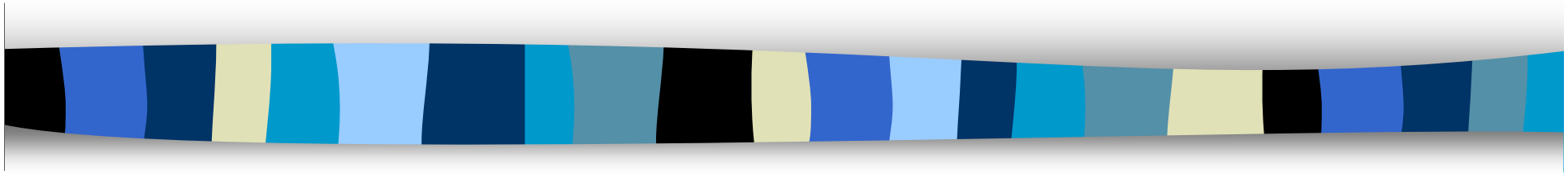
SSSII

Gwilym Pryce

www.gpryce.com



L9: Revision Lecture

Plan:

- (I) F-Tests
- (II) Heteroscedasticity
- (III) Multicolinearity
- (IV) Modelling Strategy

(I) F-Tests

- When should you use F-tests in regression analysis?
  - Most common:
    a) To test whether all coefficients equal zero
    b) To test whether a group of coefficients equal zero
    c) Chow tests for structural breaks
  - Also used for:
    d) testing other linear restrictions:
      - » E.g. $b_1 + b_2 = 4$
      - » E.g. $b_1 = b_2$
    - Only of interest if your theoretical model gives you specific hypotheses about the values of coefficients.
    - This is not unusual in economics, but very unusual in other social science disciplines so we won't spend more time on it here.

(a) To test that all coefficients equal zero:

- A special case of homogenous restrictions is where we test for the *existence of a relationship*
  - I.e. H_0: all slope coefficients are zero
- Unrestricted regression:

$$y = b_1 + b_2x_2 + b_3x_3 + u$$

- Restricted regression:
  - H_0: $b_2 = b_3 = 0$;
  - If H_0 is true, then $y = b_1$
- In this case, Restricted regression does no explaining at all and so $R_R^2 = 0$

And the *homogenous restriction* F-ratio test statistic reduces to:

$$F^{df_{\text{numerator}}}_{df_{\text{denominator}}} = F^{r}_{df_U} = \frac{(R_U^2 - 0)/r}{1 - R_U^2/df_U}$$

$$= \frac{(R_U^2)/r}{1 - R_U^2/df_U}$$

where,

$r = k - 1$

$df_U = n - k$

- This is the F-test we came across in MII Lecture 2, and is the one automatically calculated in the SPSS ANOVA table

(b) Testing whether a subset of coefficients equal zero

- Suppose we want to test whether there are any country specific effects in the relationship between inflation and the money supply:

$$\text{INFL} = a + b\,\text{MS} + g_1\,\text{COUNTRY}_1 + \ldots + g_{42}\,\text{COUNTRY}_2$$

  - I.e. we want to test the following null hypothesis:
    - H_0: $g_1 = g_2 = g_3 = \ldots = g_{42} = 0$
- Then we can think of this as being equivalent to comparing two regressions, one restricted and one unrestricted:

- The *Unrestricted* regression is:

  $$\text{INFL} = a + b\,\text{MS} + g_1\,\text{COUNTRY}_1 + \ldots + g_{42}\,\text{COUNTRY}_2$$

- The *Restricted* regression is:

  $$\text{INFL} = a + b\,\text{MS}$$

- We can test whether all the *g* coefficients equal zero using the F-test:

The General formula for F:

$$F^{df_{\text{numerator}}}_{df_{\text{denominator}}} = F^{r}_{df_U} = \frac{(RSS_R - RSS_U)/r}{RSS_U/df_U}$$

NB RSS_R is always greater than RSS_U since imposing a restriction on an equation can never reduce the RSS

Where:

RSS_U = restricted residual sum of squares

= RSS under H_1

RSS_R = unrestricted residual sum of squares

= RSS under H_0

r = number of restrictions = diff. in no. parameters between restricted and unrestricted equations

df_u = *df* from unrestricted regression = n - k where k is all coefficients including the intercept.

Using the F-test:

- If the null hypothesis is true (i.e. restrictions are satisfied) then we would expect the restricted and unrestricted regressions to give similar results
  - I.e. RSS_R and RSS_U will be similar
  - so we **accept H_0** when the test statistic gives a **small** value for F.
- But if one of the restrictions does not hold, then the restricted regression will have had an invalid restriction imposed upon it and will be mis-specified.
  - $\Rightarrow$ higher residual variation $\Rightarrow$ higher RSS_R
  - so we **reject H_0** when the test stat. gives a **large** value

Test Procedure:

- (i) Compute RSS_U
  - Run the *unrestricted* form of the regression in SPSS and take a note of the residual sum of squares = RSS_U
- (ii) Compute RSS_R
  - Run the *restricted* form of the regression in SPSS and take a note of the residual sum of squares = RSS_R
- (iii) Calculate r and df_U
- (iv) Substitute RSS_U, RSS_R, r and df_U in the equation for F and find the significance level associated with the value of F you have calculated.

Example 1: H_o: no country effects

*(R and U regressions have the **same** dependent variable)*

Step (i) $RSS_U = 1835.811$

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	296.772	4	74.193	20.652	.000[a]
	Residual	1835.811	511	3.593		
	Total	2132.583	515			

a. Predictors: (Constant), CNTRY_3, CNTRY_2, CNTRY_1, money supply

b. Dependent Variable: inflation

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.373[a]	.139	.132	1.895

a. Predictors: (Constant), CNTRY_3, CNTRY_ CNTRY_1, money supply

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.281	1.219		2.692	.007
	money supply	3.787E-02	.055	.037	.693	.489
	CNTRY_1	-1.716	.692	-.127	-2.479	.014
	CNTRY_2	-4.446	.562	-.330	-7.914	.000
	CNTRY_3	-1.384	.573	-.103	-2.413	.016

a. Dependent Variable: inflation

Step (ii) $RSS_R = 2097.722$

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	34.860	1	34.860	8.542	.004[a]
	Residual	2097.722	514	4.081		
	Total	2132.583	515			

a. Predictors: (Constant), money supply

b. Dependent Variable: inflation

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.128[a]	.016	.014	2.020

a. Predictors: (Constant), money supply

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.031	1.000		1.031	.303
	money supply	.132	.045	.128	2.923	.004

a. Dependent Variable: inflation

Step (iii) r and df_u

- r = number of restrictions

  = difference in no. of parameters between the restricted and unrestricted equations

  = **3**

- df_u = df from unrestricted regression = $n_U - k_U$

  where k is total number of all coefficients including the intercept

- = 516 - 5 = **511**

(iv) Substitute RSS_U, RSS_R, r and df_U in the formula for F

- $$F = \frac{(RSS_R - RSS_U) / r}{RSS_U / df_U} = \frac{(2097.722 - 1835.811)/3}{1835.811 / 511}$$

$$= \frac{87.304}{3.593}$$

$$= \mathbf{24.3}$$

- $df_{numerator} = r = \mathbf{3}$
- $df_{denominator} = df_U = \mathbf{511}$
- From Tables, we know that at P = 0.01, the value for $F[3,511]$ would be 3.88 (I.e. Prob(F > 3.88) = 0.01)
- F we have calculated is > 3.88, so we know that P < 0.01

(I.e. Prob(F > 24.3) <0.01) ∴ **Reject Ho**

F Critical Values

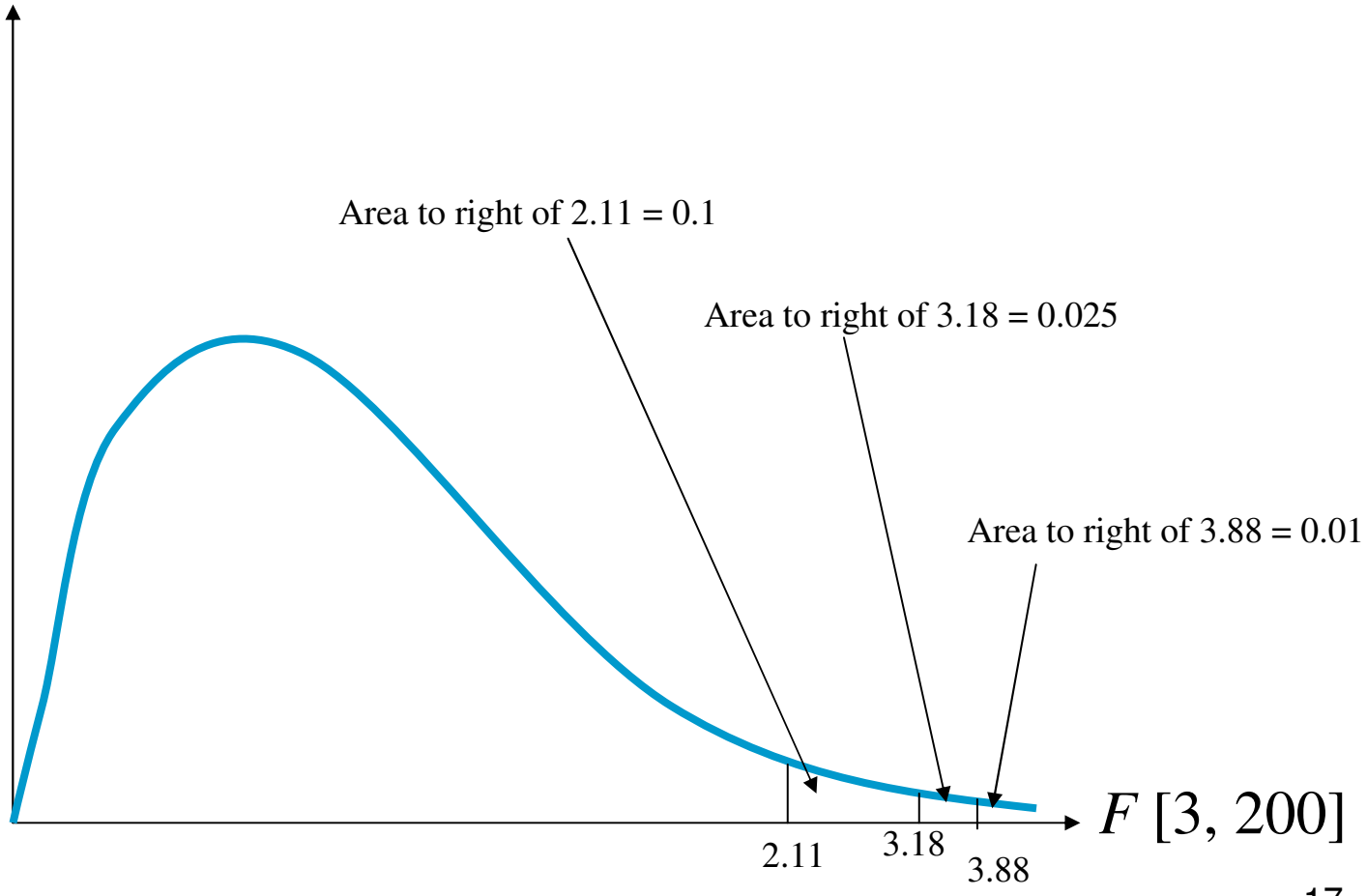
			Degrees of freedom in the numerator						
		p	1	2	3	4	5	6	7
Degrees of freedom in the denomin	200	0.100	2.73	2.33	2.11	1.97	1.88	1.80	1.75
		0.050	3.89	3.04	2.65	2.42	2.26	2.14	2.06
		0.025	5.10	3.76	3.18	2.85	2.63	2.47	2.35
		0.010	6.76	4.71	3.88	3.41	3.11	2.89	2.73
		0.001	11.15	7.15	5.63	4.81	4.29	3.92	3.65
	1000	0.100	2.71	2.31	2.09	1.95	1.85	1.78	1.72
		0.050	3.85	3.00	2.61	2.38	2.22	2.11	2.02
		0.025	5.04	3.70	3.13	2.80	2.58	2.42	2.30
		0.010	6.66	4.63	3.80	3.34	3.04	2.82	2.66
		0.001	10.89	6.96	5.46	4.65	4.14	3.78	3.51

F Critical Values

			Degrees of freedom in the numerator						
		p	1	2	3	4	5	6	7
Degrees of freedom in the denomir	200	0.100	2.73	2.33	2.11	1.97	1.88	1.80	1.75
		0.050	3.89	3.04	2.65	2.42	2.26	2.14	2.06
		0.025	5.10	3.76	3.18	2.85	2.63	2.47	2.35
		0.010	6.76	4.71	3.88	3.41	3.11	2.89	2.73
		0.001	11.15	7.15	5.63	4.81	4.29	3.92	3.65
	1000	0.100	2.71	2.31	2.09	1.95	1.85	1.78	1.72
		0.050	3.85	3.00	2.61	2.38	2.22	2.11	2.02
		0.025	5.04	3.70	3.13	2.80	2.58	2.42	2.30
		0.010	6.66	4.63	3.80	3.34	3.04	2.82	2.66
		0.001	10.89	6.96	5.46	4.65	4.14	3.78	3.51

Stylised F-Distribution:

$df_1 = 3$; $df_2 = 200$



Alternatively use Excel calculator:

F-Tests.xls

First Paste ANOVA tables of U and R models:

Restricted Model

ANOVA

Model		Sum of Sq	df	Mean Squa	F	Sig.
1	Regressior	34.86042	1	34.86042	8.541767	0.003624
	Residual	2097.722	514	4.081172		
	Total	2132.583	515			

a Predictors: (Constant), money supply

b Dependent Variable: inflation

Unrestricted Model

ANOVA

Model		Sum of Sq	df	Mean Squa	F	Sig.
1	Regressior	296.7719	4	74.19296	20.65169	0
	Residual	1835.811	511	3.592585		
	Total	2132.583	515			

a Predictors: (Constant), CNTRY_3, CNTRY_2, CNTRY_1, money supply

b Dependent Variable: inflation

Second, check cell formulas, & let Excel do the rest:

F Test

$$F^{r}_{df_U} = \frac{(RSS_R - RSS_U)/r}{RSS_U/df_U}$$

				r	=	$k_U - k_R$	=	3
				k_u			=	5
				n			=	516
				df_u	=	$n - k_u$	=	511
							=	511
				F			=	24.30111338
				Sig F			=	1.02857E-14

Example 1: H_o: no country effects

*(R and U regressions have the **same** dependent variable)*

- Our approach to this restriction when we tested it above was to use the RSSs as follows:

$$F = \frac{(RSS_R - RSS_U)\,/\,r}{RSS_U/df_U} = \frac{(2097.722 - 1835.811)/3}{1835.811\,/\,511}$$

$$= \mathbf{24.301}$$

$$Prob(F > F[3,511]\ 24.298) = 1.028E\text{-}14 \quad \therefore \mathbf{Reject\ H_0}$$

- Since it is a homogenous restriction (I.e. dep var is same in restricted and unrestricted models), we shall now attempt the same test but using the R^2 formulation of the *F*-ratio formula:

- Unrestricted model: $R_U^2 = 0.139$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.373[a]	.139	.132	1.895

a. Predictors: (Constant), CNTRY_3, CNTRY_ CNTRY_1, money supply

- Restricted model: $R_R^2 = 0.016$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.128[a]	.016	.014	2.020

a. Predictors: (Constant), money supply

$$F = \frac{(R_U^2 - R_R^2) / r}{(1-R_U^2) / df_U} = \frac{(0.139 - 0.016)/3}{(1- 0.139) / 511} = \frac{0.041}{0.0017} = \mathbf{24.301}$$

Restricted Model

ANOVA

Model		Sum of Sq	df	Mean Squa	F	Sig.
1	Regressior	34.86042	1	34.86042	8.541767	0.003624
	Residual	2097.722	514	4.081172		
	Total	2132.583	515			
a	Predictors: (Constant), money supply					
b	Dependent Variable: inflation					

Model Summary

Model	R	R Square	Adjusted F	Std. Error of the Estimate
1	0.127854	0.016347	0.014433	2.020191
a	Predictors: (Constant), money supply			

Unrestricted Model

ANOVA

Model		Sum of Sq	df	Mean Squa	F	Sig.
1	Regressior	296.7719	4	74.19296	20.65169	0
	Residual	1835.811	511	3.592585		
	Total	2132.583	515			
a	Predictors: (Constant), CNTRY_3, CNTRY_2, CNTRY_1, money supply					
b	Dependent Variable: inflation					

Model Summary

Model	R	R Square	Adjusted F	Std. Error of the Estimate
1	0.373043	0.139161	0.132422	1.895412
a	Predictors: (Constant), CNTRY_3, CNTRY_2, CNTRY_1, money supply			

F Test Statistic for homogenous restrictions:

$$F^{df_{\text{numerator}}}_{df_{\text{denominator}}} = F^{r}_{df_U} = \frac{(R_U^2 - R_R^2)/r}{1 - R_U^2/df_U}$$

			r	=	$k_U - k_R$	=	3
			k_u			=	5
			n			=	516
			df $_u$	=	$n - k_u$	=	511
						=	511
			F			=	24.30111
			Sig F			=	1.03E-14

(c) Testing for Structural Breaks

- The F-test also used to test whether the estimated coefficients change significantly if we split the sample in two at a given point
- These tests are sometimes called “Chow Tests” after one of its proponents.
- There are actually two versions of the test:
  - Chow’s first test
  - Chow’s second test

(i) Chow's First Test – Best to use this one if possible

Use where $n_2 > k$

- (1) Run the regression on the first set of data (i = 1, 2, 3, … n_1) & let its RSS be RSS_{n1}
- (2) Run the regression on the second set of data (i = n_1+1, n_1+2, …, end of data) & let its RSS be RSS_{n2}
- (3) Run the regression on the two sets of data combined (i = 1, …, end of data) & let its RSS be $RSS_{n1 + n2}$

- (4) Compute RSS_U, RSS_R, r and df_U:
  - $RSS_U = RSS_{n1} + RSS_{n2}$
  - $RSS_R = RSS_{n1 + n2}$
  - $r = k$ = total no. of coeffts including the constant
  - $df_U = n_1 + n_2 - 2k$
- (5) Use RSS_U, RSS_R, r and df_U to calculate F using the general formula for F and find the sig. Level:

$$F^{df_{\text{numerator}}}_{df_{\text{denominator}}} = F^{r}_{df_U} = \frac{(RSS_R - RSS_U)/r}{RSS_U / df_U}$$

(ii) Chow's Second Test

Only use where $n_2 < k$ (I.e. when you have insufficient observations on 2nd subsample to do Chow's 1st test) – less reliable, particularly if you have heteroscedasticity.

- (1) Run the regression on the first set of data ($i = 1, 2, 3, \ldots n_1$) & let its RSS be RSS_{n1}
- (2) Run the regression on the two sets of data combined ($i = 1, \ldots,$ end of data) & let its RSS be $RSS_{n1 + n2}$

- (3) Compute RSS_U, RSS_R, r and df_U:
  - $\boldsymbol{RSS_U = RSS_{n1}}$
  - $RSS_R = RSS_{n1 + n2}$
  - $r = n_2$
  - $df_U = n_1 - k$
- (4) Use RSS_U, RSS_R, r and df_U to calculate F using the general formula for F and find the sig.:

$$F^{df_{\text{numerator}}}_{df_{\text{denominator}}} = F^{r}_{df_U} = \frac{(RSS_R - RSS_U)/r}{RSS_U / df_U}$$

Example of Chow's 1st Test:

n_1: before 1986:

Model		Unstand Coeffic B
1	(Constant)	-8.183
	MS_GDP	3.648
	MP_GDP	8.767
	CNTRY_1	-3.406
	CNTRY_2	-7.164
	CNTRY_3	-3.585
	CNTRY_4	.214
	CNTRY_5	.320
	CNTRY_6	.873
	CNTRY_7	-7.85E-02
	CNTRY_8	.764E-02
	CNTRY_9	.878E-02

a. Dependent Variable: inflat

n_2: 1986 and after

Model		Unstanda Coeffic B
1	(Constant)	16.393
	MS_GDP	-5.806
	MP_GDP	-6.240
	CNTRY_1	-.167
	CNTRY_2	-.633
	CNTRY_3	1.943
	CNTRY_4	-.195
	CNTRY_5	-.400
	CNTRY_6	-.559
	CNTRY_7	-.130
	CNTRY_8	-.111
	CNTRY_9	6.658E-02

a. Dependent Variable: inflation

ANOVA from n 1						
ANOVA						
Model		Sum of Sq	df	Mean Squa	F	Sig.
1	Regressior	670.1102	11	60.91911	31.2705	4.53E-43
	Residual	563.0107	289	1.948134		
	Total	1233.121	300			
a	Predictors: (Constant), CNTRY_9, CNTRY_8, CNTRY_7, CNTRY_					
b	Dependent Variable: inflation					
ANOVA from n 2						
ANOVA						
Model		Sum of Sq	df	Mean Squa	F	Sig.
1	Regressior	65.9947	11	5.999519	3.428609	0.000217
	Residual	355.2176	203	1.749841		
	Total	421.2124	214			
a	Predictors: (Constant), CNTRY_9, CNTRY_8, CNTRY_7, CNTRY_					
b	Dependent Variable: inflation					
ANOVA from n1 + n2						
ANOVA						
Model		Sum of Sq	df	Mean Squa	F	Sig.
1	Regressior	300.8519	11	27.35017	7.525389	0
	Residual	1831.731	504	3.634387		
	Total	2132.583	515			

k			=	12
r	=	*k*	=	12
n1			=	301
n2			=	215
df $_{u}$	=	*n - k*	=	492
RSS $_{R}$	= RSS_{n1+n2}		=	1831.730825
RSS_U	= $RSS_{n1} + RSS_{n2}$		=	918.2283021
F			=	40.78898826
Sig F			=	4.18611E-66

$$F_{df_U}^{r} = \frac{(RSS_R - RSS_U)/r}{RSS_U / df_U}$$

Summary of F-Tests:

- Use F-tests to:
  a) test for a relationship – i.e. whether all coefficients equal zero.
  b) test whether a subset of coefficients equal zero (e.g. whether there are country effects)
  c) test for structural breaks

(II) Heteroscedasticity

a) Definition
   - Non-constant variance

b) Causes
   - Misspecification or aggregated data

c) Consequences
   - Incorrect t-ratios.

d) Detection
   - B-P/Koenker test

e) Remedy
   - White's standard errors

(a) Definition:

- Error term does not have constant variance.

(b) Causes:

i. non-constant coefficients
- Unless your model allows for varying coefficients, error becomes correlated with one or more of the independent variables.

ii. Omitted variables
- Error term becomes correlated with the omitted variable

iii. Non-linearity
- If your model does not capture the non-linear relationship between y and x, the error term will – i.e. it will become vary non-linearly with x

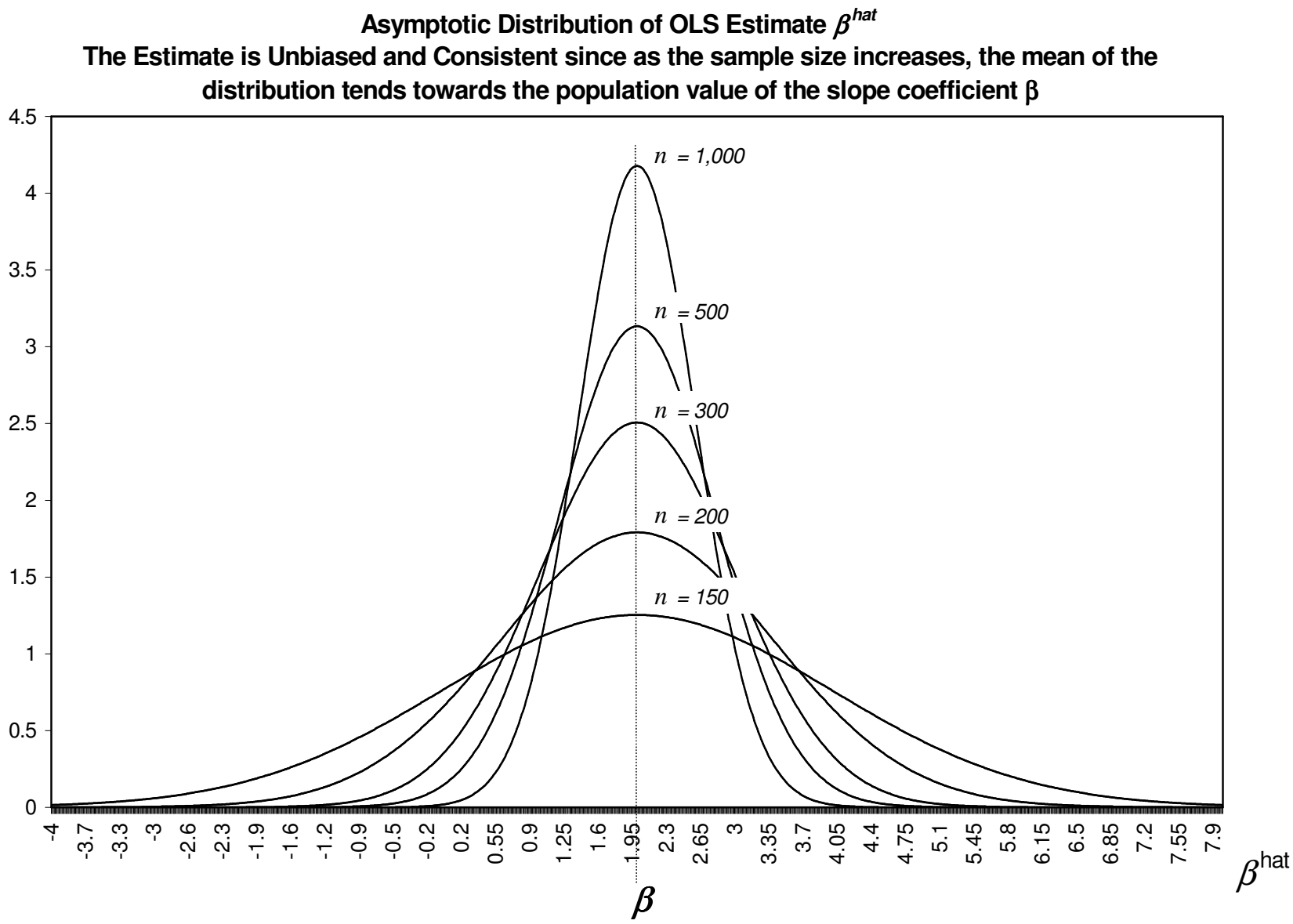
iv. Aggregation
- Often arises if use variables based on group averages where sample size varies
  - Different sample sizes lead to different sampling distributions with standard errors of the mean
- If your variables are averages across groups (e.g. average income per post code sector), and the sample size used to calculated those averages varies (e.g. some postcode sectors have more observations on individual income)

(c) Consequences

- Heteroscedasticity causes bias in the OLS estimated standard errors for the estimated coefficients:
  - which means that the *t* tests will not be reliable:
    $t = b^{hat} / \mathrm{SE}(b^{hat})$.
  - F-tests are also no longer reliable
    - e.g. Chow's second Test no longer reliable (Thursby)
- But Heteroscedasticity by itself does not cause OLS **estimators** (i.e. slope coefficients) to be biased or inconsistent*
  - NB neither bias nor consistency are determined by the covariance matrix of the error term.
- However, if heteroscedasticity is a symptom of omitted variables, measurement errors, or non-constant parameters,

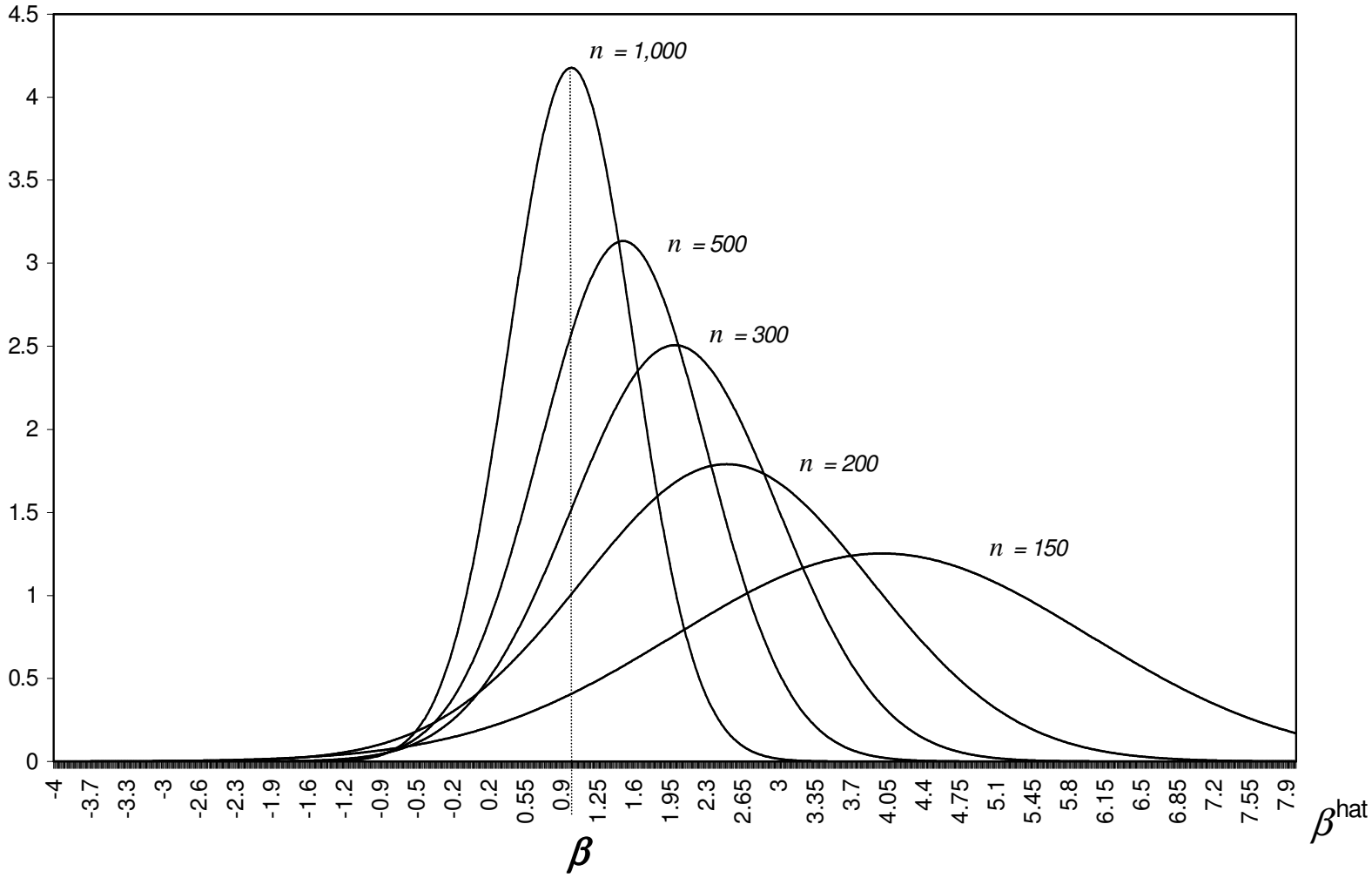
  $\Rightarrow$ OLS estimators will be biased and inconsistent.

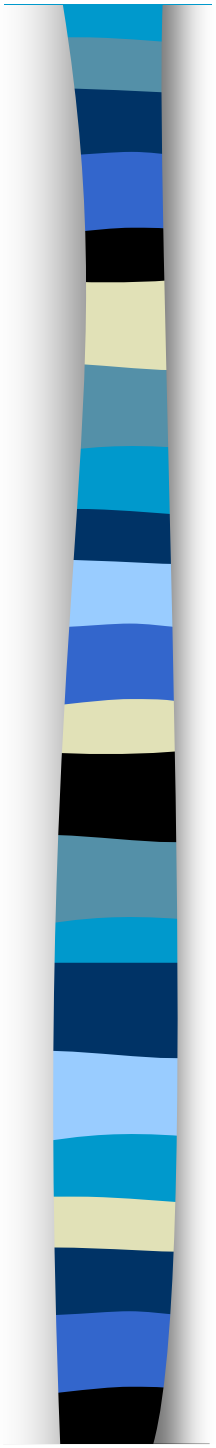
Unbiased and Consistent Estimator



Biased but Consistent Estimator

Asymptotic Distribution of OLS Estimate β^{hat}
The Estimate is Biased but Consistent since as the sample size increases, the mean of the distribution tends towards the population value of the slope coefficient β





- NB not heteroscedasticity that causes the bias,
- but failure of one of the other assumptions that happens to have hetero as the side effect.

  ⇒ testing for hetero. is closely related to tests for misspecification generally.

  - Unfortunately, there is usually no straightforward way to identify the cause

(d) Detection

- Specific tests:
  - Early tests for heterosc. (e.g. Goldfield-Quandt) have gone out of fashion because you rarely know the cause of the heterosc.
- General tests:
  - B-P and White's test
    - Best to use B-P test unless you are using a programme with White's test built in as it is a pain to do manually.
  - Disadvantage with general tests is that they do not give you any clues about the cause of heterosc.
    - You can view them as a general test for mispecification.
    - E.g. if B-P is large, then need to think about whether your model has been misspecified in some way.

Breusch-Pagan Test :

- Assumes that:

$$\sigma_i^2 = a_1 + a_2 z_1 + a_3 z_3 + a4 z_4 \ldots a_m z_m \qquad [1]$$

where *z*'s are all independent variables. *z*'s can be some or all of the original regressors or some other variables or some transformation of the original regressors which you think cause the heteroscedasticity:

$$\text{e.g. } \sigma_i^2 = a_1 + a_2 \exp(x_1) + a_3 x_3^2 + a_4 x_4$$

Procedure for B-P test:

(i) Obtain OLS residuals u_i^{hat} from the original regression equation and construct a new variable g:

$g_i = u^{hat\,2} / \sigma_i^{hat\,2}$

where $\sigma_i^{hat\,2}$ = RSS / n

(ii) Regress g_i on the z's (include a constant in the regression)

(iii) $B = {}^1/_2$(REGSS) from the regression of g_i on the z's,

where B has a Chi-square distribution with m-1 degrees of freedom.

Problems with B-P test:

- B-P test is not reliable if the errors are not normally distributed and if the sample size is small
- Koenker (1981) offers an alternative calculation of the statistic which is less sensitive to non-normality in small samples:

$$B^{\text{Koenker}} = nR^2 \sim \chi^2_{m\text{-}1}$$

where n and R^2 are from the regression of $u^{hat\,2}$ on the z's, where B^{Koenker} has a Chi-square distribution with m-1 degrees of freedom.

(e) Remedy

- **White's Standard Errors**
  - White (op cit) developed an algorithm for correcting the standard errors in OLS when heteroscedasticity is present.
  - The correction procedure does not assume any particular form of heteroscedasticity and so in some ways White has "solved" the heteroscedasticity problem.

(III) Multicolinearity

a) Definition
   - correlation between linear combinations of the explanatory variables

b) Causes
   - (i) Coincidence; (ii) x variables are different measures or aspects of the same variable;

c) Consequences
   - Increased standard errors, volatile coefficients

d) Detection
   - Check t-values; sensitivity;

e) Remedy
   - (i) do nothing; (ii) use groups of coefficients; (iii) create composites.

(d) Detection

- *Check for unstable parameter values across subsamples/specifications*
- *Check the t ratios*
- *Check the Tolerance and VIF*
  - If R_k^2 is the R^2 from a regression of x_k on the other explanatory variables:
    - e.g. R^2 from the regression: $x_1 = a_1 + a_2x_2 + a_3x_3$
  - Then 1 - R_k^2 is called the *Tolerance* of x_k.
    - A *tolerance* close to 1 means there is little multicolinearity, whereas a value close to **0** suggests multicolinearity may be a threat.
  - The reciprocal of the tolerance is known as the *Variance Inflation Factor (VIF)*
    - The *VIF* shows us how much the variance of the coefficient estimate is being inflated by multicolinearity.
    - A *VIF* near to one suggests there is no multicolinearity, whereas a VIF near **5** suggests multicolinearity.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-5863.031	674.178		-8.697	.000		
	Bedrooms	12885.593	280.324	.344	45.967	.000	.730	1.371
	PublicRooms	26431.578	439.243	.434	60.175	.000	.785	1.273
	HasGarden	347.620	516.156	.005	.673	.501	.843	1.186
	Time on the Market (number of days)	-15.213	1.289	-.076	-11.798	.000	.997	1.003

a. Dependent Variable: SellingPrice

- All the *VIF* levels in this regression are not close to 5 so there is no real problem.

More sophisticated approach:

- *Check the Eigenvalues and Condition Index:*
  - eigenvalues indicate how many distinct dimensions there are among the regressors
    - Dimensions are estimated using principle components
      - If you have ten variables but only two principle components, then most of your variables are not really independent – only really two independent drivers of y.
    - when several eigenvalues are close to zero, there may be a high level of multicolinearity.
  - Condition Indices are the square roots of the ratio of the largest eigenvalue to each successive eigenvalue.
    - **CI > 15 suggests multicolinearity** high
    - CI > 30 => multicolinearity very high
  - Only a problem if:
    - component has both a high CI **AND** contributes substantially to variance of two or more variables
    - i.e. if a variable has large CI AND large variance proportion in two or more variables, it suggests that those two variables are being driven by one component.

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions				
				(Constant)	Bedrooms	PublicRooms	HasGarden	Time on the Market (number of days)
1	1	4.028	1.000	.01	.01	.01	.01	.02
	2	.597	2.597	.00	.01	.01	.04	.88
	3	.211	4.370	.04	.03	.09	.92	.05
	4	8.810E-02	6.761	.12	.39	.86	.03	.01
	5	7.643E-02	7.259	.83	.57	.03	.01	.05

a. Dependent Variable: SellingPrice

- Two of the eigenvalues are pretty small, but:
- The dimensions All have Condition Indices are all below 10 so there is unlikely to be a problem with multicolinearity here.
- None of the dimensions constitute a high variance proportion of more than one explanatory variable.

- Conclusion:
  - Multicolinearity not a problem

- Problems with the Condition Index Approach:
  - the condition number can change by a reparametrization of the variables: “it can be made equal to one with suitable transformations of the variables” (Maddala, p. 275)
  - such transformations can be meaningless
  - does not tell you whether the multicolinearity is actually causing problems or how to go about resolving the problems if they exist.
- i.e. multicolinearity only a problem if it is causing instability, wrong signs etc.
- Also, only drop a variable if it is not theoretically relevant.

(IV) Modelling Strategy

a) Start with good theory & hypothetical model
   - Y caused by 3 or 4 key dimensions

b) Then create a "general" model with lots of variables, >1 measure of each dimension

c) Then refine the model using White's SEs
   - Dropping out insignificant variables but keeping in at least one measure of each dimension.
   - Think about non-linearities, structural breaks

d) Interpret & Apply the refined model
   - Interpret the results carefully and fully
     - Which variables are most sig and which most important? Is the model a good fit?
   - Test your theoretical presuppositions/hypoths
   - Use the model to simulate different scenarios