Problem Set 6

Artschil Okropiridse*

February 23, 2024

1 Lecture 10 Multivariate and non-linear regression

Exercise 1

(Ch. 11, ex. 11.1) Show $\Omega = \mathbb{E}\left[\bar{X}_i'\Sigma\bar{X}_i\right]$ (11.10) when errors are conditionally homoskedastic (11.8).

Exercise 2

(Ch. 11, ex. 11.15) The observations are iid. $(y_{1i}, y_{2i}, X_i : \ i = 1, ..., n)$. The dependent variables y_{1i} and y_{2i} are real-valued. The regressor X_i is a k-vector. The model is the two equation system

$$y_1 = X'\beta_1 + e_1, \quad \mathbb{E}[X_i e_{1i}] = 0$$
$$y_2 = X'\beta_2 + e_2, \quad \mathbb{E}[X_i e_{2i}] = 0.$$

(a) What are the appropriate estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ for β_1 and β_2?

(b) Find the joint asymptotic distribution of $\hat{\beta}_1$ and $\hat{\beta}_2$.

(c) Describe a test for $H_0 : \beta_1 = \beta_2$.

Exercise 3

(Ch. 23, ex. 23.2) Take the model $y(\lambda) = \beta_0 + \beta_1 X + e$ with $\mathbb{E}[e|X] = 0$ where $y(\lambda)$ is the Box-Cox transformation of y. Is this a nonlinear regression model in the parameters $(\lambda, \beta_0, \beta_1)$? (Careful this is tricky.)

*many thanks to Jakob Beuschlein

Exercise 4

(Ch. 23, ex. 23.10) In Exercise 9.26, you estimated a cost function on a cross-section of electric companies. Consider the nonlinear specification

$$\log \mathrm{TC} = \beta_1 + \beta_2 \log \mathrm{Q} + \beta_3 (\log \mathrm{PL} + \log \mathrm{PK} + \log \mathrm{PF}) + \beta_4 \frac{\log \mathrm{Q}}{1 + \exp(-(\log \mathrm{Q} - \gamma))} + e.$$

This model is called a smooth threshold model. For values of $\log \mathrm{Q}$ much below γ, the variable $\log \mathrm{Q}$ has a regression slope of β_2. For values much above β_7, the regression slope is $\beta_2 + \beta_4$. The model imposes a smooth transition between these regimes.

(a) The model works best when γ is selected so that several values (in this example, at least 10 to 15) of $\log \mathrm{Q}$ are both below and above γ. Examine the data and pick an appropriate range for γ.

(b) Estimate the model by NLLS using a global numerical search over $(\beta_1, \beta_2, \beta_3, \beta_4, \gamma)$.

(c) Estimate the model by NLLS using a concentrated numerical search over γ. Do you obtain the same results?

(d) Calculate the standard errors for all the parameter estimates $(\beta_1, \beta_2, \beta_3, \beta_4, \gamma)$.

Lecture 11 Instrumental Variables

Exercise 5

(Ch 12, ex. 12.1) Consider the single equation model $y = Z\beta + e$ where y and Z are both real valued ((1×1)). Let $\hat{\beta}$ denote the IV estimator of β using as an instrument a dummy variable D (takes only values 0 and 1). Find a simple expression for the IV estimator in this context.

Exercise 6

(Ch. 12, ex. 12.3) Take the linear model $y = X'\beta + e$. Let the estimator for β be $\hat{\beta}$ with OLS residual $\hat{e}_i$. Let the IV estimator for β using some instrument Z be $\tilde{\beta}$ with IV residual $\tilde{e}_i = y_i - X_i\tilde{\beta}$. If X_i is indeed endogenous, will IV fit better than OLS in the sense that $\sum_i \tilde{e}_i^2 \leq \sum_i \hat{e}_i^2$ at least in large samples?

Exercise 7

(Ch. 12, ex. 12.10) Consider the model

$$y = X'\beta + e$$
$$X = \Gamma Z + u,\ \mathbb{E}[Ze] = 0,\ \mathbb{E}[Zu] = 0$$

with y scalar and X and Z each k vector. You have a random sample $(y_i, X_i, Z_i : \ i = 1, ..., n)$. Take the control function equation $e = u'\gamma + \nu$ with $\mathbf{e}[u\nu] = 0$ and assume for simplicity that u is observed. Inserting into the structural equation we find $y = X'\beta + u'\gamma + \nu$. The control function estimator $(\hat{\beta}, \hat{\gamma})$ is the OLS estimation of this equation.

(a) Show that $\mathbb{E}[X\nu] = 0$ (algebraically).

(b) Derive the asymptotic distribution of $(\hat{\beta}, \hat{\gamma})$.

Exercise 8

(Ch. 12, ex. 12.11) Consider the structural equation

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + e \qquad (*)$$

$X \in \mathbb{R}$ treated as exogenous so that $\mathbb{E}[\mathbf{X}e] \neq 0$. We have an instrument $Z \in \mathbb{R}$ which satisfies $\mathbb{E}[e|Z] = 0$ so in particular $\mathbb{E}[e] = 0, \mathbb{E}[Ze] = 0$ and $\mathbb{E}[Z^2 e] = 0$.

(a) Should X^2 be treated as endogenous or exogenous?

(b) Suppose we have a scalar instrument Z which satisfies

$$X = \gamma_0 + \gamma_1 Z + u \qquad (+)$$

with u independent of Z and mean zero.
Consider using $(1, Z, Z^2)$ as instruments. Is this a sufficient number of instruments? Is the model (*) just-identified, over-identified or under-identified?

(c) Write out the reduced form equation for X^2. Under what condition on the reduced form parameters in (+) are the parameters in (*) identified?

Exercise 9

(Ch. 12, ex. 12.22) You will replicate and extend the work reported in Acemoglu, Johnson, and Robinson (2001). The authors provided an expanded set of controls when they published their 2012 extension and posted the data on the AER website.This dataset is AJR2001on the textbook website.

(a) Estimate the OLS regression (12.86), the reduced form regression(12.87), and the 2SLS regression (12.88). (Which point estimate is different by 0.01 from the reported values? This is a common phenomenon in empirical replication).

(b) For the above estimates calculate both homoskedastic and heteroskedastic-robust standard errors. Which were used by the authors (as re-ported in (12.86)-(12.87)-(12.88)?)

(c) Calculate the 2SLS estimates by the Indirect Least Squares formula. Are they the same?

(d) Calculate the 2SLS estimates by the two-stage approach. Are they the same?

(e) Calculate the 2SLS estimates by the control variable approach. Are they the same?

(f) Acemoglu, Johnson, and Robinson (2001) reported many specifications including alternative regressor controls, for example *latitude* and *africa*. Estimate by least squares the equation for *log(GDP)* adding *latitude* and *africa* as regressors. Does this regression suggest that *latitude* and *africa* are predictive of the level of GDP?

(g) Now estimate the same equation as in (f) but by 2SLS using *log(mortality)* as an instrument for risk. How does the interpretation of the effect on *latitude* and *africa* change?

(h) Return to our baseline model (without including *latitude* and *africa*). the authors' reduced from equation uses *log(mortality)* as the instrument, rather than, say, the level of mortality. Estimate the reduced from for risk with mortality as the instrument. (This variable is not provided in the daatset so you need to take the exponential of *log(mortality)*.) Can you explain why the authors preferred the equation with *log(mortality)*?

(i) Try an alternative reduced form including both *log(mortality)* and the square of *log(mortality)*. Interpret the results. Re-estimate the structural equation by 2SLS using both *log(mortality)* and its square as instruments. How do the results change?

(j) Calculate and interpret a test for exogeneity of the instruments.

(k) Estimate the equation by LIML using the instruments *log(mortality)* and the square of *log(mortality)*.