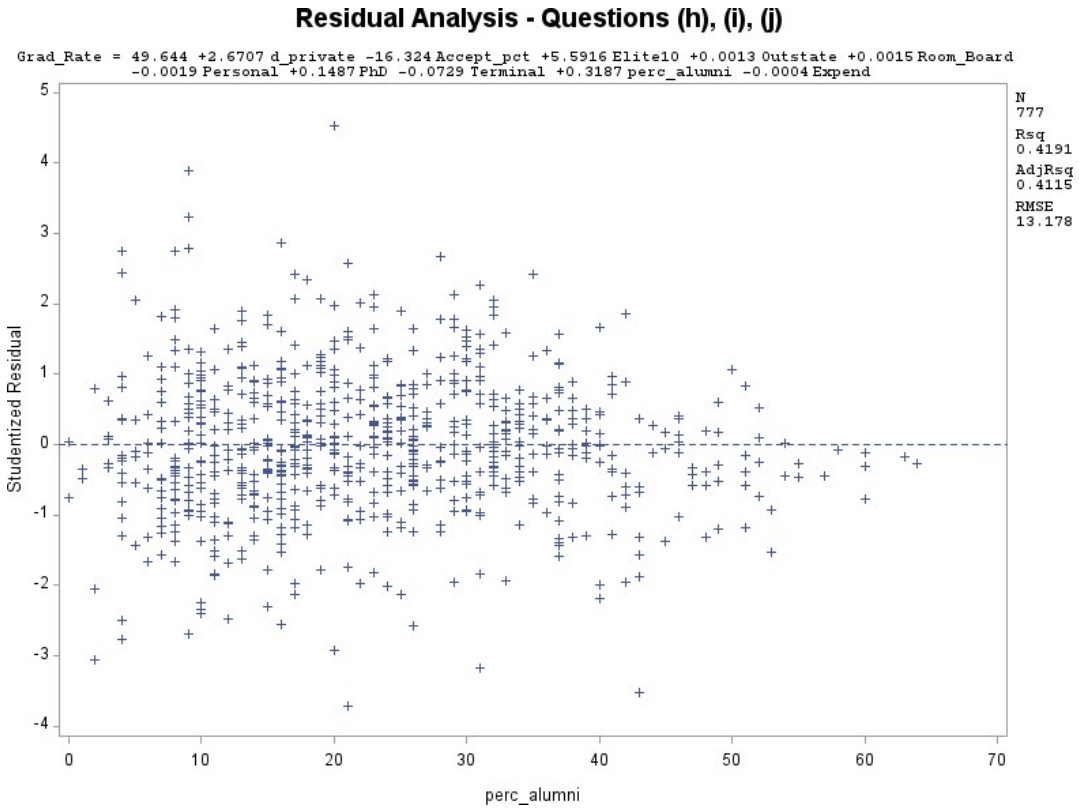

Figure 14: Parameter Estimates with Standardized Coefficients

Predictors Ranked by Importance

Table 10: Predictors Ranked by Absolute Standardized Coefficient

Rank	Variable	Standardized Estimate	Interpretation
1	Outstate	0.30358	Most important
2	perc_alumni	0.22994	Second most important
3	PhD	0.14131	Third most important
4	Accept_pct	−0.13979	Fourth most important
5	Expend	−0.12778	Fifth most important
6	Elite10	0.09789	
7	Room_Board	0.09339	
8	Personal	−0.07578	
9	d_private	0.06930	
10	Terminal	−0.06250	Least important

Interpretation

Using standardized coefficients allows direct comparison of predictor importance regardless of their original scales:

1. **Outstate (0.304)** is the most important predictor — a 1 standard deviation increase in out-of-state tuition is associated with a 0.304 standard deviation increase in graduation rate.
2. **perc_alumni (0.230)** is second — alumni giving rate strongly predicts graduation rate.
3. **PhD (0.141)** is third — higher percentage of faculty with PhDs is associated with higher graduation rates.

The negative coefficients for **Accept_pct (−0.140)** and **Expend (−0.128)** indicate inverse relationships — more selective schools (lower acceptance rate) have higher graduation rates.

SAS Code for Question (k)

```
proc reg data=college2;
    model Grad_Rate = d_private Accept_pct Elite10 Outstate Room_Board
                      Personal PhD Terminal perc_alumni Expend / stb
                         vif;
    title "Standardized Coefficients";
run;
quit;
```

Question (l): Final Model Interpretation

Regression Equation

$$\widehat{\text{Grad\_Rate}} = 49.64 + 2.67(\text{d\_private}) - 16.32(\text{Accept\_pct}) + 5.59(\text{Elite10}) + 0.0013(\text{Outstate}) + 0.0015(\text{Room\_Board}) - 0.0019(\text{Personal}) + 0.15(\text{PhD}) - 0.073(\text{Terminal}) + 0.32(\text{perc\_alumni}) - 0.00042(\text{Expend}) \quad (1)$$

Key Coefficient Interpretations

1. **perc_alumni (0.32):** For each 1% increase in alumni giving rate, the predicted graduation rate increases by 0.32 percentage points, holding all other variables constant.
2. **Accept_pct (−16.32):** For each 1-unit increase in acceptance rate (e.g., from 0.70 to 0.80), the predicted graduation rate decreases by approximately 1.63 percentage points. More selective schools have higher graduation rates.

3. **Elite10 (5.59):** Elite universities (top 10% acceptance) have graduation rates that are approximately 5.59 percentage points higher than non-elite universities, holding other factors constant.

4. **d_private (2.67):** Private universities have graduation rates approximately 2.67 percentage points higher than public universities, controlling for other factors.

SAS Code for Question (l)

```
/* Same model as Question (g) - interpret the coefficients */
proc reg data=college2;
    model Grad_Rate = d_private Accept_pct Elite10 Outstate Room_Board
                      Personal PhD Terminal perc_alumni Expend;
    title "FINAL MODEL M1 - Coefficient Interpretation";
run;
quit;
```

Question (m): Prediction for New University

Given Values for New University

Table 11: Predictor Values for New University

Variable	Value
Private (d private)	1 (Yes)
Accept_pct	0.87
Elite10	0 (No)
F_Undergrad	3,000
P_Undergrad	524
Outstate	$6,500
Room_Board	$3,300
Books	$250
Personal	$1,350
PhD	40%
Terminal	34%
S_F_Ratio	30.2
perc_alumni	13%
Expend	$5,201

SAS Output - Prediction

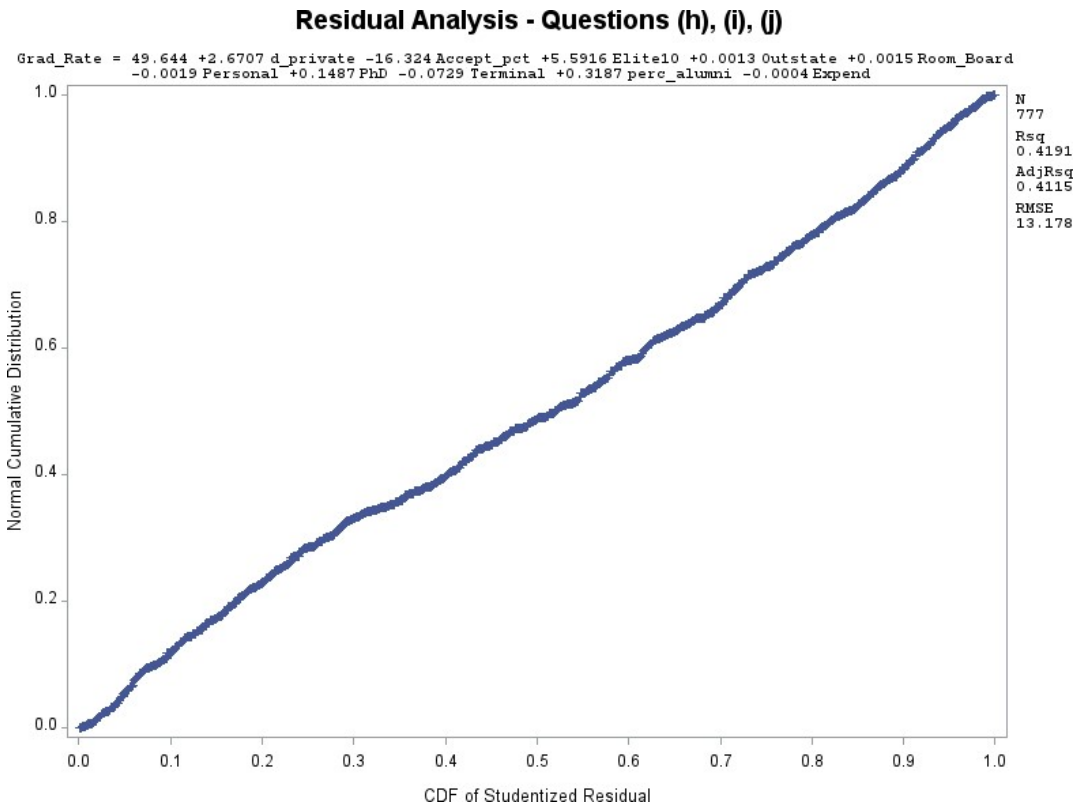

Figure 15: Prediction with Confidence and Prediction Intervals

Prediction Results

Table 12: Prediction Results for New University

Measure	Value
Predicted Graduation Rate	**54.19%**
95% Confidence Interval (CLM)	[50.89%, 57.50%]
95% Prediction Interval (CLI)	[28.12%, 80.27%]

Interpretation

For a private university with the given characteristics, we predict:

1. **Point Estimate:** The expected graduation rate is approximately **54.19%**.
2. **95% Confidence Interval (50.89% to 57.50%):** We are 95% confident that the **AVERAGE graduation rate for ALL universities** with these characteristics falls between 50.89% and 57.50%.
3. **95% Prediction Interval (28.12% to 80.27%):** We are 95% confident that the graduation rate for **THIS SPECIFIC university** falls between 28.12% and 80.27%. The wider interval accounts for individual variation beyond the model.

Note: The prediction interval is much wider than the confidence interval because it accounts for both uncertainty in estimating the mean and natural variation among individual universities.

SAS Code for Question (m)

```
data pred;
    input d_private Accept_pct Elite10 F_Undergrad P_Undergrad
          Outstate Room_Board Books Personal PhD Terminal S_F_Ratio
          perc_alumni Expend Grad_Rate;
    datalines;
1 0.87 0 3000 524 6500 3300 250 1350 40 34 30.2 13 5201 .
;
run;

data college_pred;
    set pred college2;
run;

proc reg data=college_pred;
    model Grad_Rate = d_private Accept_pct Elite10 Outstate Room_Board
                      Personal PhD Terminal perc_alumni Expend / p clm
                         cli alpha=0.05;
    title "Prediction with Intervals";
run;
quit;
```

Problem 2: Model Selection Method Explanation (Graduate Students Only)

I chose **STEPWISE SELECTION** and **BACKWARD ELIMINATION** selection methods for the following reasons:

Stepwise Selection

Stepwise selection is an iterative approach that both adds and removes variables at each step. Starting with no variables in the model:

- At each step, it adds the variable that is most significant (lowest p-value below entry threshold, default 0.15)
- After adding a variable, it checks if any previously entered variables should be removed (p-value above stay threshold, default 0.15)
- This process continues until no variables meet the criteria for entry or removal

Advantages:

- Identifies a parsimonious model with only significant predictors
- Handles situations where adding new variables changes the significance of previously included variables
- Provides clear sequential steps showing variable importance

Backward Elimination

Backward elimination starts with ALL variables in the model:

- At each step, it removes the least significant variable (highest p-value above the threshold)
- The process continues until all remaining variables are significant

Advantages:

- Considers all variables simultaneously from the start
- May retain variables that are only significant in the presence of others
- Less likely to miss important predictors compared to forward selection

Why Both Methods

Using both methods provides validation — if both approaches select similar final models, we have more confidence in the results. In this analysis:

- **Stepwise selected:** Outstate, perc.alumni, Accept.pct, P.Undergrad, F.Undergrad, Room.Board, Expend, Personal, d.private, PhD
- **Backward removed:** S.F.Ratio, Books, Terminal, Elite10

Both methods identified that S.F.Ratio and Books are not significant predictors, confirming they can be excluded from the final model.

Appendix: SAS Code

```
/*================================================================================

  PROBLEM 1: College Graduation Rate Analysis
  Dataset: College.csv
  Dependent Variable (Y): Grad_Rate
================================================================================*/


/* STEP 0: IMPORT DATA */
proc import datafile="\\tsclient\C\Users\xyz\Downloads\College (2).csv"
    out=college
    dbms=csv
    replace;
    getnames=yes;
run;

proc contents data=college;
run;

proc print data=college (obs=10);
run;

/* QUESTION (a): Distribution of Grad_Rate */
proc means data=college n mean median std min max q1 q3 skewness
   kurtosis;
    var Grad_Rate;
    title "Descriptive Statistics for Grad_Rate";
run;

proc sgplot data=college;
    histogram Grad_Rate;
    density Grad_Rate;
    title "Distribution of Graduation Rate";
run;

/* QUESTION (b): Scatterplots */
proc sgscatter data=college;
    title "Scatterplot Matrix for College Data";
    matrix Grad_Rate Accept_pct F_Undergrad P_Undergrad Outstate
           Room_Board Books Personal PhD Terminal S_F_Ratio
           perc_alumni Expend;
run;

/* QUESTION (c): Boxplots */
proc sgplot data=college;
    vbox Grad_Rate / category=Private;
    title "Graduation Rate by University Type (Private vs Public)";
run;

proc sgplot data=college;
    vbox Grad_Rate / category=Elite10;
    title "Graduation Rate by Elite Status";
run;

/* CREATE DUMMY VARIABLES */
```

```
data college2;
    set college;
    d_private = 1;
    if Private = 'No' then d_private = 0;
run;

proc print data=college2 (obs=10);
    title "College Data with Dummy Variables";
run;

/* CORRELATION MATRIX */
proc corr data=college2;
    var d_private Accept_pct Elite10 F_Undergrad P_Undergrad
        Outstate Room_Board Books Personal PhD Terminal
        S_F_Ratio perc_alumni Expend Grad_Rate;
    title "Correlation Matrix";
run;

/* QUESTION (d): FULL MODEL */
proc reg data=college2;
    model Grad_Rate = d_private Accept_pct Elite10 F_Undergrad
        P_Undergrad
                      Outstate Room_Board Books Personal PhD Terminal
                      S_F_Ratio perc_alumni Expend;
    title "FULL MODEL";
run;
quit;

/* QUESTION (e): VIF for Multicollinearity */
proc reg data=college2;
    model Grad_Rate = d_private Accept_pct Elite10 F_Undergrad
        P_Undergrad
                      Outstate Room_Board Books Personal PhD Terminal
                      S_F_Ratio perc_alumni Expend / vif tol;
    title "FULL MODEL with VIF";
run;
quit;

/* QUESTION (f): SELECTION METHODS */
proc reg data=college2;
    model Grad_Rate = d_private Accept_pct Elite10 F_Undergrad
        P_Undergrad
                      Outstate Room_Board Books Personal PhD Terminal
                      S_F_Ratio perc_alumni Expend / selection=stepwise
                          ;
    title "STEPWISE Selection";
run;
quit;

proc reg data=college2;
    model Grad_Rate = d_private Accept_pct Elite10 F_Undergrad
        P_Undergrad
                      Outstate Room_Board Books Personal PhD Terminal
                      S_F_Ratio perc_alumni Expend / selection=backward
                          ;
    title "BACKWARD Elimination";
run;
quit;
```

```
105
106 /* QUESTION (g): FINAL MODEL */
107 proc reg data=college2;
108     model Grad_Rate = d_private Accept_pct Elite10 Outstate Room_Board
109                       Personal PhD Terminal perc_alumni Expend;
110     title "FINAL MODEL M1";
111 run;
112 quit;
113
114 /* QUESTIONS (h), (i), (j): RESIDUAL ANALYSIS */
115 proc reg data=college2;
116     model Grad_Rate = d_private Accept_pct Elite10 Outstate Room_Board
117                       Personal PhD Terminal perc_alumni Expend /
                             influence r;
118     plot student.*predicted.;
119     plot student.*d_private;
120     plot student.*Accept_pct;
121     plot student.*Elite10;
122     plot student.*Outstate;
123     plot student.*Room_Board;
124     plot student.*Personal;
125     plot student.*PhD;
126     plot student.*Terminal;
127     plot student.*perc_alumni;
128     plot student.*Expend;
129     plot npp.*student.;
130     title "Residual Analysis";
131 run;
132 quit;
133
134 /* QUESTION (k) and (l): Standardized Coefficients */
135 proc reg data=college2;
136     model Grad_Rate = d_private Accept_pct Elite10 Outstate Room_Board
137                       Personal PhD Terminal perc_alumni Expend / stb
                             vif;
138     title "Standardized Coefficients";
139 run;
140 quit;
141
142 /* QUESTION (m): PREDICTION */
143 data pred;
144     input d_private Accept_pct Elite10 F_Undergrad P_Undergrad
145           Outstate Room_Board Books Personal PhD Terminal S_F_Ratio
146           perc_alumni Expend Grad_Rate;
147     datalines;
148 1 0.87 0 3000 524 6500 3300 250 1350 40 34 30.2 13 5201 .
149 ;
150 run;
151
152 data college_pred;
153     set pred college2;
154 run;
155
156 proc reg data=college_pred;
157     model Grad_Rate = d_private Accept_pct Elite10 Outstate Room_Board
158                       Personal PhD Terminal perc_alumni Expend / p clm
                             cli alpha=0.05;
159     title "Prediction with Intervals";
```

```
run;
quit;
```

Listing 1: Complete SAS Code for College Graduation Rate Analysis