

Midterm Exam 2

Read through the entirety of this exam before getting started

Exam Logistics

Reprinted from course website

1. The exam will be based on material from Chapters 2, 3, and 5 of ISLR, along with topics we discussed in class through Friday 10-1.
2. The exam will be made available on GitHub at 5pm PDT on Friday, 10-8. A link to the exam will be posted in the `#announcements` channel on Slack.
3. You must submit your completed exam via pushing any commits to GitHub prior to 9:00am PDT on Monday, 10-11.
4. This is a timed exam. You may take up to 3 hours to work on the exam between Friday and Monday. This time does not need to be spent consecutively. However, if you take a break from the exam, you should not spend the time during that break working on the exam, reviewing notes, or actively thinking about the problems.
5. You are responsible for keeping track of your own time on the exam and will be asked to provide an estimate for the total amount of time you spent.
6. You may freely consult the following references during the exam:
 - Your ISLR and Applied Predictive Modeling textbooks
 - Any course notes **you** have taken for this class
 - Cheatsheets or notes from other classes, provided **you** were the one to create the notes
 - Lecture slides on the course website
 - Homework problems you have submitted
 - Built-in RStudio help files and cheatsheets
7. You may not consult any other resources, including (but not limited to):
 - Classmates
 - Tutors
 - Other faculty
 - Other textbooks
 - Online help (stackexchange, message boards, slack, etc.)
8. If you run into problems while taking the exam, document the problem in your exam and message me on slack. I will try to respond as soon as I can, but can't guarantee I will be available at that moment.

Instructions

Each of the following 4 problems will be worth approximately equal number of points. Compose your answer to each problem between the bars of red stars. Show your work and justify your answers.

Problem 1

In this problem, you are asked to give an empirical demonstration of the bias-variance tradeoff.

The following code chunk generates 20 data points from a model with formula $y = 1 + x_1 + x_2 + \epsilon$, where $\epsilon \sim N(0, 1)$.

```
set.seed(1010)

n <- 20

x1<-rnorm(n,1,1)
x2<-rnorm(n,2,1)
e<-rnorm(n,0,1)
y<-1 + x1+x2+e

sim_data <- data.frame(x1,x2,y)
```

- a) In 2 - 4 sentences, define the bias-variance trade-off and describe why it is an important consideration in statistical modeling.
- b) Create a sequence of at least 5 models of varying flexibility that predict y based on x_1 and/or x_2 (these models do not need to be linear). Use these models to provide an explicit demonstration of the bias-variance tradeoff by computing MSE on both test and training data. Be sure to explain how your models demonstrate the BV-tradeoff, and support your answer with appropriate graphics.

Problem 2

Gordon Moore, co-founder of Intel, conjectured in the 1960s that the processing power of integrated circuits would double roughly every two years, an observation that has been codified as “Moore’s law.”

The following data set contains transistor counts for CPUs appearing between 1971 and 2015.

```
moore <- read_delim("data/moore.csv", ";",
  escape_double = FALSE, trim_ws = TRUE)
```

- a. Create a scatterplot showing the relationship between the number of `Transistors` and the `Date` the CPU was introduced. Comment on the relationship.
- b. Create a simple linear model predicting transistors as a function of date, and use your model to predict the number of transistors in 1970, 2000, and 2020. Discuss whether you feel these predictions are accurate.
- c. Create the diagnostic plot quartet for the model and discuss any concerns you have with the model based on the plots.
- d. Explain why if Moore’s law is true, we would expect to see a linear relationship between date and the base-2 logarithm of the number of transistors.
- e. Fit a linear model for the base-2 log of `Transistors` and `Date` (hint: the `log2` function computes the base-2 logarithm of a number). What is the coefficient on the slope of this model, and what does it represent in context?
- f. Create and analyze the diagnostic plot quartet for this model. Does it appear the conditions for inference are satisfied?
- g. Find the 95% confidence interval for the slope of the model.
- h. Based on your confidence interval, does Moore’s law seem plausible?

Problem 3

In 1995, Orley Ashenfelter published a study on the quality of Bordeaux wine vintages, which received considerable public attention after disputes with prominent wine critics, as shown in this video from an ABC interview (Watching the video is not necessary to solve this problem; however, it is only 5 minutes long, and don't need to count time spent watching the video towards your overall exam time).

Red Bordeaux wines are produced in the Bordeaux region of France, one of the most well-known wine growing regions in the world, and often command a high price. However, the quality of wines from this region may vary considerably due to many factors, including weather conditions.

In this problem, you will reproduce some of Ashenfelter's analysis to predict the quality of a Bordeaux vintage.

Load the data with the following code:

```
wine <- read_csv("data/wine.csv")
```

The data includes measurements of the following variables on 27 wines:

- Year: year in which grapes were harvested to make wine.
- Price: logarithm of the average market price for Bordeaux vintages according to 1990–1991 auctions. The price is relative to the price of the 1961 vintage, regarded as the best one ever recorded.
- WinterRain: winter rainfall (in mm).
- AGST: Average Growing Season Temperature (in Celsius degrees).
- HarvestRain: harvest rainfall (in mm).
- Age: age of the wine measured as the number of years stored in a cask.
- FrancePop: population of France at Year (in thousands).

In this problem, you will use `Price` as a surrogate measure for the quality of a wine.

- Create scatterplots and compute correlations for all pairs of variables in the `wine` data (hint: use `GGally`). Identify pairs of variables that have especially strong linear relationships, and give a context-based explanation for why these variables may be strongly correlated.
- Fit a multilinear model to predict `Price` as a function of **at least 2** predictors. Briefly explain why you chose to include / exclude the predictors you did.
- Interpret the coefficients in your model. What type of conditions lead to the highest quality wine?
- What is the p-value for the F-test for your linear model? What are the null and alternative hypotheses for the associated hypothesis test? What does the particular p-value you obtained indicate about your linear model?
- What are the values of the residual standard error for your linear model? Do you expect this number to be equal to, greater than, or less than the corresponding value calculated on a test set? Explain.
- Perform 10-fold cross-validation to estimate the root mean-squared error of your model.
- Based on your analysis in the previous parts of this question, do you feel that the **quality** of a Bordeaux wine can be accurately predicted based on weather? Answer in 3 - 5 sentences.

Problem 4

While I was Commons the other day, I saw the following note:

```
include_graphics("img/pumpkin_comp.jpg")
```



Your task is to describe a process for building a model to win this competition, using only the photo below (for reference, the Bon Appetit floor tiles are 12" x 12"). You cannot directly weigh the pumpkin in this picture, but you may assume you can go to the store to measure and weigh other pumpkins. (*Of course, I don't actually expect you to build a model for this midterm and/or to actually go to the store to weigh pumpkins, although you are welcome to do so after the exam in order to actually win the competition!*)

```
include_graphics("img/pumpkin.jpg")
```



- a. Is the goal of this task regression or classification? What is the response variable for this problem?
- b. Write down several predictors (at least 5, with at least one quantitative and one categorical predictor) that you could use to make your prediction. These predictors need to be ones that you can find the value of based on the picture alone. Which of these predictors are quantitative and which are categorical? For example, one predictor that you **cannot** use is *odor* (on a scale from 0 = **fresh** to 5 = **rancid**), since you can't discern this from the picture.
- c. For each of the predictors you identified in the previous part, describe the approximate relationship between that predictor and the response. For example, using the *odor* predictor, we might guess that stinky pumpkins weigh less, since some mass is lost due water evaporation during the rotting process.
- d. Describe any correlations you might expect to find between any of the predictors you listed in part (b).
- e. What is one *interaction* effect you may expect to see between predictors? Explain.
- f. What is one predictor with a *non-linear* relationship to the response. Explain.
- g. Suppose you can go to the store and measure data on 20 pumpkins. Describe how you could use information from this data estimate the variability in MSE for your model. Be sure to include specific terminology from our course.
- h. Describe how you could use the data from the store pumpkins in order to compare the quality of several candidate models. Be sure to include specific terminology from our course.
- i. Suppose each of the 19 students in our class (and me) independently build a model and submit a prediction based on the model. We agree to split the winnings equally if any prediction wins. Do you think we would be better off if we all collected data from the same 20 store pumpkins, or if we each collected data on different sets of 20 pumpkins each (assume we are not able to pool our data together to get a set of 400 pumpkins). Explain your reasoning using terminology from the course.
