Midterm

STAT 471/571/701

Nov 1, 2016, 6 - 8 PM

Table of Contents

[Instructions 1](#_Toc465756681)

[Question 1: The College Dropout 3](#_Toc465756682)

[Question 2: School Type and State vs Graduation Rate 3](#_Toc465756683)

[Question 3: Faculty effects 4](#_Toc465756684)

[Question 4: Parsimonious models (all subsets) (We could take this question out?????) 4](#_Toc465756685)

[Question 5: Parsimonious models (LASSO) 5](#_Toc465756686)

[Question 6: Graduation Evaluation 5](#_Toc465756687)

[Question 7: Freedom of the Press 6](#_Toc465756688)

**Name:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Instructions

This exam requires you to use R. It is completely open book/notes. Write your answers in a Word (.docx) or R Markdown(.rmd) format. It is not a good idea to use .rmd if you are not familiar with it yet. Always attach the plots or R output where needed. On the other hand, hide/control unnecessary code or output for knitr users. If you have trouble to format the plots, don’t worry about it. We are not looking for especially pretty solutions, but rather to see if you could make sense out of the data using R.

**Data needed**

Available on Canvas: /canvas/Data/Newsweek\_graduation\_subset.csv

**Electronic Submission**

In the ‘Assignments’ section of Canvas, go to the ‘Midterm’ assignment and you will be able to upload your completed files. If you used knitr, upload your .rmd and a compiled file (e.g. .pdf) for knitr users. If you didn't, attach your writeup with R code attached as an appendix for non-knitr users. Make it clear which parts of your code correspond with which parts of your analysis. **A write up inside R with comments is not acceptable!**

***The submisison assignment will close at 8:10 pm.***

**FAQ**

* If you have trouble uploading your files, email them to lzhao@wharton.upenn.edu.
* As always skip any part you have trouble with and you may come back to finish it if you have time. Ask one of us for help if you are stuck somewhere.
* I have written a few lines introducing the data, helping out with data downloading, and extracting a subset.
* None of the questions are related, so skip any part you have trouble with and finish it later.

Graduation rate is a key measurement of success among colleges. Identifying what affects graduation rate will help schools to improve this important outcome for their students.

Newsweek Magazine collected a data set about college outcomes in 1995. The data includes graduation rates for US colleges, various metrics describing the student body and faculty members, as well as several other descriptors of college life. After intensive sanitizing work, we cleaned the original data set and created some additional useful variables. We will work with this revised version of the data, available in /canvas/Data/Newsweek\_graduation\_subset.csv.

Out of all variables, we are in particularly interested in the following 13 variables. Note that most of them are self-explanatory.

Grad.rate: Graduation Rate

Name: Name of the school

State: Location

Schooltype: 1 = Public, 2 = Private

All.test.std: A standardized summary of SAT, ACT scores for admitted students

App.accept: # of accepted applicants that year

Acc.Rate: The percent of applicants for admission, who were accepted by the college

Pct.Yield: The percent of accepted students, who actually enrolled

Total.students: Total number of students in the school

Student.Faculty: Student-faculty ratio

Pct.fac.degree: Percentage of faculty with highest degrees

In.Tuition: In state tuition

Room.board: Room and board costs

We first do some basic data exploration to familiarize ourselves with the data set. We then use this data to develop statistical models for exploring factors related to the graduation rate and for predicting graduation rate.

# Question 1: The College Dropout

Read the Newsweek dataset into R; here are a few lines to aid you. Notice that Penn is in **row 820**.

# Required libraries - you may add more packages as desired  
library(leaps) # regsubsets() for model selection  
library(car) # Anova()  
library(glmnet) # glmnet() and cv.glmnet()

rm(list = ls()) # Remove all the existing variables  
  
college\_data <- read.csv("Newsweek\_graduation\_subset.csv")  
str(college\_data)  
college\_data$Schooltype <- as.factor(college\_data$Schooltype)

We also identify and isolate Penn within this dataset, to use for prediction purposes later.

penn\_loc <- which(college\_data$Name == "University of Pennsylvania") # identify Penn  
Penn <- college\_data[penn\_loc, ]

**a)** How many colleges are included in this dataset? How many variables are there in this data set? List all the variable names. Apart from continuous variables indicate which variables are categorical variables. Make sure they are treated as factors in college\_data.

**b)** Which school has the highest graduation rate, and what is that rate? Which school has the lowest graduation rate, and what is that rate? What was Penn's graduation rate in 1995? What is the mean graduation rate, across all schools? *Do not actually include the summary statistics.*

**c)** Write a very short (max 3 sentences) summary about the distribution of graduation rate, and provide a histogram of graduation rate.

**Assume all linear model assumptions are met in the following analyses.**

# Question 2: School Type and State vs Graduation Rate

**a)** Make a back to back boxplot of graduation rate vs school type. Does one type seems to have higher a graduation rate compared to the other type? Write a short (max 3 sentences) summary of this finding. Does that agree with your intuition about private schools (Schooltype = 2) vs. public schools (Schooltype = 1)?

**b)** fit1: Grad.rate vs. Schooltype

Perform a test to determine if the mean Grad.rate between the two school types is different at .01 level. Which type has a higher Grad.rate? Produce a 95% confidence interval for the mean difference.

**c)** fit1.1: Grad.rate vs. State

Can we prove that the mean graduation rates are different, at the 0.01 level, among all the states? Which state appears to have the highest graduation rate, and which state appears to have the lowest graduation rate? Note that AK/Alaska is the base case in this analysis.

**d)** fit1.2: Grad.rate vs. Schooltype and State

Controlling for school type, is the school's state a useful factor at the .01 level?

**e)** Write a few (max 3 lines) sentences to summarize your findings in Question 2.

# Question 3: Faculty effects

Consider the variable Pct.fac.degree, which summarizes the percent of faculty members who hold higher education degrees.

Construct fit2: Grad.rate vs. Pct.fac.degree

**a)** Report the summary of your linear model. Is Pct.fac.degree a significant variable in this model at .05 level? How does Pct.fac.degree affect Grad.rate?

**b)** Make a scatter plot with = Grad.rate and = Pct.fac.degree. Overlay fit2 onto the plot.

Construct fit2.1: Grad.rate vs. Pct.fac.degree + All.test.std

**c)** Is Pct.fac.degree still a significant variable in this model at the .05 level?

**d)** Interpret the coefficient of Pct.fac.degree in fit2.1.

**e)** Why might the two beta's for Pct.fac.degree differ?

# Question 4: Parsimonious models (all subsets) (We could take this question out?????)

Construct fit3: a model with all available sensible variables

Based on fit3, answer the following questions:

**a)** Is State a significant variable at .01 level after controlling for all other variables in the model? Provide an appropriate test.

**b)** If you were to kick one variable out from this model such that the resulting model would have the smallest possible RSS, which variable would you choose and why? (State will be considered as one variable.)

Construct fit4: a parsimonious model, using exhaustive subset search. Remove State from the data under consideration but include all other variables.

**c)** Show the Cp plot and also show the BIC plot. Based on the two plots which model size is most desirable to choose. Why?

**d)** Regardless of your answer in c), report the size 4 variable chosen by regsubsets, list those top four variables. To save time we will not pursue further.

# Question 5: Parsimonious models (LASSO)

Use LASSO for model selection, again making sure to do so without including the State variable in the LASSO process.

**a)** Run cv.glmnet() with set.seed(12). Plot cmv vs. lambda.

**b)** What is the lambda.1se value? Under the lambda.1se criterion, list the non-zero variables returned.

**c)** fit5: Run OLS with all the variables returned from part b), *and include State.* Are all the variables included here significant at the .01 level? If not, perform backward elimination (manually) until all the p-values for the remaining variables are less than .01. Show your model building process and report the final LS equations. *Note:* for this problem, force State as a whole, into the final model, i.e., do not remove State.

# Question 6: Graduation Evaluation

Independent from Question 5 - assume that we've decided to use fit6 as our final model.

fit6: Grad.rate ~ State + Schooltype + All.test.std

**a)** Are all three variables significant at the .01 level?

**b)** Provide the residual plot.

**c)** Provide the qqnorm plot of the residuals.

**d)** Are the linear model assumptions met?

**e)** Amy Gutmann's asked you to present a (short - max 6 sentences) executive summary of your findings to her, and suggest ways to improve Penn's graduation rate. Focus in particular on how each factor in fit6 affects Grad.Rate.

**f)** Finally, using fit6, provide a 95% prediction interval for Penn's graduation rate. Based on Penn's actual graduation rate, how would you consider our prediction performance?

# Question 7: Freedom of the Press

Newsweek did a great job of collecting granular data, but some schools are unhappy with their exact graduation figures being reported. They've lobbied Newsweek's publisher to report only whether a school's graduation rate is either High (Grad.rate >= 70) or Low (Grad.rate < 70), and the "journalists" have acquiesced to their corporate overlords. From now on, the only graduation rate available to you is in that high/low form.

**a)** Create a new categorical variable Grad.rate.2 in college\_data that fits the new Newsweek data specification. What proportion of the schools are categorized as "High Graduation", that is, Grad.rate.2 == "1"?

**b)** How well can we predict Grad.rate.2, with only three variables: State, Schooltype and All.test.std. Run a logistic regression of Grad.rate.2 vs. State, Schooltype and All.test.std. Is every variable significant at .01 level, while controlling for the other 2 variables?

**c)** Let us fix our classification threshold to 0.5, that is, we will classify each school to be "High Graduation" if the estimated probability of being "High Graduation" is greater than 0.5. Under this framework, what is the in-sample mis-classification error? Show your work.