

Homework 3

Statistics 109

Due March 30, 2022 at 1:00 pm EST

Homework policies. Please provide concise, clear answers for each question. Note that only writing the result of a calculation (e.g., " $SD = 3.3$ ") without explanation is not sufficient. For problems involving **R**, include the code in your solution, along with any plots.

Please submit your homework assignment via Canvas as a PDF file.

We encourage you to discuss problems with other students (and, of course, with the course head and the TAs), but you must write your final answer in your own words. Solutions prepared "in committee" are not acceptable. If you do collaborate with classmates on a problem, please list your collaborators on your solution.

Max points: 100

Student Name:

PART 1 (25 points)

Perform the following commands in R:

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

The last line corresponds to creating a linear model in which y is a function of x_1 and x_2 .

- What is the correlation between x_1 and x_2 ? Create a scatterplot displaying the relationship between the variables.
- Using this data, fit a least squares regression to predict y using x_1 and x_2 . Describe the results obtained. What are b_0 , b_1 , and b_2 ? How do these relate to the true β_0 , β_1 , and β_2 ? Can you reject the null hypothesis $H_0 : \beta_1 = 0$? How about the null hypothesis $H_0 : \beta_2 = 0$?
- Now fit a least squares regression to predict y using only x_1 . Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?
- Now fit a least squares regression to predict y using only x_2 . Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?
- Do the results obtained in (b)–(d) contradict each other? Explain your answer.
- Now suppose we obtain one additional observation, which was unfortunately mismeasured.

```
> x1 <- c(x1, 0.1)
> x2 <- c(x2, 0.8)
> y <- c(y, 6)
```

Re-fit the linear models from (b) to (d) using this new data. What effect does this new observation have on each of the models? In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers.

PART 2 (25 points)

(Use TermLife.csv data file) Term Life Insurance: Here we examine the 2004 Survey of Consumer Finances (SCF), a nationally representative sample that contains extensive information on assets, liabilities, income, and demographic characteristics of those sampled (potential U.S. customers). We study a random sample of 500 families with positive incomes. From the sample of 500, we initially consider a subsample of $n = 275$ families that purchased term life insurance.

Note: For $n = 275$, we want you to subset the data so that you are only looking at rows where $FACE > 0$. Also, variable $LNFACE = \log$ of the face variable and $LNINCOME = \log$ of the income variable.

- (a) Fit a linear regression model of $LNINCOME$, $EDUCATION$, $NUMHH$, $MARSTAT$, AGE , and $GENDER$ on $LNFACE$.
- (b) Check if multicollinearity is present.
- (c) Briefly explain the idea of collinearity and a variance inflation factor. What constitutes a large variance inflation factor?
- (d) Supplement the variance inflation factor statistics with a table of correlations of explanatory variables. Given these statistics, is collinearity an issue with this fitted model? Why or why not?

PART 3 (25 points)

(Use condo.csv data file) A real estate agent wishes to determine the selling price of residences using the size (square feet) and whether the residence is a condominium or a single- family home.

- (a) Fit a regression model to predict the selling price for residences and provide the regression equation.
- (b) Interpret the parameters β_1 and β_2 in the model given in part (a).
- (c) Fit a new regression model now including the interaction term $x_1 * x_2$ and provide the regression equation.
- (d) Describe what including this interaction term accomplishes.
- (e) Conduct a test of hypothesis to determine if the relationship between the selling price and the square footage is different between condominiums and single-family homes.

PART 4 (25 points)

The data set fat (Library: UsingR) contains several body measurements that can be done using a scale and a tape measure. These can be used to predict the body-fat percentage (body.fat). Measuring body fat requires a special apparatus; if our resulting model fits well, we have a low-cost alternative.

- (a) Partition the data into 60% for training and 40% for testing. Use `set.seed(25)` before data partition.
- (b) Use training data to develop a multiple linear regression model with `body.fat` as response variable and age, weight, height, BMI, neck, chest, abdomen, hip, thigh, knee, ankle, bicep, forearm, and wrist as independent variables.
- (c) Use the `stepAIC` function to select a model. Report model summary and provide equation for this model.
- (d) What are the top three contributors to the body-fat percentage? Provide an interpretation for these three coefficients.
- (e) Develop a scatter plot for predicted and fitted response values using the testing data. Obtain R^2 using testing data based on predicted and fitted response values?