

Problem Set 4

Due: December 7th, 2016 (Wednesday)

Problem sets are due at 5PM in the GSIs mailbox (commented code and execution files should be e-mailed to the GSI prior to that time). You may work in groups, but each student should turn in their own write-up (including individually commented and executed code).

1 Average regression: identification

Let Y be a scalar outcome interest, X a $K \times 1$ vector of regressors with a constant as its first element (the other elements may be discretely- or continuously-valued) and $W \in \{w_1, \dots, w_L\}$ a discretely-valued ‘proxy variable’ with L points of support. For a random draw from the population Y is generated according to

$$Y = X'B, \tag{1}$$

where B is a $K \times 1$ vector of random coefficients. Assume that

$$\mathbb{E}[B|X, W = w] = \mathbb{E}[B|W = w] = \beta(w). \tag{2}$$

[a] Outline a concrete economic model which fits into the general set-up of (1) and (2). Assess the plausibility of condition (2) for your chosen example. One possibility is to discuss this set-up in light of the Card and Krueger (1996) schooling model, but you may choose another model if you like.

[b] Show that, for $l = 1, \dots, L$

$$\beta(w_l) = \mathbb{E}[XX'|W = w_l]^{-1} \times \mathbb{E}[XY|W = w_l].$$

You may assume all of the relevant matrices are well-defined.

[c] Consider the **average linear regression**

$$m^{\text{ar}}(x) = x'\bar{\beta}$$

for $\bar{\beta} = \mathbb{E}[\beta(W)]$. Interpret this function; outline a policy question for which knowledge of $m^{\text{ar}}(x)$ might be useful.

[d] Let D be a $L \times 1$ vector with a 1 in the l^{th} row if $W = w_l$ and zeros elsewhere. Let $R = (D \otimes X)$ and $\beta = \left(\beta(w_1)', \dots, \beta(w_L)'\right)'$. Show that

$$\beta = \mathbb{E}[RR']^{-1} \times \mathbb{E}[RY],$$

and also, for $S = (I_K \otimes D)$, that

$$\bar{\beta} = \mathbb{E}[S'\beta].$$

[e] Assume that conditional on the event $W = w_l$ the distribution of X is degenerate. What problems might

such a situation create? Comment in light of your empirical example of part **[a]** above.

[f] Let $\underline{0}$ be a $K \times 1$ vector of zeros and

$$\mathbf{R} = \begin{pmatrix} R' & \underline{0}_{1 \times K} \\ S' & -I_K \end{pmatrix} \quad \mathbf{Z} = \begin{pmatrix} R' & \underline{0}_{1 \times K} \\ \underline{0}_{K \times KL} & -I_K \end{pmatrix}$$

and $\mathbf{Y} = (Y, \underline{0}')'$. Show that

$$\begin{pmatrix} \beta \\ \bar{\beta} \end{pmatrix} = \mathbb{E}[\mathbf{Z}'\mathbf{R}]^{-1} \times \mathbb{E}[\mathbf{Z}'\mathbf{Y}].$$

2 Average linear regression: estimation and inference.

Let $\{(Y_i, X_i, W_i)\}_{i=1}^N$ be a random sample of size N draw from a population in which (1) and (2) and additional ‘regularity conditions’ hold.

[a] Show that

$$\hat{\theta} = \left[\frac{1}{N} \sum_{i=1}^N \mathbf{Z}'_i \mathbf{R}_i \right]^{-1} \times \left[\frac{1}{N} \sum_{i=1}^N \mathbf{Z}'_i \mathbf{Y}_i \right]$$

consistently estimates $\theta = (\beta', \bar{\beta}')'$. Briefly discuss any needed regularity conditions on $F_{Y,X,W}$.

[b] Let $\mathbf{U}_i = \mathbf{Y}_i - \mathbf{R}_i' \theta$ and

$$\Gamma = \mathbb{E}[\mathbf{Z}'\mathbf{R}] \quad \Omega = \mathbb{E}[\mathbf{Z}'\mathbf{U}\mathbf{U}'\mathbf{Z}],$$

show that, for $\Lambda = \Gamma^{-1}\Omega\Gamma^{-1'}$,

$$\sqrt{N}(\hat{\theta} - \theta) \xrightarrow{D} \mathcal{N}(0, \Lambda).$$

Briefly discuss any needed regularity conditions on $F_{Y,X,W}$.

Bonus (5 points of aggregate homework score): Provide an ‘elegant’ expression for the lower-right-hand $K \times K$ block of Λ .

3 Average linear 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). 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.

[a] Let $X = (1, \text{YRSSCH}, \text{AgeInDays}, \text{AgeInDays}^2)'$ and $W = \text{ParentsSchooling}$, using the results derived above compute an estimate of $\bar{\beta}$ and as well as a set of estimated standard errors. Compare your results with those of parts **[a]** and **[b]**. Provide an explanation for any differences found.

[b] Using the Bayes Bootstrap to approximate a posterior distribution for $\bar{\beta}$. How does this posterior distribution compare with the estimated asymptotic sampling distribution calculated in part **[a]**.

[c] Compare your results with those calculated in Problem Set 3.