Midterm Exam 1

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
- Cheatsheats 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 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 $\epsilon \sim N(0, 1)$.

```r
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.

---

a. The mean-squared error for a prediction on a test set can be decomposed as

$$\text{MSE}(\hat{f}(x_0)) = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2$$

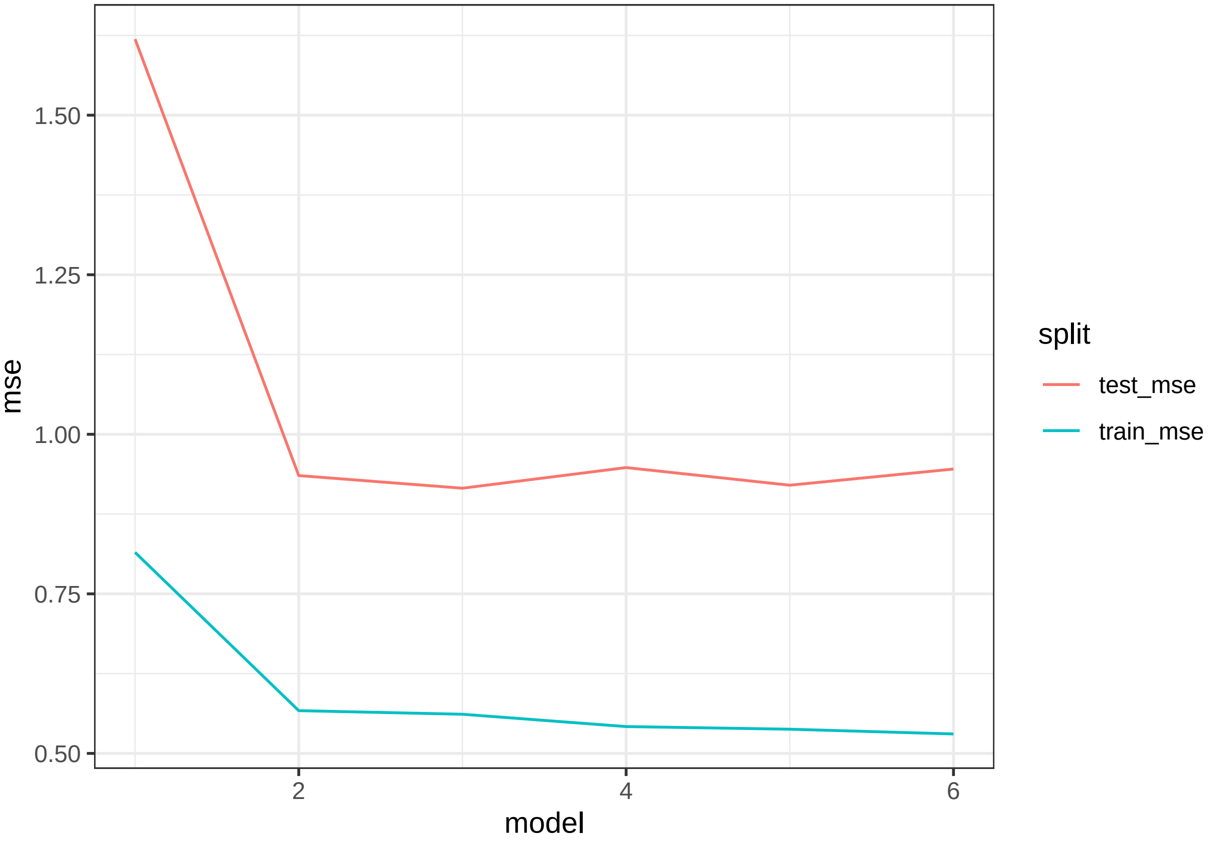
where variance indicates the variability in prediction at x_0 across a variety of different trainin sets, and where bias indicates the difference in the average prediction at x_0 from the true value at x_0. Minimal MSE is optained by simultaneously minimizing bias and variance. However, in many cases, decreasing the bias of a model comes at the cost of increasing variance, so the minimal MSE is not obtained by minimizing each separately. Generally, models with high flexibility tend to have lower bias and higher variance, while more rigid models have higher bias and lower variance.

b. We demonstrate the BV-tradeoff with the following sequence of models, which predict y as a function of x_1 and increasing polynomial powers of x_2. As the degree of the polynomial increases, the bias of the model decreases while the variance increases.

In particular, in the final plot which shows test mse and training mse for each model, we see that models with both high and low complexity have higher test mse than a model with moderate complexity. This is despite the fact that higher complexity models have consistently lower training mse. It is also worth noting that the model with the lowest test MSE is exactly the model which mimics the functional form of the true relationship between y and x_1, x_2.

```r
model_list <- list()
model_list[[1]] <- lm(y ~ x1, data = sim_data)
for (i in 2:6){
  model_list[[i]] <- lm(y ~ x1 + poly(x2, degree = i-1), data = sim_data)
}

train_mse <- c()
for (i in 1:6){
  preds <- predict(model_list[[i]], sim_data)
  train_mse[i] <- mean((sim_data$y - preds)^2)
}

set.seed(10)
e<-rnorm(n,0,1)
y<-1 + x1+x2+e

test_data <- data.frame(x1,x2,y)

test_mse <- c()
for (i in 1:6){
  preds <- predict(model_list[[i]], test_data)
  test_mse[i] <- mean((test_data$y - preds)^2)
}

bv_trade <- data.frame(model = 1:6, train_mse, test_mse)
```

```
bv_trade %>% pivot_longer(!model, values_to = "mse", names_to = "split") %>% ggplot(aes(x = model, y= ms
```



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 seen been codified as “Moore’s law.”

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

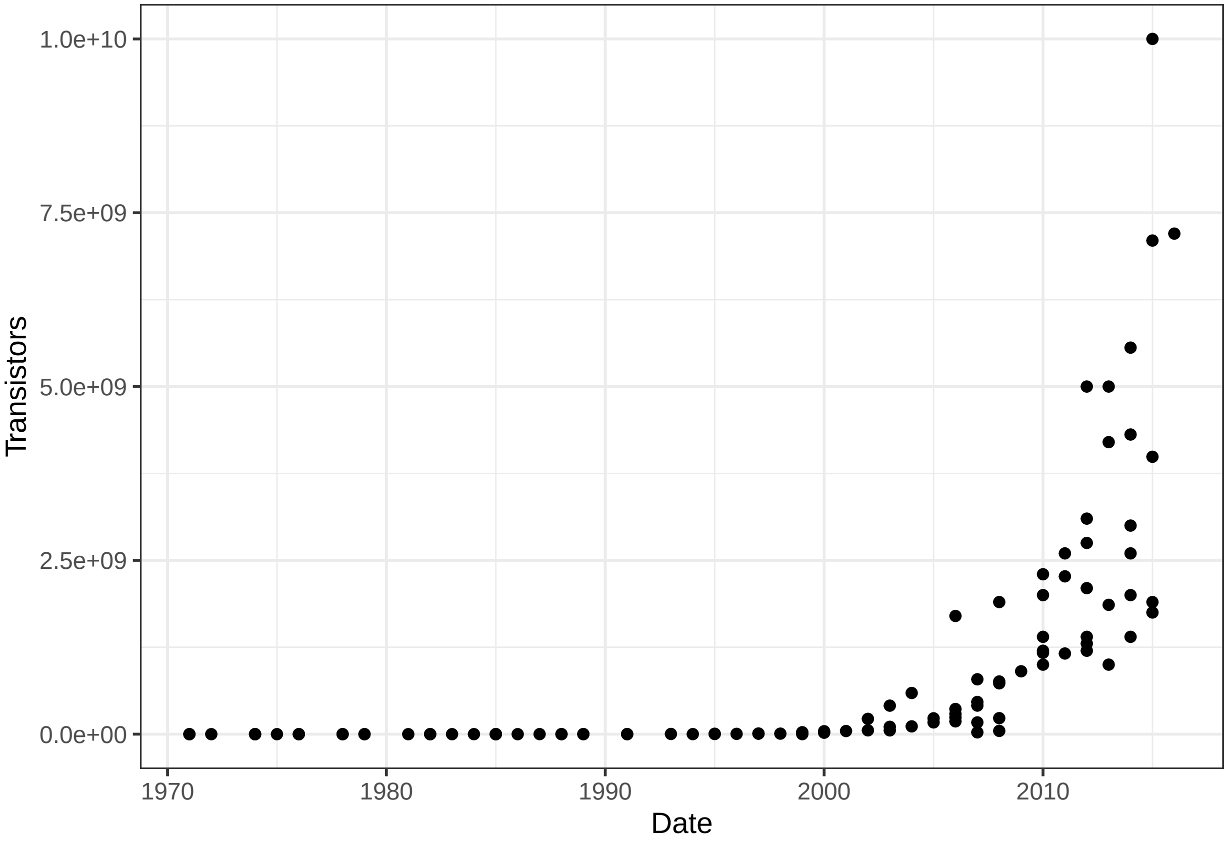
```
moore <- read_delim("data/moore.csv", ";",
    escape_double = FALSE, trim_ws = TRUE)
```

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 quarter 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?

---

a. The scatterplot shows a positive, but non-linear relationship between `Date` and `Transistors`.

```
ggplot(moore, aes(x = Date, y = Transistors))+geom_point()+theme_bw()
```



b. The model predicts -12.7×10^7 transistors in 1970, 10.7×10^8 transistors in 2000, and 26.3×10^8 transistors in 2020. The final seems too small, while the first number is negative (an impossibility)! These predictions are unlikely to be accurate.

```r
simple_mod_p2 <- lm(Transistors ~ Date, data = moore)

pred_set <- data.frame(Date = c(1970, 2000, 2020))
predict(simple_mod_p2, pred_set)
```

```
##           1           2           3 
## -1265241824  1069486032  2625971270
```

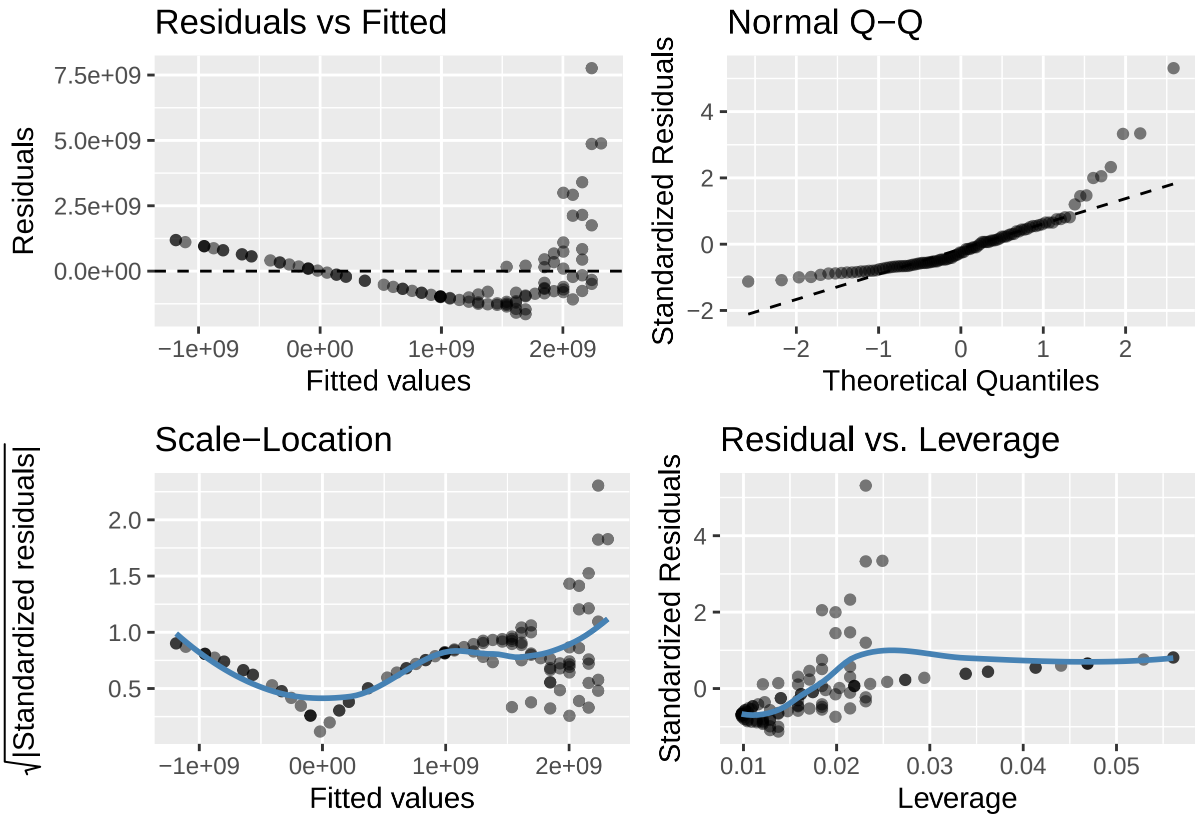
c. Each of the diagnostic plots shows some evidence of problem.

The residual plot shows strong evidence of non-lineary, the qq plot shows significant deviation from Normality in the tails of the distribution, the scale-location plot shows significant pattern to the size and location of residuals, and the leverage plot shows some influential points.

```r
library(gglm)
```

```
## Warning: package 'gglm' was built under R version 3.6.2
```

```r
gglm(simple_mod_p2)
```



d. If Moore's Law were true, then the relationship between transistors y and date x would take the form

$$y = a2^{x/2}$$

and so

$$\log_2 y = \log_2 a + \frac{1}{2}x$$

which is linear in x.

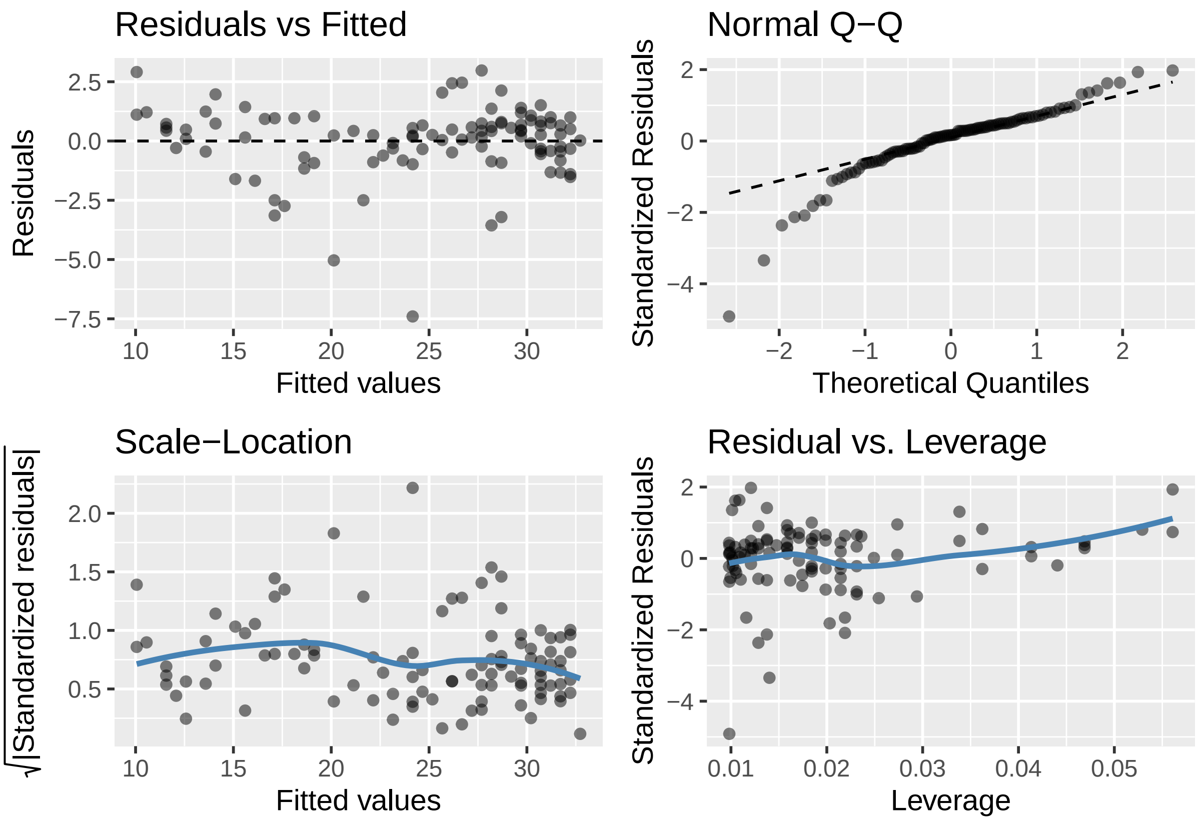
e. The coefficient on Date is 0.5, and indicates that every year the number of transistors increases by approximately 50%. Moreover, this would mean that every 2 years, the number of transistors increases by about 100% (aka the number doubles every two years).

```
log_mod_p2 <- lm(log2(Transistors) ~ Date, data = moore)
summary(log_mod_p2)
```

```
## 
## Call:
## lm(formula = log2(Transistors) ~ Date, data = moore)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -7.4011 -0.4731  0.2497  0.7531  2.9754 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -982.81973   22.73378  -43.23   <2e-16 ***
## Date           0.50374    0.01137   44.31   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.513 on 100 degrees of freedom
## Multiple R-squared:  0.9515, Adjusted R-squared:  0.9511 
## F-statistic:  1963 on 1 and 100 DF,  p-value: < 2.2e-16
```

f. The diagnostic plots show that a linear model is significantly more appropriate for the log number of tranistors. There are a few points, however, that appear to have sizable influence.

```
gglm(log_mod_p2)
```



g. The 95% confidence interval is (0.481, 0.526).

```
confint(log_mod_p2)
```

```
##                     2.5 %       97.5 %
## (Intercept) -1027.9229010 -937.7165600
## Date            0.4811872    0.5262976
```

h. The confidence interval does contain the value 0.5, which is slope we would expect to see if Moore's Law were true.

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: the average market price for Bordeaux vintages in dollars, based on 1990–1991 auctions.
- 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.

a. 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.

b. 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.

c. Choose one of the coefficients in your model and interpret it. In general, what type of conditions lead to the highest quality wine?

d. 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?

e. 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.

f. Perform 10-fold cross-validation to estimate the root mean-squared error of your model.

g. 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.

---

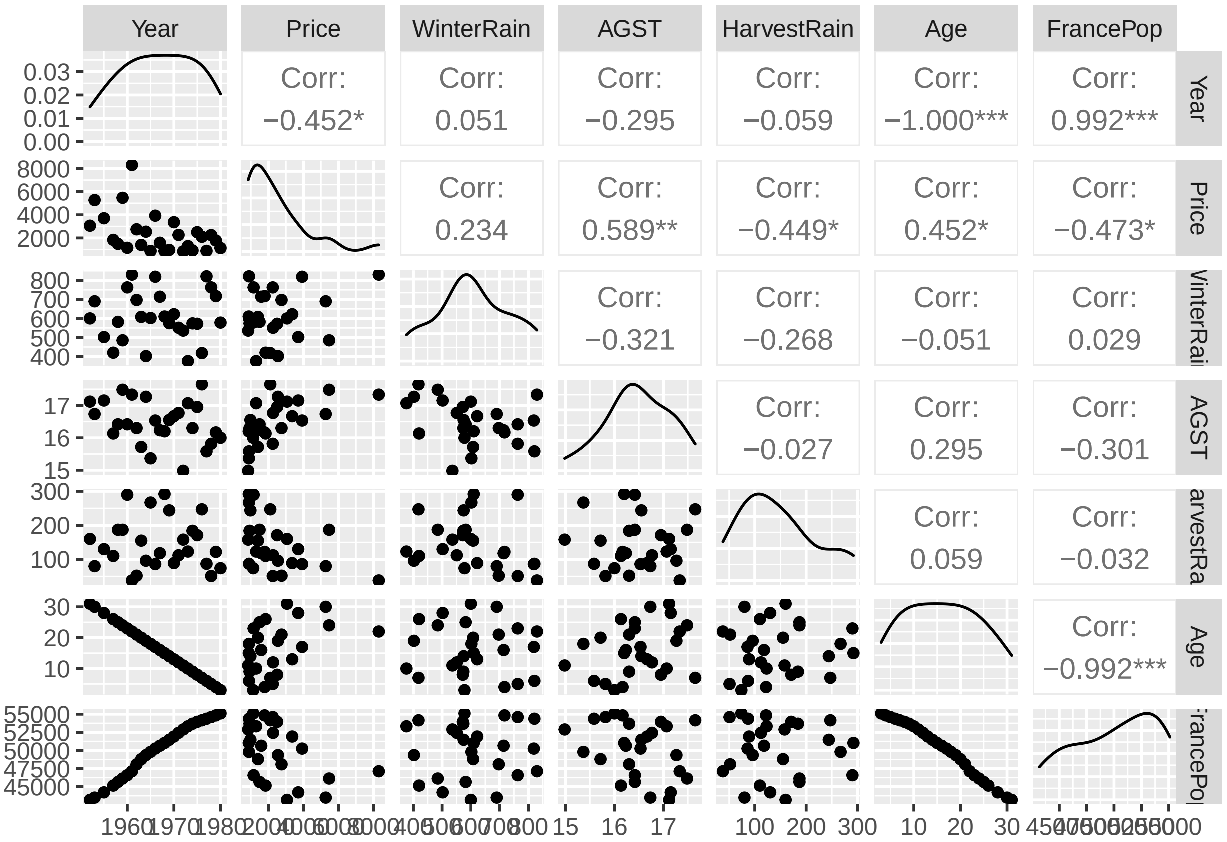
a. Based on scatterplots and correlation coefficients, we see that `Age`, `FrancePop` and `Year` all have extremely high pairwise correlation. This is not surprising, since age is just 1983 - Year, and the population of France is nearly a linear function of year.

```r
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 3.6.2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```r
ggpairs(wine)
```



b. Since Year is perfectly correlated with Age, we remove it. Moreover, since there is no plausible reason the population of France should influence wine price, we remove it as well, and fit the model on remaining variables.

```
mod_p3 <- lm(Price ~ . - Year - FrancePop, data = wine)
```

c.

- Every increase in the age of the wine increases its price by 66.6 dollars.
- Every increase in mm of Winter rain increases price by 4.67 dollars.
- Every increase in mm of Harvest rain decreases price by 8.58 dollars.
- Every increase in average growing temperature (in C) increases price by 1582

The highest value wines are those that are aged the longest, had rainy winters preceding growing season, had dry harvest seasons, and high average temperatures.

```
summary(mod_p3)
```

```
##
## Call:
## lm(formula = Price ~ . - Year - FrancePop, data = wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1160.48  -726.20   -91.91   279.47  2218.67
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -26365.316   5745.774  -4.589 0.000143 ***
## WinterRain       4.677      1.641   2.850 0.009304 **
## AGST          1582.338    323.948   4.885 6.97e-05 ***
## HarvestRain     -8.580      2.749  -3.122 0.004969 **
## Age             66.661     24.393   2.733 0.012149 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 975.2 on 22 degrees of freedom
## Multiple R-squared:  0.7361, Adjusted R-squared:  0.6882
## F-statistic: 15.34 on 4 and 22 DF,  p-value: 3.927e-06
```

d. The p-value for the F-test was 2.927×10^{-6}, which is significant at the 0.001 level, and gives good evidence against the hypothesis that there is no linear relationship between predictors and response.

e. The residual standard error was 975.2, and indicates that typically, the true price and the price predicted by the model differed by about 975 dollars. This is optimistic, since it is based on training data, and the corresponding measure would likely be higher on test data.

f. Based on 10-fold cv, the test rmse was 1065.042 dollars.

```r
set.seed(10)
library(rsample)
```

```
## Warning: package 'rsample' was built under R version 3.6.2
```

```r
prob3_cv <- vfold_cv(wine, v= 10)
```

```r
get_rmse <- function(split){
  train <- analysis(split)
  mod <- lm(Price ~ . - Year - FrancePop, data = train)
  val <- assessment(split)
  preds <- predict(mod, val)
  rmse <- sqrt(
    mean(
      (val$Price - preds)^2
    )
  )
}
```

```r
prob3_cv$rmse <- map_dbl(prob3_cv$splits, get_rmse)

mean(prob3_cv$rmse)
```

```
## [1] 1065.042
```

g. The model diagnostics, F-statistic, adjusted Rˆ2, and rmse all suggest that the model is relatively accurate in predicting the price of a bottle of wine. However, it is worth considering the relationship between price and quality. Without knowing about market efficiency, we do not have a good idea about whether price is a good indication of quality.

Problem 4

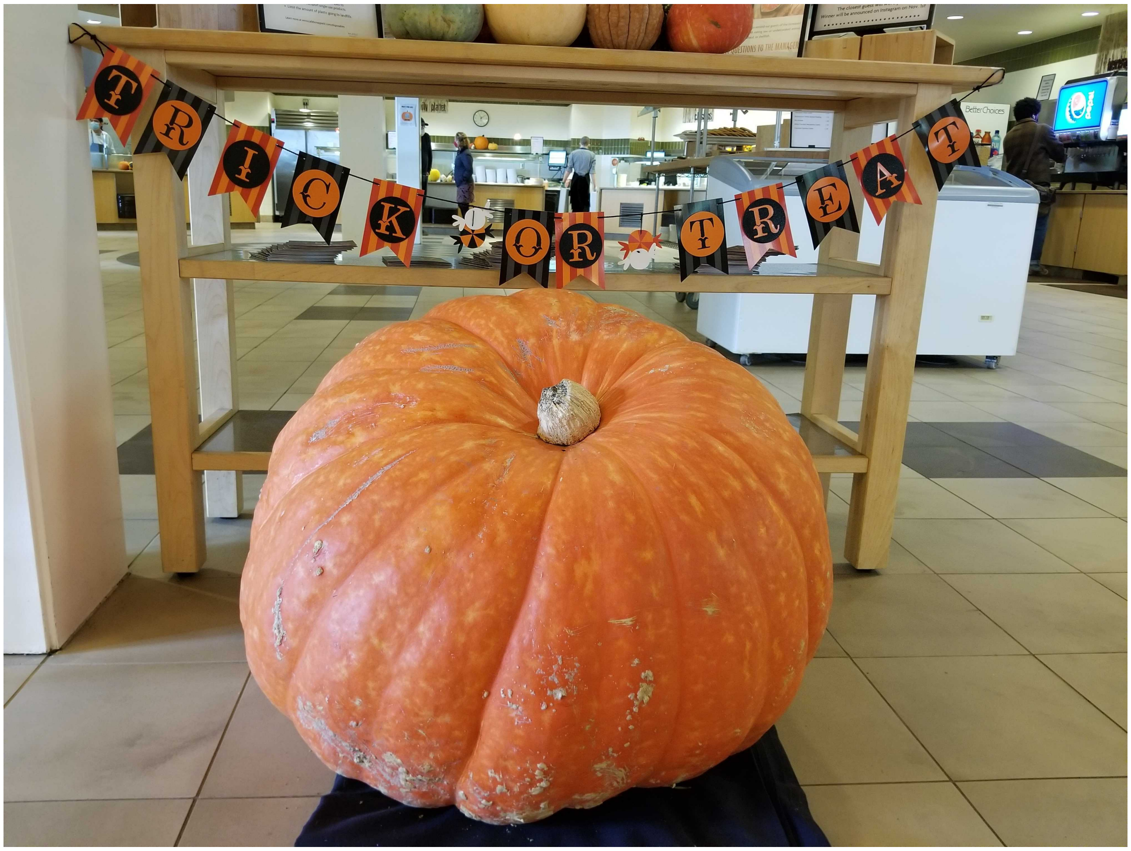
While I was Commons the other day, I saw the following note:

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

WHAT'S YOUR GUESS
Can you guess the weight of our pumpkin?
ENTER HERE
The closest guess will win a prize!
Winner will be announced on Instagram on Nov. 1st

Your task to describe a process for building a model to win this competition, using only the photo below (for reference, the Bon Appetit floor tiles ares 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? Is the goal primarily prediction or inference?

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.

a. This is a regression task, since the response variable `weight` is quantitative. This tasks primarily a prediction task, since we are interested in accurately guessing the weight of this particular pumpkin.

b.
- Pumpkin height in inches (from base to tip of stem); numeric
- Pumpkin width in inches (from left to right at widest point); numeric
- Pumpkin irregularity (spherical, irregular); categorical
- Color (orange, white, green, multicolored); categorical
- Wrinkles (few, some, lots); categorical
- Brightness (dull, moderate, bright); categorical

c.
- Pumpkin height: larger pumpkins, larger weight
- Pumpkin width: larger pumpkins, larger weight
- Pumpkin irregularity: more regular pumpkins, larger weight
- Color: orange pumpkins, larger weight (no idea)
- Wrinkles: fewer wrinkles, larger weight
- Brightness: bright pumpkins, larger weight

d. We may expect pumpkin height and pumpkin width to be positively correlated. Additionally, we expect pumpkins with height and weight roughly equal to be more regular.

e. Pumpkin color and height may have an interaction effect, since color may indicate subvariety of pumpkin, which may grow at different densities.

f. Pumpkin height likely has a non-linear relationship with weight, since weight depends on the volume of a pumpkin which is often related to the cube of its height.

g. After measuring the weights and values of other predictors for 20 pumpkins at the store, we create 1000 bootstrap samples, build an MLR model on each, compute the MSE for each, and then compute the standard deviation of the collection of MSEs.

h. After measuring the weights and values of other predictors for 20 pumpkins at the store, we can perform k-fold CV for each model to estimate the MSE and compare the results.

i. With only 20 observations, each model is very susceptible to differences between the training set and the population of pumpkins. We are almost certainly better off collecting data on different sets of 20 pumpkins, to effectively boost our sample size.