Final - Solutions

Applied Statistics and Econometrics II
Spring 2018, NYU
Ercan Karadas

1. 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.

The linear model is used here to explain a binary dependent variable. This may result in implied probabilities outside the [0,1] interval. Further, it leads to heteroskedastic error terms.

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

In a probit model, the probability that $y_i = 1$ is modeled as a function of characteristics $\mathbf{x}_i$ (including a constant), as $P\{y_i = 1 \mid \mathbf{x}_i\} = \Phi(\mathbf{x}_i'\boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is a vector of unknown coefficients and Φ is the standard normal distribution function.

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

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{N} y_i \log \Phi(\mathbf{x}_i'\boldsymbol{\beta}) + \sum_{i=1}^{N} (1 - y_i) \log[1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta})]$$

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

Ceteris paribus, a women with a higher level of education is more likely to participate in market employment.

(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?

With $\mathbf{x}_i'\boldsymbol{\beta} = 2$, we obtain $P\{y_i = 1\} = \Phi(2) \approx 0.95$. Thus, the implied probability of this women having a paid job is a large as 95%, so it seems reasonable to predict that this person is working.

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

The logit model uses a different function for transforming $\mathbf{x}_i'\boldsymbol{\beta}$ into a probability, namely the logistic distribution function. Because the scaling of this distribution is different, the coefficients $\boldsymbol{\beta}$ will also have a different scaling. Empirically, *the logit and probit models typically yield similar results in terms of signs, statistical significance and implied probabilities, unless the tails of the distribution (probabilities less than 0.05 or more than 0.95) are very important.*

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

No. The linear model suffers from similar problems as discussed under (a).

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

An ordered probit model

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

A positive coefficient for education means that higher education, ceteris paribus, leads to a lower probability of not working and a higher probability of full-time working. The impact on the probability of part-time working is ambiguous and depends upon the other characteristics and co- efficients.

(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?

Yes. This is appropriate but it would result in less efficient estimates. This is because the underlying latent model is the same, but less information is used to estimate it.

2. (a) Explain the random utility model briefly.

In random utility model, the utility of each alternative is a linear function of observed characteristics (individual and/or alternative specific) plus an additive unobservable disturbance term and individuals are assumed to choose the alternative that has the highest utility. This is a popular model because with appropriate distributional assumptions on the disturbance terms, this approach leads to manageable expressions for the probabilities implied by the model.

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.

We need to determine the utility functions as in the lecture notes. In this case,

$$\begin{aligned} U_{iC} &= \beta_1 T_{iC} + \beta_2 M_C + \varepsilon_{iC}, \\ U_{iB} &= \beta_1 T_{iB} + \beta_2 M_B + \varepsilon_{iB}, \qquad i \in \{a, b\} \end{aligned}$$

(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.

$$\begin{aligned} U_{iC} &= \beta_1 T_{iC} + \beta_2 M_C + k + \varepsilon_{iC}, \\ U_{iB} &= \beta_1 T_{iB} + \beta_2 M_B + \varepsilon_{iB}, \qquad i \in \{a, b\} \end{aligned}$$

or

$$\begin{aligned} U_{iC} &= \beta_1 T_{iC} + \beta_2 M_C + \varepsilon_{iC}, \\ U_{iB} &= \beta_1 T_{iB} + \beta_2 M_B + k + \varepsilon_{iB}, \qquad i \in \{a, b\} \end{aligned}$$

(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.

$$\begin{aligned} U_{iC} &= \beta_1 T_{iC} + \beta_2 M_C + \beta_3 Y_i + k + \varepsilon_{iC}, \\ U_{iB} &= \beta_1 T_{iB} + \beta_2 M_B + \varepsilon_{iB}, \qquad i \in \{a, b\} \end{aligned}$$

or

$$\begin{aligned} U_{iC} &= \beta_1 T_{iC} + \beta_2 M_C + \varepsilon_{iC}, \\ U_{iB} &= \beta_1 T_{iB} + \beta_2 M_B + \beta_3 Y_i + k + \varepsilon_{iB}, \qquad i \in \{a, b\} \end{aligned}$$

or there are other possible combinations but the important thing is that Y_i should be included only in one of the utilities (see the lecture notes).

3. 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 $N \to \infty$.

The between estimator is just the OLS estimator from the cross-sectional regression of $\bar{y}_i$ on $\bar{x}_i$ (including an intercept). Because we just have a single explanatory variable $\bar{x}_i$ and the error term is $a_i + \bar{u}_i$, we have,

$$plim(\tilde{\beta}_1) = \beta_1 + \frac{Cov(\bar{x}_i, a_i + \bar{u}_i)}{Var(\bar{x}_i)}$$

But $E(a_i + \bar{u}_i) = 0$ so $Cov(\bar{x}_i, a_i + \bar{u}_i) = E(\bar{x}_i a_i)$ because $E(\bar{x}_i \bar{u}_i) = Cov(\bar{x}_i, \bar{u}_i) = 0$ by assumption. Now $E(\bar{x}_i a_i) = T^{-1} \sum_{t=1}^{T} E(x_{it} a_i) = \sigma_{xa}$. Therefore,

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

which is what we wanted to show.

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

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

If $\{x_{it}\}$ is serially correlated with constant variance σ_x^2, then $Var(\bar{x}_i) = \sigma_x^2/T$, and so

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

(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?

As part (b) shows, when the x_{it} are pairwise uncorrelated, the magnitude of the inconsistency actually increases linearly with T. The sign depends on the covariance between x_{it} and a_i.

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) The orthogonality conditions are $E(X_i'\epsilon_i) = 0$ and $E(X_{IV}'\epsilon_i) = 0$. Together these can be written as $E(X'\epsilon_i) = 0$, where $X = [X_i : X_{IV}]$. Substituting out ϵ_i and putting in the observed y_i vector gives the sample moment function $\overline{m}(\delta_i) = X'(y_i - Z_i\delta_i)$.

The GMM estimator $\hat{\delta}_i$ minimizes $\overline{m}(\delta_i)'W\overline{m}(\delta_i)$:

$$\hat{\delta}_i \in \arg\min_{\delta_i} \ \overline{m}(\delta_i)'W\overline{m}(\delta_i)$$

(b) For W, note that when there is no auto or het, then $Var(X'\epsilon) = X'Var(\epsilon)X = \sigma^2(X'X)$, so we can use $W = \sigma^{-2}(X'X)^{-1}$. Although σ^{-2} is unknown, we don't need to know it in order to do the minimization. $\hat{\delta}_i$ is the value of δ_i that minimizes

$$\begin{aligned} \overline{m}(\delta_i)'W\overline{m}(\delta_i) &= \left(X'(y_i - Z_i\delta_i)\right)'\sigma^{-2}(X'X)^{-1}X'(y_i - Z_i\delta_i) \\ &= \sigma^{-2}(y_i - Z_i\delta_i)'X(X'X)^{-1}X'(y_i - Z_i\delta_i) \end{aligned}$$

This is minimized at the 2SLS estimator

$$\hat{\delta}_i = \left[Z_i'X(X'X)^{-1}X'Z_i\right]^{-1} Z_i'X(X'X)^{-1}X'y_i$$

(c) The test statistic is the minimized value of the GMM criterion. From (b), this is

$$\sigma^{-2}\left(y_i - Z_i\hat{\delta}_i\right)'X(X'X)^{-1}X'\left(y_i - Z_i\hat{\delta}_i\right) = \sigma^{-2}e_i'X(X'X)^{-1}X'e_i$$

where $e_i = \left(y_i - Z_i\hat{\delta}_i\right)$ is the 2SLS residual. We need to estimate σ^2. A commonly used one is $\hat{\sigma}^2 = e_i'e_i/n$. Substituting this above gives

$$n\left(e_i'X(X'X)^{-1}X'e_i/e_i'e_i\right)$$

This is $nR^2 = n(ESS/TSS)$, where TSS is the total uncentered sum of squares of the elements of e_i and ESS is the explained sum of squares $\hat{e}_i'\hat{e}_i$, where $\hat{e}_i = X(X'X)^{-1}X'e_i$.

5. (a) Since the cointegrating vector is known a priori, this reduces to a seemingly unrelated regression. Furthermore, since the right-hand variables are the same in all equations, the model can be estimated by equation-by-equation OLS. The estimates are asymptotically normal, $\sqrt{T}\left(\hat{\theta}-\theta_0\right) = N(0, V_\theta)$, where θ now includes both conditional mean and variance parameters. (An explicit formula for V_θ is unnecessary, especially for the block corresponding to variance parameters.)

(b) This is open ended. Because of balanced growth, I would impose long-run neutrality a la Blanchard and Quah, but other choices are possible. Transform the reduced-form VECM into a VAR representation

$$\Psi(L)x_t = \epsilon_t \,,$$

where $\Psi(1) = B_0 a'$. Since we can not identify more shocks than variables, I assume two structural shocks η_t such that $E(\eta_t \eta_t') = I$ (a normalization) and $C\eta_t = \epsilon_t$. The structural form is

$$\Psi(L)x_t = C\eta_t$$

This implies $CC' = \Sigma$, where $\Sigma = var(\epsilon_t)$. Since there are 3 free parameters in Σ and 4 in C, we need one more restriction. Cointegration implies that a given structural shock has the same long-run effect on $\ln y_t$ and $\ln c_t$, but this restriction is enforced in specifying the VECM and is implicit in $\Psi(L)$, so it doesn't restrict C (One can also work this out from the Beveridge-Nelson representation. For an example in which cointegration coexists with long-run non-neutrality of both shocks, see Cochrane (1994)). If we impose the additional restriction that one of the shocks is neutral in the long run for $\ln y_t$ (hence for $\ln c_t$ as well, in view of cointegration), we get an extra restriction that identifies η_t.

(c) For a just-identified model, I would estimate the unrestricted model and then solve the equations $CC' = \Sigma$ plus the long-run restriction. This is an indirect least squares estimator. Since the VECM parameters are asymptotically normal and the identified-VAR parameters are nonlinear functions of the VECM parameters, one can use the delta method to find the asymptotic distribution of the identified VAR parameters.

(d) The IRFs are the coefficients in the structural moving average representation for x_t. To find that representation, I would invert the estimated structural VAR representation, $\Psi(L)x_t = C\eta_t$. This makes the IRF parameters nonlinear functions of $\Psi(L)$ and C. Hence one can use the delta method to derive asymptotic distribution of IRF parameters from that of $\Psi(L)$ and C.