Ec240a, Fall 2018

Professor Bryan Graham

Problem Set 2

Due: November 2nd, 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 if applicable). Please also e-mail a copy of any Jupyter Notebook to the GSI (if applicable).

1 Hajek Projection

Let $X_1, ..., X_K$ be a set of regressors with the property that $\mathbb{C}(X_k, X_l) = 0$ for all $k \neq l$. Show that

$$\mathbb{E}^*[Y|X_1, \ldots, X_K] = \sum_{k=1}^{K} \mathbb{E}^*[Y|X_k] - (K-1)\mathbb{E}[Y].$$

HINT: First show that

$$\mathbb{E}^*[\mathbb{E}^*[Y|X_k]|X_l] = \mathbb{E}[Y]$$

for every $k \neq l$. Second verify the orthogonality conditions

$$\mathbb{E}[UX_l] = 0$$

for $U = \left(Y - \sum_{k=1}^{K} \mathbb{E}^*[Y|X_k] + (K-1)\mathbb{E}[Y]\right)$ and $l = 1, \ldots, K$.

2 Frisch-Waugh Theorem

[a] Let Y be a scalar random variable, X a K vector of covariates (which includes a constant), and W a vector of additional covariates (which excludes a constant). Consider the long (linear) regression

$$\mathbb{E}^*[Y|W, X] = X'\beta_0 + W'\gamma_0. \tag{1}$$

Next define the short and auxiliary regressions

$$\mathbb{E}^*[Y|X] = X'b_0 \tag{2}$$

$$\mathbb{E}^*[W|X] = \Pi_0 X. \tag{3}$$

[a] Let $V = W - \mathbb{E}^*[W|X]$ be the projection error associated with the auxiliary regression. Show that

$$\begin{aligned}\mathbb{E}^*[Y|V, X] &= \mathbb{E}^*[Y|X] + \mathbb{E}^*[Y|1, V] - \mathbb{E}[Y] \\ &= \mathbb{E}^*[Y|X] + \mathbb{E}^*[Y|V]\end{aligned}$$

where $\mathbb{E}^*[Y|1, V]$ denotes the linear regression of Y onto a constant and V, while $\mathbb{E}^*[Y|V]$ denotes the corresponding regression without a constant (HINT: Observe that $\mathbb{C}(X, V) = 0$).

[b] Next show that $\mathbb{E}^*[Y|V,X] = \mathbb{E}^*[Y|W,X]$ and hence that the coefficient on V in $\mathbb{E}^*[Y|V,X]$ coincides with that on W in $\mathbb{E}^*[Y|W,X]$.

[c] Let $U = Y - \mathbb{E}^*[Y|X]$ be the projection error associated with the short regression. Derive the coefficient on V in the linear regression of U onto V (excluding a constant).

[d] Discuss the possible practical value of the results shown in [b] and [c] above.

3 Linear regression: (application #1)

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

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

[a] Show that

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

with

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

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

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

[d] Your research assistant computes an estimate of $\mathbb{E}^*[Y|X]$ 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?

[e] Read the paper "Racial segregation and the black–white test score gap" in the *Journal of Public Economics* by David Card and Jesse Rothstein. Discuss the relationship between their paper and the analysis completed in parts [a] to [c] above.

4 Linear regression (application #2)

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. Compare your point estimates with those of the statsmodels OLS implementation. In a future problem set you will further develop this function so that it returns standard error estimates.

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).