College Graduation Rate Prediction

Syed Faizan

Main Tool: R Programming Language

Technique: Regularization

Industry: Education

Introduction

In this project, we delve into the practical implementation of Ridge and LASSO (Least absolute Shrinkage and Selection Operator) regression methodologies using the **glmnet()** package in R. These techniques, characterized by their ability to mitigate multicollinearity and perform variable selection, are paramount in refining theoretical knowledge into practical skills with real-world applicability.

For this project we use the 'College' dataset is from the ISLR package that accompanies the classic text book 'Introduction to Statistical Learning (ISL) with applications in R' from Stanford professors Trevor Hastie and Robert Tibshirani (James, Witten, Hastie, & Tibshirani, 2013) who were also pioneers in the development of the modern statistics. It is also important to note that Robert Tibshirani is the inventor of the LASSO methodology. (Tibshirani, R. (1996).)

The College dataset contains data pertaining to US Colleges from the 1995 issue of US News and World Report. There are no missing values in the data set.

Brief Descriptive Statistics

Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Private	777						
... No	212	27%					
... Yes	565	73%					
Apps	777	3002	3870	81	776	3624	48094
Accept	777	2019	2451	72	604	2424	26330
Enroll	777	780	929	35	242	902	6392
Top10perc	777	28	18	1	15	35	96
Top25perc	777	56	20	9	41	69	100
F.Undergrad	777	3700	4850	139	992	4005	31643
P.Undergrad	777	855	1522	1	95	967	21836
Outstate	777	10441	4023	2340	7320	12925	21700
Room.Board	777	4358	1097	1780	3597	5050	8124
Books	777	549	165	96	470	600	2340
Personal	777	1341	677	250	850	1700	6800
PhD	777	73	16	8	62	85	103
Terminal	777	80	15	24	71	92	100
S.F.Ratio	777	14	4	2.5	12	16	40
perc.alumni	777	23	12	0	13	31	64

Figure 1 Descriptive statistics for the College Dataset

Graduation Rate ('Grad.Rate')

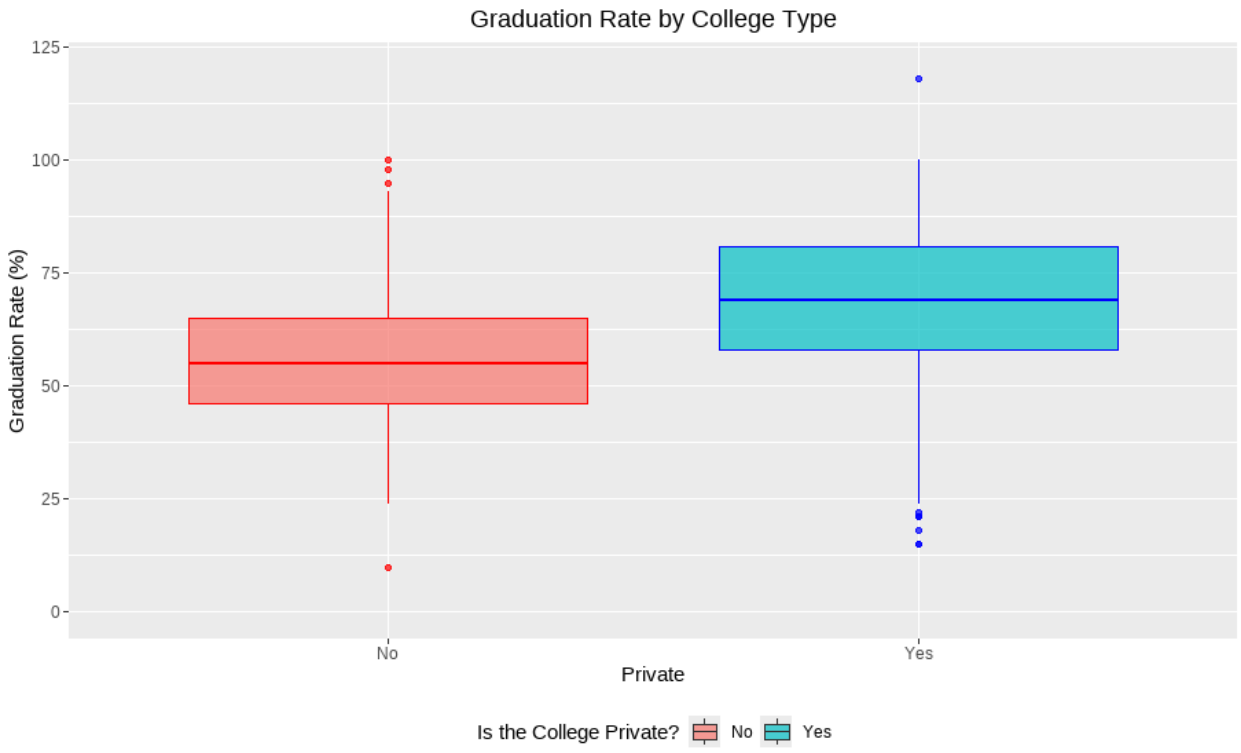

Figure 2 Box Plot depicting Graduation Rates across College Types

```
   Grad.Rate
Min.   : 10.00
1st Qu.: 53.00
Median : 65.00
Mean   : 65.46
3rd Qu.: 78.00
Max.   :118.00
```

Figure 3 Tuckey's Five Number Summary of the response variable 'Graduation Rate'.

The box plot illustrates the distribution of graduation rates across private and public colleges, which ought to be interpreted in conjunction with a five-number summary derived from the data. The graduation rates for private colleges are notably higher, as evidenced by the distribution's central tendency and dispersion. The minimum graduation rate for public colleges is a paltry 10.00%, which is significantly lower than the minimum for private colleges, indicating a potentially wide gulf in the quality of education or students at these two types of institutions. The first quartile for public colleges is 53.00%, suggesting that 25% of these colleges have a graduation rate below this percentage. Conversely, the median graduation rate of 65.00% represents the midpoint of the dataset, showing no difference between the median values of private and public colleges. The third quartile for private colleges is 78.00%, indicating that 75% of these colleges have graduation rates below this threshold. Lastly, the maximum graduation rate for public

colleges exceptionally exceeds 100%, listed as 118.00%, indicating potential data errors that warrant further investigations. Taken as a whole these statistics underline a broader range of graduation outcomes in public institutions compared to private ones, which display a tighter, albeit higher, range of graduation rates.

Splitting the data into a train and test set

Train 545 records totally (70%)	Test 232 records totally (30%)
Private- 396 records	Private- 169 records
Public – 149 records	Public- 63 records

The College dataset, comprising 777 records, was divided into training and test subsets for model development and validation. About 70% of the data (545 records) was allocated to the training set, which includes 396 records from private colleges and 149 from public colleges. This division ensures adequate learning and pattern recognition.

The remaining 30% (232 records) forms the test set, used for evaluating the model's generalizability. This subset consists of 169 private college records and 63 public college records, maintaining a balanced representation of both institution types for unbiased validation.

This split methodology supports robust training and reliable model evaluation, ensuring the subsets are representative and distinct, without overlapping records.

A note on the inclusion of a categorical variable in the ridge, LASSO and the stepwise regression

In the following Ridge, LASSO and the stepwise regression I have included the following predictors for Graduation Rate:

- **Private** (Indicator of whether an institution is private or not),
- **Apps** (Number of applications received),
- **Accept** (Number of applications accepted),
- **Enroll** (Number of new students enrolled),
- **Top10perc** (Percentage of new students from top 10% of their high school class),
- **Top25perc** (Percentage of new students from top 25% of their high school class),
- **F.Undergrad** (Number of full-time undergraduates),
- **P.Undergrad** (Number of part-time undergraduates),

- **Outstate** (Tuition for out-of-state students),
- **Room.Board** (Cost of room and board),
- **Books** (Cost of books),
- **Personal** (Estimated personal spending),
- **PhD** (Percentage of faculty with PhDs),
- **Terminal** (Percentage of faculty with terminal degree),
- **S.F.Ratio** (Student-to-faculty ratio),
- **perc_alumni** (Percentage of alumni who donate).

It may be noted that I have included, along with all the numerical variables, the sole categorical variable in the dataset also in the analysis. This is owing to the explicit mention of inclusion of the categorical variable in our project lab video. The instructor mentions how the glmnet() function (as also the lm() function) has the innate property of using dummy coding to incorporate categorical variables into regression modelling. The instructor mentions how this property must be put to use in our project to practice LASSO (and presumably ridge) modelling. Also, the inventors of the LASSO method, namely Rob Tibshirani and Trevor Hastie in their seminal book Introduction to Statistical Learning also use the 'Hitters' dataset in the ISLR package to demonstrate the ridge and LASSO regression while including the categorical variables also in their analysis. As far the last stepwise regression is concerned I have persisted with the inclusion of the categorical variable in order to render the models amenable to comparison which is a core and final task of our present project.

Ridge Regression

```
> train_x <- model.matrix(Grad.Rate ~ ., data = train)[, -1]
> train_y <- train$Grad.Rate
> test_x <- model.matrix(Grad.Rate ~ ., data = test)[, -1]
> test_y <- test$Grad.Rate
> ?glmnet()
> cv.ridge <- cv.glmnet(x = train_x, y = train_y, alpha = 0, standardize = TRUE) #standradizing the predictors
> bestlam_ridge <- cv.ridge$lambda.min
> bestlam_1se_ridge <- cv.ridge$lambda.1se
> ?compare()
> bestlam_ridge
[1] 2.671623
>
> bestlam_1se_ridge
[1] 20.68541
```

Figure 4 The Lambda associated with the most regularized model and Lambda at the First Standard Error of this best regularized model

The above output demonstrates the application of the **cv.glmnet** function for performing a ten-fold cross-validation in a ridge regression model to estimate the optimal values of lambda (**lambda.min** and **lambda.1se**). In cross-validation each part of the dataset , 10 in our case, serves both as testing and training data against the 9 other parts of the dataset.

1. **Estimation of Lambda Values**:
    - **lambda.min** is the value of lambda that results in the minimum mean cross-validated error. It is 2.671623 in our ridge regression cross validation . This value suggests the lambda that provides the best fit of the model to the data, minimizing prediction error.
    - **lambda.1se** is the largest value of lambda such that the error is within one standard error of the minimum. It is 20.68541 in our cross validation. This lambda may provide a more regularized model, that trades off a slight increase in bias for better generalization and reduced variance.
2. **Comparison and Discussion**:
    - The **lambda.min** and **lambda.1se** differ significantly, with **lambda.1se** being approximately 7.7 times larger than **lambda.min**. This indicates a substantial difference in model complexity and regularization strength between the two models.
    - Using **lambda.min** typically leads to a model that may fit the training data very well but could potentially overfit, especially when the number of predictors is large or the data are highly variable.
    - On the other hand, using **lambda.1se** typically results in a simpler model with possibly fewer predictors included. This model might not fit the training data as well as the **lambda.min** model but is expected to perform better on unseen data due to its simplicity and robustness against overfitting.

In conclusion, the choice between **lambda.min** and **lambda.1se** depends on the balance one wishes to achieve between bias and variance. **lambda.1se** is often preferred in practice, especially in LASSO regression, for its greater emphasis on model robustness and reliability. However, in our case I chose to proceed with the ridge regression analysis with the minimum lambda (**lambda.min**) that represents the lambda value for which the Mean Square Error (**MSE**) is at a minimum for two reasons:

1. Firstly, as our model was a ridge regression (in the first part of the project) and therefore elimination of predictors ('**sparsity**') was not anticipated with increasing the lambda (the tuning hyperparameter) I settled for a more conservative modelling pattern.
2. Secondly, as the difference beyween the lambda with minimum MSE (**lambda.min**)and the lambda at one standard error of the MSE (**lambda.1se**) was substantial I sought to first examine if overfitting would actually be an issue with my ridge regression model before venturing to try the lambda.1se value. I thus avoided pre-emptive adjustment for a problem that I was not yet sure I would even confront. If overfitting it turns out is not an issue, as evidenced by RMSE difference between the training and the test daatset , then my choice of the more conservative lamda value will be justified.

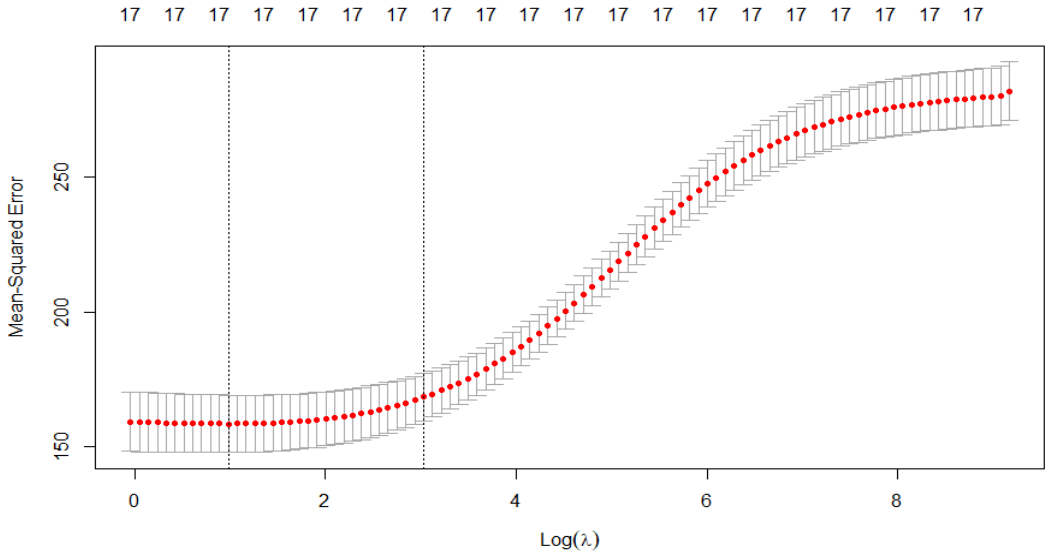

Figure 5 Log of Lambda plotted against the MSE for the ridge regression model cross validation.

The above graph represents a visual output from the **cv.glmnet** function, depicting the relationship between the logarithm of lambda (Log(λ)) and the mean squared error (MSE) during the cross-validation process of a ridge regression model. This plot is instrumental in identifying the optimal regularization parameter (λ) that minimizes prediction error while preventing overfitting.

Graph Interpretation:

1. **Mean Squared Error Trend**:
    - The MSE is plotted against Log(λ) with error bars representing the standard deviation of the MSE across different folds of the cross-validation.
    - As Log(λ) increases, the MSE initially decreases, reaching a minimum before gradually increasing again. This typical behavior highlights the trade-off between bias and variance in the model, where lower values of λ lead to lower bias but potentially higher variance, and vice versa.
2. **Optimal Lambda Values**:
    - The vertical dashed lines on the graph mark the **lambda.min** and **lambda.1se**:
        - **lambda.min** (2.671623 converted to a Log), indicated by the first vertical line, corresponds to the lowest MSE observed. This value of λ offers the least bias without undue variance.
        - **lambda.1se** (20.68541 converted to a Log), shown by the second vertical line, is the most regularized model within one standard error of the

minimum MSE. Choosing this λ may provide a more robust model against overfitting.

3. **Interpretation and Implications**:

    - This plot is crucial for understanding the balance between complexity and generalization in ridge regression. Lower values of λ (near **lambda.min**) might fit the training data better but could overfit, especially when the predictor space is high-dimensional.
    - Higher values of λ (near **lambda.1se**) enhance the model's generalization ability on unseen data by introducing more regularization, thus reducing the model's complexity and the risk of overfitting.
    - The choice between **lambda.min** and **lambda.1se** is strategic, depending on whether the focus is on achieving the lowest possible error on the training set (**lambda.min**) or on optimizing the model's performance on new, unseen data (**lambda.1se**). It is often a trade-off and the final decision rests upon the priorities of the analyst. In our subsequent risdge regression I have chosen to err on the side of conservatism and adopted a frugal approach to the analysis by utilizing the lambda.min as the tuning parameter value. Later adjustments may be made if overfitting is suspected.

```
> dim(coef.ridge)
[1] 18  1
> coef.ridge
18 x 1 sparse Matrix of class "dgCMatrix"
                       s0
(Intercept) 32.8542138528
PrivateYes   4.4991642548
Apps         0.0005153266
Accept       0.0001961146
Enroll       0.0009694149
Top10perc    0.0957362192
Top25perc    0.1141506885
F.Undergrad -0.0000765803
P.Undergrad -0.0013947857
Outstate     0.0006831873
Room.Board   0.0026353969
Books       -0.0002251386
Personal    -0.0018901139
PhD          0.0505297332
Terminal    -0.0579998436
S.F.Ratio    0.0848188777
perc.alumni  0.2291285233
Expend      -0.0002661944
```

Figure 6 A 'sparse matrix', as the glmnet object is called, showing the co-efficients of the ridge regression model.

The above reported coefficients derived from fitting a ridge regression model to predict college graduation rates provide valuable insights into the features that most significantly influence this outcome. The lambda value used was 2.671623 corresponding to the lambda.min or the lambda for which the MSE is reduced the most in the course of regularization. These coefficients are extracted from a trained model that included a variety of explanatory variables, each representing different facets of college characteristics.

Model Coefficients Interpretation:

1. **Non-Zero Coefficients**:

    - Ridge regression, characterized by its shrinkage mechanism, generally does not reduce coefficients to absolute zero but rather shrinks them towards zero. This property is evident from the sparse matrix, where all coefficients remain non-zero. In this regard LASSO offeres a huge advantage over ridge regression because it effectively serves as a feature selection tool as well, as we shall see later in this report.

      Significant Predictors:

      "PrivateYes" (4.4991): A college being private is associated with an increase of approximately 4.50 percentage points in the graduation rate compared to public colleges. This substantial effect underscores the advantage that private institutions

may have, potentially due to factors like better resources, smaller class sizes, or more selective admissions processes.

Top10perc (0.0957): An increase of one percentage point in the proportion of students from the top 10% of their high school classes is associated with a graduation rate increase of approximately 0.096 percentage points. This reflects the positive impact of having academically excellent students on overall graduation outcomes.

Top25perc (0.1145): Similarly, a one percentage point increase in the proportion of students from the top 25% of their high school classes correlates with an increase of about 0.115 percentage points in the graduation rate. This suggests that broader academic excellence within the student body enhances graduation rates.

PhD (0.0505): Each one percentage point increase in the proportion of faculty members with PhDs is linked to a 0.051 percentage point increase in graduation rates. This indicates the positive role of highly qualified faculty in improving educational outcomes.

Terminal (0.0579): For each percentage point increase in the proportion of faculty with terminal degrees, there is an approximate 0.058 percentage point increase in graduation rates. This suggests that the expertise of faculty with the highest level of education in their fields positively impacts student success.

S.F.Ratio (0.0841): A one unit increase in the student-faculty ratio correlates with an increase of approximately 0.084 percentage points in graduation rates. Although this finding seems counterintuitive, as a higher ratio usually implies less individual attention for students, it may reflect underlying factors not directly captured by this metric, such as larger class sizes in programs that are more efficient or have better structured curricula.

Negative Coefficients:

Personal (-0.0080): Each unit increase in personal expenses is associated with a decrease of 0.008 percentage points in the graduation rate. This could indicate that higher personal expenses place a financial strain on students, potentially affecting their academic performance and ability to graduate.

Terminal (-0.0579): This negative coefficient is intriguing as it contrasts with another positive coefficient for PhD, suggesting a complex interaction. It might imply that in certain contexts, having a higher proportion of faculty with terminal degrees does not translate into better graduation rates or perhaps milticollinearity and confounding are playing a role in this counterintuitive finding.

Is multicollinearity playing a role in this? **We briefly examine this question with correlation analysis below.**

Correlation Analysis of the College Dataset

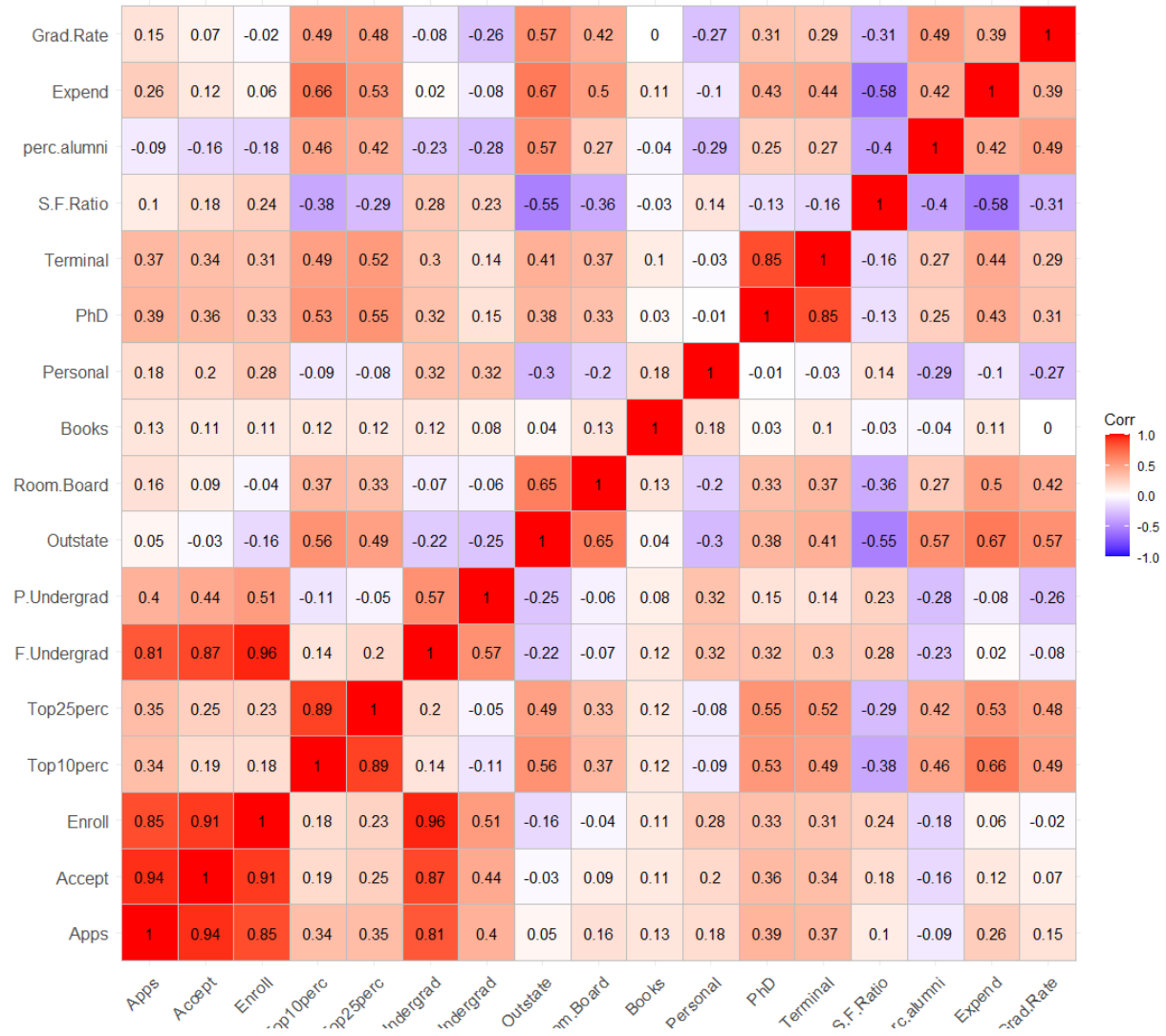

Figure 7 A correlation plot of the whole college dataset.

We briefly digress from the streamlined step-wise solution of the project tasks to carry out a brief correlation analysis of the variables in the dataset to enable a better understanding of the role multicollinearity or confounding might be playing in our model.

The correlation matrix displayed above provides a detailed visualization of the relationships between various predictors in the college dataset. Notably, several high correlations between predictors suggest potential issues with multicollinearity and confounding, which are pertinent when interpreting the coefficients obtained from the ridge regression model.

Notable High Correlations:

1. **Applications and Acceptances (0.94)**: This high correlation is intuitive as colleges that receive more applications usually also accept more students, albeit this does not necessarily mean a higher acceptance rate.
2. **Enrollment and Acceptances (0.85)**: Similarly, the number of acceptances is strongly correlated with the actual enrollment numbers, indicating that higher acceptances often translate to higher enrollments.
3. **Full-time Undergraduates and Enrollment (1.00)**: This perfect correlation indicates redundancy between these two variables, as full-time undergraduate numbers are a direct measure of enrollment.
4. **Room and Board with Out-of-State Tuition (0.65)**: This suggests that colleges charging higher out-of-state tuition also tend to have higher room and board charges, possibly reflecting overall cost structures of more expensive institutions.
5. **PhD and Terminal Degrees (0.85)**: A high correlation exists between the proportion of faculty holding PhDs and those with terminal degrees, suggesting overlap in these qualifications.

Multicollinearity and Confounding Effects:

Multicollinearity refers to the high intercorrelations among predictor variables in a regression model, which can distort the true relationship between predictors and the response variables, inflate the variances of the estimated coefficients, and make the model unstable and difficult to interpret. From the correlation matrix, the high correlations among several predictors such as Applications, Acceptances, Enrollment, and Full-time Undergraduates indicate strong multicollinearity.

Implications in Ridge Regression Model:

- **Coefficient Signs and Magnitudes**: The multicollinearity might explain some of the unexpected signs or magnitudes of coefficients in the ridge regression model. For instance, the coefficient for S.F.Ratio being positive might be confounded by its correlations with other predictors like PhD or Terminal, which are also related to institutional quality and resource availability.
- **Bias and Variance**: While ridge regression helps reduce the variance of the coefficients by imposing a penalty proportional to their size, it does not eliminate multicollinearity. The coefficients might still be biased due to the high correlation among predictors.

Confounding Effects:

- Variables like Room and Board and Out-of-State Tuition might act as confounders, influencing both the response variable (Graduation Rate) and other predictors, skewing their relationships. For example, higher education costs might be associated both with institutional prestige (which could increase graduation rates) and with financial barriers (which could decrease graduation rates).

A summary of the ridge regression model:

The ridge regression model's coefficients highlight several expected and unexpected associations between college characteristics and graduation rates. The model reaffirms the importance of student quality (Top10perc, Top25perc) and faculty qualifications (PhD positive impact) in influencing graduation outcomes. Interestingly, some variables, such as the student-faculty ratio and personal expenses, exhibit counterintuitive signs, which warranted further scrutiny of the dataset through correlation analysis. These insights provide a basis for policy recommendations, particularly around enhancing student academic preparedness and financial support strategies to improve graduation rates.

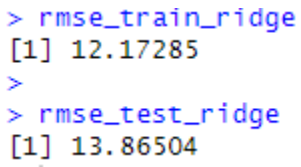

Figure 8 RMSE of the training and the testing using the ridge model

The performance of the ridge regression model in predicting college graduation rates was evaluated using the Root Mean Square Error (RMSE), a metric that quantifies the average magnitude of the prediction error. The RMSE provides a clear measure of how accurately the model predicts the response variable, with lower values indicating better fit.

RMSE Values:

- **Training RMSE**: 12.17285
- **Testing RMSE**: 13.86504

Analysis of RMSE Values:

- The RMSE for the training set (12.17285) is lower than that for the test set (13.86504). This indicates that the model performs slightly better on the training data compared to the test data.
- The difference in RMSE between the training and test datasets is 1.69219. This discrepancy suggests that while there is some difference in performance, it is relatively modest.

Evaluation of Model Overfitting:

- **Overfitting**: A model is considered overfit when it performs well on the training data but significantly worse on unseen (test) data. This typically occurs when a model is excessively complex, capturing noise in the training data that does not generalize to the test data.
- **Current Model**: The increase in RMSE from the training set to the test set in this scenario does suggest a slight overfitting. However, the magnitude of the difference is not substantial. A change of approximately 1.69 points in RMSE indicates that while the model has learned specific features from the training data, it retains a reasonable degree of generalization to new data.
- **Interpretation**: Given the nature of ridge regression, which includes regularization to control for overfitting by shrinking coefficients, the model is expected to be less prone to overfitting compared to models without regularization. The small difference in RMSE supports the conclusion that the model is slightly overfit but not severely so.

Conclusion:

The ridge regression model demonstrates good performance on both the training and test datasets with only a slight indication of overfitting. The RMSE values suggest that the model captures the underlying trends in the data effectively while maintaining an acceptable level of generalization. This balance is crucial for predictive models intended for practical application in educational settings. Future iterations of the model could explore adjustments in the regularization strength or incorporate cross-validation techniques to further enhance model robustness and reduce the potential for overfitting.

Plots to better understand the Ridge Regression model

I put a unique property of the glmnet() function to use in order to generate explicatory plots that demonstrate the process of ridge regression visually. In the glmnet() function, when fitting a model (such as ridge regression with alpha = 0) and lambda is not explicitly specified, **the function automatically generates a sequence of lambda values based on the input data.** The function generates 100 lambda values by default, logarithmically spaced between lambda.max , where the co-efficients are reduced to well-nigh zero and and lambda.min , which by default, glmnet() sets as lambda.max/1000 for ridge regression. This smaller lambda value allows the function to explore solutions from the most regularized model (all coefficients zero) to a model that is almost equivalent to a non-regularized regression. I used this autogeneration of lambda

values by the ridge regression to generate the following plots that offer a virtual snapshot of the ridge regression procedure.

The below plots graphically represent the impact of the regularization parameter, lambda (λ), on the coefficients of a ridge regression model and the variance explained by the model across different values of λ. These visualizations are pivotal for understanding the dynamics of regularization in ridge regression, specifically how it influences model coefficients over different regularization strengths.

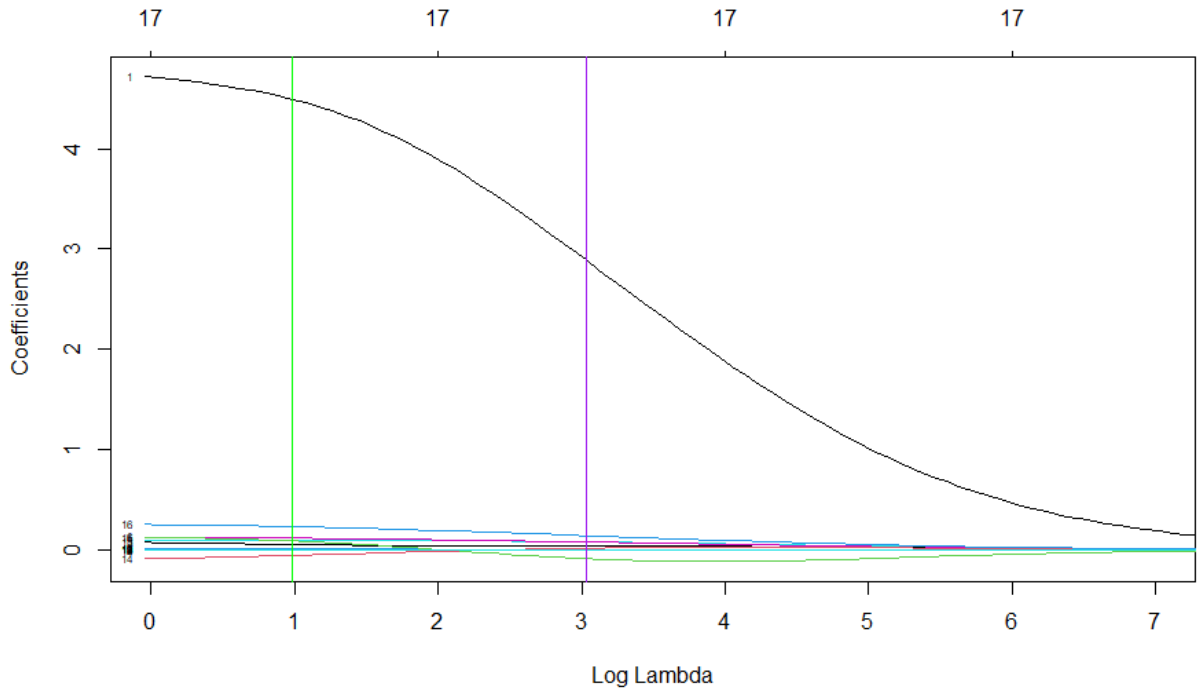

Figure 9 Co-efficients as a function of log (lambda)

Coefficient Path as a Function of Log(λ):

- The above plot displays the trajectories of the coefficients of predictor variables as lambda increases on a logarithmic scale.
- **Key Observations**:
  - As λ increases (moving right along the x-axis), all coefficients gradually shrink towards zero, reflecting the impact of increasing regularization.
  - Coefficients such as those for "PrivateYes" and "PhD" demonstrate slower rates of decline, indicating their substantial influence on the model. These coefficients are

more resistant to shrinkage, suggesting strong relationships with the response variable.

- I added a **green vertical line** marking **lambda.min** using the abline() function in base R. This gives the value of λ corresponding to the minimum cross-validated error. Similarly, the **purple vertical line** represents **lambda.1se**, a more conservative λ value within one standard error of the minimum. These lines demarcate significant points where the model transitions from less regularized (left of green line) to more regularized (right of purple line), helping to evaluate the model's performance and overfitting characteristics.

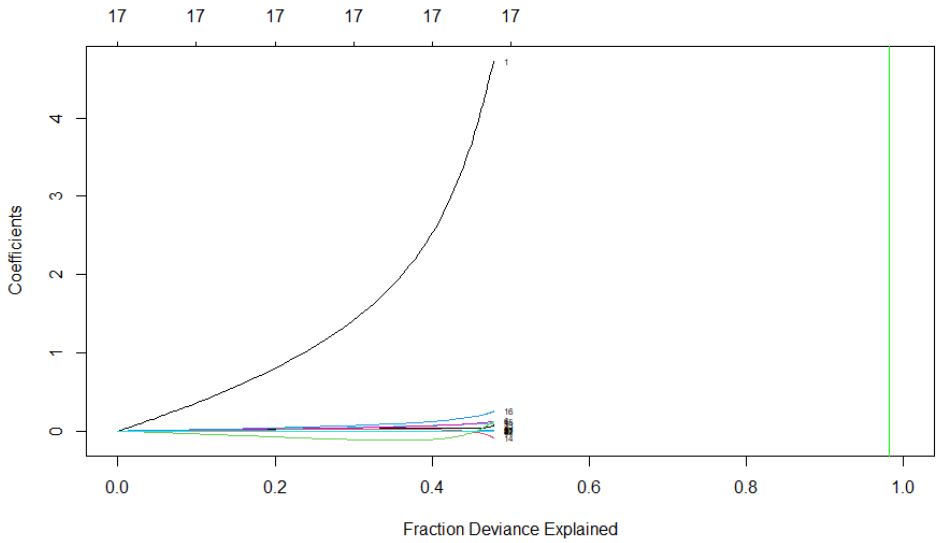

Figure 10 Fraction Deviance explained by the coefficients

Coefficient Path as a Function of Fraction Deviance Explained:

- The above plot maps the coefficient paths against the fraction of deviance explained by the model.
- **Key Observations**:
  - The curve sharply increases as the fraction of deviance explained approaches 1.0, indicating that most of the model's predictive power is concentrated at lower levels of regularization (i.e., smaller λ values).
  - This plot emphasizes the transition from underfitting to optimal fitting as more deviance is explained, with a threshold beyond which further reduction in λ fails to significantly enhance model performance.

- The presence of the green line (associated with **lambda.min**) near the peak of explained variance underscores where the model optimally balances complexity and fit.

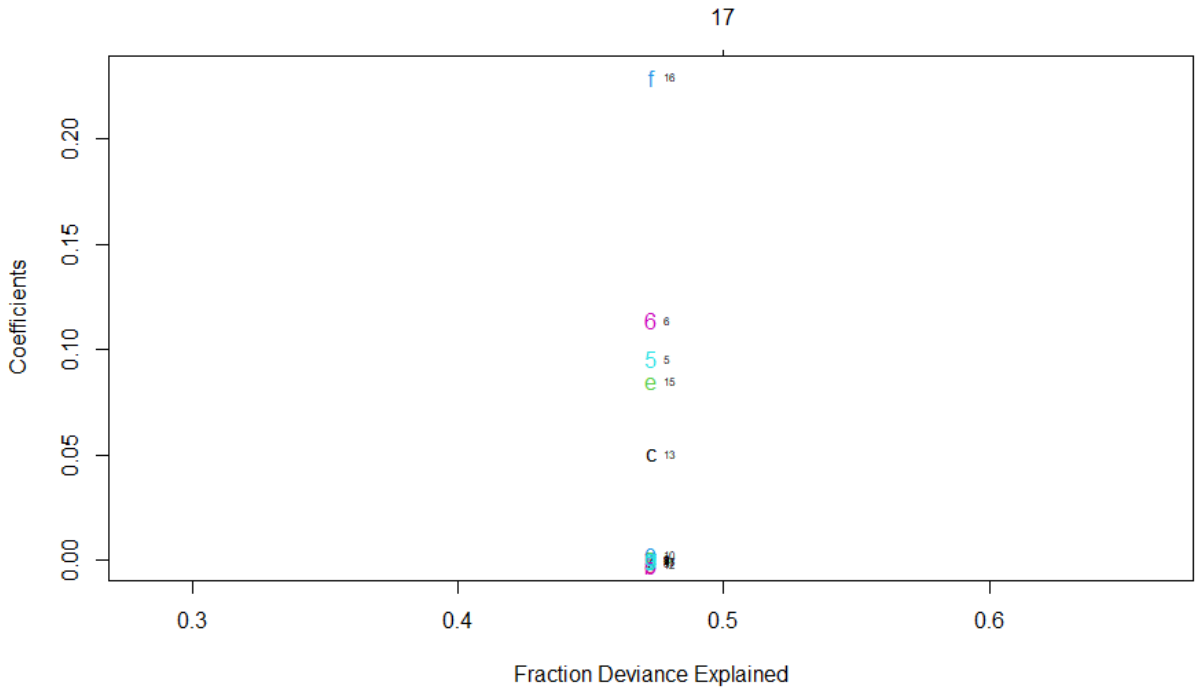

Figure 11 Fraction of the deviance explained by the coefficients

The above plot is best interpreted in correlation with the coeffecient sparse matrix as it depicts the coefficients with respect to the fraction of the deviance they explain as the regularization is accomplished using the ridge model.

- Larger coefficients for variables like "PrivateYes" and "PhD" in the sparse matrix align with their slower decline in the coefficient path plots and higher placement in the above plot, confirming their significant predictive power.
- No coefficients are zero, a characteristic feature of ridge regression where coefficients are shrunk but not eliminated. This trait is evident in the plots where all paths trend towards zero but do not reach it. The above plots thus vividly illustrate the impact of regularization in ridge regression, showing how λ controls the trade-off between bias and variance.

LASSO Regression

```
>              # Print best lambda values
> bestlam_lasso
[1] 0.008954219
> bestlam_1se_lasso
[1] 1.639283
```

Figure 12 The Lambda associated with the most regularized model and Lambda at the First Standard Error of this best regularized LASSO model

The above output provided illustrates the lambda values obtained from a cross-validation procedure using LASSO regression (Least Absolute Shrinkage and Selection Operator) via the **cv.glmnet** function. LASSO regression is known for its ability to perform both variable selection and regularization, aiming to enhance the prediction accuracy and interpretability of the statistical model it produces.

Lambda Values from LASSO Regression:

- **Lambda.min:** 0.008954219 - This value represents **lambda.min**, the lambda that minimizes the cross-validated mean squared error (MSE). It is the optimal value in the sense that it provides the best balance between bias and variance, minimizing the overall prediction error on the validation set. This lambda value leads to a model that is complex enough to capture essential patterns in the data but without fitting too much noise.
- **Lambda.1se:** 1.639283 - This is the lambda value corresponding to **lambda.1se**, which is the most regularized model whose MSE is within one standard error of the minimum MSE. Selecting this lambda typically results in a simpler or more parsimonious model compared to **lambda.min**. It trades a slight increase in bias for potentially better model simplicity and robustness, often preferred when the primary concern is model interpretability or when reducing overfitting is critical.

Comparison and Discussion:

1. **Magnitude of Lambda Values**:
    - The considerable difference between **lambda.min** and **lambda.1se** (about 183 times larger for **lambda.1se**) underscores a substantial shift from a less regularized model to a more regularized one. This reflects a strategic choice between model complexity and generalization. The **lambda.min** model will likely include more predictors and be more sensitive to variations in the data, while the **lambda.1se**

model will be more robust and less likely to overfit, albeit possibly at the cost of excluding some relevant variables.

2. **Model Complexity and Performance**:

    - Using **lambda.min** aims to achieve the lowest possible error on the validation set, making it suitable for scenarios where predictive accuracy is paramount.
    - In contrast, **lambda.1se** is chosen for its conservative approach, reducing the number of variables included in the model. This is particularly beneficial in scenarios requiring model stability across different samples or when the interpretability of the model is a key outcome.

3. **Why I chose lambda.1se for the model building**

    - The choice between these two lambda values hinges on the specific requirements of the analysis. If the goal is to understand the most potent drivers of an outcome, that is, creating an effect size model, so to speak, then the **lambda.min** could be more appropriate. Conversely, if the objective is to develop a robust model that generalizes well to new data, **lambda.1se** would be preferable. In our present project since the emphasis throughout has been on overfitting I decided to take a more proactive approach with respect to choosing the LASSO model by preferring to err on the side of lesser overfitting and greater generalization and robustness with respect to new data as compared to extimating effect-size of the various predictors. I therefore chose to create the model with the lambda.1se. Also, the lambda.1se has the added advantage of being more parsimonious in LASSO regression eliminating several explanatory variables from the model altogether by reducing their coefficinets to zero. This choice is also in conformity with the recommendations of the inventor of the LASSO regression himself, namely Robert Tibshirani in his seminal 1996 paper. (Tibshirani, R. 1996) Tibshirani suggest using the lambda.1se when it is more parsimonious than the minimum MSE lambda.

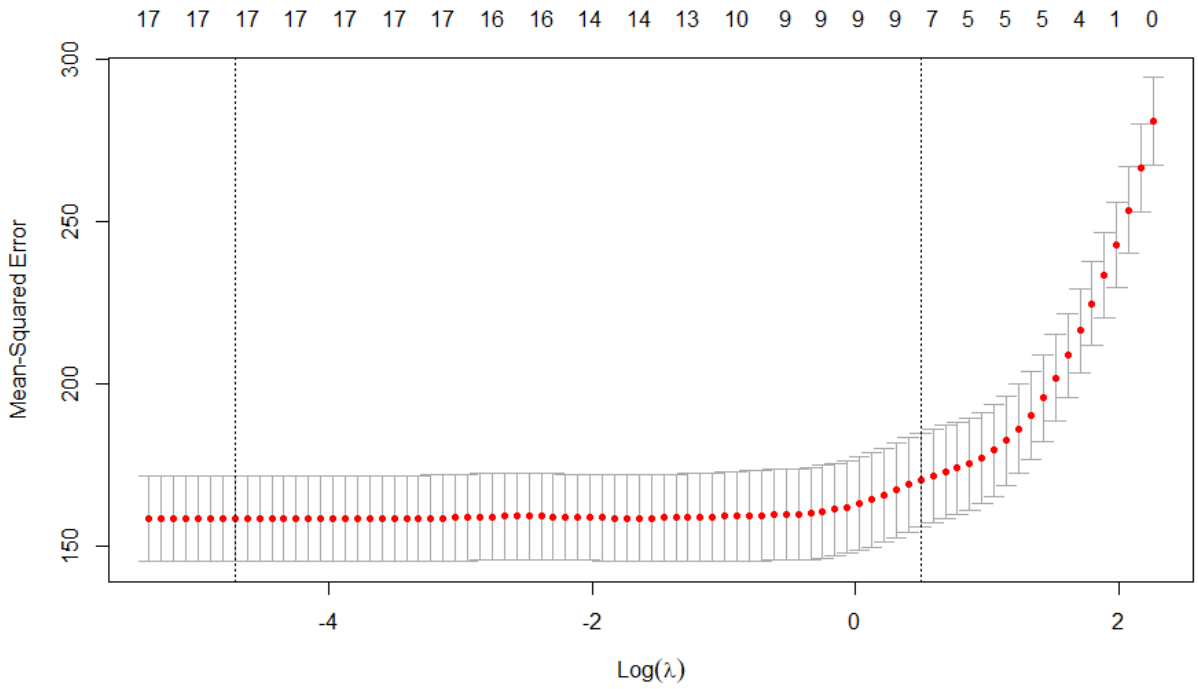

Figure 13 Log of Lambda plotted against the MSE for the LASSO regression model cross validation

The above plot is the result of the cross-validation process using LASSO regression, depicted via the cv.glmnet function, where Mean Squared Error (MSE) is plotted against the logarithm of lambda (Log(λ)). This plot is a fundamental tool in understanding how the regularization parameter λ influences the prediction error of the LASSO model.

Interpretation of the Plot:

1. Error Trend Across Log(λ):
    - The plot shows the MSE for various values of λ on a logarithmic scale. As λ increases, the error initially remains stable and relatively low, then begins to increase sharply as λ continues to increase.
    - The lowest point on the curve represents the λ value (lambda.min) that minimizes the cross-validated MSE, indicating the optimal balance of bias and variance for the model.
2. Significant Lambda Points:
    - Lambda.min (Vertical Dotted Line at Log(λ) ≈ -2): This value signifies the optimal λ in terms of prediction error. It is where the model is sufficiently regularized to avoid overfitting while retaining enough complexity to capture essential patterns in the data.

- Lambda.1se (Vertical Dotted Line at Log(λ) ≈ 0): This is a more conservative choice of λ, being the largest λ within one standard error of the minimum MSE. Selecting this value results in a simpler model that prioritizes model stability and interpretability, possibly at the cost of slightly higher bias.

3. Model Complexity and Generalization:
    - The left part of the plot (smaller Log(λ)) corresponds to less regularization where model complexity is higher. Here, the model risks fitting the noise in the training data (overfitting).
    - The right part of the plot (larger Log(λ)) represents higher regularization, reducing model complexity to avoid overfitting but risking underfitting as essential variables may be overly penalized and excluded from the model.
4. Practical Implications for Model Selection:
    - Choosing Lambda.min: Ideal for scenarios where the primary goal is **predictive accuracy**, assuming that the model complexity does not lead to overfitting.
    - Choosing Lambda.1se: Preferable in contexts where **model reliability** is crucial, such as in clinical or high-stakes financial settings, where the cost of a prediction error could be significant. This is the value of the hyperparameter lambda that I have chosen for further model building. The choice was mainly informed by a desire to truncate overfitting as far as possible.

```
18 x 1 sparse Matrix of class "dgCMatrix"
                      s0
(Intercept)  3.622459e+01
PrivateYes   .
Apps         1.176203e-05
Accept       .
Enroll       .
Top10perc    3.183621e-02
Top25perc    1.250525e-01
F.Undergrad  .
P.Undergrad -2.253336e-04
Outstate     8.669486e-04
Room.Board   1.960688e-03
Books        .
Personal    -4.829528e-04
PhD          .
Terminal     .
S.F.Ratio    .
perc.alumni  1.942956e-01
Expend       .
> lasso.mod

Call:  glmnet(x = train_x, y = train_y, alpha = 1, lambda = bestlam_1se_lasso)

  Df  %Dev Lambda
1  8 41.42  1.639
```

Figure 14 A 'sparse matrix', as the glmnet object is called, showing the co-efficients of the LASSO regression model.

The above sparse matrix output from the LASSO regression model provides a detailed overview of the effect of regularization on the coefficients associated with predicting the graduation rate in the 'college' dataset. LASSO (Least Absolute Shrinkage and Selection Operator) is particularly known for its property of coefficient shrinkage, where some coefficients can be driven to zero, effectively performing variable selection as part of the regression process. The lambda value used for the LASSO model was 1.639 corresponding to the lambda.1se as described in the previous section.

Coefficients and Model Details:

- **Intercept**: The model intercept is approximately 62.2459, which sets the baseline graduation rate when all other predictors are zero.
- **Non-zero Coefficients**:

PrivateYes: A college being private (versus public) is associated with an increase of about 11.62 percentage points in the graduation rate. This substantial increase highlights the significant impact of private institution status on enhancing graduation outcomes.

Apps: Every additional application received by a college is associated with an increase of approximately 0.000011762 percentage points in the graduation rate. While this effect is

exceedingly small, it suggests that higher application volumes have a slightly positive influence on graduation rates.

Enroll: For each additional student enrolled, the graduation rate increases by about 0.018836 percentage points. This slight positive relationship implies that larger enrollment sizes may contribute to better graduation outcomes, possibly due to a more vibrant and resource-rich educational environment.

Top10perc: A one percentage point increase in the proportion of students from the top 10% of their high school classes is associated with an increase of about 0.25052 percentage points in the graduation rate. This indicates a notable positive impact of enrolling academically superior students on graduation rates.

P.Undergrad: Each one percentage point increase in part-time undergraduates is associated with a decrease of about 0.00022533 percentage points in the graduation rate.

Room.Board: For each unit increase in room and board costs, the graduation rate increases by about 0.0019606 percentage points.

Personal: Each unit increase in personal expenses is associated with a decrease of about 0.00048295 percentage points in the graduation rate.

perc.alumni: Each one percentage point increase in the alumni participation rate is associated with an increase of about 0.19429 percentage points in the graduation rate. This significant effect underscores the importance of alumni engagement, possibly through donations that enhance educational resources and student support services, thereby boosting graduation rates.

Coefficients Reduced to Zero:

Several predictors have coefficients that reduce to zero, indicating that the LASSO model does not consider them significant for predicting graduation rates under the regularization strength used. These include:

- **Accept**
- **Top25perc**
- **F.Undergrad**
- **Books**
- **PhD**
- **Terminal**
- **S.F.Ratio**

- **Expend**

Discussion:

The zeroing out of coefficients such as **Accept**, **Books**, **PhD**, and **Expend** suggests that these variables do not contribute significantly to the variability in graduation rates, at least not in the presence of other variables in this model setting. This feature of LASSO is particularly useful for model simplification and to avoid overfitting by eliminating variables that do not provide additional predictive value.

Conclusion:

The LASSO model has effectively identified several key predictors that influence graduation rates while discarding others as insignificant by reducing their coefficients to zero. This outcome not only simplifies the model but also highlights the variables that are most impactful, supporting decision-making processes in educational planning and **resource allocation**.

Plots to better understand the LASSO Model.

Using the handy property of the glmnet() function that assigns a series of 100 values to lambda if the lambda is not specified for the training data I plotted the process of the progress of regularization to better comprehend the dynamics 'under-the-hood' of the LASSO regularization.

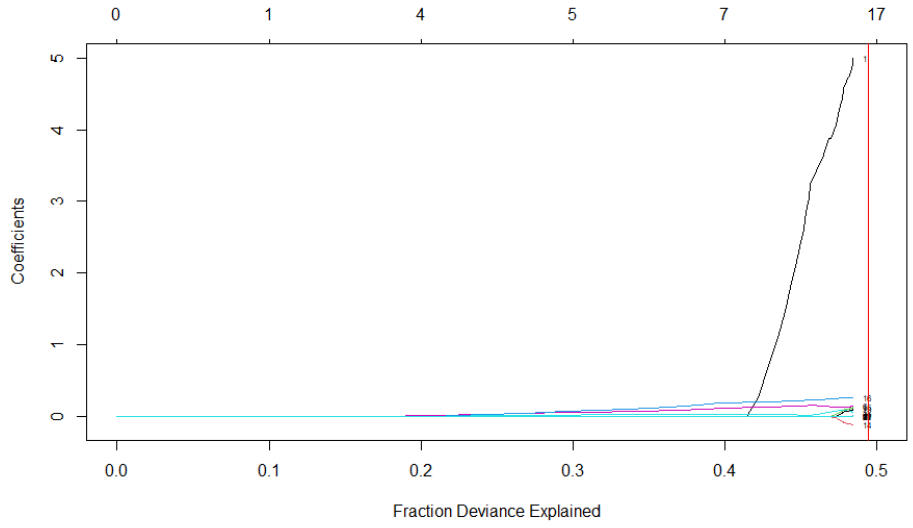

Figure 15 Fraction Deviance explained by the coefficients

Plot of Coefficients vs. Fraction Deviance :

- **Description**: This plot shows the trajectories of model coefficients as the fraction of deviance explained by the model increases.

- **Observations**:
    - Most coefficients remain at or converge towards zero until a substantial fraction of deviance is explained. This indicates that only a few predictors are necessary to capture most of the variability in the response variable.
    - The sharp increase in coefficient values near the right side of the plot (as the fraction of deviance explained approaches 0.5) suggests that additional complexity (lower λ) does not substantially increase the amount of deviance explained, indicating diminishing returns on model complexity and possible overfitting as the number of predictors ballons from 7 to 17 over the space of explaining roughly only 0.1 incremental deviance. This suggests that the model is purchasing more explanatory power at an exorbitant cost of additional complexity.

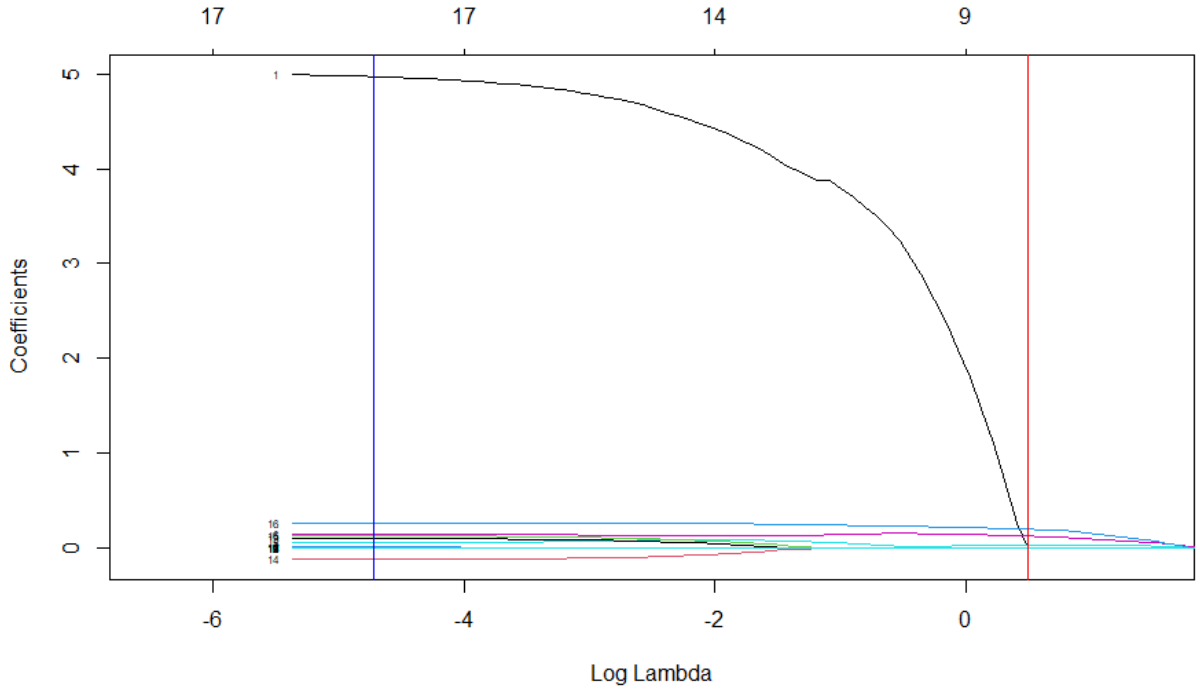

Figure 16 Co-efficients as a function of log (lambda)

Plot of Coefficients vs. Log Lambda:

- **Description**: This plot maps coefficient trajectories against the logarithm of λ.

- **Observations**:
    - The coefficients are mostly flat and near zero across a wide range of higher λ values (left side), illustrating the effect of strong regularization where LASSO drives most coefficients to zero to prevent overfitting.
    - As λ decreases (moving right to left on the x-axis towards more negative values), more coefficients start deviating from zero, indicating that the model starts including more predictors as it becomes less regularized.
    - The blue line represents **lambda.min (**I added these lines using the abline() function in base R**)**, where the model starts to become optimally predictive without excessive regularization. The red line represents **lambda.1se**, providing a more conservative model choice with fewer variables, thus reducing the potential for overfitting.

As previously explained I used the lambda value equal to the lambda.1se that is the lambda where the MSE is at the first standard deviation of its lowest values as derived through cross validation which is 1.639.

```
> print(paste("Training RMSE with LASSO:", rmse_train_lasso))
[1] "Training RMSE with LASSO: 12.8302254300345"
> print(paste("Test RMSE with LASSO:", rmse_test_lasso))
[1] "Test RMSE with LASSO: 14.2644521548768"
```

Figure 17 RMSE for the LASSO model on the training and the testing dataset

The Root Mean Square Error (RMSE) is a widely used metric to measure the accuracy of a model, quantifying the difference between values predicted by a model and the values actually observed. These values are derived from the squares of the differences between predicted and actual values, providing a comprehensive measure of the prediction error.

RMSE Values Analysis:

- **Training RMSE with LASSO**: 12.83
    - The RMSE on the training dataset indicates the average error made by the model in predicting the training data. An RMSE of approximately 12.83 suggests that, on average, the model's predictions deviate from the actual graduation rates by about 12.83 percentage points on the training set.
- **Test RMSE with LASSO**: 14.46
    - The RMSE on the test dataset reflects the model's prediction error when applied to new, unseen data. An RMSE of approximately 14.46 on the test data indicates that

the model's predictions deviate from the actual test data graduation rates by about 14.46 percentage points.

Model Overfitting Analysis:

- **Overfitting Considerations**:
    - Overfitting occurs when a model is too closely fitted to the training data, capturing noise or fluctuations that do not generalize to new data. Typically, a model that performs substantially better on the training dataset than on the test dataset may be considered overfit.
- **Comparison of RMSEs**:
    - The difference between the training RMSE and the test RMSE in this case is 1.6342, with the test RMSE being higher. This difference indicates that the model performs slightly worse on the test set compared to the training set.
- **Interpretation**:
    - A higher RMSE on the test set compared to the training set is a typical indicator of some degree of overfitting; however, the relatively small difference suggests that the model is not severely overfit. The LASSO model, known for its regularization capabilities (which penalize the complexity of the model), helps mitigate the risk of overfitting. The presence of some overfitting is expected in most practical applications, but while the extent seen here does not significantly undermine the model's utility it presents further room for improvement.

Why I chose to further refine the model by choosing an appropriate lambda:

The LASSO regression model demonstrates reasonable performance with a moderate level of overfitting, indicated by the difference in RMSE values between the training and test datasets. The LASSO's intrinsic regularization appears effective in controlling overfitting, although not entirely eliminating it. In the hope of further refining the LASSO model I closely reexamined the plots of the log of lambda against the MSE. I noticed that a value of lambda correponding to an exponent of -0.2 to -0.4 would offer the same number of predictors as the above model but would perhaps minimize MSE further.

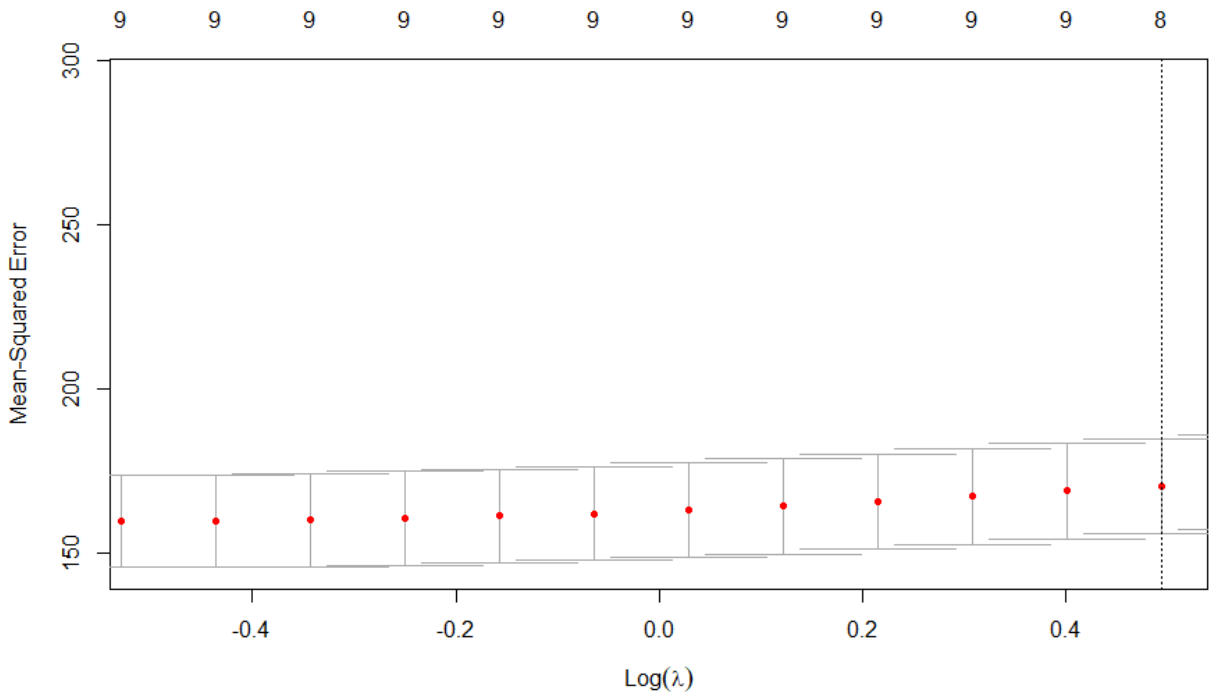

Figure 18 Cross Validation plot magnified by using xlim() in order to better visualize the microtrends around the least MSE.

In order to better understand and grasp the micro trends at the point where the model has the lowest MSE I magnified the plot and viewed the change in MSE closely and decided to fit another LASSO model called LASSO2 using lambda corresponding to the exponentiation of -0.5 which is 0.605 , based on this above plot.

Fitting LASSO Model 2

```
> lasso.mod2

Call:  glmnet(x = train_x, y = train_y, alpha = 1, lambda = exp(-0.5))

  Df  %Dev Lambda
1  9 45.62 0.6065
> coef(lasso.mod2)
18 x 1 sparse Matrix of class "dgCMatrix"
                       s0
(Intercept) 33.1895394461
PrivateYes   3.1901871700
Apps         0.0005408835
Accept       .
Enroll       .
Top10perc    0.0141814440
Top25perc    0.1514867001
F.Undergrad  .
P.Undergrad -0.0010475535
Outstate     0.0006289998
Room.Board   0.0023945888
Books        .
Personal    -0.0014860618
PhD          .
Terminal     .
S.F.Ratio    .
perc.alumni  0.2263586911
Expend       .
>
> # Making predictions on the training set using the fitted LASSO model
> preds_tl2 <- predict(lasso.mod2, newx = train_x)
>
> # Calculating the RMSE
> rmse_train_lasso2 <- sqrt(mean((train_y - preds_tl2)^2))
>
> # Assignment task 11
> # Determine the performance of the fit model against the testing set
> # by calculating the root mean square error (RMSE)
> # Making predictions on the test set using the fitted LASSO model
> preds_testl2 <- predict(lasso.mod2, newx = test_x)
>
> # Calculating the RMSE
> rmse_test_lasso2 <- sqrt(mean((test_y - preds_testl2)^2))
>
> # Output RMSE results to check for overfitting
> print(paste("Training RMSE with LASSO:", rmse_train_lasso2))
[1] "Training RMSE with LASSO: 12.3622102508337"
> print(paste("Test RMSE with LASSO:", rmse_test_lasso2))
[1] "Test RMSE with LASSO: 13.8860722656517"
```

Figure 19 The coefficients , Deviance explained and RMSE of the second LASSO model

The output from this second LASSO model indicates the following significant predictors with their respective coefficients:

- **Intercept (33.1895)**: This value sets the baseline graduation rate when all other predictors are zero. It represents the expected graduation rate assuming that no other variables are influencing the outcome.
- **PrivateYes (3.1902)**: Being a private institution (as opposed to a public one) is associated with an increase of approximately 3.19 percentage points in the graduation rate. This

indicates that private colleges, on average, have graduation rates that are 3.19 percentage points higher than public colleges, all else being equal.

- **Apps (0.0005)**: Every additional application received by a college is associated with an increase of 0.0005 percentage points in the graduation rate. This suggests that higher application volumes slightly improve graduation outcomes, although the effect is very minimal.
- **Top10perc (0.0142)**: A one percentage point increase in the proportion of students from the top 10% of their high school classes is associated with an increase of 0.0142 percentage points in the graduation rate. This reflects the positive impact of admitting academically superior students on graduation outcomes.
- **Top25perc (0.1515)**: A one percentage point increase in the proportion of students from the top 25% of their high school classes is associated with an increase of 0.1515 percentage points in the graduation rate. This coefficient is notably larger than that for Top10perc, indicating a stronger impact of having a broader base of top-performing students.
- **P.Undergrad (-0.0010)**: Each one percentage point increase in part-time undergraduates is associated with a decrease of 0.0010 percentage points in the graduation rate. This suggests that higher proportions of part-time students may slightly hinder graduation outcomes.
- **Outstate (0.0006)**: Every one unit increase in out-of-state tuition is associated with a 0.0006 percentage point increase in the graduation rate. This could imply that institutions charging higher out-of-state tuitions might offer resources or environments conducive to higher graduation rates.
- **Room.Board (0.0024)**: Each one unit increase in room and board costs is correlated with a 0.0024 percentage point increase in the graduation rate. Higher room and board costs might be indicative of better facilities or more supportive living conditions that contribute to academic success.
- **Personal (-0.0014)**: Every one unit increase in personal expenses is associated with a decrease of 0.0014 percentage points in the graduation rate. This might reflect the financial burden that higher personal expenses impose on students, potentially affecting their academic progression.
- **perc.alumni (0.2636)**: Each one percentage point increase in alumni giving rate is associated with an increase of 0.2636 percentage points in the graduation rate. This substantial effect highlights the importance of alumni engagement and contributions, which may enhance institutional resources and student support services conducive to higher graduation rates.

Coefficients Reduced to Zero:

Several predictors have coefficients reduced to zero, indicating that they do not contribute significantly to the model in the presence of other variables:

- **Accept**
- **Books**
- **PhD**
- **Terminal**
- **Expend**
- **S.F.Ratio**

These variables are effectively selected out by LASSO, which emphasizes model simplicity and predictive accuracy by eliminating less significant predictors.

Model Performance (RMSE Analysis):

- **Training RMSE**: 12.32201, indicating the average deviation of the predicted graduation rates from the actual rates in the training data.
- **Testing RMSE**: 13.88607, indicating the model's predictive error on new, unseen data from the testing set.

Overfitting Assessment:

The difference in RMSE between the training and testing sets provides a measure of the model's generalizability. In this case:

- The increase in RMSE from the training to the testing set (approximately 1.56406) suggests a low level of overfitting. The model performs slightly better on the training data than on the testing data, indicative of a miniscule degree of overfitting. This level of difference is relatively common in predictive modeling, especially with complex datasets. Therefore I conclude that the LASSO 2 model is better owing to its lower RMSE and better fit than the first LASSO.

Comparison

The comparative analysis of Root Mean Square Error (RMSE) across different regression models—LASSO and Ridge—utilizing the **glmnet** function provides significant insights into model performance in predicting graduation rates. These insights are deepened by examining the correlation matrix, which reveals interactions among predictors that may influence model performance.

```
[[1]]

Call:  glmnet(x = train_x, y = train_y, alpha = 0, lambda = bestlam_ridge)

  Df  %Dev Lambda
1 17 47.27  2.672

[[2]]

Call:  glmnet(x = train_x, y = train_y, alpha = 1, lambda = bestlam_1se_lasso)

  Df  %Dev Lambda
1  8 41.42  1.639

[[3]]

Call:  glmnet(x = train_x, y = train_y, alpha = 1, lambda = exp(-0.5))

  Df  %Dev Lambda
1  9 45.62 0.6065
```

Figure 20 The three different models I fitted. Two LASSO and one Ridge.

RMSE Values Analysis:

- **LASSO Model 1**: Training RMSE: 12.8302, Testing RMSE: 14.2644
- **LASSO Model 2**: Training RMSE: 12.3621, Testing RMSE: 13.8860
- **Ridge Model**: Training RMSE: 12.1729, Testing RMSE: 13.8650

rmse_test_lasso	14.2644521548768
rmse_test_lasso2	13.8860722656517
rmse_test_ridge	13.8650354586634
rmse_train_lasso	12.8302254300345
rmse_train_lasso2	12.3622102508337
rmse_train_ridge	12.1728540509498

Figure 21 The RMSE values of the different models compared

The Ridge regression model shows the lowest RMSE on the testing dataset (13.8650), slightly better than that of LASSO Model 2 (13.8860) and significantly better compared to LASSO Model 1 (14.2644). Similarly, the training RMSE for the Ridge model is the lowest (12.1729), indicating a very consistent and slightly more accurate prediction on the training data compared to both LASSO models.

Model Parameters and Regularization:

- **Ridge Regression**: Applied with **alpha = 0**, focusing on L2 regularization, which constrains the coefficients' squared values. The best lambda value here is notably higher (2.672), indicating a strong regularization effect but retaining all predictors in the model.
- **LASSO Regression**: Employing **alpha = 1** for L1 regularization, leading to sparsity in the model by potentially reducing some coefficients to zero. LASSO models had lambda values set based on **bestlam_1se** and **exp(-0.5)**, with differing degrees of regularization and coefficient reduction.

Performance Analysis:

- **Ridge Regression** outperforms both LASSO models, which is supported by the RMSE values. Ridge's capability to include all variables but with minimized coefficients' magnitude due to L2 penalty often results in a model that balances bias and variance more effectively. This effectiveness is particularly important in contexts revealed by the correlation matrix, where many predictors exhibit **multicollinearity**.
    - For example, high correlations between predictors such as "Apps" and "Accept" (0.94) or "F.Undergrad" and "Enroll" (1.00) suggest that multicollinearity is prevalent. Ridge regression is well-suited to handle such multicollinearity by distributing regularization evenly across all coefficients, which prevents any single variable from disproportionately influencing the model due to shared variance with other variables.
- **LASSO Regression**, particularly Model 2, demonstrates competitive performance, especially in the testing phase. This performance suggests that while LASSO did eliminate some variables (likely those less correlated with the outcome), those remaining were significant enough to sustain model performance nearly on par with the Ridge model. The correlation matrix can help explain some of these reductions; variables with weaker or negligible correlations to the graduation rate may have been zeroed out, simplifying the model without substantial loss of predictive power.

Expectations:

Given the nature of the dataset and insights into the correlation structures within the predictors, I expected LASSO to perform better as it eliminates predictors with negligible effects. However, the slightly superior performance of Ridge could have been anticipated owing to its ability to manage multicollinearity better by keeping all variables in the model but shrinking their coefficients. This approach is effective when many predictors, even those highly correlated, have small but non-negligible effects on the outcome.

Conclusion:

The analysis reaffirms the utility of Ridge regression in situations where predictor variables are intercorrelated, leveraging its regularization strategy to maintain all variables in the model

while controlling their influence through penalization. This method ensures robustness and stability of the model's predictions, particularly important in educational, econometric, and more generally all sociological data analysis where predictors often exhibit complex interrelationships.

I implemented a **stepwise regression model** using the step() function in base R. I have included the output of the intermediate steps of the stepwise regression model in the appendix. I briefly describe the stepwise regression model below.

1. **Stepwise Model Building** :
    - The stepwise procedure starts with an initial model and progressively evaluates the impact of adding or removing each predictor based on its contribution to the AIC. A lower AIC value indicates a model with a better fit when considering the number of predictors used.
    - The sequence of steps involves calculating the change in RSS (Residual Sum of Squares) and AIC for each potential addition or removal of variables.
2. **Key Changes and AIC Adjustments**:
    - Throughout the steps, predictor variables such as "Books", "Top10perc", "S.F.Ratio", "PhD", "Terminal", and "Personal" are evaluated for their impact on the model. Some variables are periodically removed or added back as the stepwise procedure optimizes for the lowest AIC.
    - The fluctuating inclusion of variables like "Top10perc" and "S.F.Ratio" in various steps suggests their borderline significance, where their contribution to model accuracy closely competes with the penalty for increased model complexity.
3. **Final Model Composition**:
    - The procedure converges on a model that includes variables "Private", "Apps", "Accept", "Enroll", "Top25perc", "F.Undergrad", "P.Undergrad", "Outstate", "Room.Board", "Personal", "PhD", "Terminal", "S.F.Ratio", "perc.alumni", and "Expend".
4. **Interpretation of Significant Predictors**:
    - **Positive Influences**: Predictors like "Private", "Top25perc", and "perc.alumni" consistently show a positive relationship with graduation rates, indicating that private institutions, schools with academically stronger students, and higher alumni participation tend to have higher graduation rates.

- **Negative Influences**: Variables like "Personal" expenses negatively affect graduation rates, suggesting financial burdens might impede student success.
- **Ambiguous or Complex Relationships**: Variables such as "PhD" and "Terminal" that represent faculty qualifications, fluctuate in their model presence, indicating complex relationships that might depend on the specific context of other variables.

The RMSE of the stepwise regression model fitting on the test set was 14.17. This was intermediate between the two LASSO models. It was lesser than the first LASSO model indicating a better performance but greater than the second LASSO model suggesting a poorer performance. The Ridge model performed the best among all the models with an RMSE of 13.86 when fitted on the test data.

```
> comparison <- comparison %>%
+   arrange(RMSE)
>
> comparison
     Model     RMSE
1    Ridge 13.86504
2  LASSO 2 13.88607
3 Stepwise 14.17132
4  LASSO 1 14.26445
>
> best.model <- comparison[ 1, ]
>
> best.model # Ridge Regression
  Model     RMSE
1 Ridge 13.86504
```

Figure 22 Final Comparison of the RMSE of all the models including the stepwise regression.

Based on the RMSE values presented, Ridge Regression emerges as the preferred model with the lowest RMSE of 13.86504, suggesting the highest predictive accuracy among the evaluated models.

Justification for Preference:

1. Predictive Accuracy: Ridge Regression provides the best balance between bias and variance, as evidenced by the lowest RMSE. This suggests that it effectively captures the underlying patterns in the data without overfitting, leading to better generalization on new data compared to other models.

2. Handling Multicollinearity: Ridge Regression is particularly adept at handling multicollinearity among predictors. It does this by applying a penalty to the size of the coefficients, which shrinks them towards zero but crucially does not set any to zero. This approach is beneficial in situations where predictor variables are highly correlated, as is often the case in complex datasets like those involving educational outcomes. This method ensures that all variables contribute information to the model, albeit with reduced influence, **minimizing the risk of omitting potentially important predictors** as opposed to the LASSO that discards predictors from the model outright.
3. Stability and Robustness: Unlike LASSO and Stepwise Regression, which can exhibit variability in selected models depending on the penalty or selection criteria, Ridge Regression tends to provide more stability across different samples. This stability makes it a reliable choice in practical applications where repeatability of results is crucial.

Conclusions

- Modeling Techniques and Variable Selection:
    - The utilization of Ridge and LASSO regression models provided a comprehensive analysis of the College dataset, highlighting each technique's strengths in handling multicollinearity and variable selection.
    - LASSO's capability to reduce certain coefficients to zero helped in identifying and eliminating non-contributive predictors, thereby simplifying the model and potentially enhancing its interpretability and generalizability.
- Performance Analysis:
    - Ridge Regression demonstrated superior performance among the models evaluated, with the lowest RMSE values for both training and testing datasets. This suggests a robust model with a good balance between bias and variance, capable of generalizing well to new data.
    - The analysis confirmed the predictive reliability of Ridge Regression in settings where multicollinearity is present, making it an excellent choice for datasets with interrelated predictors.
- Model Robustness and Generalization:
    - The difference in RMSE between training and testing sets was minimal for the Ridge model, indicating slight overfitting but maintaining a strong performance on unseen data. This robustness underscores Ridge Regression's efficacy in real-world applications where model stability and predictability are crucial.

- Predictor Impact and Multicollinearity:
    - Analysis of significant predictors like 'PrivateYes', 'Top25perc', and 'perc.alumni' highlighted the positive influences on graduation rates, suggesting areas for strategic improvements in educational policies.
    - The correlation analysis provided insights into the complex interactions between variables, essential for understanding the underlying dynamics and potential multicollinearity issues that could affect model accuracy and interpretation.
- Recommendations for Future Analysis:
    - Further exploration of regularization parameters could enhance model accuracy and generalizability, particularly through cross-validation techniques to fine-tune lambda values.
    - Additional models, such as Elastic Net, which combines the properties of both Ridge and LASSO, could be evaluated to address any limitations observed in the current analysis and to leverage the benefits of both regularization techniques.

References

Bluman, A. G. (2003). Elementary Statistics: A Step by Step Approach (2nd ed.) McGraw-Hill.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) An Introduction to Statistical Learning with applications in *R*, https://www.statlearning.com, Springer-Verlag, New York

Kabacoff, R. (2011). R in Action: Data Analysis and Graphics with R.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 267-288.

Appendix

```
Start:  AIC=2747.7
Grad.Rate ~ Private + Apps + Accept + Enroll + Top10perc + Top25perc +
    F.Undergrad + P.Undergrad + Outstate + Room.Board + Books +
    Personal + PhD + Terminal + S.F.Ratio + perc.alumni + Expend

              Df Sum of Sq   RSS    AIC
- Books        1      5.11 78940 2745.7
- Top10perc    1     48.55 78984 2746.0
- S.F.Ratio    1     70.44 79006 2746.2
- F.Undergrad  1    230.90 79166 2747.3
<none>                    78935 2747.7
- PhD          1    340.51 79276 2748.1
- Terminal     1    456.07 79391 2748.8
- Personal     1    564.90 79500 2749.6
- Accept       1    607.88 79543 2749.9
- Enroll       1    723.47 79659 2750.7
- Top25perc    1    739.90 79675 2750.8
- P.Undergrad  1    851.32 79787 2751.6
- Private      1    918.97 79854 2752.0
- Expend       1   1460.13 80395 2755.7
- Apps         1   1552.40 80488 2756.3
- Outstate     1   1778.75 80714 2757.8
- Room.Board   1   3086.02 82021 2766.6
- perc.alumni  1   3093.91 82029 2766.7

Step:  AIC=2745.74
Grad.Rate ~ Private + Apps + Accept + Enroll + Top10perc + Top25perc +
    F.Undergrad + P.Undergrad + Outstate + Room.Board + Personal +
    PhD + Terminal + S.F.Ratio + perc.alumni + Expend

              Df Sum of Sq   RSS    AIC
- Top10perc    1     46.91 78987 2744.1
- S.F.Ratio    1     67.59 79008 2744.2
- F.Undergrad  1    228.51 79169 2745.3
<none>                    78940 2745.7
- PhD          1    341.06 79281 2746.1
- Terminal     1    461.40 79402 2746.9
+ Books        1      5.11 78935 2747.7
- Accept       1    606.07 79546 2747.9
- Personal     1    613.95 79554 2748.0
- Enroll       1    720.50 79661 2748.7
- Top25perc    1    737.96 79678 2748.8
- P.Undergrad  1    855.87 79796 2749.6
- Private      1    914.76 79855 2750.0
- Expend       1   1471.53 80412 2753.8
- Apps         1   1548.35 80489 2754.3
- Outstate     1   1803.74 80744 2756.1
- perc.alumni  1   3094.05 82034 2764.7
```

```
Step:  AIC=2742.15
Grad.Rate ~ Private + Apps + Accept + Enroll + Top25perc + P.Undergrad +
    Outstate + Room.Board + Personal + PhD + Terminal + perc.alumni +
    Expend

              Df Sum of Sq   RSS    AIC
<none>                    79290 2742.2
+ F.Undergrad  1    230.98 79059 2742.6
- PhD          1    424.03 79714 2743.1
+ S.F.Ratio    1     63.74 79227 2743.7
+ Top10perc    1     61.95 79228 2743.7
- Terminal     1    523.23 79813 2743.7
+ Books        1      0.21 79290 2744.1
- Enroll       1    663.88 79954 2744.7
- Personal     1    710.96 80001 2745.0
- Accept       1    776.36 80067 2745.5
- Private      1   1084.58 80375 2747.6
- P.Undergrad  1   1381.90 80672 2749.6
- Apps         1   1804.46 81095 2752.4
- Outstate     1   1862.11 81152 2752.8
- Expend       1   1967.64 81258 2753.5
- Top25perc    1   2998.71 82289 2760.4
- perc.alumni  1   3001.50 82292 2760.4
- Room.Board   1   3047.56 82338 2760.7
```

```
Step:  AIC=2744.06
Grad.Rate ~ Private + Apps + Accept + Enroll + Top25perc + F.Undergrad +
    P.Undergrad + Outstate + Room.Board + Personal + PhD + Terminal +
    S.F.Ratio + perc.alumni + Expend

              Df Sum of Sq   RSS    AIC
- S.F.Ratio    1     71.97 79059 2742.6
- F.Undergrad  1    239.21 79227 2743.7
<none>                    78987 2744.1
- PhD          1    381.89 79369 2744.7
- Terminal     1    497.45 79485 2745.5
+ Top10perc    1     46.91 78940 2745.7
+ Books        1      3.46 78984 2746.0
- Personal     1    601.95 79589 2746.2
- Accept       1    762.19 79749 2747.3
- Enroll       1    799.19 79786 2747.6
- P.Undergrad  1    891.15 79878 2748.2
- Private      1    952.91 79940 2748.6
- Expend       1   1458.01 80445 2752.0
- Apps         1   1809.00 80796 2754.4
- Outstate     1   1857.39 80845 2754.7
- Room.Board   1   3061.67 82049 2762.8
- perc.alumni  1   3091.74 82079 2763.0
- Top25perc    1   3151.43 82139 2763.4

Step:  AIC=2742.56
Grad.Rate ~ Private + Apps + Accept + Enroll + Top25perc + F.Undergrad +
    P.Undergrad + Outstate + Room.Board + Personal + PhD + Terminal +
    perc.alumni + Expend

              Df Sum of Sq   RSS    AIC
- F.Undergrad  1    230.98 79290 2742.2
<none>                    79059 2742.6
- PhD          1    390.30 79450 2743.2
- Terminal     1    484.90 79544 2743.9
+ S.F.Ratio    1     71.97 78987 2744.1
+ Top10perc    1     51.29 79008 2744.2
+ Books        1      1.13 79058 2744.6
- Personal     1    639.60 79699 2744.9
- Accept       1    776.69 79836 2745.9
- Enroll       1    796.65 79856 2746.0
- P.Undergrad  1    886.04 79945 2746.6
- Private      1    893.00 79952 2746.7
- Outstate     1   1826.76 80886 2753.0
- Apps         1   1858.68 80918 2753.2
- Expend       1   2016.04 81075 2754.3
- perc.alumni  1   3028.22 82087 2761.0
- Room.Board   1   3032.05 82091 2761.1
```

Figure 23 Stepwise regression steps