Ec240a, Fall 2015

Professor Bryan Graham

Problem Set 3

Due: November 18th, 2015

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 individually commented and executed code if applicable). A copy of any computer code should also be e-mailed to the GSI prior to the due date.

1 Linear regression (application)

To complete this problem use the NSLY79 extract of 1,906 white male respondents constructed in iPython Notebook # 1; HGC_Age28 equals the highest grade completed by age 28; 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 (you may find the ivreg() function defined in iPython Notebook # 3 helpful). 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.

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.

2 Linear Regression (theory)

1. Show that if $\mathbb{E}^*\left[W|\,X\right] = \mathbb{E}^*\left[W\right]$ then

$$\mathbb{E}^*\left[Y|\,X,W\right] = \mathbb{E}^*\left[Y|\,W\right] + \mathbb{E}^*\left[Y|\,X\right] - \mathbb{E}\left[Y\right].$$

2. Let

$$\mathbb{E}^*\left[Y|\,X\right] = \alpha_0 + \beta_0 X.$$

   Prove that if $\mathbb{E}\left[Y\right] = \mathbb{E}\left[X\right]$ and $\mathbb{V}\left(Y\right) = \mathbb{V}\left(X\right)$, then $|\beta_0| \leq 1$. Let Y denote son's height and X father's height. Interpret your result.

3. Derive the analysis of variance decomposition

$$\mathbb{V}\left(Y\right) = \mathbb{E}\left[\mathbb{V}\left(Y|\,X\right)\right] + \mathbb{V}\left(\mathbb{E}\left[Y|\,X\right]\right).$$

4. Let $m\left(Z\right) = \mathbb{E}\left[X|\,Z\right]$ and consider the linear regression

$$\mathbb{E}^*\left[Y|\,X, m\left(Z\right), A\right] = \alpha_0 + \beta_0 X + \gamma_0 m\left(Z\right) + A.$$

   (a) Show that

$$\mathbb{E}^*\left[m\left(Z\right)|\,X\right] = \delta_0 + \xi_0 X$$

   with

$$\begin{aligned}\delta_0 &= \left(1-\xi_0\right)\mathbb{E}\left[X\right]\\ \xi_0 &= \frac{\mathbb{V}\left(\mathbb{E}\left[X|\,Z\right]\right)}{\mathbb{E}\left[\mathbb{V}\left(X|\,Z\right)\right] + \mathbb{V}\left(\mathbb{E}\left[X|\,Z\right]\right)}.\end{aligned}$$

   (b) Assume the population under consideration is working age adults who grew up in the San Francisco Bay Area. Let Y denote a adult log income, let X denote the log income of one's parents as a child and let Z be a vector of dummy variables denoting an individual's neighborhood of residence as a child. Provide an interpretation of ξ_0 as a measure of residential stratification by income.

   (c) Establish the notation $\rho = \operatorname{corr}\left(A, X\right)$, $\mu_A = \mathbb{E}\left[A\right]$, $\mu_X = \mathbb{E}\left[X\right]$, $\sigma_A^2 = \mathbb{V}\left(A\right)$ and $\sigma_X^2 = \mathbb{V}\left(X\right)$. Show that

$$\mathbb{E}^*\left[Y|\,X\right] = \alpha_0 + \gamma_0\left(1-\xi_0\right)\mu_X + \left(\mu_A - \rho\frac{\sigma_A}{\sigma_X}\right) + \left\{\beta_0 + \gamma_0\xi_0 + \rho\frac{\sigma_A}{\sigma_X}\right\}X.$$

   (d) Your research assistant computes an estimate of $\mathbb{E}^*\left[Y|\,X\right]$ using random sample from San Francisco. She computes a separate estimate using a random sample from New York City. Assume that there is more residential stratification by income in New York than in San Francisco. How would you expect the intercept and slope coefficients to differ across the two regression fits?

3 Problems from Hansen textbook

Complete the following problems from the current version on the Hansen textbook: 2.16, 2.18, 2.19, 3.7, 3.8, 3.9, 3.10, 5.2, 5.3