Homework 9 (problem 3 and 5)
Statistics 200B
Due Apr 25, 2019

1. Following the notation from class, define $RSS = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, $ESS = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$, and $TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$. Show that $TSS = ESS + RSS$.

2. Show that under the assumption of normality, the likelihood ratio test for $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ has the same form as the Wald test.

3. Consider the **regression through the origin** model:

$$Y_i = \beta X_i + \epsilon_i$$

   (a) Find the least squares estimate for β.

   (b) Find the standard error of the estimate.

   (c) Find conditions that guarantee that the estimator is consistent.

4. Read in the data file `cars.dat` from bCourse.

   (a) Make a scatterplot of HP (horsepower) against MPG (miles per gallon); that is, with HP on the y-axis and MPG on the x-axis. Experiment with taking logs of one or both variables until you find a combination that looks appropriate for the simple linear regression model. Turn in your plot, along with an explanation of how you evaluated the assumptions of the model.

   (b) Using the transformations you chose in (a), fit a simple linear regression model. Report $\hat{\beta}_0$ and $\hat{\beta}_1$, and carry out a Wald test of size $\alpha = 0.05$ for $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$.

   (c) Make two diagnostic plots as follows, and turn in each one with a one sentence interpretation.

   - Plot the x values (for whatever transformation you used) on the x-axis, and the residuals on the y-axis. Do you see evidence against the assumption of constant variance?
   - Make a plot comparing the quantiles of the residuals to the quantiles of a normal distribution. (See the help for `R` functions `qqnorm` and `qqline`.) Do you see evidence against the assumption of normality?

5. In this question we take a closer look at prediction intervals. Let $\theta = \beta_0 + \beta_1 X_*$, and let $\hat{\theta} = \hat{\beta}_0 + \hat{\beta}_1 X_*$. Thus, $\hat{Y}_* = \hat{\theta}$ while $Y_* = \theta + \epsilon$. Now, $\hat{\theta} \approx N(\theta, se^2)$, where

$$se^2 = V(\hat{\theta}) = V(\hat{\beta}_0 + \hat{\beta}_1 x_*).$$

Note that $V(\hat{\theta})$ is the same as $V(\hat{Y}_*)$. Now, $\hat{\theta} \pm 2\sqrt{V(\hat{\theta})}$ is an appropriate 95 percent confidence interval for $\theta = \beta_0 + \beta_1 x_*$ using the usual argument for a confidence interval. But, as you shall now show, it is not a valid confidence interval for Y_*.

(a) Let $s = \sqrt{V(\hat{Y}_*)}$. Show that

$$P(\hat{Y}_* - 2s < Y_* < \hat{Y}_* + 2s) \approx P\left(-2 < N\left(0, 1 + \frac{\sigma^2}{s^2}\right) < 2\right) \neq 0.95$$

(b) The problem is that the quantity of interest Y_* is equal to a parameter θ plus a random variable. We can fix this by defining

$$\xi_n^2 = V(\hat{Y}_*) + \sigma^2 = \left[\frac{\sum_i (x_i - x_*)^2}{n \sum_i (x_i - \bar{x})^2} + 1\right] \sigma^2$$

In practice, we substitute $\hat{\theta}$ for θ and we denote the resulting quantity by $\hat{\psi}_n$. Now consider the interval $\hat{Y}_* \pm 2\hat{\xi}_n$. Show that

$$P(\hat{Y}_* - 2\hat{\xi}_n < Y_* < \hat{Y}_* + 2\hat{\xi}_n) \approx P(-2 < N(0, 1) < 2) \approx 0.95.$$