2 College Scorecard Data

The US Department of Education collects data from every "college" level institution in America and makes a lot of data available under the College Scorecard.

- This question uses a curated extract of the college scorecard data. The variable names and definitions are at the end.
- This dataset has 23 variables of data on 1,695 four-year colleges.
- This dataset is on Canvas or at "https://raw.githubusercontent.com/AU-datascience/data/main/427-627/college_scorecard_extract_sep_2023.csv".

We want to predict the Endowment of a new colleges given the other variables as potential predictors.

In the following steps, build models to predict the College Endowment (`ENDOWBEGIN`) and use $K = 10$-fold cross-validation to tune and evaluate predictive performance with **`set.seed(123)` as appropriate.**

- For each problem, describe your approach, your R code, the most important results, and your interpretation of the results.
- You may write your responses on this document by hand after running code in R and/or submit a file on Canvas with your approach, code, results, and interpretation of results.

2.1 Multiple Linear Regression Regularization

- Load the data and assign the name `college` to it. Get rid of any records with `NA`s and divided `ENDOWBEGIN` by 1 million to reduce the scale. Glimpse `college`.

````
        ```{r}
        #| message: false
        library(tidyverse)
        college <- read_csv("https://raw.githubusercontent.com/AU-
datascience/data/main/427-627/college_scorecard_extract_sep_2023.csv")
        college <- na.omit(college)
        college$ENDOWBEGIN <- college$ENDOWBEGIN/1000000
        #college <-
read_csv("./data/college_scorecard_extract_sep_2023.csv")
        #glimpse(college)
        ```
````

- Fit a multiple linear regression of Endowment ($ENDOWBEGIN$) on all the other variables as a full model.
- How many predictors appear important with a p value less than 0.1?

Do any of the variables have a high generalized variance inflation factor GVIF? If any, which ones and do they make sense as having high GVIF given the other variables?

Refit a reduced model without `MN_EARN_WNE_P10` and `PCT_WHITE`. Are there any changes in significant variables?

- Check the GVIF again and comment on any changes.
- Create a new data frame with the variables below (you can use the following code).
    - Remove the rows with `REGION = "Outlying Regions"`.
    - Convert all character variables to factors.

````
        ```{r}
        college |>
          dplyr::select(
             ENDOWBEGIN, ADM_RATE, AVGFACSAL, CONTROL,FIRST_GEN,
GRAD_DEBT_MDN, PCIP27,
        PCT_ASIAN, REGION
          ) |>
          filter(REGION != "Outlying Regions") |>
          mutate(across(where(is.character), as.factor)) ->
        college2
        ```
````

- Use the {boot} package with `college2` to report the prediction MSE for a full model (`ENDOWBEGIN` on the other data) based on K-10 fold cross-validation adjusted deviance.

2.2 Regularization via Shrinkage

Use LASSO with cross validation to model `ENDOWBEGIN` on the other variables in `college2` and find the best lambda.

- Create model matrices for `X` and `Y`.
- Use `set.seed(123)` for the cross validation.
- Plot the result of the cross validation.
- Show the result and identify whether `lambda-min` or `lambda.1se` has fewer non-zero variables?
- Show the coefficients for `lambda.1se` and discuss which were driven to zero if any.
- Do any of the +/- signs of the coefficients for the variables surprise you?

2.3 Principal components.

Calculate the principal components using the `X` model matrix you created earlier, with scaling, and show the scree plot.

- Interpret the scree plot

Given the scree plot, choose to create **either** a PCR **or** a PLSR model.

Create the model with scaling and K=10 fold cross-validation. (Use seed 123)

How many principal components are

- Needed to explain 90% of the total variation *among X-variables*?
- Needed to explain 40% of the total variation *of the response*, `ENDOWBEGIN`?
- What is the optimal number of PCs based on adjusted Cross-Validation RMSEP?
- What is the adjusted Cross Validation MSEP for the optimal number of PCs?
- Show the validation plot.

2.4 Summary

Create a summary table showing the method, the MSE, and the number of predictors.

- Recommend a model for predicting `ENDOWBEGIN` for new observations and explain your choice.

Method	Predicted MSE	Number of Predictors
Linear Model (Reduced)		
LASSO lambda.1se		
PCR		
PLSR		

2.5 Classification with SVM (Optional Extra Credit 4 points)

We now want to predict whether a new college is Private or Public based on the data in `college2`.

Tune a Support Vector Machine model to find the best cost and kernel.

- Use the range of costs in `seq(4.0, 6.0, 0.25)` and the linear and radial kernels.

What is the best cost value and the best kernel and the cross-validated error rate?

- How many support vectors are there?

Plot the results looking at `ADM_RATE` and `AVGFACSAL`.

- Comment on the plot

2.5.1 College Scorecard Data

Variable	Definition
ADM_RATE	Admission Rate
AGE_ENTRY	Average age of entry
AVGFACSAL	Average Faculty Salary
CONTROL	Public or Private Non-Profit
COSTT4_A	Cost of an Academic Year
ENDOWBEGIN	Endowment at the Beginning of the year
FEMALE	Percent Female Students
FIRST_GEN	Percent First Generation Students
GRAD_DEBT_MDN	Median Debt at Graduation
LOCALE	City, Suburban, Town or Rural
MD_EARN_WNE_P10	Median Earnings 10 years after enrollment
MN_EARN_WNE_P10	Mean earnings 10 years after enrollment
PCIP14	Percent Engineering Degrees
PCIP27	Percent Math Stat Degrees
PFTFAC	Percent Full Time Faculty
PCT_ASIAN	Percent Asian in Home Zip Code
PCT_BLACK	Percent Black in Home Zip Code
PCT_WHITE	Percent White in Home Zip Code
PCT_HISPANIC	Percent Hispanic in Home Zip Code
PCTPELL	Percent with Pell Grant
REGION	Location in United States
SAT_AVG	Average SAT Score
UGDS	Total Undergraduates