Ec240a, Fall 2021

Professor Bryan Graham

Problem Set 4

Due: December 3rd, 2021

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).

1 Linear regression (Example 1)

To complete this problem use the NSLY79 extract of 1,906 white male respondents placed in the problem sets folder on GitHub (nlsy79.csv). This is a comma delimited text file; the HGC_Age28 column gives the highest grade completed by age 28 for each respondent; AFQT, a respondent's (national) percentile on the Armed Forces Qualification Test; Earnings, average annual earnings over the 1997, 1999, 2001 and 2003 calendar years in 2010 prices. Define LogEarn to be the natural logarithm of Earnings.

1. Compute the least squares fit of LogEarn onto a constant and HGC_Age28. 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 Python StatsModels OLS implementation.

2. Compute the least squares fit of LogEarn on a constant, HGC_Age28 and AFQT. Use your results to construct/predict the coefficient on HGC_Age28 in a linear regression of AFQT on a constant and HGC_Age28 (show your work clearly). Numerically compute this auxiliary least squares fit to verify your answer.

3. Show how you can compute the coefficient on HGC_Age28 in (2) by a least squares fit of LogEarn on a single variable. Describe this variable, construct it and calculate the least squares fit to check your answer.

4. Estimate the parameters of the following linear regression model by the method of least squares

$$\mathbb{E}^*[\texttt{LogEarn}|\,\texttt{X}] = \alpha_0 + \beta_0\texttt{HGCAge28} + \gamma_0\texttt{HGCAge28} \times (\texttt{AFQT} - 50) + \delta_0\texttt{AFQT}$$

where $X = (\texttt{HGCAge28}, \texttt{HGCAge28} \times (\texttt{AFQT} - 50), \texttt{AFQT})'$.

(a) Provide a semi-elasticity interpretation of β_0.

(b) Provide a semi-elasticity interpretation of $\beta_0 + \gamma_0 (\texttt{AFQT} - 50)$.

(c) Interpret the null hypothesis $H_0 : \gamma_0 = 0$.

(d) Plot your estimate of $\beta_0 + \gamma_0 (\texttt{AFQT} - 50)$ as a function of AFQT (for AFQT from zero to one hundred). Use the Bayesian Bootstrap to construct and plot a 95 percent credibility band for this line.

(e) Construct an asymptotic pointwise confidence band for $\beta_0 + \gamma_0 (\texttt{AFQT} - 50)$ and plot it on the same figure.

(f) Dicuss the interpretation of your two confidence bands. Which do you prefer? Why?

Linear regression (Example 2)

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?