Etc2410 exam notes

Introductory Econometrics (Monash University)

ETC2410 Exam Notes

Week 1: Introduction

Data structures

Cross-sectional data	Observations on one or more variables taken at the **same point in time** i.e Inflation in 2022 for Australia, UK, US
Time series data	Observations on one or more variables taken at **different points in time** i.e Inflation in Australia for 1990, 1991, 1992 etc.
Pooled cross-sectional data	Observations on **cross-sectional** units at **different points in time** i.e Earnings in Victoria (worker 1, 2 and 3 in 2022 worker 4, 5 and 6 in 2023)
Panel data	Observations on the **same cross-section** units at **different points in time** i.e Earnings in Victoria (worker 1 in 2022, worker 1 in 2023, worker 2 in 2022, worker 2 in 2023)

The linear regression (sample) model:

$$y_i = \beta_0 + \beta_1 x_i + u_i.$$

u = errors: a random variable which captures the effect on y of factors other than x (unobserved)
B0 = y intercept (unobserved)
B1 = slope between y and x (unobserved)

- The data (x1, y1), (x2, y2), . . . ,(xn, yn) is observed.

E.g.

1) How do certain factors affect wage rate (dependent variable)?
2) Factors: education = x1, expertise = x2, training = x3.

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 training + u,$$

Why is this a linear model?

- Relations in the real world are not usually linear, this is just an assumption we make

$$\widehat{birth} = \underset{(0.374)}{5.572} - \underset{(0.013)}{0.128}age - \underset{(0.054)}{0.127}north - \underset{(0.068)}{0.217}east - \underset{(0.051)}{0.166}south$$

(Note: include a hat or include an error term with i subscript)

Figure analysis:

Intercept	When [x variables] are equal to zero, the predicted [y variable] will be equal to [number and unit]. e.g: if the median age in the state was 0 and a woman was from a western state, the average birth rate for a woman in that region would be equal to 5.572 births per woman.
Slope	Controlling for [other variables], the **predicted** [y variable], **on average** [increases/decreases] with [number and unit] for each increase in [one unit] of the [x variable]. e.g: Controlling for region, the predicted birth rate in a state on average decreases with 0.128 births per woman for each year increase in the median age in that state. * ensure beta is in the same unit as x
Error term	Controlling for [other variables], the [y variable] that is unexplained by these factors is equivalent to [number and unit].
Standard error	The standard error for each value is the standard deviation of the residuals. A smaller residual standard error means the regression model is a better fit.

Matrices

A (m x n) where m = row, n = column

(1 x n) → row vector → denoted by a’
(m x 1) → column vector → denoted by a

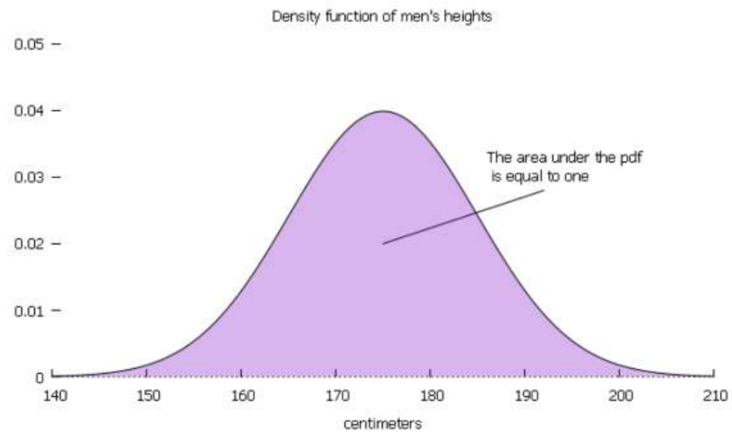
Addition and subtraction (conformable)	The two matrices must have the same number of rows and columns. Add/subtract each element in the matrix
Multiplication	The column of A must equal the rows of B I.e (2x2) and (2x3) → run and dive
Transpose	Interchanging the rows and columns → A’ - Can be done for any matrix - If a transpose is the same as the original, it is a symmetric matrix

Trace	For square matrices only → sum of the elements on the principal diagonal → denoted by tr(A)
Identity matrix	Has 1 along the principal diagonal - Pre/post multiplying by I has no effect
Diagonal matrix	A square matrix which is not a null (zero) matrix and the off-leading diagonal elements are all zero. - All identity matrices are diagonal matrices - $A = diag(a_{11}, a_{22}, ..., a_{nn}).$
Orthogonal vectors	Where two vectors multiplied equals zero
Norm	$\|\|a\|\| = (a'a)^{1/2}$ - The norm of the vector a is the square root of the scalar product of a with itself.
Determinant	$\|A\|$ = ad - bc - Square matrix only - $\|AB\| = \|A\|\|B\| = \|B\|\|A\| = \|BA\|$. $\|kA\| = k^n \|A\|$ → factor k as a constant into the matrix
Inverse	$\underset{(nxn)}{A}\underset{(nxn)}{A^{-1}} = \underset{(nxn)}{A^{-1}}\underset{(nxn)}{A} = I_n.$ $A^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$ - If A has an inverse matrix, it is nonsingular if and only if $\|A\| \neq 0$
Linearly dependent	If one of the columns can be expressed as a linear function of the remaining column $\begin{bmatrix} 4 \\ 2 \end{bmatrix} = 2\begin{bmatrix} 2 \\ 1 \end{bmatrix}.$ $A = \begin{bmatrix} 4 & 2 \\ 2 & 1 \end{bmatrix}$ column 1 = 2 column 2. - If linear it is singular

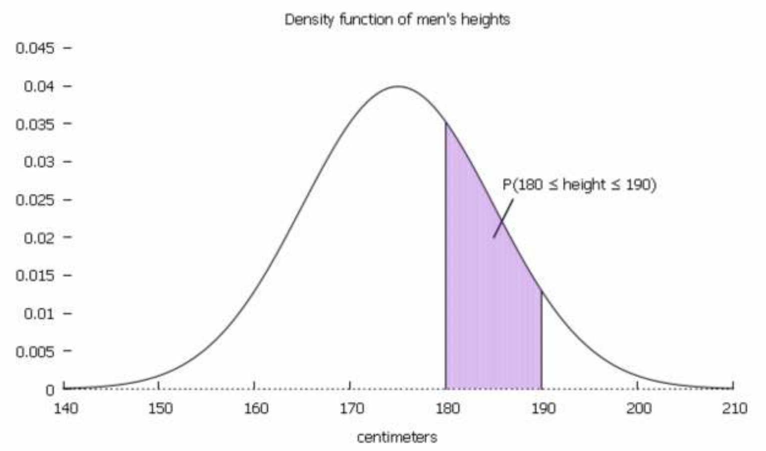
Week 2: Review of Statistical Concepts

Random variable: set of possible outcomes from a random experiment (e.g. flipping a coin)

- Can be put into a PDF (probability density function)
- P(X = x1) = p1, P(X = x2) = p2 …



The probability of that the height of a randomly selected man lies in a certain interval is the area under the pdf over that interval.



Expected value E(X):

E(X): is a weighted average of all possible values of X, with weights determined by the probability density function

$$E(X) = \mu(X) = p_1 x_1 + p_2 x_2 + \dots + p_m x_m$$

Properties of expected values:
1. For any constant c, E(c) = c.
2. For any constants a and b, E(aX + b) = aE(X) + b.
3. The expected value of the sum of several variables is the sum of their expected values: E(X + Y + Z) = E(X) + E(Y) + E(Z).
4. The above three properties imply that for any constants a, b, c and d and random variables X, Y and Z: E(a + bX + cY + dZ) = a + bE(X) + cE(Y) + dE(Z).
5. Non-linear transformations DO NOT WORK But E(XY) = E(X)E(Y) when X and Y are independent e.g. $E(X^2) \neq (E(X))^2$, $E(\log X) \neq \log(E(X))$, $E(XY) \neq E(X)E(Y)$.
6. Medium of X is such that P(X ≤ xmed) = 0.5
7. The mode of X: the most likely value of X, which is the outcome with the highest probability.

E.g. X is a discrete random variable with possible values {−3, −1, 0, 1, 5}. The expected value of X is?

A: We are not told that the outcomes are equally likely (c) We cannot compute E(X) because we do not know the probability of each outcome.

E.g. X and Y are random variables. If E(X) = 3 and E(Y) = 4 what is E(2X − Y)?

A: E(2X − Y) = 2E(X) − E(Y) = 2 × 3 − 4 = 2

Standard deviation and variance:

Variance: the expected distance from X to its mean

$$\sigma_X^2 = Var(X) = E\{(X - \mu_X)^2\}.$$

$$\sigma_X = sd(X) = \sqrt{E\{(X - \mu_X)^2\}}.$$

$$\begin{aligned} Var(X) &= E(X-\mu)^2 \\ &= E(X^2 - 2\mu X + \mu^2) \\ &= E(X^2) - 2\mu E(X) + \mu^2 \\ &= E(X^2) - 2\mu^2 + \mu^2 \\ &= E(X^2) - \mu^2. \end{aligned}$$

E.g. You have $1 to bet. You can either bet it all at once or play 20 cents for 5 rounds. Compare the risk and return for each strategy.

For the $1 bet:

E(5X) = 5E(X)
Var(5X) = 25Var(X)

For the 5x 20 cent bets:

E(X1 + X2 + X3 + X4 + X5) = 5E(X)
Var(X1 + X2 + X3 + X4 + X5) = 5Var(X)

Note: here we are not averaging returns/var

E.g. We can take a sample of one observation from the random variable and use that as the estimate of the mean, or we can take a sample of 5 observations and take the average of those 5 observations as the estimate of the mean.

1 Observation:

E(X) = mu
Var(X) = sigma squared

5 Observations:

E(X bar) = ⅕ (X1 + X2 + X3 + X4 + X5) = mu
Var(X bar) = Var (⅕(X1 + X2 + X3 + X4 + X5)) = 1/25[(Var(X1) + Var(X2)...]
= 1/25[⅕(sigma squared)] = ⅕ (sigma squared).

Note: here we ARE averaging returns/var

Covariance

Covariance: measures how two variables change together (proportional or inversely proportional)

- Covariance can be misleading if X and Y have different units (i.e hundreds vs thousands)

$$Cov(X, Y) = E((X - \mu_X)(Y - \mu_Y)) = E(XY) - \mu_X \mu_Y.$$

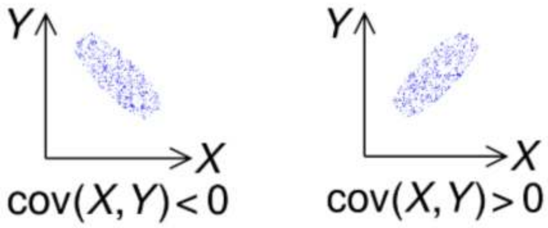
If $Cov(X, Y) > 0$, then on average
$X > E[X] \Rightarrow Y > E[Y]$ and $X < E[X] \Rightarrow Y < E[Y]$.

If $Cov(X, Y) < 0$, then on average
$X > E[X] \Rightarrow Y < E[Y]$ and $X < E[X] \Rightarrow Y > E[Y]$.

- If X and Y are independent Cov(X, Y) = 0.
- But, if X and Y are dependent Cov(X, Y) does not necessarily ≠ 0

Where a and b are constant → Cov(aX, bY) = abCov(X, Y)

Properties of variance:
8. For any constant c, Var(c) = c.
9. For any constants a and b, Var(aX + b) = a^2 Var(X)
10. For any constants a and b, Var(aX + bY) = a^2 Var(X) + 2abCov(X, Y) + b^2 Var(Y)
11. For any constants a and b, Var(aX - bY) = a^2 Var(X) - 2abCov(X, Y) + b^2 Var(Y)



E.g. X and Y are statistically independent random variables. If Var(X) = 4 and Var(Y) = 9 what is Var(2X − Y)? *A: Var(2X − Y) = 4Var(X) + Var(Y) = 4 × 4 + 9 = 25*

Correlation

Correlation: measures the strength and direction of the linear relationship between variables

$$Corr(X, Y) = \frac{Cov(X, Y)}{sd(X)sd(Y)}.$$

$-1 \leq Correlation \leq 1$

Correlation does not change if we change the units of measurement
→ no linear transformation → Corr(a1X + b1, a2Y + b2) = Corr(X, Y)

Properties of correlation
12. Corr(X, Y) = 1: there is an exact positive linear relationship between X and Y.
13. Corr(X, Y) = − 1: there is an exact negative linear relationship between X and Y.

Statistical dependence:

Two random variables that have non-zero covariance or correlation are statistically dependent, meaning that knowing the outcome of one of the two random variables gives us useful information about the other.

Joint, marginal and conditional distributions

Joint probability density function: f (x, y) = P(X = x, Y = y)

$y \downarrow, x \rightarrow$	1	2	3	marginal f_y
1	0.40	0.24	0.04	0.68
2	0.00	0.16	0.16	0.32
marginal f_x	0.40	0.40	0.20	

LIE Unconditional mean: $E(Y) = E_X[E(Y|X)].$

$$E(Y) = E(Y|X = 1)P(X = 1) + E(Y|X = 2)P(X = 2) + E(Y|X = 3)P(X = 3)$$

Weighted value

E[Y] = y1 f(y1) + y2 f(y2)

e.g. E[Y = 2] = x=1*0 + x=2*0.16 + x=3*0.16 = 0.8

Conditional distribution

Conditional probability density function: f (y|x) = f (x and y)/f (x)

E.g.

P(y = 1 | x = 1) = 0.4/0.4 = 1
P(y = 2 | x = 1) = 0/0.4 = 0

Condititional expectation function

E[Y|X] = y1 f(y1|X) + y2 f(y2|X)

e.g. E[Y|X = 1] = 1* f(y = 1| X = 1) + 2 f(y = 2 | X = 1)

= 1*1 + 0 = 1

Joint, marginal and conditional distributions

$$E(y \mid x) = \beta_0 + \beta_1 x.$$

An individual observation for Y can be written as random variations around this central value:

$y = \beta_0 + \beta_1 x + u,$ where u is a random variable with E(u | x) = 0.

Population parameters vs sample statistics

Difference between $u \; and \; \hat{u}$

"Population" = the universal reality
"Sample" = the observable universe → hat

The sample mean is an *estimator* for the population mean

$$y = \beta_0 + \beta_1 x_1 + u$$

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_1$$

$$y = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \hat{u}$$

Week 3: Linear Regression

$$E(y_i|x_i) = \beta_0 + \beta_1 x_i,$$

$$E(y_i|x_i+1) - E(y_i|x_i) = \beta_1.$$

- This makes sense because the coefficient β1 measures the average change in Y in response to a one unit change X

$$y_i = \beta_0 + \beta_1 x_i + u_i \text{ with } E(u_i|x_i) = 0.$$

$$u_i = y_i - E(y_i|x_i).$$

- The error term ui represents factors other than xi that affect yi

The OLS estimator

The OLS estimators of β0 and β1 are the values of b0 and b1 which minimise the sum of squared residuals (SSR).

The residual $\hat{u}_i = y_i - b_0 - b_1 x_i$ is the prediction error when we use $\hat{y}_i = b_0 + b_1 x_i$ as our prediction of y_i.

OLS estimators

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}.$$

$$\widehat{\beta}_1 = \frac{\widehat{Cov(x,y)}}{\widehat{Var(x)}} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2}.$$

Predicted or fitted values: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Prediction errors or residuals: $\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$.

R-squared

Total sum of squares (SST): measure of total sample variation in y
Explained sum of squares (SSE): measure of sample variation in $\hat{y}$

Residual sum of squares (SSR): measure of sample variation in $\widehat{u}$

- When you add more variables, SSR always decreases regardless of the quality of the variable

R^2: denotes the proportion of the sample variation in y that is explained by x

- R squared can only be used for comparison for the same dependent variable

SST = SSE + SSR

R^2 = SSE/SST = 1 − SSR/SST

$$\hat{\sigma} = \sqrt{\frac{SSR}{n-k-1}}$$ = Ser

SSR = Variance x (n - k - 1)

Adjusted R squared (R bar squared): Adjusted R-squared tells us how well a set of predictor variables is able to explain the variation in the response variable, *adjusted for the number of predictors in a model*.

Note: R-squared tends to reward you for including too many independent variables in a regression model, and it doesn't provide any incentive to stop adding more. Adjusted R-squared and predicted R-squared use different approaches to help you fight that impulse to add too many.

$$\text{Adjusted } R^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)}.$$

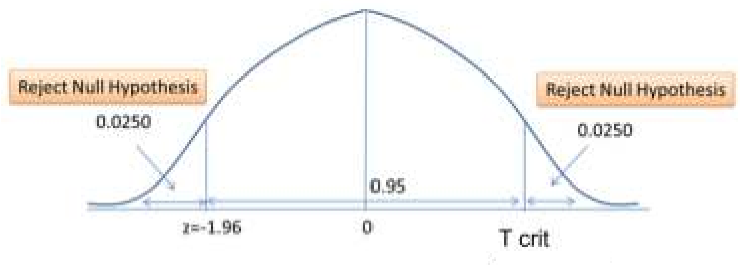
95% confidence interval (table G2 - critical values of t distribution):

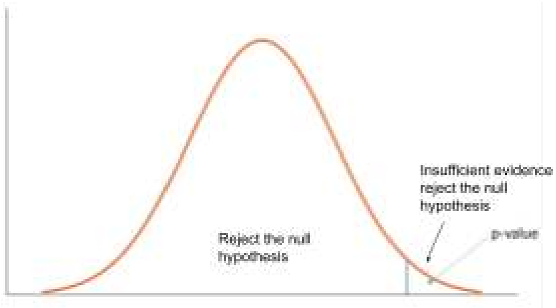
We are 95% confident that the population parameter lies between two values.

- If 0 is not included in the confidence interval → the x factor is statistically significant

$$\hat{\beta}_j \pm c_{\alpha/2} * se(\hat{\beta}_j)$$

Week 4: Inference (Hypothesis Testing)





T-test (1 co-efficient involved)

1. H0: β1 = a
2. H1: β1 ≠ a
3. Test statistic & null distribution: $\frac{\widehat{\beta_1} - a}{se(\widehat{\beta_1})} \sim t_{n - coefficients - 1}$ under H0
4. T calc =
5. T crit = $t_{n - coefficients - 1}$ =
6. Decison rule: reject H0 if |t calc| > t crit
7. We [can/cannot] reject the null hypothesis (in favour of the alternative hypothesis) at a 5% significance that ______ (holding all other factors equal).
8. [Since we reject/do not reject], [variable] [is/is not] satistically significant at the 5% level.

F-test (2+ regressors involved)

Note: an F distribution required an unrestricted (original) model and a restricted model (which assumes the null hypotheses are true)

1. H0: β2 = β3 = β4
2. H1: β2 ≠ β3 and/or β3 ≠ β4 and/or β2 ≠ β4
3. Test statistic & null distribution: $\frac{(R^2_{UR} - R^2_{R})/q}{(1 - R^2_{UR}) / (n - coefficients - 1)} \sim F_{q, n - coefficients - 1}$

   under H0

Test statistic & null hypothesis: $\frac{(SSRr - SSRur)/q}{(SSRur)/(n-k-1)}$

4. F calc =
5. F crit = $F_{n - coefficients - 1}$ =
6. Decision rule: F(calc) > F(crit).

q = number of restrictions, k = coefficients, n = observations

Degrees of freedom = n - k - 1

<u>Answer using P value:</u>

P-value < 0.05 → reject the null hypothesis (statistically significant factor)

I.e there is sufficient evidence to say that the null hypothesis is rejected

P-value > 0.05 → fail to reject the null hypothesis (statistically insignificant factor)

I.e there is insufficient evidence to say that the null hypothesis is rejected

Week 5: Dummy Variables

Null hypothesis = 0 → that the regressors are insignificant/jointly insignificant

Slope dummy variable

$$\begin{aligned} E(wage_i \mid female_i = 0, educ_i) &= \beta_0 + \beta_1 educ_i, \\ E(wage_i \mid female_i = 1, educ_i) &= (\beta_0 + \delta_0) + \beta_1 educ_i. \end{aligned}$$

→ this imposes a different intercept for dummy on and dummy off

Note: For including an attribute that has only two possibilities, we only need one dummy variable

Slope dummy variable

$$E(wage_i \mid female_i, educ_i) = \beta_0 + \delta_0 female_i + \beta_1 educ_i + \delta_1 female_i \times educ_i.$$

- This imposes different intercepts and slopes for men and women:

$$\begin{aligned} E(wage_i \mid female_i = 0, educ_i) &= \beta_0 + \beta_1 educ_i, \\ E(wage_i \mid female_i = 1, educ_i) &= (\beta_0 + \delta_0) + (\beta_1 + \delta_1) educ_i. \end{aligned}$$

- δ_1 measures the difference in slopes between women and men.

<u>If F test and P test give different results:</u>

- Consider the correlation between e.g. FEMALE and FEMALE*EDUC (i.e they move positively together).
- Makes it hard to determine if the intercept is different or the slope

Dummy for multiple categories

MUST OMIT 1 OF THE DUMMIES AS THE “BASE” CATEGORY
→ the coefficient of each dummy variable will show the difference between the intercept in that category and the benchmark.

E.g

$$\begin{aligned} \log(salary_i) = \beta_0 + \beta_1 finance_i + \beta_2 consprod_i + \beta_3 utility_i \\ + \beta_4 \log(sales_i) + \beta_5 ROE_i + u_i. \end{aligned}$$

- β0 measures the average log salary in the transport industry (i.e the base dummy)
- β1 measures the difference between the average log salary in the finance industry and the transport industry (i.e finance against the base)

- β2 − β1 measures the difference between the average log salary in the consumer product firms and finance firms.

Interpretation log-level model

$$\widehat{log(y)} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_k x_k.$$

- βb1(hat) is the chantge in log(y)(hat) as x1 increases by 1 unit, all else constant.

Predicted change in y for a dummy measured against the base
= $100(e^{Beta} - 1)\%$

Week 6: Model Selection

1. **Log-level model → % change in y = 100*B%**

$$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u.$$

β1 measures the change in log(y) as x1 increases by 1 unit.

Slope: The % change in y as x1 increases by 1 unit = $100(e^{Beta} - 1)\%$
- %Δy = 100β1% for a small β

Intercept: e to the power of B0 in the units of y

When to use:

- When the dependent variable can only take positive values
- The change in the level of the dependent variable is not really of interest.
- It is more interesting to estimate the marginal effect of the explanatory variables on the % change in the value of the dependent variable.

Note: The r squared for a log regression for y is different to the original

2. **Level-log model → unit change in y for 1% x = B/100**

$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + u.$$

β1 measures the change in log(x) as it increases by 1%

Change in y based on a 1% change in the level of x1 = $\frac{\beta 1}{100}$

3. **Log-log model → % change in y = B%**

$$\log(y) = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + u.$$

% change in the level of y in respond to a 1% change in the level of x1 = $\%\beta 1$

- Use that $100\Delta \log(x_1) \approx \%\Delta x_1$ if $\%\Delta x_1$ is small,
- and that $100\Delta \log(y) \approx \%\Delta y$ if $\%\Delta y$ is small.

Note:

- Years are not logged
- Variables already in percentages are not logged
- If a variable is positively skewed, logarithmic transformations make it less skewed

Which scale is more appropriate?

GDP and unemployment → log log
Total debt and education → log level
Wage and GDP → level log

Quadratic models

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u.$$

Derivative: The marginal effect of x on y is $\beta_1 + 2\beta_2 x$.

The coefficients of x and x squared on their own do not have meaningful interpretation.

Find when derivative = 0
Maximum when B2 < 0, minimum when B2>0
Min or max = -a/2b

- When in doubt about whether to add a quadratic term, we can add it and check its statistical significance or see if it improves the adjusted R squared.

Model selection criteria

Information criteria:

$$IC = c + \ln(SSR) + \frac{P(k)}{n}$$

We choose the model with the smallest IC and larged adjusted R squared.

1. Akaike Information Criteria (AIC)

$$AIC = c_1 + \ln(SSR) + 2k/n.$$

2. Hannan-Quinn Criterion (HQ)

$$HQ = c_2 + \ln(SSR) + 2k\ln(\ln(n))/n.$$

3. Schwarz or Bayesian Information Criterion (SIC or BIC)

$$BIC = c_3 + \ln(SSR) + k\ln(n)/n.$$

BIC: The BIC will never choose a model which has more regressors than a model chosen by the AIC, as the BIC places a larger penalty on extra regressors.

Week 7: Properties OLS

The multiple linear regression model

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \beta_0 + \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix} \beta_1 + \cdots + \begin{pmatrix} x_{1k} \\ x_{2k} \\ \vdots \\ x_{nk} \end{pmatrix} \beta_k + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

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

3 features of the OLS matrix notation:

▶ $R^2 = 1 - SSR/SST.$	
▶ $\hat{\beta} = (X'X)^{-1}X'y.$ $\hat{\beta}_1 = \frac{\widehat{Cov(x,y)}}{\widehat{Var(x)}} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2}.$	OLS estimator of β Note: the difference between the population parameter β and its OLS estimator β hat. - β is constant and does not change. - B hat is a function of sample and its value changes for different samples.
▶ $X'\hat{u} = 0.$	The vector of residuals must be orthogonal to every column of the X (i.e when multiplied must equal 0).

The OLS is 'BLUE: best linear unbiased estimator' for Beta

Unbiased

Definition: an estimator is an unbiased estimator of a parameter of interest if its expected value is the parameter of interest.

- If the assumptions below do not hold, the null distribution of the t statistics are no longer reliable
- X bar does not equal the expected value of x bar

$$E(\hat{\beta}) = E[(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}] = \beta,$$

A1	The population moodel is linear in parameters: y = Xβ + u.	

A2	Columns of X are linearly independent	
A3	conditional mean of errors is zero:	E(u|X) = 0
A4	Homoskedasiticty and no serial correlation	$Var(u\|X) = \sigma^2 I_n$
A5	Errors are normally distributed	$u\|X \sim N(0, \sigma^2 I_n)$.

The law of iterated expectations (LIE): $E(Y) = E_X[E(Y|X)]$. → LIE

Variance of the OLS estimator

- Choose the estimator with the smallest variance

► The covariance matrix of z is

$$Var(z) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix}.$$

Sigma squared = variance, sima = covariance between the two values

$$Var(\hat{\beta}) = \begin{bmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) \\ Cov(\hat{\beta}_1, \hat{\beta}_0) & Var(\hat{\beta}_1) \end{bmatrix}$$

Variance of the OLS estimator

$$\hat{\beta} \sim N(\beta, \sigma^2 (X'X)^{-1}).$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n \hat{u}_i^2}{n-k-1} = \frac{\hat{\mathrm{u}}'\hat{\mathrm{u}}}{n-k-1}$$

Efficient estimator: choose the estimator with the lowest variance

Week 8: Heteroskedasticity

The multiple regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i, \ i = 1, 2, ...n.$$

A1 model is linear in parameters: $y = X\beta + u$.

A2 columns of X are linearly independent.

A3 conditional mean of errors is zero: $E(u|X) = 0$.

A4 homoskedasticity and no serial correlation: $Var(u|X) = \sigma^2 I_n$.

A5 errors are normally distributed: $u|X \sim N(0, \sigma^2 I_n)$.

- Homoskedasticity: $Var(u_i) = \sigma^2$ for all $i = 1, 2, \ldots, n$.
- No serial correlation: $Cov(u_i, u_j) = 0$ for all $i \neq j$.

Heteroskedasticity: where the variation of y given x differs for different values of x.

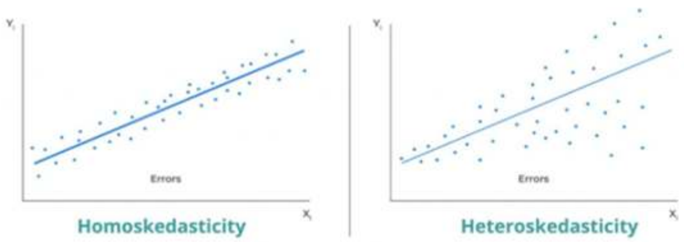
- **The OLS remains unbiased**

Consequences:

1. The t and F tests are incorrect
2. The OLS estimator of beta is no longer BLUE (but remains unbiased)

$$Var(\hat{\beta}) \neq \sigma^2 (X'X)^{-1}$$

Test 1: Informal analysis



Heteroskedasticity is an unequal scatter of the residual in comparison to measured values.

E.g. The variation in expenditure on food for the population of high income households is greater than the variation for the population of low income households.

Test 2: Bruesch-Pagan test

$$H_0 : E(u_i^2 \mid x_{i1}, x_{i2}, \ldots, x_{ik}) = \sigma^2 \text{ for } i = 1, \ldots, n.$$

1. Obtain the OLS residuals $\widehat{u}_i$ for $i = 1, \ldots, n$ from the model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i, \ i = 1, \ldots, n.$$

2. Obtain the R-squared $R^2_{\widehat{u}^2}$ from the auxiliary regression:

$$\widehat{u}_i^2 = \gamma_0 + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \cdots + \gamma_q z_{iq} + v_i, \ i = 1, \ldots, n.$$

3. Under $H_0 : \delta_1 = \cdots = \delta_q = 0$, we have the test statistic:

$$BP = nR^2_{\widehat{u}^2} \overset{asy}{\sim} \chi^2(q).$$

4. Reject H_0 in favor of $H_1 : \delta_j \neq 0$ for some j $= 1, \ldots, q$, if

$$BP_{calc} > \chi^2_{crit}(q).$$

Test 3: White test

$$H_0 : E(u_i^2 \mid x_{i1}, x_{i2}, \ldots, x_{ik}) = \sigma^2 \text{ for } i = 1, \ldots, n.$$

$$H_1 : \text{ the variance is a smooth unknown function of } x_{i1}, \ldots, x_{ik}.$$

1. Obtain the OLS residuals $\widehat{u}_i$ for $i = 1, \ldots, n$ from the model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i, \ i = 1, \ldots, n.$$

2. Obtain the R-squared $R^2_{\widehat{u}^2}$ from the auxiliary regression:

$$\begin{aligned}\widehat{u}_i^2 &= \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 x_{i3} + \alpha_1 x_{i1}^2 + \alpha_2 x_{i2}^2 + \alpha_3 x_{i3}^2 \\ &\quad + \lambda_1 x_{i1} x_{i2} + \lambda_2 x_{i1} x_{i3} + \lambda_3 x_{i2} x_{i3} + v_i, \ i = 1, \ldots, n.\end{aligned}$$

3. Under $H_0 : \gamma_1 = \gamma_2 = \gamma_3 = \alpha_1 = \alpha_2 = \alpha_3 = \lambda_1 = \lambda_2 = \lambda_3 = 0$:

$$W = nR^2_{\widehat{u}^2} \overset{asy}{\sim} \chi^2(9).$$

4. Reject H_0 in favor of H_1 : conditional heteroskedasticity, if

$$W_{calc} > \chi^2_{crit}(9).$$

Special white test:

- Omit the cross-terms from the auxiliary regression

2. Estimate the following auxiliary regression in step 2 instead:

$$\widehat{u}_i^2 = \gamma_0 + \gamma_1 \widehat{y}_i + \gamma_2 \widehat{y}_i^2 + v_i, \ i = 1, \ldots, n,$$

where $\widehat{y}_i$ is the predicted value of y_i from the model in step 1.

3. Step 3 and 4 test $H_0 : \gamma_1 = \gamma_2 = 0$ versus $H_1 : \gamma_1$ and/or $\gamma_2 \neq 0$.

How can we fix heteroskedasticity?

1. Heteroskedasticity-robust t tests

the White t test statistic for testing $H_0 : \beta_j = 0$ is

$$\frac{\widehat{\beta}_j}{se^w(\widehat{\beta}_j)} \overset{asy}{\sim} t_{(n-k-1)}.$$

2. Heteroskedasticity-robust F tests (Wald statistic)

Why not always use heteroskedastic-robust test statistics?

- OLS standard errors are only valid under homoskedasticity.
- White's standard errors are valid with homo- or heteroskedasticity.
- We are never certain whether homoskedasticity holds.

However,

- The tests based on OLS standard errors are exact tests.
- Heteroskedasticity-robust tests are asymptotic.
- In small samples, heteroskedasticity-robust tests may be misleading.

How can we fix heteroskedasicity?

1. Heteroskedasticity robust standard errors (HAC SE) + this fixes serial correlation as well (large sample size)
2. Log transformation (but can only be done on a positive value)
3. WLS (Weighted LS)*Variance (but also cannot be negative and variance must be proportional to a single explanatory variable)

Week 9: Serial Correlation (not 'BEST')

Definition: covariance of two errors ≠ 0

$$Cov(u_i, u_j|X) = 0 \text{ for all } i \neq j.$$

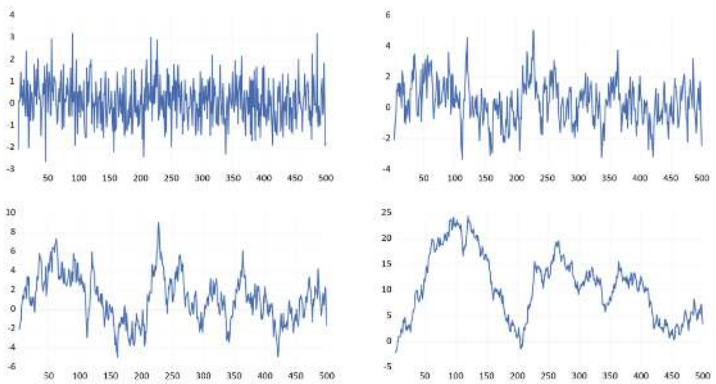
- Happens if you omit an important variable
- Likely to be the case for time series data

E.g. If x variables 3-4 weeks apart individually add predictive power to y. What problem might arise in terms of the errors?

- *This would imply that the error term is no longer white noise but exhibits some time dependence.*
- *Problem of serial correlation in the error term as this would affect the validity of OLS standard errors and by consequence inference based on this OLS regression.*

Test 1: Visual detection

Line graph of the residuals

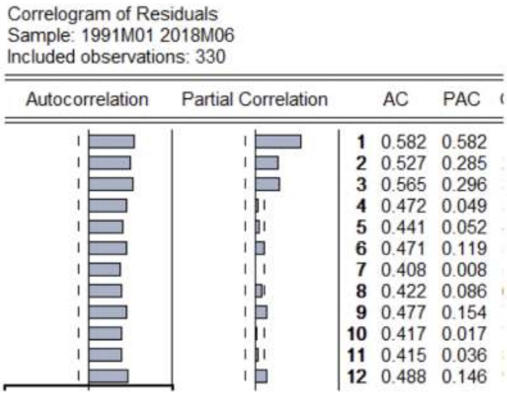


- Graphs go from least serial correlation to most serially correlated.
- First graph is totally random as the residuals randomly vary around zero
- In the last graph there is a pattern in the residuals around zero

- Column 3: Autocorrelation (AC) $\hat{\rho}_j = Corr(\hat{u}_t, \hat{u}_{t-j})$.
- Column 1: Bar charts $\hat{\rho}_j$ with 95% confidence bands.
  - If $\hat{\rho}_j$ outside the bands, reject $H_0 : \rho_j = Corr(u_t, u_{t-j}) = 0$.
- Column 4: Partial autocorrelation coefficients (PAC):
  - Coefficient estimates final lagged error terms:

$$u_t = \phi_1 u_{t-1} + e_t,$$
$$u_t = \phi_1 u_{t-1} + \phi_2 u_{t-2} + e_t,$$
$$u_t = \phi_1 u_{t-1} + \phi_2 u_{t-2} + \phi_3 u_{t-3} + e_t.$$

- Column 2: Bar charts $\hat{\phi}_j$ with 95% confidence bands.
  - If $\hat{\phi}_j$ outside the bands, reject $H_0 : \phi_j = 0$.


Correlogram of Residuals
Sample: 1991M01 2018M06
Included observations: 330

Autocorrelation	Partial Correlation		AC	PAC
		1	0.582	0.582
		2	0.527	0.285
		3	0.565	0.296
		4	0.472	0.049
		5	0.441	0.052
		6	0.471	0.119
		7	0.408	0.008
		8	0.422	0.086
		9	0.477	0.154
		10	0.417	0.017
		11	0.415	0.036
		12	0.488	0.146

Correlogram

3rd order lag: regressing ut over ut-3

- Choose the PAC level that is significant - above it is the first three so we construct a model of order 3

- All $\hat{\rho}_j s$ are outside their confidence bands, so reject

$$H_0 : \rho_j = 0, \ j = 1, 2, ..., 12.$$

- This suggests serially correlated errors in the linear regression

$$\log(Vic_t) = \beta_0 + \beta_1 time_t + \sum_{i=1}^{11} \lambda_i Q_{ti} + u_t.$$

- The first three $\hat{\phi}_j s$ are outside their confidence bands, so reject

$$H_0 : \phi_j = 0, \ j = 1, 2, 3.$$

- This suggests an AR(3) process of the form

$$u_t = \phi_0 + \phi_1 u_{t-1} + \phi_2 u_{t-2} + \phi_3 u_{t-3} + e_t.$$

Test 2: Bruesch-Godfrey

1. Obtain the OLS residuals $\hat{u}_t$ for $t = 1, \ldots, n$ from the model:

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + u_t, \ t = 1, \ldots, n.$$

2. Obtain the R-squared $R^2_{\hat{u}}$ from the auxiliary regression:

$$\hat{u}_t = \alpha_1 + \alpha_2 x_{t2} + \ldots . + \alpha_k x_{tk} + \phi_1 \hat{u}_{t-1} + \ldots + \phi_q \hat{u}_{t-q} + e_t.$$

3. Under $H_0 : \phi_1 = \phi_2 = \ldots = \phi_q = 0$, we have the test statistic:

$$BG = (n - q) R^2_{\hat{u}} \overset{asy}{\sim} \chi^2(q).$$

4. Reject H_0 in favor of $H_1 : \phi_j \neq 0$ for at least one $j = 1, 2, ..., q$, if

$$BG_{calc} > \chi^2_{crit}(q).$$

A formal test for no serial correlation in errors against an AR(8) alternative starts with specifying an AR equation for errors:

$$dlgdp_t = \beta_0 + \beta_1 dldgp_{t-1} + \beta_2 dlgdp_{t-2} + u_t$$

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \cdots + \rho_8 u_{t-8} + e_t$$

$$H_0 \; : \; \rho_j = 0 \text{ for } j = 1, 2, \ldots, 8$$

$$H_1 \; : \; \text{at least one } \rho \text{ is not zero}$$

$$BG = (n-8)R_{\hat{u}}^2 \sim \chi_8^2 \;\; \text{under } H_0$$

where $R_{\hat{u}}^2$ is the R^2 of the regression of residuals on a constant, $dldgp_{t-1}$, and $dlgdp_{t-2}$ and 8 lags of residuals. From the Eviews output we get

$$BG_{calc} = 8.972$$

with p-value of 0.345, which is much larger than 0.05. This means that we cannot reject the null at the 5% level of significance. Hence the AR(2) model seems adequate. We can then proceed to use t-test to see if both lags are needed, or an AR(1) would be sufficient.

Note: we lose q observations when we form q lags

Table 1

t	$\{\widehat{u}_t\}$	$\{\widehat{u}_{t-1}\}$	$\{\widehat{u}_{t-2}\}$
1	$\widehat{u}_1$	-	-
2	$\widehat{u}_2$	$\widehat{u}_1$	-
3	$\widehat{u}_3$	$\widehat{u}_2$	$\widehat{u}_1$
4	$\widehat{u}_4$	$\widehat{u}_3$	$\widehat{u}_2$
5	$\widehat{u}_5$	$\widehat{u}_4$	$\widehat{u}_3$

HAC standard error

- *How do we correct serial correlation/heteroskedasticity?*
- Alternative hypothesis tests which are valid in large samples, even when serial correlation and heteroskedasticity are present
- This allows different standard errors for each variable coefficient

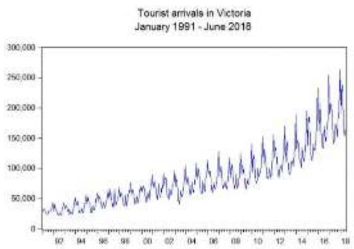
Week 10: Modelling Dynamics

- Static time series models
  - Trend
  - Seasonality
  - Structural breaks
- Dynamic time series models
  - Autoregressive models
  - Stationary time series models
  - The autoregressive distributed lag model

STATIC TIME SERIES MODELS:

1) Trend

- A log transformation is used to correct for an exponential trend



2) Seasonality

- A set of seasonal dummy variables can account for seasonality
- No seasonality if the null hypothesis of insignificance cannot be rejected

Model log(VIC) with a trend and dummies for 11 out of 12 months.

$$\log(VIC_t) = \beta_0 + \delta_1 feb_t + \delta_2 mar_t + \delta_3 apr_t + \ldots + \delta_{11} dec_t + \beta_1 t + u_t,$$

where $feb_t, \ldots, dec_t$ are dummies equal to 1 if the observation is from the month indicated by their names and 0 otherwise.

Here, January is the base month, and β_0 is the intercept for January.

One can test for joint significance of $\delta_1, \ldots, \delta_{11}$ via an F test. Once the trend is controlled for, if the null $H_0 : \delta_1 = \ldots = \delta_{11} = 0$ cannot be rejected, then this is indication of no seasonality.

3) Structural break

- I.e GFC, Covid etc.
- Insert a dummy of 0 before and a dummy of 1 after the event

Static model: no lags included as regressors in the model, i.e a change in x in time period t only affects y in time period t

$$mr_t = \beta_0 + \beta_1 cr_t + u_t,$$

where a change in cr_t only has an effect on mr_t.

Dynamic model: a change in time in time period t can affect y in different time periods

$$mr_t = \beta_0 + \beta_1 cr_t + \beta_2 cr_{t-1} + u_t$$

where a change in cr_t has an effect on mr_t and mr_{t+1}.

Why may variables have an effect over several time periods?

- Habit persistence.
- Institutional arrangements.
- Administrative lags.
- Optimizing behavior.

DYNAMIC TIME SERIES MODELS:

1) Autoregressive models (white noise properties, AR)

The simplest dynamic time series model is the AR(p) model

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots.. + \phi_p y_{t-p} + u_t,$$

A time series is white noise $u_t \sim WN(0, \sigma^2)$ if:

$$E(u_t) = 0 \text{ for all } t,$$
$$Var(u_t) = \sigma^2 \text{ for all } t,$$
$$Cov(u_t, u_{t-j}) = 0 \text{ for } j \neq 0.$$

The simplest AR model is the AR(1) model

$$y_t = \varphi_0 + \varphi_1 y_{t-1} + u_t.$$

- This model can produce:
  1. an uncorrelated sequence when $\varphi_1 = 0$
  2. a stationary process when $|\varphi_1| < 1$
  3. a random walk when $\varphi_0 = 0$ and $\varphi_1 = 1$
  4. a random walk with drift when $\varphi_0 \neq 0$ and $\varphi_1 = 1$
  5. an explosive process when $\varphi_1 > 1$

2) Stationary time series models

A time series $\{y_t\}$ is called stationary if it has the properties:

P1

$$E(Y_t) = \mu \ < \infty \ \text{ for all t.}$$

(The mean is finite and time invariant)

P2

$$Var(Y_t) = E[(Y_t - \mu)^2] = \gamma_0 \ < \infty \text{ for all t.}$$

(The variance is finite and time invariant)

P3

$$Cov(Y_t, Y_{t-j}) = E[(Y_t - \mu)(Y_{t-j} - \mu)] = \gamma_j \ < \infty \text{ for all t and j.}$$

(The covariance is finite and depends only on the time interval)

Covariance stationary time series ^^

- fytg is covariance stationary if its mean (1 mark) and variance (1 mark) are Önite and do not depend on t; and the covariance between yt and any of its lags only depend on the lag and does not depend on t

- If y_t is generated by an AR(1) model

$$y_t = \varphi_0 + \varphi_1 y_{t-1} + u_t, \text{ with } |\varphi_1| < 1 \text{ and } \{u_t\} \sim WN(0, \sigma^2)$$

then we have:

P1 Mean of y

$$E(y_t) = \frac{\varphi_0}{1 - \varphi_1} \text{ for all } t$$

P2 Variance of y

$$Var(y_t) = \frac{\sigma^2}{1 - \varphi_1^2} \text{ for all } t$$

P3 Autocovariances and autocorrelations of y

$$Cov(y_t, y_{t-j}) = \gamma_j = \varphi_1^j Var(y_t), \text{ for all } t \text{ and } j$$

$$Corr(y_t, y_{t-j}) = \rho_j = \frac{\gamma_j}{\gamma_0} = \varphi_1^j, \text{ for all } t \text{ and } j.$$

3) Autoregressive distributed lag model (ARDL)

OLS and standard errors and tests are reliable if:

1) X and y are stationary
2) The residuals show no sign of autocorrelation
3) The sample is large

Consider the ARDL(1,1) model

$$y_t = \varphi_0 + \varphi_1 y_{t-1} + \alpha_0 x_t + \alpha_1 x_{t-1} + u_t.$$

Suppose there is a one unit increase in x at time t.

- Contemporaneous change in y is α_0.
- At time $t+1$, this change is x will still affect y_{t+1} in two ways:
  - directly because x_t still appears in the equation for y_{t+1}
  - and again because y_t also appears in the equation for y_{t+1} (and y_t was already influenced by the change in x_t).
- The long-run effect on y of a one unit change in x_t equals

$$\frac{\alpha_0 + \alpha_1}{1 - \varphi_1}.$$

$$= \frac{\text{sum of the coefficients } x_t \text{ and its lags}}{1 - \text{sum of the coefficients of lags of } y_t}.$$

But, a restricted autoregressive distributed lag model:

$$y_t = \beta_0(1 - \rho) + \beta_1 x_t - \rho\beta_1 x_{t-1} + \rho y_{t-1} + e_t$$

- A one unit increase in x changes y by Beta1 in the short term and long run.
- This shows that the regression with AR errors imposes that all impact of x on y is realised immediately. All the dynamics in these models are in the error (the part of y that is not explained by x).

Summary

- As long as the dependent and independent variables are stationary and errors are white noise, the OLS estimator of the parameters of a dynamic model is reliable and we can use t and F tests provided the sample size is large.

AR models: only use the history of a time series to predict its future.
ARDL models: measure immediate and long-run effects.

Week 11: Large Sample Properties of OLS

The OLS estimator of the parameters of a dynamic model is reliable and we can use t and F tests provided that the sample size is large.

The AR(1) model

$$y_t = \beta_0 + \beta_1 y_{t-1} + u_t, \; t = 1, 2, ...n.$$

- It follows that $y_{t-1} = \beta_0 + \beta_1 y_{t-2} + u_{t-1}$.
- Hence $corr(u_{t-1}, y_{t-1}) \neq 0$.
- This violates A3: $corr(u_{t-1}, y_{t-1}) = 0$.

Consistency

A3 conditional mean of errors is zero: $E(u_t|x_t) = 0$ (NEW).

Consider a set of i.i.d. random variables $(X_1, \ldots, X_n)$ with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$ for $i = 1, \ldots, n$.

- Define $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
- Recall that $E(\bar{X}) = \mu$ and $Var(\bar{X}) = \sigma^2/n$.
- $\bar{X}$ is an unbiased estimator of the parameter μ.

The sample mean is also a consistent estimator of μ:

- As $n \to \infty$, $Var(\bar{X}) \to 0$.
- The chance of $\bar{X}$ being anything other than μ goes to zero.
- We say that $\bar{X}$ converges in probability to μ.
- We write that $\text{plim}(\bar{X}) = \mu$ or $\bar{X} \xrightarrow{p} \mu$.

If an estimator converges in probability to the population parameter that it estimates, we say that the estimator is consistent.

A consistent estimator can be biased: since the estimator converges in probability to the population parameter that it estimates, we say that the estimator is consistent.

Asymptotic normality

A4 homoskedasticity and no serial correlation:
$Var(u_t|x_t) = \sigma^2$ for all t and $E(u_t u_s|x_t, x_s) = 0$ for all $t \neq s$ (NEW).

Under these assumptions, the OLS estimator is asymptotically normal:

$$\widehat{\beta} \overset{a}{\sim} N(\beta, \sigma^2(X'X)^{-1}).$$

As $n \to \infty$,

$$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \xrightarrow{d} N(0,1),$$

since t with a large degree of freedom is approximately a $N(0,1)$.

Homoskedasticity and serial correlation

Where (NEW) A4 does not hold → use HAC standard errors

Under these assumptions, the OLS estimator is asymptotically normal:

$$\widehat{\boldsymbol{\beta}} \overset{a}{\sim} N(\boldsymbol{\beta}, \sigma^2(\mathrm{X}'\mathrm{X})^{-1}).$$