Regression

Dr. Arief Gusnanto

Department of Statistics
University of Leeds

IMforFUTURE Network meeting

Leeds 15-18 October 2018

Outline

Introduction

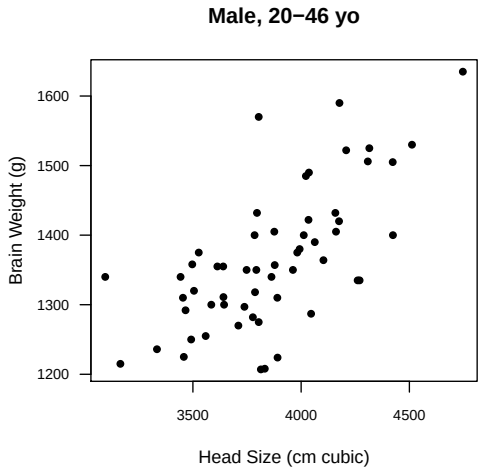
- Regression is the main workhorse in statistical modelling
- It postulates a relationship between a response variable (or a function of response variable) and some predictor variables
- Purpose: modelling and/or prediction

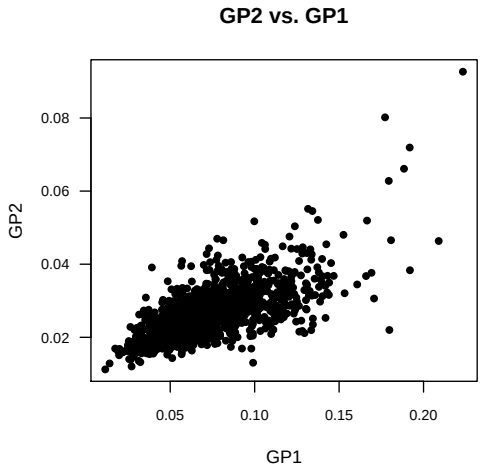
Introduction

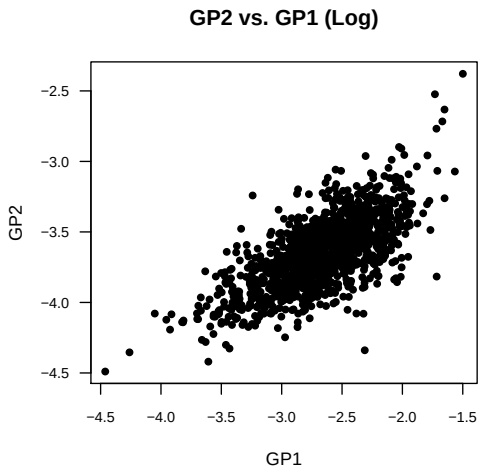
- In the previous training, we have learned about different types of data
- Linear Regression: Continuous response variable
- Generalised Linear Models: Non-normal response variable
  - Logistic regression: Binary
  - Poisson regression: Count
- Survival analysis: time-to-event data
  - Cox regression: modelling hazard

Simple linear regression

- We are interested in the *linear* relationship between a *continuous* response variable y and a predictor variable x
- First steps: explore the data, scatter plots (!)







Simple linear regression

- Suppose we have the i-th pair of data (y_i, x_i), $i = 1, 2, \ldots, n$
- Consider a *response variable* y_i
- We wish to model y_i as a function of *predictor* x_i

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

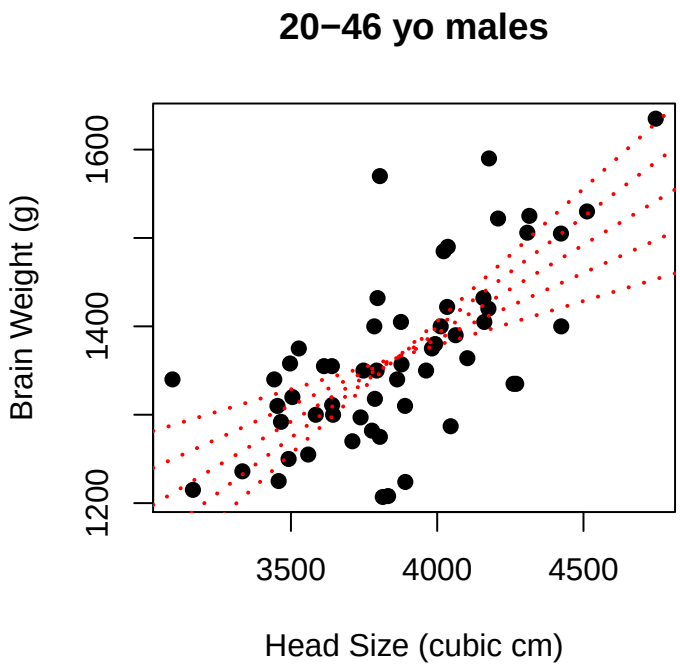
where

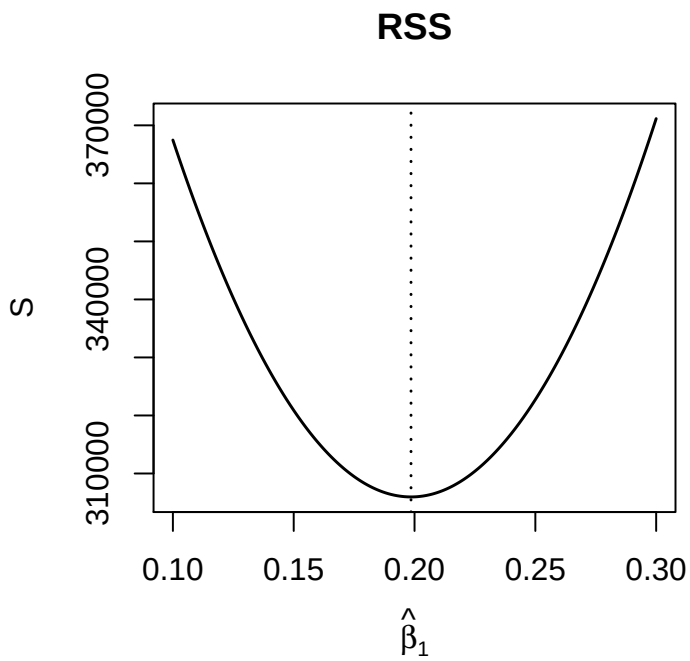
  - β_0 is an intercept or 'constant'
  - β_1 is a 'slope' – it determines how 'fast' any changes in x would affect the mean of y
  - ε_i is the error term

- The error term is assumed to follow a normal distribution with mean zero and a variance

Simple linear regression: parameter estimation

- Estimation of β_0 and β_1 is based on the principle of minimising the sum of squared residuals
- Residuals are the distance between the points and the linear regression line





Simple linear regression

- In the head size data, 20-46 y.o. males, suppose we model

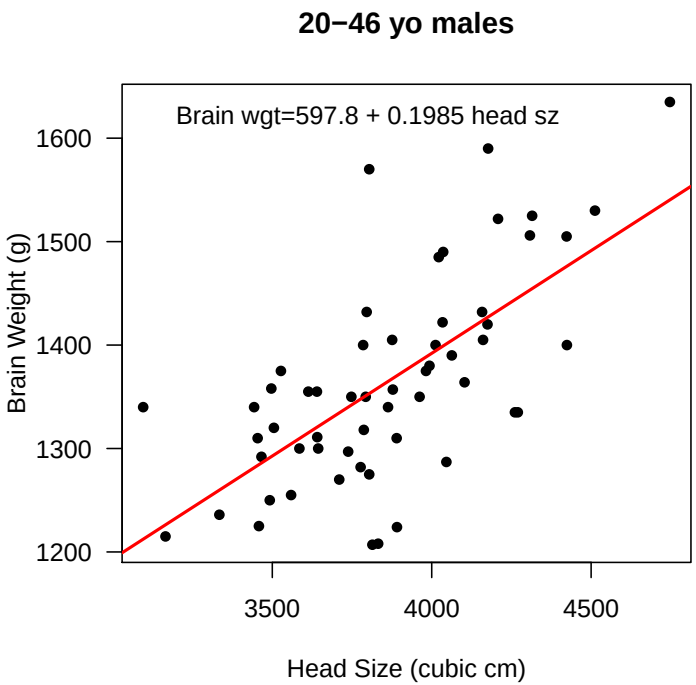
$$i\text{-th brain weight} = \beta_0 + \beta_1 \times i\text{-th head size} + i\text{-th } \varepsilon$$

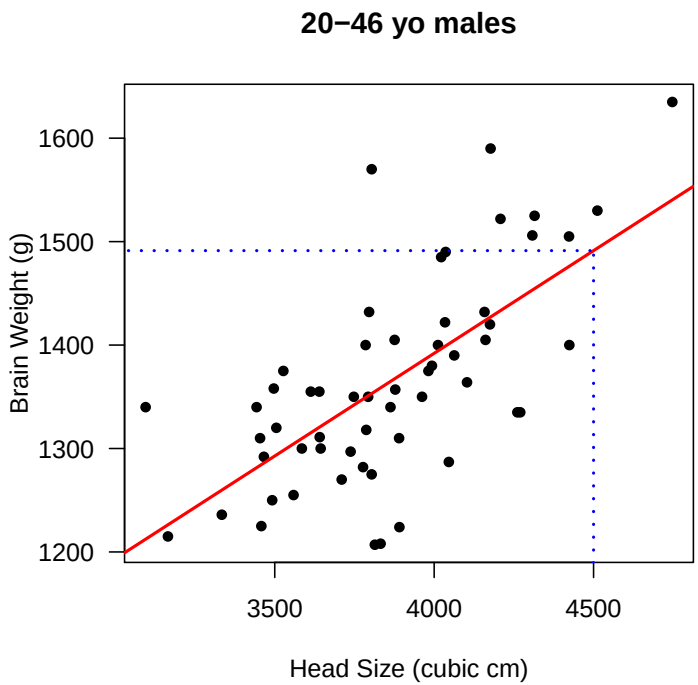
- It can be shown that the (resulting) fitted regression line is given by

$$\text{brain weight} = 597.8 + 0.1985 \times \text{head size}$$

- We can predict the mean brain weight at head size 4,500cm^3 (say) as

$$597.8 + 0.1985 \times 4500$$





- The estimation of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ is done *analytically* and not on "trial and error"
- The residuals are set as a function of the parameters and solve
- Interpretation of $\widehat{\beta}_0$: the expected brain weight for head size=0 (it is not sensible in this case, but may be sensible in other cases)
- Interpretation of $\widehat{\beta}_1$: the expected *increase* in the mean of brain weight for an increase in one unit (cm^3) of head size
  - You may also consider it as a 100 cm^3 increase of head size is expected to increase the mean of brain weight by $100\widehat{\beta}_1$

Simple linear regression: Inference

- From the brain weight data, we have $\widehat{\beta}_1 = 0.1985$
- Important question: Is it large enough? (from what value?)
- The modelling is meaningful if head size is significantly has an effect on brain weight
  - No effect $\rightarrow \beta_1 = 0$
- In our context, we are interested in $H_0 : \widehat{\beta}_1 = 0$
  - In other contexts, you may have a prespecified value δ^0
  - I.e. $H_0 : \beta_1 = \delta^0$

- The test statistic is given by

$$t = \frac{\widehat{\beta}_1 - \delta^0}{\text{se}(\widehat{\beta}_1)}$$

  where $\text{se}(\widehat{\beta}_1)$ is the standard error of $\widehat{\beta}_1$
- $\text{se}(\widehat{\beta}_1)$ is a measure of uncertainty in the estimation of β_1
- It can be shown that the standard error depends on the error variance
- ... and error variance depends on the residual sum of squares

Simple linear regression: Inference

- Under the null hypothesis, the test statistic t would follow a t-distribution with $n-2$ degrees of freedom (in the brain weight data)
- That is *residual* degrees of freedom
- The degrees of freedom of *fit* is 2 because there are two model parameters to estimate (β_0 and β_1)
- We reject H_0 if the p-value associated with the test is less than the significance level α

Simple linear regression: Example

- In R, you can use the command `lm` to run linear regression, which produce the following (edited) output

```
> res <- lm(brain.weight~head.size, data=dat)
> summary(res)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 597.83865  114.96426   5.200 3.02e-06 ***
head.size     0.19854    0.02964   6.699 1.17e-08 ***

Residual standard error: 74.58 on 55 degrees of freedom
Multiple R-squared:  0.4493,Adjusted R-squared:  0.4393
F-statistic: 44.88 on 1 and 55 DF,  p-value: 1.172e-08
```

Simple linear regression: Example

Few points to note

1. The column 'Estimate' contains the parameter estimates. From the R output, we have $\widehat{\beta}_0 = 597.8$ and $\widehat{\beta}_1 = 0.1985$
2. The column 'Std. Error' contains their respective standard errors (of the parameter estimates)
3. The column 't value' contains their respective t statistic **when** $\delta^0 = 0$ **in the hypothesis testing**. I.e. $t = \widehat{\beta}/\text{se}(\widehat{\beta})$
4. The column 'Pr($>|t|$)' contains their respective p-values
5. The 'Residual standard error' is the estimated standard deviation of error (square root of error variance)
6. The degrees of freedom in the output is the residual df (**not** df of fit)
7. The F-statistic and its p-value (at the bottom of the output) will be covered later

Simple Linear Regression: Interpretation

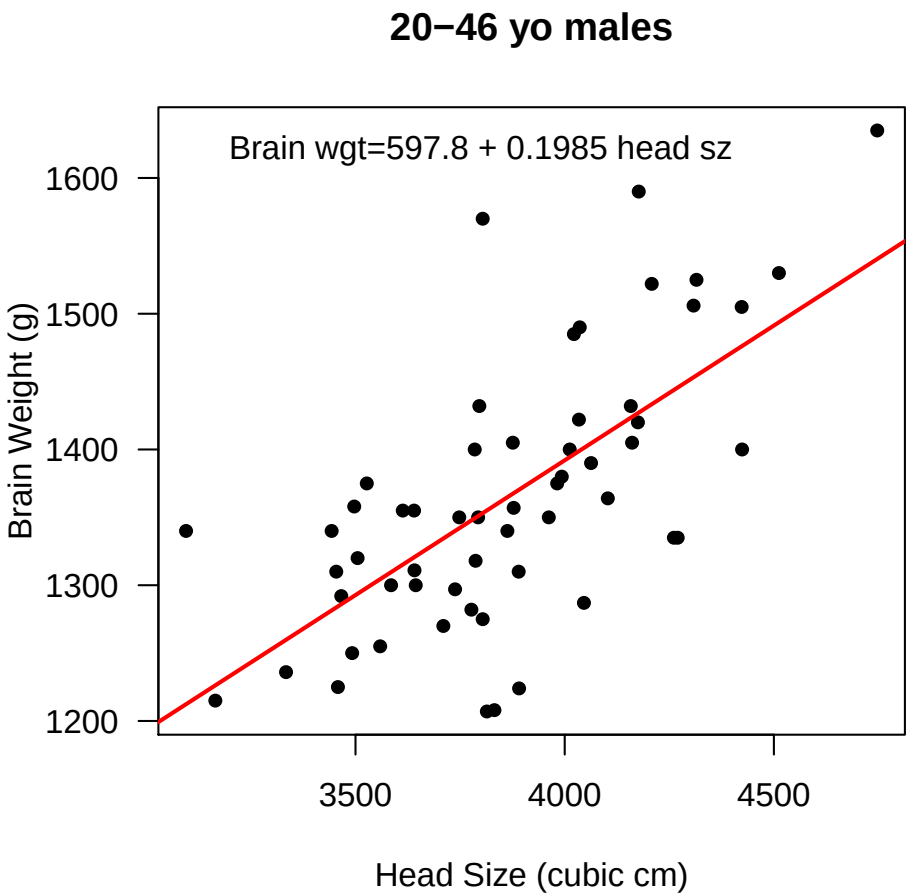
- The brain weight is significantly associated with head size
  - p-value for the effect of head size is less than 0.05
- For an increase of 1 unit (cm^3) in head size we expect to have an increase of 0.1985 unit (gram) of the brain weight
- Important: We only interpret this **within** the range of the data only

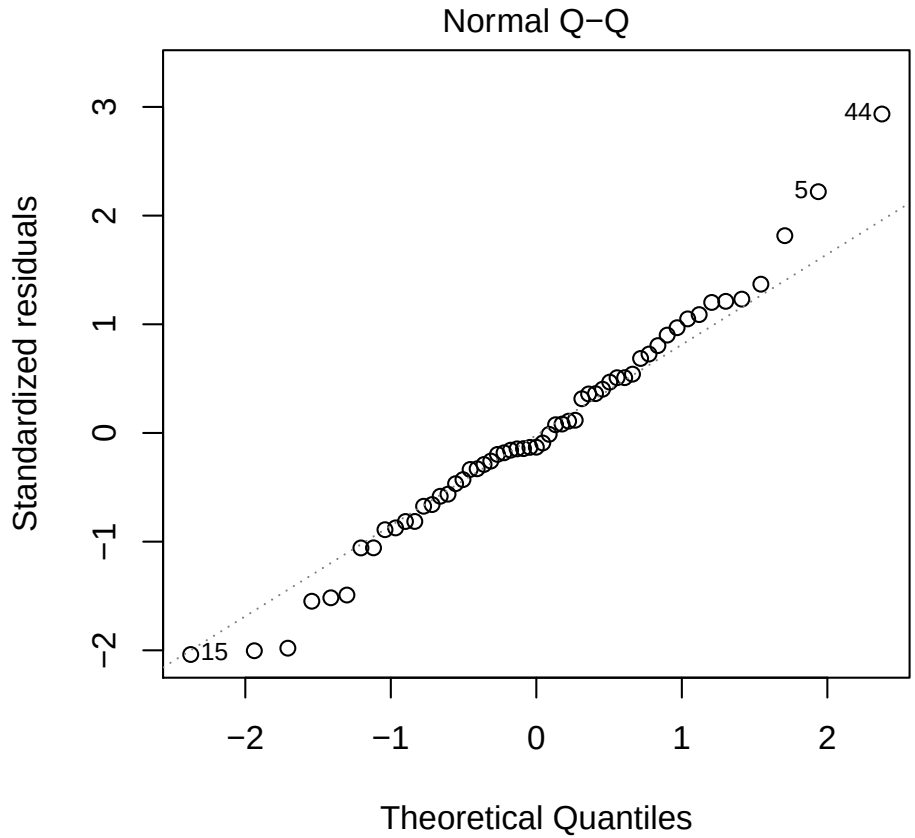
Simple Linear Regression: Model Diagnostics

- Model diagnostics is critical in any modelling – it checks whether the model is appropriate for the data
- Our assumptions in the modelling: error follows a normal distribution independently and the error variance is constant
- We can investigate the residuals to check those assumptions
- To check the normality, we consider the QQ plot
- To check the other assumptions, we can check the plot of residuals against fitted values (the points on the regression line for each data point)

Simple Linear Regression: Model Diagnostics

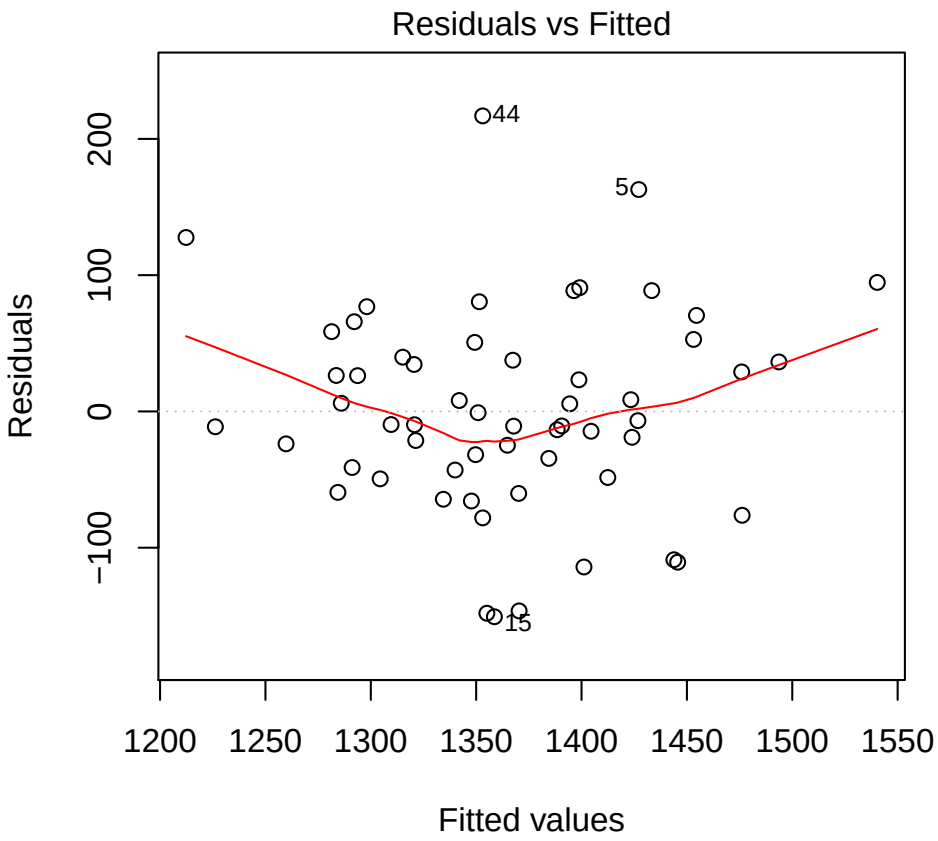
```
plot(res,2)
```

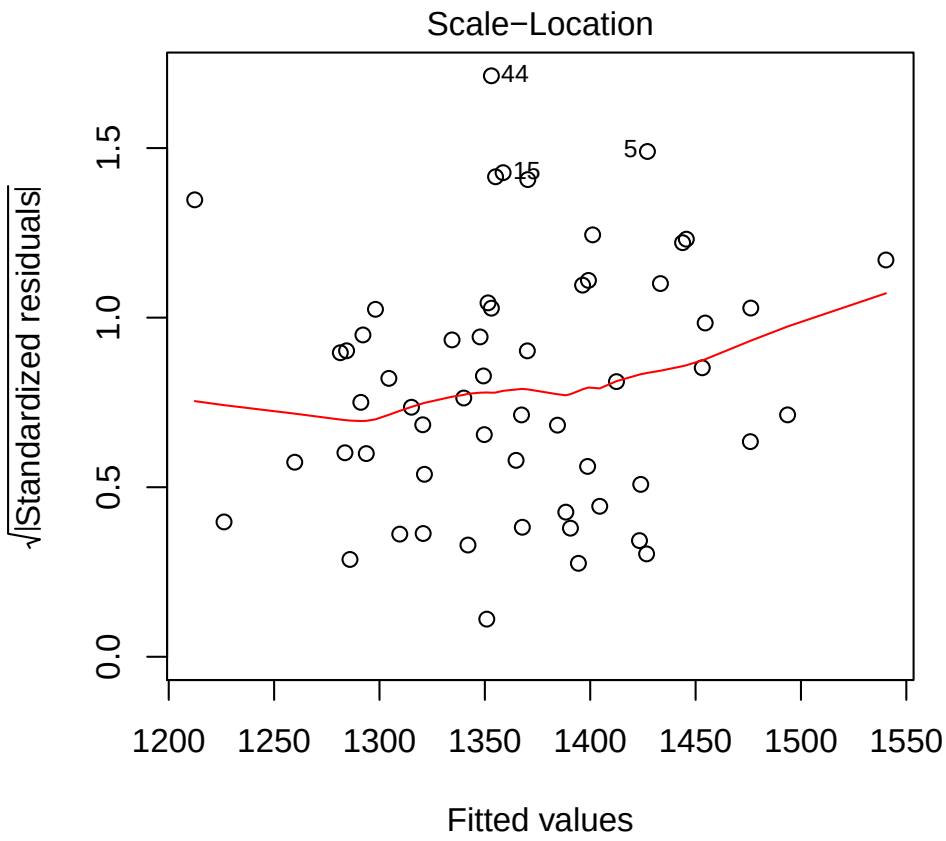




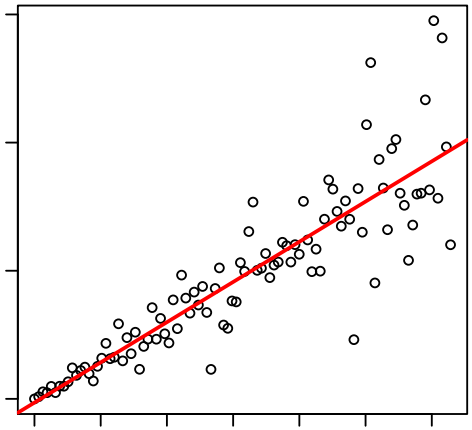
Simple Linear Regression: Model Diagnostics

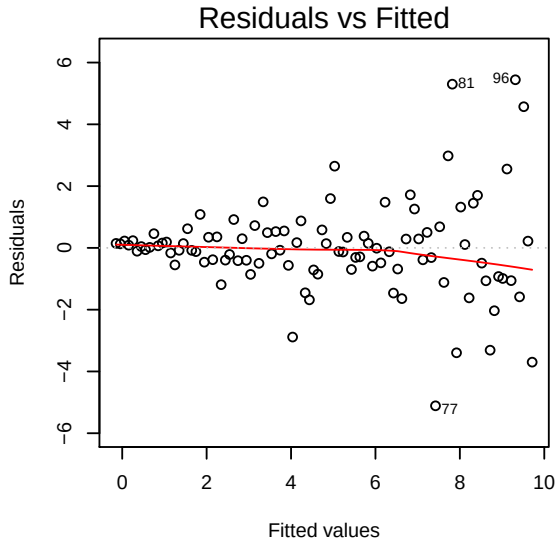
```
plot(res,1)
plot(res,3)
```

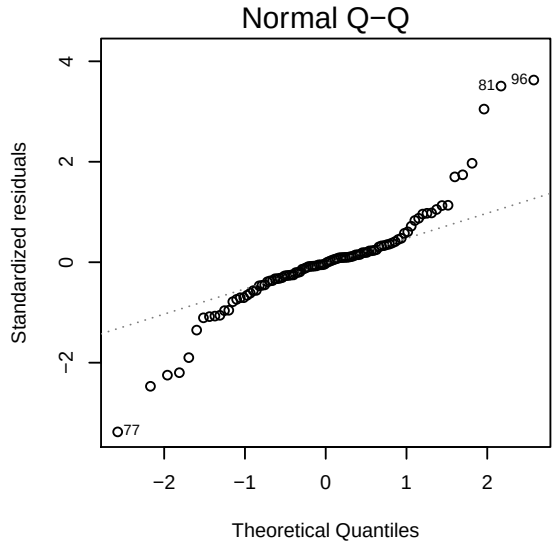


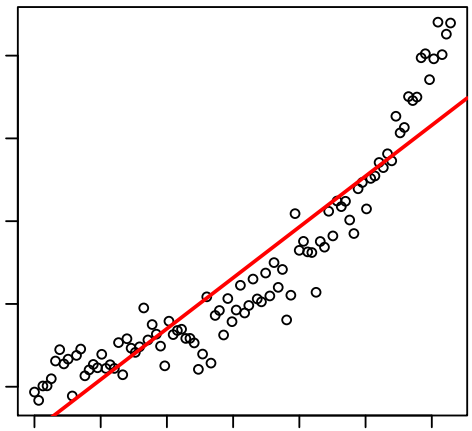


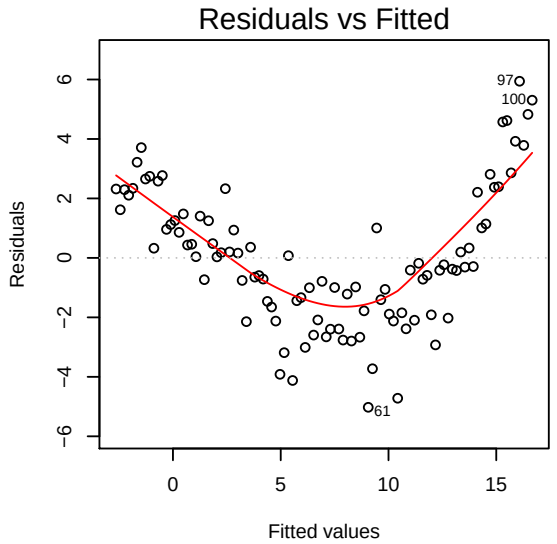
Simple Linear Regression: Model Diagnostics

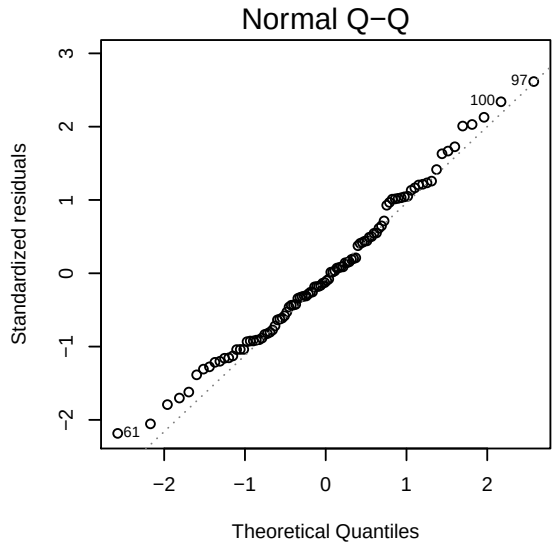












What to do?

- What to do when there is a problem?
- First, always check the **cause** of the problem
- (A well-defined problem is half-solved!)
- If a higher polynomial order is needed, then include quadratic (or possibly cubic) terms
  - It is still a *linear* regression (!)
  - The model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

    can be written as

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 w_i + \varepsilon_i$$

    where $w_i = x_i^2$

What to do?

- If transformation is needed, then perform transformation to the data
  - You may consider transformation to the response variable only or together with the predictor variables
  - Check and recheck (!) - modelling is a 'dynamic' process
- Consider simple (!) transformation
- For example: log (either base e, base 2 or base 10)
  - Usually works best when we have right skewed data with no (or little) zeros
  - Not recommended where many zeros are present in the data

What to do?

- Another example: square root transformation
  - May be considered when the data are also right skewed
  - Suitable when you have many zeros
- Reciprocal transformation $(1/y_i)$
  - When the data is wide spread, right skewed, with more concentration close to zero
- These transformations are generally suitable for many cases
- (For those who are from maths/stats background) You may study about the Box-Cox transformation as a general transformation tool

What to do?

- What if transformation is not enough?
- There is a rich class of models
- For data with increasing/decreasing error variance ('heteroscedastic'), you may consider a *general linear model* (**not** *generalised linear model*)
  - General linear model is still working with continuous response variable
  - The heteroscedasticity is taken into account in the model
- Generalised linear model (GLM) is a class of models where the response variable is not continuous and cannot be assumed to follow a normal distribution
  - Binary response: success/failure, dead/alive
  - Count response: number of items/events per unit time/space
- Any question?

Multiple linear regression

- In multiple linear regression, we have several (more than one) predictors in linear regression
- Suppose we have the i-th point of data $(y_i, x_{1i}, \ldots, x_{pi})$, $i = 1, 2, \ldots, n$
- Consider a *response variable* y_i
- We wish to model y_i as a function of *predictor* x_i

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} \varepsilon_i$$

- The error term is assumed to follow a normal distribution with mean zero and a variance
- Note: we have p predictor variables and $p + 1$ model parameters (!)
  - Don't forget β_0

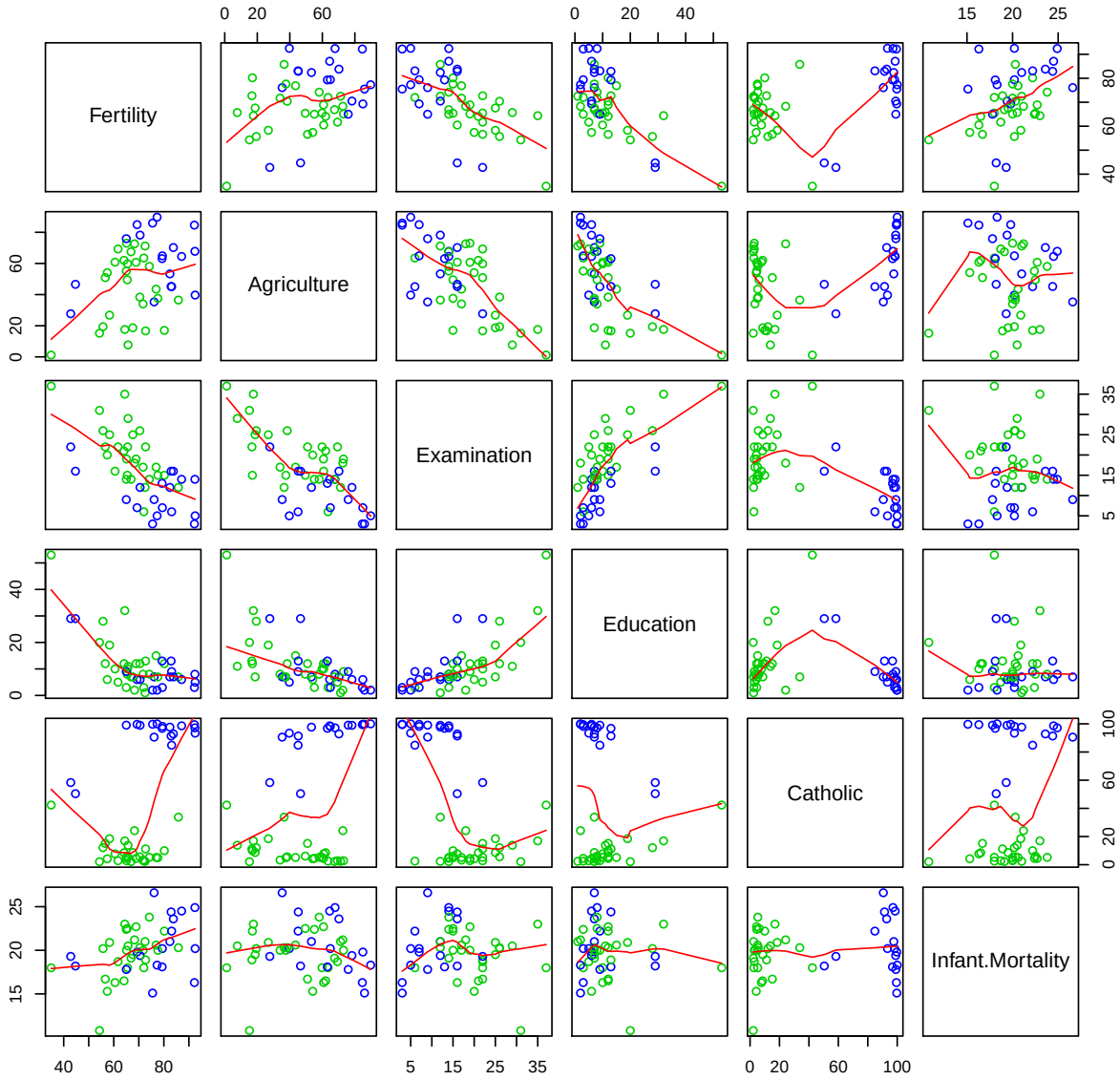
Multiple linear regression

- Consider the `swiss` dataset in `R`
  - This is a built-in dataset in `R`
  - Call it using the command `data(swiss)`
  - A data frame with 47 observations on 6 variables, each of which is in percent, i.e., in $[0, 100]$.
- Standardized fertility measure and socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888
- Variables:
  1. Fertility: a common standardised fertility measure
  2. Agriculture: Percentage of males involved in agriculture as occupation
  3. Examination: Percentage of draftees receiving highest mark on army examination
  4. Education: Percentage of education beyond primary school for draftees
  5. Catholic: Percentage of 'catholic' (as opposed to 'protestant')
  6. Infant.Mortality: Percentage of live births who live less than one year

Multiple linear regression

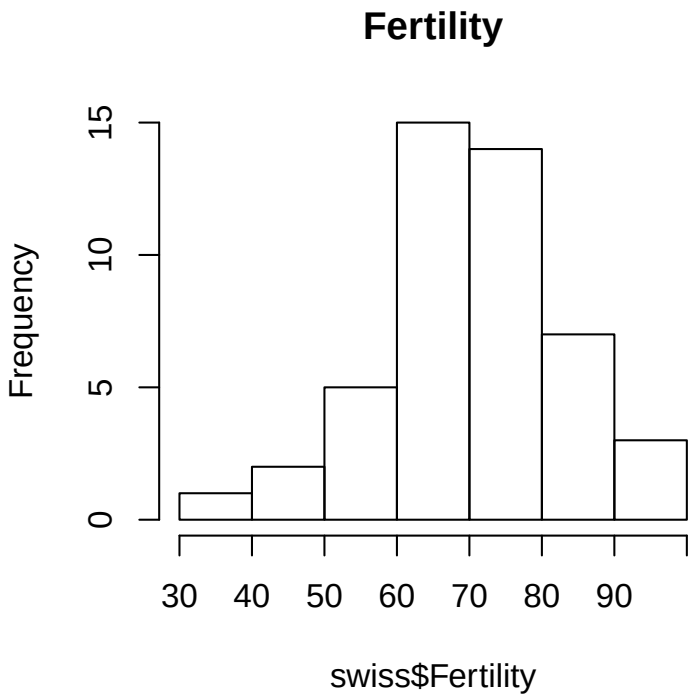
```
> data(swiss)
> summary(swiss)
   Fertility      Agriculture     Examination      Education
 Min.   :35.00   Min.   : 1.20   Min.   : 3.00   Min.   : 1.00
 1st Qu.:64.70   1st Qu.:35.90   1st Qu.:12.00   1st Qu.: 6.00
 Median :70.40   Median :54.10   Median :16.00   Median : 8.00
 Mean   :70.14   Mean   :50.66   Mean   :16.49   Mean   :10.98
 3rd Qu.:78.45   3rd Qu.:67.65   3rd Qu.:22.00   3rd Qu.:12.00
 Max.   :92.50   Max.   :89.70   Max.   :37.00   Max.   :53.00
    Catholic       Infant.Mortality
 Min.   :  2.150   Min.   :10.80
 1st Qu.:  5.195   1st Qu.:18.15
 Median : 15.140   Median :20.00
 Mean   : 41.144   Mean   :19.94
 3rd Qu.: 93.125   3rd Qu.:21.70
 Max.   :100.000   Max.   :26.60

> pairs(swiss, panel = panel.smooth, main = "swiss data",
  col = 3 + (swiss$Catholic > 50))
```

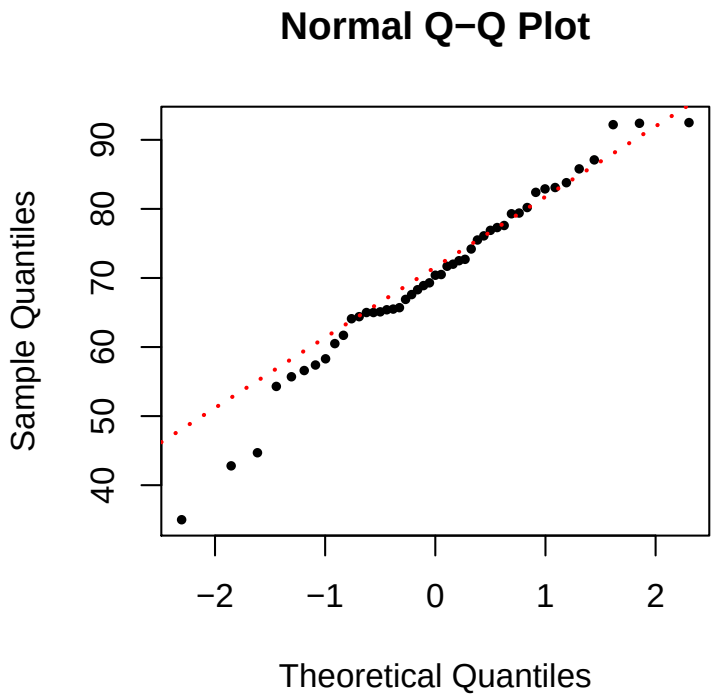
Multiple linear regression



Our interest: modelling fertility as a function of socio-economic indicators

```
hist(swiss$Fertility, main="Fertility")
qqnorm(swiss$Fertility, pch=19, cex=0.5)
qqline(swiss$Fertility, lty=3, col=2, lwd=2)
```





Multiple linear regression: parameter estimation

- We are interested in the model

  $$\text{Fertility} = \beta_0 + \beta_1\ \text{Agriculture} + \beta_2\ \text{Examination} + \beta_3\ \text{Education} + \beta_4\ \text{Catholic} + \beta_5\ \text{Infant.Mortality}$$

- The parameters are estimated using the same principle: minimising the residual sum of squares

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
Agriculture      -0.17211    0.07030  -2.448  0.01873 *
Examination      -0.25801    0.25388  -1.016  0.31546
Education        -0.87094    0.18303  -4.758 2.43e-05 ***
Catholic          0.10412    0.03526   2.953  0.00519 **
Infant.Mortality  1.07705    0.38172   2.822  0.00734 **

Residual standard error: 7.165 on 41 degrees of freedom
Multiple R-squared:  0.7067,Adjusted R-squared:  0.671
F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

Multiple linear regression: Inference

- We still use the same principle in inference
- There are two types of test:
  1. We are interested to test individual variables whether they are significantly associated with the response variable
     I.e. $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$ for any given j
  2. We are interested to test whether *any* predictor is significantly associated with the response variable
     I.e. $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ against $H_1 : \beta_j \neq 0$ for *at least one* j
- The first type focuses on the contribution of each variable
- The second type focuses whether the model is worthwhile (not necessarily interested in a particular variable)
- Of course, when we reject the first type's H_0, we also reject the second type's H_0 (provided the same set of variables are in the model)

- To test the null hypothesis $H_0 : \beta_j = 0$ for any given j (any variable), then we perform the test

$$t = \frac{\widehat{\beta}_j}{\text{se}(\widehat{\beta}_j)}$$

  where $\text{se}(\widehat{\beta}_j)$ is the standard error for $\widehat{\beta}_j$
- Under H_0 ('if H_0 is true'), then t would follow a t-distribution with $n - (p + 1)$ degrees of freedom
- Remember, we have p predictors and $p + 1$ parameters because we include the intercept
- At significance level 5%, we reject H_0 if the corresponding p-value is less than 5%
- (If H_0 is rejected) We conclude that the variable is significantly associated with the response variable

Multiple linear regression: Inference

- To test the null hypothesis $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$, we perform anova (ANalysis Of VAriance) test
- We basically decompose the variability in the data into two parts: variability due to the model (the systematic part of the model) and the variability due to error
- If the variability due to the model is 'big enough' compared to the variability due to the error, then we reject the H_0
- We then use the F test to assess this, and we can obtain a p-value of the test
- At significance level 5%, we reject H_0 if the corresponding p-value is less than 5%
- (If H_0 is rejected) We conclude that at least one variable is significantly associated with the response variable

Multiple linear regression: Inference

- We review the R output again:

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
Agriculture      -0.17211    0.07030  -2.448  0.01873 *
Examination      -0.25801    0.25388  -1.016  0.31546
Education        -0.87094    0.18303  -4.758 2.43e-05 ***
Catholic          0.10412    0.03526   2.953  0.00519 **
Infant.Mortality  1.07705    0.38172   2.822  0.00734 **

Residual standard error: 7.165 on 41 degrees of freedom
Multiple R-squared:  0.7067,Adjusted R-squared:  0.671
F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

- Agriculture, Education, Catholic, and Infant Mortality are significantly associated with Fertility
- Examination is not significant (or, is it? See next slide)
- Overall model is significant, p-value of 5.6×10^{-10}

Multiple linear regression: Interpretation

- Interpretation of the estimates cannot be separated from the set of variables in the model
- Consider the case where we have one predictor in the model: Agriculture

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 60.30438    4.25126  14.185   <2e-16 ***
Agriculture  0.19420    0.07671   2.532   0.0149 *
```

- When we include Infant.Mortality,

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      21.95462   11.52852   1.904  0.06341 .
Agriculture       0.20892    0.06864   3.044  0.00394 **
Infant.Mortality  1.88563    0.53522   3.523  0.00101 **
```

- Note: The sign of $\widehat{\beta}$ for Agriculture differs from that in the full model(!)

Multiple linear regression: Interpretation

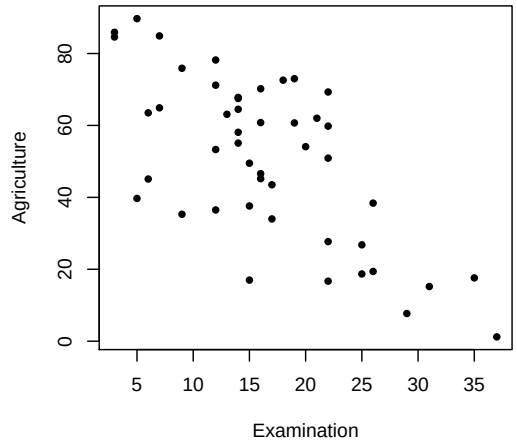
- When a new variable is included in the model, then the estimate of the parameter is adjusted
- The adjusted estimate can be higher or lower, or even change sign
- It can be shown that this adjustment depends on the correlation between the newly entered variable with the existing variable
- $\rightarrow$ correlation $\equiv$ shared information
- If the new variable is orthogonal to the existing variables, then the new variable contains all new information
- $\rightarrow$ The estimates of the 'old' variables are not adjusted

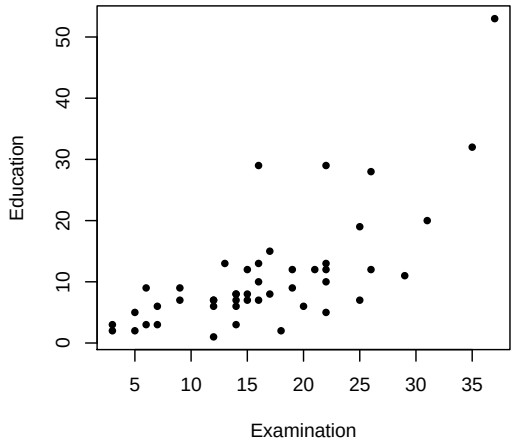
Multiple linear regression: Interpretation

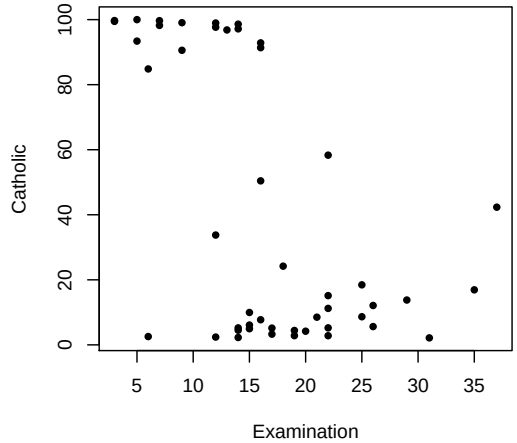
- In linear regression, our interpretation on each variable is 'ignoring' the other variables
- When there are two correlated variables, both the parameter estimates and the standard error will be affected
- Even when each of them is significantly associated with response variable, one of them will show as insignificant when put together in a model

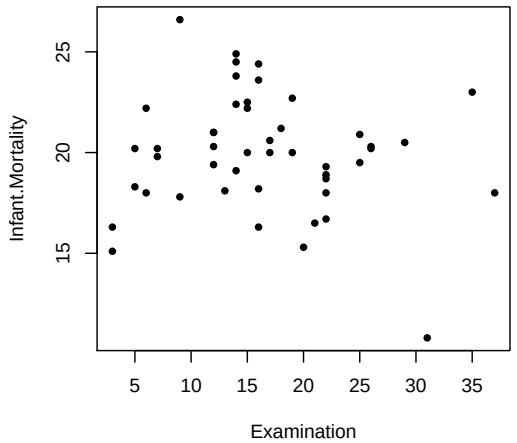
Multiple linear regression: Interpretation

Is examination not significant?









Multiple linear regression: Interpretation

- Examination on its own

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   86.8185     3.2576  26.651  < 2e-16 ***
Examination   -1.0113     0.1782  -5.675 9.45e-07 ***
```

- With all variables in the model

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
Agriculture      -0.17211    0.07030  -2.448  0.01873 *
Examination      -0.25801    0.25388  -1.016  0.31546
Education        -0.87094    0.18303  -4.758 2.43e-05 ***
Catholic          0.10412    0.03526   2.953  0.00519 **
Infant.Mortality  1.07705    0.38172   2.822  0.00734 **
```

Multiple linear regression: Interpretation

- We remove Education from the full model

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      59.60267   13.04246   4.570 4.25e-05 ***
Agriculture      -0.04759    0.08032  -0.593 0.556688
Examination      -0.96805    0.25284  -3.829 0.000423 ***
Catholic          0.02611    0.03843   0.679 0.500551
Infant.Mortality  1.39597    0.46259   3.018 0.004315 **
```

- Which model is the best?

Multiple linear regression: Interpretation

- The Akaike's Information Criterion (AIC) measures the relative quality of statistical models (model fit and prediction) for a given set of data
- The magnitude (in absolute sense) is not meaningful
- (Given the same data) Model with lower AIC has better quality than that with higher AIC
- It can be shown that the optimal model is given by

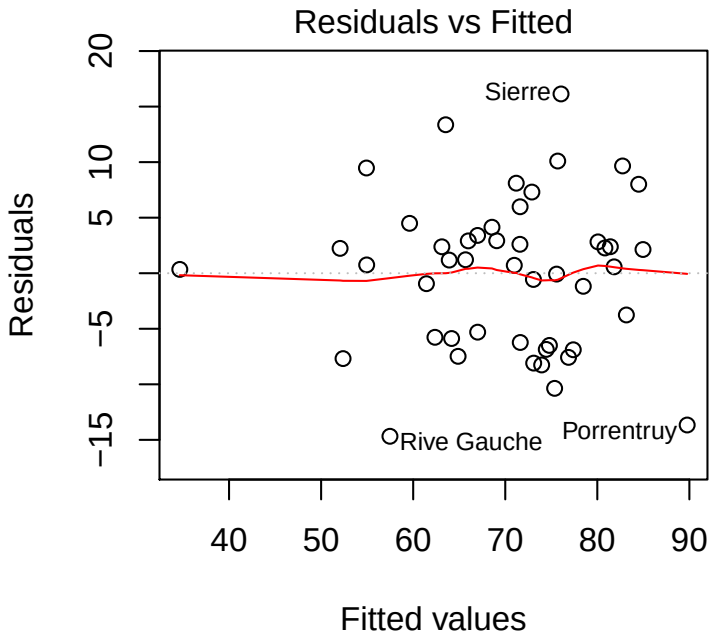
```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       62.10131    9.60489   6.466 8.49e-08 ***
Agriculture       -0.15462    0.06819  -2.267  0.02857 *
Education         -0.98026    0.14814  -6.617 5.14e-08 ***
Catholic           0.12467    0.02889   4.315 9.50e-05 ***
Infant.Mortality   1.07844    0.38187   2.824  0.00722 **
```

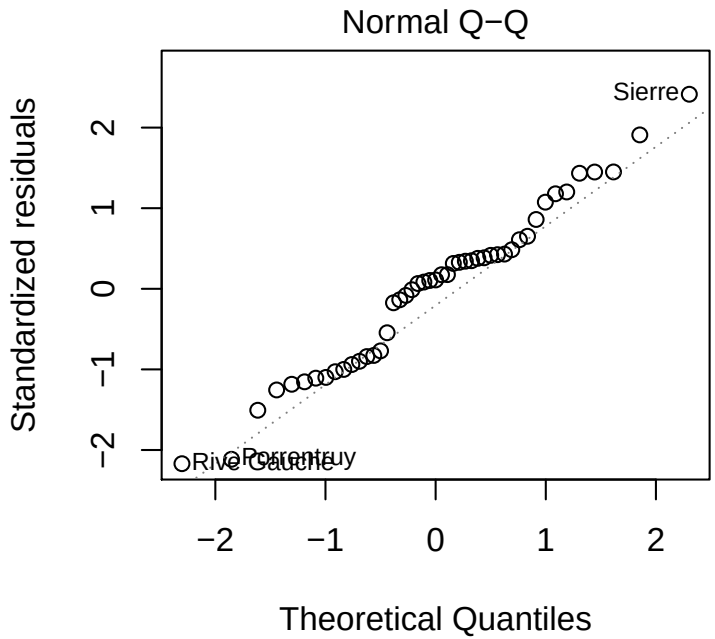
What to do when we have correlated variables?

- Identify *the reason* for correlation
- If the concerned variables are repeated measurements or have some sort of structure, then consider linear mixed models (not covered)
- Create a new variable , e.g. $z_i = x_{1i}/x_{2i}$
  - If relevant and appropriate
  - In a study involving children, age will be correlated with other variables that improve with age
- Other models: ridge regression, penalised regression etc.

Multiple linear regression: Model diagnostics

- Model diagnostics carry the same principle as the simple linear regression





Multiple linear regression: A note on categorical variables

- When a predictor is categorical (with t levels)
- Then there are $t - 1$ parameters to estimate (for that variable)
- By default, `R` uses the first level as the baseline (part of intercept)
- In terms of interpretation, the parameter estimates as basically compared to the baseline
- Consider the following example, where the variable `Agriculture` is categorised to $t = 3$ levels
  - Low: $< 30\%$
  - Medium: $> 30\%$ but $< 70\%$
  - High: $> 70\%$

```
swiss$Agri.factor <- ifelse(swiss$Agriculture>=30, "Medium", "Low")
swiss$Agri.factor[swiss$Agriculture>=60] <- "High"
swiss$Agri.factor <- factor(swiss$Agri.factor)
```

Multiple linear regression: A note on categorical variables

- R output:

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       51.29239    8.78368   5.840 7.95e-07 ***
Agri.factorLow     5.01018    4.26848   1.174 0.247432
Agri.factorMedium  2.41187    2.74031   0.880 0.384039
Education         -0.78146    0.19104  -4.091 0.000202 ***
Examination       -0.18052    0.28284  -0.638 0.526952
Catholic           0.09281    0.03820   2.430 0.019697 *
Infant.Mortality   1.23602    0.40653   3.040 0.004155 **
```

- There are two parameters being estimated by R for variable Agriculture: "Low" and "Medium"
- Level "High" is part of the baseline/intercept
- "H" comes before either "L" or "M", hence "High" is automatically selected by R as the baseline

Multiple linear regression: A note on categorical variables

- You can change this (to make "Low" the baseline, for example), using the command `relevel()`

```
> table(swiss$Agri.factor)# Before

  High    Low Medium
    20     10     17

> swiss <- within(swiss, Agri.factor <- relevel(Agri.factor, ref="Low"))

> table(swiss$Agri.factor)# After

   Low   High Medium
    10     20     17
```

Multiple linear regression: A note on categorical variables

- R output:

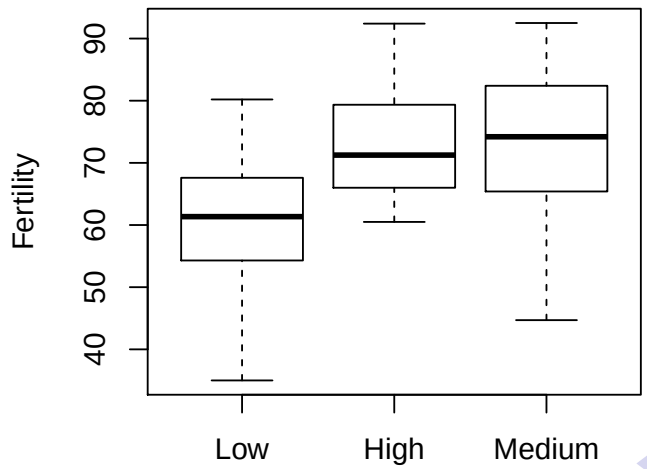
```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       56.30257   10.25207   5.492 2.44e-06 ***
Agri.factorHigh   -5.01018    4.26848  -1.174 0.247432
Agri.factorMedium -2.59831    3.95520  -0.657 0.514985
Education         -0.78146    0.19104  -4.091 0.000202 ***
Examination       -0.18052    0.28284  -0.638 0.526952
Catholic           0.09281    0.03820   2.430 0.019697 *
Infant.Mortality   1.23602    0.40653   3.040 0.004155 **
```

- Now, "Low" is the baseline, and none of the other levels/groups are significantly different from the baseline
- Suppose, for our illustration purposes, the estimates of "High" and "Medium" are significant
- Then we can interpret that regions with medium (high) level of agriculture are expected to have 2.6 (5) percent less fertility than those in the low level

Multiple linear regression: A note on categorical variables

- Note that the interpretation of the categorical variable is still dependent on the other variables present in the model
- We have seen that there are some correlations between Agriculture and the other variables (the correlations remain!)
- When only Agriculture (factor) in the model:

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         59.670      3.621  16.480  < 2e-16 ***
Agri.factorHigh     13.920      4.434   3.139  0.00302 **
Agri.factorMedium   12.577      4.563   2.756  0.00848 **
```



Multiple linear regression: A note on categorical variables

- When only Agriculture (factor) in the model:

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         59.670      3.621  16.480  < 2e-16 ***
Agri.factorHigh     13.920      4.434   3.139  0.00302 **
Agri.factorMedium   12.577      4.563   2.756  0.00848 **
```

- The parameter estimates for "High" and "Medium" are simply the differences with the intercept/baseline "Low"

```
> by(swiss$Fertility, swiss$Agri.factor, mean)
swiss$Agri.factor: Low
[1] 59.67
------------------------------------------------------------
swiss$Agri.factor: High
[1] 73.59
------------------------------------------------------------
swiss$Agri.factor: Medium
[1] 72.24706
```

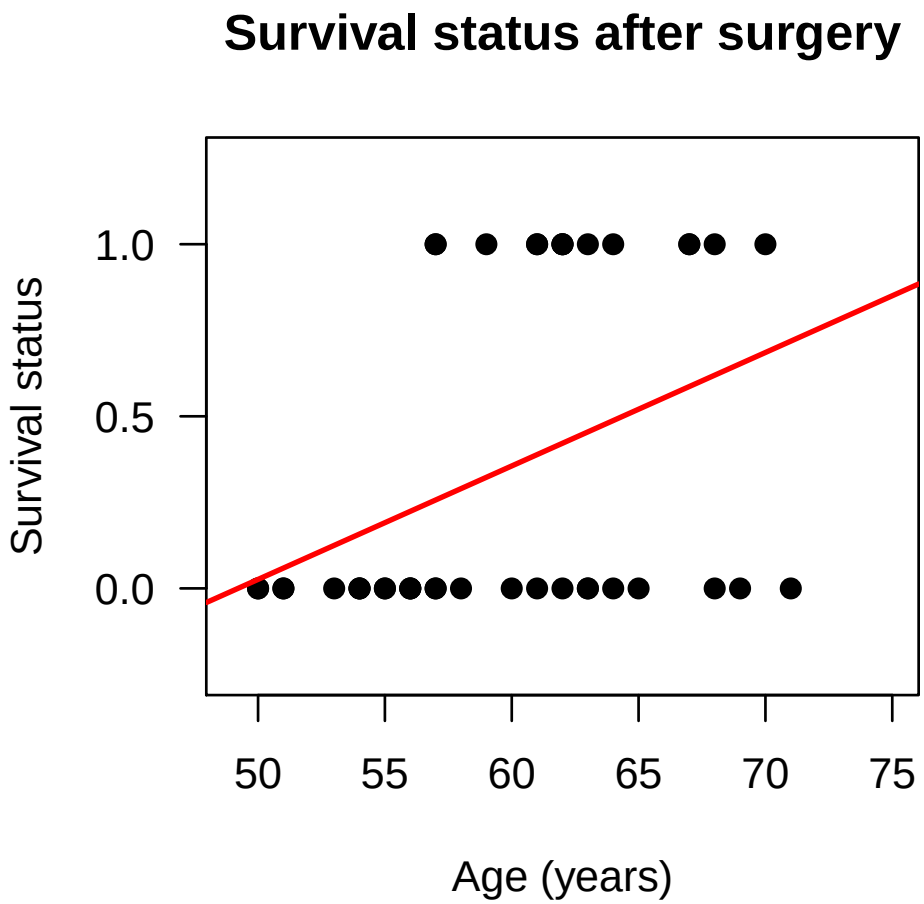
Any Questions?

Generalised linear models: Motivating example

- Consider the following example from Pawitan's *In All Likelihood* Chapter 6 (2001)
- We record the survival status (alive or dead) of 40 people within 30 days after surgery
- Note (!) this is not survival analysis with time-to-event data
- Each person is either dead or alive 30 days after their surgery
- Our interest: Is age a risk factor in the survival?
  In other words: *Do older people have higher probability to die after surgery (may be due to complication)?*

Generalised linear models: Motivating example

- The data (red line: linear regression line)



- Interpretation??

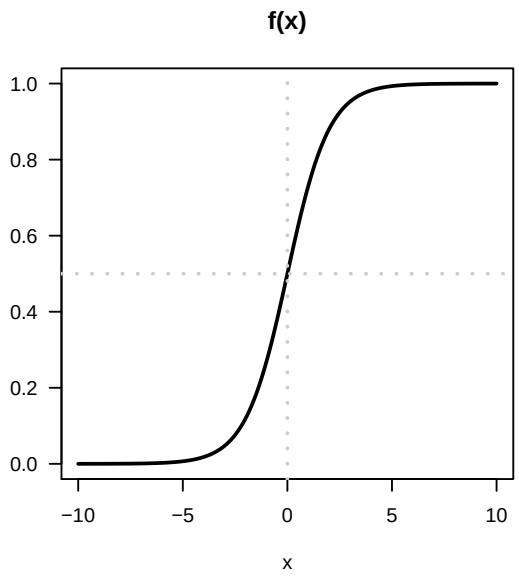
- Modelling the value (0 or 1) directly (i.e. linear regression) is not meaningful
- Consider the probability of death as π (ranges from 0 to 1)
- Predictor ranges from $-\infty$ to ∞ (in principle)
- We need a function that "translate" the value of predictors $(-\infty, \infty)$ into probability $(0, 1)$
- In mathematics, there is a function called *logistic function*

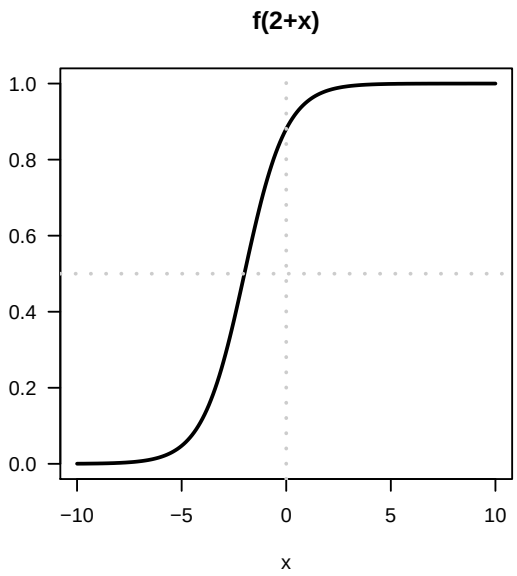
$$f(x) = \frac{\exp\{x\}}{1 + \exp\{x\}}$$

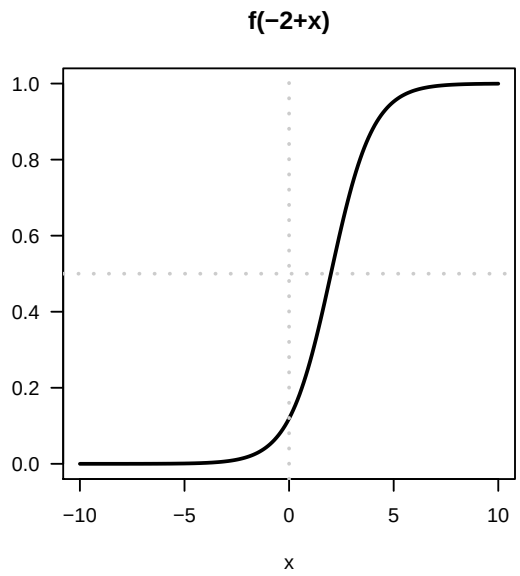
(for a general x)

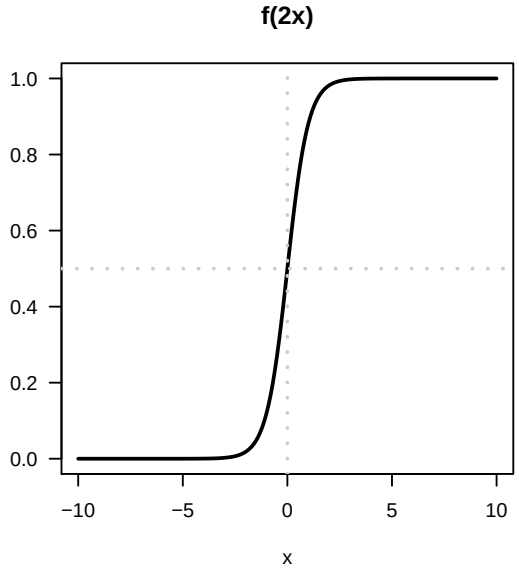
- (Hence, the name *logistic regression*)

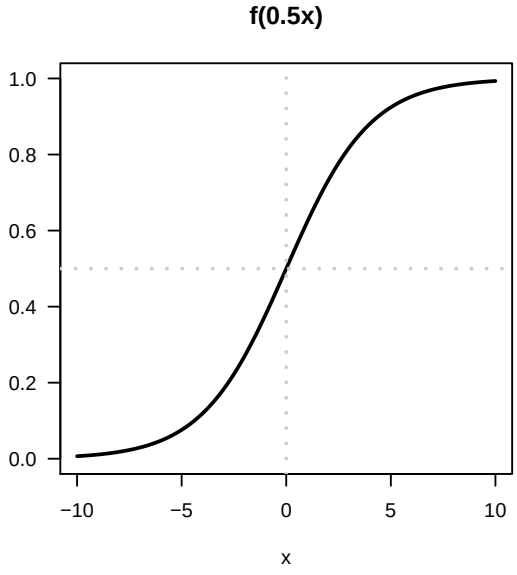
Generalised linear models: Logistic regression

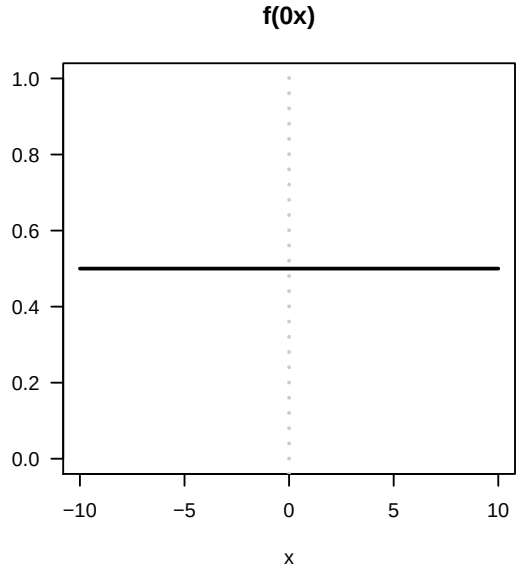












- For our purpose, we utilise this function to make relationship between the probability of death of individual i, π_i, and the predictor x_i

$$\pi_i = \frac{\exp\{\beta_0 + \beta_1 x_i\}}{1 + \exp\{\beta_0 + \beta_1 x_i\}}$$

- We model the *probability* of death, and **not** the death/alive status
- It can be shown that the above can be reformulated as

$$\log\left\{\frac{\pi_i}{1-\pi_i}\right\} = \beta_0 + \beta_1 x_i$$

  i.e. modelling log odds of death
- The function $\log\left\{\frac{\pi_i}{1-\pi_i}\right\}$ is called *logit*

Logistic regression: Interpretation

- If β_1 is zero, then the logistic curve (as a function of x) will be flat
- $\rightarrow$ x does not affect (increase/decrease) the probability of death
- $\rightarrow$ log odds, i.e. $\log\left\{\frac{\pi_i}{1-\pi_i}\right\}$, is constant
- For an increase of $\widehat{\beta}_1$ by one unit, the log odds is expected to increase by one unit
- Or, the odds is expected to increase by exp(1) (natural number, approx. 2.72)

Logistic regression: Parameter estimation

- The parameters β_0 and β_1 are estimated using numerical techniques as the function is non linear
  Exercises for those who are from Maths/Stats/Physics background
- Iterative method: Iterative Weighted Least Squares (IWLS)
- The standard error of the estimates can be obtained as an approximation

Logistic regression: Surgery data

- Analysis of the surgery data gives the following R output

```
sur <- read.table("logisticreg-surgery-data.txt", header=T)
res <- glm(y~age, family="binomial", data=sur)
summary(res)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.48174    4.30409  -2.435   0.0149 *
age           0.16295    0.07018   2.322   0.0202 *

    Null deviance: 51.796  on 39  degrees of freedom
Residual deviance: 45.301  on 38  degrees of freedom
AIC: 49.301
```

- The estimate for age, $\widehat{\beta}_1$, is 0.16295
- This indicates that an increase of one year in age, the odds to die within 30 days after surgery is expected to increase by $\exp(0.16295) \approx 1.177$ times

Logistic regression: Surgery data

- The fitted regression line is given by

$$\widehat{\pi}(\text{Age}) = \frac{\exp\{-10.48 + 0.163 \times \text{Age}\}}{1 + \exp\{-10.48 + 0.163 \times \text{Age}\}}$$

Logistic regression: Surgery data

- The estimated probability of death for someone whose age is 70 years old is
$$\frac{\exp\{-10.48 + 0.163 \times 70\}}{1 + \exp\{-10.48 + 0.163 \times 70\}} = 0.7159732$$
- (This means) the odds ratio is given by $0.71597/(1 - 0.71597) = 2.520795$
- For someone whose age is 71 years old, $\widehat{\pi}(\text{Age} = 71) = 0.7479147$, with odds ratio $0.7479/(1 - 0.7479) = 2.966912$
- For an increase of one unit of age, the odds ratio increases by $2.966912/2.520795 \approx 1.177$ times
- Note that $\log(1.177) = 0.16295$ $(= \widehat{\beta}_1!)$

Logistic regression: Multiple predictors

- When we have multiple predictors, $x_1, x_2, \ldots, x_p$, we can enlarge the above model into something like

$$\log\left\{\frac{\pi_i}{1-\pi_i}\right\} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

- The interpretation of each parameter remains consistent (given the others are fixed)
- The issue of correlation (between predictors) in the linear regression is also present in generalised linear models

Logistic regression: Inference

- Similar to the context of linear regression, there are two types of test:
  1. We are interested to test individual variables whether they are significantly associated with the response variable
     I.e. $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$ for any given j
  2. We are interested to test whether *any/some* predictor(s) is/are significantly associated with the response variable
     I.e. $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ against $H_1 : \beta_j \neq 0$ for *at least one* j

Logistic regression: Inference

- To test the null hypothesis $H_0 : \beta_j = \delta^0$, for any j, we calculate the test statistic

$$z_j = \frac{\widehat{\beta}_j - \delta^0}{\text{se}(\widehat{\beta}_j)}$$

- Under H_0, z_j approximately follow a standard normal distribution
- At significance level 5%, we reject H_0 if the corresponding p-value is less than 5%
- (If H_0 is rejected) We conclude that the variable is significantly associated with the log odds

Logistic regression: Inference

- To test the null hypothesis $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$, we can perform either **likelihood ratio test** (an exercise for those who are from Maths/Stats/Physics background) or **deviance test**
- *Deviance* is a measure of lack of fit in generalised linear models
- Similar to the concept of residual sum of squares in the (normal) linear regression models

- The test statistic is calculated as

$$X = \text{Null deviance} - \text{Model deviance}$$

- Under $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$, X approximately follows χ^2 distribution with p degrees of freedom
- p being the difference in degrees of freedom (between the two deviances)
- ... or the number of parameters tested at the same time
- Model deviance (or *residual deviance*) is the deviance of the fitted model
- Null deviance is the deviance when there is only intercept in the model

Logistic regression: Example

- Wisconsin Diagnostic Breast Cancer (WDBC) study (Wolberg et al, 1994)
- Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass
- They describe characteristics of the cell nuclei present in the image
  1. radius (mean of distances from center to points on the perimeter)
  2. texture (standard deviation of gray-scale values)
  3. perimeter
  4. area
  5. smoothness (local variation in radius lengths)
  6. compactness ($\text{perimeter}^2/\text{area} - 1.0$)
  7. concavity (severity of concave portions of the contour)
  8. concave points (number of concave portions of the contour)
  9. symmetry
  10. fractal dimension ("coastline approximation" - 1)

Logistic regression: Example

- The response variable: whether the breast cancer is benign ($y = 0$) or malignant ($y = 1$)
- The output in R gives

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -22.97705   16.03990  -1.432   0.1520
radius.mean         -1.63760    2.01183  -0.814   0.4157
area.mean            0.03239    0.02288   1.416   0.1568
concavity.mean      17.72956   12.43666   1.426   0.1540
fractal.dim.mean   -44.43326  102.17617  -0.435   0.6637
perimeter.mean       1.46141    0.35358   4.133 3.58e-05 ***
compactness.mean   -89.95446   30.90153  -2.911   0.0036 **
symmetry.mean       61.46146   60.17448   1.021   0.3071
texture.mean         0.34164    0.06606   5.171 2.32e-07 ***
smoothness.mean     63.13449   23.87973   2.644   0.0082 **
concave.points.mean 54.20848   17.36747   3.121   0.0018 **
---
    Null deviance: 751.440  on 568  degrees of freedom
Residual deviance:  78.016  on 558  degrees of freedom
AIC: 100.02
```

Logistic regression: Example

- Each row of the outcome table contains the estimate, standard error, test statistic, and p-value to test $H_0 : \beta_j = 0$ (for each variable)
- The results indicate that perimeter, compactness, texture, smoothness, and concave points are significant
- For testing $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$, the difference of the null and residual deviance is $X = 751.440 - 78.016 = 673.424$
- Under the H_0, $X \sim \chi^2_{10}$
- The degrees of freedom (10) is the difference of degrees of freedom of the deviances (568-558)
- The 5% critical value is 18.3 – The difference is significant (673¿18.3)
- (p-value is approximately 3.175661e-138)
- Conclusion: at least one of the variables is significant

Logistic regression: Example

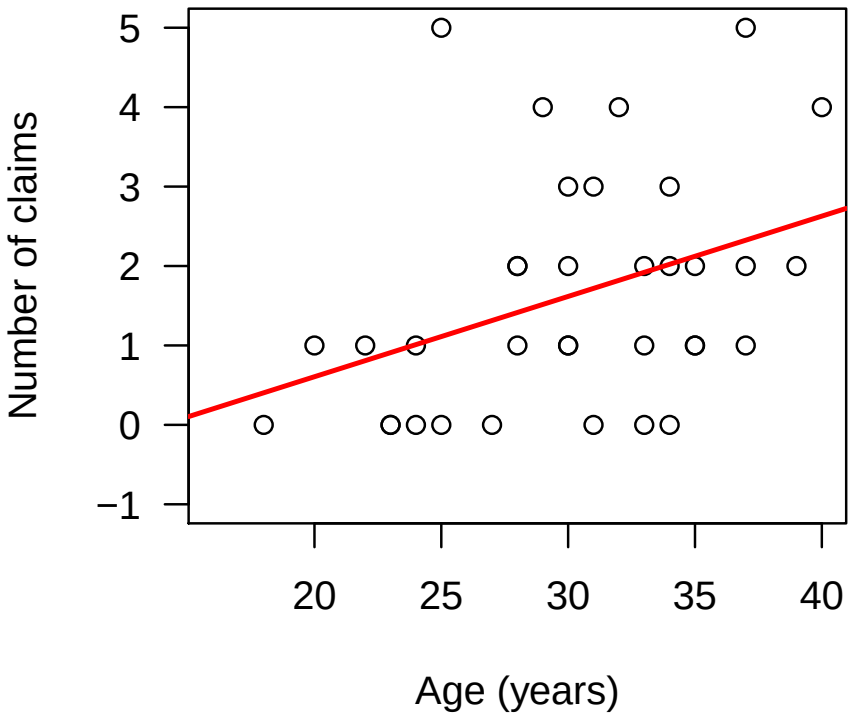
- In terms of modelling, you need to consider that some variables are also correlated
- You may consider the variable selection method we discussed earlier to arrive at the best model (I will leave these two issues as exercise)

Any Questions?

Generalised Linear Models: Poisson regression

- Other non-normal data that are quite common is count data
- A main characteristic: count is non-negative
- The minimum value is zero and take more positive values
- The logistic function cannot be used

- Health insurance claim data (Pawitan, 2001, Chapter 6)



- Interpretation?

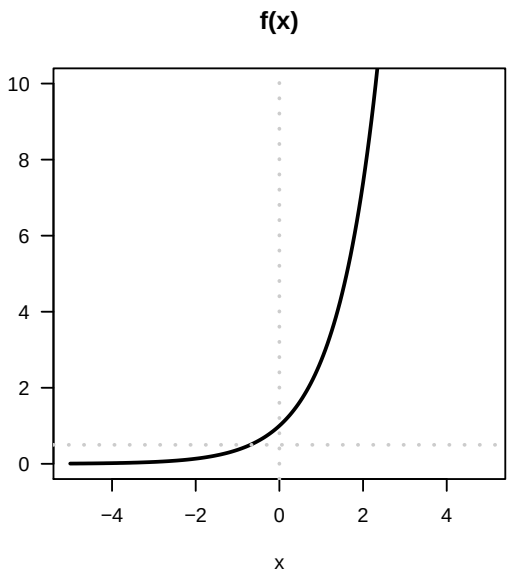
Generalised linear models: Poisson regression

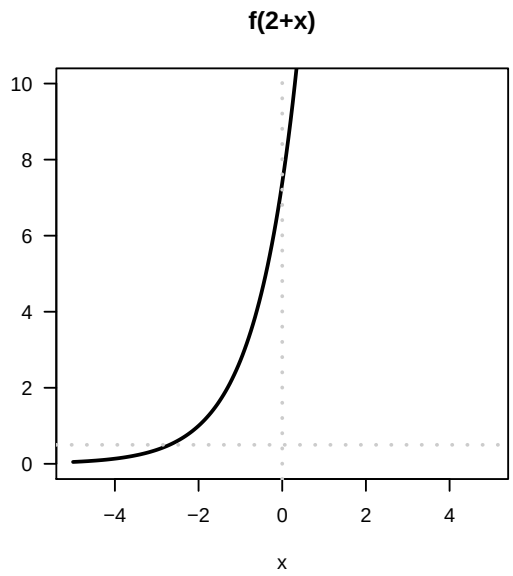
- Modelling the number of claims directly (i.e. linear regression) is not meaningful (negative prediction!)
- Consider the mean of number of claims as μ (non-negative)
- Predictor ranges from $-\infty$ to ∞ (in principle)
- We need a function that "translate" the value of predictors $(-\infty, \infty)$ into non-negative values
- We can consider the exponent function

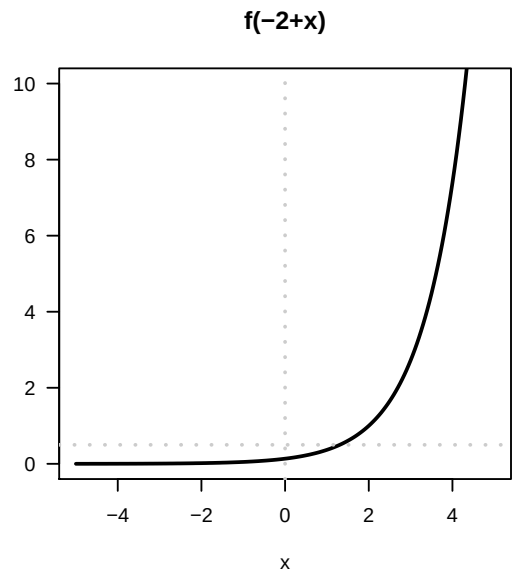
$$f(x) = \exp\{x\}$$

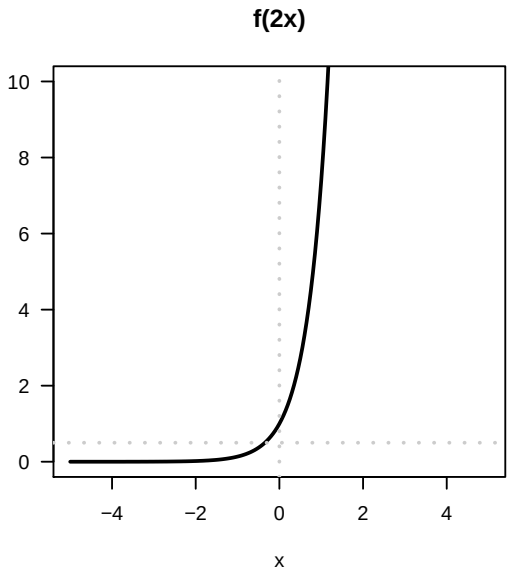
(for a general x)

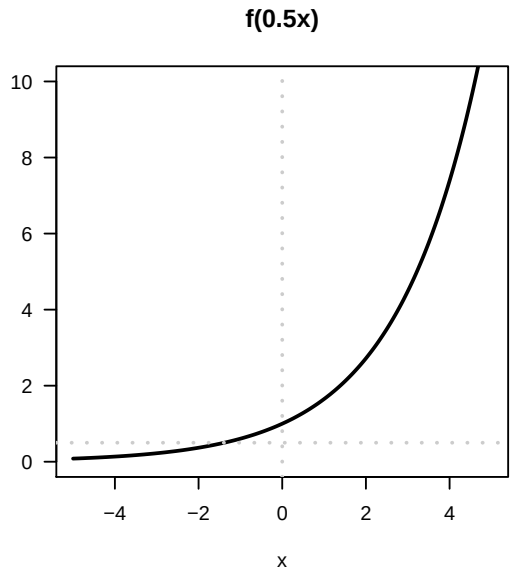
Generalised linear models: Logistic regression

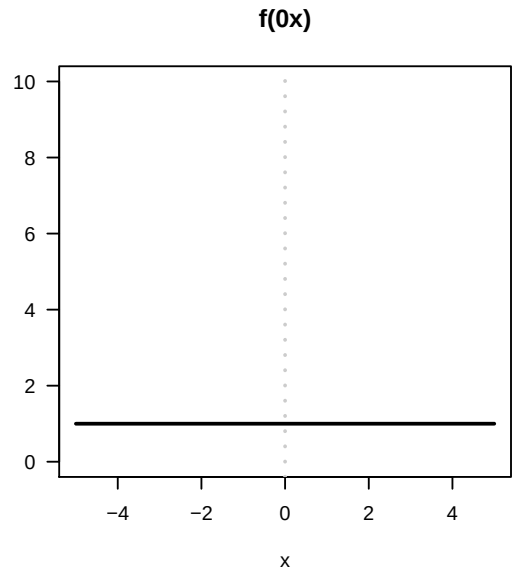












- For our purpose, we utilise this function to make relationship between the mean number of claims for individual i, μ_i, and the predictor x_i

$$\mu_i = \exp\{\beta_0 + \beta_1 x_i\}$$

- We model the *mean* number of claims, and **not** the number of claims
- It can be shown that the above can be reformulated as

$$\log\{\mu_i\} = \beta_0 + \beta_1 x_i$$

i.e. modelling log mean of number of claims

Poisson regression: Interpretation

- If β_1 is zero, then the log curve (as a function of x) will be flat
- $\rightarrow$ x does not affect (increase/decrease) the mean count (mean of the number of claims)
- For an increase of $\widehat{\beta}_1$ by one unit, the log mean count is expected to increase by one unit
- Or, the mean is expected to increase by exp(1) (natural number, approx. 2.72)

Poisson regression: Parameter estimation

- Similar to the logistic regression, the parameters β_0 and β_1 are estimated using numerical techniques as the function is non linear
  Exercises for those who are from Maths/Stats/Physics background
- Iterative method: Iterative Weighted Least Squares (IWLS)
- The standard error of the estimates can be obtained as an approximation

Poisson regression: Health claims data

- Analysis of the health claims data gives the following R output

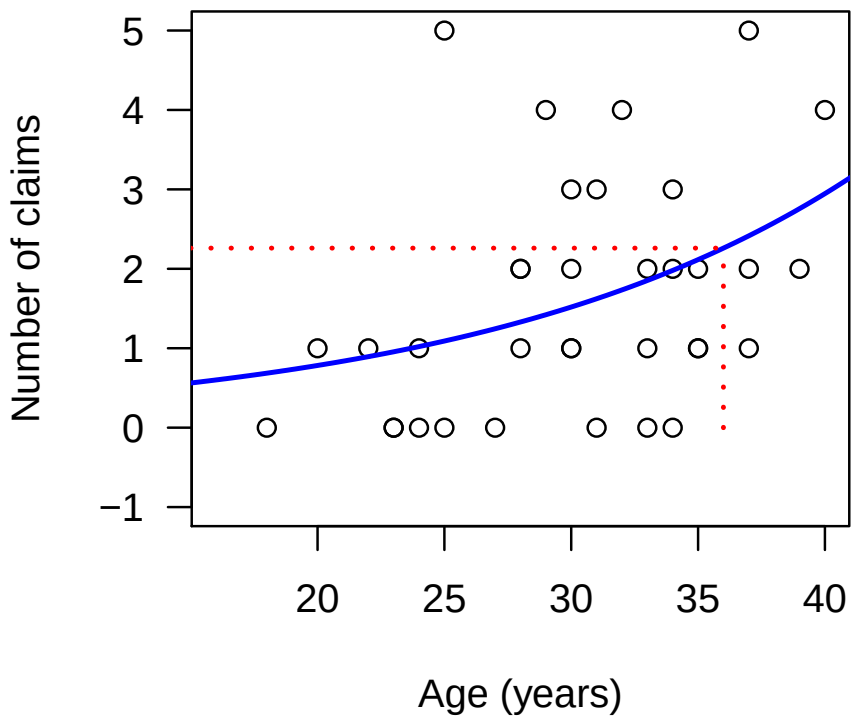
```
ins.claims=read.table("ins-claims.txt", header=T)
res <- glm(claims~age, family="poisson", data=ins.claims)
summary(res)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.56988    0.84788  -1.852   0.0641 .
age          0.06626    0.02621   2.528   0.0115 *

    Null deviance: 51.817  on 34  degrees of freedom
Residual deviance: 45.062  on 33  degrees of freedom
AIC: 115.71
```

- The estimate for age, $\widehat{\beta}_1$, is 0.06626
- This indicates that an increase of one year in age, the mean number of claims is expected to increase by $\exp(0.06626) \approx 1.07$ times

Poisson regression: Health claims data

- The fitted regression line is given by

$$\widehat{\mu}(\text{Age}) = \exp\{-1.57 + 0.066 \times \text{Age}\}$$



Poisson regression: Health claims data

- The estimated mean of number of claims for someone whose age is 36 years old is
$$\exp\{-1.57 + 0.066 \times 36\} = 2.26$$
- For someone whose age is 37 years old, $\widehat{\mu}(\text{Age} = 37) = 2.415$
- For an increase of one unit of age, the expected mean increases by $2.415/2.26 \approx 1.07$ times
- Note that $\log(1.07) = 0.066$ $(= \widehat{\beta}_1!)$

Poisson regression: Multiple predictors

- When we have multiple predictors, $x_1, x_2, \ldots, x_p$, we can enlarge the above model into something like

$$\log\{\mu_i\} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

- The interpretation of each parameter remains consistent (given the others are fixed)
- The issue of correlation (between predictors) in the linear regression is also present in generalised linear models

Poisson regression: Inference

- Similar to the context of logistic regression, there are two types of test:
  1. We are interested to test individual variables whether they are significantly associated with the response variable
     I.e. $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$ for any given j
  2. We are interested to test whether *any/some* predictor(s) is/are significantly associated with the response variable
     I.e. $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ against $H_1 : \beta_j \neq 0$ for *at least one* j
- The testing of those hypotheses use the same principle as those in the logistic regression (not elaborated again here)

Poisson regression: Some notes

- When you model counts, it is possible to have *over-dispersion*
- What it means is that the variability in the data is actually higher than that we expect under the Poisson regression model
- This is quite common
- The (mean) estimate is not affected (robust)
- But the inference is badly affected

Poisson regression: Some notes

- Identify the reason for over dispersion
  - Groupings
  - Problems in the measurement
  - Heterogeneity
- There is no single solution for everything
- Need careful investigation

Any Questions?

Survival analysis: background

- In logistic regression, we are interested to understand whether a predictor (or 'risk factor') is associated with an event (e.g. death)
- Sometimes we are interested in the *time* to an event
  - Time to death after treatment intervention
  - Time until disease recurrence
  - Time to AIDS for HIV patients
- Some individuals may drop out from study or die due to other causes
- Logistic regression is not suitable

Survival analysis: background

- In survival analysis, we are dealing with time-to-event data
- First rule: you need to define properly what do you consider as an event
- For example, *death due to disease* or just *death*?
- Characteristics of data:
  - Non-negative values
  - Continuous
  - We know the exact time for some individuals, and only the minimum time
- Incompletely observed times are **censored**

Survival analysis: background

- When all times are exact, standard regression procedures could be used
- However, this may be inadequate because
  - Time is positive and rightly skewed
  - Probability of surviving past a certain point in time may be of more interest than the expected (mean) time of event
  - The hazard function (later on this) can be more informative (on mechanism) than regression function

Survival analysis: Censoring

- Censoring happens when we only have some partial information on the time, but not the exact time
- Censoring mechanism must be independent of the survival mechanism for survival analysis to be valid (there are some exceptions)
- There are generally three reasons why censoring might occur
  - The study ends before we can observe the event for some individuals
  - Lost to follow up
  - Withdrawal (e.g. withdraw consent)

Survival analysis: motivation

- Two groups of rats were exposed to carcinogen DBMA and the number of days to death due to cancer was recorded (Kalbfleisch and Prentice, 1980)

```
rat <- read.table("rat-survival.txt", header=T)
y.rat= Surv(rat$times, event=rat$status)
> y.rat
 [1] 143  164  188  188  190  192  206  209  213  216  220  227  230  234  246
[16] 265  304  216+ 244+ 142  156  163  198  205  232  232  233  233  233  233
[31] 239  240  261  280  280  296  296  323  204+ 344+
```

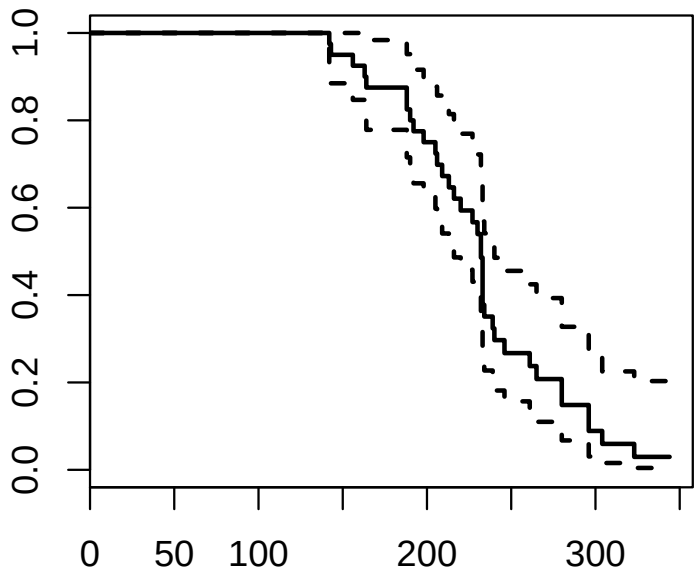
- 40 rats, first 19 are in Group 1, second 21 are in Group 2
- "+" indicates censored times

Survival analysis: survivor and hazard functions

- Survivor function, $S(t)$, gives the probability of an individual to survive past time t
- t ranges fom 0 to ∞
- Survivor function has the properties:
  - Non-increasing
  - At time $t = 0$, $S(0) = 1$
  - At time $t = \infty$, $S(\infty) = 0$
- In theory, the survivor function is smooth
- In practice, we observe events on a discrete time scale (days, weeks, etc.)

Survival analysis: survivor and hazard functions

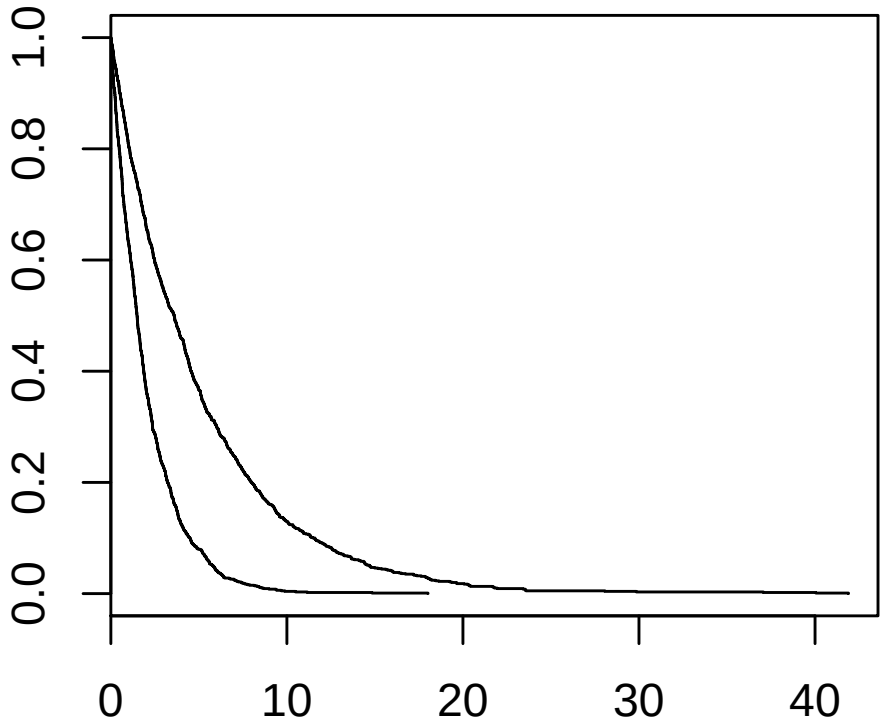
```
> summary(survfit(y.rat~1))

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  142     40       1   0.9750  0.0247      0.92780        1.000
  143     39       1   0.9500  0.0345      0.88480        1.000
  156     38       1   0.9250  0.0416      0.84687        1.000
  163     37       1   0.9000  0.0474      0.81167        0.998
  164     36       1   0.8750  0.0523      0.77829        0.984
  188     35       2   0.8250  0.0601      0.71527        0.952
  190     33       1   0.8000  0.0632      0.68517        0.934
```



Survival analysis: survivor and hazard functions

- The hazard function, $h(t)$, is the instantaneous *rate* at which events occur, given no previous events
- Instantaneous rate of death, given the individual survives until time t
- It can be shown that this is the negative of 'slope' of $\log S(t)$
  - The figure in the previous page is $S(t)$
- Higher hazard corresponds to 'faster death' ($S(t)$ goes down rapidly)
  - Lower hazard corresponds to 'slower death' ($S(t)$ goes down slowly)

Survival analysis: survivor and hazard functions



Cox regression

- Cox proportional hazard (PH) model, or Cox regression, models the hazard of individual i as a function of covariate

$$h_i(t; X) = h_0(t) \exp\{x_i \beta\}$$

- No intercept (!)
- $h_0(t)$ is the baseline hazard function
  - Does not take any form
  - A 'typical' hazard function for the sample
  - Only a function of time, and not covariate
- The function of covariates ($\exp\{x_i \beta\}$) modifies this 'typical' hazard function
  - This 'modifier' does not depend on time (does not contain t)

- When $\beta = 0$, the hazard of all individuals are the same (as the baseline hazard)
- The covariate x is not able to distinguish different hazard for groups of individuals
- $\rightarrow$ The covariate has no effect (to modify the baseline hazard function)
- Note that the model can be written as

$$\log \left\{ \frac{h_i(t)}{h_0(t)} \right\} = x_i \beta$$

Cox regression

- The term "proportional hazard" comes from the fact that the ratio of hazard for two individuals i and j are given by

$$\frac{h_i(t;X)}{h_j(t;X)} = (x_i - x_j)\beta$$

  as $h_0(t)$ (baseline hazard) cancels out
- The hazard for any individual is a fixed proportion of the hazard for any other individual
- In the Cox regression, the hazard functions for individuals are proportional
- β is interpreted as the log increase in the hazard for an increase of one unit of x
- As x increases by one unit, the hazard is expected to increase by $\exp(\beta)$ times
  - $\beta = 0 \rightarrow$ hazard remains the same (as $h_0(t)$)

Cox regression: parameter estimation

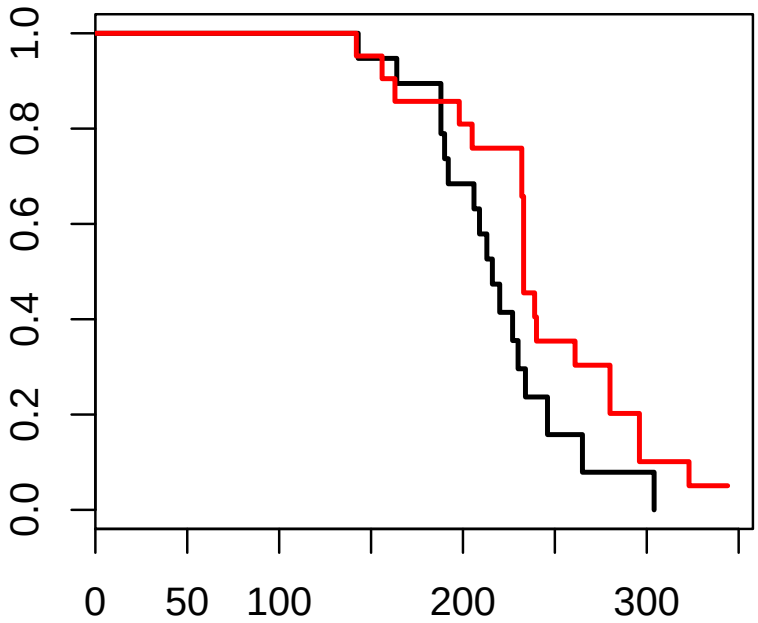
- Estimation of parameter β is performed using computation procedures, such as Newton-Raphson method
- Maximising log partial likelihood
- The estimation of standard error of $\widehat{\beta}$ can also be obtained
- The estimation of $h_0(t)$ and $S_0(t)$ (the corresponding baseline survivor function) is done separately

Rat data example

```
> res = coxph(y.rat ~ group, data=rat)
> summary(res)

  n= 40, number of events= 36

          coef exp(coef) se(coef)      z Pr(>|z|)
group2 -0.5686    0.5663   0.3472 -1.638    0.101

       exp(coef) exp(-coef) lower .95 upper .95
group2    0.5663      1.766    0.2868     1.118
```

- Estimate $\widehat{\beta} = -0.5686$
- The hazard of the second group is approximately $\exp(-0.5686) = 0.5663$ times that of the first group
- However, this is not significant (next)

Rat data example



- Survivor functions for the two groups
- Group 2 has lower hazard (which one?)

Cox regression: Multiple predictors

- We can incorporate more predictors in the model

$$h_i(t; X) = h_0(t) \exp\{x_{i1}\beta_1 + \ldots + x_{ip}\beta_p\}$$

- The interpretation of each parameter remains consistent (given the others are fixed)
- The issue of correlation (between predictors) in the linear regression is also present in the Cox regression

Cox regression: Inference

- Similar to the context of generalised linear models, there are two types of test:
  1. We are interested to test individual variables whether they are significantly associated with the response variable
     I.e. $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$ for any given j
  2. We are interested to test whether *any/some* predictor(s) is/are significantly associated with the response variable
     I.e. $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ against $H_1 : \beta_j \neq 0$ for *at least one* j

Cox regression: Inference

- For testing individual parameter, we use the test statistic

$$\frac{\widehat{\beta}_j - 0}{\text{se}(\widehat{\beta}_j)}$$

  which, under $H_0 : \beta_j = 0$, would follow approximately standard normal distribution
- At the 5% significance level, we reject the null hypothesis if the corresponding p-value is less than 5%
- If H_0 is rejected, we conclude that the corresponding variable is significantly affecting the hazard
  - Modify significantly the baseline hazard function

Cox regression: Inference

- To test the null hypothesis $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ against $H_1 : \beta_j \neq 0$, we can consider the likelihood ratio test
- Under the null hypothesis, the test statistic would follow χ^2_p distribution
- At the 5% significance level, we reject the null hypothesis if the corresponding p-value is less than 5%
- If H_0 is rejected, we conclude that *at least one (some)* predictors are significantly affecting the hazard

Cox regression: Inference

```
> res = coxph(y.rat ~ group, data=rat)
> summary(res)

  n= 40, number of events= 36

          coef exp(coef) se(coef)      z Pr(>|z|)
group2 -0.5686    0.5663   0.3472 -1.638    0.101

       exp(coef) exp(-coef) lower .95 upper .95
group2    0.5663      1.766    0.2868     1.118

Concordance= 0.581  (se = 0.05 )
Rsquare= 0.064   (max possible= 0.994 )
Likelihood ratio test= 2.64  on 1 df,   p=0.1041
Wald test            = 2.68  on 1 df,   p=0.1015
Score (logrank) test = 2.75  on 1 df,   p=0.09751
```

Cox regression: Prediction

- In the previous models, we discuss a prediction of a new data point (given a fitted model)
- The predicted value is a point (mean value)
- In Cox regression, the prediction of a new data point x^* is a function

$$\widehat{h}(t; x = x^*) = \widehat{h}_0(t) \exp\{x^* \widehat{\beta}\}$$

where

$$\widehat{h}_{0i}(t) = \frac{1}{\sum_{j \in R(t_i)} \exp\{x_j \widehat{\beta}\}}$$

and $R(t_i)$ is a 'risk set' at time t_i

- Since our interest is $\widehat{S}(t; x = x^*)$, this is translated into

$$\widehat{S}(t; x = x^*) = \exp\{-H(t; x = x^*)\}$$

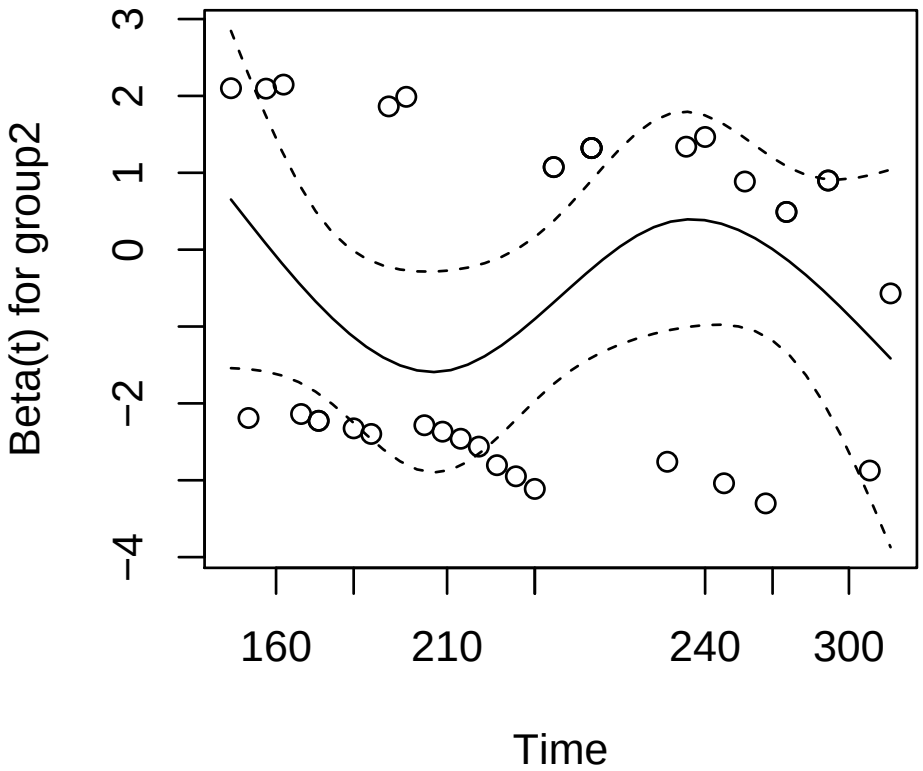
where $H(t; x = x^*)$ is the cumulative hazard (integrated over time)

Cox regression: Model diagnostics

- To check whether the proportional hazard assumption holds, we usually investigate through the so called 'scaled Schoenfeld residuals'
- Residuals for each predictor
- If the proportional hazard holds, the scaled residuals will be constant across times

Cox regression: Model diagnostics

```
res = coxph(y.rat ~ group, data=rat)
res2 <- cox.zph(res)
> print(res2)
          rho  chisq     p
group2 0.0491 0.0786 0.779
> plot(res2)
```



END

Any Questions?