ECMA31000: Introduction to Empirical Analysis

Linear Regression; IV

Joe Hardwick

University of Chicago

Autumn 2021

Outline

- Last time:
  - OLS Estimation
  - Partitioned Regression
- Today:
  - Properties of the OLS estimator
  - Measures of Fit
  - Hypothesis testing in the linear regression model
  - Introduction to IV: Endogeneity.

Definitions

- Suppose (y, x, u) satisfy

$$y = x'\beta + u; \quad \mathrm{E}(xu) = 0.$$

- Suppose we have an iid sample $\{y_i, x_i\}_{i=1}^n$.
- Assume a unique ordinary least squares estimator exists:

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

Finite Sample properties of OLS

- Now make the additional assumption that $\mathrm{E}(u_i|x_i) = 0$.
- Since (u_i, x_i) is independent of x_j for $j \neq i$, we obtain for each $i \leq n$:

$$0 = \mathrm{E}(u_i|x_i) = \mathrm{E}(u_i|x_1, \ldots x_n).$$

- This is equivalently written as

$$\mathrm{E}(U|X) = 0.$$

- This is in turn equivalent to

$$\mathrm{E}(Y|X) = X\beta.$$

Finite Sample properties of OLS: Bias

- We have

$$\begin{aligned}\mathrm{E}\left(\hat{\beta}_n|X\right) &= \mathrm{E}\left(\left(X'X\right)^{-1}X'Y|X\right)\\ &= \left(X'X\right)^{-1}X'\mathrm{E}\left(Y|X\right)\\ &= \left(X'X\right)^{-1}X'X\beta\\ &= \beta.\end{aligned}$$

- By the Law of Iterated Expectation:

$$\mathrm{E}\left(\hat{\beta}_n\right) = \mathrm{E}\left(\mathrm{E}\left(\hat{\beta}_n|X\right)\right) = \beta,$$

so $\hat{\beta}_n$ is unbiased.

Finite Sample properties of OLS: Variance

- Suppose in addition that $Var\left(u_i|x_i\right) = \sigma^2$.
- Since the error variance does not depend on the value of x_i, it is called homoskedastic.
- If the error variance depends on x_i, u_i is heteroskedastic.
- Under homoskedasticity,

$$Var\left(U|X\right) = \sigma^2 I_n,$$

because the (i, j) entry is given by

$$Var\left(U|X\right)_{i,j} = \mathrm{E}\left(u_i u_j|X\right) = \begin{cases} \mathrm{E}\left(u_i^2|x_i\right) & i = j \\ \mathrm{E}\left(u_i u_j|x_i, x_j\right) & i \neq j \end{cases}$$

Finite Sample properties of OLS: Variance

- We have $\mathrm{E}\left(u_i^2|x_i\right) = \sigma^2$, and

$$\begin{aligned}\mathrm{E}\left(u_i u_j|x_i, x_j\right) &= \mathrm{E}\left(u_i \mathrm{E}\left(u_j|u_i, x_i, x_j\right)|x_i, x_j\right) \\ &= \mathrm{E}\left(u_i \mathrm{E}\left(u_j|x_j\right)|x_i, x_j\right) = 0.\end{aligned}$$

- Under heteroskedasticity, the off diagonal elements will still be 0 but $\mathrm{E}(u_i^2|x_i) = \sigma\left(x_i\right)^2$, so

$$Var\left(U|X\right) = \begin{pmatrix} \sigma\left(x_1\right)^2 & 0 & 0 & 0 \\ 0 & \sigma\left(x_2\right)^2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \sigma\left(x_n\right)^2 \end{pmatrix} := \Omega.$$

Finite Sample properties of OLS: Variance

- We now conclude that

$$\begin{aligned} Var\left(\hat{\beta}_n|X\right) &= Var\left(\beta + \left(X'X\right)^{-1}X'U|X\right) \\ &= \left(X'X\right)^{-1}X'\Omega X\left(X'X\right)^{-1}. \end{aligned}$$

- Under homoskedasticity, $\Omega = \sigma^2 I_n$, and so the variance becomes

$$\begin{aligned} Var\left(\hat{\beta}_n|X\right) &= \sigma^2\left(X'X\right)^{-1}X'I_nX\left(X'X\right)^{-1} \\ &= \sigma^2\left(X'X\right)^{-1}. \end{aligned}$$

Finite Sample properties of OLS: Variance

- Suppose $\mathrm{E}(u_i|x_i) = 0$ and the error term u_i is homoskedastic.
- Under these assumptions, the OLS estimator is the "best linear unbiased estimator", which means it has the "smallest" variance in the class of linear estimators that are also unbiased conditional on X. This result is known as the *Gauss-Markov Theorem*.
- Our goal is to show that

$$\sigma^2 \left(X'X\right)^{-1}$$

  is "smaller" than the conditional variance of any other estimator $\tilde{\beta} = A(X)\, Y$ which also satisfies $\mathrm{E}\left(\tilde{\beta}|X\right) = \beta$.
- Precisely: $Var\left(\tilde{\beta}|X\right) - Var\left(\hat{\beta}_n|X\right) = D$ is a positive-semidefinite matrix.

Gauss-Markov Theorem

- In particular, let r be a $(k+1) \times 1$ vector, and let $r'\beta$ denote a linear combination of the β_j.
- The Gauss-Markov theorem implies that $r'\hat{\beta}$ is the Best Linear Unbiased Estimator of $r'\beta$, since

$$\begin{aligned} Var\left(r'\tilde{\beta}|X\right) - Var\left(r'\hat{\beta}|X\right) &= r'Var\left(\tilde{\beta}|X\right)r - r'Var\left(\hat{\beta}|X\right)r \\ &= r'\left[Var\left(\tilde{\beta}|X\right) - Var\left(\hat{\beta}|X\right)\right]r \\ &= r'Dr \geq 0. \end{aligned}$$

- For example, if $r = \left(0, 0 \ldots, 0, \underbrace{1}_{(j+1)st \text{ position}}, 0, \ldots 0\right)$, then $r'\hat{\beta} = \hat{\beta}_j$, and so $Var\left(\hat{\beta}_j|X\right) \leq Var\left(\tilde{\beta}_j|X\right)$.

Gauss-Markov Theorem

- First note that a linear estimator $\tilde{\beta}$ of β satisfies

$$\tilde{\beta} = A(x_1, \ldots, x_n)\, Y = AY,$$

  for some $(k+1) \times N$ matrix $A = A(x_1, \ldots, x_n)$ depending only on the sample $\{x_i\}_{i=1}^n$.
- No matter what the true value of the unknown parameter β is, $\tilde{\beta}$ must also satisfy

$$\mathrm{E}\left(\tilde{\beta}|X\right) = \beta.$$

- This implies

$$\mathrm{E}\left(\tilde{\beta}|X\right) = \mathrm{E}\left(AY|X\right) = AX\beta = \beta.$$

Gauss-Markov Theorem

- Since the final equality must hold for any β, it must be the case that $AX = I_{k+1}$.
- Next, we compute the variance of AY conditional on X:

$$\begin{aligned} Var\left(AY|X\right) &= AVar\left(Y|X\right)A' \\ &= AVar\left(U|X\right)A' \\ &= \sigma^2 AA'. \end{aligned}$$

- If $A = \left(X'X\right)^{-1}X'$, we obtain the OLS estimator, with variance $\sigma^2\left(X'X\right)^{-1}$.

Gauss-Markov Theorem

- It remains to show that $Var\left(\hat{\beta}_{OLS}|X\right) \leq Var\left(\tilde{\beta}|X\right)$, where $\leq$ means that
$$\sigma^2 AA' - \sigma^2 \left(X'X\right)^{-1}$$
is a positive semi-definite matrix for any A such that $AX = I_{k+1}$.
- To this end, define
$$C = A - \left(X'X\right)^{-1} X'.$$

Gauss-Markov Theorem

- Note that

$$\begin{aligned} AA' - (X'X)^{-1} &= \left(C + (X'X)^{-1} X'\right)\left(C + (X'X)^{-1} X'\right)' \\ &\quad - (X'X)^{-1} \\ &= CC' + (X'X)^{-1} X'C' + CX (X'X)^{-1} \\ &= CC', \end{aligned}$$

  where the final equality holds since

$$CX = AX - (X'X)^{-1} X'X = I_{k+1} - I_{k+1} = 0.$$

- The conclusion follows because CC' is always positive semi-definite.

Questions?

Large Sample properties of OLS

- Now drop the assumptions that $\mathrm{E}(u_i|x_i) = 0$ and $Var(u_i|x_i) = \sigma^2$. Rewrite the model as

$$y_i = x_i'\beta + u_i; \qquad \mathrm{E}(u_i x_i) = 0,$$

  and suppose $\mathrm{E}(xx')$ exists and is invertible.

- The OLS estimator is consistent, because

$$\frac{1}{n}\sum_{i=1}^{n} x_i x_i' \xrightarrow{a.s.} \mathrm{E}\left(xx'\right);$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i \xrightarrow{a.s.} \mathrm{E}(xy),$$

  by the SLLN. As we know, the convergence is joint, so

$$\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} x_i y_i \xrightarrow{a.s.} \mathrm{E}\left(xx'\right)^{-1}\mathrm{E}(xy) = \beta.$$

Asymptotic Normality of OLS

- Maintain the assumption that $\mathrm{E}(xx')$ exists, and assume $Var(xu) = \mathrm{E}(u^2 xx')$ exists also. Then:

$$\sqrt{n}\left(\hat{\beta}_n - \beta\right) \xrightarrow{d} \mathcal{N}(0, \Sigma),$$

  where $\Sigma = \mathrm{E}(xx')^{-1} Var(xu) \mathrm{E}(xx')^{-1}$.
- This follows because

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

Asymptotic Normality of OLS

- Since the sequence of vectors $\left\{\left(y_i, x_i'\right)'\right\}_{i\geq 1}$ is iid, the sequence $\left\{x_i\left(y_i - x_i'\beta\right)\right\}_{i\geq 1}$ is iid. Therefore, the CLT implies:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i u_i \xrightarrow{d} \mathcal{N}\left(0, Var\left(xu\right)\right).$$

- Applying Slutsky's Theorem gives

$$\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i u_i \xrightarrow{d} \mathcal{N}\left(0, \Sigma\right),$$

  as desired.

Estimation of Σ

- Deriving asymptotic normality of $\hat{\beta}_n$ will enable us to test hypotheses about the unknown parameter β.
- Since we do not know Σ, we must construct a consistent estimator of it to yield an asymptotic distribution whose quantiles are known so we may conduct tests.
- First, suppose $\mathrm{E}(u|x) = 0$ and $Var(u|x) = \sigma^2$. Then

$$\begin{aligned} Var(xu) &= \mathrm{E}\left(u^2 xx'\right) \\ &= \mathrm{E}\left(\mathrm{E}\left(u^2|x\right) xx'\right) \\ &= \sigma^2 \mathrm{E}\left(xx'\right). \end{aligned}$$

- It follows that

$$\Sigma = \sigma^2 \mathrm{E}\left(xx'\right)^{-1}.$$

Estimation of Σ

- A natural estimator of $\mathrm{E}\left(xx'\right)^{-1}$ is $\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}$.
- It remains to find a consistent estimator of σ^2. We use

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2 = \frac{1}{n}\left\|M_X Y\right\|^2,$$

  since $M_X Y$ is the vector of residuals of a regression of y on x.
- Next, note that $M_X Y = M_X\left(X\beta + U\right) = M_X U$,so

$$\begin{aligned}\left\|M_X Y\right\|^2 &= \left\|M_X U\right\|^2 \\ &= U' M_X' M_X U \\ &= U' M_X U \\ &= U'U - U' P_X U.\end{aligned}$$

Estimation of Σ

- Finally, note that

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} u_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} u_i x_i'\right)\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i u_i\right) \\ &\xrightarrow{a.s.} \sigma^2 - 0 \cdot \mathrm{E}\left(xx'\right)^{-1} \cdot 0 \\ &= \sigma^2,\end{aligned}$$

  by the continuous mapping theorem.

- In summary, a consistent estimator of Σ is given by

$$\hat{\Sigma} = \hat{\sigma}^2 \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}.$$

Estimation of Σ

- If we do not assume $\mathrm{E}(U|X) = 0$ and $Var(U|X) = \sigma^2 I_n$, Σ does not simplify, and we use

$$\hat{\Sigma} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2 x_i x_i'\right) \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}.$$

- Proving consistency boils down to showing that

$$\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2 x_i x_i' \xrightarrow{p} \mathrm{E}\left(u^2 x x'\right).$$

- First, decompose this quantity as

$$\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2 x_i x_i' = \frac{1}{n}\sum_{i=1}^{n} u_i^2 x_i x_i' + \frac{1}{n}\sum_{i=1}^{n} \left(\hat{u}_i^2 - u_i^2\right) x_i x_i'.$$

Estimation of Σ

- We have

$$\frac{1}{n}\sum_{i=1}^{n} u_i^2 x_i x_i' \xrightarrow{a.s.} \mathrm{E}\left(u^2 x x'\right),$$

  so it remains to show that

$$\frac{1}{n}\sum_{i=1}^{n} \left(\hat{u}_i^2 - u_i^2\right) x_i x_i' = o_p(1).$$

- We prove that every element of this matrix is $o_p(1)$.

Estimation of Σ

- Fix j, k element $\frac{1}{n}\sum_{i=1}^{n}\left(\hat{u}_i^2 - u_i^2\right)x_{i,j}x_{i,k}$, and observe that

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left(\hat{u}_i^2 - u_i^2\right)x_{i,j}x_{i,k}\right| \leq \frac{1}{n}\sum_{i=1}^{n}\left|\left(\hat{u}_i^2 - u_i^2\right)\right|\left|x_{i,j}x_{i,k}\right|$$

$$\leq \max_{1\leq i\leq n}\left|\hat{u}_i^2 - u_i^2\right|\frac{1}{n}\sum_{i=1}^{n}\left|x_{i,j}x_{i,k}\right|.$$

- Since

$$\frac{1}{n}\sum_{i=1}^{n}\left|x_{i,j}x_{i,k}\right| \xrightarrow{a.s.} \mathrm{E}\left(\left|x_{i,k}x_{i,k}\right|\right),$$

this term is $O_p(1)$. Therefore, it is sufficient to prove

$$\max_{1\leq i\leq n}\left|\hat{u}_i^2 - u_i^2\right| = o_p(1).$$

Estimation of Σ

- To this end, note that

$$\hat{u}_i - u_i = x_i' \left(\beta - \hat{\beta}_n \right),$$

so since $\hat{u}_i = y_i - x_i' \hat{\beta}_n = x_i' \left(\beta - \hat{\beta}_n \right) + u_i$, we have:

$$\begin{aligned}
\left| \hat{u}_i^2 - u_i^2 \right| &= \left| x_i' \left(\beta - \hat{\beta}_n \right) (\hat{u}_i + u_i) \right| \\
&= \left| x_i' \left(\beta - \hat{\beta}_n \right) \left(x_i' \left(\beta - \hat{\beta}_n \right) + 2u_i \right) \right| \\
&= \left| \left(x_i' \left(\beta - \hat{\beta}_n \right) \right)^2 + 2u_i x_i' \left(\beta - \hat{\beta}_n \right) \right| \\
&\leq \left| \left(x_i' \left(\beta - \hat{\beta}_n \right) \right)^2 \right| + 2 \left| u_i x_i' \left(\beta - \hat{\beta}_n \right) \right|.
\end{aligned}$$

Estimation of Σ

- Next, using the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}\max_{1\leq i\leq n}\left|\hat{u}_i^2-u_i^2\right| &\leq \left\|\beta-\hat{\beta}_n\right\|^2 \max_{1\leq i\leq n}\|x_i\|^2 \\ &\quad + 2\left\|\beta-\hat{\beta}_n\right\| \max_{1\leq i\leq n}\|x_i u_i\| \\ &= \left\|\sqrt{n}\left(\beta-\hat{\beta}_n\right)\right\|^2 \frac{\max_{1\leq i\leq n}\|x_i\|^2}{n} \\ &\quad + 2\left\|\sqrt{n}\left(\beta-\hat{\beta}_n\right)\right\| \frac{\max_{1\leq i\leq n}\|x_i u_i\|}{\sqrt{n}}.\end{aligned}$$

Estimation of Σ

- Since $\sqrt{n}\left(\beta - \hat{\beta}_n\right) = O_p(1)$, we must show:

$$\frac{\max_{1\leq i\leq n}\|x_i\|^2}{n} = o_p(1);$$
$$\frac{\max_{1\leq i\leq n}\|x_i u_i\|}{\sqrt{n}} = o_p(1).$$

Lemma

Let $\{Z_i\}_{i\geq 1}$ be a sequence of identically distributed random vectors such that $\mathrm{E}(\|Z_i\|^r) < \infty$. Then

$$\frac{\max_{1\leq i\leq n}\|Z_i\|}{n^{1/r}} \xrightarrow{p} 0.$$

Estimation of Σ

Proof.

Fix $\epsilon > 0$ and note that

$$
\begin{aligned}
\mathrm{P}\left(\max_{1\le i\le n}\|Z_i\| > \epsilon n^{1/r}\right) &= \mathrm{P}\left(\cup_{i=1}^{n}\left\{\|Z_i\|^r > \epsilon^r n\right\}\right) \\
&\le \sum_{i=1}^{n}\mathrm{P}\left(\|Z_i\|^r > \epsilon^r n\right) \\
&= \sum_{i=1}^{n}\mathrm{P}\left(\|Z_i\|^r \mathbf{1}\left(\|Z_i\|^r > \epsilon^r n\right) > \epsilon^r n\right) \\
&\le \frac{1}{n\epsilon^r}\sum_{i=1}^{n}\mathrm{E}\left(\|Z_i\|^r \mathbf{1}\left(\|Z_i\|^r > \epsilon^r n\right)\right) \\
&= \frac{1}{\epsilon^r}\mathrm{E}\left(\|Z_i\|^r \mathbf{1}\left(\|Z_i\|^r > \epsilon^r n\right)\right) \\
&\to 0.
\end{aligned}
$$

□

Estimation of Σ

- The second equality holds because

$$\left\{\|Z_i\|^r > \epsilon^r n\right\} = \left\{\|Z_i\|^r \mathbf{1}\left(\|Z_i\|^r > \epsilon^r n\right) > \epsilon^r n\right\}.$$

- The second inequality follows by Markov's inequality and the third because the Z_i have identical distribution.
- Convergence to 0 holds because $\mathrm{E}\left(\|Z_i\|^r\right) < \infty$.
- Finally, since $\mathrm{E}\left(\|x\|^2\right) < \infty$ and $\mathrm{E}\left(\|ux\|^2\right) < \infty$, we conclude that

$$\max_{1 \le i \le n} \left|\hat{u}_i^2 - u_i^2\right| = o_p(1),$$

so

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{u}_i^2 - u_i^2\right) x_i x_i' = o_p(1).$$

Questions?

Coefficient of Determination

- The coefficient of determination, or R^2, is a measure of how well a linear model estimated by OLS fits a given dataset.
- R^2 is defined by

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(y_i - x_i' \hat{\beta} \right)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{\|M_X y\|^2}{\|M_c y\|^2},$$

  where M_c is the residual maker of the matrix c consisting of a single column of ones. (Regressing y on a constant yields the OLS estimate $\bar{y}$).
- R^2 measures how well the model fits the data relative to a model with only a constant and no other regressors.

Coefficient of Determination

- We may equivalently write

$$R^2 := 1 - \frac{\|M_X y\|^2}{\|M_c y\|^2} = 1 - \frac{\|M_X M_c y\|^2}{\|M_c y\|^2},$$

since $M_X y = M_X M_c y$. To see this note that

$$\begin{aligned} M_X M_c y &= M_X (I - P_c) y \\ &= M_X y - M_X P_c y \\ &= M_X y, \end{aligned}$$

where the last equality follows because $P_c y$ is just the projection of y onto the span of the first column of X (which is a column of ones), it follows that $P_c y \in s(X)$. Therefore, $M_X P_c y = 0$, since the residual of a projection of $P_c y \in s(X)$ onto $s(X)$ is 0.

Coefficient of Determination

- It follows that we can also write

$$R^2 = 1 - \frac{\|M_X M_c y\|^2}{\|M_c y\|^2} = \frac{\|P_X M_c y\|^2}{\|M_c y\|^2}$$

- The second equality follows by the pythagorean theorem since $P_X M_c y$ is orthogonal to $M_X M_c y$:

$$\begin{aligned} \|P_X M_c y\|^2 + \|M_X M_c y\|^2 &= \|P_X M_c y + M_X M_c y\|^2 \\ &= \|M_c y\|^2 . \end{aligned}$$

- These formulae show that $0 \leq R^2 \leq 1$.

Coefficient of Determination

- $\|M_c y\|^2 = \sum_{i=1}^n (y_i - \bar{y})^2 := TSS$ is called the total sum of squares.
- $\|M_X y\|^2 = \sum_{i=1}^n \left(y_i - x_i'\hat{\beta}\right)^2 := SSR$ is the sum of squared residuals.
- $\|P_X M_c y\|^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 := ESS$ is the explained sum of squares.
- $R^2 = 1$ iff $SSR = 0$, which means $\hat{U} = 0$.
- $R^2 = 0$ iff $SSR = SST$, which says $\hat{\beta} = (\bar{y}, 0, \ldots 0)$, and the model fits the data no better than a constant would.

Coefficient of Determination

- High R^2 does not mean the linear model is "correct" in the sense that it provides a causal explanation of the relationship between x and y.
- On the other hand, low R^2 does not preclude a causal relationship between x and y in a linear model.
- R^2 is an estimator of the population R^2, defined by

$$R^2_{pop} = 1 - \frac{Var(u)}{Var(y)},$$

since $\frac{SSR}{n}$ is an estimate of $Var(u)$ and $\frac{TSS}{n}$ is an estimate of $Var(y)$.

Coefficient of Determination

- Note that R^2 and R^2_{pop} always weakly increase when a regressor is added.
- Adjusted R^2 penalizes additional regressors, and is defined by

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS} \leq R^2.$$

- Adding in a regressor which does not improve in-sample fit will cause $\bar{R}^2$ to decrease.

Questions?

Finite Sample Inference

- Suppose that $Y|X \sim \mathcal{N}\left(X\beta, \sigma^2 I_n\right)$, so that

$$Y = X\beta + U; \quad \mathrm{E}\left(U|X\right) = 0, Var\left(U|X\right) = \sigma^2 I_n.$$

- These assumptions (linear conditional mean, homoskedasticity) would be satisfied if the iid observations of (y_i, x_i) are multivariate normal.
- We showed in PSET 5 that

$$\hat{\beta}_n|X \sim \mathcal{N}\left(\beta, \sigma^2 \left(X'X\right)^{-1}\right).$$

Finite Sample Inference

- Let e_j be a $(k+1) \times 1$ column vector consisting of zeroes, except in position j, where there is a 1.
- Then

$$e_j' \hat{\beta}_n | X = \hat{\beta}_{n,j} | X \sim \mathcal{N}\left(\beta_j, \left[\sigma^2 \left(X'X\right)^{-1}\right]_{j,j}\right),$$

  where the subscript j,j indicates the (j,j) entry of the associated matrix.
- We can also derive the distribution of

$$\hat{\sigma}^2 = \frac{SSR}{n-k-1}.$$

  This estimate is unbiased and consistent (see PSET 7).

Finite Sample Inference

- We will also see that $\hat{\beta}_n$ and $\hat{\sigma}^2$ are independent conditional on X, and:

$$\frac{(n-k-1)\,\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-k-1}.$$

- If σ^2 were known, we could write

$$\frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{e_j'\left(X'X\right)^{-1}e_j}} \sim \mathcal{N}\left(0,1\right),$$

though typically we must replace it with $\hat{\sigma}^2$, which yields:

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{e_j'\left(X'X\right)^{-1}e_j}} \sim t_{n-k-1}.$$

- $\hat{\sigma}\sqrt{e_j'\left(X'X\right)^{-1}e_j}$ is called the standard error of $\hat{\beta}_j$, denoted $se\left(\hat{\beta}_j\right)$.

Finite Sample Inference

- To test $H_0 : \beta_j = \beta_j^0$, we can form the t-statistic:

$$T_n = \left| \frac{\hat{\beta}_j - \beta_j^0}{se\left(\hat{\beta}_j\right)} \right|,$$

  which has a t_{n-k-1} distribution under H_0.

- Let $t_{n-k-1,1-\alpha/2}$ be the $1 - \alpha/2$ quantile of the t_{n-k-1} distribution. The test

$$\phi_n = \mathbf{1}\left(T_n > t_{n-k-1,1-\alpha/2}\right)$$

  has rejection probability equal to α when H_0 is true, and rejection probability strictly greater than α when H_0 is false.

- Therefore ϕ_n is of size α.

Finite Sample Inference

- Let F denote the CDF of t_{n-k-1}. Conditional on the data, the p-value is given by

$$\begin{aligned}\hat{p}_n &= \inf_{\alpha \in (0,1)} \left\{ T_n > t_{n-k-1,1-\alpha/2} \right\} \\ &= \inf_{\alpha \in (0,1)} \left\{ T_n > F^{-1}\left(1-\alpha/2\right) \right\}.\end{aligned}$$

  Since F^{-1} is strictly increasing and continuous, we obtain:

$$\begin{aligned}\alpha &= 2\left[1 - F\left(T_n\right)\right] \\ &= 2F\left(-T_n\right).\end{aligned}$$

Questions?

Large sample inference

- Joint normality allows us to obtain exact finite sample distributions for these statistics, but we may also appeal to asymptotic normality.
- A test is called asymptotically of size α if

$$\lim_{n\to\infty} \beta_n(\theta) \leq \alpha \quad \text{for all } \theta \in \Theta_0.$$

- Now suppose

$$\sqrt{n}\left(\hat{\beta}_n - \beta\right) \xrightarrow{d} \mathcal{N}(0, V),$$

where $V = \mathrm{E}(xx')^{-1}\mathrm{E}\left(u^2xx'\right)\mathrm{E}(xx')^{-1}$. Suppose also that V is non-singular and $\hat{V}_n \xrightarrow{p} V$ is a consistent estimator of V.

Testing a single linear restriction

- Consider testing

$$H_0 : r'\beta = c \quad vs. \quad H_1 : r'\beta \neq c,$$

where r is some specified vector in $\mathbb{R}^{k+1}$, and c is a scalar.

- By the CMT:

$$\sqrt{n}\left(r'\hat{\beta}_n - r'\beta\right) \xrightarrow{d} \mathcal{N}\left(0, r'Vr\right),$$

and so by Slutsky's theorem:

$$\frac{\sqrt{n}\left(r'\hat{\beta}_n - r'\beta\right)}{\sqrt{r'\hat{V}_n r}} \xrightarrow{d} \mathcal{N}\left(0,1\right).$$

Testing a single linear restriction

- It follows that under H_0, the test statistic

$$T_n = \frac{\sqrt{n}\left(r'\hat{\beta}_n - c\right)}{\sqrt{r'\hat{V}_n r}} \xrightarrow{d} \mathcal{N}(0,1).$$

- The test we use is $\phi_n = \mathbf{1}\left(|T_n| > z_{1-\alpha/2}\right)$. It is of asymptotic size α, because under H_0 :

$$\begin{aligned} Pr\left(\text{Reject } H_0\right) &= \mathrm{P}\left(|T_n| > z_{1-\alpha/2}\right) \\ &= \mathrm{P}\left(T_n < -z_{1-\alpha/2}\right) + \mathrm{P}\left(T_n > z_{1-\alpha/2}\right) \\ &\to \Phi\left(-z_{1-\alpha/2}\right) + 1 - \Phi\left(z_{1-\alpha/2}\right) \\ &= \frac{\alpha}{2} + 1 - \left(1 - \frac{\alpha}{2}\right) = \alpha. \end{aligned}$$

- Use $\phi_n = \mathbf{1}\left(T_n > z_{1-\alpha}\right)$ for testing $H_0 : r'\beta \leq c$ vs. $H_1 : r'\beta > c$.

Asymptotic Confidence Set for $r'\beta$

- It follows by definition of convergence in distribution that for any value of $r'\beta$:

$$\lim_{n\to\infty} \mathrm{P}_{r'\beta}\left(\left|\frac{\sqrt{n}\left(r'\hat{\beta}_n - r'\beta\right)}{\sqrt{r'\hat{V}_n r}}\right| \leq z_{1-\alpha/2}\right) = 1-\alpha.$$

- Since $z_{\alpha/2} = -z_{1-\alpha/2}$ by symmetry of the standard normal about 0, rearranging yields that

$$C_n = \left[r'\hat{\beta}_n - z_{1-\alpha/2}\sqrt{\frac{r'\hat{V}_n r}{n}}, r'\hat{\beta}_n + z_{1-\alpha/2}\sqrt{\frac{r'\hat{V}_n r}{n}}\right]$$

  is an asymptotic $1-\alpha$ confidence interval for $r'\beta$.

Testing Multiple Linear Restrictions

- Consider testing

$$H_0 : R\beta = c \quad vs. \quad R\beta \neq c,$$

where R is a $p \times (k+1)$ −dimensional matrix of full row rank and c is a $p \times 1$ vector.

- The full rank condition means none of our restrictions are redundant.
- By the CMT:

$$\sqrt{n}\left(R\hat{\beta}_n - R\beta\right) \xrightarrow{d} \mathcal{N}\left(0, RVR'\right),$$

where RVR' is full rank (because R and V are), and hence positive definite (because V is).

Testing Multiple Linear Restrictions

- To see that RVR' is positive definite, note that if $a \neq 0$, $R'a \neq 0$, so
$$\left(R'a\right)' V \left(R'a\right) > 0,$$
because V is positive definite.
- A positive definite and symmetric matrix A has a square root $A^{1/2}$ with inverse $A^{-1/2} = \left(A^{-1}\right)^{1/2}$.
- It follows by Slutsky's Theorem that
$$\left(R\hat{V}_n R'\right)^{-1/2} \sqrt{n}\left(R\hat{\beta}_n - R\beta\right) \xrightarrow{d} \mathcal{N}\left(0, I_p\right).$$

Testing Multiple Linear Restrictions

- It follows that

$$n \cdot \left(R\hat{\beta}_n - R\beta\right)' \left(R\hat{V}_n R'\right)^{-1} \left(R\hat{\beta}_n - R\beta\right) \xrightarrow{d} \chi^2_p.$$

- Under H_0,

$$T_n = n \cdot \left(R\hat{\beta}_n - c\right)' \left(R\hat{V}_n R'\right)^{-1} \left(R\hat{\beta}_n - c\right) \xrightarrow{d} \chi^2_p,$$

and so we reject iff $T_n > \chi^2_{p,1-\alpha}$.

Asymptotic Confidence Set for $R\beta$

- It follows that

$$C_n = \left\{ c \in \mathbb{R}^p : n \cdot \left(R\hat{\beta}_n - c \right)' \left(R\hat{V}_n R' \right)^{-1} \left(R\hat{\beta}_n - c \right) \leq \chi^2_{p,1-\alpha} \right\}$$

  is an asymptotic $1-\alpha$ confidence set for $R\beta$.
- This set is an ellipsoid centered at $R\hat{\beta}_n$, and satisfies

$$\mathrm{P}_{R\beta} \left(R\beta \in C_n \right) \to 1 - \alpha.$$

- Taking $R = I_{k+1}$ yields an asymptotic $1-\alpha$ confidence set for β.

Tests of Non-Linear Restrictions

- Finally, consider testing

$$H_0 : f(\beta) = 0 \quad vs. \quad H_1 : f(\beta) \neq 0,$$

  where $f : \mathbb{R}^{k+1} \to \mathbb{R}^p$ is continuously differentiable at β.
- Let $D_\beta f(\beta)$ denote the $p \times (k+1)$ dimensional matrix of partial derivatives of f evaluated at β.
- The Delta Method implies

$$\sqrt{n}\left(f\left(\hat{\beta}_n\right) - f(\beta)\right) \xrightarrow{d} \mathcal{N}\left(0, D_\beta f(\beta) V D_\beta f(\beta)'\right)$$

Tests of Non-Linear Restrictions

- The continuous mapping theorem implies that

$$D_\beta f\left(\hat{\beta}_n\right) \hat{V}_n D_\beta f\left(\hat{\beta}_n\right)' \xrightarrow{p} D_\beta f\left(\beta\right) V D_\beta f\left(\beta\right)'.$$

- Now assume $D_\beta f\left(\beta\right)$ is full row rank. We can construct a statistic with asymptotic χ^2_p distribution as before.
- Note that $f\left(\beta\right) = R\beta$ yields our previous analysis as a special case, since $D_\beta f\left(\beta\right) = R$.

Questions?

Introduction to IV

- Let (y, x, u) be a random vector such that y and u are scalar random variables and $x \in \mathbb{R}^{k+1}$.
- Assume the first component of x equals 1:

$$x = (x_0, x_1 \ldots, x_k),$$

  where $x_0 = 1$.
- Let $\beta = (\beta_0, \ldots, \beta_k) \in \mathbb{R}^{k+1}$ be a constant vector of unknown parameters such that

$$y = x'\beta + u.$$

- We *no longer* assume $\mathrm{E}(ux) = 0$, so β may not represent the best linear predictor, and therefore not the best predictor either.

Introduction to IV

- We are therefore interpreting this regression as a causal model.
- If $\mathrm{E}(ux_j) = 0$ for some j, x_j is exogenous.
- If $\mathrm{E}(ux_j) \neq 0$ for some j, x_j is endogenous.
- x_0 can always be made exogenous by shifting β_0 such that $\mathrm{E}(x_0 u) = \mathrm{E}(u) = 0$.
- Multiply the model by x and take expectations:

$$\mathrm{E}(xy) = \mathrm{E}\left(xx'\right)\beta + \mathrm{E}(xu).$$

Introduction to IV

- It follows that

$$\mathrm{E}\left(xx'\right)^{-1}\mathrm{E}\left(xy\right) = \beta + \mathrm{E}\left(xx'\right)^{-1}\mathrm{E}\left(xu\right).$$

- Therefore,

$$\hat{\beta}_n^{OLS} = \left(\frac{X'X}{n}\right)^{-1}\frac{X'Y}{n} \xrightarrow{a.s.} \beta + \mathrm{E}\left(xx'\right)^{-1}\mathrm{E}\left(xu\right) \neq \beta.$$

- The OLS estimator is now an inconsistent estimator of β under endogeneity.

Instrumental Variables

- Our goal is to use a random vector $z \in \mathbb{R}^{l+1}$ such that $\mathrm{E}(zu) = 0$ to identify β.
- The condition $\mathrm{E}(zu) = 0$ is called instrument validity. (Multivariate version of $Cov(z, u) = 0$)
- First, note that any exogenous component of x is included in z. These components of x are called included instruments.
- The constant 1 is included, since we can always set $\mathrm{E}(u) = 0$. So, letting $z_0 = 1$:

$$z = (z_0, z_1, \ldots, z_l) \in \mathbb{R}^{l+1}.$$

Instrumental Variables

- How to get β as a function of quantities we can estimate? Model
$$y = x'\beta + u.$$
- Pre-multiply by z:
$$zy = zx'\beta + zu.$$
- Take expectations:
$$\begin{aligned} \mathrm{E}(zy) &= \mathrm{E}\left(zx'\right)\beta + \mathrm{E}(zu) \\ &= \mathrm{E}\left(zx'\right)\beta. \end{aligned}$$
- If $l = k$ (exactly as many instruments as regressors), $\mathrm{E}(zx')$ is square, so
$$\beta = \left[\mathrm{E}\left(zx'\right)\right]^{-1}\mathrm{E}(zy).$$

Instrumental Variables

- The components of z are called instrumental variables.
- We further assume that $\mathrm{E}(zx')$ has rank $k+1$. (Instrument relevance/rank condition) (Multivariate version of $Cov(z,x) \neq 0$).
- Finally, we assume $\mathrm{E}(zz') < \infty$ and that there is no perfect collinearity in z.
- A necessary condition for the rank condition is $l \geq k$. This is called the order condition. In other words, we must have as many valid instruments as we have endogenous regressors.

Instrumental Variables: Order Condition

- If $l = k$, the system is exactly identified.
- If $l > k$, the system is overidentified, since we have more instruments than we need to identify β.
- Notice: If x_j is endogenous, it is not an included instrument.
- Given the order condition holds, the rank condition is necessary and sufficient to uniquely determine β.
- Later: What to do with extra instruments? Could throw them out and get an IV estimate, but this is inefficient.

IV Estimator

- We showed under validity and relevance assumptions:

$$\beta = \mathrm{E}\left(zx'\right)^{-1}\mathrm{E}\left(zy\right).$$

- The sample analog principle yields

$$\frac{1}{n}\sum_{i=1}^{n} z_i\left(y_i - x_i'\hat{\beta}_{IV}\right) = 0,$$

or

$$\begin{aligned}\hat{\beta}_{IV} &= \left(\frac{1}{n}\sum_{i=1}^{n} z_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} z_i y_i\right) \\ &\xrightarrow{p} \mathrm{E}\left(zx'\right)^{-1}\mathrm{E}\left(zy\right) = \beta.\end{aligned}$$

using the LLN and continuous mapping theorem, so the IV estimator is consistent.

Questions?