Analysis of Economics Data
Appendix C: Properties of OLS, IV and ML

© A. Colin Cameron
Univ. of Calif. Davis

November 2022

APPENDIX C: Properties of OLS, IV and ML

- Appendix C considers properties of OLS and related estimator.
- C.1 OLS with Independent Homoskedastic Errors
- C.2 Robust Standard errors
- C.3 Instrumental Variables Estimation
- C.4 OLS with Matrix Algebra
- C.5 Maximum Likelihood Estimation

C.1: OLS with Independent Homoskedastic Errors

- Simplify model to make algebra easier by dropping intercept

$$y_i = \beta x + u_i.$$

- Then OLS estimator is

$$b = \left(\sum_i x_i^2\right)^{-1} \sum_i x_i y_i.$$

- Also simplify by assume x_i is a fixed regressor. Then assume.

1. Model: $y_i = \beta x_i + u_i$.
2. Zero error mean: $\mathrm{E}[u_i] = 0$.
3. Constant error variance: $\mathrm{Var}[u_i] = \sigma_u^2$.
4. Uncorrelated errors: $\mathrm{Cov}[u_i, u_j] = 0, \quad i \neq j$.

A key result

- Given assumption 1 it is always the case that

$$b = \beta + (\textstyle\sum_i x_i^2)^{-1}(\textstyle\sum_i x_i u_i).$$

- To obtain this result, note that

$$\begin{aligned} b &= (\textstyle\sum_i x_i^2)^{-1}(\textstyle\sum_i x_i y_i) \\ &= (\textstyle\sum_i x_i^2)^{-1}(\textstyle\sum_i x_i(\beta x_i + u_i)) \qquad \text{assuming } y_i = \beta x_i + u_i. \\ &= (\textstyle\sum_i x_i^2)^{-1}(\textstyle\sum_i \beta x_i^2 + x_i u_i) \\ &= (\textstyle\sum_i x_i^2)^{-1}(\textstyle\sum_i \beta x_i^2) + (\textstyle\sum_i x_i^2)^{-1}(\textstyle\sum_i x_i u_i) \\ &= \beta + (\textstyle\sum_i x_i^2)^{-1}(\textstyle\sum_i x_i u_i), \end{aligned}$$

- The mean and variance of b will depend on assumptions about u_i.

Mean of the OLS Estimator

- Since $b = \beta + (\sum_i x_i^2)^{-1}(\sum_i x_i u_i)$ we have

$$\begin{aligned} \mathrm{E}[b] &= \mathrm{E}\left[\beta + (\sum_i x_i^2)^{-1}(\sum_i x_i u_i)\right] \\ &= \mathrm{E}[\beta] + \mathrm{E}\left[(\sum_i x_i^2)^{-1}(\sum_i x_i u_i)\right] \\ &= \beta + (\sum_i x_i^2)^{-1} \times \mathrm{E}[\sum_i x_i u_i] \\ &= \beta + (\sum_i x_i^2)^{-1} \times \sum_i \mathrm{E}[x_i u_i] \\ &= \beta \qquad \text{if } \mathrm{E}[x_i u_i] = 0. \end{aligned}$$

- $\mathrm{E}[x_i u_i] = 0$ given assumption 2 that $\mathrm{E}[u_i] = 0$
  - since x_i is fixed so $\mathrm{E}[x_i u_i] = x_i \mathrm{E}[u_i] = x_i \times 0 = 0$ assuming $\mathrm{E}[u_i] = 0$.

Variance of OLS with Independent Homoskedastic Errors

- Since $b = \beta + (\sum_i x_i^2)^{-1}(\sum_i x_i u_i)$
  the variance of b is simply the variance of $(\sum_i x_i^2)^{-1}(\sum_i x_i u_i)$.
- Given independent and homoskedastic errors and fixed x_i

$$\begin{aligned}
\text{Var}[b] &= \text{Var}[(\sum_i x_i^2)^{-1}(\sum_i x_i u_i)] \\
&= \{(\sum_i x_i^2)^{-1}\}^2 \times \text{Var}\left[\sum_i x_i u_i\right] \text{ as Var}[aY] = a^2\text{Var}[Y] \\
&= (\sum_i x_i^2)^{-2} \times \sum_i \text{Var}\left[x_i u_i\right] \qquad \text{by independence} \\
&= (\sum_i x_i^2)^{-2} \times \sum_i x_i^2 \text{Var}[u_i] \qquad \text{as fixed } x_i \\
&= (\sum_i x_i^2)^{-2} \times \sum_i x_i^2 \times \sigma_u^2 \text{ for homoskedastic errors.} \\
&= \sigma_u^2 (\sum_i x_i^2)^{-1} \qquad \text{simplifying.}
\end{aligned}$$

- We estimate σ_u^2 using $s_e^2 = \frac{1}{n-1}\sum_i e_i^2$ where $e_i = y_i - \widehat{y}_i$. Then

$$\text{Estimated Var}[b] = \frac{s_e^2}{\sum_i x_i^2}.$$

- With an intercept $\widehat{\text{Var}}[b] = \frac{s_e^2}{\sum_i (x_i - \bar{x})^2}$ where $s_e^2 = \frac{1}{n-2}\sum_i e_i^2$.

C.2 Robust Standard Errors Summary

- Since $b = \beta + (\sum_i x_i^2)^{-1}(\sum_i x_i u_i)$ some algebra yields

$$\begin{aligned} \mathrm{Var}[b] &= (\sum_i x_i^2)^{-2} \times \mathrm{Var}\left[\sum_i x_i u_i\right] \\ &= (\sum_i x_i^2)^{-2} \times \sum_i \mathrm{Var}\left[x_i u_i\right] \qquad \text{if errors are independent} \\ &= (\sum_i x_i^2)^{-2} \times \sum_i \sum_j x_i x_j \mathrm{E}[u_i u_j] \qquad \text{in general.} \end{aligned}$$

- This leads to robust standard error estimates where $e_i = y_i - \widehat{y}_i$.
- Heteroskedastic independent errors

$$\widehat{\mathrm{Var}}_{het}[b] = (\sum_i x_i^2)^{-2} x_i^2 e_i^2.$$

- Clustered (and heteroskedastic errors) where $\delta_{ij} = 1$ if i and j in same cluster

$$\widehat{\mathrm{Var}}_{clu}[b] = (\sum_i x_i^2)^{-2} \times \sum_i \sum_j \delta_{ij} x_i x_j e_i e_j.$$

- Autocorrelated errors (to m periods apart)

$$\begin{aligned} \widehat{\mathrm{Var}}_{HAC}[b] &= (\sum_t x_t^2)^{-2} \times \{\sum_{t=1}^T x_t^2 e_t^2 + \tfrac{2m}{m+1} \sum_{t=2}^m x_t x_{t-1} e_t e_{t-1} \\ &\quad + \cdots + \tfrac{2}{m+1} \sum_{t=m}^T x_t x_{t-m} e_t e_{t-m}\}, \end{aligned}$$

Robust Standard Errors Algebra

- Since $b = \beta + (\sum_i x_i^2)^{-1}(\sum_i x_i u_i)$

the variance of b is simply the variance of $(\sum_i x_i^2)^{-1}(\sum_i x_i u_i)$.

$$\begin{aligned} \text{Var}[b] &= \text{Var}\left[\left\{(\textstyle\sum_i x_i^2)^{-1}\sum_i x_i u_i\right\}\right] \\ &= \left\{(\textstyle\sum_i x_i^2)^{-1}\right\}^2 \times \text{Var}\left[\textstyle\sum_i x_i u_i\right] \\ &= (\textstyle\sum_i x_i^2)^{-2} \times \text{Var}\left[\textstyle\sum_i x_i u_i\right] \end{aligned}$$

- In general $\text{Var}[\sum_{i=1}^n Y_i] = \sum_i \sum_j \text{Cov}[Y_i, Y_j]$. So

$$\begin{aligned} \text{Var}\left[\textstyle\sum_i x_i u_i\right] &= \textstyle\sum_i \sum_j \text{Cov}[x_i u_i, x_j u_j] \\ &= \textstyle\sum_i \sum_j x_i x_j \text{Cov}[u_i, u_j] \\ &= \textstyle\sum_i \sum_j x_i x_j \text{E}[u_i u_j] \text{ using } \text{E}[u_i] = 0. \end{aligned}$$

- So

$$\begin{aligned} \text{Var}[b] &= \left(\sum_i x_i^2\right)^{-2} \times \text{Var}\left[\textstyle\sum_i x_i u_i\right] \\ &= \left(\sum_i x_i^2\right)^{-2} \times \textstyle\sum_i \sum_j x_i x_j \text{E}[u_i u_j] \qquad \text{in general.} \end{aligned}$$

Variance with Heteroskedastic Independent Errors

- Given independent and heteroskedastic errors and fixed x_i

$$
\begin{aligned}
\mathrm{E}[u_i u_j] &= \mathrm{E}[u_i^2] \text{ if } i = j \\
&= 0 \text{ if } i \neq j.
\end{aligned}
$$

- Then

$$
\begin{aligned}
\mathrm{Var}[b] &= (\textstyle\sum_i x_i^2)^{-2} \times \sum_i \sum_j x_i x_j \mathrm{E}[u_i u_j] \qquad \text{in general} \\
&= (\textstyle\sum_i x_i^2)^{-2} \times \sum_i x_i^2 \mathrm{E}[u_i^2].
\end{aligned}
$$

- We estimate $\sum_i x_i^2 \mathrm{E}[u_i^2]$ using $\sum_i x_i^2 e_i^2$. Then

$$
\widehat{\mathrm{Var}}_{het}[b] = \frac{\sum_i x_i^2 e_i^2}{(\sum_i x_i^2)^2}.
$$

Variance with Clustered Errors

- Define

$$\begin{aligned} \delta_{ij} &= 1 \text{ if observations } i \text{ and } j \text{ are in the same cluster} \\ \delta_{ij} &= 0 \text{ otherwise.} \end{aligned}$$

- Then with clustered errors we assume

$$\begin{aligned} \text{Cov}[u_i, u_j] &= \text{E}[u_i, u_j] \neq 0 \text{ if } \delta_{ij} = 1 \\ \text{Cov}[u_i, u_j] &= 0 \text{ if } \delta_{ij} = 0 \end{aligned}$$

- So given clustered (and heteroskedastic) errors and fixed x_i

$$\begin{aligned} \text{Var}[b] &= (\sum_i x_i^2)^{-2} \times \sum_i \sum_j x_i x_j \text{E}[u_i u_j] \qquad \text{in general} \\ &= (\sum_i x_i^2)^{-2} \times \sum_i \sum_j \delta_{ij} x_i x_j \text{E}[u_i u_j]. \end{aligned}$$

- We estimate $\sum_i \sum_j \delta_{ij} x_i x_j \text{E}[u_i u_j]$ by $\sum_i \sum_j \delta_{ij} x_i x_j e_i e_j$.

$$\widehat{\text{Var}}_{clu}[b] = \frac{\sum_i \sum_j \delta_{ij} x_i x_j e_i e_j}{(\sum_i x_i^2)^2}.$$

Variance with Autocorrelated Errors

- Use subscript t for time series (rather than subscript i).
- Assume that errors are uncorrelated after m periods

$$\begin{aligned} Cov[u_t, u_s] &\neq 0 \text{ for } |t-s| \leq m \\ &= 0 \text{ for } |t-s| > m, \end{aligned}$$

- Then

$$\begin{aligned} & \mathrm{Var}[\textstyle\sum_t x_t u_t] = \sum_t \sum_t x_t x_s \mathrm{E}[x_t u_t u_t u_s] \qquad \text{in general} \\ = \; & \textstyle\sum_t \mathrm{E}[x_t u_t] + 2\sum_t \mathrm{E}[x_t u_t x_{t-1} u_{t-1}] \\ & + \cdots + 2\textstyle\sum_t \mathrm{E}[x_t u_t, x_{t-m} u_{t-m}] \text{ as correlated up to } m \text{ periods} \end{aligned}$$

- We estimate $\mathrm{Var}[b] = (\sum_t x_t^2)^{-2} \times \mathrm{Var}[\sum_t x_t u_t]$ with

$$\begin{aligned} \widehat{\mathrm{Var}}_{HAC}[b] \; = \; & (\textstyle\sum_t x_t^2)^{-1} \times \{\sum_{t=1}^{T} x_t^2 \widehat{u}_t^2 + \frac{2m}{m+1} \sum_{t=2}^{m} x_t x_{t-1} \widehat{u}_t \widehat{u}_{t-1} \\ & + \cdots + \tfrac{2}{m+1} \textstyle\sum_{t=m}^{T} x_t x_{t-m} \widehat{u}_t \widehat{u}_{t-m}\} \times (\sum_t x_t^2)^{-1}, \end{aligned}$$

C.3 Instrumental Variables

- Consider model without intercept: $y_i = \beta x_i + u_i$.
- Suppose $\text{Cov}(x_i, u_i) \neq 0$. Then **OLS is biased and inconsistent** as

$$\begin{array}{rcll} \mathrm{E}[b] & = & \beta + (\sum_i x_i^2)^{-1} \times \sum_i \mathrm{E}[x_i u_i] & \text{from earlier OLS results} \\ & \neq & \beta \quad \text{because } \mathrm{E}[x_i u_i] \neq 0. & \end{array}$$

- Instead **assume there exists an instrument z_i that is uncorrelated with the error**. Specifically $\text{Cov}(z_i, u_i) = 0$ which implies the average $(\frac{1}{n}\sum_i z_i u_i) \to 0$ as $n \to \infty$.
- The instrumental variables estimator of β is

$$b_{IV} = (\textstyle\sum_{i=1}^n z_i x_i)^{-1} \sum_{i=1}^n z_i y_i.$$

- The instrumental variables estimator is **consistent** for β since

$$\begin{array}{ll} b_{IV} & = \beta + (\sum_i z_i x_i)^{-1} (\sum_i z_i u_i) \qquad \text{by algebra similar to OLS} \\ & = \beta + (\frac{1}{n}\sum_i z_i x_i)^{-1} (\frac{1}{n}\sum_i z_i u_i) \\ & \to \beta + (\frac{1}{n}\sum_i z_i x_i)^{-1} \times 0 \text{ as } n \to \infty \text{ as } \text{Cov}(z_i, u_i) = 0 \\ & \to \beta. \end{array}$$

C.4 OLS with Matrix Algebra

- Let $y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki} + u_i$.
- In vector notation this can be written as

$$y_i = \begin{bmatrix} 1 & x_{2i} & \cdots & x_{ki} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + u_i.$$

- Stacking all n equations for the n observations into vectors and matrices yields

$$\begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{21} & \cdots & x_{k1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{2i} & \cdots & x_{ki} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{2n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ \vdots \\ u_i \\ \vdots \\ u_n \end{bmatrix}.$$

OLS with Matrix Algebra (continued)

- The stacked model can be written as

$$\underset{(n\times 1)}{\mathbf{y}} = \underset{(n\times k)}{\mathbf{X}}\underset{(k\times 1)}{\boldsymbol{\beta}} + \underset{(n\times 1)}{\mathbf{u}},$$

  for $n \times 1$ vectors $\mathbf{y}$ and $\mathbf{u}$, $n \times k$ matrix $\mathbf{X}$, and $k \times 1$ vector $\boldsymbol{\beta}$.
- The OLS estimator that minimizes the sum of squared residuals $\mathbf{u}'\mathbf{u}$ solves the so-called normal equations $\mathbf{X}'\mathbf{u} = \mathbf{0}$ or

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\beta) = \mathbf{0}.$$

- Solving for $\boldsymbol{\beta}$ yields the the OLS estimator:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

  where $\mathbf{b}$ is a $k \times 1$ vector with entries $b_1, b_2, ..., b_k$.
- Under assumptions 1-4

$$\text{Var}[\mathbf{b}] = \sigma_u^2(\mathbf{X}'\mathbf{X})^{-1} \text{ and } \widehat{\text{Var}}[\mathbf{b}] = s_e^2(\mathbf{X}'\mathbf{X})^{-1}$$

  where $s_e^2 = \frac{1}{n-k}\sum_{i=1}^n \widehat{u}_i^2 = \frac{1}{n-k}\widehat{\mathbf{u}}'\widehat{\mathbf{u}}$ where $\widehat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\mathbf{b}$.

C.5 Maximum Likelihood Estimation

- For some types of data OLS is not appropriate.
- Then the **maximum likelihood (ML) method is often used.**
- This specifies a particular model for the conditional probability of the dependent variable given the regressors.
- Let $f(y_i|x_i, \theta)$ denote the model for the *ith* observation.
- The probability of observing the n independent observations is then

$$f(y_1, ..., y_n|x_1, ..., x_n, \theta) = f(y_1|x_1, \theta) \times \cdots \times f(y_n|x_n, \theta).$$

- The **likelihood function** reframes this probability as a function of the parameter(s) θ given the data $(y_1, x_1), ..., (x_1, x_n)$. Then

$$L(\theta) = L(\theta|(y_1, x_1), ..., (y_n, x_n)) = f(y_1|x_1, \theta) \times \cdots \times f(y_n|x_n, \theta).$$

- We estimate θ by the value that is most likely given the data; i.e. the **maximum likelihood estimator** maximizes $L(\theta)$.
- Equivalently use θ that maximizes the natural logarithm of $L(\theta)$

$$\ln L(\theta) = \ln f(y_1|x_1, \theta) + \cdots + \ln f(y_n|x_n, \theta) = \sum_{i=1}^{n} \ln f(y_i|x_i, \theta).$$

Maximum Likelihood Estimation Properties

- The **maximum likelihood estimator** (MLE) of θ, denoted $\widehat{\theta}_{\text{ML}}$, maximizes $\ln L(\theta) = \sum_{i=1}^{n} \ln f(y_i|x_i, \theta)$.
- For standard problems the MLE has very desirable properties.
- Assuming $f(y_i|x_i, \theta)$ is correctly specified the MLE is consistent, has asymptotic distribution that is normal, and has the smallest variance among consistent and asymptotically normal estimators.
- If inference is relaxed to allow for the possibility that $f(y_i|x_i, \theta)$ is incorrectly specified then the MLE is called the **quasi-MLE.** Then inference must be based on appropriate robust standard errors.
- In general the quasi-MLE is inconsistent for θ, though in the leading cases of logit, probit and Poisson regression, and the linear model with independent normally distributed errors, the quasi-MLE is still consistent for θ provided that the functional form for the conditional mean $\mathrm{E}[y_i|x_i]$ is correctly specified.

Maximum Likelihood Estimation Example

- Consider regression where y_i is a binary outcome with probability
$$f(y_i|p_i) = \begin{cases} p_i & \text{if } y_i = 1 \\ 1 - p_i & \text{if } y_i = 0 \end{cases}$$
.
- This can be rewritten as
$$f(y_i|p_i) = p_i^{y_i}(1-p_i)^{1-y_i}.$$
- The logit regression model specifies
$$p_i = \Lambda(\beta_1 + \beta_2 x_i) = \exp(\beta_1 + \beta_2 x_i)/\{1 + \exp(\beta_1 + \beta_2 x_i)\}.$$
- The log-likelihood function given independent observations is then
$$\begin{aligned} \ln L(\beta_1, \beta_2) &= \prod_{i=1}^n p_i^{y_i}(1-p_i)^{1-y_i} \\ &= \sum_{i=1}^n \{y_i \ln p_i + (1-y_i)\ln(1-p_i)\} \\ &= \sum_{i=1}^n \{y_i \ln \Lambda(\beta_1 + \beta_2 x_i) + (1-y_i)\ln(1 - \Lambda(\beta_1 + \beta_2 \end{aligned}$$
- The ML estimates of β_1 and β_2 maximize this function.