Final

Applied Statistics and Econometrics II
Spring 2018, NYU
Ercan Karadas

1. **(Logit-probit)** For a sample of 600 married females, we are interested in explaining participation in market employment from exogenous characteristics in $\mathbf{x}_i$ (age, family composition, education). Let $y_i = 1$ if person i has a paid job and 0 otherwise. Suppose we estimate a linear regression model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$$

by OLS.

   (a) Give two reasons why this is not really an appropriate model.

   (b) As an alternative, we could model the participation decision by a probit model. Explain the probit model very briefly.

   (c) Give an expression for the loglikelihood function of the probit model.

   (d) How would you interpret a positive β coefficient for education in the probit model?

   (e) Suppose you have a person with $\mathbf{x}_i'\boldsymbol{\beta} = 2$. What is your prediction for her labor market status y_i? Why?

   (f) To what extent is a logit model different from a probit model?

Now assume that we have a sample of women who are not working ($y_i = 0$), part-time working ($y_i = 1$) or full-time working ($y_i = 2$).

   (g) Is it appropriate, in this case, to specify a linear model as $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$?

   (h) What alternative model could be used instead that exploits the information contained in part-time versus full-time working?

   (i) How would you interpret a positive coefficient for education in this latter model?

   (j) Would it be appropriate to pool the two outcomes $y_i = 1$ and $y_i = 2$ and estimate a binary choice model? Why or why not?

2. **(Random Utility Model)**

   (a) Explain the random utility model briefly.

   Consider two consumers $\{a, b\}$ who can take either a Car or a Bus to work. There are two attributes that the consumers care about: Time (T) and Money (M).

   (b) Suppose that the price does not change across the consumers but time does. Formulate the random utility model for this case. In particular, pay special attention to the determination of the observation vector and the parameter vector in the utility functions.

   (c) Now suppose that the researcher thinks that there is a minimum utility that each alternative brings to the consumer. Redo the previous part for this case.

   (d) Now extend the model to include the effect of a person's income (Y) on the decision whether to take bus or car to work.

3. **(Panel Data)** With a single explanatory variable, the equation used to obtain the between estimator is

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_i + a_i + \bar{u}_i$$

   where the overbar represents the average over time. Assume $E(a_i) = 0$ and suppose that $\bar{u}_i$ is uncorrelated with $\bar{x}_i$, but $\text{Cov}(x_{it}, a_i) = \sigma_{xa}$ for all t (and i because of random sampling in the cross section).

   (a) Letting $\tilde{\beta}_1$ be the between estimator, that is, the OLS estimator using the time averages, show that

$$plim(\tilde{\beta}_1) = \beta_1 + \frac{\sigma_{xa}}{Var(\bar{x}_i)}$$

   where the probability limit is defined as $N \to \infty$.

   (b) Assume further that the x_{it}, for all $t = 1, 2, \ldots, T$ are uncorrelated with constant variance σ_x^2. Show that

$$plim(\tilde{\beta}_1) = \beta_1 + T\frac{\sigma_{xa}}{\sigma_x^2}$$

   (c) If the explanatory variables are not very highly correlated across time, what does part (b) suggest about whether the inconsistency in the between estimator is smaller when there are more time periods?

4. **(GMM-2SLS)** Consider the model

$$\begin{aligned} y_i &= Z_i\delta_i + \epsilon_i \\ &= Y_i\gamma_i + X_i\beta_i + \epsilon_i \end{aligned}$$

   where $E(\epsilon_i) = 0$, $E(\epsilon_i\epsilon_i') = \sigma^2 I$, $E(X_i'\epsilon_i) = 0$, $E(Y_i'\epsilon_i) \neq 0$ and $E(Y_i'\epsilon_i)$ is unknown. Suppose that there are some instrumental variables X_{IV} such that $E(X_{IV}'\epsilon_i) = 0$ and X_{IV} has more columns than Y_i.

(a) What are the orthogonality conditions? Using these conditions, write down the GMM minimization problem.

(b) Applying the GMM method, derive the two-stage least squares estimator of δ_i.

(c) Show that the GMM test of overidentifying restrictions is equal to n times the uncentered R^2 from regressing the 2SLS residuals on X.

5. **(Time Series)** Suppose you have data on GDP (y) and aggregate consumption (c), $x_t = [\ln y_t, \ln c_t]'$. Assume that both series are $I(1)$ and cointegrated with cointegrating vector $a' = [1, -1]$. We want to estimate a VEC model

$$\Delta x_t = B_0 z_{t-1} + B_1 \Delta x_{t-1} + B_2 \Delta x_{t-2} + \epsilon_t \,,$$

where $z_t = a' x_{t-1}$ and constants are omitted for convenience.

(a) How would you estimate the VECM? What is the asymptotic distribution of the estimates?

(b) Suppose you want to identify structural shocks. What kind of restrictions seem promising for this model? Why? (There is no need to derive the restrictions fully. Just describe them qualitatively.)

(c) How would you estimate the parameters of the identified VAR? What is the asymptotic distribution of the estimates?

(d) How would you estimate impulse response functions? What is the asymptotic distribution of the IRFs?