IE 5561, Spring 2019
Homework 1
Due February 8 at 11:55 pm

Instructions:

- You can use notes, books, the Internet, or any other source. You can talk to anyone, but you must write and submit your own code.

Problem 1

Suppose we have a data set with five predictors, X_1 = GPA, X_2 = IQ, X_3 = Gender (1 for Female and 0 for Male), X_4 = Interaction between GPA and IQ, and X_5 = Interaction between GPA and Gender. The response is starting salary after graduation, in thousands of dollars. Suppose we fit a regression model to the data and get $\hat{\beta}_0 = 50,\ \hat{\beta}_1 = 20,\ \hat{\beta}_2 = 0.07,\ \hat{\beta}_3 = 35,\ \hat{\beta}_4 = 0.01,\ \hat{\beta}_5 = -10$.

1. For a fixed value of IQ and GPA, who makes more money? Males or females? Can you answer this question? If not, describe when one gender makes more money?
2. Predict the salary of a female with an IQ of 110 and GPA of 4.0.
3. True or false: Since $\hat{\beta}_4$ is very small, there is very little evidence that the interaction between IQ and GPA is relevant to starting salary. Justify your answer.

Problem 2

Use Equation (3.4) from ISLR to show that for single variable regression, the line of best fit always goes through the point $(\bar{x}, \bar{y})$. (Those are the means of the x and y data.)

Problem 3

We will now investigate multi-colinearity. Run the following lines of code in R:

```
set.seed(1)
x1 = runif(100)
x2 = 0.5*x1 + rnorm(100)/10
y = 2 + 2*x1 + 0.3*x2 + rnorm(100)
```

Here by making `x2` depend on `x1` we have created multi-colinearity between `x1` and `x2`.

1. If `y` is the response variable and `x1, x2` are the regressors, what should the coefficients of the regression be?
2. What is the correlation between `x1` and `x2`?
3. Fit a multiple regression model where `y` is the response variable and `x1, x2` are the regressors. Describe the results. What are the estimated coefficients? Are these close the the true values? Are all of the coefficients statistically significant? Why or why not?

Problem 4

Read section 1.1 of the Analytics Edge textbook (attached here). Download the files `Wine.csv` and `WineTest.csv`. The first file is what we will use to train our model. We will then use the second file to test our model out-of-sample.

1. Inspect the data by making summaries and plots.
2. Re-create Figure 1.1 in the Analytics Edge textbook. First identify the observations in the training set that are above the average price. Then plot growing season temperate vs. harvest rain amount and color the observations by whether they are above average. Add a legend to the plot.
3. Fit a linear regression model with price as the response variable and growing season temperature, winter rain amount, harvest rain amount, and age (i.e. vintage) as the predictor variables.
4. Do the error terms in the model appear to have a Normal distribution? Justify your answer with some graphs.
5. In the summary output of the fitted model, the residual standard error is reported to be 0.295. Compute this quantity "by hand". In other words, use the actual values from the data and the fitted values from the model to compute the residual standard error yourself, using R to do the math... The formula is
$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - m - 1}}$$
6. What is interpretation of the coefficient on age?
7. Plot `Price` as a function of `AGST`. Overlay the regression line onto the plot. You will need to choose reasonable values for the other variables in the model, such as the mean.
8. Use the fitted model to make predictions for the test data (`WineTest.csv`).
9. Assess the accuracy of the predictions by computing the mean absolute prediction error. If y_i and $\hat{y}_i$ are the actual and predicted prices for observation i, respectively, then the mean absolute prediction error is
$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$
10. *Note:* The model in the book uses the natural logarithm of price as the dependent variable. Re-fit the model using `log(Price)`. Do you see any difference in model fit or in the quality of the plots?
11. How do you feel about using a model to make predictions that are outside the range of the data that was used to fit the model?

Some helpful R functions for this problem: `residuals()`, `coef()`, `curve()`, `qqnorm()`, `qqline()`, `sqrt()`, `predict()`.

Create a single RMarkdown file with the answers to all questions. Be sure to include your graphs and discussion. Submit the html file on Canvas.

On March 4, 1990, the *New York Times* announced that a Princeton University economics professor can predict the quality of wine without tasting a single drop. This professor believes that his predictions are more accurate than those of the world's most influential wine critic, Robert Parker, who called the predictions "ludicrous and absurd." These predictions have nothing to do with assessing the aroma, looking at the color, or determining the flavor profile of the wine; they are the results of a mathematical model.

Meanwhile, in a completely unrelated field, a study was being performed that ultimately reported shocking statistics about the quality of healthcare in the United States. One of the results of this study was the discovery that over half of 2,000 diabetic adults did not have a dilated eye exam in the past year, although diabetes is the leading cause of new cases of blindness among adults. They stated that "the dominant finding of our review is that there are large gaps between the care people should receive and the care they do receive." In research at the Massachusetts Institute of Technology, two analytics professors and a physician developed a mathematical model to assess the quality of care of patients in a fraction of the time that it takes a physician.

Mathematical models have even disrupted the fields of political science and law. In 2004, two law professors and two political science professors published an article claiming that a statistical model is better at predicting the results of Supreme Court cases than the collective opinions of experts. They predicted the affirm/reverse decisions of every Supreme Court case in the October 2002 term, cases including deep issues such as the constitutionality of affirmative action and various free speech rights. Lawyers, legal academics, and specialized journals that follow the Court closely have had little success in consistently predicting Supreme Court decisions in advance. However, these professors claim that the unbiased and unemotional nature of a model can better capture the decisions of the court.

In this chapter, we explore the possibility that predictive mathematical models can outperform expert human judgment by analyzing three different examples. While people are not always consistent in their opinions, are often emotional, and get tired, models are consistent, unemotional, and fast. We will show in this chapter that these characteristics lead to a competitive edge when models are used, especially if they are constructed using human expertise and judgment.

1.1 Predicting the Quality and Prices of Wine

Orley Ashenfelter made his prediction for red wine that is produced in the Bordeaux region of France, commonly referred to as "Bordeaux wine." He sought to address the mystery that while this wine has been produced in much the same way for hundreds of years, there are differences in price and quality from year to year that are sometimes very significant. Bordeaux wines are widely believed to taste better when they are older, so there is an incentive

to store young wines until they are mature. The main problem is that it is hard to determine the quality of the wine when it is so young just by tasting it since the taste will change so significantly by the time it will actually be consumed. This is why wine tasters and experts are helpful; they taste the wines and then predict which ones will be the best several years into the future. However, Ashenfelter observed that "bad" vintages are usually overpriced when they are young, and "good" vintages may sometimes be underpriced when they are young. Realizing that the advice of experts was making the market for young wines inefficient, he developed a different system for judging the wines.

Ashenfelter discovered two explanations for the variation in prices: the age of the vintage, and the weather. He hypothesized that since older wines have been held longer, they must be more expensive than younger wines. The fact that the quality of the grapes depends on the weather was widely understood. However, Ashenfelter pointed out that weather in the Bordeaux region can vary dramatically from one year to the next. Figure 1.1 is a scatterplot of the average growing season temperature (in degrees Celsius) versus the amount of harvest rain (in milliliters) in Bordeaux from 1952-1980. Each point is a year, and the years with higher than average prices are shown as triangles. (A year is marked as higher than average if the price is higher than the average price across all years in this dataset.) In this figure, and throughout the rest of this section, the price for a year is computed according to a price index developed by Ashenfelter. His price index took into account the results of several thousand auction sales for wines from many different wineries in the corresponding year. For more information, see the references in Section 1.5.

Figure 1.1 establishes that it is hot, dry summers that produce vintages in which the mature wines obtain the higher prices. Additionally, the data is fairly consistent; there are very few cases that do not adhere to this rule. This observation led Ashenfelter to build a linear regression model for the price of mature wine as a function of the age of the vintage and the weather. For more about linear regression, see Chapter 21.

A sample of the data he used is in Table 1.1. The regression equation is for the logarithm of the average price of the year, normalized relative to the highest selling year in the data, 1961. For more about why Ashenfelter used the logarithm of price, see Section 1.5. The "Price" variable referenced here is the price index developed by Ashenfelter.

Using this data, Ashenfelter published what is now known as the "Bordeaux equation":

$$\begin{aligned}\text{Log(Price)} = &-12.145 + 0.001173 \times (\text{Winter Rainfall}) \\ &+ 0.616 \times (\text{Average Growing Season Temperature}) \\ &- 0.00386 \times (\text{Harvest Rainfall}) + 0.0238 \times (\text{Age of Vintage}).\end{aligned}$$

The units for each of the independent variables are given in Table 1.1. This regression equation has an R^2 value of 0.83 and all variables are statistically significant. Ashenfelter experimented with additional variables, but the predictions were remarkably robust to the addition of any other variables.

While this regression equation predicts the logarithm of the average price index of the wine when it is mature, it can be thought of as an equation to predict the quality of the wine. By the time the wine is mature, the quality has been realized, and it is generally believed that the price of the wine reflects the true quality.

In response to Ashenfelter's equation, Britain's *Wine* magazine said "the formula's self-evident silliness invites disrespect." The Bordeaux wine industry was outraged; how can this man claim to know anything about wine that he has never tasted? You would never listen to the opinion of a food critic who had never tasted the food, or to a movie critic who had never seen the movie. To almost everyone, Ashenfelter's approach was just as absurd.

The Weather Makes the Wine

To the surprise of many, Ashenfelter went on to prove that his predictions for the quality of wine are surprisingly accurate. Using his Bordeaux equation, he was able to predict the quality of vintages that had just been bottled and had never even been tasted by comsumers. In 1991, he predicted that the vintages

Figure 1.1: Average growing season temperature and harvest rain in Bordeaux 1952-1980. The years with an above average price index are shown as triangles. The price index was constructed by Ashenfelter to represent the results of several thousand auctions sales for many different wineries in the Bordeaux region.

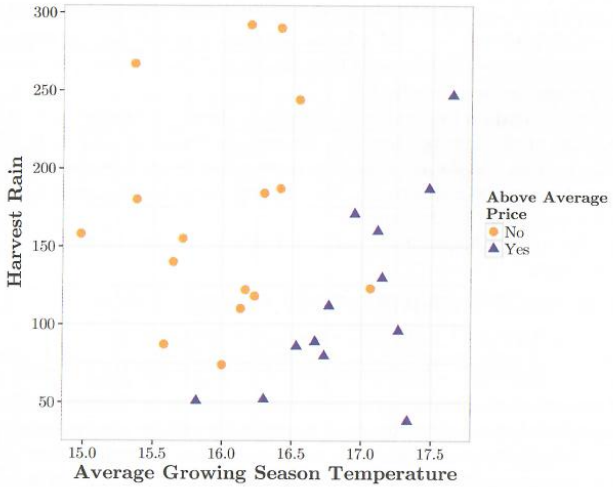

Table 1.1: Vintage and weather data used for the regression equation. The years 1954 and 1956 do not appear here because they were generally considered the worst in their decade and were no longer sold at the time Ashenfelter built his model.

Vintage	Log of Price	Winter Rain (ml)	Average Growing Season Temp (°C)	Harvest Rain (ml)	Age of Vintage (yrs)
1952	-0.99868	600	17.1167	160	31
1953	-0.45440	690	16.7333	80	30
1954		430	15.3833	180	29
1955	-0.80796	502	17.1500	130	28
1956		440	15.6500	140	27
1957	-1.50926	420	16.1333	110	26
1958	-1.71655	582	16.4167	187	25
1959	-0.41800	485	17.4833	187	24
1960	-1.97491	763	16.4167	290	23
1961	0.00000	830	17.3333	38	22
1962	-1.10572	697	16.3000	52	21
1963	-1.78098	608	15.7167	155	20
1964	-1.18435	402	17.2667	96	19
1965	-2.24194	602	15.3667	267	18
1966	-0.74943	819	16.5333	86	17
1967	-1.65388	714	16.2333	118	16
1968	-2.25018	610	16.2000	292	15
1969	-2.14784	575	16.5500	244	14
1970	-0.90544	622	16.6667	89	13
1971	-1.30031	551	16.7667	112	12
1972	-2.28879	536	14.9833	158	11
1973	-1.85700	376	17.0667	123	10
1974	-2.19958	574	16.3000	184	9
1975	-1.20168	572	16.9500	171	8
1976	-1.37264	418	17.6500	247	7
1977	-2.23503	821	15.5833	87	6
1978	-1.30769	763	15.8167	51	5
1979	-1.53960	717	16.1667	122	4
1980	-1.99582	578	16.0000	74	3

of 1989 and 1990 would be exceptional. Many professional wine critics did not agree with him at the time, but there is now a virtually unanimous agreement that 1989 and 1990 are two of the best vintages of the last 50 years. Ashenfelter further predicted that the 2000 and 2003 vintages are in the same league as the 1989 and 1990 vintages. These years have also been praised by the influential Robert Parker who has said "2000 is the greatest vintage Bordeaux has ever

produced." He has equally praised the 2003 vintage. Without tasting a single drop of wine, Ashenfelter was able to come to the same conclusion about the quality of the vintages as the man considered to be the most influential wine critic of our time. Additionally, his method does not require any tasting of the wine; just the use of a simple equation.

1.2 Assessing Quality in Healthcare

Perhaps no other domestic policy topic in the United States has spawned more debate in recent years than healthcare. However, even though there is a significant amount of disagreement about healthcare policies, the ultimate goal of politicians, physicians, hospitals, and patients is the same: *quality* healthcare. But what exactly is quality in healthcare? How is it defined, measured, and improved? One possibility is for an expert physician with many years of experience to look at cases and assess the quality of healthcare that patients have received. Clearly, this is impractical if one wants to make this assessment available for every patient in the healthcare system, as it takes an expert physician approximately an hour to read and understand each case before making an appropriate assessment. Also, how does the physician get the information to make this assessment?

Dimitris Bertsimas and David Czerwinski (two analytics professors), together with Michael Kane (a physician), built a model that assesses the quality of healthcare received by a group of patients. In this work, they define good quality care as healthcare that improves outcomes, educates patients, coordinates care among all doctors that see a patient, and controls costs. The goal is to capture the concerns of the patient, physician, and hospital.

This model is built using the opinion of a physician, and it is intended to accurately predict good or bad quality of care on cases not seen by the physician. With this method, an expert's opinion is used to assess the quality of care, but a model is developed to extend this opinion to all patients, without requiring the expert to evaluate each individual case. For this model, the researchers set the goal of predicting the opinion of Dr. Michael Kane, an internal medicine physician with over 40 years of experience. Keep in mind that the model could easily be extended to predict the average opinion of a committee of physicians, or to predict the opinion of a set of guidelines (this is discussed more later in this section). The key observation here is that the model predicts the opinion of a *domain expert*; this concept could be extended to many other applications and problems.

Claims data, the data that healthcare providers submit to insurance companies to be paid for various services, provides the data for this model. This was the most easily accessible and sufficiently large set of data available electronically and up to date. This data is not 100% accurate, and under-reporting is common, but other data sources are not very accessible. With the increasing use of electronic medical records, there is potential for more accurate and complete data in future calibrations of the model. We will also see claims data used for different applications later in this book.

The first step in building the model was asking a physician, in this case Dr. Kane, to rate the quality of care of a set of 101 patients. The patients selected were diabetic and between the ages of 35 and 55 with annual healthcare costs between $10,000 and $20,000. The creators of the model decided to test it with diabetic patients since there is a wide range of tests, medications, and complications associated with diabetes. The age range was to ensure that the patients should receive similar care (an 18 year old and an 81 year old will probably need very different care, partially due to their age difference), and the cost range was to insure that the patient had enough data to make an accurate assessment, but not so much data that it was impractical for Dr. Kane to review.

Dr. Kane rated the quality of care for each patient on a two point scale: poor care, or good care. He also gave his level of confidence that the patient was indeed receiving that quality of care. These ratings are summarized in Table 1.2. He also wrote a paragraph for each patient, explaining his reasoning. The following paragraph gives an example:

> Male on glucophage, had sporadic medical visits and labs. Did have eye exam 4/05. His primary problems were back pain and narcotic use. He had monthly percocet prescriptions in addition to an NSAID and a muscle relaxer. No diagnostic studies. Had a few physical therapy visits in October and November 2003. No other significant diagnostic or therapeutic initiatives. Poor care with high confidence.

From Dr. Kane's assessments, 80 different variables were defined that fell into six different categories: (1) diabetes treatment (e.g. number of glycated hemoglobin tests); (2) utilization (e.g. number of office visits); (3) markers of good care (e.g. mammogram); (4) markers of poor care (e.g. narcotics); (5) providers (e.g. number of doctors); and (6) prescriptions (e.g. number of different drugs). Since blindly classifying every patient as receiving good care gives an accuracy percentage of 78%, the goal was to be more accurate than this. In addition to accuracy, the types of errors that occur are also significant: the percentage of cases classified as poor care that are actually good care and the percentage of cases classified as good care that are actually poor care. Errors of the first type (good care mistakenly classified as poor care) may cause unnecessary expenses to investigate and treat perfectly fine patients. But errors of the second type (poor care mistakenly classified as good care) can be very dangerous. These errors mean that we are overlooking patients that need help, and might cause serious health issues in the future. Understanding the trade-off between the different types of errors for any model is critical in deciding how useful the model will be in practice.

The model uses logistic regression to predict the binary variable "Quality," which takes value 1 if the patient received good quality care according to the