Ch 5: The extra sum of squares principle and regression modelling assumptions

Outline

1. Extra sum of squares principle
2. F-tests and mathematical formulae
3. Examples
4. Regression modelling assumptions
5. Graphical detection of heteroscedasticity
6. Numerical testing of heteroscedasticity

1.1 Extra sum of squares principle

- In the last lecture we saw that R automatically produces an F-statistic to test the overall level of fit
- In formal terms we are comparing the models

Model 0

$$Y_i = \beta_1 + u_i$$

Model 1

$$Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \ldots + \beta_p X_{p,i} + u_i$$

- **The basic idea can be greatly expanded upon**

1.2 In summary

- t-**test**
  - Test the significance of individual parameters
- F-**test**
  - Test the joint significance of multiple parameters
  - Evaluate competing models that are **nested**
- **E.g. the overall F-test in R compares the models Model 0**

$$Y_i = \beta_1 + u_i$$

Model 1

$$Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \ldots + \beta_p X_{p,i} + u_i$$

1.3 F-test: Extra sum of squares principle

- **Nested models**

Model 0

$$Y_i = \beta_1 + \beta_2 X_{2,i} + \ldots + \beta_{p-m} X_{p-m,i} + u_i$$

Model 1

$$Y_i = \beta_1 + \beta_2 X_{2,i} + \ldots + \beta_{p-m} X_{p-m,i} + \ldots + \beta_p X_{p,i} + u_i$$

- **The overall aim is to test whether the extra variation in the data that Model 1 explains – the Extra Sum of Squares – is statistically significant**

1.4 Extra sum of squares principle: generality of approach

• Using the extra sum of squares principle we saw that on the previous slide comparing Model 0 and Model 1 was equivalent to testing the m **linear** restrictions

$$\beta_{p-m+1} = \beta_{p-m+2} = \ldots = \beta_p = 0. \quad (1)$$

• In terms of formal mathematics (Bingham and Fry, 2010 Ch. 6) it is possible to show that you can use the same approach to test more general **linear constraints**

- There is potentially some difficult mathematics involving Lagrange multipliers but the underlying ideas remain relatively simple

• Equation (1) gives the simplest linear restriction but other hypotheses may be possible e.g.

$$\begin{aligned} \text{Model 0} &: Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_2 X_{3,i} + u_i \\ \text{Model 1} &: Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + u_i \end{aligned}$$

• **But the same basic testing procedure applies in each case!**

2.1 Setting up the F-test

- **Nested models**
- Model 0 has m constraints or equivalently m fewer parameters to estimate
- Model 1 is larger and has no parameter constraints

1. Easy case

$$\begin{aligned} \text{Model 0} &: Y_i = \beta_1 + \beta_2 X_{2,i} + \ldots + \beta_{p-m} X_{p-m,i} + u_i \\ \text{Model 1} &: Y_i = \beta_1 + \beta_2 X_{2,i} + .. + \beta_{p-m} X_{p-m,i} + .. + \beta_p X_{p,i} + u_i \end{aligned}$$

2. More complicated e.g.

$$\begin{aligned} \text{Model 0} &: Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_2 X_{3,i} + u_i \\ \text{Model 1} &: Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + u_i \end{aligned}$$

   - m=1 constraint

2.2 F-statistic always has the same form

- **Is the extra sum of squares or variation in the data explained by the larger (unconstrained) model statistically significant?**
- How I remember the F-statistic

$$F = \frac{\frac{\text{Difference in residual SS}}{\text{Difference in residual df}}}{\frac{\text{Residual SS (larger model)}}{\text{Residual df (larger model)}}}$$

$$F = \frac{\frac{\text{Difference in residual SS}}{m}}{\frac{\text{Residual SS (larger model)}}{n-p}} \tag{2}$$

2.3 Developing more applicable formulae

- As we saw in the last lecture the F-statistic can be written in terms of the R^2 statistic.
- Tells you exactly the same information as before...
- But results in formulae that are simpler and easier to apply
- F-test for the extra sum of squares principle becomes

$$F = \frac{\frac{\text{Difference in residual SS}}{m}}{\frac{\text{Residual SS (larger model)}}{n-p}}$$

$$F = \frac{\frac{TSS(R_1^2 - R_0^2)}{m}}{\frac{TSS(1-R_1^2)}{n-p}} = \frac{\frac{R_1^2 - R_0^2}{m}}{\frac{1-R_1^2}{n-p}} \sim F_{m,n-p} \qquad (3)$$

2.4 Equivalent applicable formulae

- Equivalently equation (3) on Slide 2.3 can be re-written as

$$F = \frac{\frac{\text{Difference in } R^2}{m}}{\frac{1-R^2(\text{larger model})}{n-p}} \tag{4}$$

- Equations (3) and (4) are equivalent – it is up to you which formula you remember and use in your exam...
- Will give some practical examples but first I wanted to show where the F-test automatically reported by R comes from

2.5 F-Rtest for overall significance revisited

Model 0

$$Y_i = \beta_1 + u_i$$

Model 1

$$Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \ldots + \beta_p X_{p,i} + u_i$$

$$F = \frac{\frac{\text{Difference in residual SS}}{m}}{\frac{\text{Residual SS (larger model)}}{n-p}} = \frac{\frac{TSS(R^2-0)}{p-1}}{\frac{TSS(1-R^2)}{n-p}} = \frac{\frac{R^2}{p-1}}{\frac{1-R^2}{n-p}} \sim F_{p-1,n-p}$$

- **Gives the same formula as given in the previous lecture but shows that this fits into a much more general way of thinking...**

3.1 Illustrative examples

• We will illustrate the F-test and the extra sum of squares principle with two examples:

1. F-test for joint significance
2. F-test in polynomial regression

3.2 Multiple linear regression example

• To show you how to interpret the results from a multiple linear regression model use an example from the classical Longley dataset
• Overall aim is to explain the number of employed people in the US in terms of

1. X_2, GNP
2. X_3 the number of unemployed
3. X_4 the unemployment rate
4. X_5 the "non-institutionalised" population over the age of 14
5. X_6 the yearly trend

• **Based on the results of the last lecture want to see if BOTH X_2 and X_5 can be excluded from the model**

3.3 R commands for reading in the data

- Data in the file `longley.txt`

```
longley<-read.table(‘‘E:longley.txt")
x2<-longley[,1]
x3<-longley[,2]
x4<-longley[,3]
x5<-longley[,4]
x6<-longley[,5]
y<-longley[,6]
```

3.4 R code

• Once the data has been entered in fit the full unconstrained model using

`a.lm<-lm(y~x2+x3+x4+x5+x6)`

• Then fit the constrained model simply by not including x_2 and x_5 in the above and calling it something else

`b.lm<-lm(y~x3+x4+x6)`

• The F-test using the extra sum of squares principle can then be performed using the command `anova`:

```
anova(a.lm, b.lm, test=‘‘F")
```

3.5 Results in R

```
anova(a.lm, b.lm, test=‘‘F")
Model 1:  y ~ x2 + x3 + x4 + x5 + x6
Model 2:  y ~ x3 + x4 + x6
Res.Df RSS Df Sum of Sq F Pr(>F)
1 10 0.83935
2 12 1.32336 -2 -0.48401 2.8833 0.1026
```

- Since the result is non-significant $p = 0.1026 > 0.05$ this gives statistical evidence that X_2 and X_5 can be excluded from the model

3.6 Reconstructing the F-test by hand

- Using `summary(a.lm)` and `summary(b.lm)` tells you that in each case the R^2 values are 0.9955 and 0.9928
- Similarly the residual degrees of freedom are 10 and 12 respectively
- The F-statistic can hence be re-constructed as

$$F = \frac{\frac{\text{Difference in } R^2}{m}}{\frac{1-R^2(\text{larger model})}{n-p}} = \frac{\frac{0.9955-0.9928}{2}}{\frac{1-0.9955}{10}}$$

$$F = \frac{0.00135}{0.00045} = 3$$

- Note that in this case heavy rounding errors mean this deviates substantially from the "exact" value of 2.8833 calculated by R above. However, this is, in principle, how the F-test can be constructed from first principles.

3.7 Example 2: Divorces

• Data from Daily Mirror gives the percentage of divorces caused by adultery as a function per year of marriage
• Original analysis claimed divorce-risk peaks at year 2 then decreases thereafter. But is this conclusion misleading?
• Healthy scepticism the most important skill in the era of Big Data?
• Wanted to test originally whether a quadratic model offers a better fit than a straight line model to this data
• This example is also interesting as it shows that a nonlinear model in X can still be treated as a linear regression model because it remains linear in the regression coefficients β

3.8 Dataset: percent of divorces caused by adultery by year of marriage

Year	1	2	3	4	5	6	7
%	3.51	9.50	8.91	9.35	8.18	6.43	5.31
Year	8	9	10	15	20	25	30
%	5.07	3.65	3.80	2.83	1.51	1.27	0.49

Table: Data on divorces caused by adultery.

3.9 Analysis in R

- **Enter the data into R**

```
year<-c(1, 2, 3, 4, 5, 6,7, 8, 9, 10, 15, 20, 25, 30)
percent<-c(3.51, 9.5, 8.91, 9.35, 8.18, 6.43, 5.31,
5.07, 3.65, 3.8, 2.83, 1.51, 1.27, 0.49)
```

- **Introduce a squared term for year**

```
yearsq<-year^2
```

- **Equivalent R commands**

`.lm` for linear model

`.glm` for generalised linear model

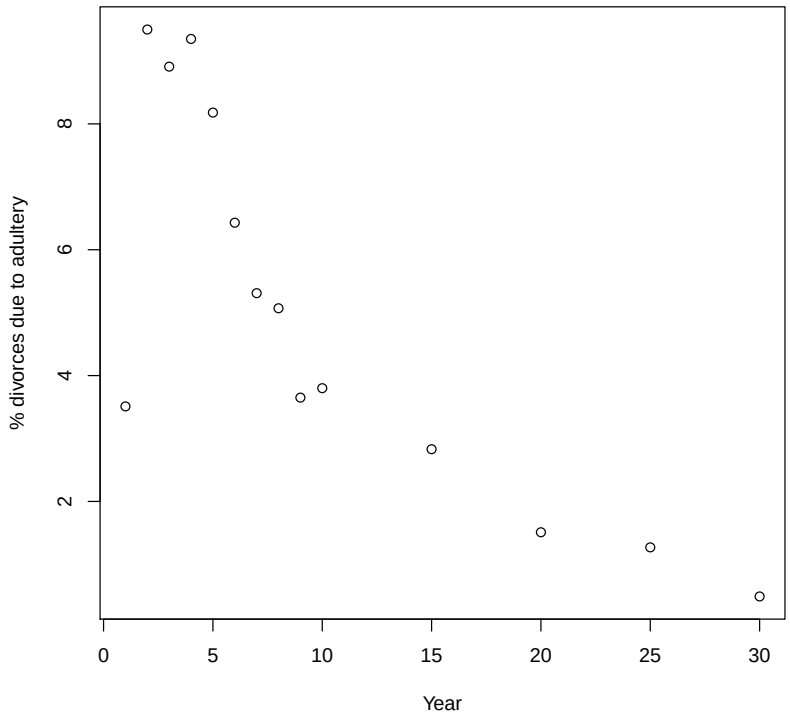
- In R a generalised linear model with Normal errors is the default.
- There is some suggestion that the regression command is better numerically and would be the best to use in practical problems

3.10 Plotting the data

- Plotting the data in R

```
plot(year, percent, xlab=‘‘Year", ylab=‘‘% divorces due to adultery")
```

- Suggestion that divorce rate unusually low in the first year then decreases steadily over time?



3.11 Analysis in R

- **Fit a linear regression model**

```
a.lm<-lm(percent~year)
a.glm<-glm(percent~year)
```

- **Fit a quadratic regression model**

```
b.lm<-lm(percent~year+yearsq)
b.glm<-glm(percent~year+yearsq)
summary(b.lm)
```

3.12 Regression output

```
Coefficients:
Estimate Std.  Error t value Pr(>|t|)
(Intercept) 8.751048 1.258038 6.956 2.4e-05 ***
year -0.482252 0.235701 -2.046 0.0654 .
yearsq 0.006794 0.007663 0.887 0.3943
```

- **No evidence ($p = 0.3943 > 0.05$) that the quadratic term is needed in the model. Suggestion is that the original analysis in the Daily Mirror is probably mistaken.**

3.13 Analysis in R

- **Test for model improvement using an F-test. Should get the same answer as the t-test on the previous slide.**

```
anova(a.lm, b.lm, test="F")
anova(a.glm, b.glm, test="F")
Model 1:  percent ~ year
Model 2:  percent ~ year + yearsq
Resid.  Df Resid.  Dev Df Deviance F Pr(>F)
1 12 42.375
2 11 39.549 1 2.826 0.786 0.3943
```

- **Analysis suggests the quadratic term is not needed in the model**

3.14 Final analysis and conclusions

```
summary(a.glm)
Estimate Std.  Error t value Pr(>|t|)
(Intercept) 7.88575 0.78667 10.024 3.49e-07 ***
year -0.27993 0.05846 -4.788 0.000442 ***
```

- The coefficient of year is negative and statistically significant $p = 0.000442 < 0.05$. As the number of years marriage increases the percentage of divorces caused by adultery decreases.
- Some suggestion that Daily Mirror analysis misleading? Safer to say that the divorce rate caused by adultery generally decreases over time but is unusually low in the first year?

4.1 Regression modelling assumptions (Gujarati and Porter, Ch. 7) – worth remembering!

- **The classical linear regression model is**

$$Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \ldots + \beta_p X_{p,i} + u_i$$

Assumptions

1. The model is linear in the parameters β
2. The X variables are independent of the error term
3. The disturbance has zero mean: $E[u_i] = 0$
4. Homoscedasticity or constant residual variance: $\text{Var}(u_i) = \sigma^2$
5. The u_i are normally distributed
6. The disturbances are uncorrelated

$$\text{Cor}(u_i, u_j) = 0 \quad (i \neq j)$$

7. The number of observations is greater than the number of parameters to be estimated: $n > p$

4.2 Regression modelling assumptions: continued

8. There is no exact linear relationship between any pair of X variables

9. No specification bias – the model is correctly specified

Note

• The above list is suggestive of wider problems with econometric textbooks in that they are sometimes not very clear about the modelling assumptions made and tend to emphasise applications

- May lead to significant problems when undertaking more advanced research outside of the scope of your course

4.3 Responses to failures in regression modelling assumptions

1. Should be easy to see from the specification of the model
2. May be violated in some time series problems – largely outside the scope of the course
3. The fitted residuals will automatically have a zero mean
4. Today's lecture – see below
5. May be remedied to some extent by trying to model log y instead of y
6. Chapter 6
7. $n < p$ occurs in certain specialised bioinformatics problems but this is outside the scope of your course
8. Chapter 7
9. Outside the scope of the course – but may be solved to some extent if the models used are derived from an underlying theory or literature review

4.4 Heteroscedasticity

- In this lecture we will discuss heteroscedasticity
- **The classical linear regression model is**

$$Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \ldots + \beta_p X_{p,i} + u_i$$

- In particular, the classical linear regression model assumes that for each observation i:

$$\text{Var}(u_i) = \sigma^2 \tag{5}$$

- Equation (5) defines homoscedasticity
- If equation (5) does not apply then we have heteroscedasticity

4.5 Overview

• The basic techniques of applied statistics are largely mechanical in nature
  - R^2
  - F-statistic
  - t-tests
• These alone may be enough to pass an exam – particularly if your lecturer is benevolent/not very cunning...
• **It is important to realise that it is often more difficult to use statistical methods well in applied project work. Statistics requires critical thinking as well as mechanical calculations**
  - Heteroscedasticity is one very commonly encountered problem

4.6 Consequences of heteroscedasticity

- It is important to realise that the underlying mathematics of the classical linear regression model assumes that we have heteroscedasticity as shown in equation (5) on Slide 4.4
- **If this assumption does not hold the maths does not work!**
- If we heteroscedasticity and not homoscedasticity then
  - confidence intervals
  - hypothesis tests
  - F-tests etc all break down
- Parameter estimates obtained by OLS remain unbiased but are no longer optimal

4.7 Detecting heteroscedasticity

- There are formal tests
  - Goldfeldt-Quant test
  - White's test for heteroscedasticity
- **Often more important to just perform graphical checks – simpler and more robust**
- Can show mathematically (Bingham and Fry, Ch. 3) that if a model is correctly specified then the residuals and the fitted values should be independent.
- **Judge heteroscedasticity by plotting residuals against fitted values or squared residuals against fitted values**
  - Squared residuals may give a better indication about how the residual variance changes

5.1 Analysis of regression residuals in R

- In the previous lecture example we fitted the following regression model

```
a.lm<-lm(realgoldprice~realstockprice)
summary(a.lm)
```

- Having run these commands we can then use the named regression model to obtain a residual series as follows

```
residuals<-a.lm$resid
```

- We can then use e.g. `hist(residuals)` or `ts.plot(residuals)` to obtain e.g. a histogram or a time series plot of the residuals as appropriate
- The fitted values from a regression model can similarly be obtained using

```
fitted<-a.lm$fitted
```

5.1 Graphical tests of residuals

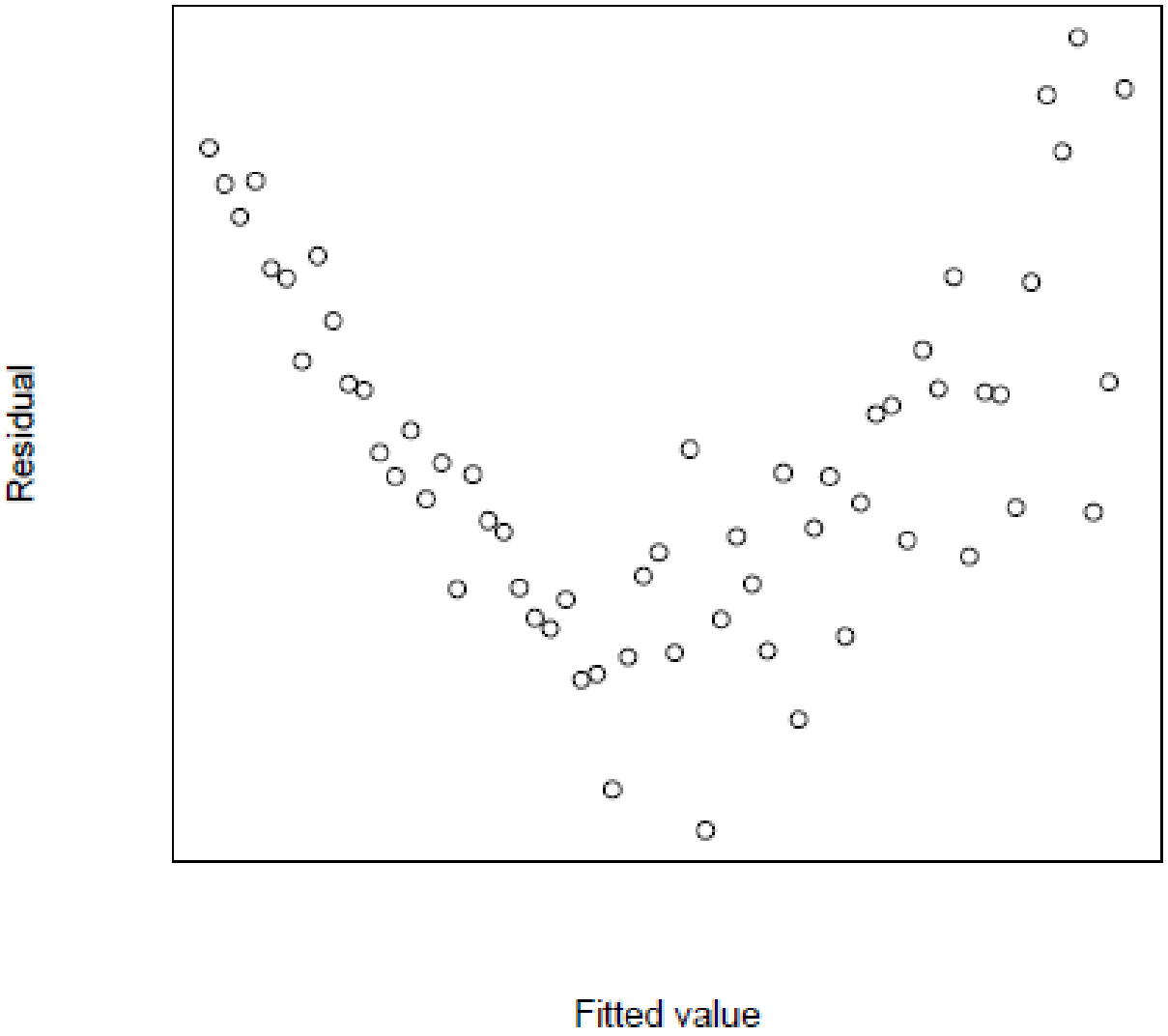
• Graphical tests of residuals are less precise than some of the mathematical techniques we have previously discussed in lectures
• There are more interesting things in life and in mathematics and statistics but graphical tests of residuals are deceptively important in applied project work
• **Because the area is so widely studied there are a number of commonly encountered patterns that it is often important to look out for...**
• These graphs are not necessarily very easy to interpret (this is discussed at length in the excellent book on applied regression by Draper and Smith). However, these graphs are easy to produce using modern software and so are perhaps over-discussed.

5.2 Interpretation of residual graphs

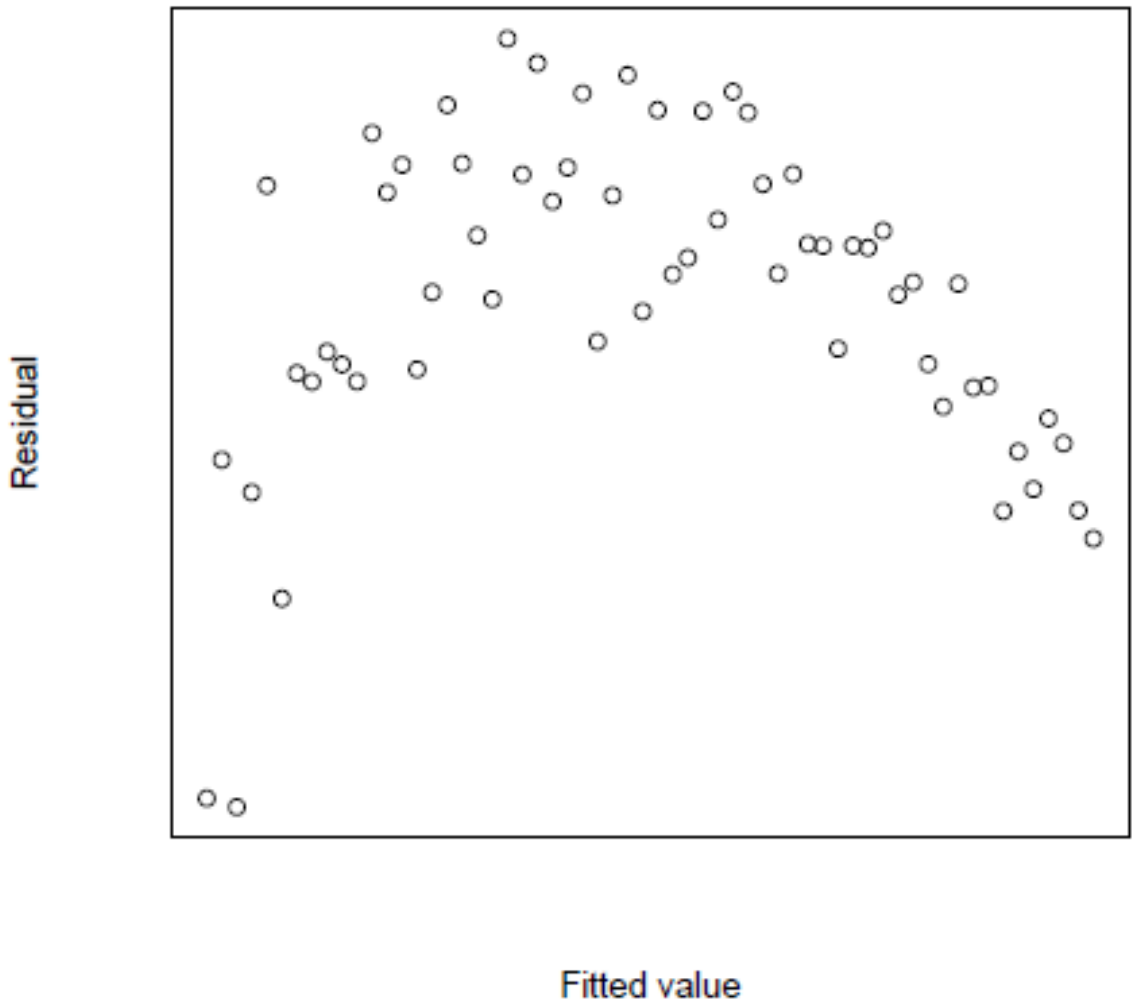
1. No systematic pattern suggesting no heteroscedasticity – this is the ideal scenario that we want to happen as this corresponds to the constant variance or homoscedasticity assumption of the classical linear model
2. Funnelling out of residuals
3. Funnelling in of residuals
4. Linear – error variance proportional to $\hat{Y}_i$
5. Quadratic – error variance proportional to $\hat{Y}_i^2$
6. Quadratic – error variance proportional to $\hat{Y}_i^2$

• **In these special cases we can transform the data to avoid the problem of heteroscedasticity**

5.3 Approximate funnelling out of residuals



5.4 Approximate funnelling in of residuals



5.5 Problems with residuals – remedial transformations

2. Fit a model for log y or $\sqrt{y}$

- In general terms fitting a model for log y may often help reduce problems with heteroscedasticity

3. Fit a model for y^2

4. Fit a model for $\sqrt{y}$

5. Fit a model for log y

6. Fit a model for log y

5.6 Problems with residuals – remedial transformations

• **It is also possible that heteroscedasticity may also be associated with some of the X-variables**

• (Plot the residuals or squared residuals against X instead of the fitted values in the above)

• Gujarati and Porter discuss two cases

1. The error variance is proportional to X_i^2

$$E[u_i^2] \approx \sigma^2 X_i^2$$

2. The error variance is proportional to X_i

$$E[u_i^2] \approx \sigma^2 X_i$$

• **In each case divide through by the square root of the offending X-term**

5.7 Error variance proportional to X_i^2

- Start with the model

$$Y_i = \beta_1 + \beta_2 X_{2,i} + u_i$$

- Divide through by X_i

$$\frac{Y_i}{X_i} = \frac{\beta_1}{X_i} + \beta_2 + \frac{u_i}{X_i} \tag{6}$$

- **Estimate equation (6) by the usual ordinary least squares regression approach**

5.8 Error variance proportional to X_i

- Start with the model

$$Y_i = \beta_1 + \beta_2 X_{2,i} + u_i$$

- Divide through by $\sqrt{X_i}$

$$\frac{Y_i}{\sqrt{X_i}} = \frac{\beta_1}{\sqrt{X_i}} + \beta_2 \sqrt{X_i} + \frac{u_i}{\sqrt{X_i}} \tag{7}$$

- **Estimate equation (7) by the usual ordinary least squares regression approach**

5.9 Weighted Least Squares (non-examinable)

• In rare circumstances we may encounter heteroscedasticity and **know** the exact form of the heteroscedasticity that occurs

• **The only example of this I have seen is using a weighted least squares approach to estimating a Generalised Linear Model**

- Since Generalised Linear Models can now be fitted using specialist modern software there is no need to use an approximate approach based on Weighted Least Squares

- So I think Weighted Least Squares has limited importance for this course

6.1 Numerical statistical tests of heteroscedasticity

• Whilst these exist these are usually thought to constitute rather separate parts of econometrics rather than statistics and are not covered further here in much detail.

1. Goldfeld-Quandt test (non-examinable)
2. White's General Heteroscedasticity test (non-examinable)

6.2 Goldfeld-Quandt test

- As we saw above often the heteroscedastic variance σ_i^2 increases as one of the X-variables increases
- One commonly-encountered case that is useful in applications is

$$\text{Var}(u_i) = \sigma^2 X_i^2 \tag{8}$$

- Equation (8) is hypothesis tested by the Goldfeld-Quandt test:

$$\begin{aligned} H_0 &: \text{Var}(u_i) = \sigma^2 \\ H_1 &: \text{Var}(u_i) = \sigma^2 X_i^2 \end{aligned}$$

6.3 Goldfeld-Quandt test (Gujarati and Porter, Ch. 11)

1. Order or rank the observations according to the values of the X_i from smallest to largest
2. Omit the central c observations and divide the sample into two groups of size $\frac{n-c}{2}$
3. Fit separate OLS regressions to each segment
4. Form the statistic

$$\lambda = \frac{\frac{RSS_2}{df}}{\frac{RSS_1}{df}} = \frac{RSS_2}{RSS_1}$$

- RSS_1=Residual SS for the first segment, RSS_2=Residual SS for the second segment
- Under the null hypothesis of homoscedasticity

$$\lambda = \frac{RSS_2}{RSS_1} \sim F_{\frac{n-c-2p}{2}, \frac{n-c-2p}{2}}$$

- **Pragmatic empirical experience suggests that for $n = 30$ take $c = 4$ and if $n = 60$ take $c = 10$**