STAT 223

Laboratory Activity #3 by Akbota Serikkyzy

Exploratory Analysis

We will be considering how the graduation rate (GradRate) is affected by a number of variables, such as Enroll, Outstate SFRatio and Expend.

Enroll: Number of new students enrolled
Outstate: Out-of-state tuition
S.F.Ratio: Student/faculty ratio
Expend: Instructional expenditure per student
Grad.Rate: Graduation rate

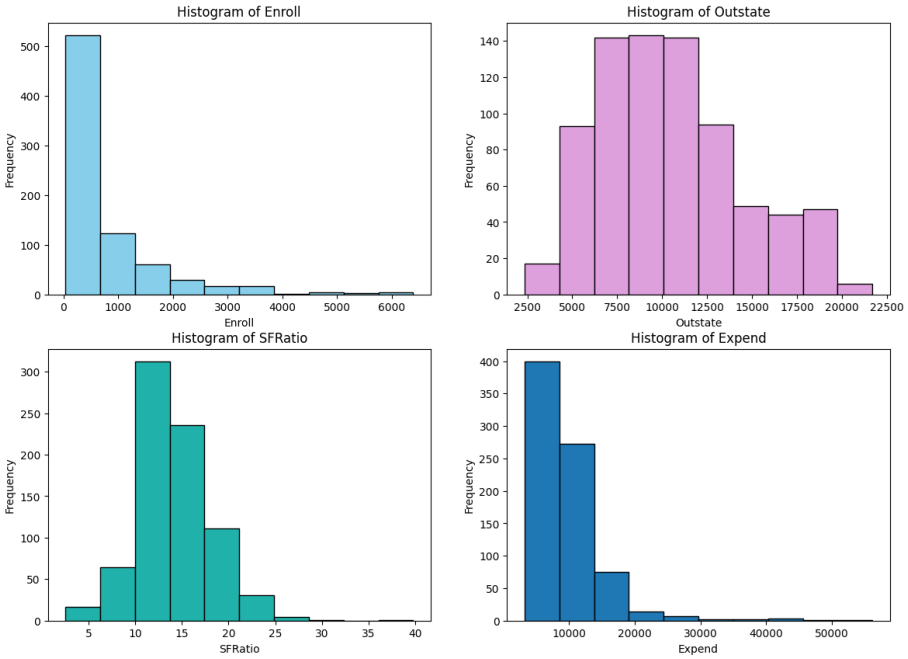

Figure 1: Histograms

Enroll

Statistic	Enroll
Mean	779.97
Median	434.00
Standard Deviation	929.18

Table 1: Descriptive Statistics for Enroll

Looking at table 1, we can see that the standard deviation is quite high, indicating a large variation in the number of students enrolled. Additionally, the mean is much higher than the median, suggesting that the distribution of enrollment is right-skewed. This can also be observed in the histogram in figure 1.

Outstate

Statistic	Outstate
Mean	10440.67
Median	9990.00
Standard Deviation	4023.02

Table 2: Descriptive Statistics for Outstate Tuition

Looking at table 2, we can see that the standard deviation is relatively high, indicating a notable variation in out-of-state tuition costs. The mean is slightly higher than the median, suggesting a slight right-skew in the distribution. This implies that while most values cluster around the median, there are some higher outliers that increase the average.

SFRatio

Statistic	SFRatio
Mean	14.09
Median	13.60
Standard Deviation	3.96

Table 3: Descriptive Statistics for Student-Faculty Ratio

Looking at table 3, we can see that the standard deviation is moderate, indicating some variation in the student-faculty ratios between institutions. The mean is slightly higher than the median, which suggests a slight right skew in the distribution. This could mean that, while many institutions have similar ratios, some have high ratios that are pulling the average up.

Expend

Statistic	Expend
Mean	9660.17
Median	8377.00
Standard Deviation	5221.77

Table 4: Descriptive Statistics for Instructional Expenditure per Student

Looking at table 4, we observe a relatively high standard deviation, indicating considerable variability in instructional expenditure per student across institutions. The mean is noticeably higher than the median, which suggests that the distribution is right-skewed. This implies that while many schools spend around the median amount, some institutions invest significantly more, pulling the average up.

Multicollinearity

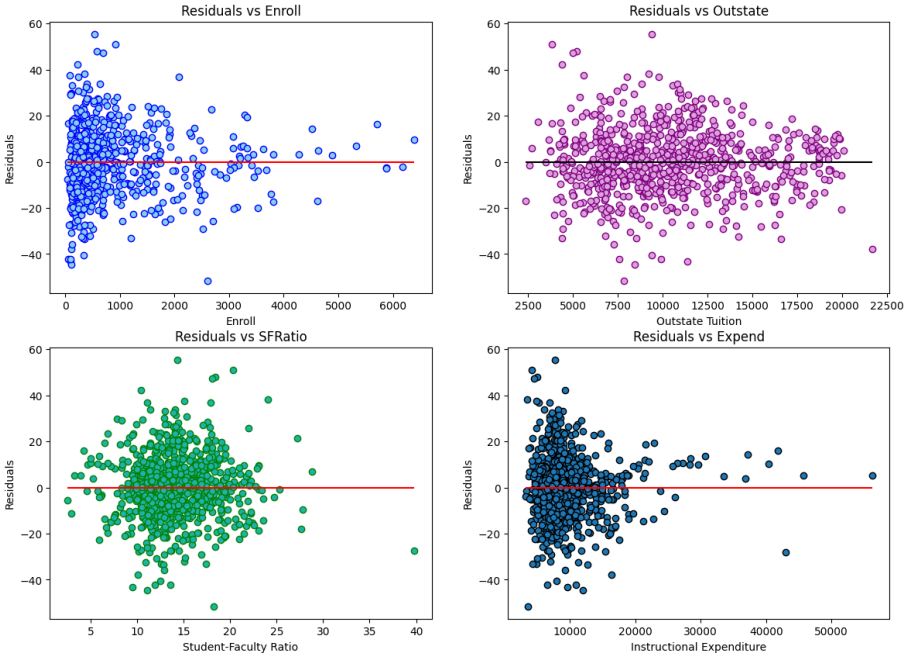

Figure 2: Residulas vs Fitted Values

Looking at figure 2, we can see that Outstate and SFRatio support the assumption of linearity. However, Enroll and Expend show signs of heteroscedasticity, as the spread of their residuals changes across the fitted values.

Variable	VIF
Enroll	1.179941
Outstate	2.049040
SFRatio	1.783476
Expend	2.301899

Table 5: VIF

Looking at table 5, we can see that the VIF values are all around 1–2, which indicates that multicollinearity is not a concern.

Simple Linear Regression

Enroll and GradRate

Variable	Coefficient	Standard Error	p-value
Intercept	65.7855	0.805	0.000
Enroll	-0.0004	0.001	0.534

Table 6: Regression Results for GradRate and Enroll

Statistic	Value
R-squared	0.000
Adjusted R-squared	-0.001
F-statistic	0.3870
Prob (F-statistic)	0.534

Table 7: Model Summary Statistics for Regression

$$\text{The regression equation: } \text{GradRate} = 65.7855 - 0.0004 \times \text{Enroll} \tag{1}$$

As shown in table 6, the enrollment coefficient is -0.0004, suggesting a slight negative relationship between enrollment and the graduation rate. However, this relationship is not statistically significant since p = 0.534, indicating that enrollment does not significantly predict the graduation rate in this model.

To support this, the model summary in table 7 shows an R-squared value of 0.000, which means that the model does not explain the variance in graduation rates. So, there is no evidence of a meaningful linear relationship between enrollment size and graduation rate based on this analysis.

Outstate and GradRate

Variable	Coefficient	Standard Error	p-value
Intercept	39.9951	1.408	0.000
Outstate	0.0024	0.000	0.000

Table 8: Regression Results for GradRate and Outstate Tuition

Statistic	Value
R-squared	0.326
Adjusted R-squared	0.326
F-statistic	375.5
Prob (F-statistic)	1.63e-68

Table 9: Model Summary Statistics for Regression

$$\text{The regression equation: } \text{GradRate} = 39.9951 + 0.0024 \times \text{Outstate} \tag{2}$$

As shown in table 8, the coefficient for out-of-state tuition is 0.0024, indicating a positive relationship between tuition and graduation rate. This relationship is statistically significant with a p-value of 0.000, suggesting that higher out-of-state tuition is associated with higher graduation rates.

To support this, the model summary in table 9 shows an R-squared value of 0.326, which means that approximately 32.6% of the variance in graduation rates can be explained by out-of-state tuition alone.

Overall, this analysis provides strong evidence for a meaningful linear relationship between out-of-state tuition and the graduation rate.

SFRatio and GradRate

Variable	Coefficient	Standard Error	p-value
Intercept	84.2168	2.171	0.000
SFRatio	-1.3310	0.148	0.000

Table 10: Regression Results for GradRate and Student-Faculty Ratio

Statistic	Value
R-squared	0.094
Adjusted R-squared	0.093
F-statistic	80.48
Prob (F-statistic)	2.18e-18

Table 11: Model Summary Statistics for Regression (GradRate and SFRatio)

$$\text{The regression equation: } \text{GradRate} = 84.2168 - 1.3310 \times \text{SFRatio} \tag{3}$$

As shown in table 10, the coefficient for student-faculty ratio is -1.3310, indicating a negative relationship between SFRatio and graduation rate. This relationship is statistically significant with a p-value of 0.000, suggesting that higher student-faculty ratios are associated with lower graduation rates.

The model summary in table 11 shows an R-squared value of 0.094, which means that about 9.4% of the variation in graduation rates can be explained by student-faculty ratio alone.

Overall, this analysis shows a statistically significant, though relatively weak, negative linear relationship between student-faculty ratio and graduation rate.

Expend and GradRate

Variable	Coefficient	Standard Error	p-value
Intercept	53.0588	1.194	0.000
Expend	0.0013	0.000	0.000

Table 12: Regression Results for GradRate and Expend

Statistic	Value
R-squared	0.152
Adjusted R-squared	0.151
F-statistic	139.3
Prob (F-statistic)	1.10e-29

Table 13: Model Summary Statistics for Regression (GradRate and Expend)

$$\text{The regression equation: } \text{GradRate} = 53.0588 + 0.0013 \times \text{Expend} \tag{4}$$

As shown in table 12, the coefficient for instructional expenditure per student is 0.0013, indicating a positive relationship between expenditure and graduation rate. This relationship is statistically significant with a p-value of 0.000, suggesting that higher instructional spending is associated with higher graduation rates.

To support this, the model summary in table 13 shows an R-squared value of 0.152, which means that approximately 15.2% of the variance in graduation rates can be explained by instructional expenditure alone.

Overall, this analysis suggests a meaningful and moderate linear relationship between instructional spending and graduation rate.

Multiple Linear Regression

GradRate with Expend, Outstate, SFRatio and Expend

Variable	Coefficient	Standard Error	p-value
Intercept	38.8918	3.565	0.000
Enroll	0.0013	0.001	0.027
Outstate	0.0025	0.000	0.000
SFRatio	-0.0197	0.171	0.908
Expend	-4.179e-05	0.000	0.776

Table 14: Regression Results for GradRate on Enroll, Outstate, SFRatio, and Expend

Statistic	Value
R-squared	0.331
Adjusted R-squared	0.328
F-statistic	95.48
Prob (F-statistic)	5.39e-66

Table 15: Model Summary Statistics for Regression (GradRate = Enroll + Outstate + SFRatio + Expend)

$$\text{GradRate} = 38.8918 + 0.0013 \times \text{Enroll} + 0.0025 \times \text{Outstate} - 0.0197 \times \text{SFRatio} - 4.179 \times 10^{-5} \times \text{Expend} \tag{5}$$

This model, made by combining four variables such as Enroll, Outstate, SFRatio, and Expend which aims to explain the variation in graduation rates using multiple linear regression. As shown in table 14, the coefficients for Enroll and Outstate are both positive and statistically significant, with p-values of 0.027 and 0.000 respectively. This suggests that higher enrollment and out-of-state tuition are associated with higher graduation rates.

In contrast, the coefficients for SFRatio and Expend are not statistically significant, with high p-values of 0.908 and 0.776. This indicates that, in the presence of the other variables, student-faculty ratio and instructional expenditure per student do not have a strong effect on graduation rate.

According to the model summary in table 15, the R-squared value is 0.331, meaning that about 33.1% of the variance in graduation rates is explained by the combination of these four predictors.

Overall, this multiple linear regression model provides a better fit than the simple linear regressions, highlighting the stronger combined explanatory power of Enroll and Outstate in predicting graduation rates.

GradRate with Enroll and Outstate

Variable	Coefficient	Standard Error	p-value
Intercept	38.5405	1.542	0.000
Enroll	0.0013	0.001	0.022
Outstate	0.0025	0.000	0.000

Table 16: Regression Results for GradRate on Enroll and Outstate

Statistic	Value
R-squared	0.331
Adjusted R-squared	0.329
F-statistic	191.4
Prob (F-statistic)	2.92e-68

Table 17: Model Summary Statistics for Regression (GradRate = Enroll + Outstate)

$$\text{GradRate} = 38.5405 + 0.0013 \times \text{Enroll} + 0.0025 \times \text{Outstate} \tag{6}$$

Since SFRatio and Expend were not statistically significant in the previous multiple regression model, they were removed to simplify the model. This new regression includes only Enroll and Outstate as predictors of graduation rate. As shown in table 16, both variables remain statistically significant with p-values of 0.022 and 0.000, respectively, indicating that they still contribute meaningfully to explaining graduation rates.

The model summary in table 17 shows an R-squared value of 0.331, which is nearly identical to the previous model that included all four variables. This suggests that removing SFRatio and Expend did not significantly reduce the model's significance.

Overall, this result indicates that Enroll and Outstate are sufficient predictors of graduation rate, and a simpler model excluding SFRatio and Expend can still provide reliable predictions.

GradRate with log(Enroll) and Outstate

$$\text{GradRate} = 26.7453 + 2.0688 \cdot \log(\text{Enroll}) + 0.0025 \cdot \text{Outstate}$$

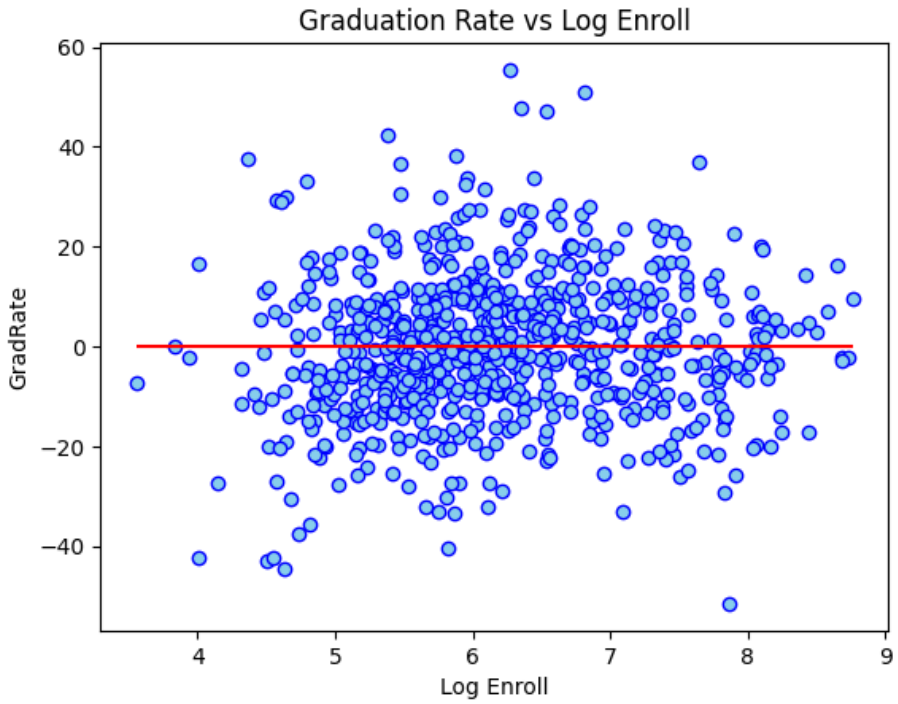

Figure 3: Residulas

Variable	Coefficient	Standard Error	p-value
Intercept	26.7453	3.662	0.000
log_Enroll	2.0688	0.529	0.000
Outstate	0.0025	0.000	0.000

Table 18: Regression Results for GradRate on log(Enroll) and Outstate

Statistic	Value
R-squared	0.339
Adjusted R-squared	0.338
F-statistic	198.9
Prob (F-statistic)	2.02e-70

Table 19: Model Summary Statistics for Regression (GradRate = log(Enroll) + Outstate)

In the beginning of exploration of the data, Enroll showed signs of heteroscedasticity. To address this, I applied a log transformation to the Enroll variable. In the revised model using log(Enroll) (figure 3), the residuals appeared more evenly spread in the residual, suggesting improved homoscedasticity and a better fit.

We can also observe slight improvements in the regression statistics. The R-squared value increased from 0.331 to 0.339, and the adjusted R-squared rose from 0.329 to 0.338, indicating a slightly better explanation power. Additionally, both log(Enroll) and Outstate remain statistically significant predictors of GradRate, with p-values being close to 0.

However, it is important to note that while the log transformation of Enroll contributes to addressing heteroscedasticity and enhancing interpretability, it comes with a trade-off. The standard error for log(Enroll) increased from 0.001 to 0.529. This indicates a loss in precision for that predictor. So, the choice between using Enroll or log(Enroll) depends on what we value more whether a marginal gain in model fit or greater precision.

Example

Suppose a relative of mine is looking at a college with 1,000 new student enrolled, an out-state tuition of $11,000, a student-to-faculty ratio of 12 and an instructional expenditure per student of $9,200. What would you predict the graduation rate of that school to be? What are the confidence and prediction intervals?

Model	Predictors	Predicted GradRate	Standard Error	95% Confidence Interval
Model 1	Expend, Outstate, SFRatio	67.22%	0.681	(65.88%, 68.56%)
Model 2	Enroll, Outstate	67.13%	0.526	(66.10%, 68.16%)
Model 3	log(Enroll), Outstate	68.37%	0.642	(67.11%, 69.63%)
Model 4	Outstate	66.83%	0.511	(65.82%, 67.83%)

Table 20: Predicted Graduation Rates with Standard Errors and Confidence Intervals

Model	Predictors	95% Prediction Interval
Model 1	Expend, Outstate, SFRatio	(39.53%, 94.90%)
Model 2	Enroll, Outstate	(39.49%, 94.77%)
Model 3	log(Enroll), Outstate	(40.90%, 95.84%)
Model 4	Outstate	(39.12%, 94.54%)

Table 21: Prediction Intervals for Graduation Rates

All four models predict a graduation rate in the narrow range of 66% to 68%, with overlapping confidence and prediction intervals. This suggests that all models perform similarly in terms of predictive power. Given this similarity, Model 1 has more predictors and therefore adds more complexity without offering a meaningful improvement in accuracy.

Model 2 and Model 3, which use fewer predictors, provide nearly identical performance while being more interpretable and simpler to implement. While Model 3 have a slightly higher predicted mean, Model 2 has a smaller standard error, offering more precise estimates. Overall, the differences between these two models are minimal, and they perform similarly on the data.

Additionally, Model 4, which uses only Outstate as a predictor, shows that Outstate tuition alone explains much of the variation in graduation rates. The predicted graduation rate, standard error, and intervals from Model 4 are remarkably close to those of the other models, suggesting that Outstate is the dominant predictor. While Enroll adds some additional explanatory power, the marginal improvement is relatively small.

Therefore, if model simplicity is the top priority, Model 4 is a strong candidate. It achieves comparable predictive performance with the least complexity and the lowest standard error, making it useful.

Conclusion

In this analysis, I explored the relationship between a college's graduation rate and several key explanatory variables: enrollment size, out-of-state tuition, student-to-faculty ratio, and instructional expenditure per student. Using three different linear regression models, we consistently predicted graduation rates in range of 66% to 68%, indicating that all models performed similarly in terms of predictive accuracy.

Our results suggest that out-of-state tuition is a strong predictor of graduation rate across all models. Additionally, enrollment, particularly when log-transformed, also proved to be statistically significant and improved model fit slightly by addressing heteroscedasticity. However, this came with a trade-off in the form of a higher standard error, making the coefficient estimate less precise. Enroll appears insignificant on its own, and even shows a negative relationship. But when Outstate is added to the model, Enroll becomes weakly significant and flips positive. This suggests Outstate is a strong predictor, and including it helps clarify the small positive effect of enrollment. Still, most of the model's explanatory power is by Outstate.

The inclusion of additional variables like student-to-faculty ratio and instructional expenditure did not lead to a meaningful improvement in prediction, which suggests that simpler models with fewer predictors can perform just as well in terms of interpretability and generalization.

Overall, the analysis indicates that graduation rates tend to be positively associated with higher out-of-state tuition. However, this reflects a correlation, so it does not not necessarily a causal relationship since other unobserved factors may also influence graduation rates.