BST 760: Advanced Regression
Breheny

Assignment 2
Due: Thursday, January 31

1. Carry out a simulation to compare the accuracy of the following estimators for the variance of a random variable Y:

$$\hat{\sigma}_1^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}$$
$$\hat{\sigma}_2^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}$$

   The first estimator is the usual estimator for the variance; the second uses n instead of $n-1$ in the denominator (and is, incidentally, the MLE if Y follows a normal distribution). In the simulation, generate $n = 10$ observations from the standard normal distribution, estimate $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$, and repeat this process 10,000 times (*i.e.*, you will end up with 20,000 estimates, 10,000 for each estimator).

   (a) Take the average of the 10,000 estimates for each estimator. Which average is closest to the true value of σ^2?

   (b) Find the average squared distance away from the true value of σ^2 (this is known as the *mean squared error*) for each estimator. In other words, find the average of $\left\{(\hat{\sigma}_{1,b}^2 - \sigma^2)^2\right\}_{b=1}^{10,000}$ and $\left\{(\hat{\sigma}_{2,b}^2 - \sigma^2)^2\right\}_{b=1}^{10,000}$, where $\hat{\sigma}_{j,b}^2$ is the bth estimate for estimator j.

   (c) Comment on your findings in (a) and (b). Is one estimator superior to the other one, or does it depend on how the estimator is evaluated?

2. The course web page contains a data set `islands.txt` that records the numbers of reptile and amphibian species and the island areas (in square miles) for seven islands in the West Indies. The data come from the book *The Diversity of Life*, by E.O. Wilson (1992).

   (a) Plot the number of species versus the island area and comment on whether a linear model would be appropriate for these data.

   (b) Plot the log of `Species` versus the log of `Area` and comment on whether a linear model would be appropriate given these transformations.

   (c) Fit a linear regression model using log transformations for both Area and Species. For islands of size 40, 400, 4,000, and 40,000, how many species does the model predict that each island will have? Provide the estimate and the 95% confidence interval. Be sure to provide the estimates and intervals on the original scale, not the log scale.

(d) Let x_1, x_2 be two different values for island area, and let y_1, y_2 be the most likely number of species on islands 1 and 2 according to the model fit in part (c). Show that

$$\frac{y_1}{y_2} = \left(\frac{x_1}{x_2}\right)^{\beta_1},$$

where β_1 is the slope of the regression line.

(e) If island A is twice as large as island B (in area), island A will likely have _____% more species than island B.

(f) Take the plot that you made in part (a) and overlay the relationship between area and number of species that is estimated by the model in part (c). If you would like to shade in upper and lower confidence limits, that would be neat, but you don't have to.