Ec142, Spring 2017

Professor Bryan Graham

Problem Set 2

Due: March 10th, 2017

Problem sets are due at 5PM in the GSIs mailbox. You may work in groups, but each student should turn in their own write-up (including a printout of a narrated/commented and executed iPython Notebook). Please also e-mail a copy of any iPython Notebook to the GSI (if applicable).

1 Quantile regression: computation/illustration

The file **brazil_pnad96_ps4.dta** contains 65,797 records drawn from the 1996 round of the Brazilian *Pesquisas Nacional por Amostra de Domicilos* (PNAD96). The population corresponds to employed males between the ages of 20 and 60. Respondents with incomplete data are dropped from the sample. Each record contains **MONTHLY_EARNINGS**, **YRSSCH** and **AgeInDays**. These variables correspond monthly earnings, years of completed schooling and age in years (but measured to the precision of a day).

For an analysis of the relationship between schooling and earnings using a closely related dataset you might read the 1991 paper "Declining inequality in schooling in Brazil and its effect on inequality in earnings," by David Lam in the *Journal of Development Economics* 37 (1-2): 199 - 225. This is available of ScienceDirect.

[a] Compute the least squares fit of Log(**MONTHLY_EARNINGS**) onto a constant **YRSSCH, AgeInDays**, and **AgeInDays** squared. Construct a 95 percent confidence interval for the coefficient on **YrsSch**. What is the expected earnings difference between an individual with no schooling (**YRSSCH** = 0) and one who has completed primary school (**YRSSCH** = 8)?

[b] Create a dummy variable for each of the 16 possible schooling levels. Compute the least squares fit of Log(**MONTHLY_EARNINGS**) onto each of the 16 dummy variables, **AgeInDays**, and **AgeInDays** squared (exclude a constant from this regression). On the basis of this regression fit discuss the appropriateness of the model fitted in (a) above. What is the expected earnings difference between an individual with no schooling (**YRSSCH** = 0) and one who has completed primary school (**YRSSCH** = 8) implied by this model?

[c] Construct two histograms. One each for the distribution of the logarithm of monthly earnings given **YRSSCH** = 0 and **YRSSCH** = 8. Comment on any differences.

[d] Consider the following $L = 7$ age ranges: $[20, 25)$, $[25, 30)$, $[30, 35)$, $[35, 40)$, $[40, 45)$, $[45, 50)$, $[55, 60$
Let $K = 16$ be the number of distinct schooling values. For each of the $K \times L = 7 \times 16 = 112$ years of schooling and age range combinations *with at least 30 observations* in the dataset estimate the 10th, 25th, 50th, 75th and 90th quantiles of the distribution of log earnings. For each conditional quantile construct a confidence interval using order statistics as described in lecture. Using this confidence interval construct a standard error estimate.

[e] Inspect your standard error estimates. Are any of them are zero. Why? Inspect the distribution of **MONTHLY_EARNINGS**. Is **MONTHLY_EARNINGS** a continuously-valued random variable? Relate what you find to the phenomena of standard error estimates of zero.

[f] Assume that, for the five estimated quantiles, the conditional quantile function of the logarithm of monthly earnings given schooling and age is a linear function of **YRSSCH**, **AgeInDays**, and **AgeInDays** squared (you may use the mid-point of each of the age ranges as your measure of "age"). Estimate the parameters indexing each of the five conditional quantile functions by minimum distance. You should *exclude* all cells with less that 30 observations and/or where the estimated standard error is zero. How does the coefficient on schooling vary with the quantile under consideration? How does it compare to that computed in part (a) above?

[g] Summarize, in words, your analysis. How do earnings vary with education in Brazil? [3 to 4 paragraphs]

[h] Repeat your analysis in part [f] for all "centiles" 5,6,7....,94,95. Plot "centile" on the x-axis and the corresponding coefficient on schooling on the y-axis. Also plot the corresponding point-wise 95 percent confidence band. Comment on your graph.