STAT 608, Spring 2022 - Assignment 3

1. Question 1, Chapter 3, p. 103.

2. Explain in words why when we create confidence intervals and prediction intervals using a transformed response variable Y, we can't simply take the inverse transformation of the endpoints to get a confidence or prediction interval in the original units of Y.

3. Recall the model with two indicator variables from question 3 of the previous homework. Calculate the hat matrix (use software if you like; it might be faster by hand). Explain what that projection matrix does and why it makes sense, as if to someone who has taken one semester of statistics.

4. For the simple linear regression model in the case that our assumption is met that the errors are independent and identically distributed with variance σ^2:

   (a) Show that the formula for the vector of residuals $\hat{\mathbf{e}}$ can be expressed compactly using the hat matrix:
   $$(\mathbf{I} - \mathbf{H})\,\mathbf{y}$$

   (b) Show that the covariance matrix of the residuals is therefore equal to
   $$(\mathbf{I} - \mathbf{H})\,\mathbf{\Sigma}\,(\mathbf{I} - \mathbf{H})',$$
   where $\mathbf{\Sigma}$ is the covariance matrix of the errors. Show that the covariance matrix of the residuals reduces to $(\mathbf{I} - \mathbf{H})\,\sigma^2$. (Please show that $\mathbf{H}$ is idempotent; that is, that $\mathbf{HH} = \mathbf{H}$.)

   (c) Conclude that $\text{Cov}\,(\hat{e}_i, \hat{e}_j) = -h_{ij}\sigma^2$, $i \neq j$.

5. For the simple linear regression model, show that the hat matrix $\mathbf{H}$ has the following properties:

   (a) $\mathbf{H}$ is symmetric.

   (b) $0 \leq h_{ii} \leq 1$, where h_{ii} is the i^{th} diagonal entry of the hat matrix. (**HINT**: First show that $h_{ii} \geq h_{ii}^2$ and note that $h_{ii} = \sum_j h_{ij}^2$.)

   (c) The off-diagonals of the hat matrix are found by the formula
   $$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})\,(x_j - \bar{x})}{SXX}.$$

   (d) Finally, the text states that "There is a small amount of correlation present in standardized residuals, even if the errors are independent." Comment on when the covariances of the residuals are close to zero, for a fixed sample size. Why does it make sense that the covariances are close to zero in those situations?

6. Under the simple linear regression model, $y_i = \beta_0 + \beta_1 x_i + e_i$, suppose the following data are collected and recorded: $x = [-3, -2, -1, 0, 6]$, $y = [10, 6, 5, 3, 12]$. Show your work, and do the following calculations by hand. (You may double-check with a computer.)

(a) First, create a quick sketch of the scatterplot of the data. Label any potential outliers or leverage points as such. Write out your design matrix.

(b) Using $\hat{y} = 7.2 + 0.5x$, calculate the residuals $\hat{e}_i$.

(c) Compute the leverage for each observation. Use the rule $h_{ii} > 4/n$ to identify potential leverage points. Are any points of high leverage "good" or "bad"?

(d) Compute the variances of the residuals. (Assume the variance of the errors to simply be σ^2.)

(e) Compute the standardized residuals ($s = 3.755$). Comment no why this answer seems to conflict a bit with the answer to part (b) above.

(f) Comment on why the point with the highest leverage in this dataset had the smallest variance.

7. 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 transformation. 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.)

8. 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 $\log(\texttt{Assets}) = \beta_0 + \beta_1 \log(\texttt{Sales})$, interpret the slope in the context of the problem.

   (g) Again using the model $\log(\texttt{Assets}) = \beta_0 + \beta_1 \log(\texttt{Sales})$, 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.