Ec240a, Fall 2023

Professor Bryan Graham

Problem Set 4

Due: December 6th, 2024

Problem sets are due at 5PM. The GSI will provide instructions on how to turn in your problem set. 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 if applicable). Please also e-mail a copy of any Jupyter Notebook to the GSI (if applicable).

Linear regression

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). 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`, `AgeInDays`, `Dad_NoSchool_c`, `Dad_1stPrim_c`, `Dad_2ndPrim_c`, `Dad_Sec_c`, `Dad_DK_c`, `Mom_NoSchool_c`, `Mom_1stPrim_c`, `Mom_2ndPrim_c`, `Mom_Sec_c`, `Mom_DK_c` and `ParentsSchooling`. The first three variables equal monthly earnings, years of completed schooling and age in years (but measured to the precision of a day). The next 5 variables are dummies for father's level of education (no school, first primary cycle completed, second primary cycle completed, secondary or more and 'don't know'). The next 5 variables are the corresponding dummies for mother's level of education. The final variable takes on 25 values corresponding to each possible combination of parent's schooling.

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 on ScienceDirect.

[a] Compute the least squares fit of ln(`MONTHLY_EARNINGS`) onto a constant `YRSSCH`, `AgeInDays`, and `AgeInDays` squared. Construct a 95 percent confidence interval for the coefficient on `YrsSch`. *Write your own Python function to complete this computation.* Your function should also construct and return a variance-covariance estimate which can be used to construct asymptotic standard errors. Compare your results – point estimates and standard errors – with those of the StatsModels OLS implementation.

[b] Compute the least squares fit of ln(`MONTHLY_EARNINGS`) onto a constant `YRSSCH`, `AgeInDays`, `AgeInDays` squared, `Dad_NoSchool_c`, `Dad_1stPrim_c`, `Dad_2ndPrim_c`, `Dad_Sec_c`, `Mom_NoSchool_c`

Mom_1stPrim_c, Mom_2ndPrim_c, and Mom_Sec_c. Compare the resulting coefficient on YRSSCH with that in part [a] above. Provide an explanation for any differences found.

[c] Show how you can compute the coefficient on YRSSCH in [b] by a least squares fit of ln(MONTHLY_EARNINGS) on a *single* variable. Describe this variable, construct it and calculate the least squares fit to check your answer.

[d] Using the Bayes' Bootstrap to approximate a posterior distribution of the coefficient on YRSSCH in the linear predictors described in parts [a] and [b]. How do these posterior distributions compare with their estimated asymptotic sampling distributions?