Homework Assignment #3

Math 437 - Modern Data Analysis

Due March 20, 2023

Instructions

You should submit either two or three files:

1. You should write your solutions to the Simulation and Applied Problems in this R Markdown file and submit the (.Rmd) file.
2. You should knit the final solution file to pdf and submit the pdf. If you are having trouble getting code chunks to run, add `eval = FALSE` to the chunks that do not run. If you are having trouble getting R Studio to play nice with your LaTeX distribution, I will begrudgingly accept an HTML file instead.
3. Solutions to the Key Terms and Conceptual Problems can be submitted in a separate Word or pdf file or included in the same files as your solutions to the Simulation and Applied Problems.

This homework assignment is worth a total of **45 points**.

Key Terms (7 pts)

Read Chapter 3 of Introduction to Statistical Learning, Second Edition. Based on your reading, answer the following questions.

1. What is the difference between the *population regression line* and the *least-squares regression line*? When do we use the notation β_1 vs. $\hat{\beta}_1$?
2. What is a *residual*? Why are the residuals important in finding the $\hat{\beta}_j$?
3. Write a sentence to interpret what β_3 means in the model given by Equation (3.20). (Make sure to use appropriate units!)
4. Give two statistics that measure the fit (or lack thereof) of a multiple linear regression model. For each statistic, as it increases, does the F-statistic for an ANOVA test with $H_0 : \beta_1 = \ldots = \beta_p = 0$ increase or decrease?
5. Explain how to turn a factor variable into one or more *dummy variables* (indicator variables).
6. Refer to the model whose coefficient estimates are given in Table 3.8. If the *baseline* (reference level) changed to West, what would be the equation of the new least-squares regression plane?
7. What is the difference between a *main effect* and an *interaction effect*?
8. What kind of patterns should you look for in the *residual plot* to identify non-linear relationships between the predictors and the response? What kind of patterns should you look for to identify heteroscedasticity?
9. With what kind of data would you expect to see correlation of error terms when fitting a linear regression model? Why?
10. Give a rule of thumb for guessing that a point has an *outlier* residual. Give a rule of thumb for guessing that a point is a *high leverage* point.

11. Explain why creating a scatterplot matrix of the response variable y and all predictor variables is insufficient for detecting outliers and high leverage points.
12. Can *collinearity* be suspected based on inspecting the scatterplot matrix and correlation matrix? What about *multicollinearity*? Explain why/why not.

13. What is the *variance inflation factor*? When/why do we use it? How is it computed?
14. Briefly explain how *k-nearest neighbors regression* works. What are its advantages and disadvantages compared to linear regression?

Conceptual Problems

Conceptual Problem 1 (3 pts)

Textbook Exercise 3.7.3.

Conceptual Problem 2 (4 pts total)

Part a (3 pts)

Consider multiple linear regression without an intercept. Suppose that x_1 and x_2 have correlation $r_{x_1x_2}$. Find a closed-form expression for $\hat{\beta}$, a vector containing the values of β_1 and β_2 when

$$RSS = \sum_{i=1}^{n}(y_i - \beta_1 x_{i1} - \beta_2 x_{i2})^2$$

is minimized. Express your final expression in terms of three standard deviations (s_{x_1}, s_{x_2}, s_y) and three correlations ($r_{x_1x_2}$, r_{x_1y}, and r_{x_2y}).

HINT 1: Remember that a function cannot be minimized unless the gradient (vector of partial derivatives) is the 0 vector. You do not need to prove that the point at which the gradient is 0 is a local minimum instead of a local maximum or saddle point.

HINT 2: It may be easiest to do a bunch of algebra to get the β_1 and β_2 coefficients outside the sum before taking partial derivatives with respect to β_j.

HINT 3: Because we are doing linear regression without an intercept, the formulas for sample variance and sample correlation simplify to:

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} x_i^2$$

and

$$r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i}{s_x}\right)\left(\frac{y_i}{s_y}\right)$$

Part b (1 pt)

In Math 338 you likely learned that in simple linear regression, the slope estimate $b_1 = \hat{\beta}_1 = r\frac{s_y}{s_x}$. Show that for $j = 1, 2$, if x_1 and x_2 are uncorrelated, then $\hat{\beta}_j = r_{x_jy}\frac{s_y}{s_{x_j}}$ $(j = 1, 2)$.

For 1 pt extra credit: Show that under the additional assumption that $\hat{\beta}_1 \neq 0$ and $\hat{\beta}_2 \neq 0$, the converse also holds; that is, if $\hat{\beta}_j = r_{x_jy}\frac{s_y}{s_{x_j}} \neq 0$ $(j = 1, 2)$, then $r_{x_1x_2} = 0$.

Simulation Problems

Simulation Problem 1 (5 pts)

Textbook Exercise 3.7.13 parts (a)-(f).

Simulation Problem 2 (4 pts total)

Part a (Code: 1 pt; Explanation: 2 pts)

Textbook Exercise 3.7.14 parts (a)-(c).

Part b (Code: 0.5 pts; Explanation: 0.5 pts)

Center `x1`, `x2`, and `y` to have mean 0 (the slopes should remain the same after centering the data). What are the values of $\hat{\beta}_1$ and $\hat{\beta}_2$ that minimize MSE/RSS? Compare these values to the true β_1 and β_2 as well as the $\hat{\beta}_1$ and $\hat{\beta}_2$ obtained in 3.7.14c.

HINT: you should have derived a formula in Conceptual Problem 2, so all you have to do is find the right values based on your simulated data and plug them into the formula.

Applied Problems

Applied Problem 1 (7 pts total)

The code chunk below creates a `mpg_new` dataset from the `mpg` dataset in the ggplot2 package. Using this new dataset:

```
library(ggplot2)
library(dplyr)
mpg_new <- mpg %>%
  filter(year == 2008, class == "compact")
```

Part a (Code: 1 pt; Explanation: 1 pt)

Fit a simple linear regression model predicting highway gas mileage (`hwy`) from the fuel type `fl` (r = regular gas, p = premium gas). Interpret the `flr` coefficient in the summary. Is this coefficient significant (at the 5% significance level)?

Part b (Code: 1 pt; Explanation: 1 pt)

Fit a multiple linear regression model predicting highway gas mileage from the fuel type and the city gas mileage. Interpret the `flr` coefficient in the summary. Is this coefficient significant (at the 5% significance level)?

Part c (Code: 1 pt; Explanation: 1 pt)

Using the code in the Multiple Linear Regression Examples file as a guide, create an interaction plot showing city gas mileage on the x-axis and using `fl` as the `trace.factor`. What does the interaction plot reveal about the relationship between `cty` and `fl` that helps explain your results from part (b)?

Part d (Explanation: 1 pt)

Argue that the error terms in this model are not independent. (HINT: actually view the dataset and see what cars are in it.)

Applied Problem 2 (15 pts total)

Mediation analysis is a technique, often used in social science and health science, that combines causal modeling with multiple linear regression to estimate the effect of a potential confounding variable on the relationship between a predictor and response variable. The steps in mediation analysis are the following:

1. Propose a causal model explicitly naming the predictor variable X, response variable Y, and mediator M. The causal model proposes that the predictor variable causes changes in *both* the mediator and the response, and that the response variable is caused by changes in *both* the predictor and the mediator.
2. Fit a simple linear regression model predicting Y from X, $Y = c_0 + c_1 X$.
3. Fit a simple linear regression model predicting M from X, $M = b_0 + b_1 X$.
4. Fit a multiple linear regression model predicting Y from *both* M and X, $Y = c_0' + c_1' X + c_2' M$.
5. Compute the estimated *direct* effect of X on Y, c_1.
6. Compute the estimated *indirect* effect of X on Y through the mediator M. We can calculate this as either $c_1 - c_1'$ or $(b_1)(c_2')$.
7. Obtain bootstrap confidence intervals for the direct and indirect effects.

Part a (Code: 3 pts; Explanation: 6 pts)

Blake and colleagues (2020) asked 63 Australian women aged 18-40 to record a short video for a mock job interview.

The researchers proposed a causal model in which `sexual_motivation` (mediator M) mediates the relationship between `beautification` (predictor X) and `assertive_behavior` (response Y). These variables can be found in the `assertive_woman` dataset.

Perform an exploratory data analysis on this dataset. You should be able to answer the following questions:

- What does each variable in the dataset mean? How was it measured? (It would be a good idea to read the *Experiment 1 Procedure and materials* section of the linked paper.)
- Is there any missing data? If so, is there an obvious explanation for any of the missingness?
- What is the (univariate) distribution of each variable?
- Which pairs of variables appear to be related? Can you create a plot that shows potential relationships between all three variables?
- Do any women appear to be particularly unusual (i.e., are there any outliers or women outside the range of sensible values for one or more variables)?

Part b (Code: 2 pts)

Create the three regression models (as defined in the problem introduction) in R and obtain point estimates of the direct and indirect effect of `beautification` on `assertive_behavior`.

Part c (Code: 2 pts)

Create a function, `assertive_indirect_effect`, that takes in a data frame `df`, runs the three regression models from Part b using the data frame `df`, and returns the indirect effect.

Then, obtain a bootstrap 95% confidence interval for the indirect effect. It is probably easiest to use the `boot` and `boot.ci` functions in the `boot` package, but you can code your own bootstrapping if you really want to. The researchers used 95% BCa confidence intervals, but you do not need to replicate their work.

Part d (Explanation: 2 pts)

Based on your bootstrap confidence interval, do you find sufficient evidence to reject H_0 : there is no indirect effect of beautification on assertiveness as mediated via sexual motivation in favor of H_a : there is such an indirect effect? Explain your reasoning.