Ec142, Spring 2018

Professor Bryan Graham

Problem Set 3

Due: April 6th, 2018

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 Jupyter Notebook). Please also e-mail a copy of any such notebooks to the GSI (if applicable). Please use markdown boxes within your Jupyter notebook for narrative answers to the questions that appear below.

1 Quantile regression: computation/illustration

The file `brazil_pnad96_ps4.out` contains 65,801 comma delimited records drawn from the 1996 round of the *Brazilian Pesquisas Nacional por Amostra de Domicilos* (PNAD96). An overview of education, earnings and inequality in Brazil is provided by Blom et al. (2001).

[a] Compute the least squares fit of ln(`MONTHLY_EARNINGS`) onto a constant `YRSSCH, AgeInDays,` and `AgeInDays` squared

[b] Create a dummy variable for each of the 16 possible schooling levels. Compute the least squares fit of ln(`MONTHLY_EARNINGS`) onto each of the 16 dummy variables, `AgeInDays,` and `AgeInDays` squared (exclude a constant from this regression).

[c] Plot the regression fits in [a] and [b] on the same figure holding `AgeInDays` fixed at 40, but varying `YRSSCH.`

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

[e] Consider the following $L = 8$ age ranges: $[20, 25)$, $[25, 30)$, $[30, 35)$, $[35, 40)$, $[40, 45)$, $[45, 50)$, $[50, 55)$, $[55, 60]$. Let $K = 16$ be the number of distinct schooling values. For each of the $K \times L = 8 \times 16 = 128$ 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.

[f] 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.

[g] 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 (b) above?

[h] Summarize, in words, your analysis. How do earnings vary with education in Brazil? [5 to 8 paragraphs]

[i] 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.

References

Blom, A., Holm-Nielsen, L., & Verner, D. (2001). Education, earnings, and inequality in brazil, 1982-1998: implications for education policy. *Peabody Journal of Education*, 76(3-4), 180 – 221.