STAT 504 - Homework 1

Due date: Wednesday, January 29. Please submit your homework to the 423/504 mailbox at Padelford B-313 on Wednesday by 5:00 PM on the designated day. Please submit the output and plots, but not your R code unless the quesiton specifically asks for it. Total possible points: 27 + 4 Bonus points.

1. (10 points) (ALR, 1.2 and 2.15). Use data `wblake` in R package `alr4`. You can obtain a description of the data set by loading package `alr4` and executing `?wblake`.
   (a) (1 point) Compute the means for each of the eight (age) subpopulations in the smallmouth bass data. Draw a plot of mean `Length` versus `Age`. Is there evidence for a linear relationship?
   (b) (1 point) Compute the standard deviations for each of the eight (age) subpopulations in the smallmouth bass data. Draw a plot of the standard deviations of `Length` in each subpopulation versus `Age`. Does the variance appear constant across the differenct age populations?
   (c) (2 points) Suppose that you want to estimate the relationship between the radius of a key scale (predictor) and the length of the fish (response). Make a scatterplot to investigate the possible linear relationship. Fit a simple linear regression *Length* $\sim$ *Scale* and print the `R` summary. What do you observe?
   (d) (3 points) Plot the histogram of the residuals from the above regression and the TA plot (Tukey-Anscombe plot, TA plot involves plotting residuals (vertical axis) vs. fitted values (horizontal axis)). Do the normality and the constant variance assumptions appear to hold? Does this linear model seem appropriate?

       `R hint:` Use

       ```
       plot(lm(wblake$Length~wblake$Scale), which = 1)
       ```
   (e) (3 points) Find the fitted value, the 95% confidence interval, and the 95% prediction interval for the new data point `Scale = 200`.

2. (12 points) (ALR, 1.1 and 2.16) The data in the file UN1 in R package `alr4` contains data on:
   - `PPgdp` - the 2001 gross national product per person in US dollars,
   - `Fertility` - the birth rate per 1000 women in the population in the year 2000,
   - `locality` - Place where the data was collected. (This variable is unlabeled in the data, but is listed as the row name.)

   The data are collected from 193 localities, mostly UN member countries. In this problem, we will study the conditional distribution of `Fertility` given `PPgdp`.
   (a) (1 point) Identify the predictor and the response.
   (b) (1 point) Draw the scatterplot of `Fertility` on the vertical axis versus `PPgdp` on the horizontal axis and summarize the information in this graph. Does linear model seem appropriate here?
   (c) (1 point) Draw the scatterplot of `log(Fertility)` versus `log(PPgdp)`, using the logarithm with base 10. Does the simple linear regression model seem plausible for a summary of this graph?

(d) (1 point) Fit a simple linear regression to the log transformed data from c and print the summary.

(e) (3 points) Look at the histogram and TA-plot of the residuals. What can you say about the assumptions on the errors?

(f) (2 points) Test the null hypothesis that the slope is zero versus the two-sided alternative at the 1% level. Give the t-value and a sentence to summarize the result.

(g) (3 points) Plot the marginal 99% confidence intervals for the intercept (β_0) and slope (β_1) in the model from c as well as the 99% confidence ellipse for vector $(\beta_0, \beta_1)^T$. Would you reject the hypothesis $(\beta_0, \beta_1)^T = (1.1, -.2)$ at the 1% level?

3. (3+2 Bonus points) (Lecture SLR I) Suppose that

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \text{and} \quad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2),$$

and let $\hat{\beta}_0$ and $\hat{\beta}_1$ be OLS estimates of β_0 and β_1 based on n samples.

(a) (1 point) Find d_i such that

$$\hat{\beta}_0 = \sum_{i=1}^{n} d_i y_i.$$

Hint: You can use the fact that $\hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i$, $c_i = \frac{x_i - \overline{x}}{SXX}$, and $SXX = \sum_i (x_i - \overline{x})^2$.

(b) (1 point) Show that

$$\mathrm{E}[\hat{\beta}_0 | X = x] = \beta_0.$$

Hint: You can use the result from the previous step.

(c) (2 Bonus points) Show that

$$\mathrm{Var}[\hat{\beta}_0 | X = x] = \sigma^2 (\frac{1}{n} + \frac{\overline{x}^2}{SXX}).$$

Hint: You can use the result from the previous two steps.

(d) (1 point) Show that

$$\sum_{i=1}^{n} (y_i - \hat{y}_i) = 0.$$

4. (2 + 2 Bonus points) (ALR 2.17, **Regression through the origin**) Occasionally, a mean function in which the intercept is known a priori to be zero may be fit. This mean function is given by

$$\mathrm{E}[Y | X = x] = \beta_1 x,$$

The residual sum of squares for this model, assuming the errors are independent with common variance σ^2, is

$$RSS = \sum_{i=1}^{n} (y_i - \hat{\beta}_1 x_i)^2.$$

(a) (1 point) Show that the least squares estimate of β_1 is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

(b) (1 point) Show that

$$\mathrm{E}[\hat{\beta}_1 | X = x] = \beta_1.$$

Hint: You can use the result from the previous step.

(c) (2 Bonus points) Show that

$$\mathrm{Var}[\hat{\beta}_1 | X = x] = \frac{\sigma^2}{\sum x_i^2}.$$

Hint: You can use the result from the previous two steps.