HOMEWORK 2 – DUE APRIL 25

1. A regression model

In this part, you will replicate Yule's regression equation for the metroplitan unions, 1871-81. See Chapter 1 of Freedman (2009) for a discussion. Fix the design matrix $\boldsymbol{X}$ at the values reported in Table 3 there, available at

http://pages.uoregon.edu/dlevin/DATA/yule.txt

(Subtract 100 from each entry to get the percent changes.) Yule assumed

$$\Delta\text{Paup}_i = a + b \cdot \Delta\text{Out}_i + c \cdot \Delta\text{Old}_i + d \cdot \Delta\text{Pop}_i + \varepsilon_i$$

for 32 metropolitan unions i. For now, suppose the errors ε_i are IID, with mean 0 and variance σ^2.

(a) Estimate a, b, c, d and σ^2.
(b) Compute the SE's.
(c) Are these SEs exact, or approximate?
(d) Plot the residuals against the fitted values. (This is often a useful diagnostic: If you see a pattern, something is wrong with the model. You can also plot residuals against other variables, or time, or ...)

2. The t-Test

Make a t-test of the null hypothesis that $b = 0$. What do you conclude? If you were arguing with Yule, would you want to take the position that $b = 0$ and he was fooled by chance variation?

In this part, you will do a simulation to investigate the distribution of $t = \hat{b}/\text{SE}$, under the null hypothesis that $b = 0$.

(a) Set the parameters in Yule'es equation as follows:

$$a = -40,\ b = 0,\ c = 0.2,\ d = -0.3,\ \sigma = 15\,.$$

Fix the design matrix $\boldsymbol{X}$ as in Part 1. Generate 32 $N(0, \sigma^2)$ errors and plug them into the equation

$$\Delta\text{Paup}_i = -40 + 0 \cdot \Delta\text{Out}_i + 0.2 \times \Delta\text{Old}_i - 0.3 \times \Delta\text{Pop}_i + \varepsilon_i\,,$$

to get simulated values for ΔPaup_i for $i = 1, 2, \ldots, 32$.

(b) Regress the simulated ΔPaup_i on $\Delta\text{Out}, \Delta\text{Pop}$ and ΔOld. Calculate $\hat{b}, \text{SE}(\hat{b})$, and t.
(c) Repeat (b) and (c) 1000 times.
(d) Plot a histrogram for the 1000 $\hat{b}$'s, a scatter diagram for the 1000 pairs $(\hat{b}, \hat{\sigma})$ and a histogram for the 1000 t's.
(e) What is the theoretical distribution of $\hat{b}$? of $\hat{\sigma^2}$? of t? How close is the theoretical dsitribution of t to normal?
(f) Calculate the mean and SD of the 1000 $\hat{b}$'s. How does the mean compare to the true b ("True" in the simulation.) How does does the SD compare to the true SD for $\hat{b}$?

(g) Would it matter if you set the parameters differently? For instance, you could try $a = 10, b = 0, c = 0.1, d = -0.5$ and $\sigma = 25$. What if $b = 0.5$? What if $\varepsilon_i \sim \sigma \cdot (\chi^2_5 - 5)/\sqrt{10}$? The simulation in this exercise is for the level of the test. How would you do a simulation to get the power of the test?

3. Balance scale

A two balance scale reports the difference between the weights of the right and left plates, plus a random measurement error.

Suppose you have 4 objects whose weights you wish to estimate with the scale, and are allowed 12 measurements. One approach is to measure each weight alone 3 times. (How would you then estimate the four weights with this information?) Is there a better way to use the 12 allowed measurements?

Suppose that

$$x_{i,j} = \begin{cases} +1 & \text{if weight } j \text{ is included on the right plate in the } i\text{-th measurement} \\ -1 & \text{if weight } j \text{ is included on the left plate in the } i\text{-th measurement} \end{cases}$$

Then the model we are investigating is $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{Y}$ is the vector of the 12 scale readings, $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)'$ is the vector of the true weight of the four objects, and $\boldsymbol{\varepsilon}$ is the vector of 12 measurement errors. The vector equation $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ is equivalent to the 12 individual equations

$$Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \varepsilon_i, \quad i = 1, 2, \ldots, 12\,.$$

For example, if in the first measurement we put weight 1 on the right plate and weight 2 on the left, then the first reading of the scale is

$$Y_1 = \beta_1 - \beta_2 + \varepsilon_1\,.$$

Find a design matrix $\boldsymbol{X}$ that does a better job of estimating $\boldsymbol{\beta}$ than the design matrix corresponding to measuring each wieght 3 times alone. (What is the former matrix?) Discuss the choice of design matrix.