Assumptions in Regression Analysis

Agenda

Sl. No.	Topics For The Agenda
1.	The Assumptions
2.	Assumption 1
3.	A Look At Normal Distributions
4.	Examining Univariate Distributions
5.	Boxplots
6.	QQ Plot Of Residuals
7.	Regression Example Using R Code
8.	Residuals Are Independent Of X
9.	What About Non-normality
10.	Assumption 2
11.	Heteroscedasticity
12.	Plot Of Pred And Res
13.	Assumption 3
14.	Additivity
15.	Assumption 4
16.	Linearity

Sl. No.	Topics For The Agenda
17.	Detecting Non-Linearity
18.	Residual Plot
19.	Linearity: A Case Of Additivity
20.	Assumption 5
21.	Independence Assumption
22.	Residual Plots
23.	How Does It Arise?
24.	Why Does It Matter?
25.	Result, With Line Of Best Fit
26.	What The Result Shows
27.	Some Difference
28.	Assumption 6
29.	Uncorrelated With The Error Term
30.	No Perfect Multicollinearity
31.	Assumption 8
32.	Mean Of The Error Term = 0

The Assumptions

- The residuals have a normal distribution for respective dependant variable values

- For respective independent variable values, the residual variance is homogenous and is independent of the Y variable

- There is no specification error with respect to the model structure and the model is additive along with the residuals

- The sum of the residuals is always zero. Ideally the expected value of the residual at respective values of Y is zero

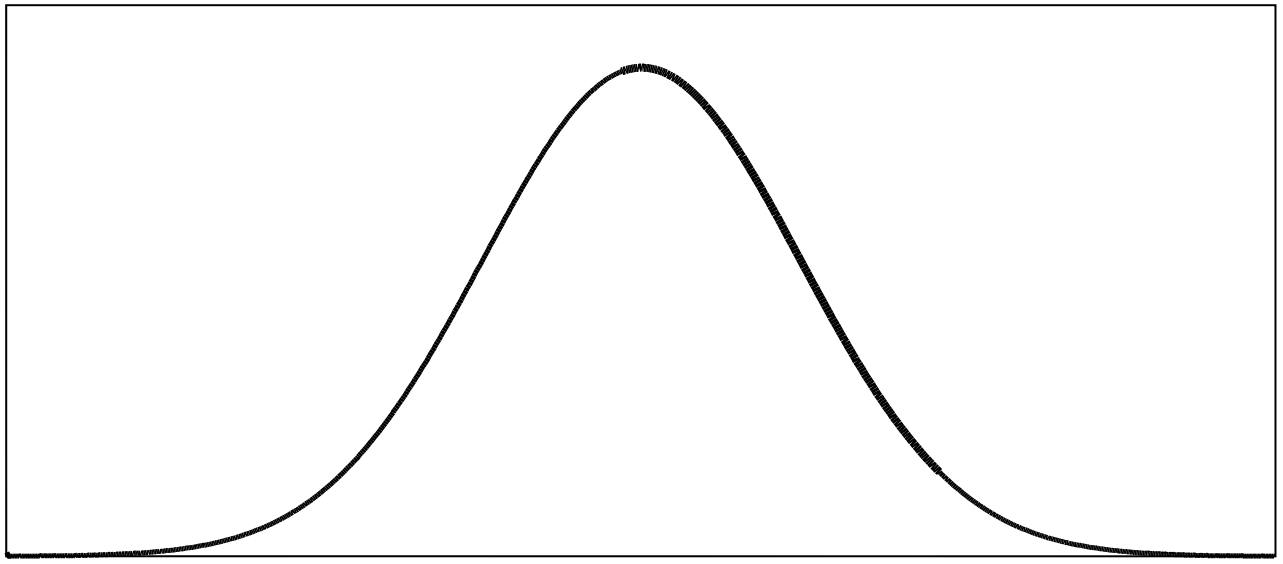
The Assumptions (Contd.)

- There is no autocorrelation between residuals
- All independent variables are uncorrelated with the error term
- There is no perfect multicollinearity between independent variables
- The mean of the error term is zero

Assumption 1

A Look At Normal Distributions

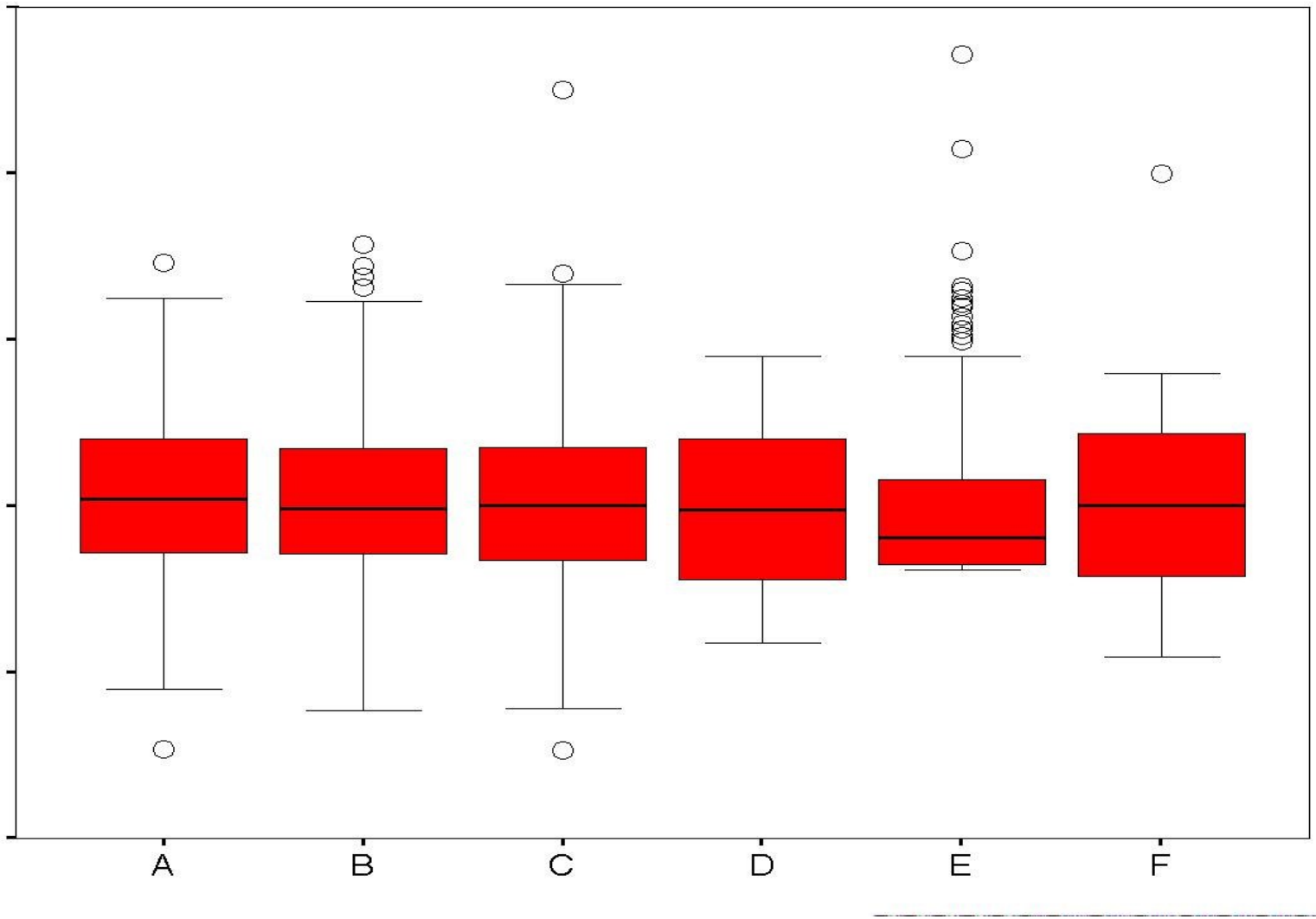
- A normal distribution is a balanced bell shaped curve with probability density values laid out such that:
  - Skewness – is zero i.e. the curve is neither towards the left nor towards the right
  - Kurtosis – the curve is neither too peaked nor too flat
  - Outliers – no outliers



Examining Univariate Distributions

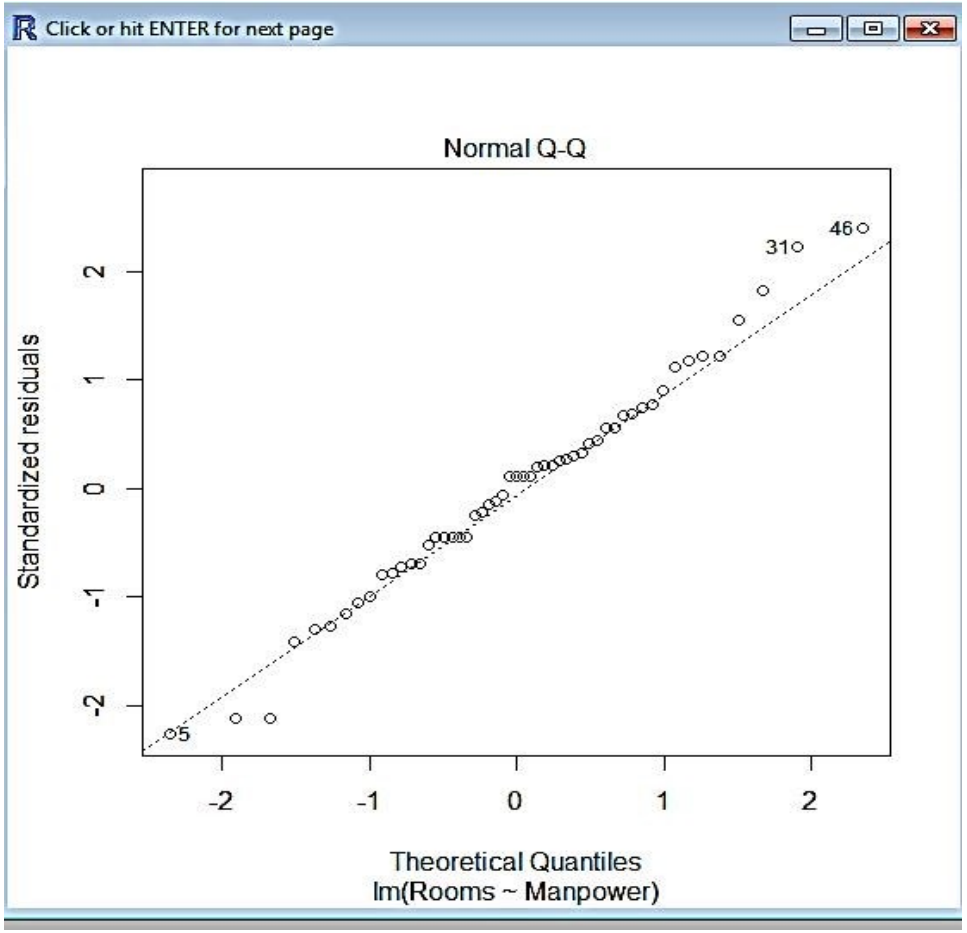
- The following diagrams are popularly used to examine distributions
  - Histograms
  - Boxplots
  - P-P Plots

Boxplots



QQ Plot Of Residuals

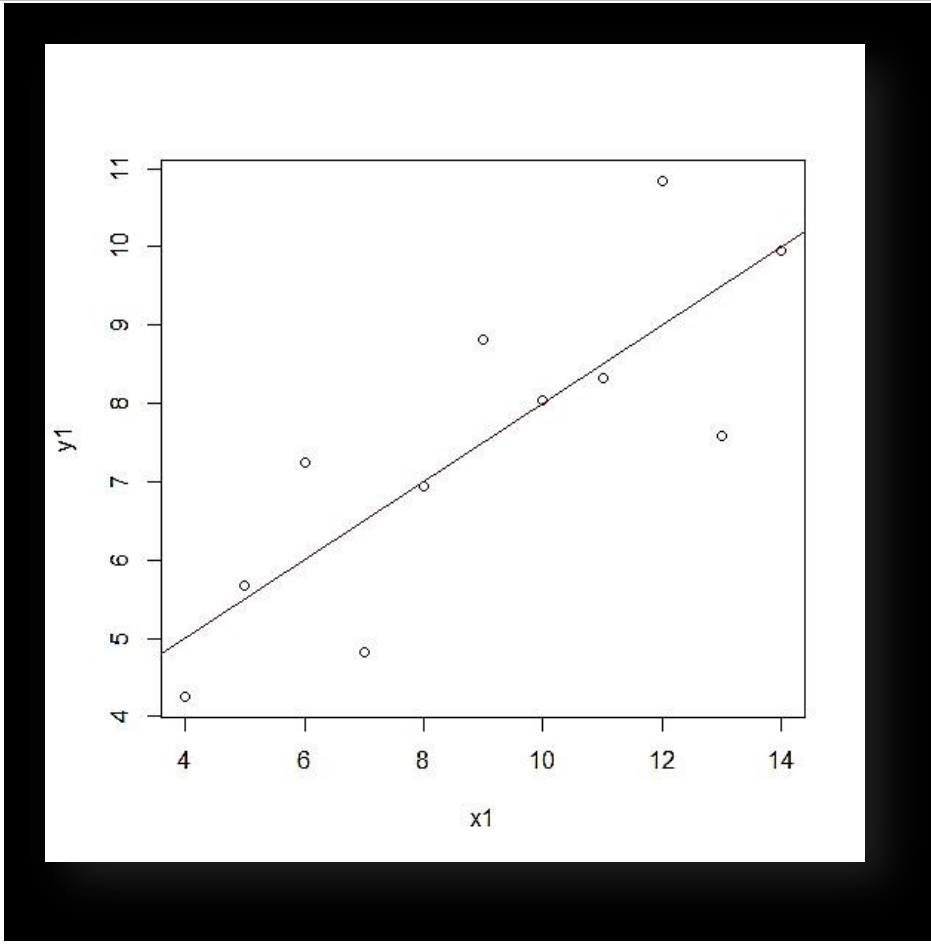
```
attach(neat)
# Regression model
neat.lm <- lm(Rooms~Manpower ,
data=neat)
# Plotting data and regression line
plot(Rooms~Manpower)
abline(neat.lm , col =2)
# Diagnostic plots
plot(neat.lm)
```



Regression Example Using R Code

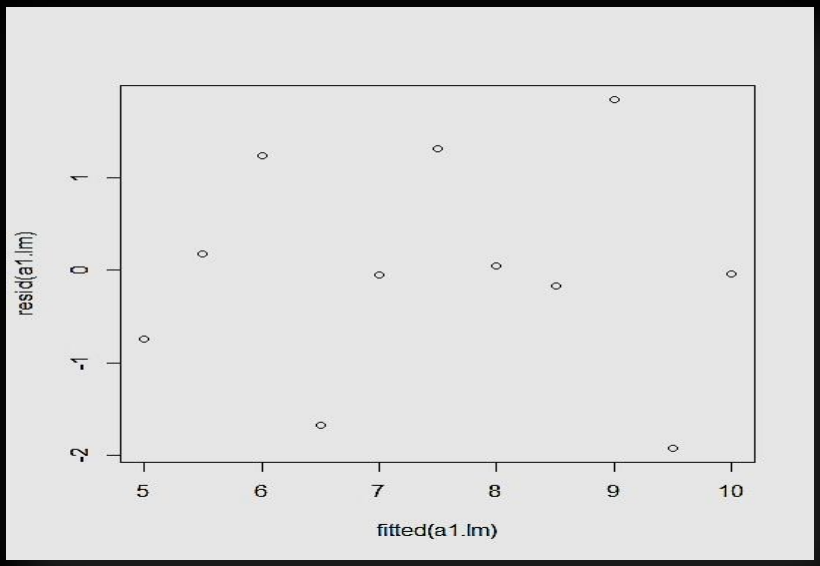
```
# Fitting the regressions
a1.lm <- lm(y1~x1 , data=proddata)
a2.lm <- lm(y2~x2 , data=proddata)
a3.lm <- lm(y3~x3 , data=proddata)
a4.lm <- lm(y4~x4 , data=proddata)

#Plotting
# For the first data set
plot(y1~x1 , data=proddata)
abline(a1.lm , col =2)
```

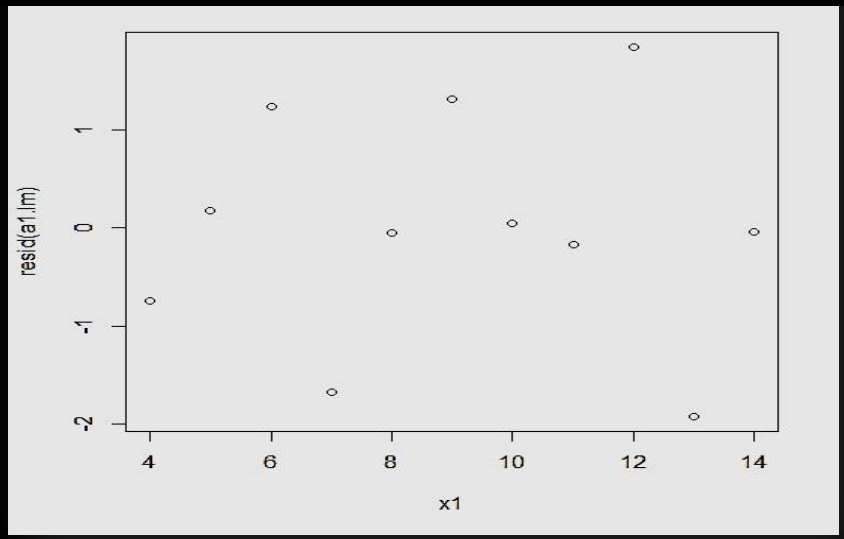


Residuals Are Independent Of X

#Residuals vs. fitted values
# For the first data set
plot(resid(a1.lm)~fitted(a1.lm))



#Checking assumptions graphically
#Residuals vs. X
# For the first data set
plot(resid(a1.lm)~x1)



What About Non-normality

Skew and Kurtosis

- Skew – much easier to deal with
- Kurtosis – less serious anyway

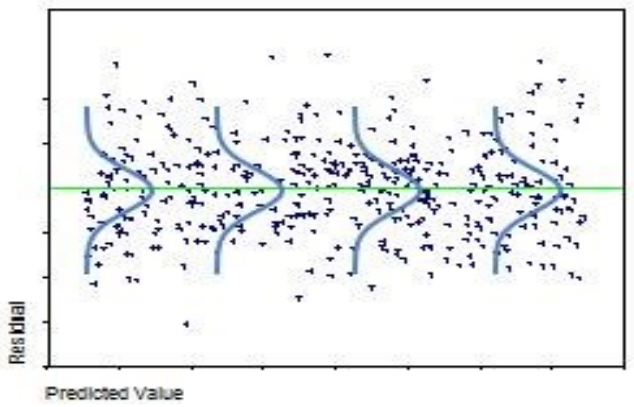
Transform data

- removes skew
- positive skew – log transform
- negative skew – square

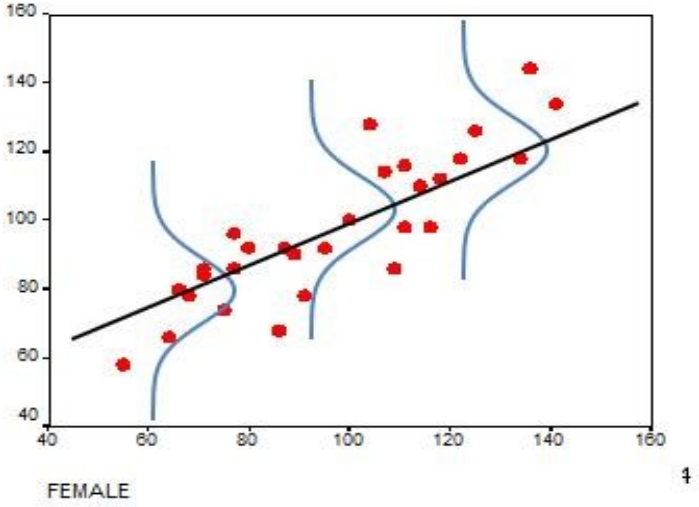
Assumption 2

Heteroscedasticity

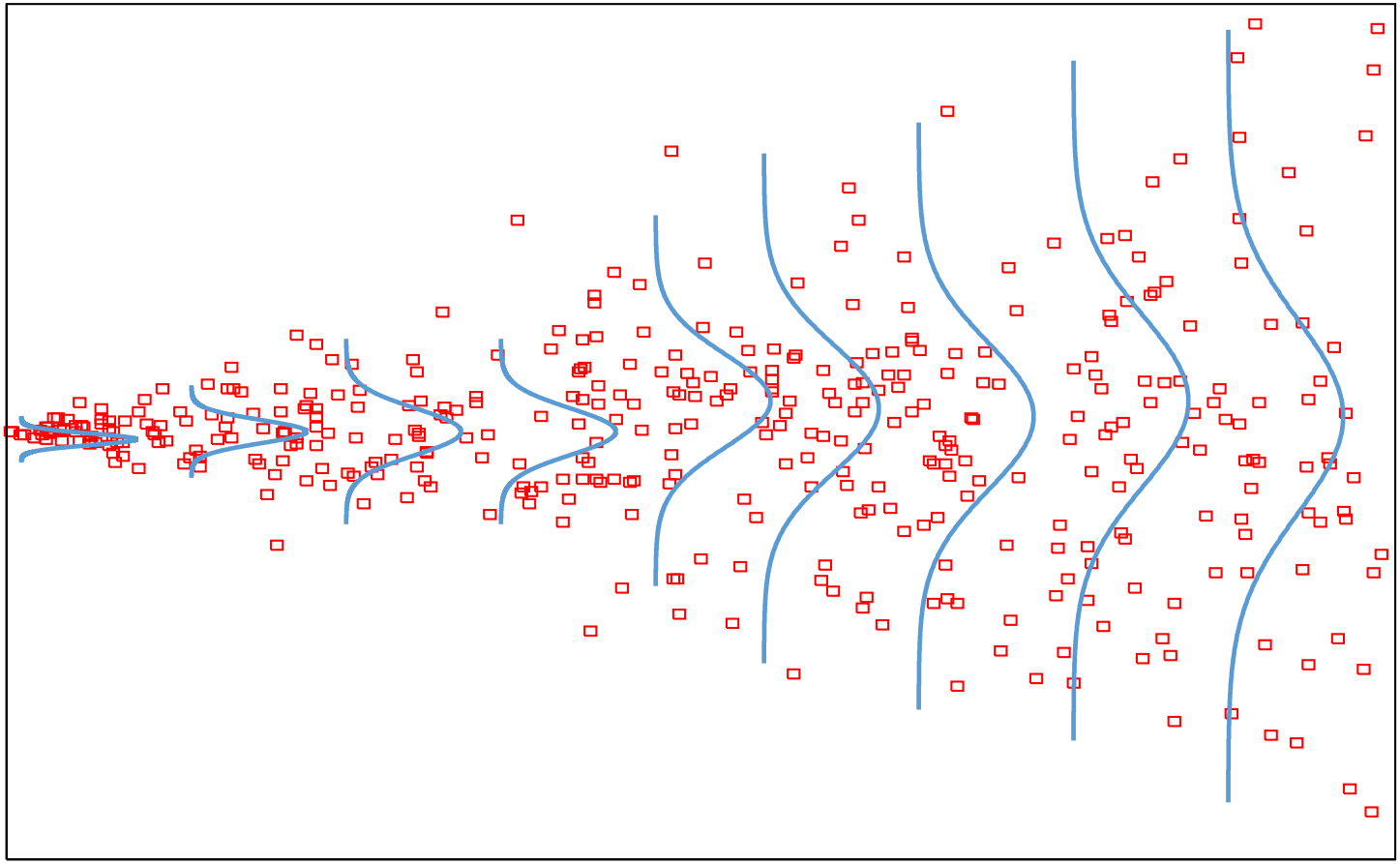
- The word Heteroscedasticity comes from Heteros -> Different
- Scedasticity -> Conditional Variance i.e. Variation in residuals is independent of X
- So if there is no Heteroscedasticity then there is Homoscedasticity i.e. Uniform Variance
- Testing of Heteroscedasticity is done by White's Test



Homogenity of variance

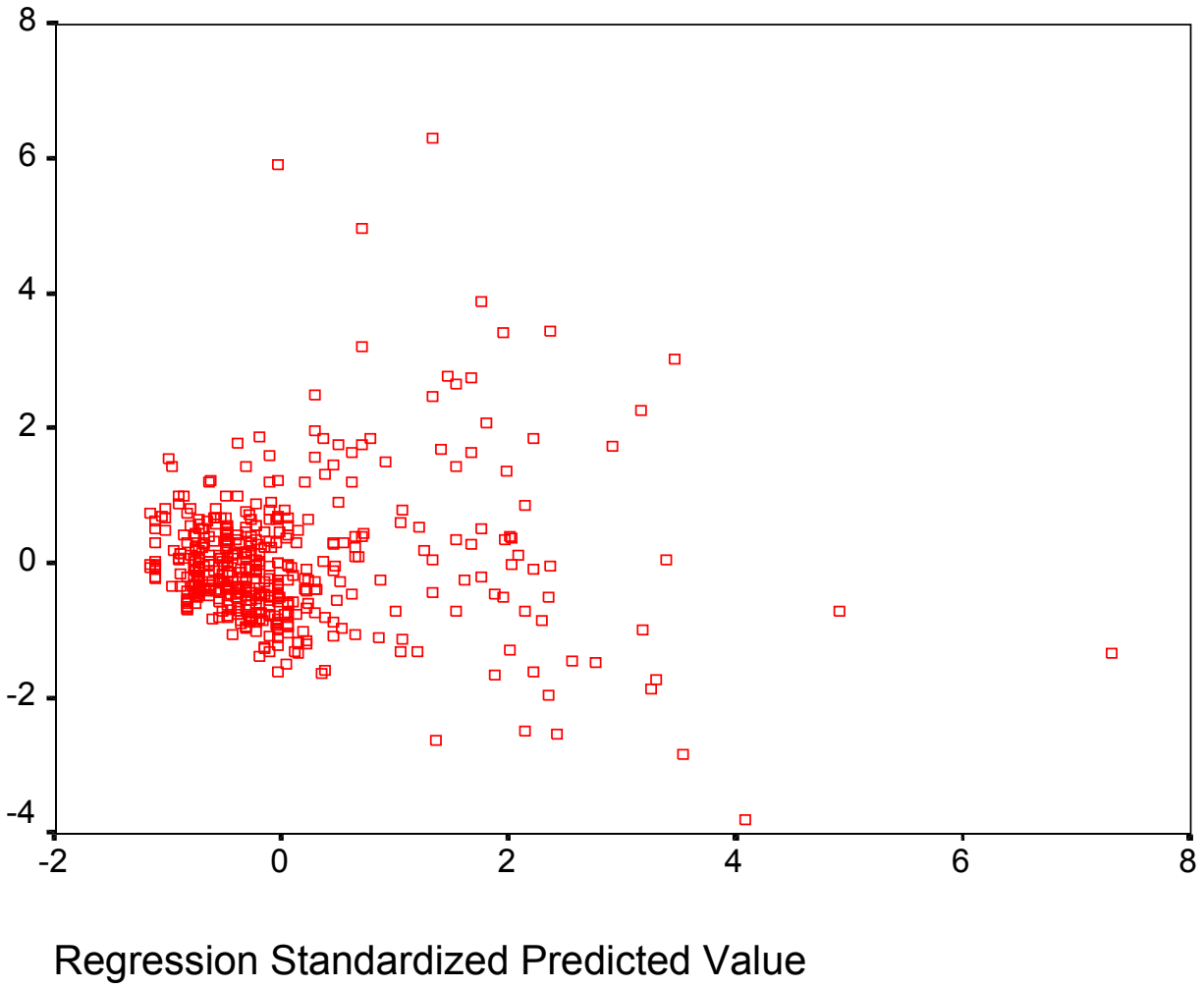


Heteroscedasticity (Contd.)



Predicted Value

Plot Of Pred And Res



Assumption 3

Additivity

- The assumption is that the equation consists of Y, expressed as a sum of X1 , X2 , X3 ….so on and not as Y, a product of X's
- But in many cases that may not be true
  - Gender impacts purchase
  - Similarly age also impacts purchase
  - But age combined with gender may not only add but also may compound

Eg.
Wife purchases separately and the husband purchases separately.
But when they go shopping together, purchase behavior will be different – eg. shopping for kid or household items

- This leads to a specification error, if we continue to add X

Assumption 4

Linearity

• Relationships between variables should be linear best represented by a straight line

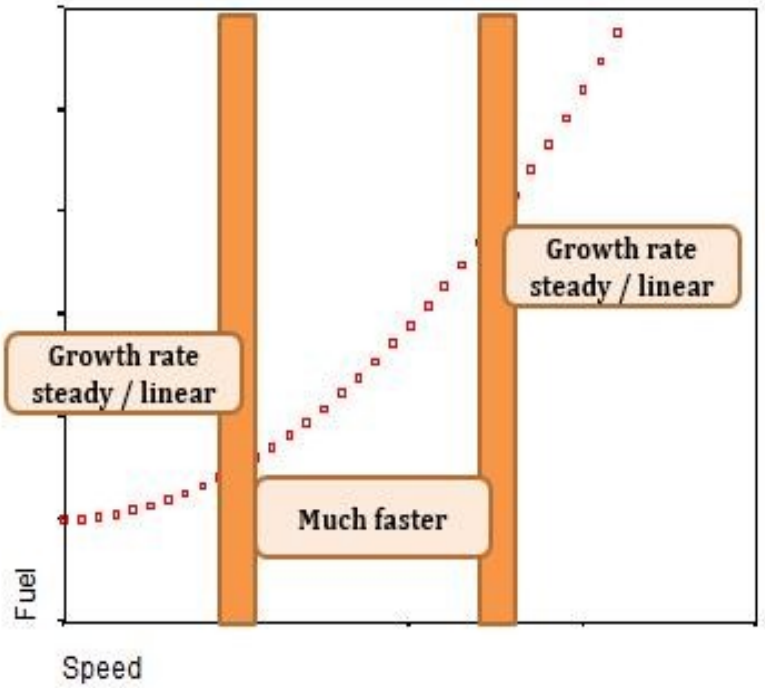
• Some times that may not be case, for e.g., in economics there is a boom in the economy and demand is not gradual but exponential

- In such scenarios , when a model is built in a linear manner the R-square value is very low
- Another measure is the Adjusted R-square which is also low
- Residual plots also help in detecting non-linear relationships

E.g. We can predict the number of flights to metros that will be booked during Diwali etc. but something unpredictable or unnatural like Nepal earthquake will cause your predictions to go for a toss.

Linearity (Contd.)

- Relationship between speed of travel and fuel used



Linearity (Contd.)

- R2 = 0.938
  - looks pretty good
  - We know the speed and hence make a good prediction of fuel

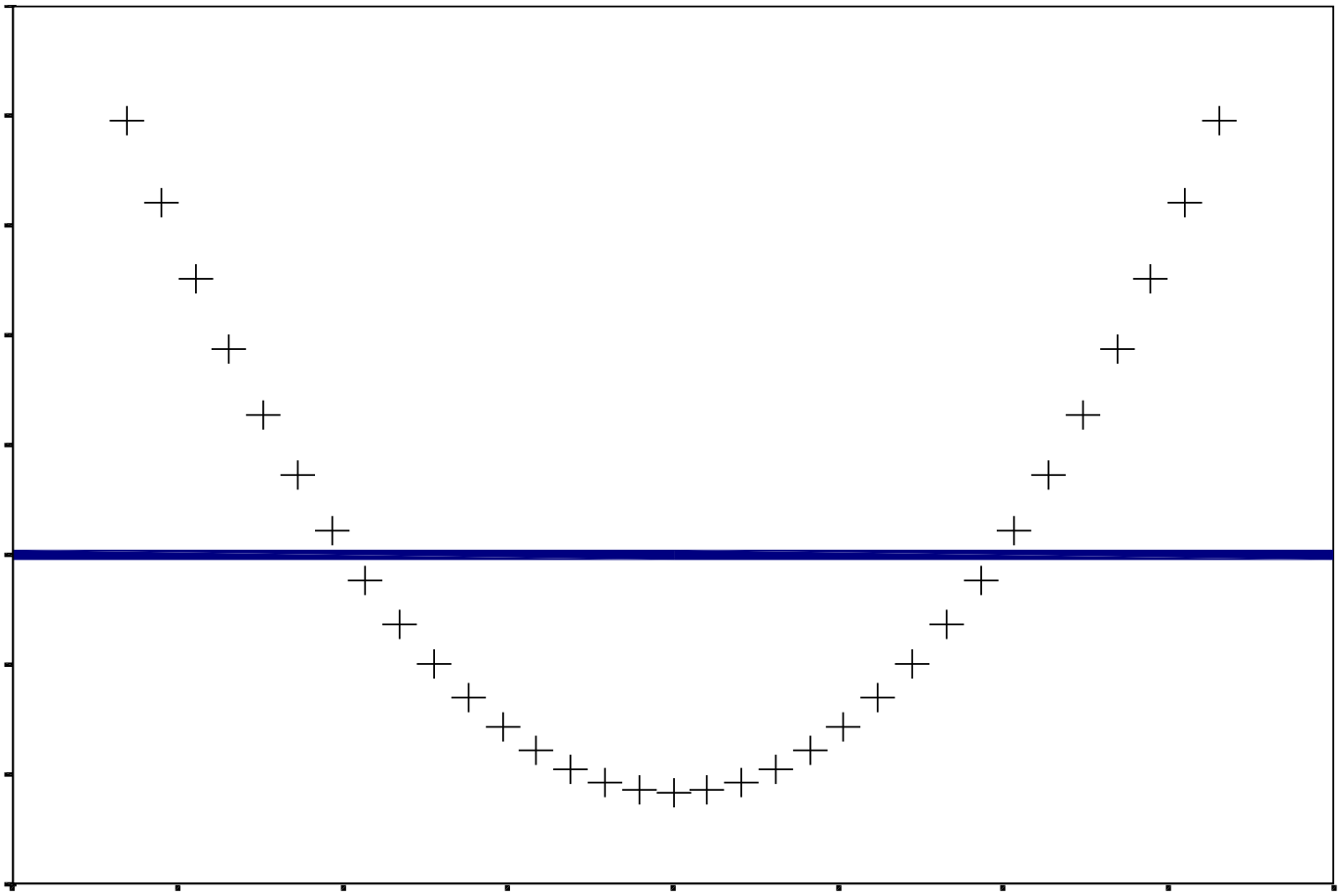
BUT

- look at the chart
- if we know speed, we can make a perfect prediction of fuel used
- R2 should be 1.00

Detecting Non-Linearity

- Residual plot
  - is just like heteroscedasticity
  - This is very clear from the example

Residual Plot



Linearity: A Case Of Additivity

- Linearity = additivity along the range of the IV

E.g. Martin rides his bike harder

- Increase in speed depends on current speed
- Not additive, multiplicative

Assumption 5

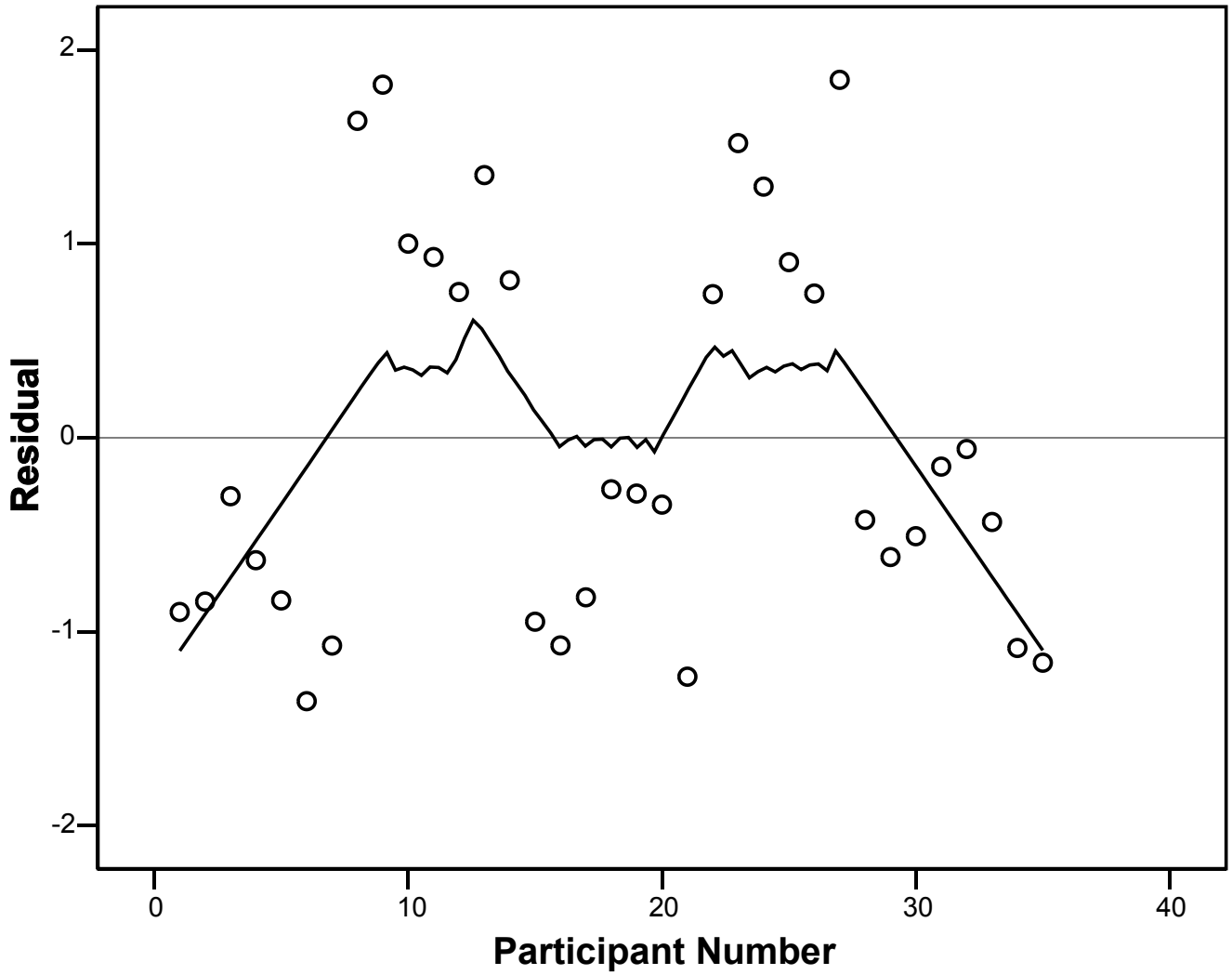
Independence Assumption

- The earlier value of residual should be a function of the next value in the series
- All cases should be independent of one another
  - knowing the value of one case will revealanything about the value of other cases
- To test this, Durbin Watson Test is used

E.g.: You feeling hungry is generally determined by when you had your last meal. But if you had spicy food late in the night then factors like acidity or an upset stomach will disrupt that normal cycle.
Hence it is important to remember that past behavior alone cannot predict future behavior.

- Graphically, Residual Plots are useful
- Like wise data should not have time component in-grained into it

Residual Plots



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

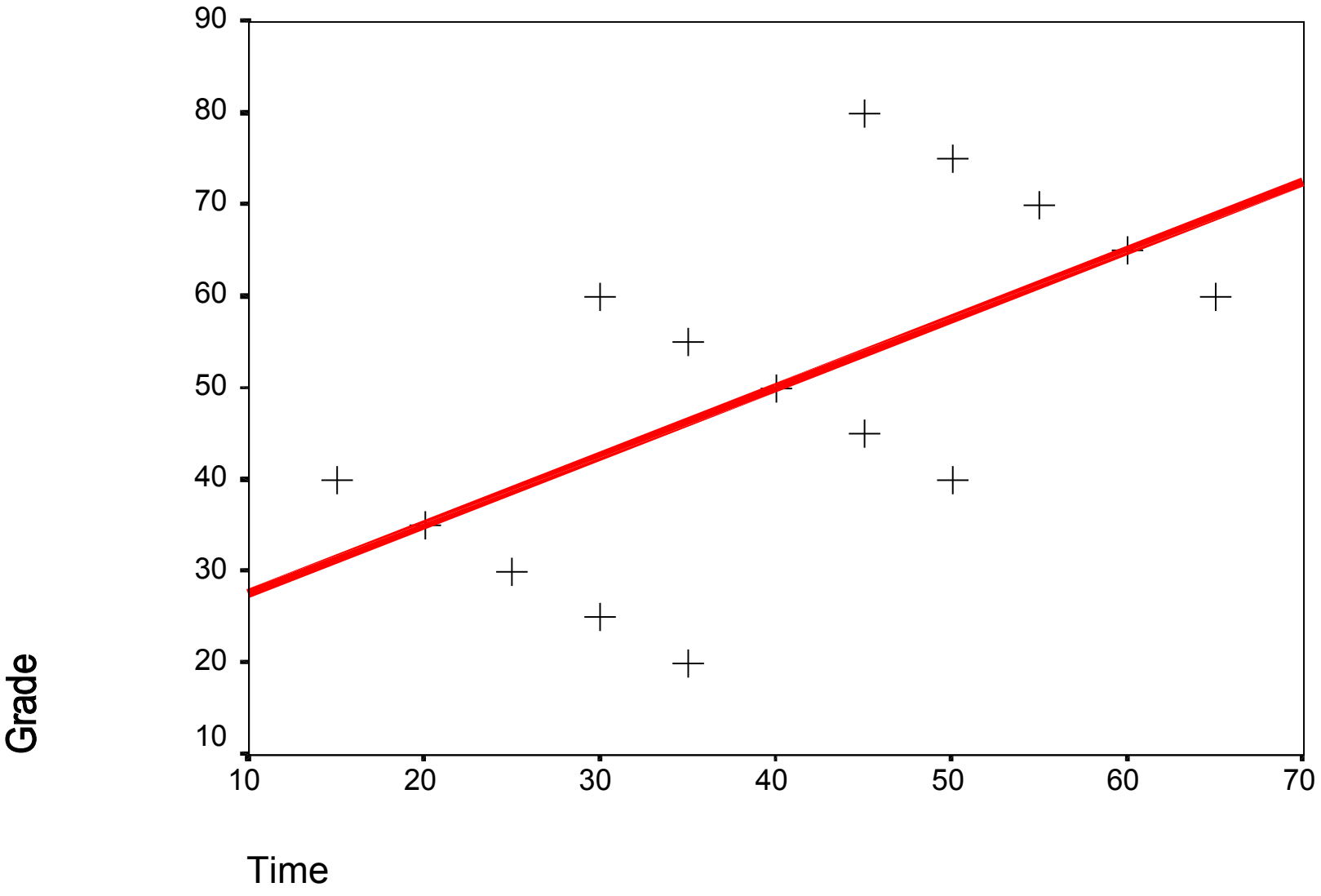
Why Does It Matter?

- Standard errors can be wrong
  - therefore significance tests can be wrong
- Parameter estimates can be wrong
  - from positive to negative

Example

- students write an exam (on statistics)
- choose one of three questions
  - IV: time
  - DV: grade

Result, With Line Of Best Fit



What The Result Shows

- Result shows that
  - people who spent long time to write the exam, achieve better grades

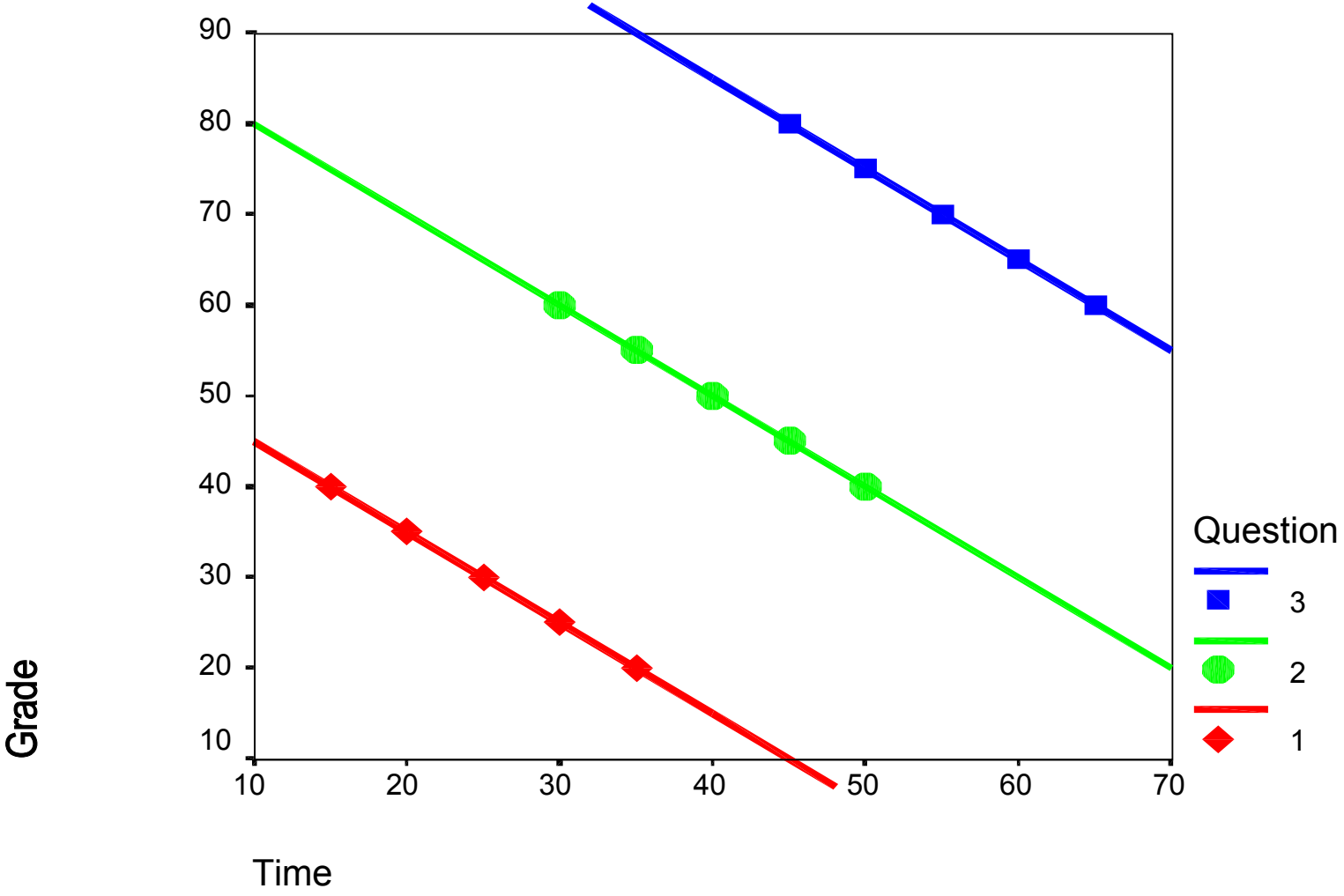
BUT

- we haven't considered which question people answered
- we might have violated the independence assumption
  - DV will be auto correlated

- Look again
  - with questions marked

What The Result Shows

- Notice the difference



Some Difference

- Now, people who spent longer time to write exam, got lower grades
  - questions differed in difficulty
  - work hard to get better grade

- Very difficult to analyse
  - need multilevel models

Assumption 6

Uncorrelated With The Error Term

- A curious assumption
  - by definition, the residuals are uncorrelated with the independent variables (try it, if you like)

- It concerns the DV
  - must have no effect (when the IVs have been removed)
  - on the DV

Uncorrelated With The Error Term (Contd.)

- Problem in economics
  - Demand increases supply
  - Supply increases wages
  - Higher wages increase demand

- OLS estimates will be biased in this case
  - need a different estimation procedure
  - two-stage least squares
    - simultaneous equation modelling

Assumption 7

No Perfect Multicollinearity

- IVs must not be linear functions of one another
  - matrix of correlations of IVs is not positively definite
  - cannot be inverted
  - hence parameter values will show extreme high importance
  - hence analysis cannot proceed

Examples

- Income & income tax

E.g.. Electricity/gas bills

The bills are payable according to the slabs. Beyond a threshold, rate payable increases.

Assumption 8

Mean Of The Error Term = 0

- The assumption is, Y should get impacted only by X values and not anything else
- Hence residuals should have minimal impact
- Thus in case of additive assumption, the sum of residuals is zero
- Hence Mean of the residuals = 0
- That is where the constant comes into picture
  - if the mean of the error term deviates from zero, the constant compensates it

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

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

Next Class
Model Selection in R

- Fitting the Model
- Diagnostic Plots
- Comparing Models
- Cross Validation
- Variable Selection
- Relative Importance
- Dummy Variable


Q & A time