Assignment 4

Question 1 [2+2+2=6 **points**]: Suppose that for the model $y_i = \alpha + e_i$, the errors are inde-pendent with mean 0. Also suppose that measurements are taken using one device for the first n_1 measurements, and then a more precise instrument was used for the next n_2 measurements. Thus $Var(e_i) = \sigma^2$, $i = 1$, 2, . .., n_1 and $Var(e_i) = \sigma^2/2$, $i = n_1 + 1$, $n_1 + 2$, . .., n.

(a) Ignore the fact that the errors have different variances, and derive the least squares estimator for $\hat{\alpha}$ using matrix notation and $\boldsymbol{\hat{\alpha} = (X'X)^{-1}X'y}$.

(b) Derive the weighted least squares estimator for α, $\hat{\alpha}_{WLS}$.

(c) Suppose that $n_1 = n_2$. Compute the expected values and variances of the two estimators above. Which is a better estimator and why? (Use theoretical $MSE(\hat{\alpha}) = Bias^2(\hat{\alpha}) + Var(\hat{\alpha})$ as your definition of "better.")

Question 2 [5 **points**]: Solve question 2, Chapter 4

Question 3 [2+2+2=6 **points**]: Solve question 3, Chapter 4

Question 4 [2+2=4 **points**]: Return to Question 4 from Homework 2 , about coins being put on a scale. Now suppose that the variance in Y is proportional to the number of coins put on the scale. I recommend double-checking using both linear algebra and (if you're working in R) the linear model function lm($y \sim x$, weight $= w$), where w is a vector of weights (the diagonal of the weight matrix) and you invent your own y.

(a) Design an appropriate matrix of weights $\mathbf{W}$.

(b) Calculate the new least-squares estimates of the weights of the coins using weighted least squares.

Question 5 [**2+2+2+2=8 points**]: For the model $y_i = \beta_0 + \beta_1 x_i + e_i$, the errors are iid with mean 0. The four observed values of x_i are $\mathbf{x} = \begin{bmatrix} 1 & 2 & 3 & 4 \end{bmatrix}'$. The estimator of β_1 is $\tilde{\beta}_1 = (y_4 + 2y_3 - 2y_2 - y_1)/5$. For this model, do the following:

(a) Is $\tilde{\beta}_1$ unbiased? Show why or why not.

(b) What is the sampling variance of $\tilde{\beta}_1$?

(c) Get the usual least squares estimator of β_1 and calculate its sampling variance.

(d) Compare the sampling variance of $\tilde{\beta}_1$ with the sampling variance of the least squares estimator of β_1

Question 6 [**2+2+2=6 points**]: A food manufacturing company is interested in modeling whether people prefer $x_1 =$ Type A or Type B hotdog buns with their hot dogs. They also want to control for $x_2 =$ different amounts of sodium in the hot dogs themselves and are testing the hot dog buns at a variety of sodium contents, giving each taster both a hot dog and a bun with no condiments. The response variable is $y =$ perceived taste of the bun, on a scale of 1 to 10 .

(a) In order to find out whether Type A or Type B is preferred, is it necessary to have an interaction term? Why or why not?

(b) Develop a linear model for this study, interpreting all parameters in the context of the problem. Write down your hypotheses to be testing in terms of your model parameters. (You don't have any data to conduct the test; just write down the hypotheses.)

(c) Whether or not you added an interaction term above, assume now that it was added and it is statistically significant. How should we interpret this interaction in context?

Question 7 [2+2=4 **points**]**:** In a one-way ANOVA model with $k = 3$ groups and 4 observations per group:

(a) Use the F-statistic in Model Reduction Method 2 to derive a statistic for testing whether the average of the means of the first two groups is the same as the mean of the third group. That is, create the F-statistic for testing $H_0 : (\mu_1 + \mu_2)/2 = \mu_3$. (Hint: Don't fit a model with a y-intercept. It makes everything easier.)

(b) Where $\hat{\mu}_1 = 5.6, \hat{\mu}_2 = 7.9, \hat{\mu}_3 = 6.1$, and SSE $= 12.8$, test your hypothesis. Use $\alpha = 0.05$. Note that the degrees of freedom for the F-statistic are r (the number of rows of your $\mathbf{A}$ matrix and $n - p - 1$.

Question 8 [1+1+1+1+1+1+1=7 **points**]**:** Download the dataset called company.csv from Canvas. The dataset contains a systematic sample (every tenth company; we'll take these as randomly selected) for the Forbes 500 list. The variables of interest are Sales and Assets of the companies (both in millions of U.S. dollars). As with many financial datasets, many of these variables are skewed. Your job is to choose appropriate power transformations such that the relationship between Assets (response variable) and Sales (explanatory) are approximately linear.

(a) Begin by creating a scatterplot of Sales and Assets and fit a simple linear regression line. What transformations does your scatterplot suggest? Create diagnostic plots for this model (Model 1). Discuss any weaknesses of this model.

(b) Choose an appropriate transformation for Sales. Explain how you made your choice. Include plots if applicable.

(c) Choose an appropriate transformation for Assets, and again explain how you made your choice. Because using an inverse response plot in this example is messy, you can just (1) fit a regression model of Assets vs. the transformed version of Sales that you chose in part (b), then (2) pass the fitted model into the powerTransform function. No plots required.

(d) Call the model with both variables transformed Model 2. Create diagnostic plots for this model, and discuss any weaknesses of this model.

(e) Compare Model 1 and Model 2. Which model is preferable?

(f) Using the model $log(Assets) = \beta_0 + \beta_1 log(Sales)$, interpret the slope in the context of the problem.

(g) Again using the model $log(Assets) = \beta_0 + \beta_1 log(Sales)$, find a 95% confidence interval for the average assets of a company with $6,571$ million in sales, as Hewlett-Packard did. Interpret your confidence interval in context.

Question 9 [4 **points**]**:** When Y has both mean and variance equal to μ, we showed in the notes that the appropriate transformation of Y for stabilizing the variance is the square root transforma-tion. Now, suppose that Y has mean equal to μ and variance equal to μ^2. Show that the appropriate transformation of Y for stabilizing variance is the log transformation. (Question 7, Chapter 3, page 112 of the textbook.)