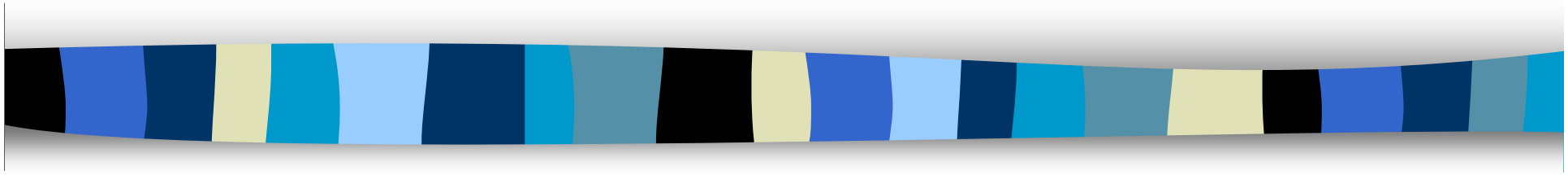
Graduate School

Social Science Statistics II

Gwilym Pryce



Lecture 6: Heteroscedasticity

Plan:

- 1. What is heteroscedasticity?
- 2. Causes
- 3. Consequences
- 4. Detection
- 5. Solutions

1. What is heteroscedasticity?

- For estimation of coefficients and standard errors (t ratios) to be correct:
  - 1. Equation is correctly specified:
  - 2. Error Term has zero expected mean
  - **3. Error Term has constant variance**
  - 4. Error Term is not autocorrelated
  - 5. Explanatory variables are fixed
  - 6. No linear relationship between RHS variables

- When assumption 3 holds,
  - i.e. the errors u_i *(sometimes denoted as* e_i*)* in the regression equation have common variance
    - i.e. constant or "scalar" variance

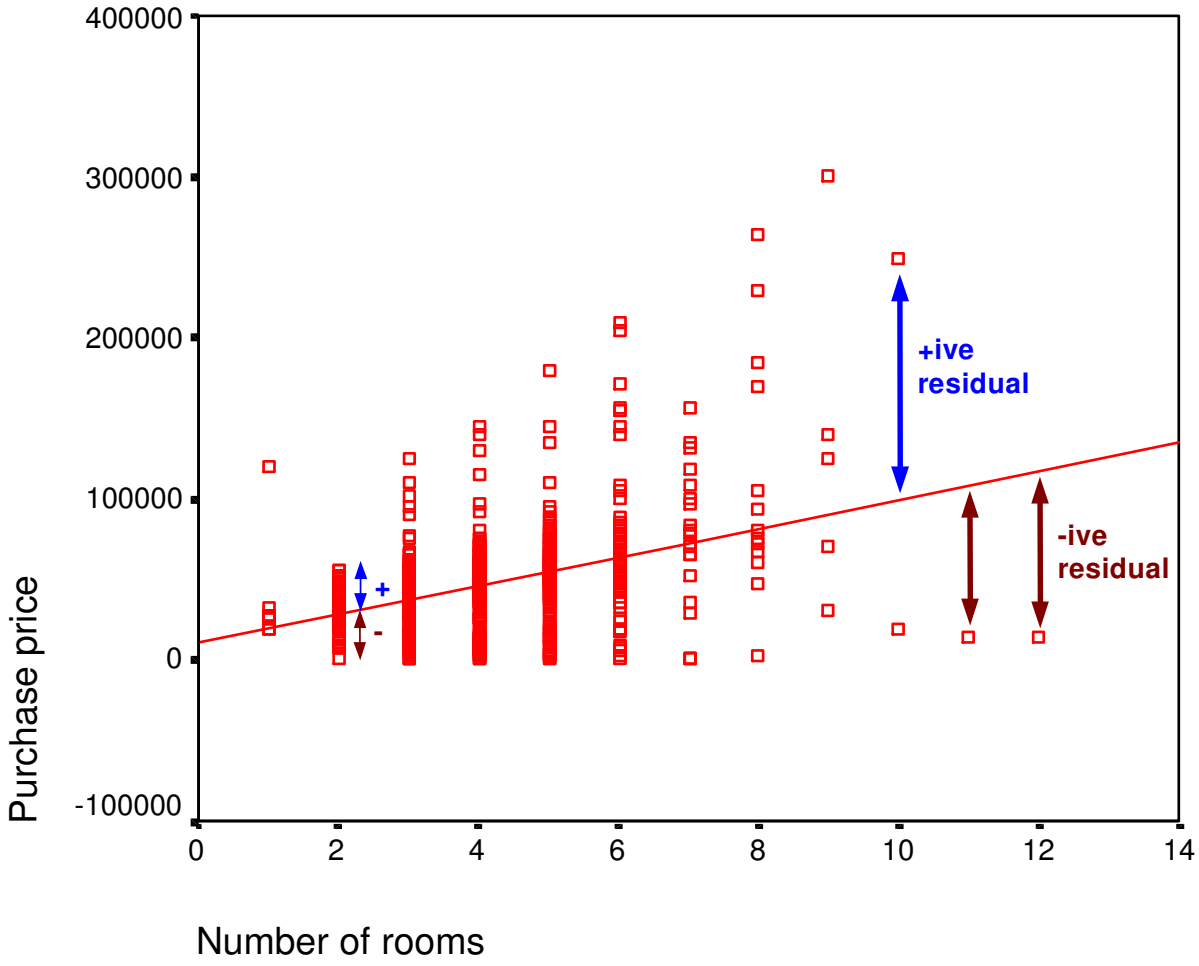
  then we have *homoscedasticity*.
  - or a "scalar error covariance matrix"
- When assumption 3 breaks down, we have what is known as *heteroscedasticity*.
  - or a "non-scalar error covariance matrix" (also caused by violation of assumption 4)

- Recall that the value of the Residual for each observation *i* is the vertical distance between the *observed* value of the dependent variable and the *predicted* value of the dependent variable
  - I.e. the difference between the observed value of the dependent variable and the line of best fit value:

Case	Price	Predicted Price	Residual
1	19000	19174.0	-174.0
2	30000	28028.7	1971.3
3	8100	45738.2	-37638.2
4	55000	36883.5	18116.5
5	130000	45738.2	84261.8
6	55000	45738.2	9261.8
7	54000	36883.5	17116.5
8	7500	45738.2	-38238.2
9	36000	36883.5	-883.5
10	32000	28028.7	3971.3

N.B. ***Predicted price*** is the value on the regression line that corresponds to the values of the dependent variables (in this case, No. rooms) for a particular observation.

(Assume that this represents multiple observations of *y* for each given value of *x*):



Homoskedasticity => variance of error term constant for each observation

- Error covariance matrix:
  - Imagine a matrix (table) which tabulates the covariance of the error term for each value of x

	u_1	u_2	u_3	u_4
u_1				
u_2				
u_3				

  - Symmetric matrix
  - Along the diagonal are the variances

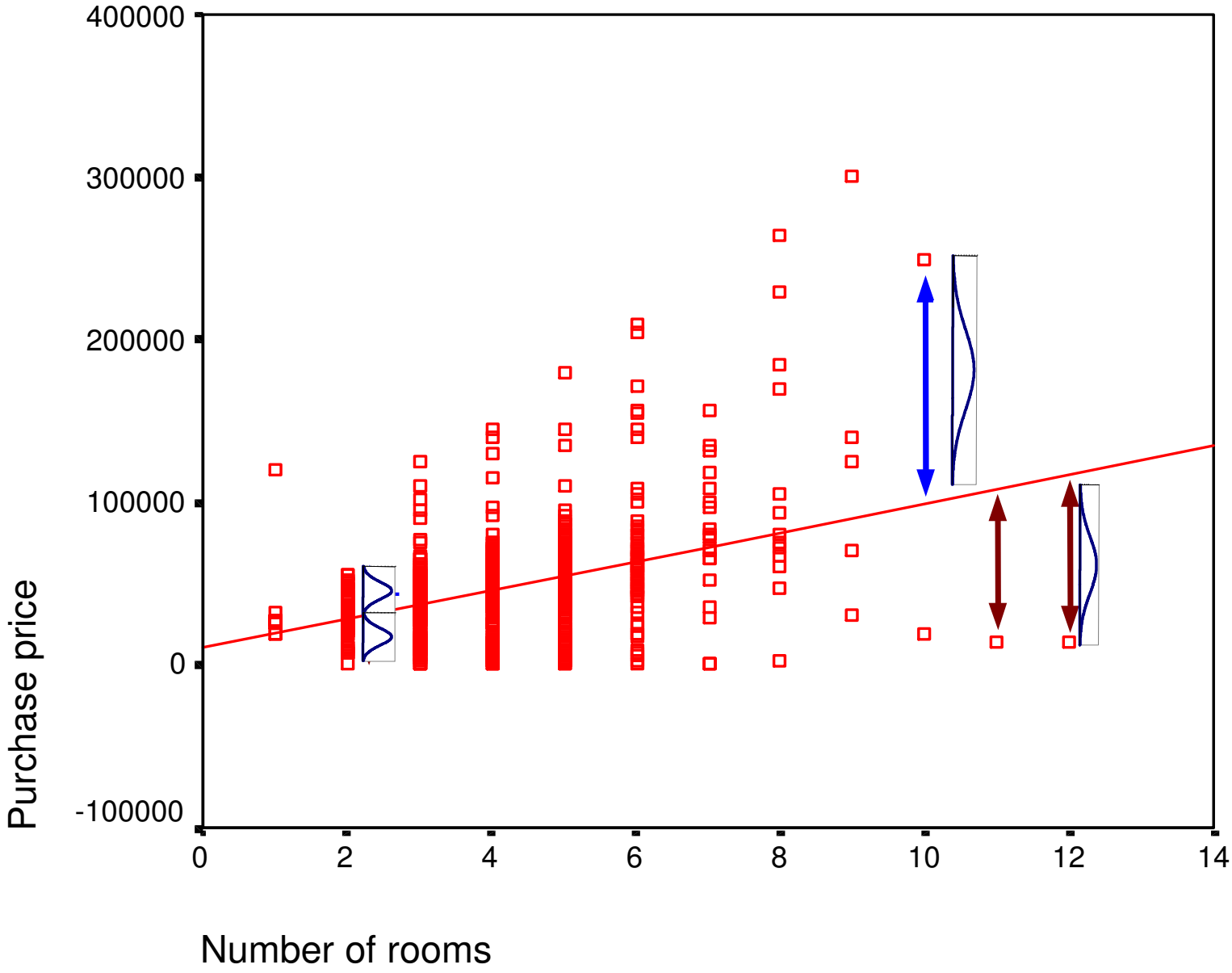
Error covariance matrix:

$$\mathrm{cov}(u_1, u_2, \ldots . u_n) = \begin{bmatrix} \mathrm{var}(u_1) & \mathrm{cov}(u_1, u_2) & \cdots & \mathrm{cov}(u_1, u_n) \\ \mathrm{cov}(u_2, u_1) & \mathrm{var}(u_2) & & \mathrm{cov}(u_2, u_n) \\ \vdots & & \ddots & \vdots \\ \mathrm{cov}(u_n, u_1) & \mathrm{cov}(u_n, u_2) & \cdots & \mathrm{var}(u_n) \end{bmatrix}$$

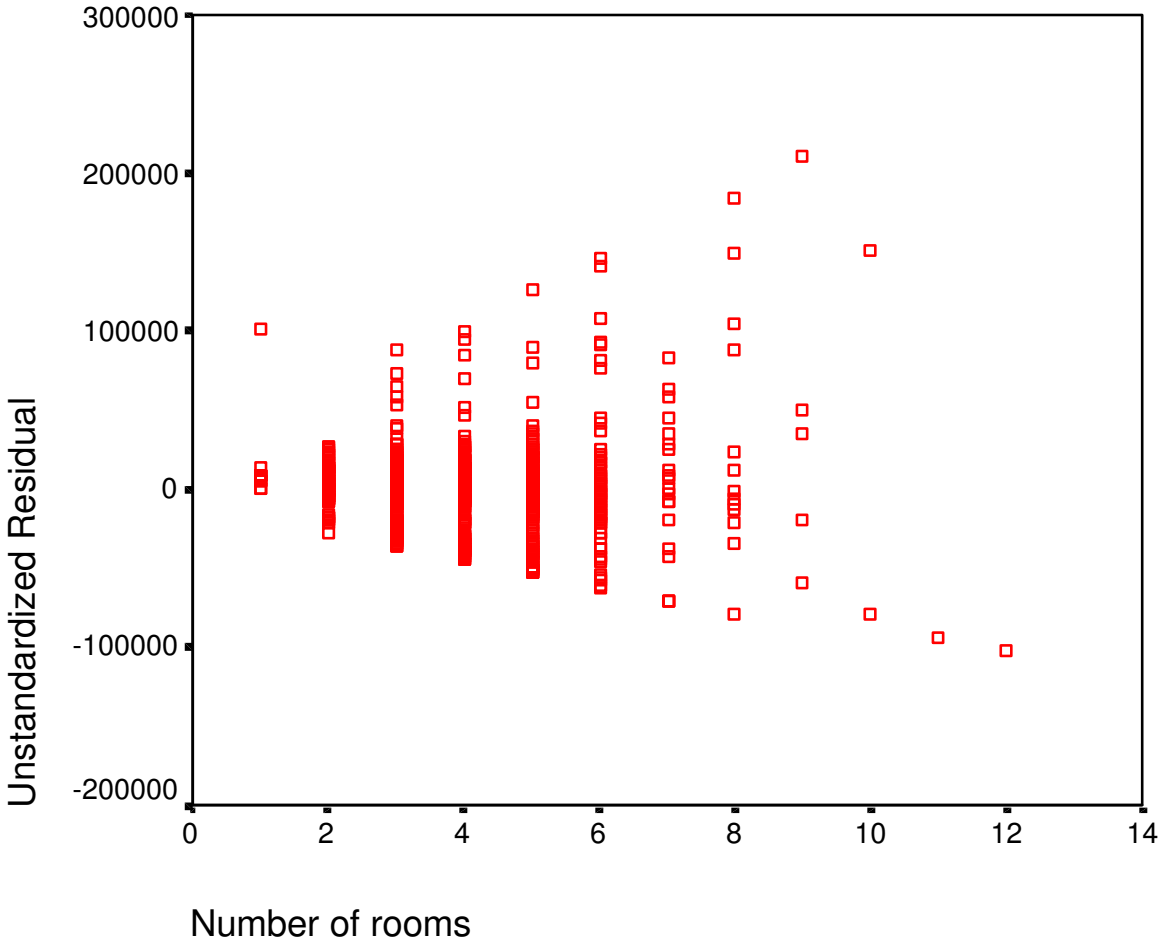
$$= \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

where σ^2 is a scalar

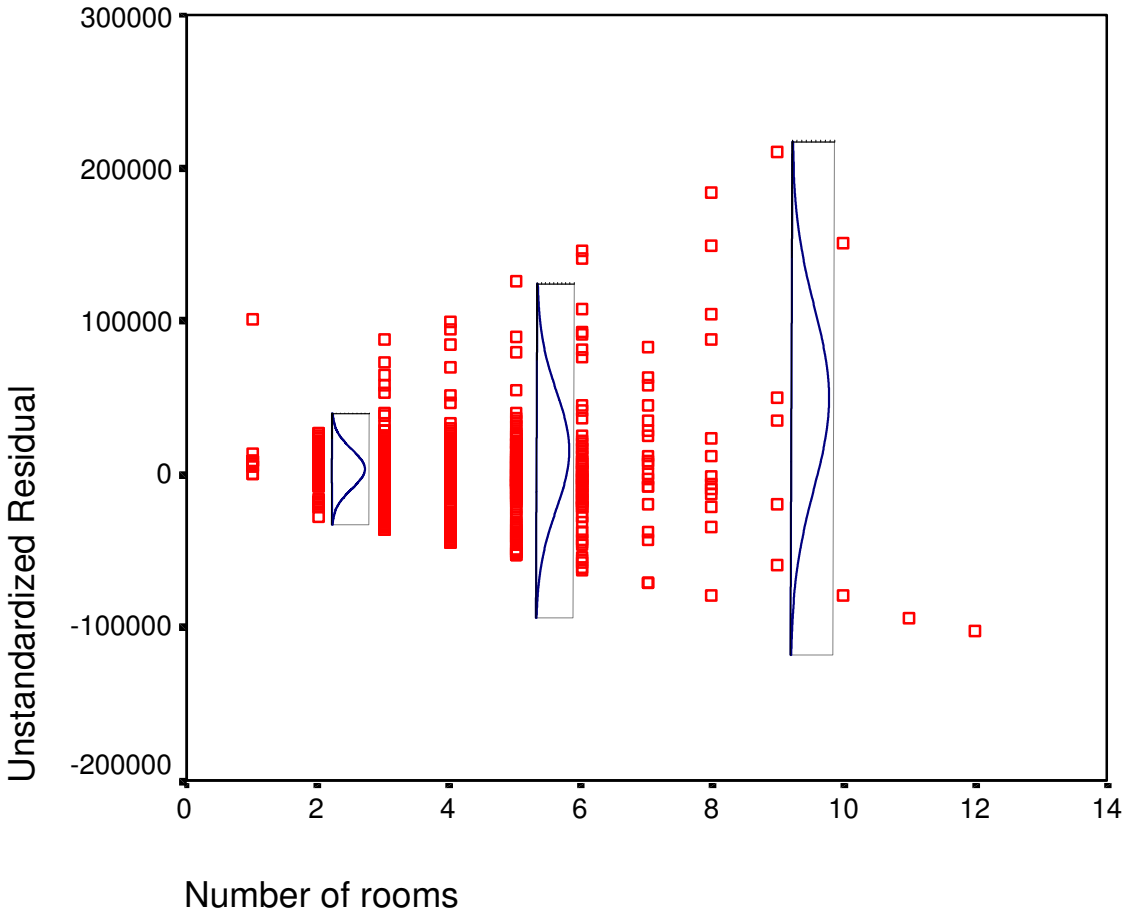
- Each one of the residuals has a sampling distribution, each of which should have the same variance -- "*homoscedasticity*"
- Clearly, this is not the case within in this sample, and so is unlikely to be true across samples:


400000
300000
200000
100000
0
-100000
Purchase price
0
2
4
6
8
10
12
14
Number of rooms

If we plot the residual against Rooms, we can see that its variance increases with No. rooms:



We can imagine the distributions of particular residuals as follows:



There is clear evidence of increasing variance here

This is confirmed when we look at the standard deviation of the residual for different parts of the sample

Group Statistics

	Number of rooms	N	Mean	Std. Deviation	Std. Error Mean
Unstandardized Residua	>= 3	669	-680.676	31647.60	1223.567
	< 3	96	4743.461	15024.51	1533.433

Group Statistics

	Number of rooms	N	Mean	Std. Deviation	Std. Error Mean
Unstandardized Residua	>= 4	452	-1575.28	36020.35	1694.255
	< 4	313	2274.843	18350.73	1037.245

Remember that these are only sample based sd's. I.e. they are only *a guide* to what the true sd's of the residuals would be like in the population.

2. Causes

- What might cause the variance of the residuals to change over the course of the sample?
  - the error term may be correlated with:
    - either the dependent variable and/or the explanatory variables in the model,
    - or some combination (linear or non-linear) of all variables in the model
    - or those that should be in the model.
  - But why?

(i) Non-constant coefficient

- Suppose that the slope coefficient varies across *i*:

$$y_i = a + b_i x_i + u_i$$

- suppose that it varies randomly around some fixed value β:

$$b_i = \beta + \varepsilon_i$$

- then the regression actually estimated by SPSS will be:

$$y_i = a + (\beta + \varepsilon_i) x_i + u_i$$

$$= a + \beta x_i + (\varepsilon_i x_i + u_i)$$

where $(\varepsilon_i x + u_i)$ **is the error term in the SPSS regression. The error term will thus vary with *x*.**

(ii) Omitted variables

- Suppose the "true" model of *y* is:

  $y_i = a + b_1 x_i + b_2 z_i + u_i$

- but the model we estimate fails to include *z*:

  $y_i = a + b_1 x_i + v_i$

- then the error term in the model estimated by SPSS (v_i) will be capturing the effect of the omitted variable, and so it will be correlated with *z*:

  $v_i = c z_i + u_i$

- **and so the variance of v_i will be non-scalar**

(iii) Non-linearities

- If the true relationship is non-linear:

$$y_i = a + b x_i^2 + u_i$$

- but the regression we attempt to estimate is linear:

$$y_i = a + b x_i + v_i$$

- then the residual in this estimated regression will capture the non-linearity and its variance will be affected accordingly:

$$v_i = f(x_i^2, u_i)$$

(iv) Aggregation (true heterosc.)

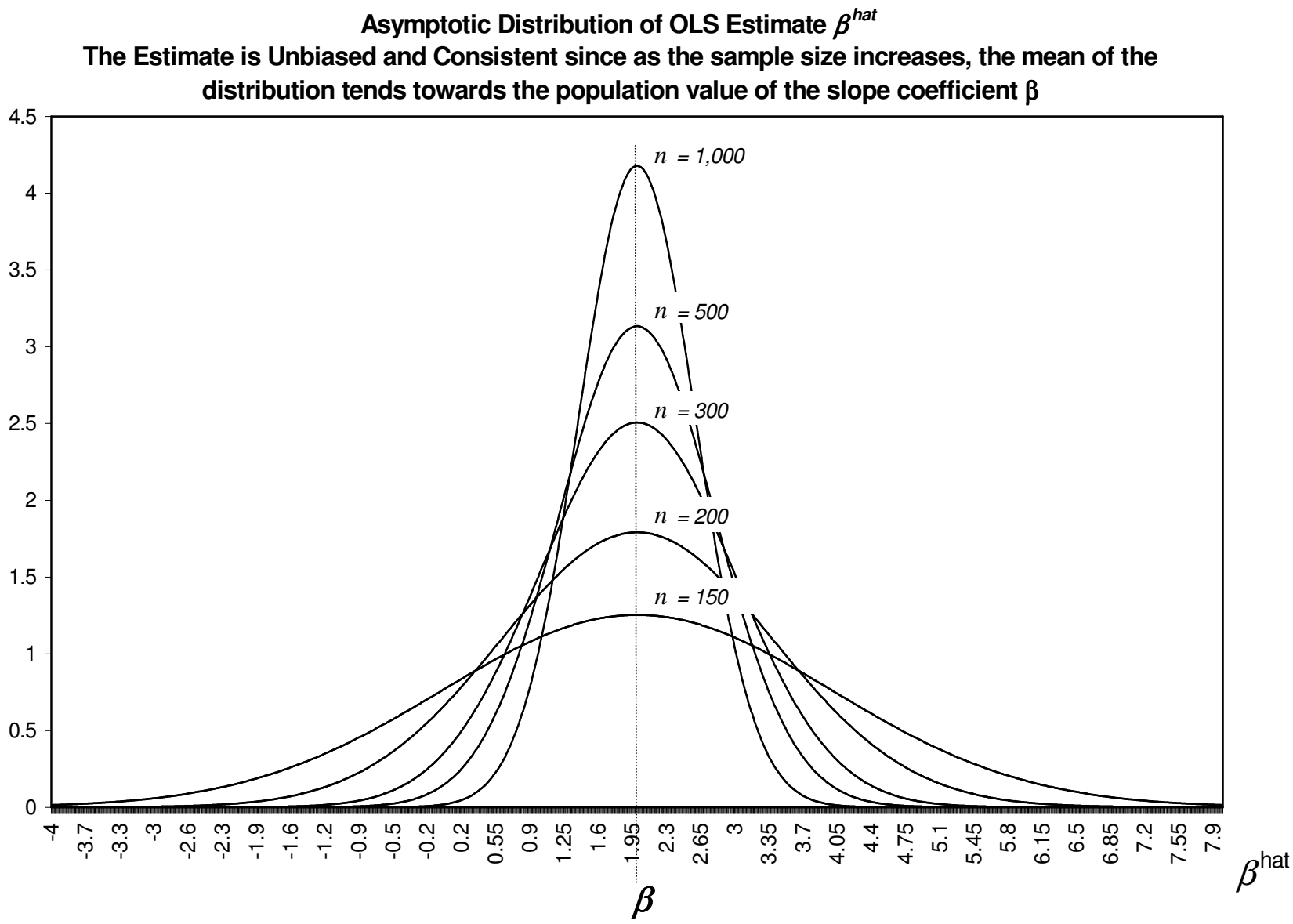
- Sometimes we aggregate our data across groups:
  - *e.g.* quarterly time series data on income = average income of a group of households in a given quarter
- if this is so, and the size of groups used to calculate the averages varies,
  - $\Rightarrow$ variation of the mean will vary
  - larger groups will have a smaller standard error of the mean.
  - $\Rightarrow$ the measurement errors of each value of our variable will be correlated with the sample size of the groups used.
- Since measurement errors will be captured by the regression residual
  - $\Rightarrow$ regression residual will vary the sample size of the underlying groups on which the data is based.

3. Consequences

- Heteroscedasticity by itself does not cause OLS estimators 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,

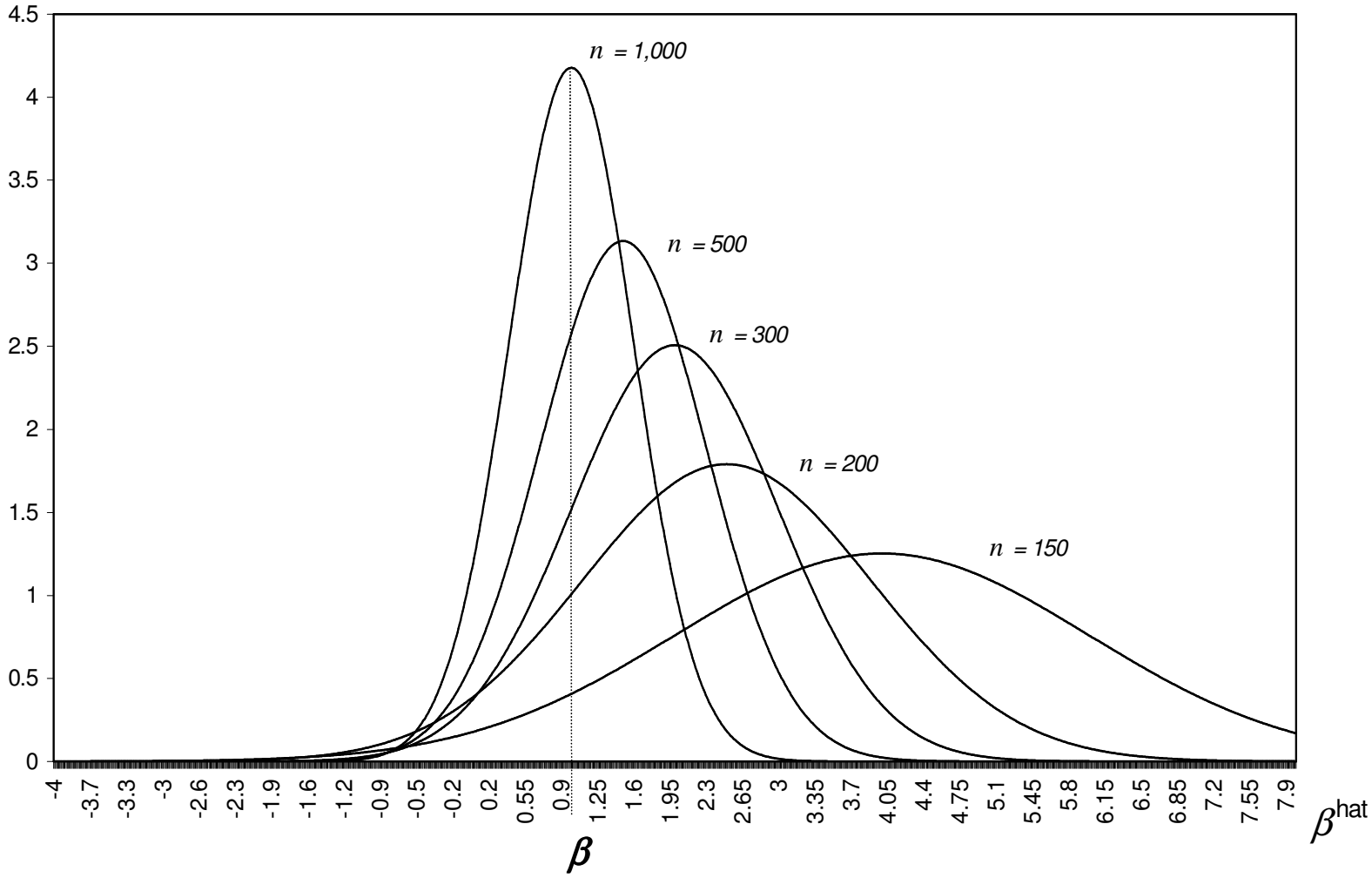
  $\Rightarrow$ OLS estimators will be biased and inconsistent.

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 heteroskedasticity that causes the bias,
- but failure of one of the other assumptions that happens to have hetero as the side effect.

  $\Rightarrow$ testing for hetero. is closely related to tests for misspecification generally.

  - Unfortunately, there is usually no straightforward way to identify the cause
- Heteroskedasticity does, however, bias 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)

4. Detection

- Q/ How can we tell whether our model suffers from heteroscedasticity?

4.1 Specific Tests/Methods

- **A. Visual Examination of Residuals**
- **B. Levene's Test**
- **C. Goldfeld-Quandt Test:**
  - S.M. Goldfeld and R.E. Quandt, "Some Tests for Homoscedasticity," *Journal of the American Statistical Society*, Vol.60, 1965.
  - H_0: σ_i^2 is not correlated with a variable z
  - H_1: σ_i^2 is correlated with a variable z

- G-Q test procedure is as follows:
  - (i) order the observations in ascending order of *z.*
  - (ii) omit *p* central observations (as a rough guide take $p \approx n/3$ where *n* is the total sample size).
    - This enables us to easily identify the differences in variances.
  - (iii) Fit the separate regression to both sets of observations.
    - The number of observations in each sample would be $(n - p)/2$, so we need $(n - p)/2 > k$ where *k* is the number of explanatory variables.
  - (iv) Calculate the test statistic *G* where:

$$G = \frac{RSS_2 / (\frac{1}{2}(n - p) - k)}{RSS_1 / (\frac{1}{2}(n - p) - k)}$$

    G has an F distribution: $G \sim F[\frac{1}{2}(n - p) - k, \frac{1}{2}(n - p) - k]$
    - NB G must be > 1. If not, invert it.
- Problem: In practice we don't usually know what *z* is.
  - But if there are various possible *z*'s then it may not matter which one you choose if they are all highly correlated which each other.

4.2 General Tests

- **A. Breusch-Pagan Test :**
  - T.S. Breusch and A.R. Pagan, "A Simple Test for Heteroscedasticity and Random Coefficient Variation," *Econometrica*, Vol. 47, 1979.
  - 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:

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

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.

B. White (1980) Test

- The most general test of heteroscedasticity
  - no specification of the form of hetero required
- (i) run an OLS regression - use the OLS regression to calculate $u^{hat\,2}$ (i.e. square of residual).
- (ii) use $u^{hat\,2}$ as the dependent variable in another regression, in which the regressors are:
  - (a) all "*k*" original independent variables, and
  - (b) the square of each independent variable, (excluding dummy variables), and all 2-way interactions (or crossproducts) between the independent variables.
    - The square of a dummy variable is excluded because it will be perfectly correlated with the dummy variable.
  - Call the total number of regressors (not including the constant term) in this second equation, *P*.

– (iii) From results of equation 2, calculate the test statistic

$$nR^2 \sim \chi^2_P$$

where n = sample size, and R^2 = unadjusted coefficient of determination.

- The statistic is asymptotically (I.e. in large samples) distributed as chi-squared with P degrees of freedom, where P is the number of regressors in the regression, not including the constant

Notes on White's test:

- The White test does not make any assumptions about the particular form of heteroskedasticity, and so is quite general in application.
  - It does not require that the error terms be normally distributed.
  - However, rejecting the null may be an indication of model specification error, as well as or instead of heteroskedasticity.
- generality is both a virtue and a shortcoming.
  - It might reveal heteroscedasticity, but it might also simply be rejected as a result of missing variables.
  - it is "nonconstructive" in the sense that its rejection does not provide any clear indication of how to proceed.
- NB: if you use White's standard errors, eradicating the heteroscedasticity is less important.

Problems:

- Note that although *t*-tests become reliable when you use White's standard errors, *F*-tests are still not reliable (so Chow's first test still not reliable).
- White's SEs have been found to be unreliable in small samples
  - but revised methods for small samples have been developed to allow robust SEs to be calculated for small *n*.

5. Solutions

- **A. Weighted Least Squares**
- **B. Maximum likelihood estimation.** (not covered)
- **C. White's Standard Errors**

A. Weighted Least Squares

- If the differences in variability of the error term can be predicted from another variable within the model, the Weight Estimation procedure (available in SPSS) can be used.
  - computes the coefficients of a linear regression model using WLS, such that the more precise observations (that is, those with less variability) are given greater weight in determining the regression coefficients.
- Problems:
  - Wrong choice of weights can produce biased estimates of the standard errors.
    - we can never know for sure whether we have chosen the correct weights, this is a real problem.
  - If the weights are correlated with the disturbance term, then the WLS slope estimates will be inconsistent.
  - Also: Dickens (1990) found that errors in grouped data may be correlated within groups so that weighting by the square root of the group size may be inappropriate. See Binkley (1992) for an assessment of tests of grouped heteroscedasticity.

C. Whites 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.

Summary

- (1) Causes
- (2) Consequences
- (3) Detection
- (4) Solutions

Reading:

- Kennedy (1998) “A Guide to Econometrics”, Chapters 5,6,7 and 9
- Maddala, G.S. (1992) “Introduction to Econometrics” chapter 12
- Field, A. (2000) chapter 4, particularly pages 141-162.
- Green, W. H. (1990) Econometric Analysis
- **Grouped Heteroscedasticity:**
  - Binkley, J.K. (1992) “Finite Sample Behaviour of Tests for Grouped Heteroskedasticity”, *Review of Economics and Statistics*, 74, 563-8.
  - Dickens, W.T. (1990) “Error components in grouped data: is it ever worth weighting?”, *Review of Economics and Statistics*, 72, 328-33.
- **Breusch Pagan critique:**
  - Koenker, R. (1981) “A Note on Studentizing a Test for Heteroskedascity”, *Journal of Applied Econometrics*, 3, 139-43.