Homework 3, Due in Class 10/28

October 23, 2014

Theory and Thinking

1. **Crickets Chirping**. Crickets make their chirping sounds by sliding one wing cover across the other. Biologists (6th grade science camp?) have long known that there is a linear relationship between the *temperature* and *frequency* with which a cricket chirps. The following table lists 15 frequency-temperature observations recorded for the striped ground cricket, *Nemobius fasciatus fasciatus*.

   - Plot these data. Do it by hand.
   - Find the equation of the least squares regression line. Do it by hand.
   - Suppose a cricket of this species is observed to chirp 18 times per second? What is the estimated temperature? How far off do you think your prediction will be, *on average*? Do it by hand.

Observation	Chirps per Second	Temperature
1	20.0	88.6
2	16.0	71.6
3	19.8	93.3
4	18.4	84.3
5	17.1	80.6
6	15.5	75.2
7	14.7	69.7
8	17.1	82.0
9	15.4	69.4
10	16.2	83.3
11	15.0	79.6
12	17.2	82.6
13	16.0	80.6
14	17.0	83.5
15	14.4	76.3

2. **College Game Day.** Why does warm beer taste "flatter" than a cold beer, all else equal? As it turns out, as beer temperature increases, the kinetic energy of the

solution increases, and more of the gas can escape from the solute phase to the gas phase. The following table gives the number of parts of CO_2 that are soluble in 100 parts of "no, it's *water*, officer."

Temperature Degrees C, (x)	**Parts Dissolved** Parts Dissolved (y)
0	66.7
4	71.0
10	76.3
15	80.6
21	85.7
29	92.9
36	99.4
51	113.6
68	125.1

Calculate the residuals for each data-point. Then plot these residuals with the Degrees (x) on the x-axis and the residuals $(y_i - \hat{y}_i)$ for each point on the y-axis. Does it look like fitting a straight line through the data is appropriate? Why or why not? Do all this work by hand too.

3. Why must a least-squares line necessarily pass through the point $(\bar{x}, \bar{y})$. (I'm asking you to prove this.)

4. In some regression situations there are *a priori* reasons to force the least squares line to pass through the origin (a within estimator comes to mind). Using the least squares optimization criteria that we demonstrated on the board – $S(A, B)$ – show that in this case

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$$

Programming, Rocking and Rolling

1. Read in the file at `http://polisci2.ucsd.edu/dhughes/dist.txt`.

2. This is data with numbers drawn from the same distribution for all rows, but the *unknown* population parameters for each of the "districts" might be different. This means that all rows might be drawn from some distribution ϕ, but District A may come from $\phi(10, 2, 0.2)$ while District B may instead be drawn from $\phi(3, 6, 9)$.

   (a) What do you think the distribution is that is generating these numbers?

   (b) What would it take for you to think your guess for the distribution is correct?

   (c) After you make a guess at the distribution, develop an estimator (or estimators) to attempt to recover *all* the sufficient parameters to describe the distribution.

(d) Describe these estimators using the four properties of estimators we care about. Does your estimator(s) meet these properties? Which?

(e) Now that you have the sufficient parameters of the distribution, test to see if you can distinguish your data from the reference distribution with these parameters.

(f) Finally, is there a difference in the medians of the districts? Use a permutation test to find out.

3. Anscombe (1981) presents and analyzes the data on the U.S. states and Washington, D.C. given in Table 5.1 and `http://socserv.socsci.mcmaster.ca/jfox/Books/Applied-Regression-2E/datasets/index.html`, for state per-capita public-school expenditures (in dollars), per-capita annual income (dollars), the proportion of residents under the age of 18 (per 1000) and the proportion of the population residing in urban areas (per 1000). Regress school expenditures on each of income, proportion under 18, and proportion urban. Plot the least-squares line for each regression on a scatterplot of the data. Does the line adequately capture the relationship between the variables shown in the plot? In each case, examine and interpret the values of A, B, S_e, and r (or r^2). A copy of the Anscombe data can be found on the textbook website.

4. Get ICPSR dataset # 9246. Think of an interesting bivariate regression, run it, present it in a clean table, and write up your question, measurement, and findings in a paragraph or so.

5. **BONUS!** Prove that the least-squares estimator is an unbiased estimator of β. That is

$$E(\hat{\beta} - \beta) = 0$$

assuming that (a) either the x's are fixed values, or (b) if they are random, that they are independent and orthogonal of the errors (ϵs), and the ϵ's, are *iid* following some distribution with finite variance.