Homework 2
36-401/607 Section B (Wasserman) Due Friday, September 15 at 3:00

1. Suppose that
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \ldots, n$$
where $\mathbb{E}[\epsilon_i] = 0$ and $\text{Var}[\epsilon_i] = \sigma^2$. Let $\widehat{\beta}_0$ and $\widehat{\beta}_1$ be the least squares estimates.

   (a) Find $\text{Var}(\widehat{\beta}_0)$ (treating the X_i's as fixed).

   (b) The observed residuals are defined by $\widehat{\epsilon}_i = Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i)$. Show that $\sum_i \widehat{\epsilon}_i = 0$.

   (c) Show that $\sum_i \widehat{Y}_i \widehat{\epsilon}_i = 0$. What is the interpretation of this result?

   (d) Suppose you did a regression of the Y_i's on the observed residuals. (In other words, treat the $\widehat{\epsilon}_i$'s as the X_i's.) What intercept and slope would you get?

2. Suppose that
$$Y_i = \beta_1 X_i + \epsilon_i, \quad i = 1, \ldots, n$$
where $\mathbb{E}[\epsilon_i] = 0$ and $\text{Var}[\epsilon_i] = \sigma^2$. In other words, there is no intercept in this model. (This is called regression through the origin.)

   (a) Find the least squares estimate $\widehat{\beta}_1$ for this model.

   (b) Show that $\widehat{\beta}_1$ is unbiased.

   (c) Suppose you use your estimator from (a) but that, in fact, the true model is $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$. Show that the estimator from part (a) is biased and find an expression for the bias.

3. Simulation problem.

   (a) Generate $n = 100$ data points as follows. Take $X_i \sim \text{Uniform}(0, 1)$. Then set
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \ldots, n$$
where $\beta_0 = 5$, $\beta_1 = 3$ and $\epsilon_i \sim N(0, 1)$. Plot the data. Fit the regression line. Add the fitted line to the plot.

   (b) Repeat the experiment in part (a) 1,000 times. You will get a different value of $\widehat{\beta}_1$ each time. Denote these by $\widehat{\beta}_1^{(1)}, \ldots, \widehat{\beta}_1^{(1000)}$. What is the mean of these values? What do you expect the mean to be? Plot a histogram of $\widehat{\beta}_1^{(1)}, \ldots, \widehat{\beta}_1^{(1000)}$.

   (c) Repeat (b) but now take ϵ_i to have a Cauchy distribution. The Cauchy distribution looks like a Normal but it has very thick tails. How does the histogram change?

   (d) Now we will investigate what happens when the X_i's are measured with error. Generate $n = 100$ data points as follows:
$$\begin{aligned} X_i &\sim \text{Unif}(0, 1) \\ W_i &= X_i + \delta_i \\ Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i \end{aligned}$$

where $\beta_0 = 5$, $\beta_1 = 3$, $\epsilon_i \sim N(0,1)$ and $\delta_i \sim N(0,2)$. Suppose we only observe $(Y_1, W_1), \ldots, (Y_n, W_n)$. (We don't get to see the X_i's.) Plot the data. Fit the regression line. Add the fitted line to the plot. Now repeat this 1000 times and find, from the simulation, $\mathbb{E}[\widehat{\beta}_1]$. Also, plot a histogram of $\widehat{\beta}_1^{(1)}, \ldots, \widehat{\beta}_1^{(1000)}$. Based on this experiment, what is the effect of having errors in the X_i's.

4. Load the dataset `airquality` using the R command: `data(airquality)`.

   (a) Use the `summary` command to summarize the data. Use the `pairs` command to plot scatterplots of all pairs of data.

   (b) Plot Ozone versus Solar Radiation. (Put Solar on the x-axis.) Describe the relationship between these variables.

   (c) Fit a least squares regression line. Add the line to the plot and report the intercept and slope.

   (d) Compute the residuals $\widehat{\epsilon}_i = Y_i - [\widehat{\beta}_0 + \widehat{\beta}_1 X_i]$. Plot the residuals versus the X_i's. Does it appear that the standard linear regression model assumptions hold?