Homework 10
Due Tuesday, December 2

1. Show that

   (a) $\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^{n} X_i(Y_i - \overline{Y})$

   (b) $\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}(X_i - \overline{X})X_i$

   (HINT: The technique used in solving Homework 1, Problem 5, may be helpful).

2. If we have a simple regression model, $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, $i = 1, \ldots, n$, we show that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \quad \text{and} \quad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

   by following the steps below:

   (a) We see in WMS, p. 570 that we obtain the values $\hat{\beta}_0$ and $\hat{\beta}_1$ by minimizing

$$\text{SSE} = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$$

   with respect to both $\hat{\beta}_0$ and $\hat{\beta}_1$ and setting them equal to zero, resulting in (first with respect to $\hat{\beta}_0$),

$$-2\left(\sum_{i=1}^{n} Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} X_i\right) = 0.$$

   Show from here that

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

   (b) Continuing, minimizing SSE with respect to $\hat{\beta}_1$ results in

$$-2\left(\sum_{i=1}^{n} X_i Y_i - \hat{\beta}_0 \sum_{i=1}^{n} X_i \; - \hat{\beta}_1 \sum_{i=1}^{n} X_i^2\right) = 0.$$

   Show, by plugging in $\hat{\beta}_0$ from part (a) to the above equation, that we obtain

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

3. If we have a simple regression model, with the additional information that ϵ_i are iid $N(0, \sigma^2)$, show that

   (a) $E(Y_i) = \beta_0 + \beta_1 X_i$

   (b) $\text{Var}(Y_i) = \sigma^2$

4. Using the information from Problems 2 and 3, show that $E(\hat{\beta}_1) = \beta_1$ (HINT: WMS, p. 578).
5. Using the information from Problems 2 and 3, show that

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

6. Suppose that we have a regression model, $Y_i = \beta_1(X_i - \overline{X}) + \epsilon_i,\ i = 1, \ldots, n$. Use the techniques outlined in Problem 2 to find $\hat{\beta}_1$ (HINT: This will be easier and have a familiar answer).
7. With `tempdata` from Lecture 21, verify the numbers on page 24 (follow the pages 7 to 15).
8. With `tempdata` from Lecture 21, (a) compute `yhat` and `ehat` (see page 8) and verify that the numbers match with `fitted(fit.tempdata)` and `resid(fit.tempdata)`, respectively, and (b) create a residual plot and a Q-Q plot, and comment on any observations.
9. Please find the dataset `simple1.csv`, where the response is `wage` and the predictor is `educ`. Perform the regression and verify the `summary` values. Provide all relevant plots. Comment.
10. With `fuel` data and `fit.fuel` from Lecture 22, create a residual plot and a Q-Q plot. Repeat with `fit.fuel2`. Try transformations (such as $\sqrt{Y}$ and $\log(Y)$) that may improve the model. Comment.
11. Please find the dataset `rat.txt` and its description `rat.description.txt`. Here, `y` is the response and all others are predictors. Perform the multiple regression analysis, providing any appropriate analyses, diagnostics and plots along the way. Comment.
12. Please find the dataset `bodyfat.txt` and its description `bodyfat.description.txt`. The purpose of this study is to see what variables (if any) contribute significantly to Percent body fat, in a multiple regression model. Perform the analysis, providing any appropriate analyses, diagnostics and plots along the way. Please present and defend your final model.