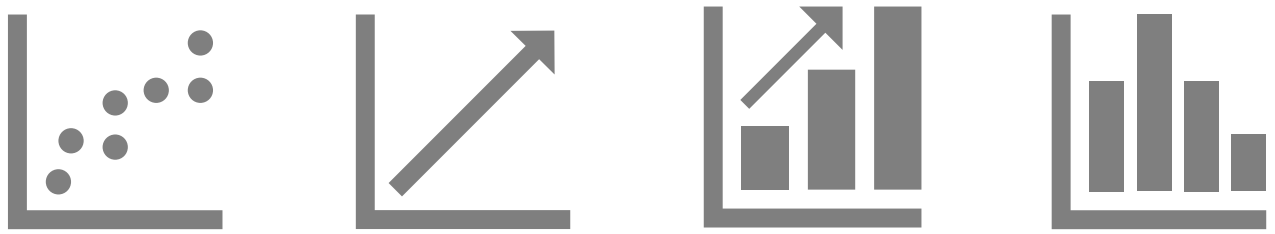
Association between variables



Clinical Oncology

Curriculum 2021

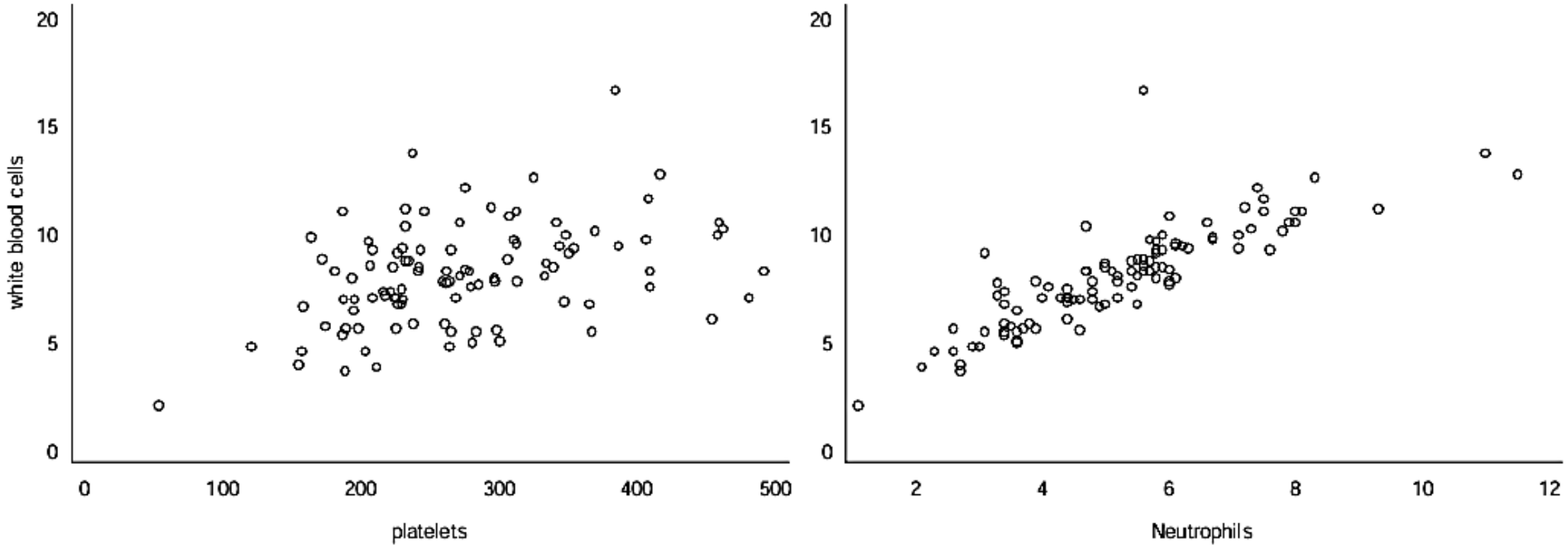
The Royal College of Radiologists

Medical statistics module

3.5 Association between variables	▪ Interpret the meaning of correlation and regression analysis ▪ Interpret the meaning of scatter plots

Correlation

- Correlation is used to measure the degree of linear association between two continuous variables

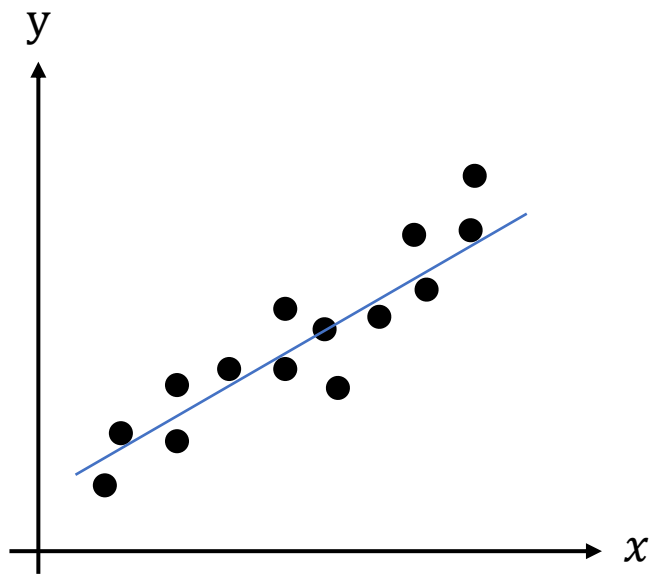


Scatter Plot

- A 2-dimensional scatter diagram

- Visualises the relationship between 2 variables
  - x → horizontal axis
  - y → vertical axis

- Plots the points of all individuals in a sample (n)



- The most appropriate approximation to the observed relationship between x and y

- If a straight line can be drawn through the midst of the points → linear relationship

Correlation coefficient

- The **strength** of the association is summarised by the correlation coefficient.
- Measures how close the observations are to the straight line that best describes their linear relationship

- There are two main types of correlation coefficients:
  - Pearson's correlation coefficient ($\boldsymbol{r}$)
    - relies on assumptions of Normality of the data

  - Spearman's rank correlation coefficient ($\boldsymbol{r_s}$)
    - Non-parametric alternative
    - Uses:
      - if data are not approximately normally distributed (x and y)
      - have extreme values (outliers)
      - the sample size is small
      - At least one of the variables is measured on an ordinal scale

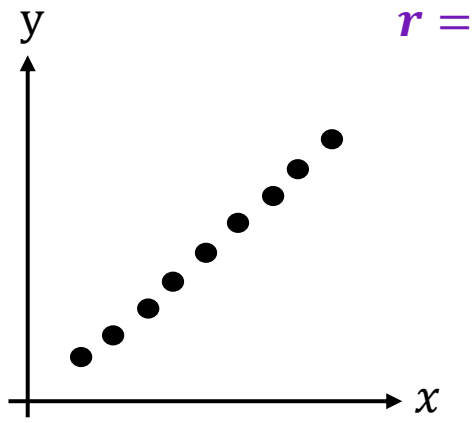
Correlation coefficient

- The correlation coefficient ($\boldsymbol{r}$) can take any value in the range -1 to +1.

- The **sign** of the correlation coefficient indicates the **direction** (+/−):
  - **+** one variable increases as the other variable increases (positive $\boldsymbol{r}$)
  - **−** one variable decreases as the other increases (negative $\boldsymbol{r}$)

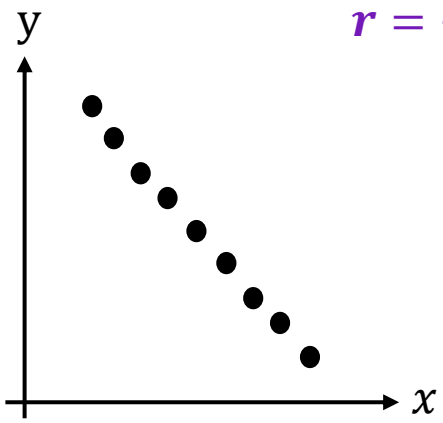
- The **magnitude** of the correlation coefficient indicates the **strength** of the linear association.
  - If $\boldsymbol{r}$ = +1 or −1 → Perfect correlation!
  - If $\boldsymbol{r}$ = 0 → No linear correlation.
  - The closer $\boldsymbol{r}$ is to -1 or +1, the greater the degree of linear association.

Examples



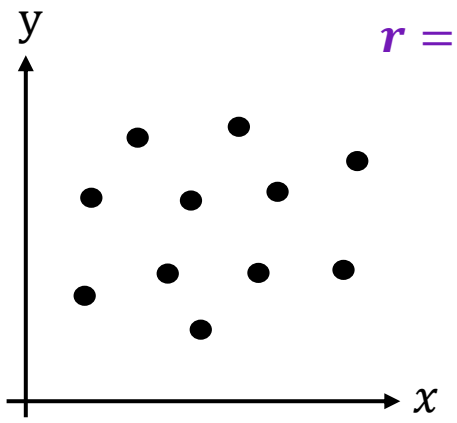
$r = 1$

Strong positive



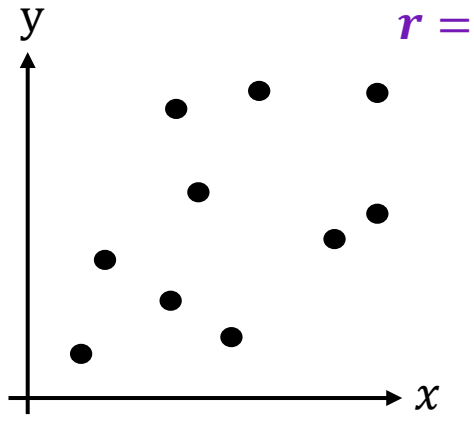
$r = -1$

Strong negative



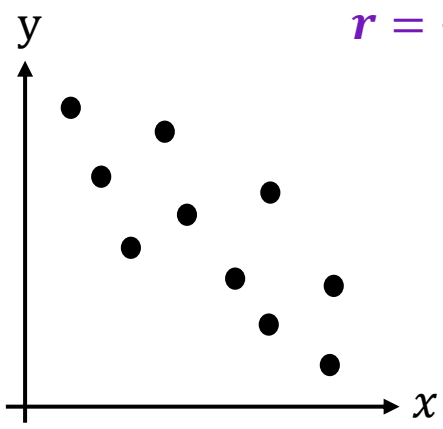
$r = 0$

No correlation



$r = 0.2$

Weak positive

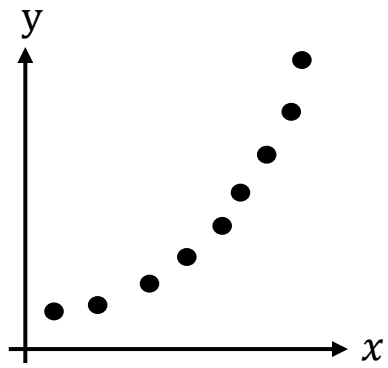


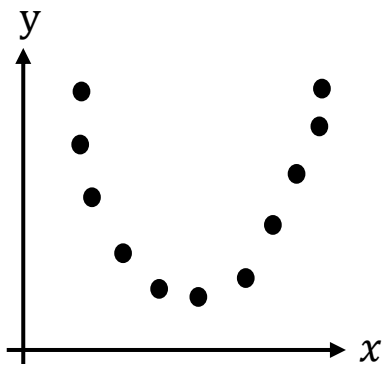
$r = -0.4$

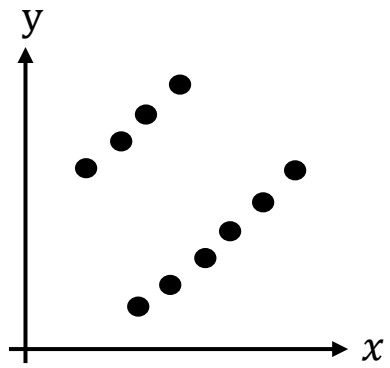
Weak negative

When not to calculate r?

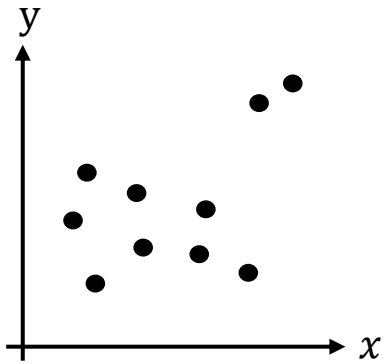
- When there is a non-linear relationship between two variables





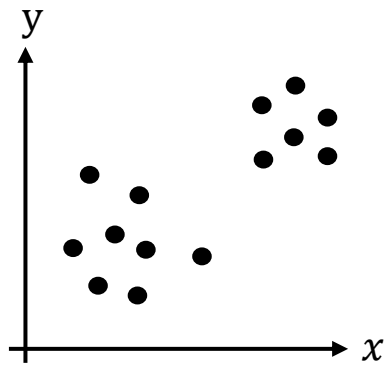


- When outliers are present



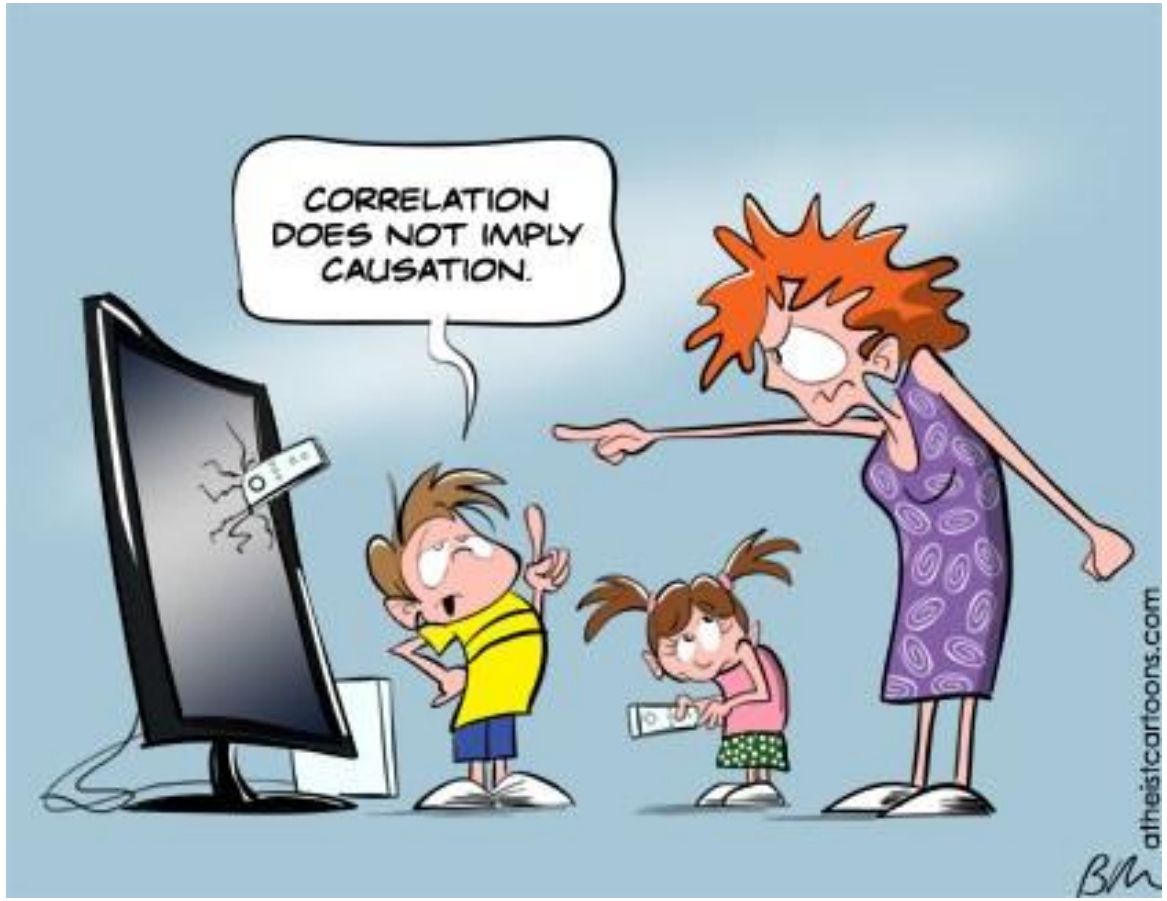
When not to calculate r?

- Data comprise subgroups of individuals for which the mean level of observations on at least one of the variables are different



- The data include more than one observation for each individual

Correlation between two variables does not necessarily imply a **'cause and effect'** relationship.



Example in R

`PSAProstateCancer_df` *Factors associated with prostate specific antigen*

Description

This dataset, PSAProstateCancer_df, is a data frame containing data from a study by Stamey et al. (1989) to examine the association between prostate specific antigen (PSA) and several clinical measures in men about to receive a radical prostatectomy. The dataset includes 97 observations and 9 variables, each representing a factor potentially associated with PSA.

Usage

```
data(PSAProstateCancer_df)
```

Format

A data frame with 97 observations and 9 variables:

lcavol Logarithm of cancer volume (numeric).

lweight Logarithm of prostate weight (numeric).

age Age of the patient in years (integer).

lbph Logarithm of benign prostatic hyperplasia (numeric).

svi Seminal vesicle invasion (integer).

lcp Logarithm of cancer perineural invasion (numeric).

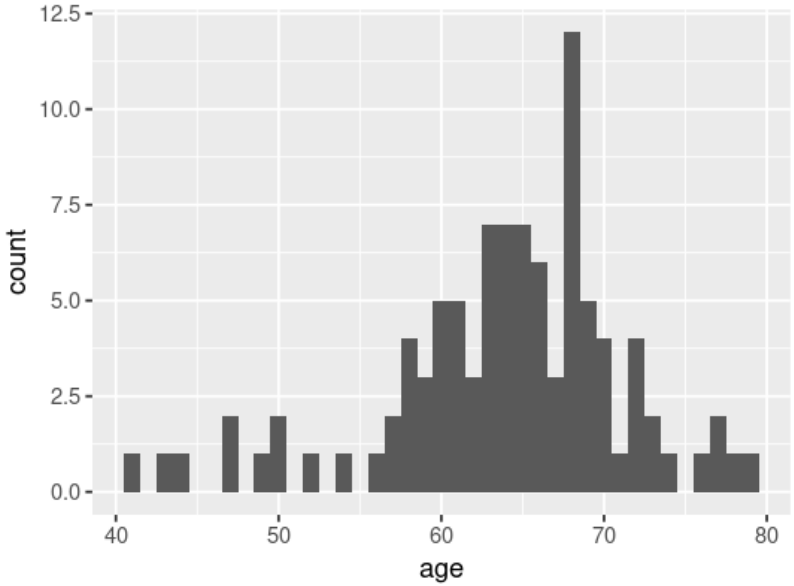
gleason Gleason score (integer).

pgg45 Percentage of cancerous tissue with Gleason score 4 or 5 (integer).

lpsa Logarithm of prostate specific antigen (PSA) (numeric).

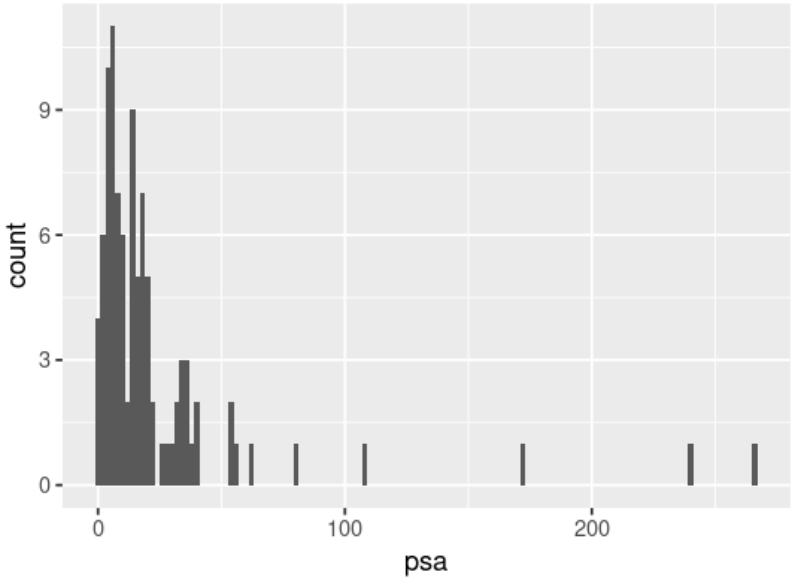
Example in R

Age (x)

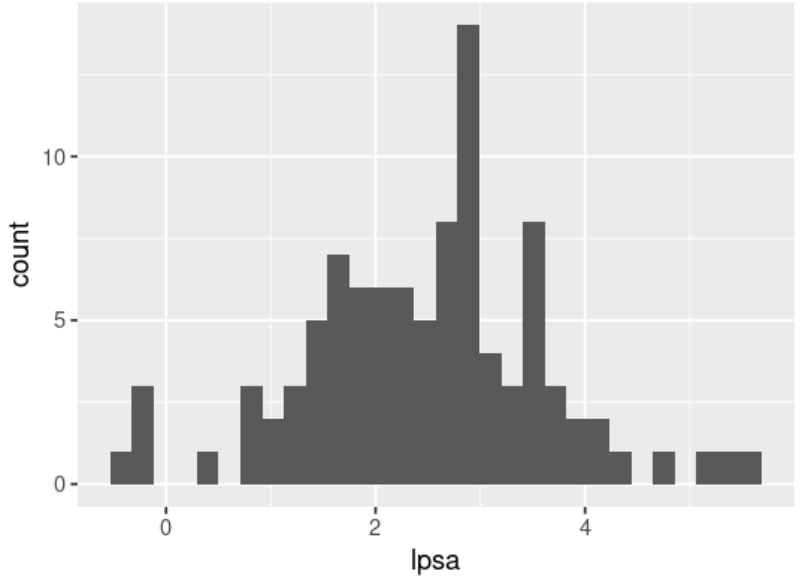



```
# Histogram for Age
PSAProstateCancer_df |>
  ggplot(aes(x = age)) +
  geom_histogram()
```

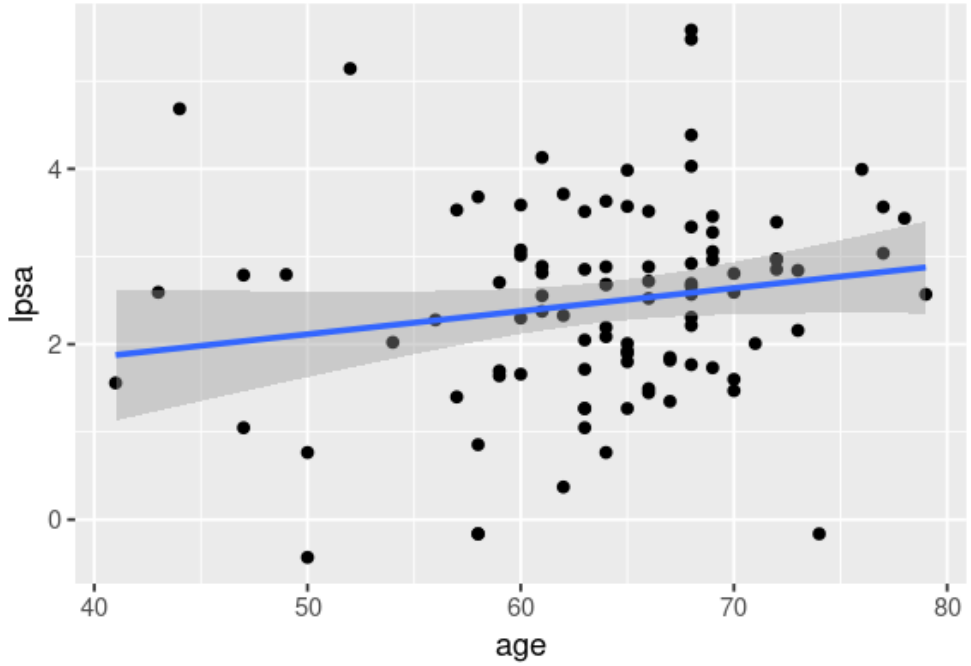
?PSA (y)



logPSA (y)



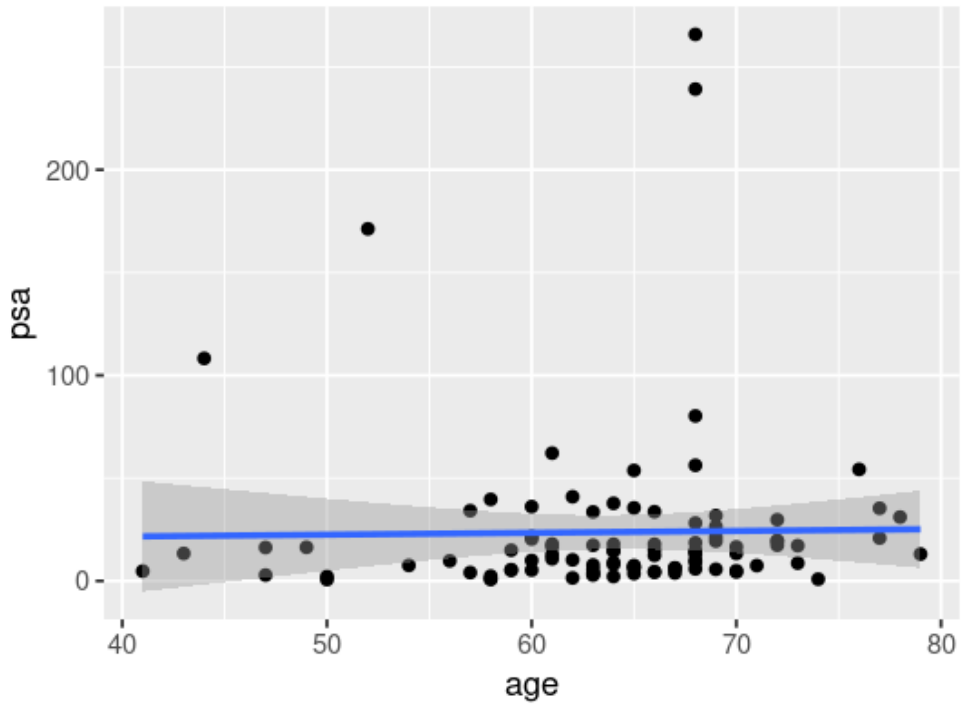
Example in R




```
> # Correlation
> cor.test(PSAProstateCancer_df$age, PSAProstateCancer_df$lpsa)

        Pearson's product-moment correlation

data:  PSAProstateCancer_df$age and PSAProstateCancer_df$lpsa
t = 1.6773, df = 95, p-value = 0.09677
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.0308976  0.3569640
sample estimates:
      cor
0.1695928
```

```
# Scatter plot (Age vs. log PSA)
PSAProstateCancer_df |>
  ggplot(aes(x = age, y = lpsa)) +
  geom_point() +
  geom_smooth(method = lm)
```

Example in R




```r
# Scatter plot (Age vs. PSA)
PSAProstateCancer_df |>
  ggplot(aes(x = age, y = psa)) +
  geom_point() +
  geom_smooth(method = lm)
```

```r
> # Correlation
> cor.test(PSAProstateCancer_df$age, PSAProstateCancer_df$psa)

        Pearson's product-moment correlation

data:  PSAProstateCancer_df$age and PSAProstateCancer_df$psa
t = 0.16088, df = 95, p-value = 0.8725
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1835457  0.2152406
sample estimates:
      cor
0.0165038
```

```r
> cor.test(PSAProstateCancer_df$age, PSAProstateCancer_df$psa,
method = "spearman")

        Spearman's rank correlation rho

data:  PSAProstateCancer_df$age and PSAProstateCancer_df$psa
S = 119372, p-value = 0.03431
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.2151562
```

Regression

- Regression looks for the dependence of one variable (the dependant variable - y) on another (the independent - x) variable.

- It quantifies the best linear relation between the variables and allows the prediction of the dependent variable when the independent variable is known.

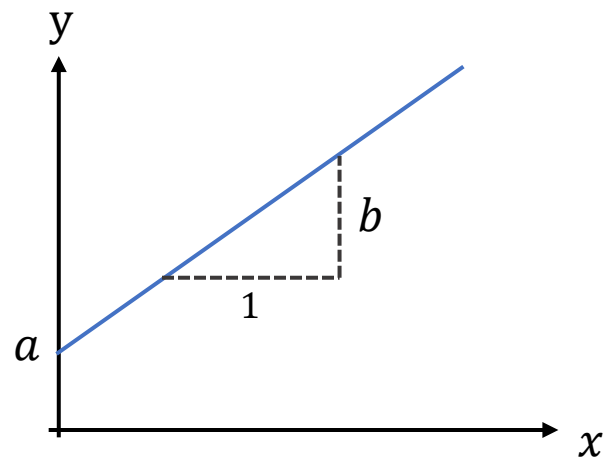
- Linear regression is used to determine the linear line (the regression of y on x) that best describes the straight-line relationship between the two continuous variables.

- Logistic regression is used when the outcome variable is binary as opposed to continuous.

Linear regression

- The equation which estimates the simple linear regression line:

$$Y = a + bx$$



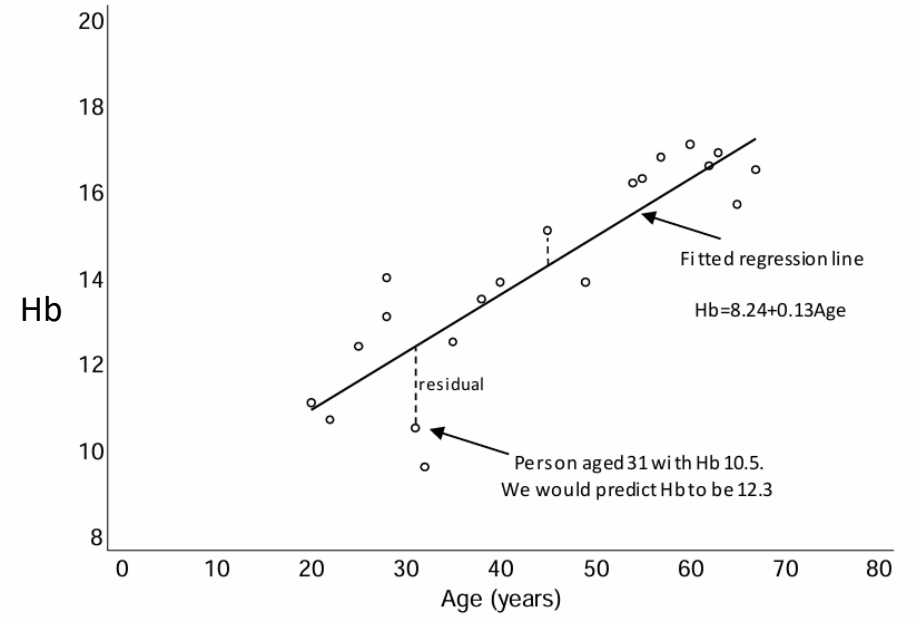
- x → independent, predictor or explanatory variable
- Y → dependent, outcome or response variable

- For a given value of x, Y is the value of y which lies on the estimated line. It is an estimate of the value we expect for y if the value of x is known.
- Y is called "the fitted" value of y.

- a → intercept of the estimated line; it is the average value of Y when $x = 0$
- b → slope/ gradient of the estimated line; it represents the amount by which Y increases on average if we increase x by one unit.

Linear regression

- The residual is the difference between the actual response y and the predicted response Y from the regression line.
- The intercept and slope are determined by the method of least squares (often called ordinary least squares, OLS).
- This method determines the line of best fit so that the sum of the squared residuals is at a minimum.
- The residuals are assumed to be Normally distributed and to have an average value of zero.



Regression coefficients

- The intercept (a) and the slope (b) are called "regression coefficients" of the estimated line. (often used for b)

Intercept

- The average value of the response when the predictor is 0.

Slope

- The average change in the response when the predictor increases by 1 unit.
- If the predictor was a binary variable the slope (b) would indicate the average difference in response between the two groups.

Confidence Intervals (CI)

- The intercept (a) and the slope (b) are sample estimates of corresponding population parameters.
- These estimates have an inherent variability which is used to provide **95% confidence intervals** for where the true population parameters may lie.
- The interval for the slope indicates the range, for the wider population, that the change in the response is likely to lie between as the predictor increases by 1 unit.
- If this interval includes 0, the coefficient is not statistically different from 0.

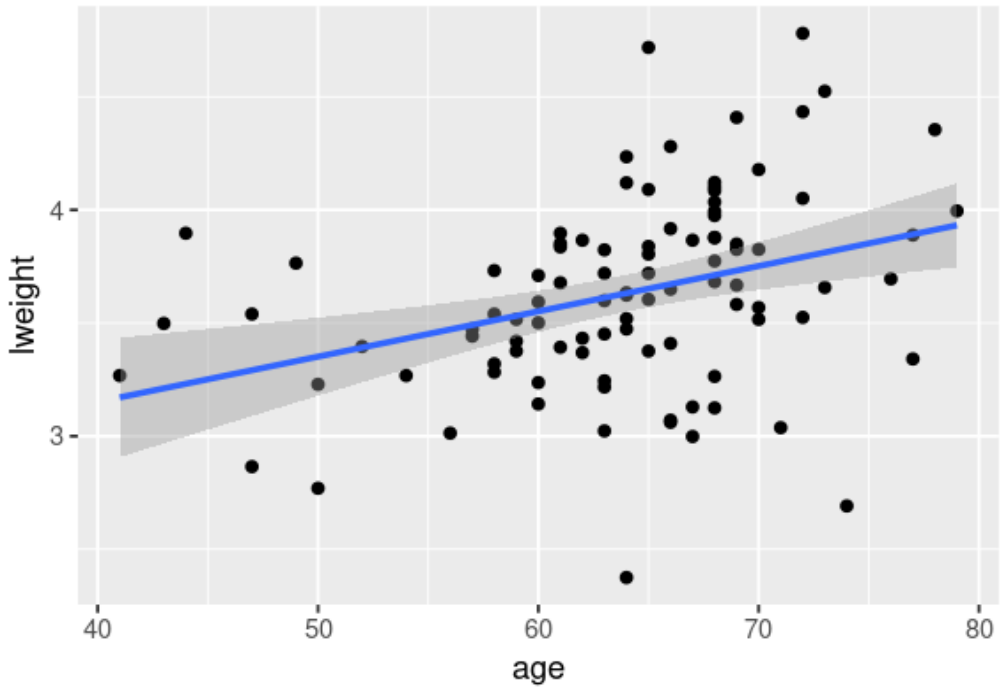
P-value

- Each coefficient has a p-value.
- The p-value relates to a test of the null hypothesis that the coefficient = 0, versus the alternative hypothesis that the coefficient ≠ 0.
- If p > 0.05 the null hypothesis cannot be discounted.

R-squared

- We can assess how well the line fits the data by calculating the "coefficient of determination" **R squared (R^2)**
- Usually expressed as a percentage ranging from 0 - 100%, which is equal to the square of the correlation coefficient.
- This represents the percentage of the variability of the response that can be explained by the predictor.
- The higher the R-squared, the better the model.

Example – Simple Linear Regression



model ← lm (y ~ x, data = data)

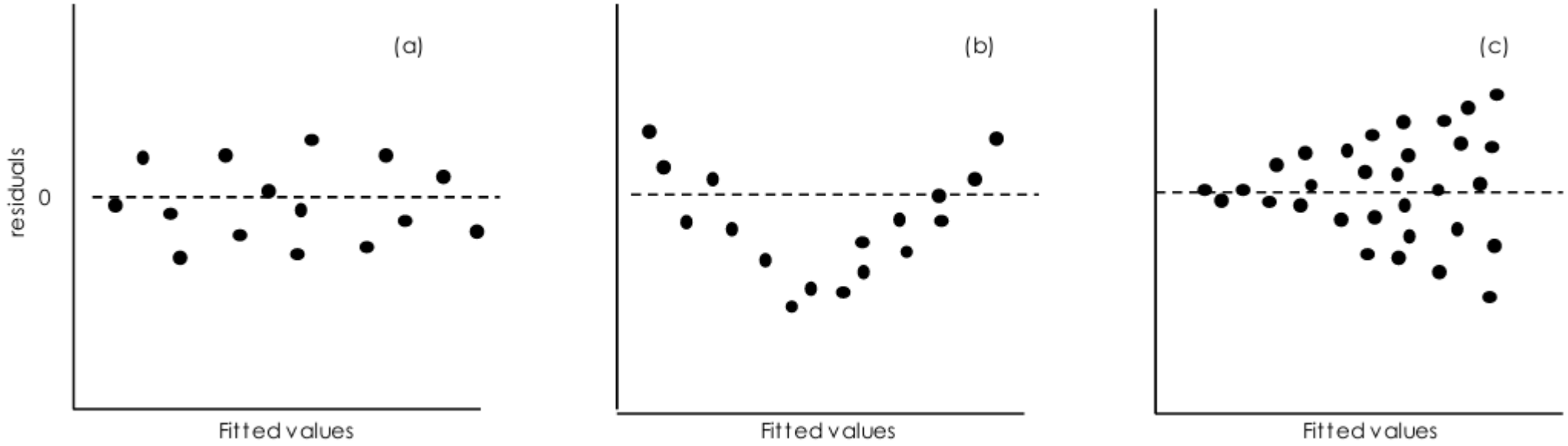
model ← lm (lweight ~ age, data = PSAProstateCancer_df)

```
> # Simple Linear regression
> model <- lm(lweight ~ age, data = PSAProstateCancer_df)
> summary(model)

Call:
lm(formula = lweight ~ age, data = PSAProstateCancer_df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.25672 -0.19186 -0.00317  0.25910  1.06640

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.350152   0.355854   6.604 2.29e-09 ***
age         0.020023   0.005535   3.618 0.000479 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4037 on 95 degrees of freedom
Multiple R-squared:  0.1211,    Adjusted R-squared:  0.1118
F-statistic: 13.09 on 1 and 95 DF,  p-value: 0.0004786
```

Assumptions of linear regression

1. **Linearity -** The relationship between the response and predictor is approximately linear.
2. **Independence -** The observations in the sample are independent.
3. **Normality -** The distribution of residuals is Normal.
4. **Homoscedasticity -** The residuals have constant variance.

- Assumptions can be checked by examining plots of the residuals.
- The most common method is to plot the residuals against the fitted values.
- This plot can show systematic deviations from a linear relationship and highlight non-constant variance.
- A Normal probability plot or histogram can be used to assess the Normality assumption of residuals.

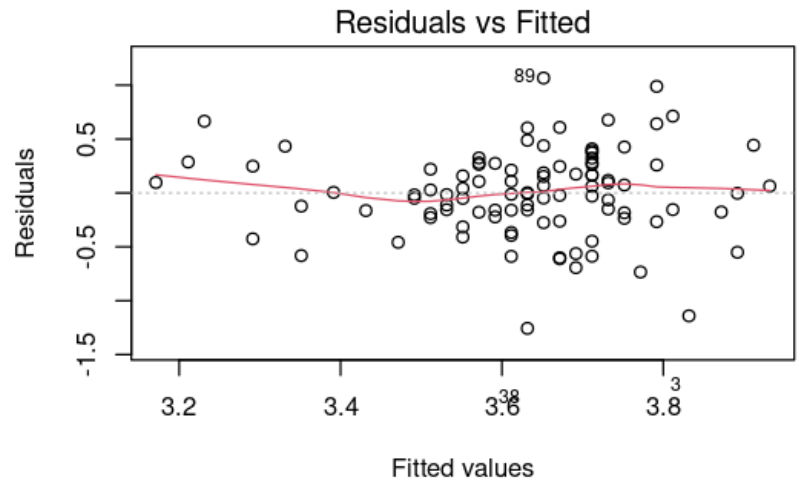


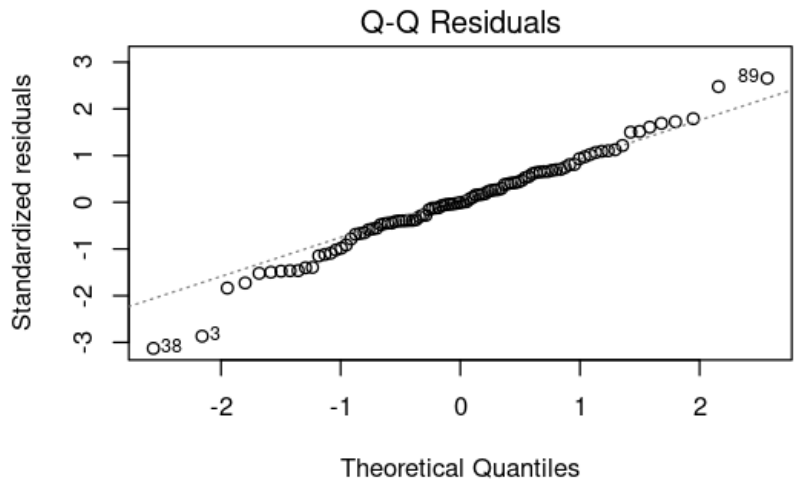
- A linear relationship means that across the range of fitted values, the residuals are spread equally above and below 0. **Figure (a)**
- The assumption does not hold in **Figure (b)**

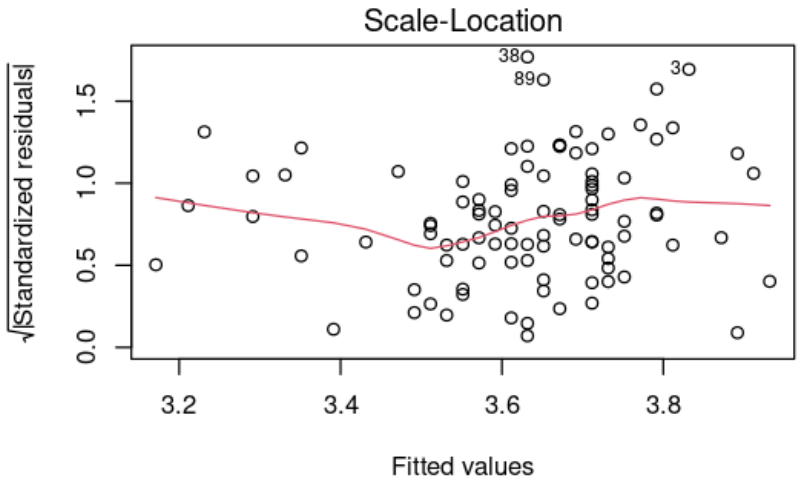
- Constant variance of the residuals means that in a plot of residuals against fitted values, the spread of the residuals doesn't change. **Figure (a)**
- The assumption does not hold if the vertical spread of the residuals changes across the plot as in **Figure (c)**

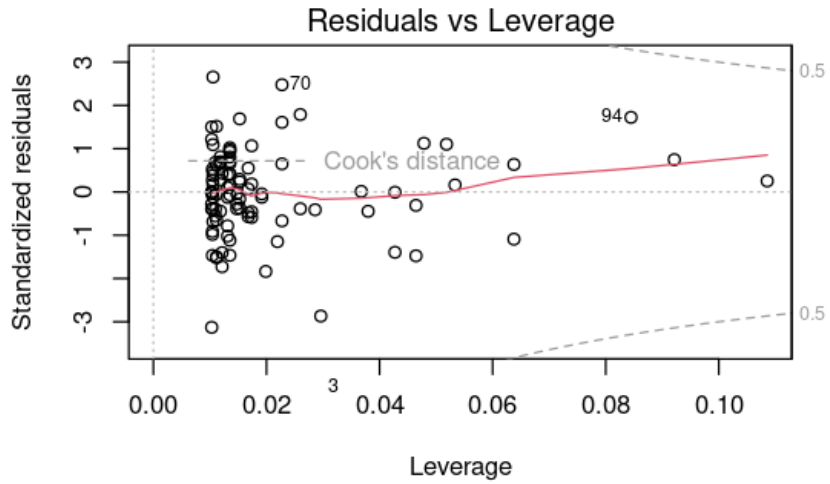
Model diagnostics

```
# Model diagnostics
par(mfrow = c(2, 2))
plot(model)
```



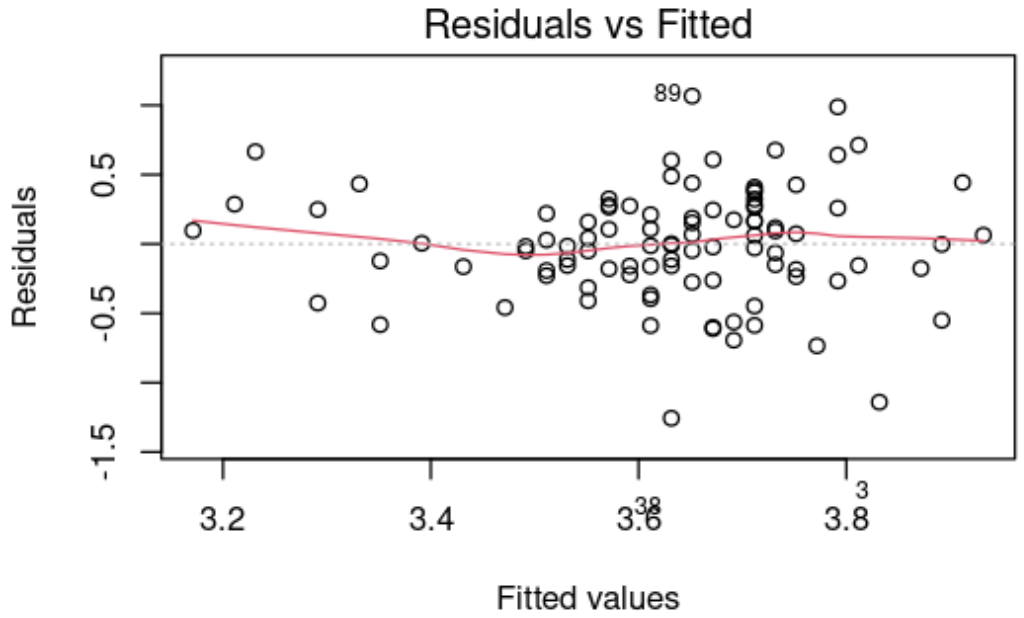






Linearity

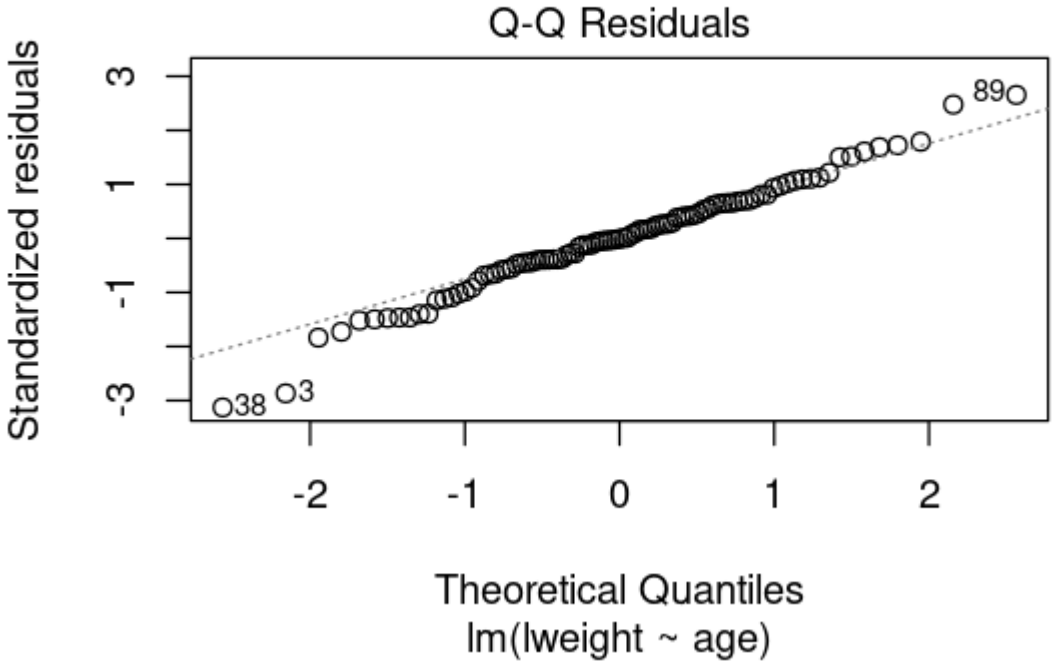
```
# Residuals vs. Fitted
plot(model,1)
```



- Ideally, the residual plot will show no fitted pattern. That is, the red line should be approximately horizontal at zero. The presence of a pattern may indicate a problem with some aspect of the linear model.
- In our example, there is no pattern in the residual plot. This suggests that we can assume linear relationship between the predictor and the outcome variables.
- Note that, if the residual plot indicates a non-linear relationship in the data, then a simple approach is to use non-linear transformations of the predictors, such as log(x), sqrt(x) and x^2, in the regression model.

Normality of residuals

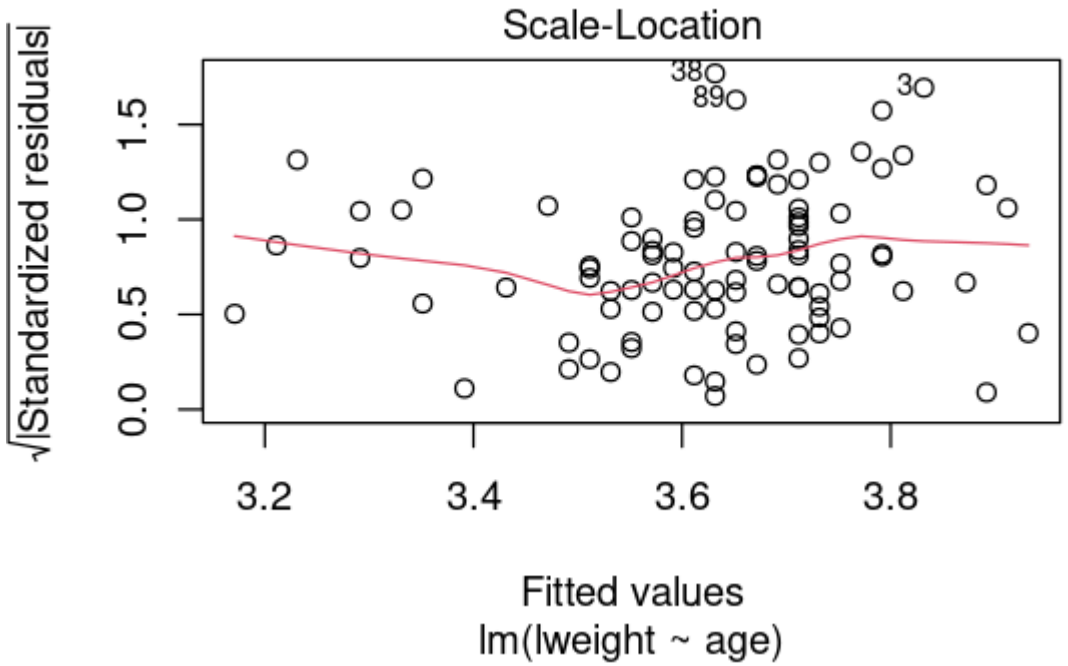
```
# Q-Q plot
plot(model,2)
```



- The QQ plot of residuals can be used to visually check the normality assumption.
- The normal probability plot of residuals should approximately follow a straight line.
- In our example, all the points fall approximately along this reference line, so we can assume normality.

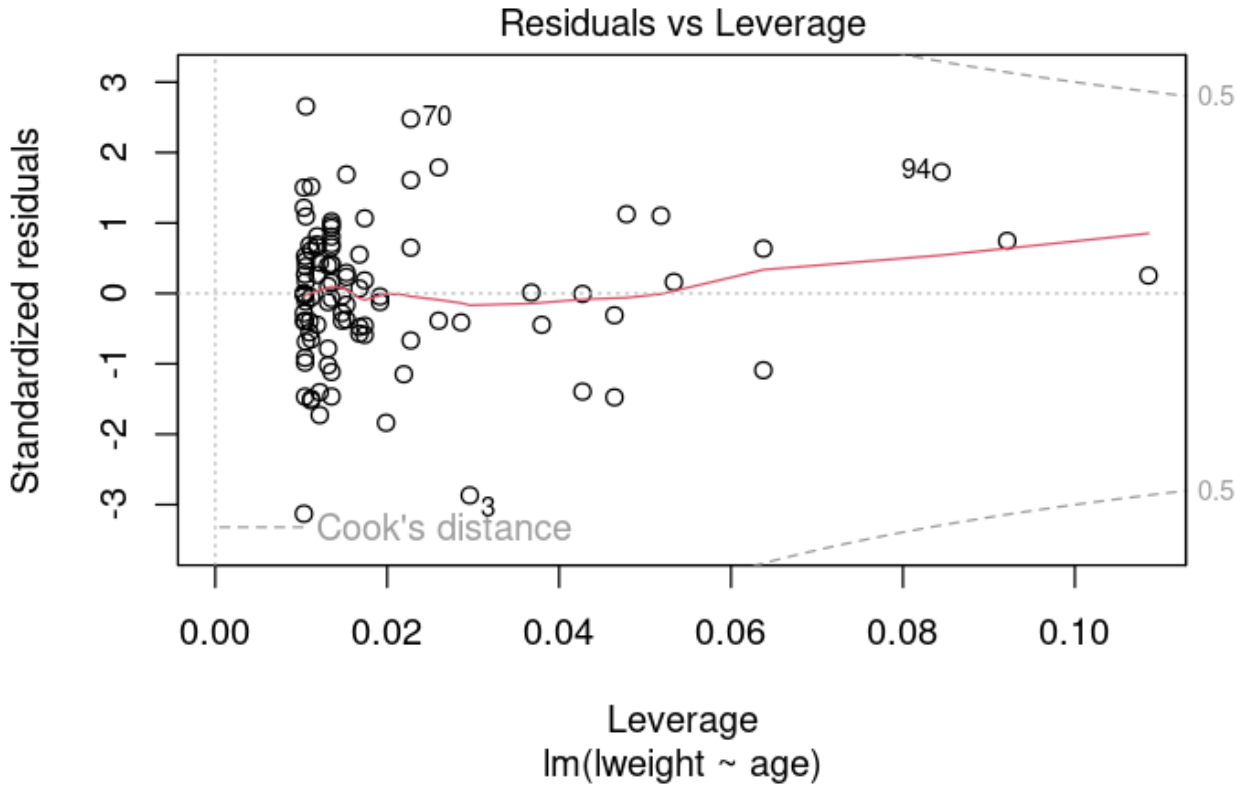
Homoscedasticity (Homogeneity of variance)

```
# Scale-location plot
plot(model,3)
```



- This assumption can be checked by examining the scale-location plot, also known as the spread-location plot.
- This plot shows if residuals are spread equally along the ranges of predictors. It's good if you see a horizontal line with equally spread points.

Outliers



- The plot above highlights the top 3 most extreme points (#3, #70 and #94), with a standardized residuals below -2 and above +2.
- However, there is no outliers that exceed 3 standard deviations, which is good.

Multiple Linear Regression

- Multiple linear regression is an extension of simple linear regression.
- Used to predict a response using several predictor variables.
- The assumptions of multiple linear regression are the same as those for simple linear regression.

Example in R

model ← lm (y ~ x1 + x2 + x3 + x4, data = data)

model ← lm (lweight ~ age + lpsa + lbph + gleason, data = PSAProstateCancer_df)

- **Interpretation:**
  - The intercept (a) is the average value of the response when all the predictors have values equal to zero.
  - If a predictor is continuous, it's slope (b) indicates the average change in the response when all the other predictors are held constant.
  - If a predictor is binary the slope (b) represents the average difference in the response between the groups when all other predictors are held constant.

- In multivariable the **adjusted R^2** is employed to assess the fit of the model.
- The **adjusted R^2** takes account of the number of predictors used in the model.
- Its interpretation is the same as for R-squared.

Example – Multiple Linear Regression

```
> # Multiple Linear Regression
> model <- lm(lweight ~ age + lpsa + lbph + gleason, data = PSAProstateCancer_df)
> summary(model)

Call:
lm(formula = lweight ~ age + lpsa + lbph + gleason, data = PSAProstateCancer_df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.05232 -0.22186  0.03222  0.20483  1.00815

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.144840   0.415853   7.562 2.90e-11 ***
age          0.012609   0.005233   2.410 0.017957 *
lpsa         0.150646   0.033300   4.524 1.81e-05 ***
lbph         0.090422   0.026213   3.449 0.000849 ***
gleason     -0.104202   0.054008  -1.929 0.056767 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3453 on 92 degrees of freedom
Multiple R-squared:  0.3776,    Adjusted R-squared:  0.3505
F-statistic: 13.95 on 4 and 92 DF,  p-value: 6.196e-09
```

Logistic Regression

- Logistic regression is similar to linear regression
- Used to predict a **binary outcome** of interest

- Simple logistic regression – One predictor
- Multiple logistic regression – Several predictors

Example in R

model ← glm (y ~ x1 + x2, data = data, family = binomial)

model ← glm (lweight ~ age + sex, data = Melanoma_df, , family = binomial)

Example in R

`Melanoma_df` *Survival from Malignant Melanoma*

Description

This dataset, Melanoma_df, is a data frame containing information about 205 patients with malignant melanoma (a type of skin cancer) who underwent a radical operation at Odense University Hospital, Denmark, between 1962 and 1977. Patients were followed up until the end of 1977. By that time, 134 patients were still alive, and 71 had died (57 due to cancer and 14 from other causes). This dataset provides detailed clinical and demographic information for studying malignant melanoma outcomes.

Usage

```
data(Melanoma_df)
```

Format

A data frame with 205 observations and 7 variables:

time Follow-up time in days (integer).

status Patient's status at the end of the study: 1 = alive, 2 = dead from cancer, 3 = dead from other causes (integer).

sex Sex of the patient: 1 = male, 2 = female (integer).

age Age of the patient at the time of surgery (integer).

year Year of surgery (integer).

thickness Tumor thickness in millimeters (numeric).

ulcer Presence of ulceration: 1 = no, 2 = yes (integer).

```r
#1. Simple logistic regression
model <- glm(ulcer ~ age, data = Melanoma_df, family = binomial)
summary(model)
exp(coef(model))
```

```
> #1. Simple logistic regression
> model <- glm(ulcer ~ age, data = Melanoma_df, family = binomial)
> summary(model)

Call:
glm(formula = ulcer ~ age, family = binomial, data = Melanoma_df)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.066856   0.482628  -2.211   0.0271 *
age          0.015578   0.008696   1.792   0.0732 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 281.13  on 204  degrees of freedom
Residual deviance: 277.84  on 203  degrees of freedom
AIC: 281.84

Number of Fisher Scoring iterations: 4

> exp(coef(model))
(Intercept)         age
  0.3440885   1.0157004
```

```r
#2. Multiple logistic regression
model <- glm(ulcer ~ age + sex, data = Melanoma_df, family = binomial)
summary(model)
# Odds Ratio
exp(coef(model))
```

```
> #2. Multiple logistic regression
> model <- glm(ulcer ~ age + sex, data = Melanoma_df, family = binomial)
> summary(model)

Call:
glm(formula = ulcer ~ age + sex, family = binomial, data = Melanoma_df)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.27759    0.49821  -2.564   0.0103 *
age          0.01457    0.00880   1.655   0.0978 .
sex          0.67257    0.29375   2.290   0.0220 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 281.13  on 204  degrees of freedom
Residual deviance: 272.55  on 202  degrees of freedom
AIC: 278.55

Number of Fisher Scoring iterations: 4

> # Odds Ratio
> exp(coef(model))
(Intercept)         age         sex
  0.2787092   1.0146744   1.9592614
```


QUIZ TIME

Select all of the following statements which you believe to be true.

Q1. The Pearson correlation coefficient between two variables, x and y:

A. Is always positive.

B. Is dimensionless.

C. Takes the same value when the variables x and y are interchanged.

D. Takes the value zero when there is no linear association between the two variables x and y.

E. Takes the value + 1 when one variable increases as the other variable decreases in value, and it is possible to draw a straight line on the scatter diagram with all the points lying on it.

Q2. Interpretation of Pearson's correlation coefficient

The estimated Pearson correlation coefficient between systolic blood pressure (mmHg) and age (years) in a sample of 30 middle-aged women from a given community was **r = 0.72** ($P < 0.001$). Hence **r^2 = 0.52.**

A. There is substantial evidence that systolic blood pressure and age in these women are linearly related.

B. 72% of the variability of systolic blood pressure in these women can be explained by its linear relationship with age.

C. 48% of the variability of systolic blood pressure in these women is unexplained by its linear relationship with age.

D. We can conclude that increasing age is a cause of rising systolic blood pressure in these women.

E. The null hypothesis that has been tested is that there is no association between systolic blood pressure and age in these women.

Q3. Spearman's correlation coefficient

50 subjects with alcoholic cirrhosis underwent an interview to assess the reliability and validity of historical variables such as duration of sobriety, duration and quantity of drinking and treatment history on the assessment of an individual's alcohol history. In addition, an alternative source close to each subject, usually a spouse, was interviewed by a second interviewer, who was blind to the subject's alcoholism history.

Duration of sobriety correlated highly between subject and the alternative source (Spearman's r = 0.96, P = 0.0001) as did the individual's score on the High-risk Alcoholism Relapse Scale (HRAR, Spearman's r = 0.72, P = 0.0001).

A. Spearman's correlation coefficient was used for these data because both variables were Normally distributed.

B. We can conclude that there is a linear relationship between the duration of sobriety as assessed by the subject and their collateral source.

C. If the authors had taken a larger sample size, they would have been able to calculate the Pearson's correlation coefficient for these data.

D. Spearman's correlation has provided a measure of association between the HRAR scores as assessed by the subject and their collateral source.

E. We can conclude that 92.2% (=0.96 x 0.96) of the variability in the duration of sobriety as assessed by the subject can be 'explained' by the variability in the duration of sobriety as assessed by their collateral sources.

Q4. The slope of the linear regression line between an explanatory variable, x, and a dependent variable, y, is:

A. The same as the gradient of the line.

B. The value of Y when x = 0, where Y is the predicted value of y.

C. The average change in Y for a unit increase in x.

D. Always positive.

E. Often called the regression coefficient.

References

- Introduction to Medical Statistics - The Beatson West of Scotland Cancer Centre
- Medical Statistics at a Glance – Petrie & Sabin 3rd Ed 2009 (Wiley-Blackwell)
- Medical Statistics at a Glance – WORKBOOK (Quiz)
  https://books.wiley.com/titles/9781119167815/multiple-choice-questions/
- Examples: OneDataSets package in R
  https://cran.r-project.org/web/packages/OncoDataSets/OncoDataSets.pdf