Instructor: **Anna Houštecká / Yaroslav Korobka** Updated: July 3, 2025

Course: **Econometrics II**

1 Panel data models

Problem 1

Suppose that the random effects model $y_{it} = x_{it}'\beta + \eta_i + v_{it}$ is to be estimated with a panel in which the groups have different numbers of observations. Let T_i be the number of observations in group i. Show that the pooled least squares estimator is unbiased and consistent despite this complication.

Solution 1

The model is equivalent to

$$y_i = X_i\beta + v_i + \eta_i\iota, \quad y_i \in \mathbb{R}^{T_i}, \quad X_i \in \mathbb{R}^{T_i \times K}, \quad v_i \in \mathbb{R}^{T_i}, \quad \iota := (1, \ldots, 1)' \in \mathbb{R}^{T_i}, \quad i = 1, \ldots, n,$$

and given the random effects model assumption, $\mathbb{E}[\eta_i x_{it}] = 0$. The pooled OLS estimator of β is

$$\hat{\beta} = \left(\sum_{i=1}^{n} X_i' X_i \right)^{-1} \sum_{i=1}^{n} X_i' y_i$$

given that $\sum_{i=1}^{n} X_i' X_i$ is invertible. To show the bias, rewrite

$$\begin{aligned}
\hat{\beta} &= \left(\sum_{i=1}^{n} X_i' X_i \right)^{-1} \sum_{i=1}^{n} X_i' \left(X_i\beta + v_i + \eta_i\iota \right) \\
&= \beta + \left(\sum_{i=1}^{n} X_i' X_i \right)^{-1} \sum_{i=1}^{n} X_i' \left(v_i + \eta_i\iota \right) \\
&= \beta + \left(\sum_{i=1}^{n} X_i' X_i \right)^{-1} \sum_{i=1}^{n} X_i' u_i
\end{aligned}$$

with $u_i := v_i + \eta_i\iota$. Hence, the bias $\mathbb{E}[\hat{\beta} - \beta | X_i]$ is zero if $\mathbb{E}[X_i' u_i | X_i] = X_i' \mathbb{E}[u_i | X_i] = X_i' \left(\mathbb{E}[v_i | X_i] + \mathbb{E}[\eta_i | X_i] \right) = 0$. It holds because the first expectation is zero by the i.i.d. independent mean zero errors v_{it}, and the second expectation is zero by the random effects assumption and the law of iterated expectations.

To show consistency, rewrite

$$\hat{\beta} - \beta = \left(\frac{1}{n} \sum_{i=1}^{n} X_i' X_i \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} X_i' u_i \right).$$

As $n \to \infty$, using the weak law of large numbers and Slutsky's theorem, we have that

$$\hat{\beta} - \beta \xrightarrow{p} Q^{-1} \mathbb{E}[X_i' u_i],$$

where $\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} X_i' X_i = \mathbb{E}[X_i' X_i] := Q$ is a non-deficient matrix with full rank. Under the random effects assumption and arguments as above, we have that $\mathbb{E}[X_i' u_i] = 0$. Hence, the estimator is consistent.

Problem 2

Consider $y_{it} = x_{it}'\beta + \eta_i + v_{it}$, $i = 1, \ldots, N$, $t = 1, \ldots, T$, where $v_{it} \sim \mathcal{N}(0, \sigma^2)$ and $\beta = 0$. Write out the likelihood for estimating η_i and σ^2, and show that the MLE estimator $\hat{\sigma}^2$ is biased when $T < \infty$.

Solution 2

From the setup, it implies that $y_{it} \sim \mathcal{N}(\eta_i, \sigma^2)$. The individual log-likelihood for each i (across T) is then

$$\log \mathcal{L}(y_{it}|\eta_i, \sigma^2) = C_0 - \frac{T}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_{it} - \eta_i)^2$$

where C_0 is some constant independent of η_i and σ^2. The ML estimator of η_i is the solution to the equation

$$\frac{\partial \log \mathcal{L}(y_{it}|\eta_i, \sigma^2)}{\partial \eta_i} = \sum_{t=1}^{T}(y_{it} - \hat{\eta}_i) = 0,$$

which is $\hat{\eta}_i = T^{-1}\sum_{t=1}^{T} y_{it} := \bar{y}_i$.

To estimate σ^2, we use the joint log-likelihood (across i and T),

$$\log \mathcal{L}(y_{it}|\eta_i, \sigma^2) = C_1 - \frac{NT}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \eta_i)^2$$

where C_1 is some constant independent of η_i and σ^2. The ML estimator of σ^2 is the solution to the equation

$$\frac{\partial \log \mathcal{L}(y_{it}|\eta_i, \sigma^2)}{\partial \sigma^2} = -\frac{NT}{2}\cdot\frac{1}{\hat{\sigma}^2} - \frac{1}{2\hat{\sigma}^4}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \hat{\eta}_i)^2 = 0.$$

Substituting for $\hat{\eta}_i$ and rearranging, we have

$$\hat{\sigma}^2 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \bar{y}_i)^2.$$

Expectation of the estimator is

$$\begin{aligned}
\mathbb{E}[\hat{\sigma}^2] &= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(y_{it} - \bar{y}_i)^2\right] \\
&= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} y_{it}^2 - \frac{2}{T}\sum_{t=1}^{T} y_{it}\bar{y}_i + \frac{1}{T}\sum_{t=1}^{T}\bar{y}_i^2\right] \\
&= \sigma^2 - \frac{2}{T}\sum_{t=1}^{T}\mathbb{E}[y_{it}\bar{y}_i] + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\bar{y}_i^2] \\
&= \sigma^2 - \frac{2}{T}\sigma^2 + \frac{1}{T}\sigma^2 = \sigma^2 - \frac{\sigma^2}{T},
\end{aligned}$$

which is not equal to σ^2 unless $T \to \infty$.

Problem 3

Consider $y_{it} = \mathbb{1}\{x_{it}\beta + \eta_i + v_{it} \geq 0\}$, where the errors v_{it} have the logisic cdf. Consider $T = 2$, $x_{i1} = 0$ and $x_{i2} = 1$, and show that the sufficient statistic for η_i is $y_{i1} + y_{i2} = 1$, i.e. conditioning on $y_{i1} + y_{i2} = 1$ implies that the MLE does not depend on η_i.

Solution 3

The log-likelihood function for two periods is given by

$$\begin{aligned}\log \mathcal{L}(y_i|x_i,\beta,\eta_i) = y_{i1}\log\Lambda\left(x_{i1}\beta+\eta_i\right) + (1-y_{i1})\log\left(1-\Lambda(x_{i1}\beta+\eta_i)\right)\\ + y_{i2}\log\Lambda\left(x_{i2}\beta+\eta_i\right) + (1-y_{i2})\log\left(1-\Lambda(x_{i2}\beta+\eta_i)\right),\end{aligned}$$

where $\Lambda(z) = 1/(1+e^{-z})$, and given known values of the covariates,

$$\log \mathcal{L}(y_i|\beta,\eta_i) = y_{i1}\log\Lambda\left(\eta_i\right) + (1-y_{i1})\log\left(1-\Lambda(\eta_i)\right) + y_{i2}\log\Lambda\left(\beta+\eta_i\right) + (1-y_{i2})\log\left(1-\Lambda(\beta+\eta_i)\right).$$

Taking the first derivative w.r.t. η_i, and using $\Lambda(z)'/\Lambda(z) = (1-\Lambda(z))$, we have

$$\frac{\partial \log \mathcal{L}(y_i|\beta,\eta_i)}{\partial \eta_i} = y_{i1}(1-\Lambda(\hat{\eta}_i)) - (1-y_{i1})\Lambda(\hat{\eta}_i) + y_{i2}(1-\Lambda(\hat{\eta}_i+\beta)) - (1-y_{i2})\Lambda(\hat{\eta}_i+\beta) = 0,$$

which implies

$$y_{i1}+y_{i2} = \Lambda(\hat{\eta}_i) + \Lambda(\hat{\eta}_i+\beta).$$

Now, we discuss three cases:

1. if $y_{i1}+y_{i2}=0$, $\hat{\eta}_i = -\infty$,
2. if $y_{i1}+y_{i2}=2$, $\hat{\eta}_i = \infty$,
3. if $y_{i1}+y_{i2}=1$, $-2\hat{\eta}_i = \beta$, and $\hat{\eta}_i = -\beta/2$.

Hence, in the case 3., it is possible to identify η_i from the estimate of β only. It implies that conditional on $\zeta_i := y_{i1}+y_{i2}=1$, the log-likelihood is independent on η_i making ζ_i a sufficient statistic.

Problem 4

Derive the bias of the OLS estimator for α in a dynamic panel of the form $y_{it} = \alpha y_{it-1} + \eta_i + v_{it}$. Are there any conditions on α that should hold for the estimator to be well-defined?

Solution 4

First, rewrite the model in recursive form,

$$\begin{aligned}y_{it} &= \alpha\left(\alpha y_{it-2} + \eta_i + v_{it-1}\right) + \eta_i + v_{it}\\ &= \alpha^2\left(\alpha y_{it-3} + \eta_i + v_{it-2}\right) + \alpha\eta_i + \alpha v_{it-1} + \eta_i + v_{it}\\ &= \dots\\ &= \alpha^t y_0 + \left(\sum_{s=0}^{t-1}\alpha^s\right)\eta_i + \sum_{s=0}^{t-1}\alpha^s v_{it-s}.\end{aligned}$$

Problem 5

Consider the panel AR(1) model with individual effects,

$$y_{it} = \alpha y_{it-1} + \eta_i + v_{it}$$

where $\eta_i \sim$ i.i.d.$(0,\sigma^2)$ and $v_{it} \sim$ i.i.d.$(0,\sigma^2)$ are mutually independent, and for all i we have $y_{i0}=0$. Derive var $[y_{it}]$ for $t=1,\dots,T$.

Problem 6

Assume that we are in the AR(1) dynamic model setup such that

$$y_{it} = \alpha y_{it-1} + \eta_i + v_{it}$$

but now our v_{it} follows an MA(1) process such that

$$v_{it} = w_{it} + b w_{it-1},$$

where $w_{it} \sim$ i.i.d.$(0,\sigma_w^2)$ (i.e. v_{it} is serially correlated). Show that in this case the instrument y_{it-2} is not a valid instrument for estimating α with GMM in first differences, while the instruments y_{it-j} for $j \geq 3$ remain valid.

Problem 7

We have data for a panel of companies on gross investment expenditures I_{it} and net capital stock K_{it}. We model the investment rate $y_{it} = I_{it}/K_{it}$ as

$$\left(\frac{I_{it}}{K_{it}}\right) = \alpha \left(\frac{I_{it-1}}{K_{it-1}}\right) + \eta_i + v_{it},$$

and Table 1 shows the results of estimating the model in *levels* by OLS and WG, and the model in *first differences* with one instrument, two instruments, and all Arellano-Bond instruments. For the last two estimators, it also shows the Sargan test statistic and the m_2 statistic for second-order serial correlation in the residuals from the estimated model.

Table 1: Estimation results (703 firms, 4966 observations)

	OLS	WG	2SLS DIF	GMM DIF	GMM DIF
	(1)	(2)	(3)	(4)	(5)
$\hat{\alpha}$	0.2669	−0.0094	0.1626	0.1593	0.1560
	(.0185)	(.0181)	(.0362)	(.0327)	(.0318)
m_2				0.52	0.46
Sargan test				0.36	0.43
Instruments			$(I/K)_{t-2}$	$(I/K)_{t-2}$	$(I/K)_{t-2}$
				$(I/K)_{t-3}$	$(I/K)_{t-3}$
					⋮
					$(I/K)_1$

a) For each of the models in columns (2) and (3), write down the estimated equation(s).

b) Comment on the estimates of α in each of the columns. Are the results in line with theory (in terms of possible bias of the different estimators)? Why do we need to use instruments?

c) Comment on the standard errors of the last three estimators. Are the results in line with theory?

d) For the two GMM estimators (column (4) and (5)), what do you conclude from the two specification tests? What are these tests' null hypotheses and why are these useful to run?

Problem 8

Formulate a linear dynamic panel regression with a single weakly exogenous regressor, and AR(2) feedback in place of AR(1) feedback (i.e. when two most recent lags of the left side variable are present at the right side). Describe the algorithm of estimation of this model.

2 Causal inference