Assumptions in linear regression analysis

Tuan V. Nguyen
Garvan Institute of Medical Research
Sydney, Australia

The linear regression model

- **Simple linear regression model**
- ***Y*** response variable, dependent variable, continuous variable
- ***X*** predictor variable, independent variable, can be continuous or categorical

The linear regression model

- The statement:

$$Y = \alpha + \beta X + \varepsilon$$

α : intercept

β : slope / gradient

ε : random error – the variatio in Y for each X value

Assumptions of regression model

- **Linearity**: The relationship between X and Y is linear
- **Normality**: For any value of X, Y is normally distributed
- **Homoscedasticity**: The variance of residual is the same for any value of X
- **Independence**: Observations are independent of each other
- *X* does not have random error

Random error ε: Normal distribution with mean 0, constant variance,

$$\varepsilon \sim N(0, \sigma^2)$$

Using R

- The linear regression model:

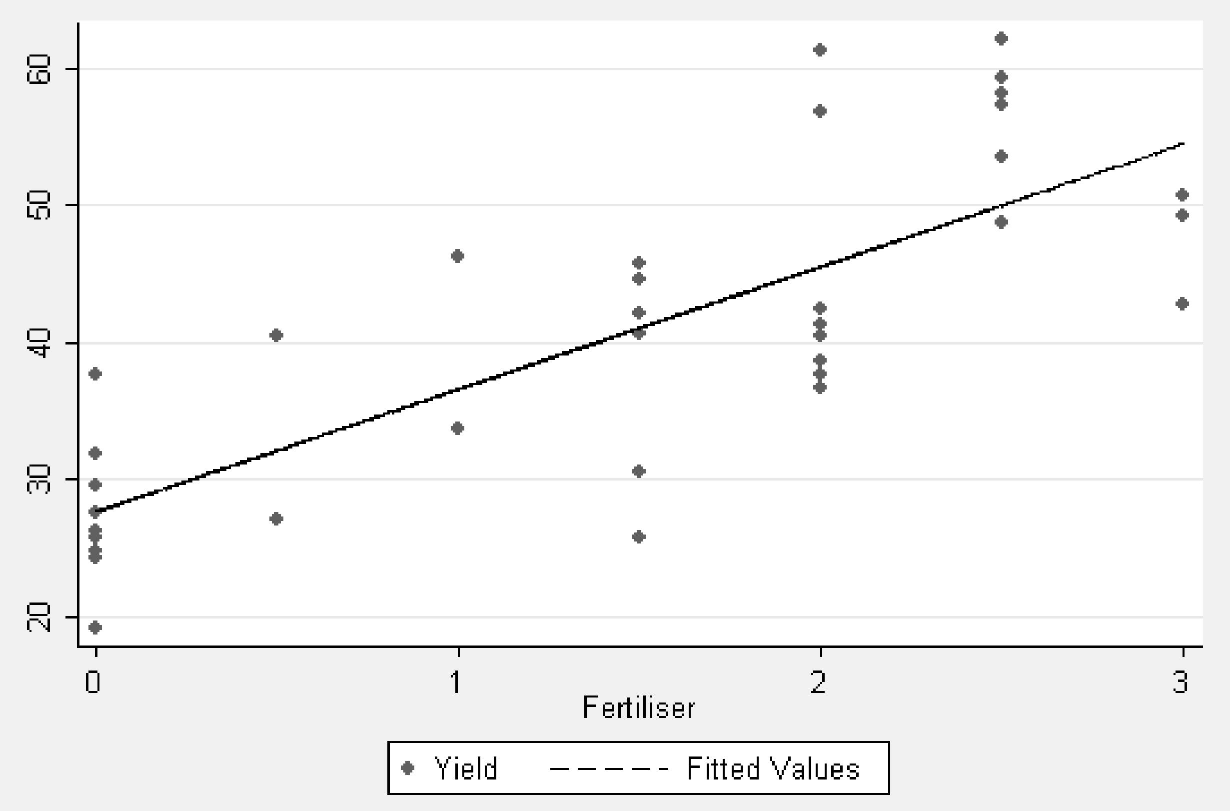
$$Y = \alpha + \beta * X + \varepsilon$$

- R codes (using function lm):

```
lm(y ~ x)
```

How to check assumptions

- Conduct a residual analysis
- Residuals are deviations of *observed values* from *model fitted values*



Residuals

- The linear regression model:

$$Y = \alpha + \beta * X + \varepsilon$$

- Predicted (fitted) values

$$\hat{Y} = a + b * X$$

- Residuals

$$e = \hat{Y} - Y$$

Residuals using R

- Predicted (fitted) values

$$\hat{Y} = a + b*X$$

- Residuals

$$e = \hat{Y} - Y$$

- Using R

```
m = lm(y ~ x)
pred = predict(m)
e = resid(m)
```

Residual plots

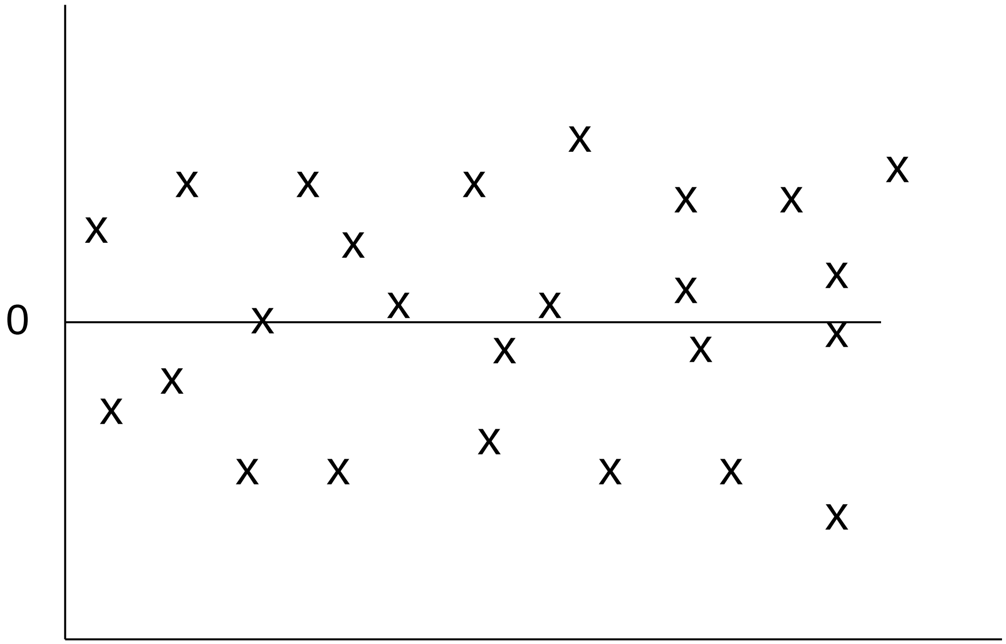
- A histogram of the residuals (provided there are enough observations) can be used to check for normality

- A *normal probability plot* of residuals. A straight line plot suggests that the normality assumption is reasonable

Residual plots

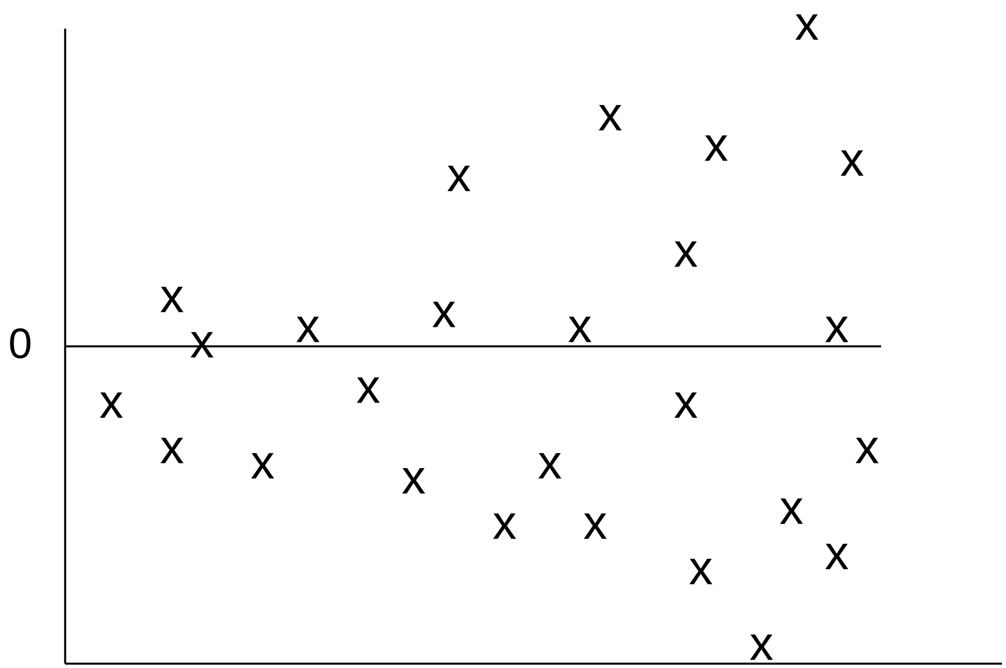
- Plot of residuals against fitted values ($\hat{y}$)
  - Detect variance homogeneity assumption
  - Identify potential outliers

Residuals (y) vs fitted values (x)



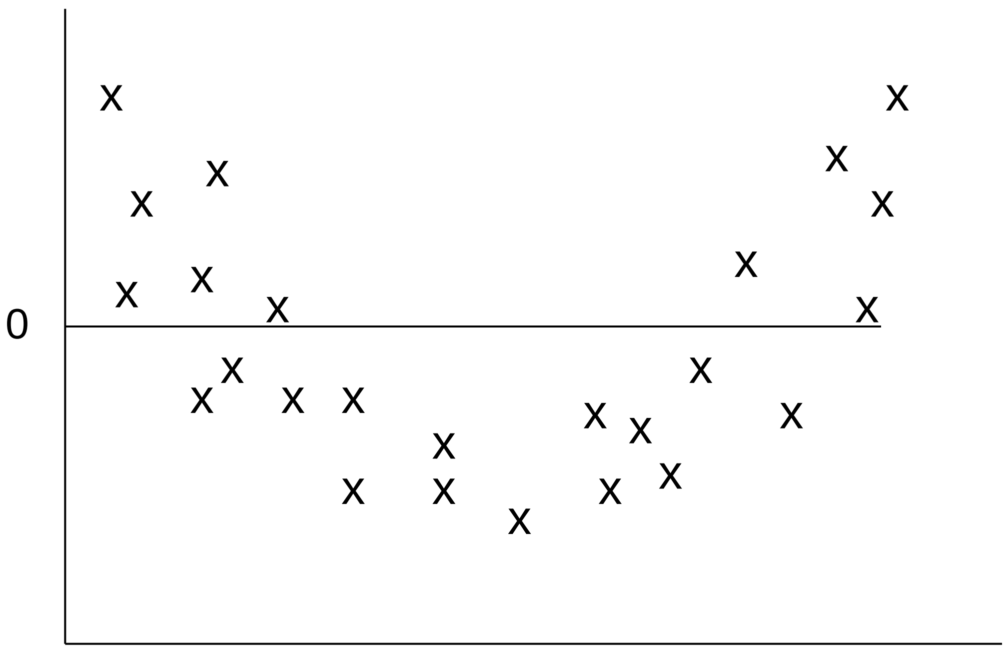
A random scatter as above is good. It shows no obvious departures of the variance homogeneity assumption.

Residuals (y) vs fitted values (x)

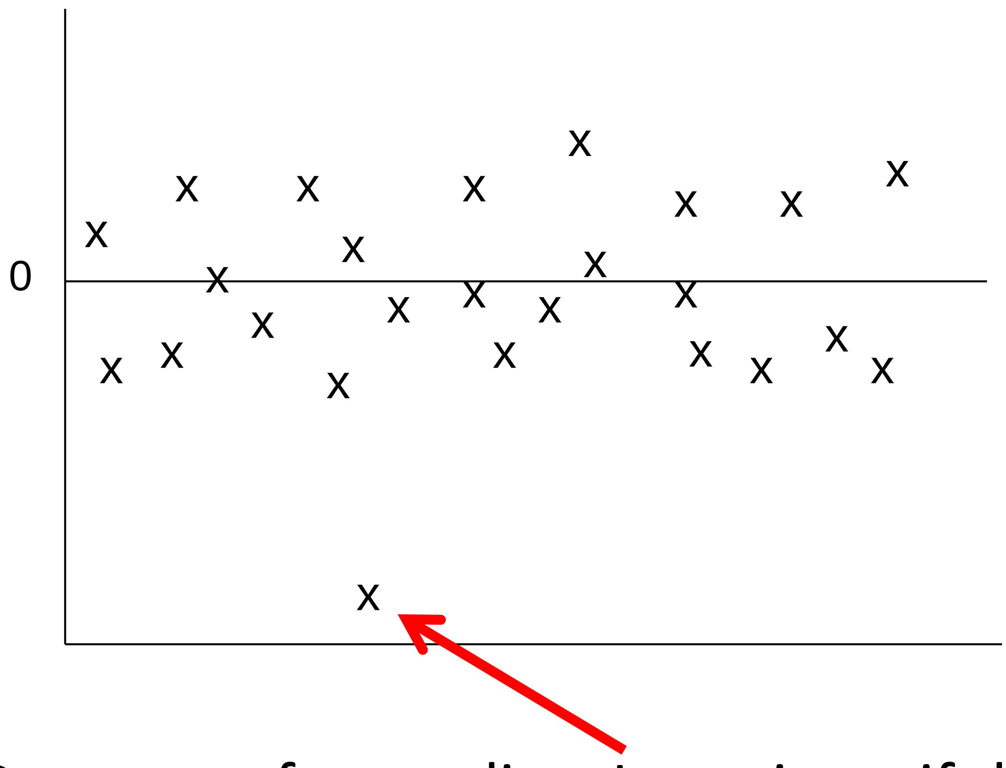


Variance increases with increasing *X*

Could try a $\log_e(y)$ transformation



Lack of linearity. Pattern indicates an incorrect model - probably due to a missing squared term.



Presence of an outlier. Investigate if there is a reason for this odd-point

Stabilization of variance

Some typical transformations are:

- taking logs (useful when there is skewness)
- square root transformation
- reciprocal transformation

Sometimes theoretical grounds will determine the transformation to use

Fix the non-independence problem

- Problem of design

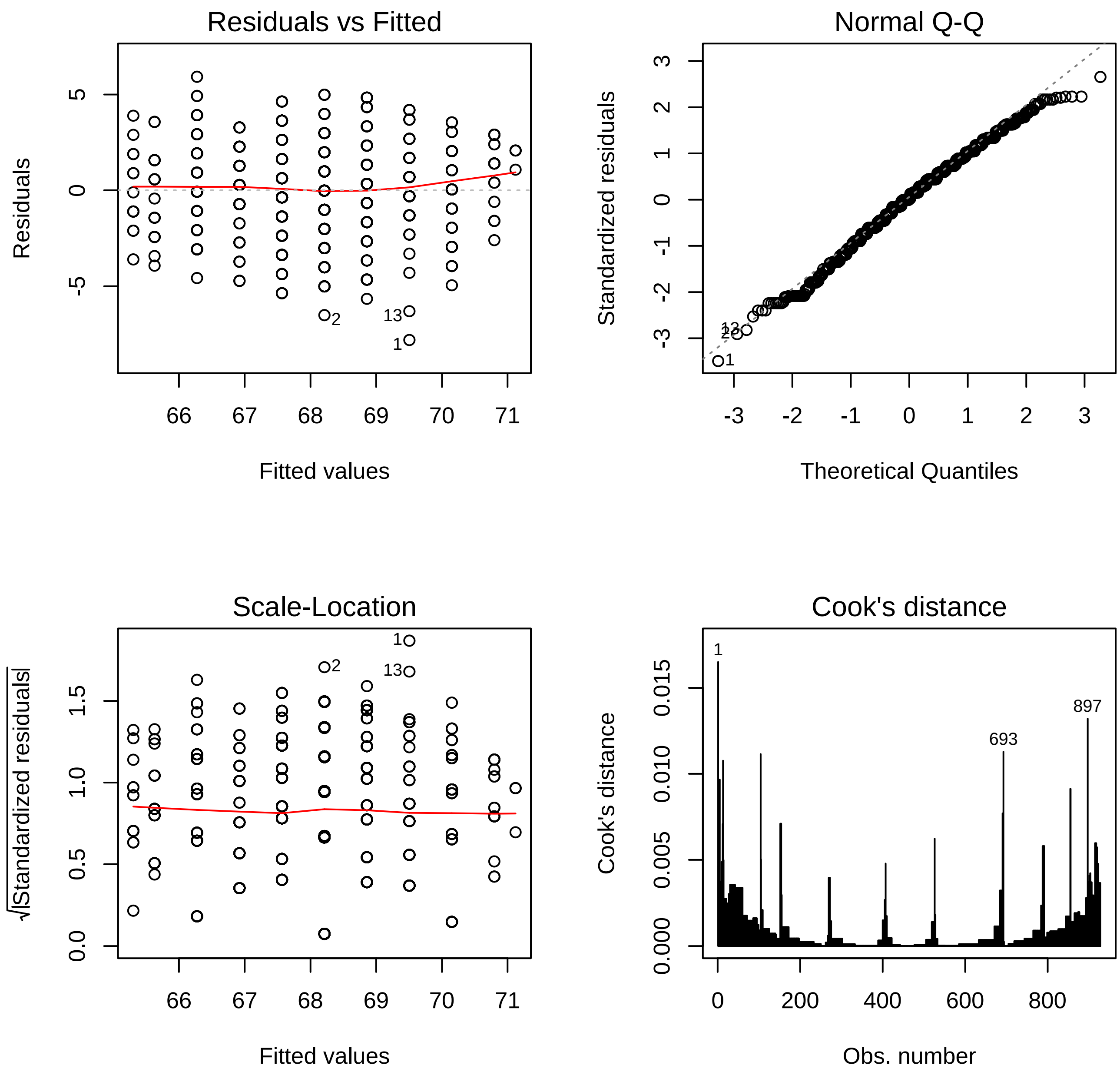
- Techniques similar to those used in *time series analysis* or *analysis of repeated measurements* data may be more appropriate

Other issues of regression analysis

- **Outliers**: observations have large residuals
- **Leverage points**: observations have X values that are far away from the mean of X
- **Influential observations**: observations that change the slope of the line
- *Outliers may or may not be influential points*

Galton's data

```
galton = read.csv("~/Google Drive/Garvan Lectures 2014/Datasets and Teaching Materials/Galton data.csv", header=T)

attach(galton)

m = lm(child ~ parent, data=galton)

# diagnostic plots

par(mfrow=c(2,2))

plot(m, which=1:4)
```



Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Cook's distance
Cook's distance
Obs. number