Week 6: Assumptions in Regression Analysis

The Assumptions

1. The distribution of residuals is normal (at each value of the dependent variable).
2. The variance of the residuals for every set of values for the independent variable is equal.
    - violation is called heteroscedasticity.
3. The error term is additive
    - no interactions.
4. At every value of the dependent variable the expected (mean) value of the residuals is zero
    - No non-linear relationships

5. The expected correlation between residuals, for any two cases, is 0.
    - The independence assumption (lack of autocorrelation)
6. All independent variables are uncorrelated with the error term.
7. No independent variables are a perfect linear function of other independent variables (no perfect multicollinearity)
8. The mean of the error term is zero.

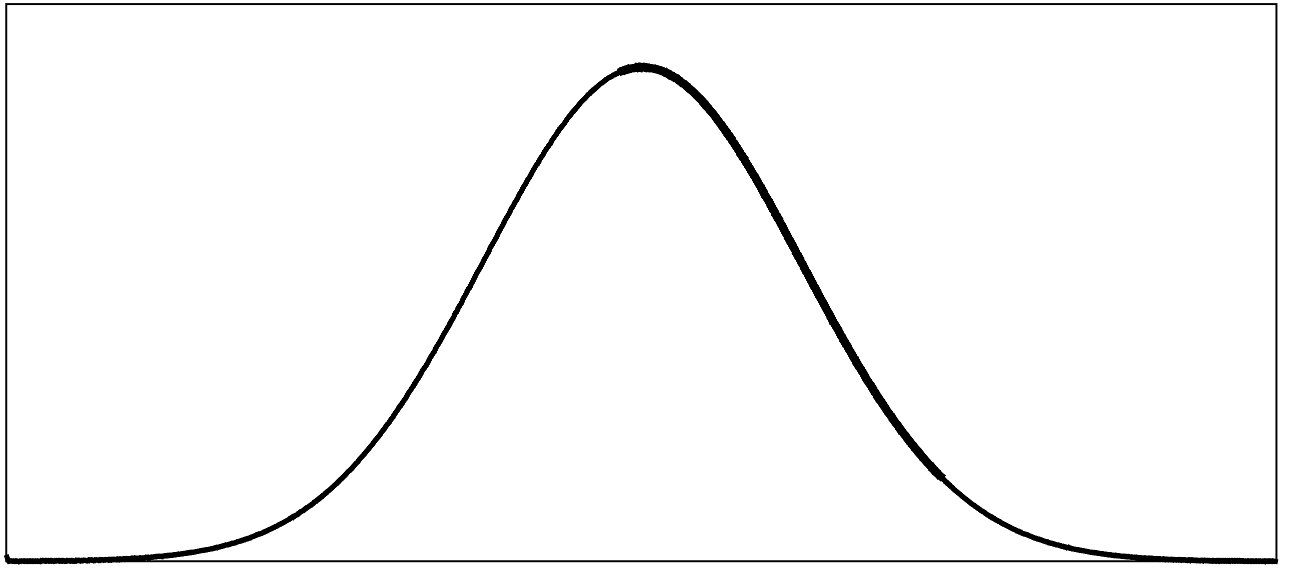
What are we going to do …

- Deal with some of these assumptions in some detail
- Deal with others in passing only

Assumption 1: The Distribution of Residuals is Normal at Every Value of the Dependent Variable

Look at Normal Distributions

- A normal distribution
  - symmetrical, bell-shaped (so they say)



What can go wrong?

- Skew
  - non-symmetricality
  - one tail longer than the other
- Kurtosis
  - too flat or too peaked
  - kurtosed
- Outliers
  - Individual cases which are far from the distribution

Effects on the Mean

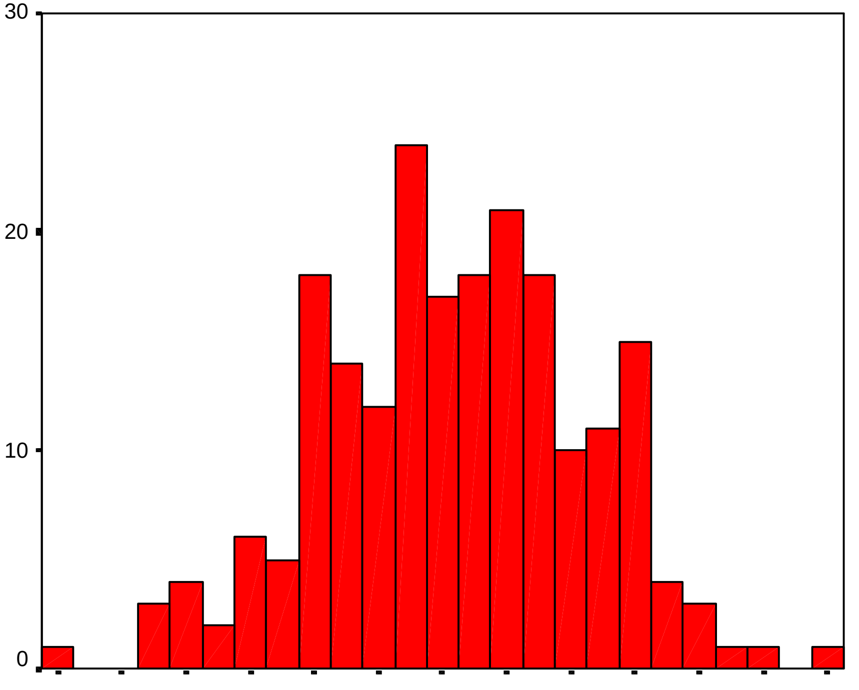
- Skew
  - biases the mean, in direction of skew
- Kurtosis
  - mean not biased
  - standard deviation is
  - and hence standard errors, and significance tests

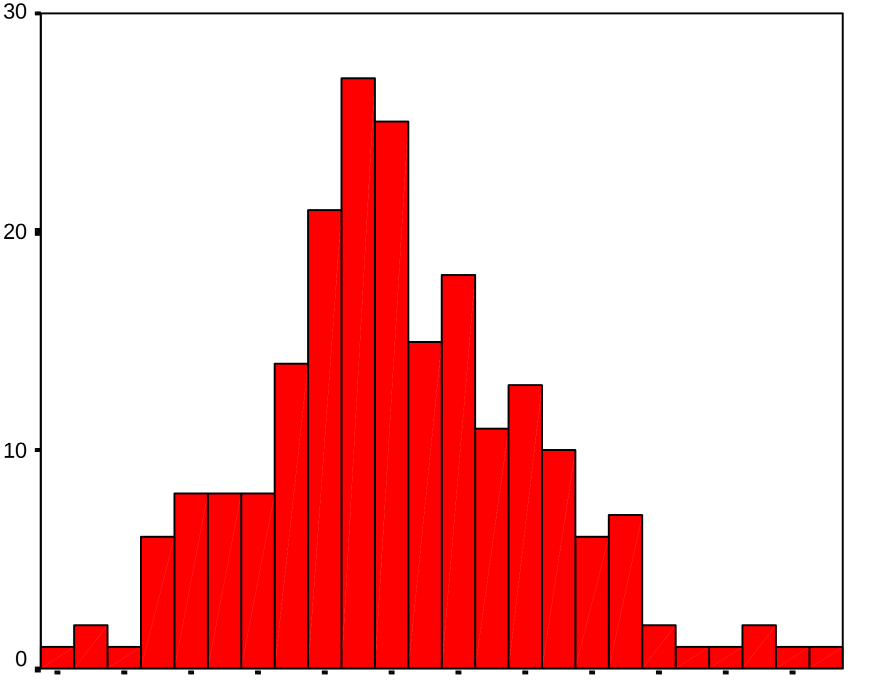
Examining Univariate Distributions

- Histograms
- Boxplots
- P-P Plots

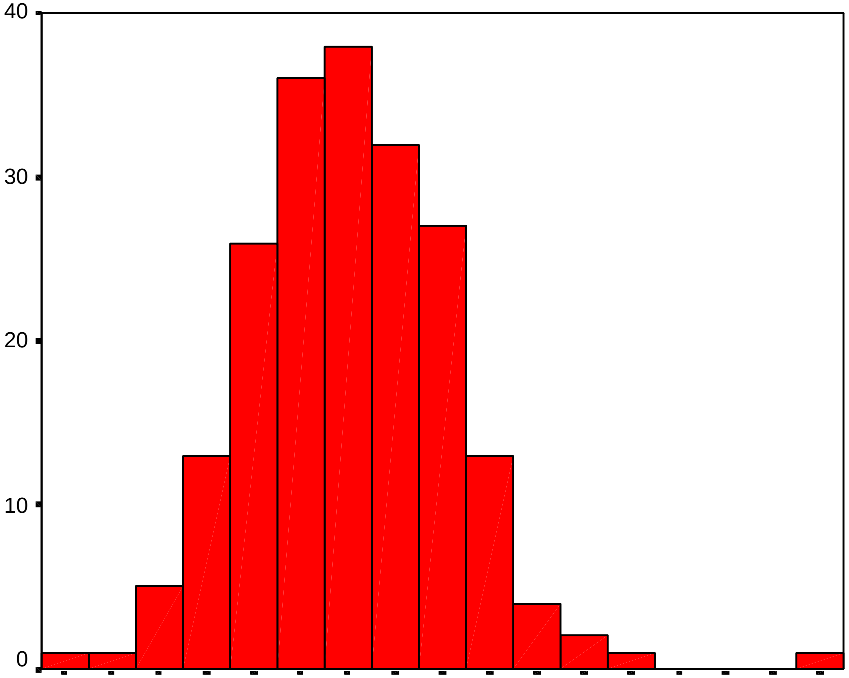
Histograms

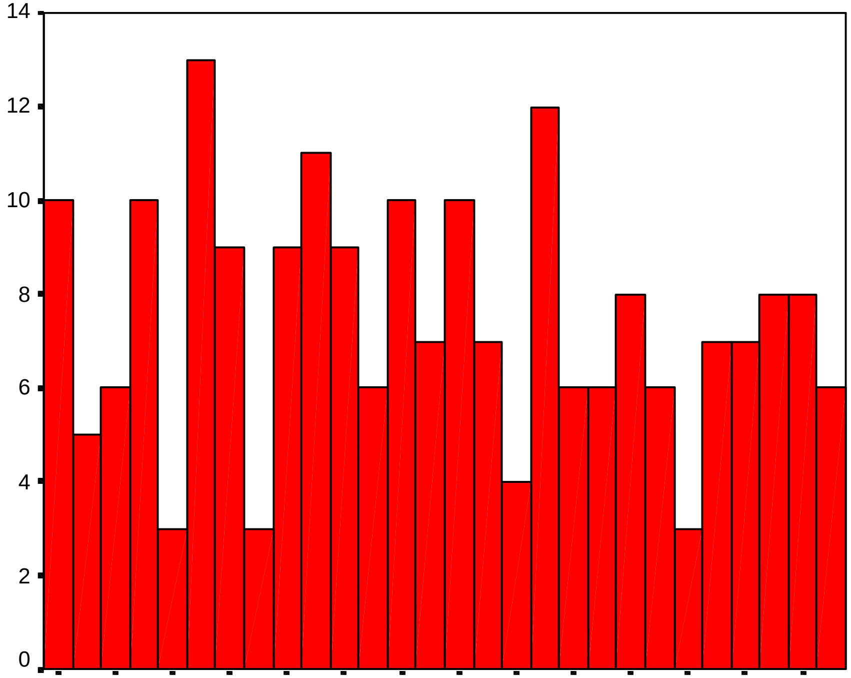
A and B



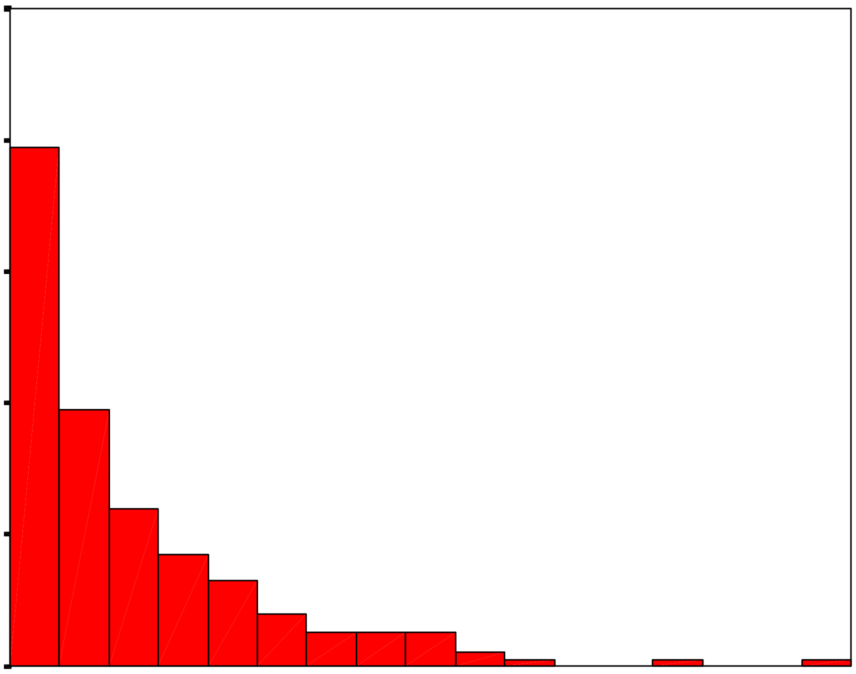


- C and D

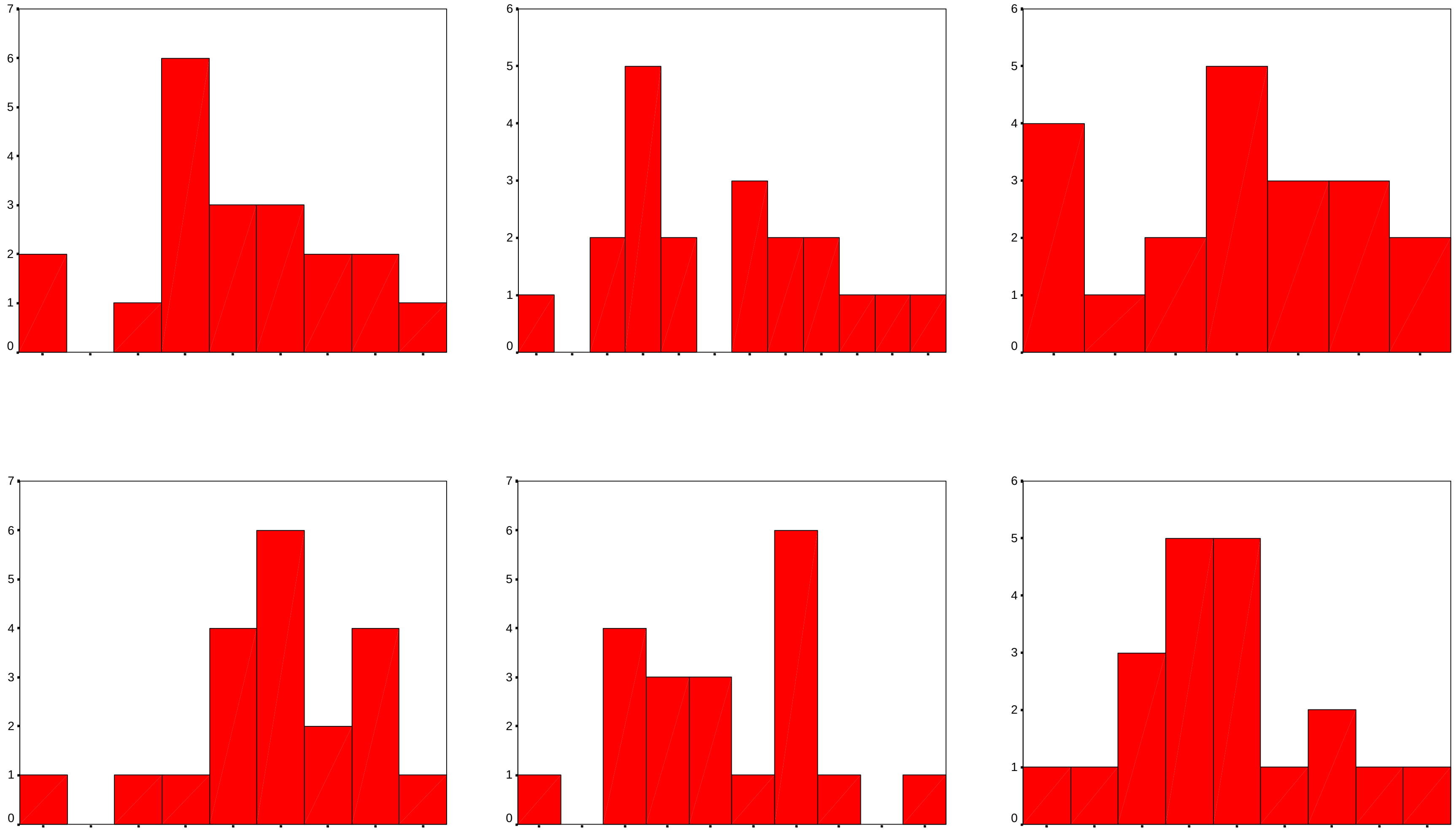




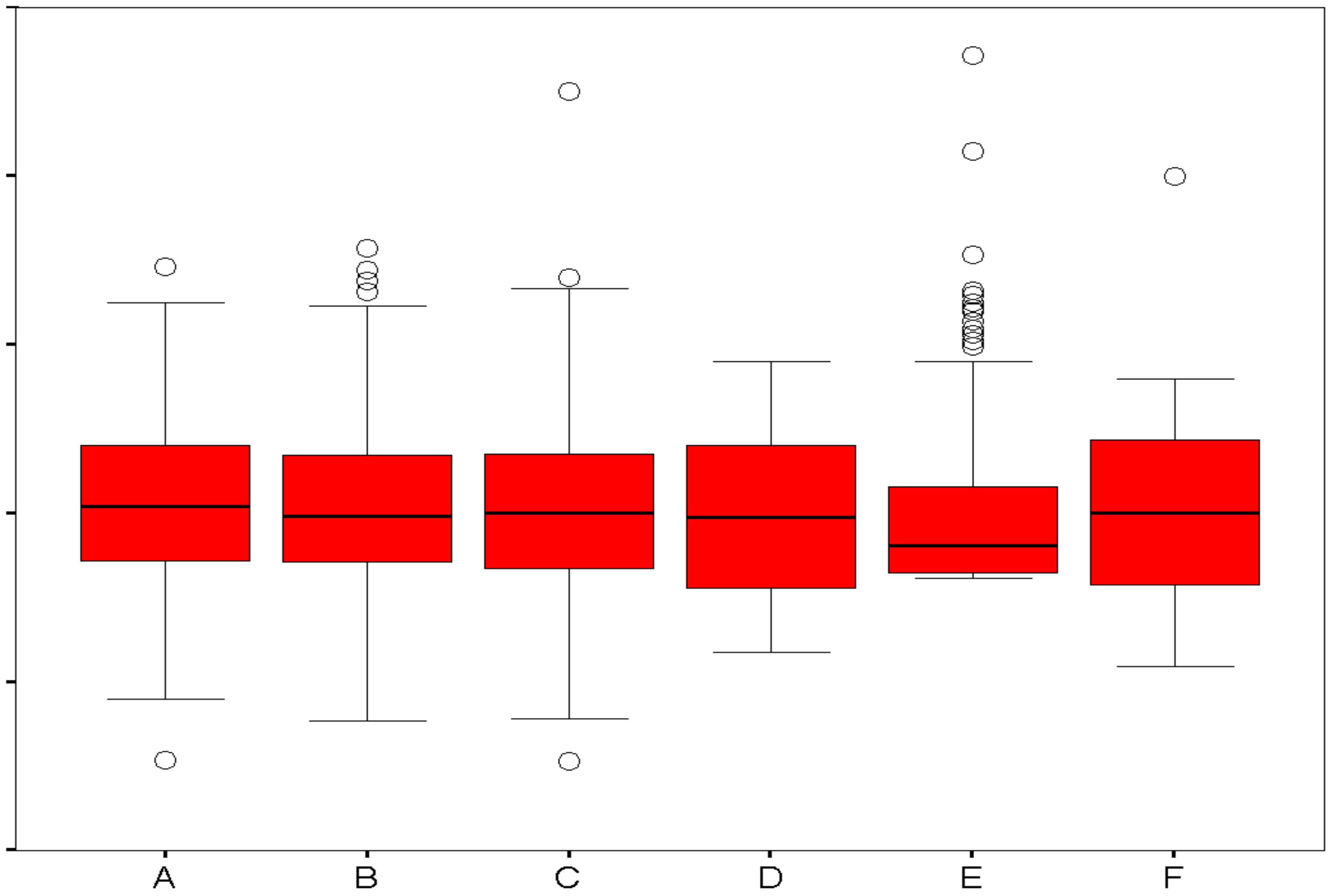
- E & F



Histograms can be tricky ….

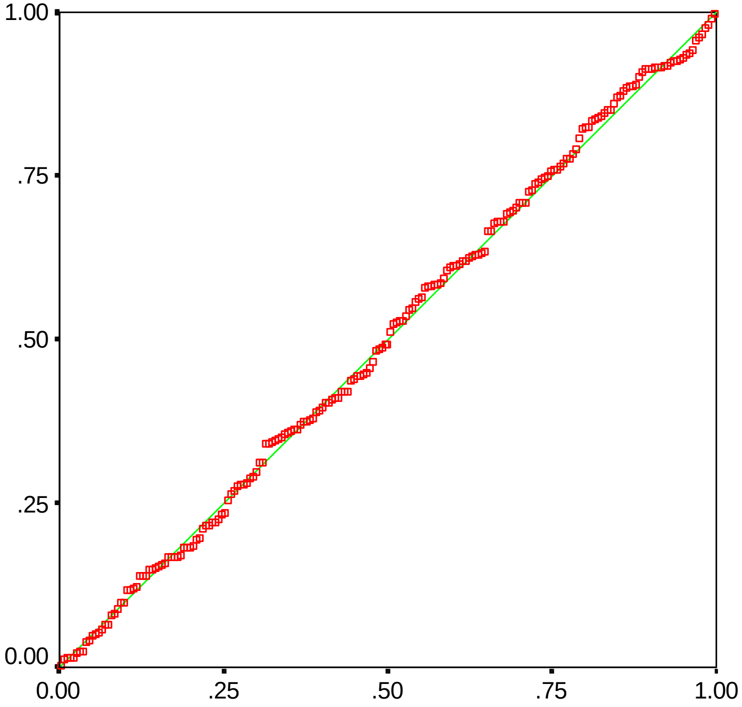


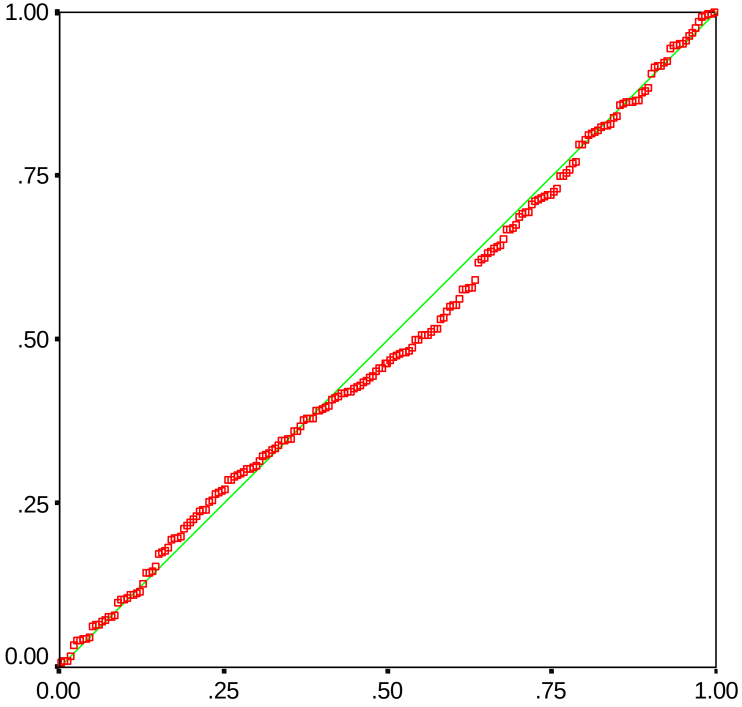
Boxplots



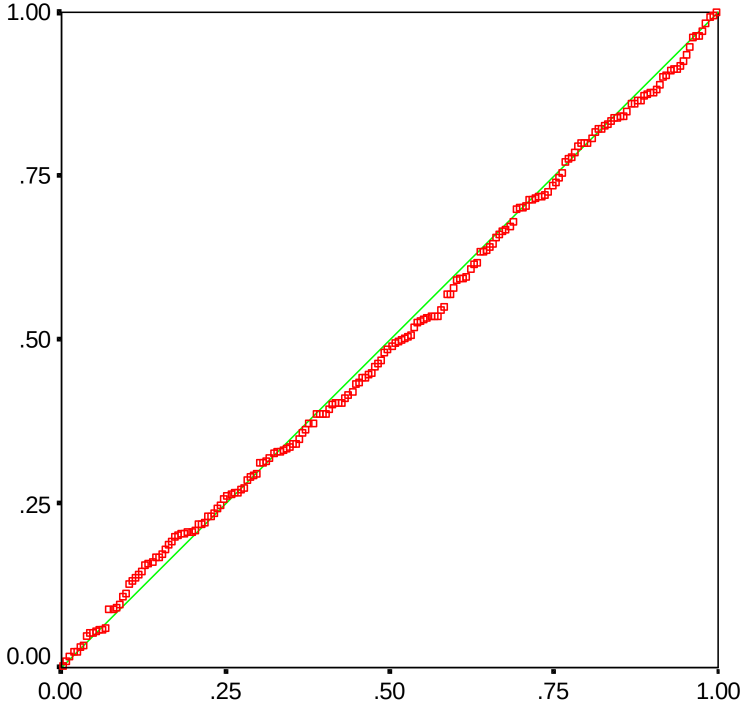
P-P Plots

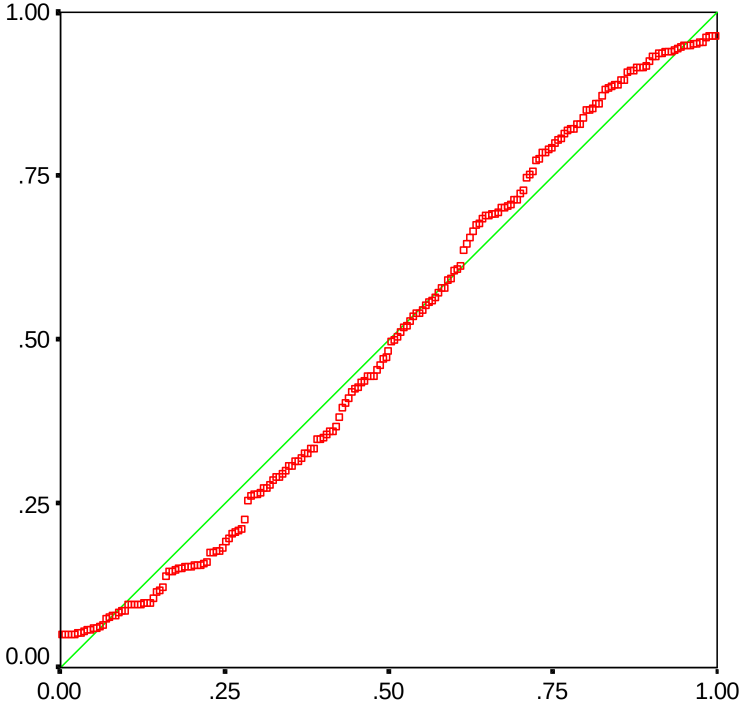
- A & B



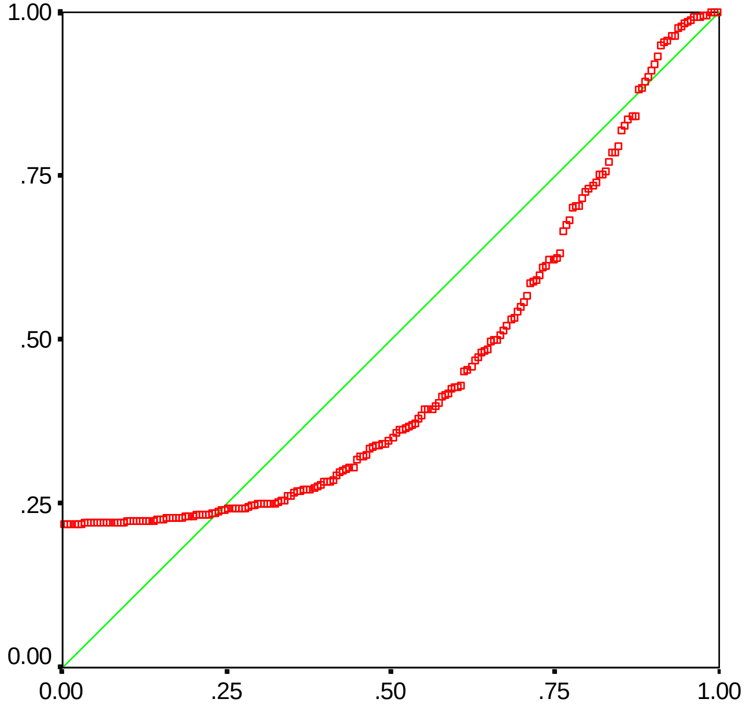


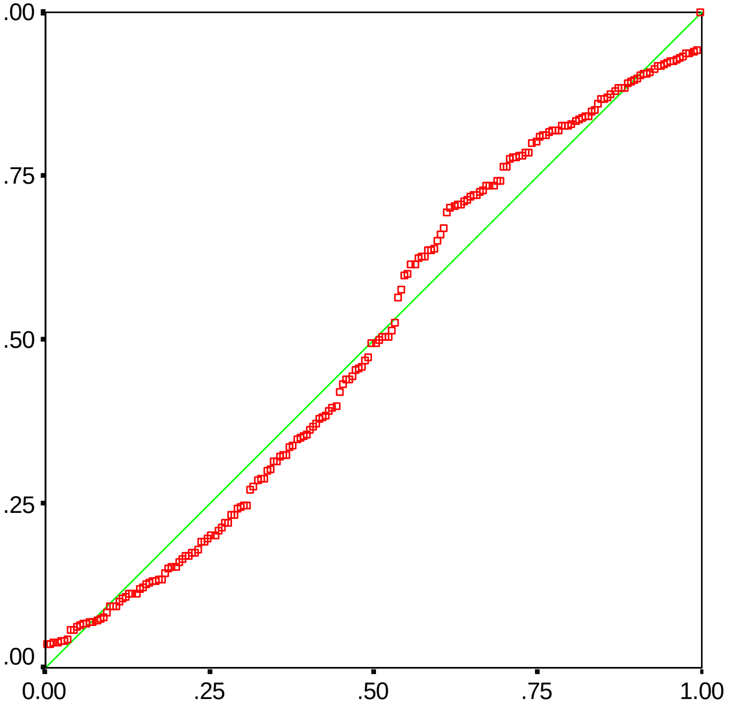
• C & D





- E & F

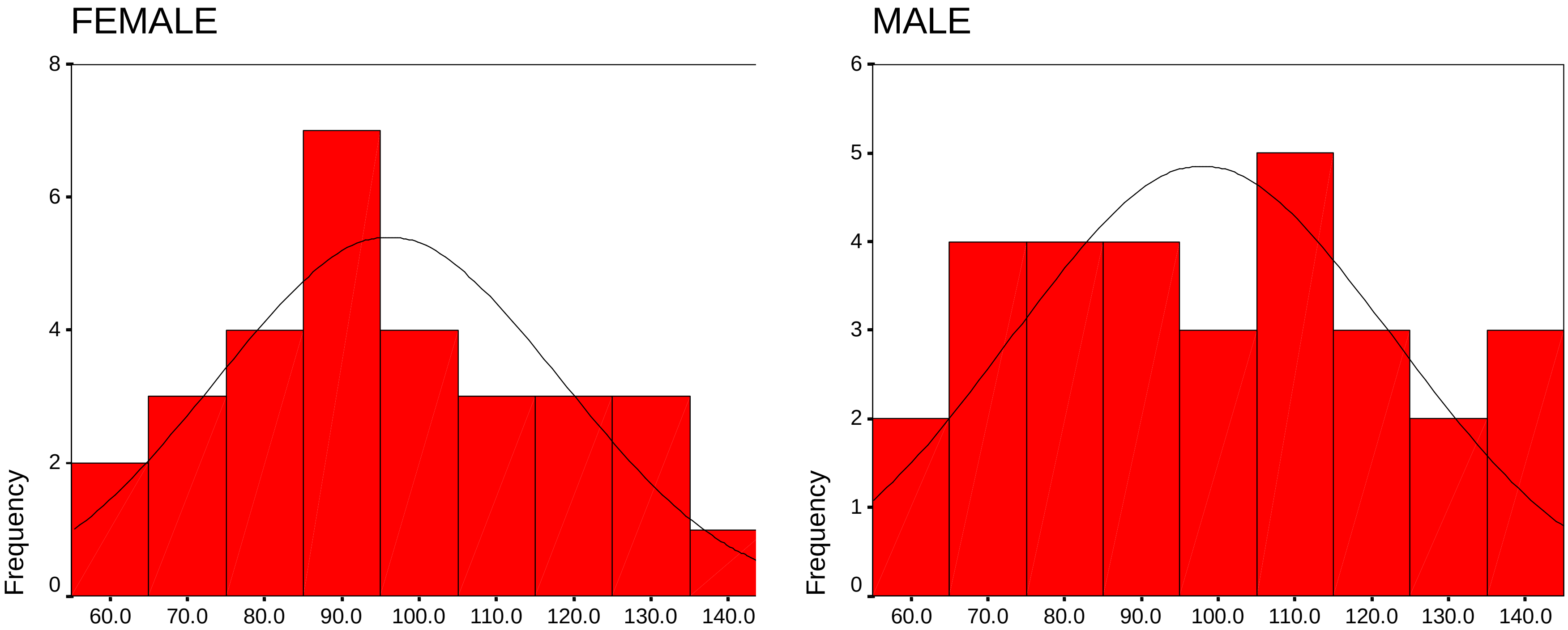




Bivariate Normality

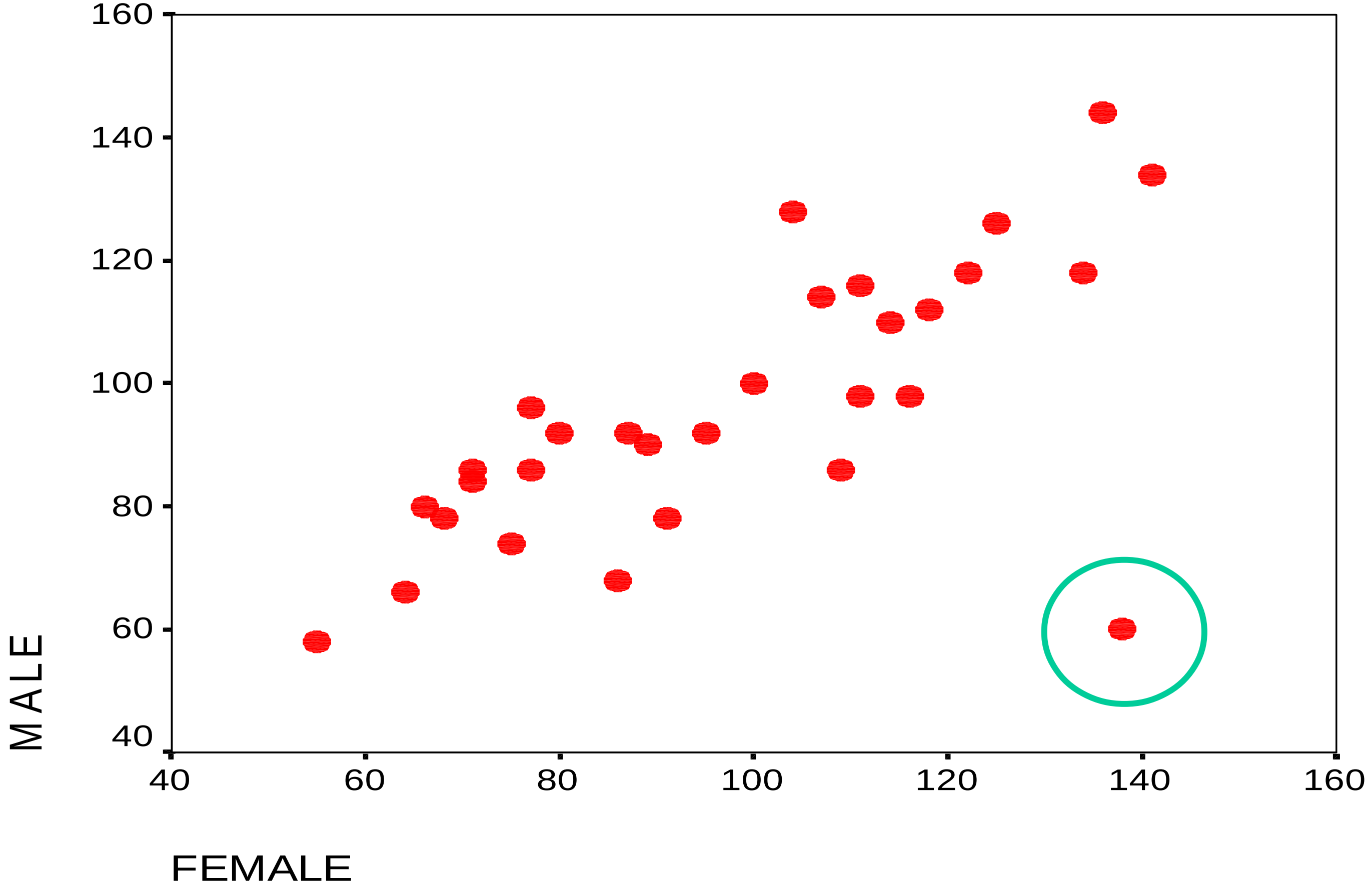
- We didn't just say "residuals normally distributed"
- We said "at every value of the dependent variables"
- Two variables can be normally distributed – univariate,
  - but not bivariate

- Couple's IQs
  - male and female



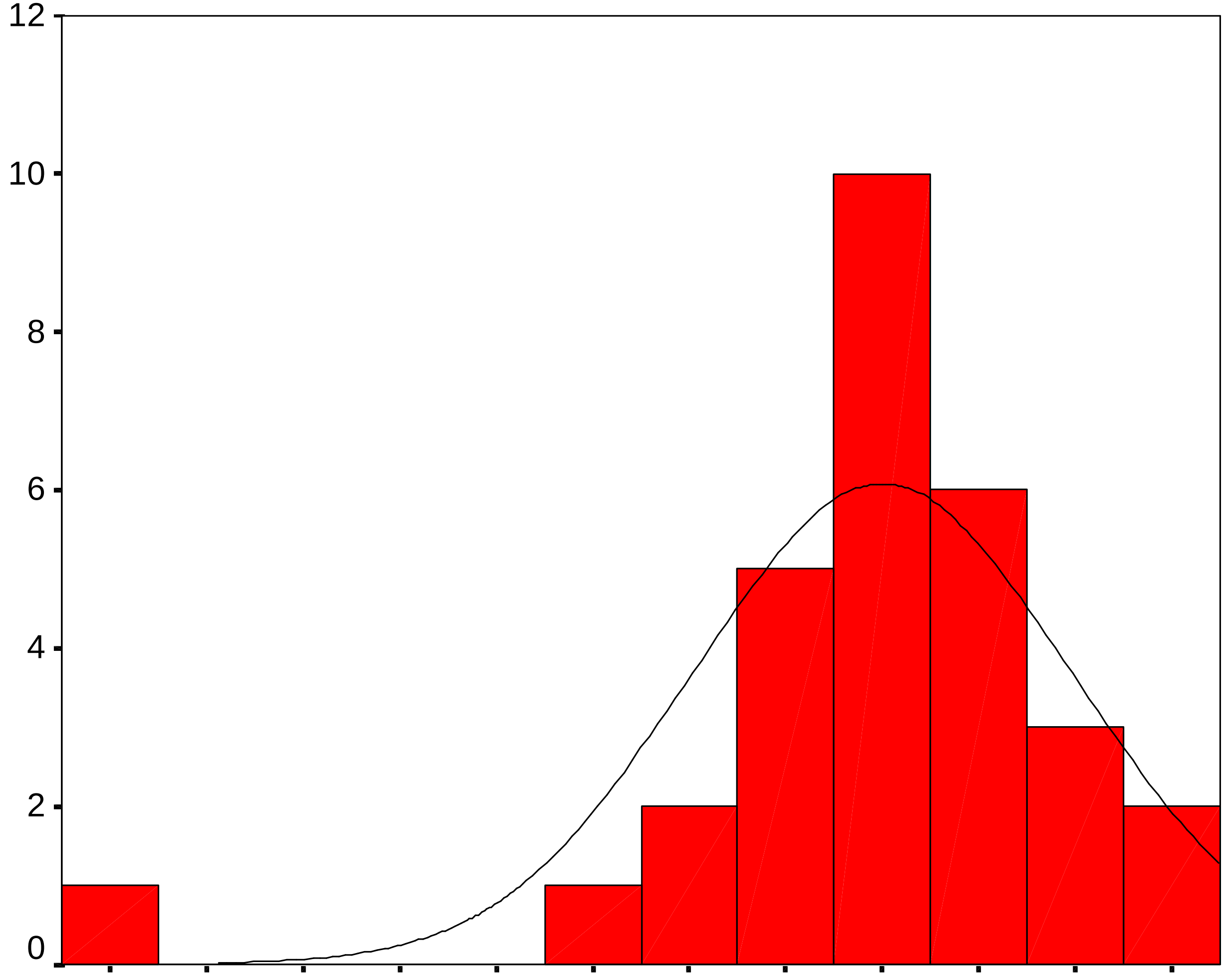
–Seem reasonably normal

- But wait!!



- When we look at bivariate normality
  - not normal – there is an outlier
- So plot X against Y
- OK for bivariate
  - but – may be a multivariate outlier
  - Need to draw graph in 3+ dimensions
  - can’t draw a graph in 3 dimensions
- But we can look at the residuals instead …

- IQ histogram of residuals



Multivariate Outliers …

- Will be explored later in the exercises

- So we move on …

What to do about Non-Normality

- Skew and Kurtosis
  - Skew – much easier to deal with
  - Kurtosis – less serious anyway
- Transform data
  - removes skew
  - positive skew – log transform
  - negative skew - square

Transformation

- May need to transform IV and/or DV
  - More often DV
    - time, income, symptoms (e.g. depression) all positively skewed
  - can cause non-linear effects (more later) if only one is transformed
  - alters interpretation of unstandardised parameter
  - May alter meaning of variable
  - May add / remove non-linear and moderator effects

- Change measures
  - increase sensitivity at ranges
    - avoiding floor and ceiling effects
- Outliers
  - Can be tricky
  - Why did the outlier occur?
    - Error? Delete them.
    - Weird person? Probably delete them
    - Normal person? Tricky.

- You are trying to model a process
  - is the data point ‘outside’ the process
  - e.g. lottery winners, when looking at salary
  - yawn, when looking at reaction time
- Which is better?
  - A good model, which explains 99% of your data?
  - A poor model, which explains all of it

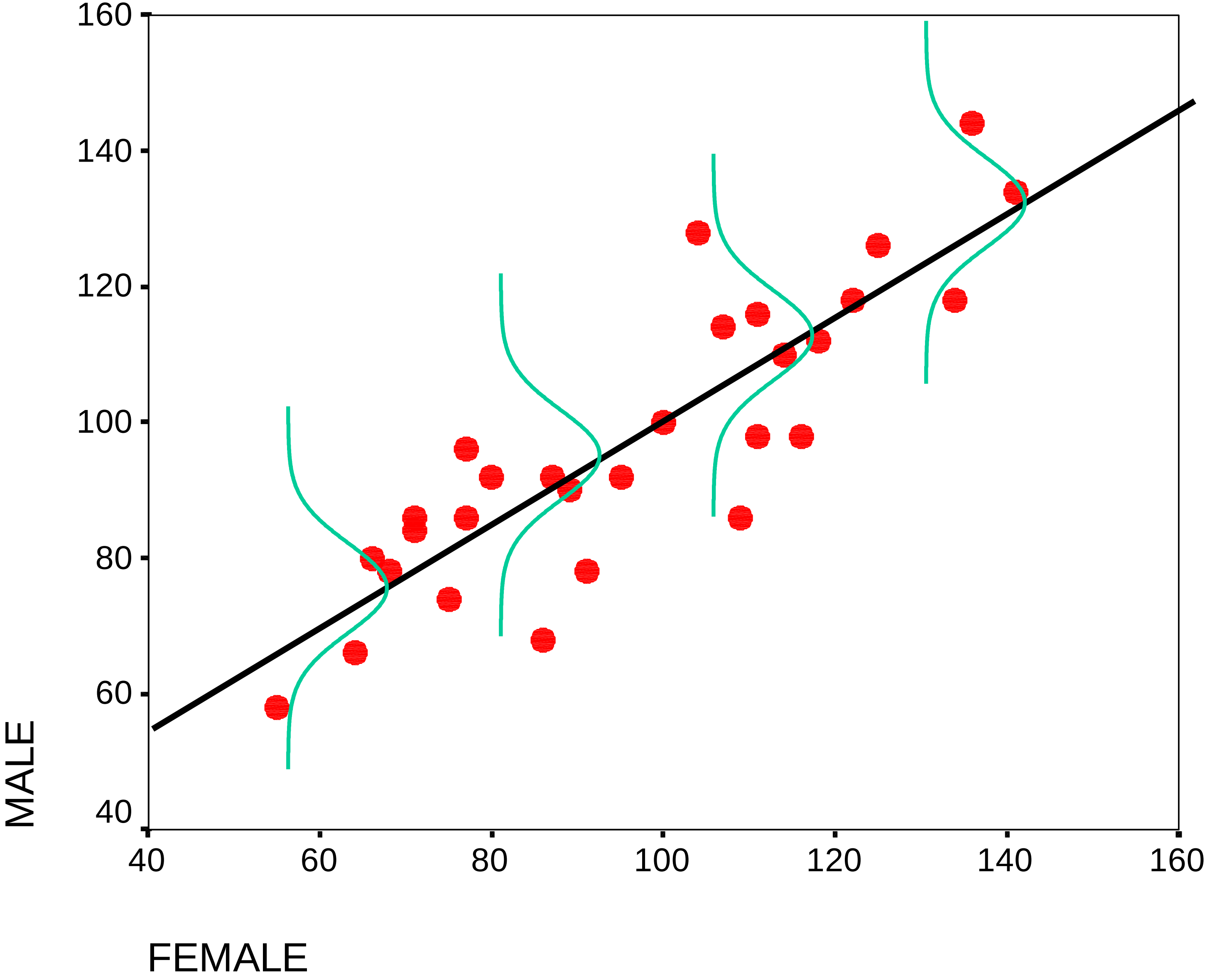
- Pedhazur and Schmelkin (1991)
  - analyse the data twice

- We will spend much less time on the other 6 assumptions

Assumption 2: The variance of the residuals for every set of values for the independent variable is equal.

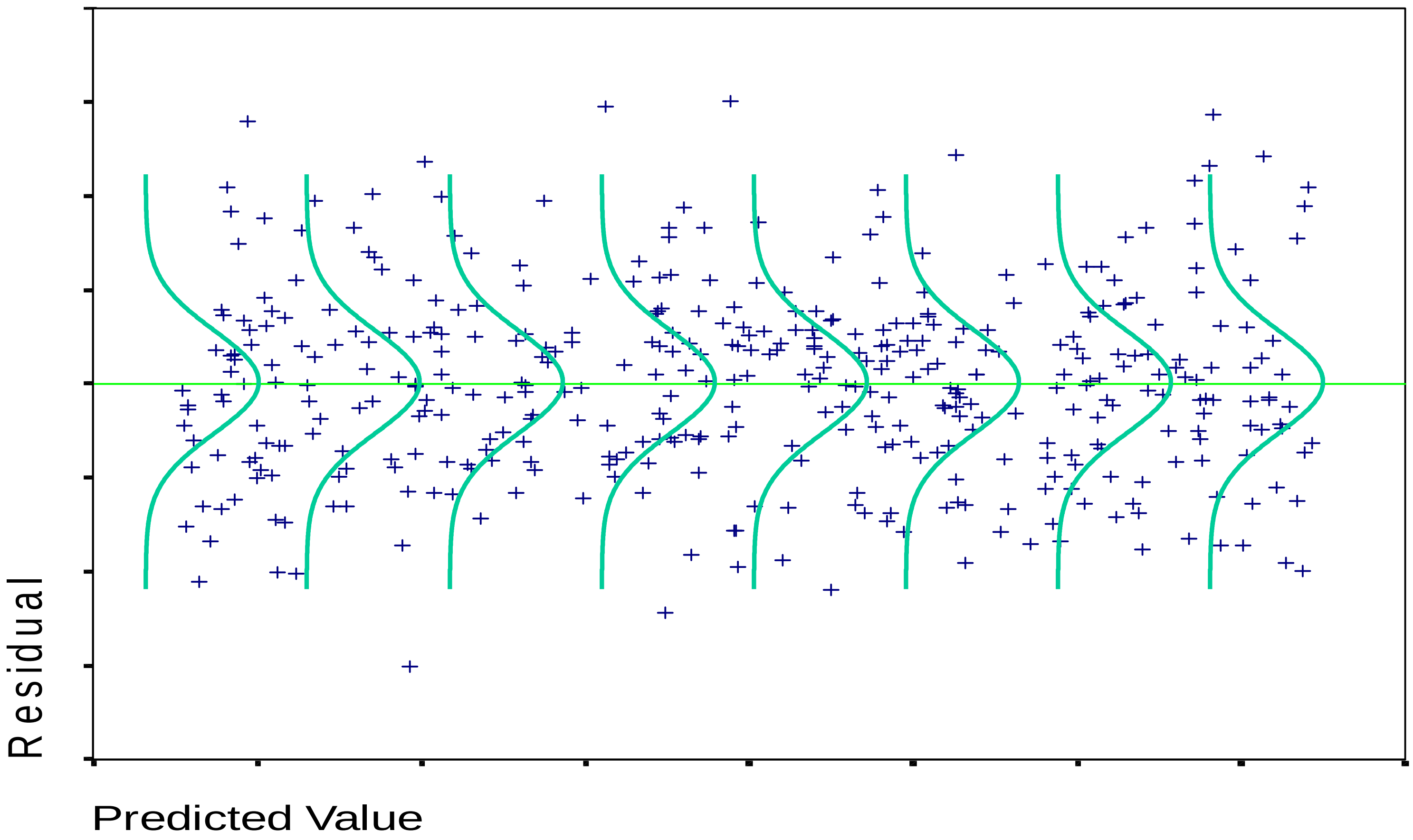
Heteroscedasticity

- This assumption is a about heteroscedasticity of the residuals
  - Hetero=different
  - Scedastic = scattered
- We don’t want heteroscedasticity
  - we want our data to be homoscedastic
- Draw a scatterplot to investigate

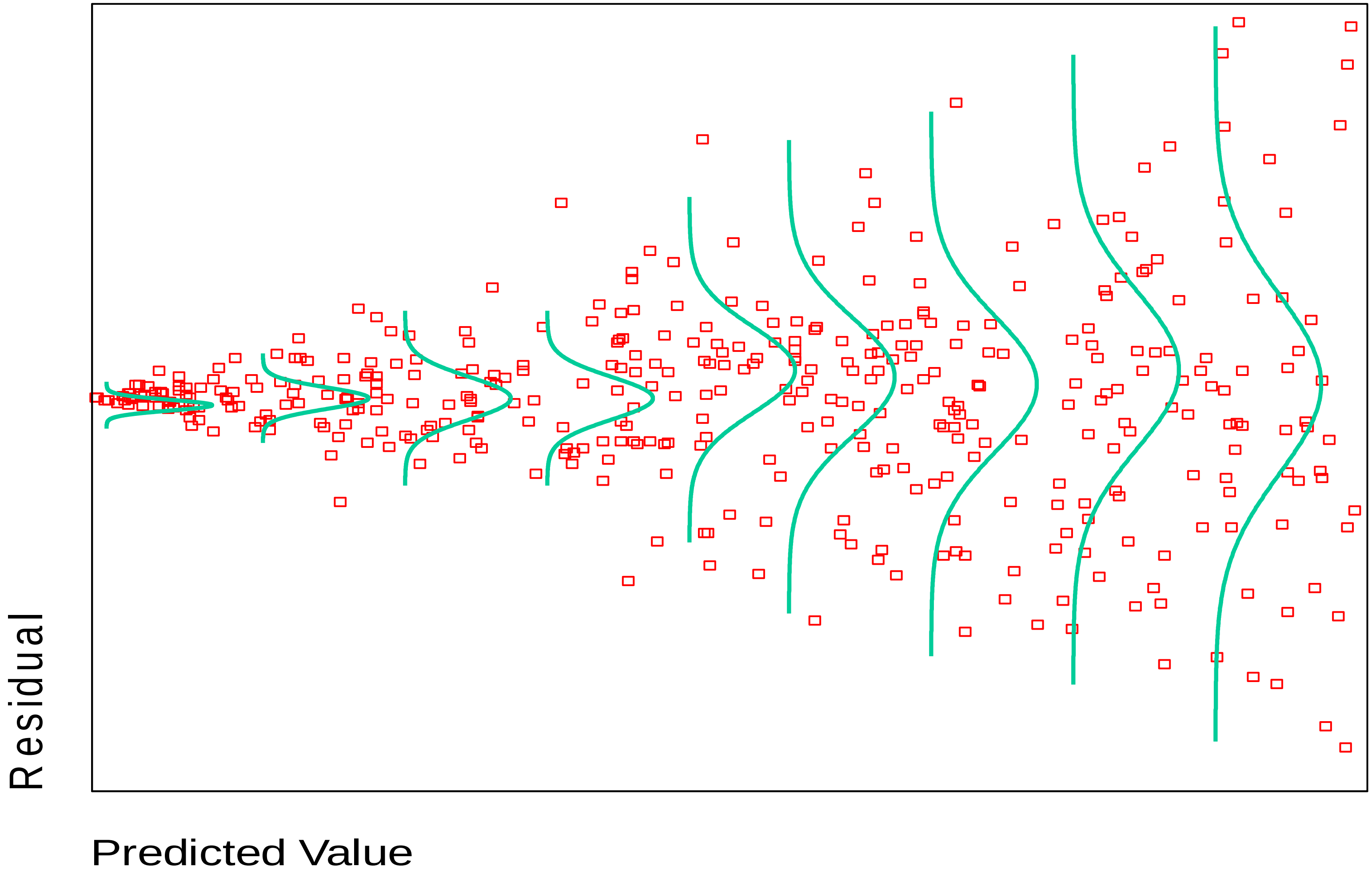

MALE
160
140
120
100
80
60
40
40
60
80
100
120
140
160
FEMALE

- Only works with one IV
  - need every combination of IVs
- Easy to get – use predicted values
  - use residuals there
- Plot predicted values against residuals
  - or standardised residuals
  - or deleted residuals
  - or standardised deleted residuals
  - or studentised residuals
- A bit like turning the scatterplot on its side

Good – no heteroscedasticity



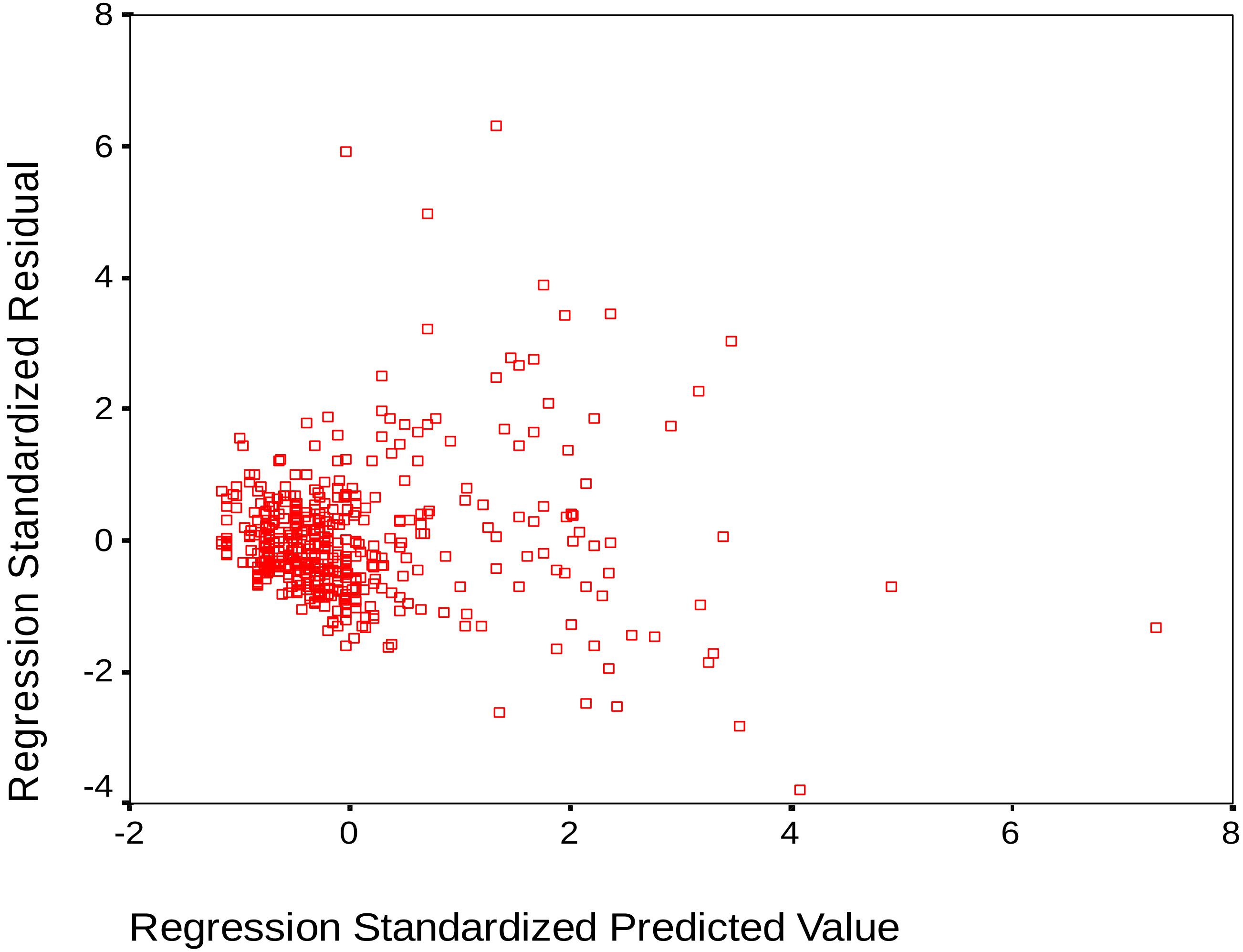
Bad – heteroscedasticity



Testing Heteroscedasticity

- White’s test
  - Not automatic in SPSS (is in SAS)
  - Luckily, not hard to do
  - More luckily, we aren’t going to do it
    - (In the very unlikely event you will ever have to do it, look it up.
    - Google: White’s test spss)

Plot of Pred and Res

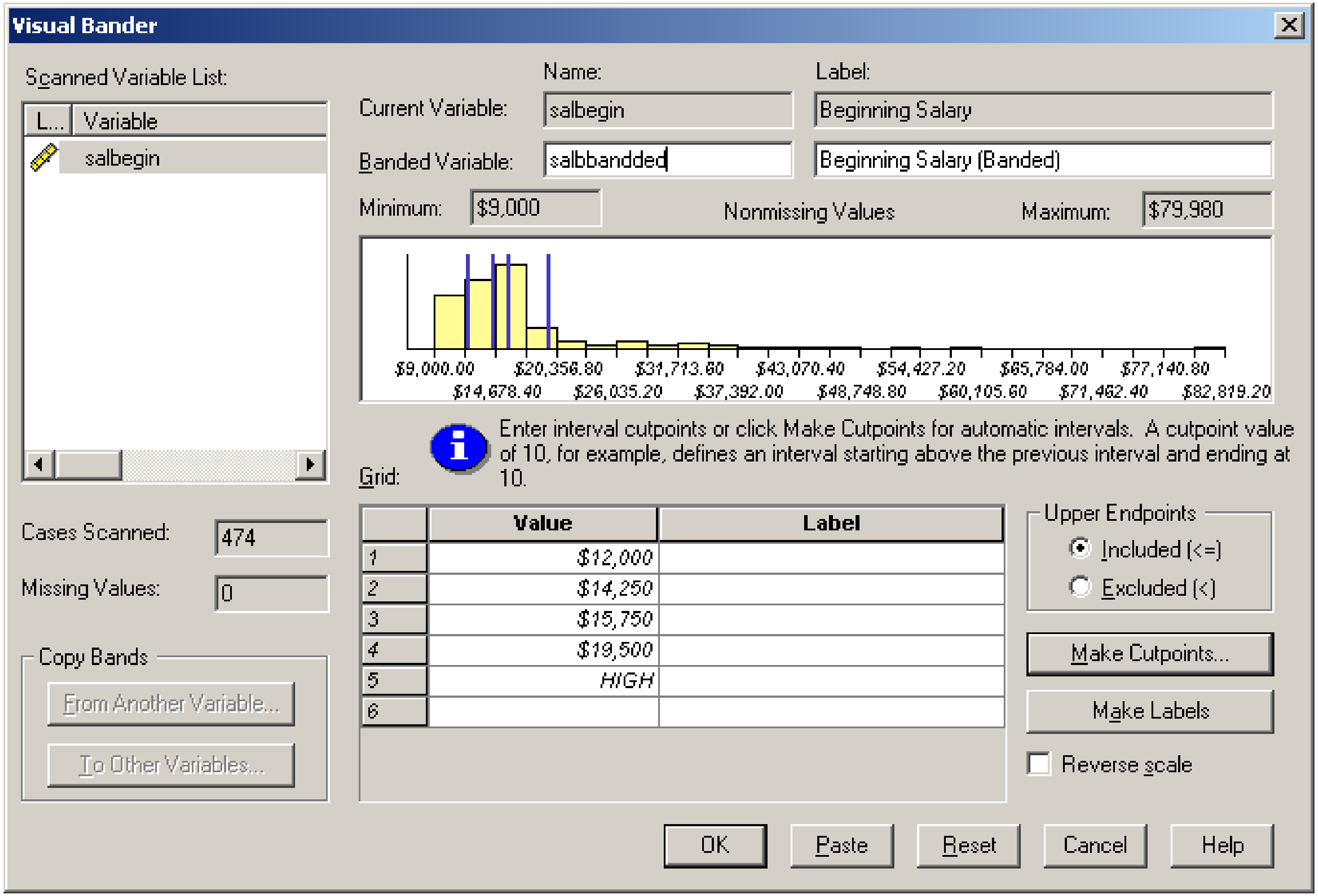


Magnitude of Heteroscedasticity

- Chop data into “slices”
  - 5 slices, based on X (or predicted score)
    - Done in SPSS
  - Calculate variance of each slice
  - Check ratio of smallest to largest
  - Less than 10:1
    - OK

The Visual Bander

- New in SPSS 12



- Variances of the 5 groups

1	.219
2	.336
3	.757
4	.751
5	3.119

- We *have* a problem
  - 3 / 0.2 ~= 15

Assumption 3: The Error Term is Additive

Additivity

- We sum the scores in the regression equation
  – Is that legitimate?
  – can test for it, but hard work
- Have to know it from your theory
- A specification error

Additivity and Theory

- Two IVs
  - Alcohol has sedative effect
    - A bit makes you a bit tired
    - A lot makes you very tired
  - Some painkillers have sedative effect
    - A bit makes you a bit tired
    - A lot makes you very tired
  - A bit of alcohol and a bit of painkiller doesn't make you very tired
  - Effects multiply together, don't add together

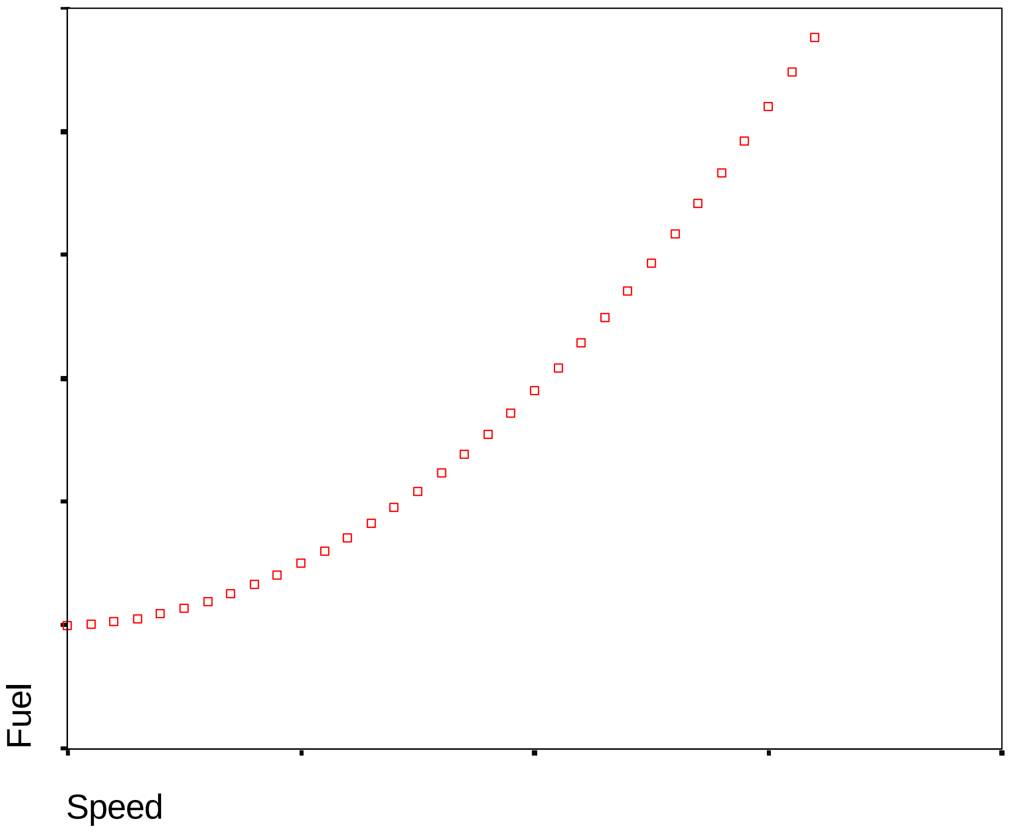
- If you don't test for it
  - It's very hard to know that it will happen
- So many possible non-additive effects
  - Cannot test for all of them
  - Can test for obvious
- In medicine
  - Choose to test for salient non-additive effects
  - e.g. sex, race

Assumption 4: At every value of the dependent variable the expected (mean) value of the residuals is zero

Linearity

- Relationships between variables should be linear
  - best represented by a straight line
- Not a very common problem in social sciences
  - except economics
  - measures are not sufficiently accurate to make a difference
    - R^2 too low
    - unlike, say, physics

- Relationship between speed of travel and fuel used

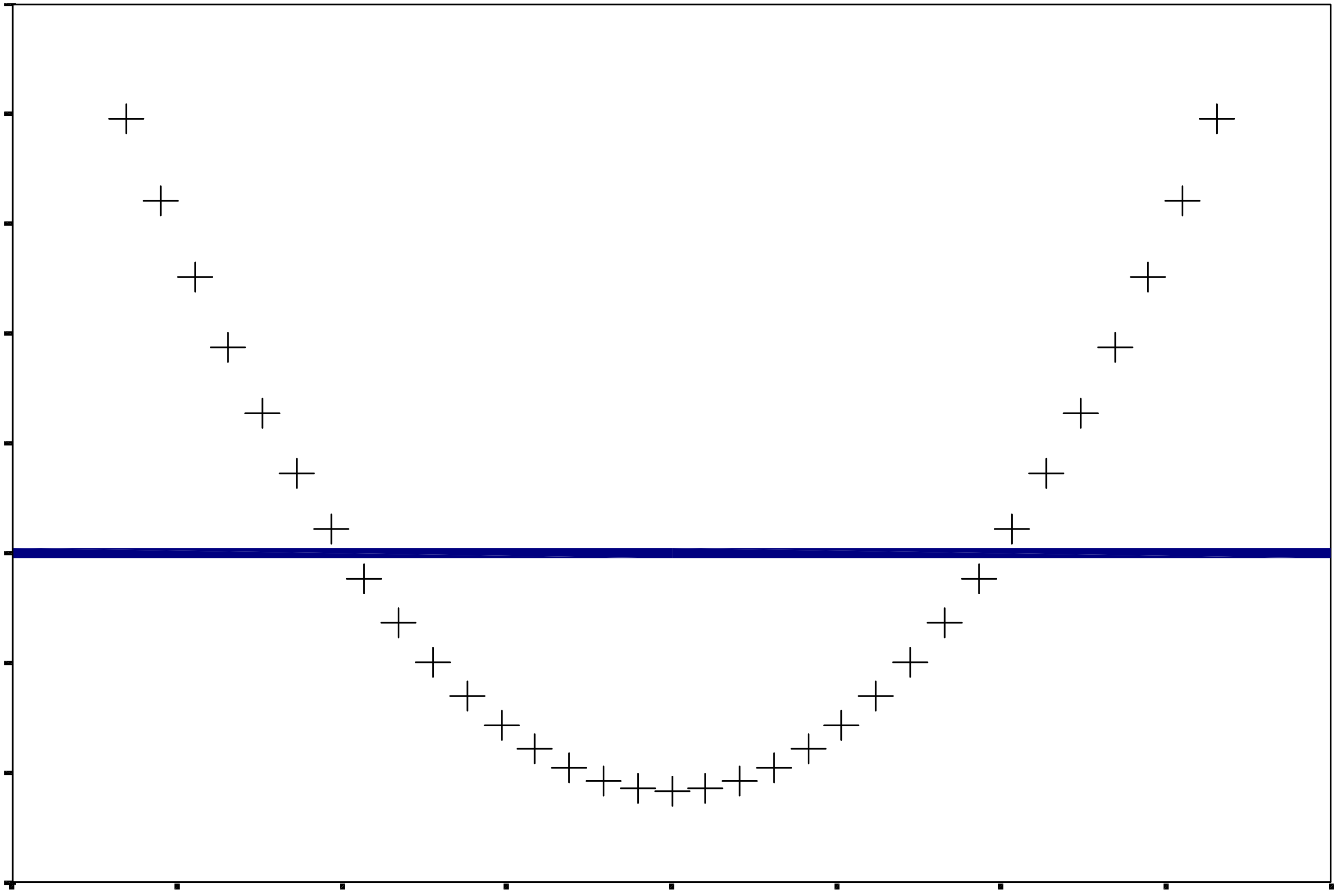


- $R^2 = 0.938$
  - looks pretty good
  - know speed, make a good prediction of fuel
- BUT
  - look at the chart
  - if we know speed we can make a perfect prediction of fuel used
  - R^2 should be 1.00

Detecting Non-Linearity

- Residual plot
  - just like heteroscedasticity
- Using this example
  - very, very obvious
  - usually pretty obvious

Residual plot



Linearity: A Case of Additivity

- Linearity = additivity along the range of the IV
- Jeremy rides his bicycle harder
  - Increase in speed depends on current speed
  - Not additive, multiplicative
  - MacCallum and Mar (1995). Distinguishing between moderator and quadratic effects in multiple regression. *Psychological Bulletin*.

Assumption 5: The expected correlation between residuals, for any two cases, is 0.

The independence assumption
(lack of autocorrelation)

Independence Assumption

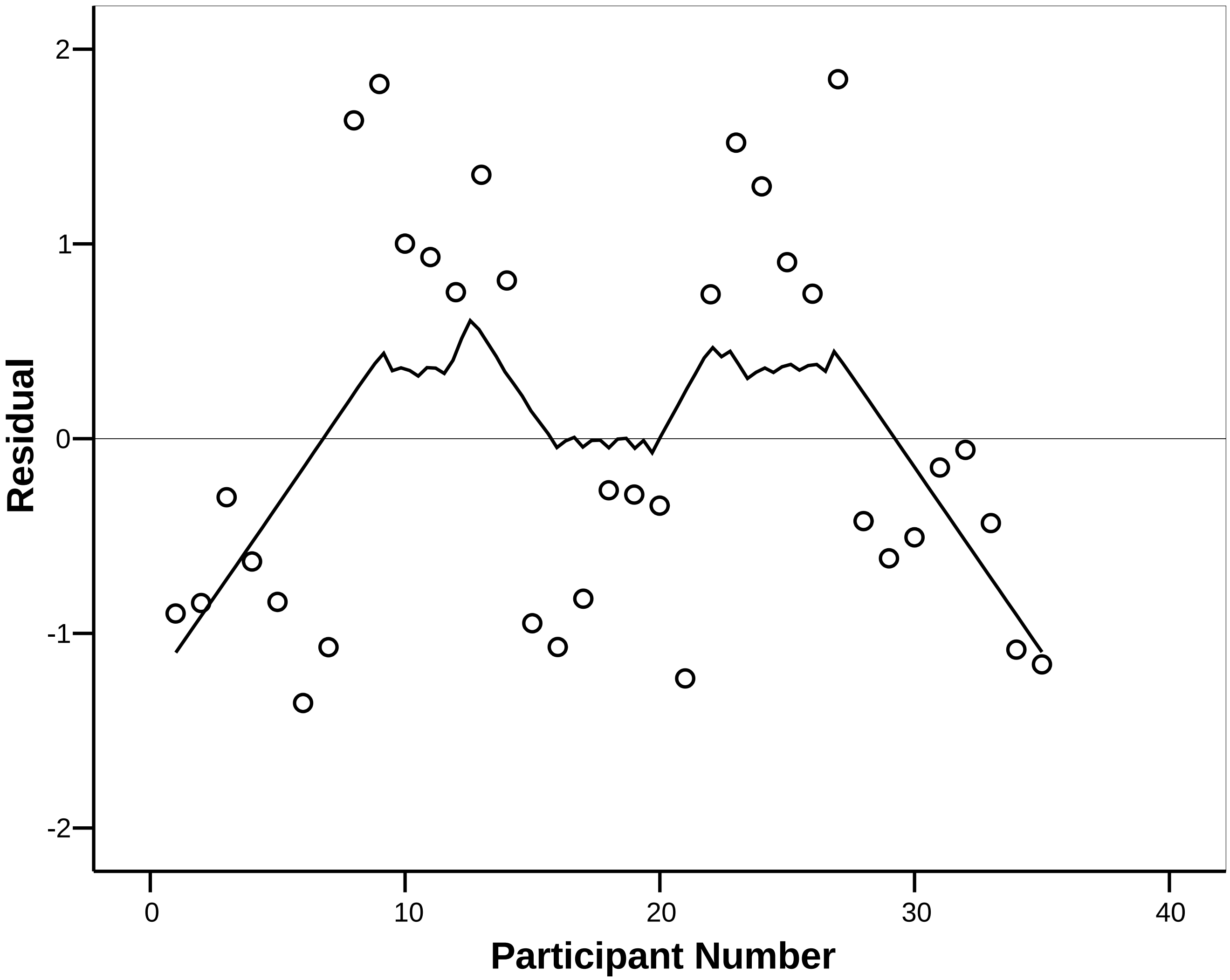
- Also: lack of autocorrelation
- Tricky one
  - often ignored
  - exists for almost all tests
- All cases should be independent of one another
  - knowing the value of one case should not tell you anything about the value of other cases

How is it Detected?

- Can be difficult
  - need some clever statistics (multilevel models)
- Better off avoiding situations where it arises
- Residual Plots
- Durbin-Watson Test

Residual Plots

- Were data collected in time order?
  - If so plot ID number against the residuals
  - Look for any pattern
    - Test for linear relationship
    - Non-linear relationship
    - Heteroscedasticity



Residual
2
1
0
-1
-2
0
10
20
30
40
Participant Number

How does it arise?

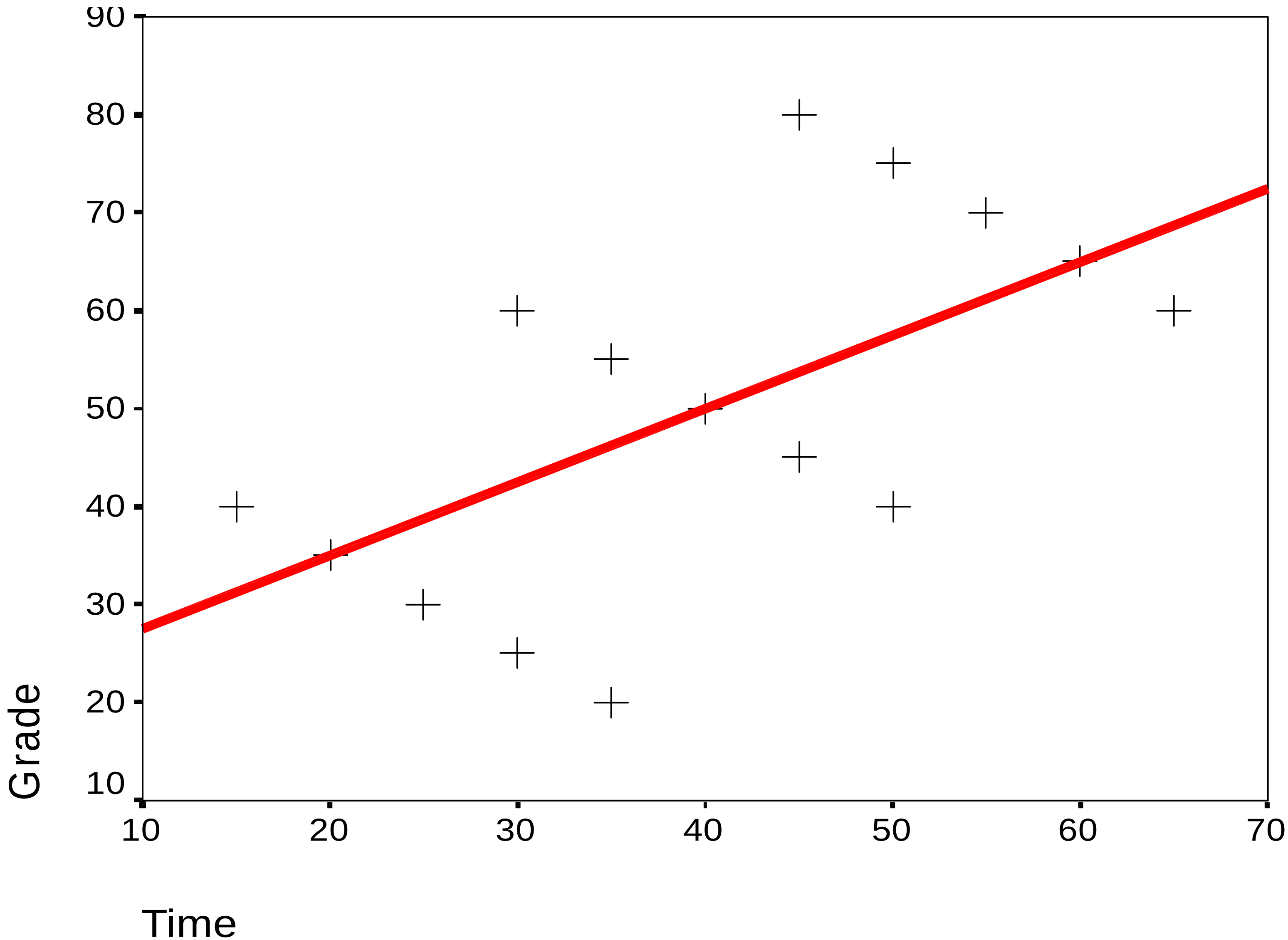
Two main ways

- time-series analyses
  - When cases are time periods
    - weather on Tuesday and weather on Wednesday correlated
    - inflation 1972, inflation 1973 are correlated
- clusters of cases
  - patients treated by three doctors
  - children from different classes
  - people assessed in groups

Why does it matter?

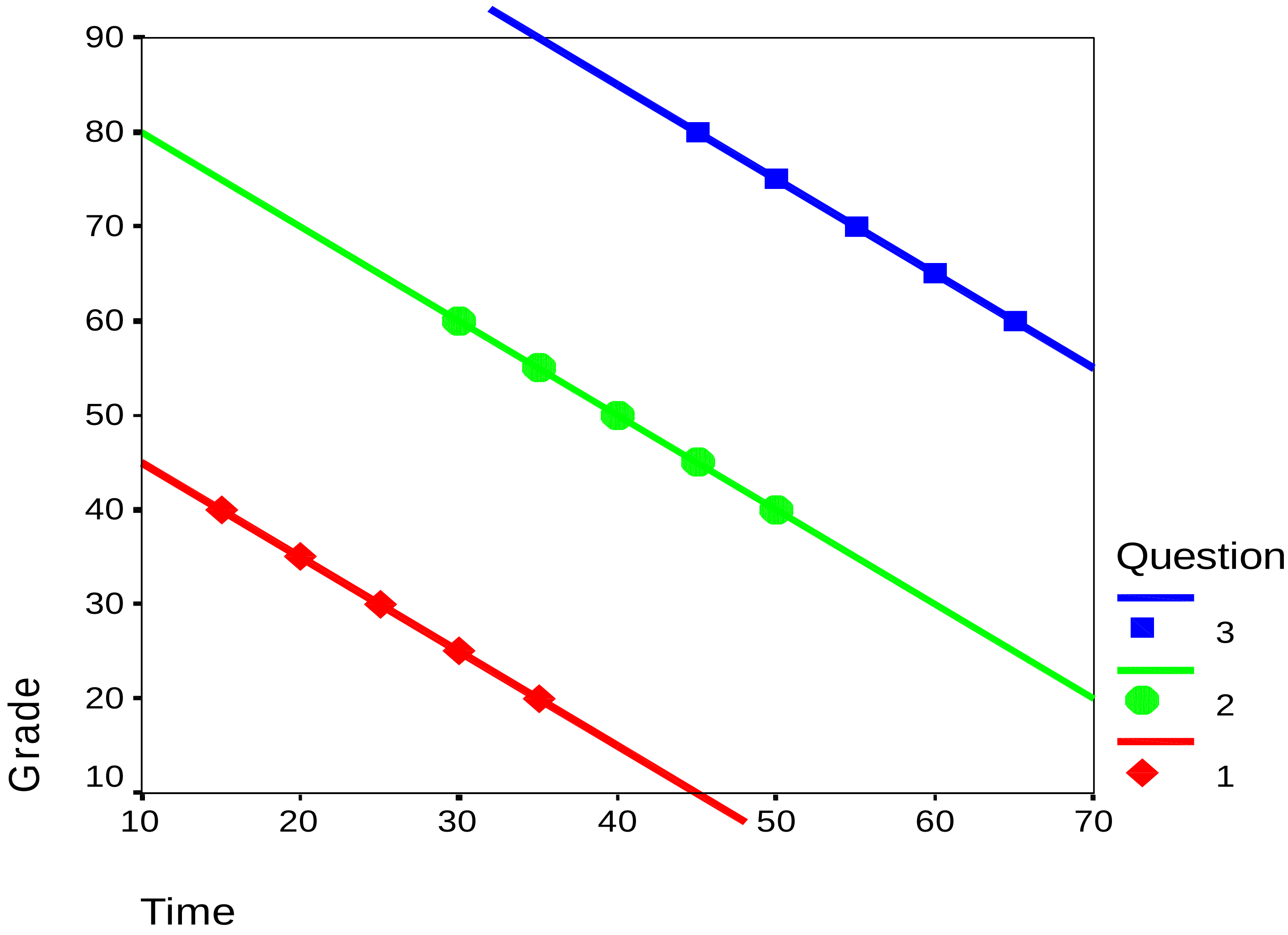
- Standard errors can be wrong
  - therefore significance tests can be wrong
- Parameter estimates can be wrong
  - really, really wrong
  - from positive to negative
- An example
  - students do an exam (on statistics)
  - choose one of three questions
    - IV: time
    - DV: grade

- Result, with line of best fit



- Result shows that
  - people who spent longer in the exam, achieve better grades
- BUT …
  - we haven’t considered which question people answered
  - we might have violated the independence assumption
    - DV will be autocorrelated
- Look again
  - with questions marked

- Now somewhat different



- Now, people that spent longer got lower grades
  - questions differed in difficulty
  - do a hard one, get better grade
  - if you can do it, you can do it quickly
- Very difficult to analyse well
  - need multilevel models

Assumption 6: All independent variables are uncorrelated with the error term.

Uncorrelated with the Error Term

- A curious assumption
  - by definition, the residuals are uncorrelated with the independent variables (try it and see, if you like)
- It is about the DV
  - must have no effect (when the IVs have been removed)
  - on the DV

- Problem in economics
  - Demand increases supply
  - Supply increases wages
  - Higher wages increase demand
- OLS estimates will be (badly) biased in this case
  - need a different estimation procedure
  - two-stage least squares
    - simultaneous equation modelling

Assumption 7: No independent variables are a perfect linear function of other independent variables

no perfect multicollinearity

No Perfect Multicollinearity

- IVs must not be linear functions of one another
  - matrix of correlations of IVs is not positive definite
  - cannot be inverted
  - analysis cannot proceed
- Have seen this with
  - age, age start, time working
  - also occurs with subscale and total

- Large amounts of collinearity
  - a problem (as we shall see) sometimes
  - not an assumption

Assumption 8: The mean of the error term is zero.

You will like this one.

Mean of the Error Term = 0

- Mean of the residuals = 0
- That is what the constant is for
  - if the mean of the error term deviates from zero, the constant soaks it up

$$Y = \beta_0 + \beta_1 x_1 + \varepsilon$$

$$Y = (\beta_0 + 3) + \beta_1 x_1 + (\varepsilon - 3)$$

- note, Greek letters because we are talking about population values