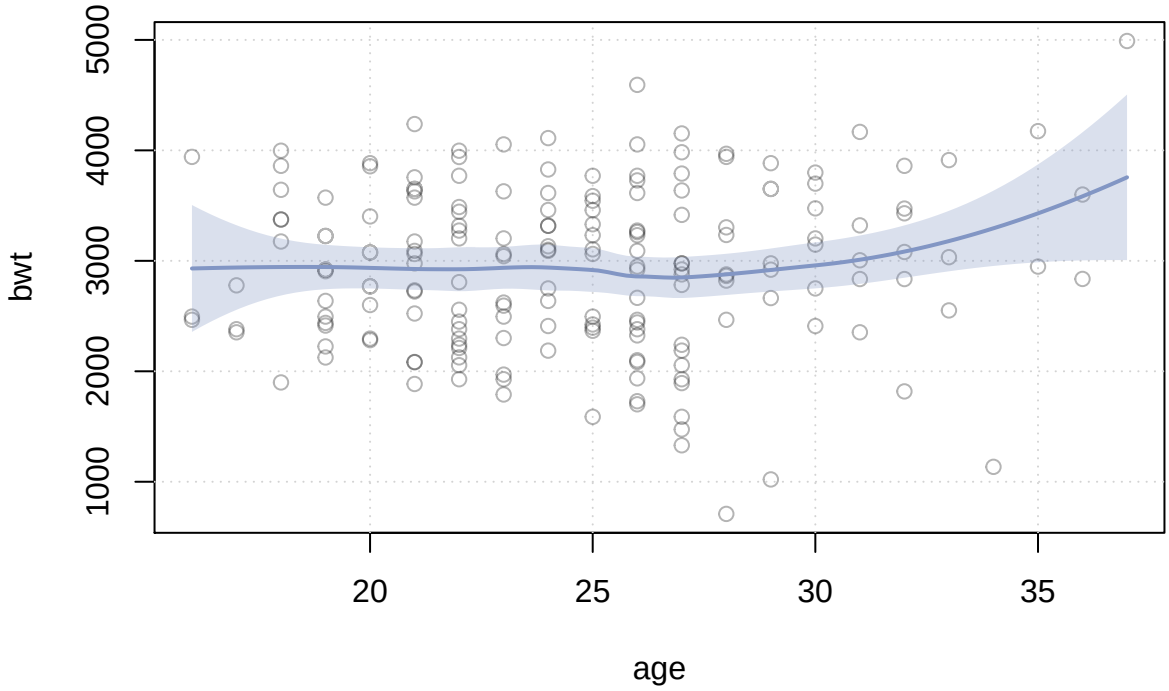
Applied Statistics using R

FACTSHEET

UNIT 4



Medical Statistics Team

University of Aberdeen

This booklet should not be reproduced without permission from the Medical Statistics Team

© Unless otherwise stated all content Copyright University of Aberdeen

Copyright Notice

Unless otherwise stated all content Copyright University of Aberdeen. The University of Aberdeen subscribes to the Copyright Licensing Agency's Higher Education Photocopying and Scanning Licence. You may access, download and print out a copy of any material included under the terms of this licence.

Any digital or print copy supplied to or made by you are for use in connection with this Course of Study. You may retain such copies after the end of the course, but strictly for your own personal use.

All copies (including electronic copies) shall include this Copyright Notice and shall be destroyed and/or deleted if and when required by the University of Aberdeen.

Except as provided by copyright law, no further copying, storage or distribution (including by e-mail) is permitted without the consent of the copyright holder.

The author has moral rights in the work. No distortion, mutilation, or other modifications of the work, or any other derogatory treatment prejudicial to the honour or reputation of the author is permitted.

Contents

Chapter 1

Correlation coefficient

This is a measure of the degree of association between two variables (x and y) initially assumed to be numerical.

1.1 Pearson's correlation coefficient

- Investigates the linear relationship between x and y (scatterplot of x and y)
- Assumes both x and y to be numerical and that at least one is normally distributed.
- Pearsons correlation is often denoted in the population by ρ and estimated from a sample by r.

H_0: There is no linear association in the population between the two variables x and y ($\rho = 0$).

H_1: There is some association in the population between the two variables x and y (2-sided test) ($\rho \neq 0$).

Interpretation

The *sign* indicates the direction of the relationship

- Positive correlation: r between 0 and 1
- No correlation: r close to 0
- Negative correlation: r between 0 and -1

The *magnitude* indicates the strength of the relationship (subject specific)

1.2 Spearman's rank correlation coefficient

- Investigates any *monotonic* relationship between x and y
- At least one variable should be continuous (the other could be ordinal) – they do not have to be normal (ranks)
- Spearman's Rank Correlation is often denoted in the population by (ρ_S) and estimated from a sample by r_S.

H_0: There is no association between the two variables x and y in the population ($\rho_S = 0$)

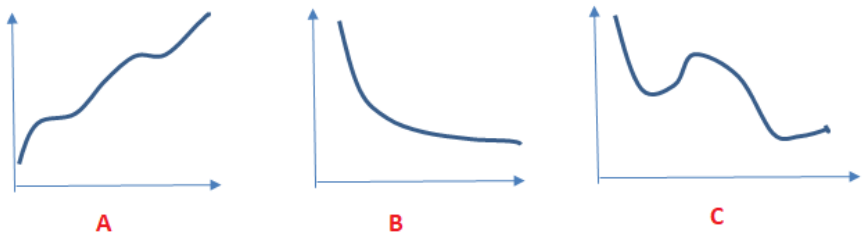
H_1: There is some association in the population between the two variables x and y in the population (2-sided test) ($\rho_S \neq 0$)

- Interpretation is same as Pearson's correlation coefficient.

Note: Correlation does NOT imply causation!

1.3 Monotonic relationship

A monotonic relationship between x and y is one where if x increases then y also will increase – perhaps not at a constant rate (i.e. is not always linear) but it does not change direction. Similarly if x increases, y may also continually decrease.



In the above figure, the relationhsip between x and y indicates (A) monotonically increasing, (B) monotonically decreasing and (C) not monotonic.

1.4 Using R

- We will use the data `cardiac.csv`. Please load the data into R as an R data.frame object `DF` if you have not loaded it already.
- The variable 'Age' is normally distributed; check the distribution of the variable.

- The *Pearson correlation coefficient* of 'Age' with either 'Triglycerides' or 'Alcohol' would be correct.

- However, both 'Triglycerides' and 'Alcohol' were skewed, and so together the better correlation would be the *Spearman correlation coefficient*.

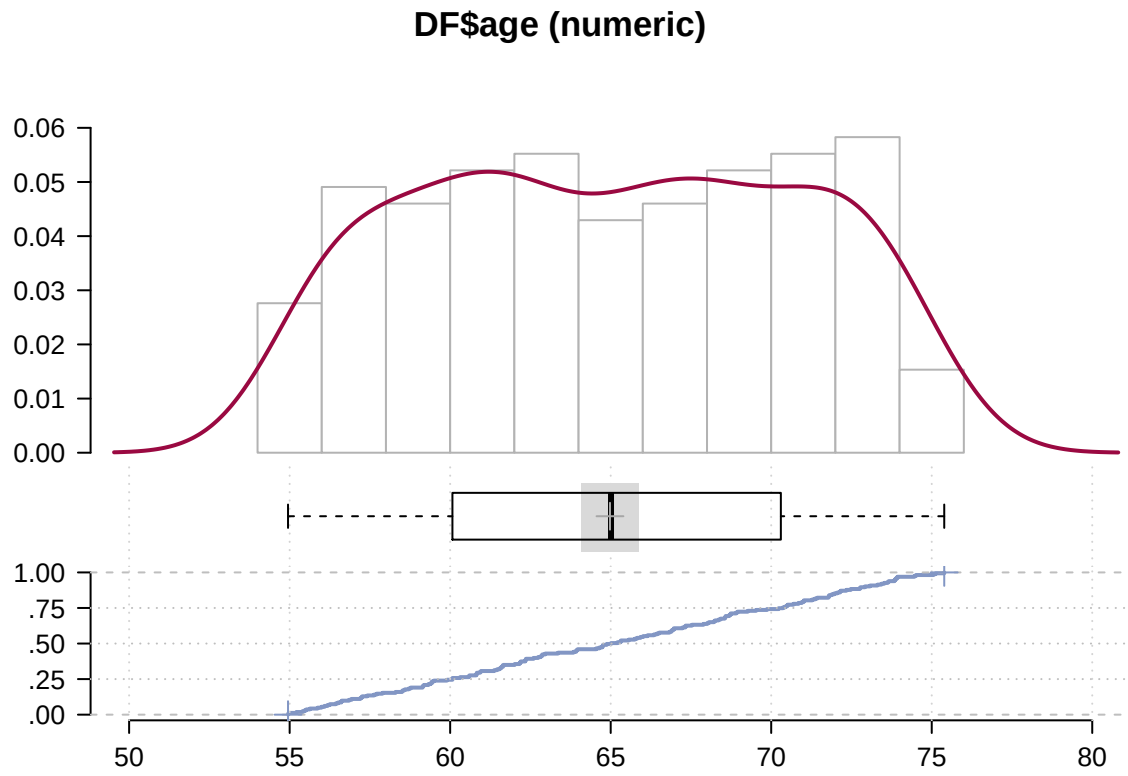
1.4.1 Summary Statistic

- It is useful to create the summary statistics to check the assumptions.

- Use the `DescTools::Desc` function to obtain the summary statistics and relevant plot.

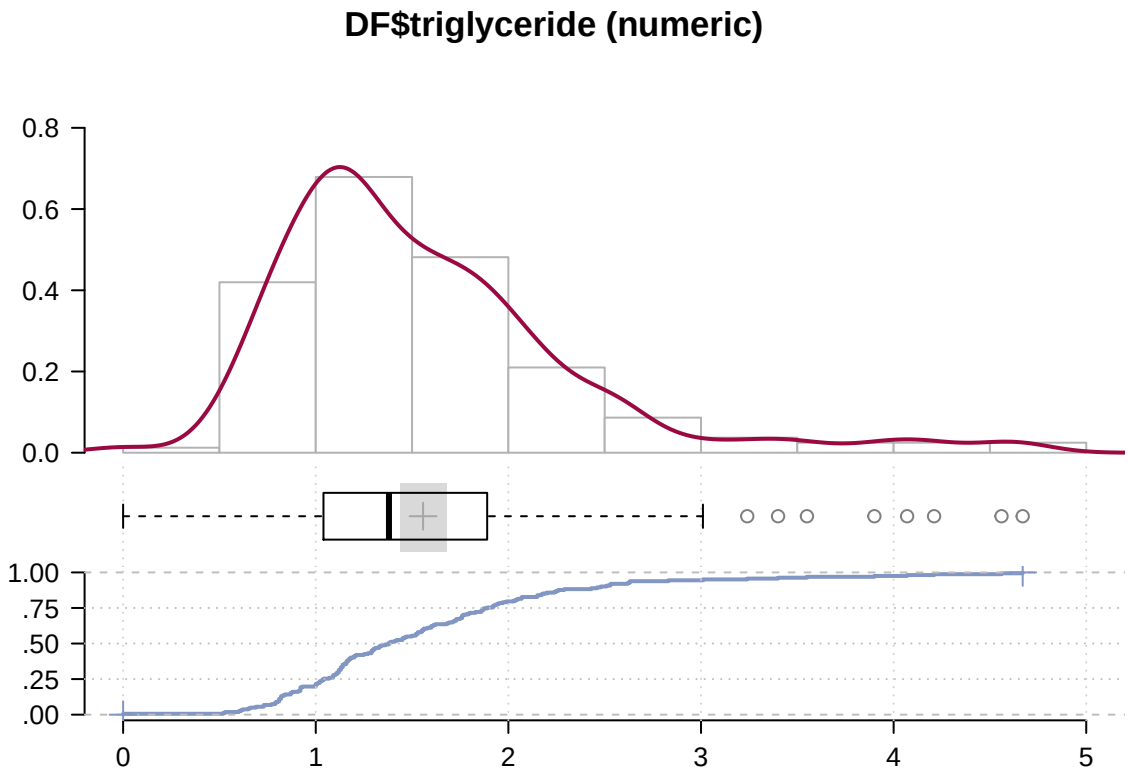
```
DescTools::Desc(DF$age, plotit = TRUE)
```

```
------------------------------------------------------------------------------
DF$age (numeric)

  length       n     NAs  unique      0s    mean  meanCI
     163     163       0     157       0  64.978  64.076
          100.0%    0.0%           0.0%          65.879

     .05     .10     .25  median     .75     .90     .95
  56.000  56.866  60.070  65.010  70.300  72.892  73.858

   range      sd   vcoef     mad     IQR    skew    kurt
  20.440   5.828   0.090   7.576  10.230  -0.002  -1.216

lowest : 54.95, 54.98, 55.21, 55.4, 55.47
highest: 74.46, 74.51, 75.05, 75.12, 75.39
```



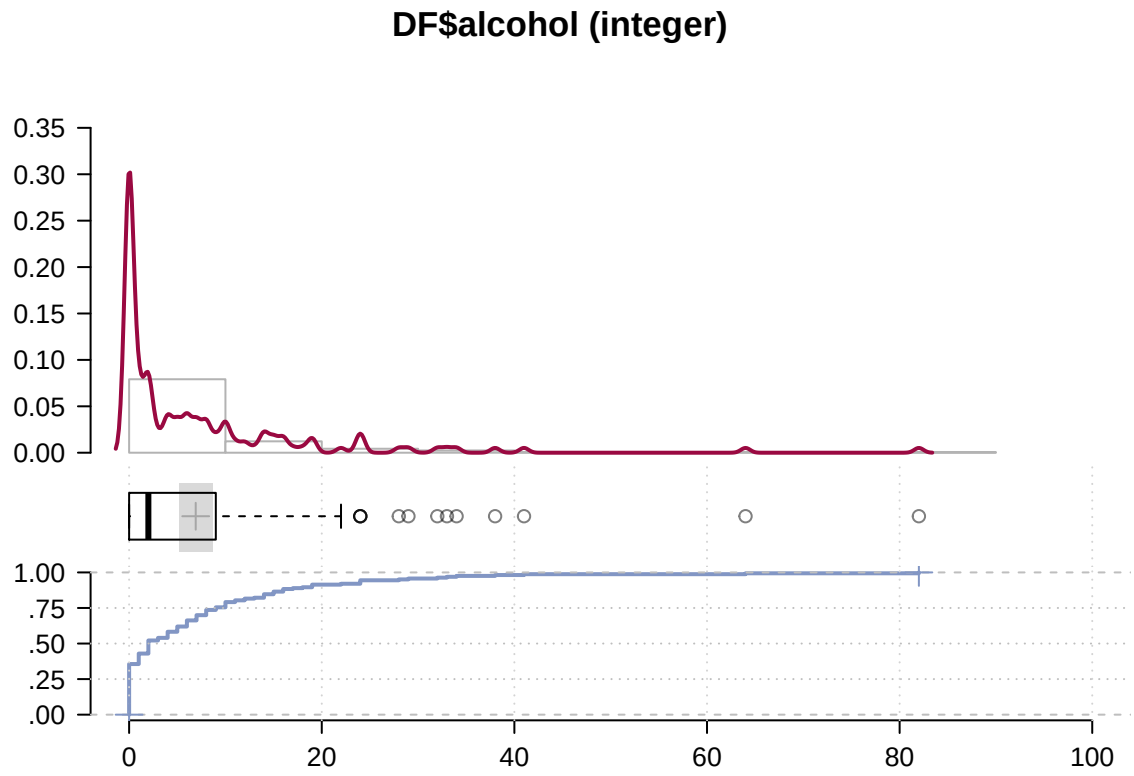

```
DescTools::Desc(DF$triglyceride, plotit = TRUE)
```

```
-------------------------------------------------------------------------------
DF$triglyceride (numeric)

  length       n     NAs  unique      0s    mean  meanCI
     163     162       1     113       1  1.5573  1.4367
           99.4%    0.6%           0.6%          1.6778

     .05     .10     .25  median     .75     .90     .95
  0.6920  0.8100  1.0475  1.3800  1.8850  2.4780  3.0010

   range      sd   vcoef     mad     IQR    skew    kurt
  4.6700  0.7769  0.4989  0.5708  0.8375  1.5543  3.2391

lowest : 0.0, 0.52, 0.53, 0.6, 0.61
highest: 3.9, 4.07, 4.21, 4.56, 4.67
```




```
DescTools::Desc(DF$alcohol, plotit = TRUE)
```

```
-------------------------------------------------------------------------
DF$alcohol (integer)

  length       n    NAs  unique     0s   mean  meanCI
     163     163      0      31     58   6.91    5.17
          100.0%   0.0%          35.6%           8.66

     .05     .10     .25  median    .75    .90     .95
    0.00    0.00    0.00    2.00   9.00  18.80   27.60

   range      sd  vcoef     mad     IQR   skew    kurt
   82.00   11.27   1.63    2.97    9.00   3.28   14.93

lowest : 0 (58), 1 (12), 2 (15), 3 (3), 4 (7)
highest: 34, 38, 41, 64, 82

heap(?): remarkable frequency (35.6%) for the mode(s) (= 0)
```




```
# pairs(DF[, c('age', 'triglyceride', 'alcohol')])
```

1.4.2 Correlation coefficient

- The R function `cor` estimates correlation coefficient using several methods between x and y if they are numeric vectors
- If x and y are matrices then the correlations between the columns of x and the columns of y are computed.
- The function can estimate the correlation coefficient using three methods through `method` argument:
  - Pearson correlation coefficient (`pearson`) - this is the default method
  - Kendal correlation coefficient (`kendall`) - not discussed in this course
  - Spearman correlation coefficient (`spearman`)
- The `kendall` and `spearman` options can consider both numeric and ordinal data.
- It provides several options to handle missing data using optional character string for the argument `use`.
  - Everything (`everything`) - a resulting value will be `NA` whenever one of its contributing observations is `NA`.

  - All observations (`all.obs`) - the presence of missing observations will produce an error
  - Complete observations (`complete.obs`) - casewise deletion (and if there are no complete cases, that gives an error)
  - Pairwise complete case (`pairwise.complete.obs`) - the correlation or covariance between each pair of variables is computed using all complete pairs of observations on those variables

- Generally, the default option `everything` or `pairwise.complete.obs` is commonly used
- The R function `cor.test` test is used to test the association between paired samples, using one of Pearson's product moment correlation coefficient, Kendall's τ or Spearman's ρ correlation coefficients.
- Check the help file for `cor` and `cor.test` for further details.

1.4.3 Pearson correlation coefficient and hypothesis testing

H_0: There is no linear association in the population between 'Age' and 'Triglyceride' ($\rho = 0$).

H_1: There is some association in the population between 'Age' and 'Triglyceride' (2-sided test) ($\rho \neq 0$).

```
cor(x = DF$age, y = DF$triglyceride,
    use = 'pairwise.complete.obs',
    method = 'pearson')
```

```
[1] -0.08025144
```

```
cor.test(x = DF$age, y = DF$triglyceride,
         alternative = 'two.sided',
         method = 'pearson')
```

```
    Pearson's product-moment correlation

data:  DF$age and DF$triglyceride
t = -1.0184, df = 160, p-value = 0.31
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2315812  0.0748705
sample estimates:
        cor
-0.08025144
```

1.4.4 Spearman correlation coefficient and hypothesis testing

H_0: There is no association in the population between between 'Age' and 'Triglyceride' ($\rho_S = 0$)

H_1: There is some association in the population between between 'Age' and 'Triglyceride' (2-sided test) ($\rho_S \neq 0$)

```
cor(x = DF$age, y = DF$triglyceride,
    use = 'pairwise.complete.obs',
    method = 'spearman')
```

```
[1] -0.07663933
```

```
cor.test(x = DF$age, y = DF$triglyceride,
         alternative = 'two.sided',
         method = 'spearman')
```

```
    Spearman's rank correlation rho

data:  DF$age and DF$triglyceride
S = 762865, p-value = 0.3324
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.07663933
```

1.4.5 Spearman correlation coefficient for three variables

```
cor(x = DF[, c('age', 'triglyceride', 'alcohol')],
    use = 'pairwise.complete.obs',
    method = 'spearman')
```

```
                    age triglyceride    alcohol
age           1.00000000  -0.07663933 -0.1178633
triglyceride -0.07663933   1.00000000  0.1330687
alcohol      -0.11786332   0.13306868  1.0000000
```

Note: The function `cor.test` does not accept a matrix input. You have to include separate vectors for testing the correlation coefficients or use the formula interface. Check the help file for `cor.test` for further details.

1.4.6 Steps to interpretation

The **Spearman rank correlation coefficient** between age and triglyceride is -0.077 which is not significantly different from zero ($p = 0.332$).

Conclusion:

There is no association between age and triglyceride values in the population.

Chapter 2

Cohen's Kappa for agreement

2.0.1 When is it appropriate?

If you wish to compare the agreement between two raters for a categorical variable.

2.0.2 Example

For this example, we will use the data file `lowbirthweight.csv`. Please read the data into R environment as a data.frame object `BW` as shown below. Check that the data file `lowbirthweight.csv` is in the current working directory while reading the file using the `read.csv` function.

```
BW <- read.csv(file = 'lowbirthweight.csv')
```

In a sample of women (N = 189), information was collected on anxiety (clinician reported or patient reported). Let's use the `table` function to find out the frequencies.

```
table(BW$AnxietyClin, BW$AnxietyPart)
```

```
   0  1
0 84 16
1 13 76
```

As observed in the above outputs, for 100 patients the clinician reported no anxiety, and the patient agreed in 84 cases, but not in 16. For the 89 patients whom the clinician reported as being anxious, 76 of the patients agreed.

Do the patient and clinician agree? We can use Cohen's Kappa to answer this question.

2.0.3 Summary Statistic

- Obtain the summary statistics to help in interpretation of the results.
- We can use the function `gmodels::CrossTable` from the `gmodels` package to obtain the summary statistics.
- Load the `library(gmodels)` if it is not already loaded
- Note that the `gmodels` package also provides the chi-squared test outputs; we do not need it for this exercise, therefore, included `FALSE` to those arguments.
- You can also use the function `DescTools::Desc` function to obtain the summary statistics.
- Use the `DescTools::Desc` function to obtain the summary statistics.

```
library(gmodels)

gmodels::CrossTable(x = BW$AnxietyClin, y = BW$AnxietyPart,
                    expected = FALSE, prop.r = TRUE, prop.c = TRUE, prop.t = FALSE,
                    prop.chisq = FALSE, chisq = FALSE, format = 'SPSS')
```

```
   Cell Contents
|-------------------------|
|                   Count |
|             Row Percent |
|          Column Percent |
|-------------------------|

Total Observations in Table:  189

                | BW$AnxietyPart
BW$AnxietyClin |         0  |         1  | Row Total |
---------------|------------|------------|-----------|
             0 |        84  |        16  |       100 |
               |    84.000% |    16.000% |    52.910% |
               |    86.598% |    17.391% |           |
---------------|------------|------------|-----------|
             1 |        13  |        76  |        89 |
               |    14.607% |    85.393% |    47.090% |
               |    13.402% |    82.609% |           |
---------------|------------|------------|-----------|
  Column Total |        97  |        92  |       189 |
               |    51.323% |    48.677% |           |
---------------|------------|------------|-----------|
```

2.0.4 Using R to calculate Kappa statistic

- The R function `DescTools::CohenKappa` from the `DescTools` library calculates the Kappa statistic.
- The input variable is a numeric vector of the corresponding columns (here, `AnxietyClin` and `AnxietyPart`)
- You can also put a table object as you have used for the `chisq.test`.
- Include the option `conf.level = 0.95` to calculate 95% confidence interval for the Cohen's Kappa statistic.
- Check the help file for `CohenKappa` for further details.

```
DescTools::KappaM(BW[, c('AnxietyClin','AnxietyPart')], method = 'Fleiss', conf.level = 0.
```

```
    kappa    lwr.ci    upr.ci
0.6925709 0.5500045 0.8351373
```

```
tab <- table(BW$AnxietyClin, BW$AnxietyPart)

tab
```

```
   0  1
0 84 16
1 13 76
```

```
CohenKappa(tab, conf.level = 0.95)
```

```
    kappa    lwr.ci    upr.ci
0.6926485 0.5897252 0.7955717
```

```
# Or, directly use the vector for each variable
# CohenKappa(x = BW$AnxietyClin, y = BW$AnxietyPart, conf.level = 0.95)
```

2.0.5 Steps to interpretation

From the cross-tabulation we can see the clinician and patient agree no 44.4%, agree yes 40.2%; so in total agree 84.6% of the time but disagree with each other 15.4% of the time.

The Kappa statistic is 0.693 with 95% interval does not include zero indicating good agreement between the clinician and patient.

Conclusion:

There is a good agreement between clinician reported anxiety and the patient reported anxiety.

Chapter 3

Bland-Altman Plot

3.1 Background

Bland-Altman plot is used to analyse the agreement between two methods or raters.

Both variables should be continuous variables, or scale variables.

We can write a simple script to produce Bland-Altman plots in the R environment.

Let us use the following sample data to create Bland-Altman plot using R.

StudyNo	Systolic1	Systolic2
1	130	125
2	125	120
3	131	127
4	132	130
5	130	130
6	119	123
7	129	128
8	137	136
9	107	114
10	109	110
11	127	138
12	136	132
13	121	108
14	118	113
15	117	104
16	138	130
17	117	115
18	125	125
19	96	105
20	133	120

3.2 Create Bland-Altman Plot

3.2.1 Step 1: Read the data

- Create your own dataset with the data shown above
- Copy the above data in Excel
- Create a csv file names 'BA.csv'
- Read the data into R environment as a data.frame object `BA` as shown below
- Note that the data file `BA.csv` is in the current working directory while using the `read.csv` function as shown below.

```
BA <- read.csv('BA.csv')
```

3.2.2 Step 2: Estimate the summary statistics

- Calculate the mean and difference of the two methods for each observation, i.e. Calculate MEAN and DIFFERENCE of `Systolic1` and `Systolic2` for each observation.

```
BA$Mean <- (BA$Systolic1 + BA$Systolic2)/2

BA$Diff <- BA$Systolic1 - BA$Systolic2
```

- Find out the mean and standard deviation of DIFFERENCE.

```
meanDiff <- mean(BA$Diff)
cat('meanDiff = ', meanDiff)
```

```
meanDiff =  2.2
```

```
meanSD <- sd(BA$Diff)
cat('meanSD = ', meanSD)
```

```
meanSD =  6.72466
```

The average difference (the mean of DIFFERENCE), in our sample data, is 2.2 and standard deviation is 6.725.

- Calculate UPPER and LOWER limits of the agreement
- The upper limits of agreement is calculated as: `2.2 + 1.96*6.725 = 15.4`
- The upper limits of agreement is calculated as: `2.2 - 1.96*6.725 = -11`

```
lower <- meanDiff - 1.96 * meanSD

upper <- meanDiff + 1.96 * meanSD
```

3.2.3 Step 3: Create the plot

- Create the scatter plot using MEAN on the X axis and DIFFERENCE on the Y axis.

- Create the graph to add the following three reference lines.

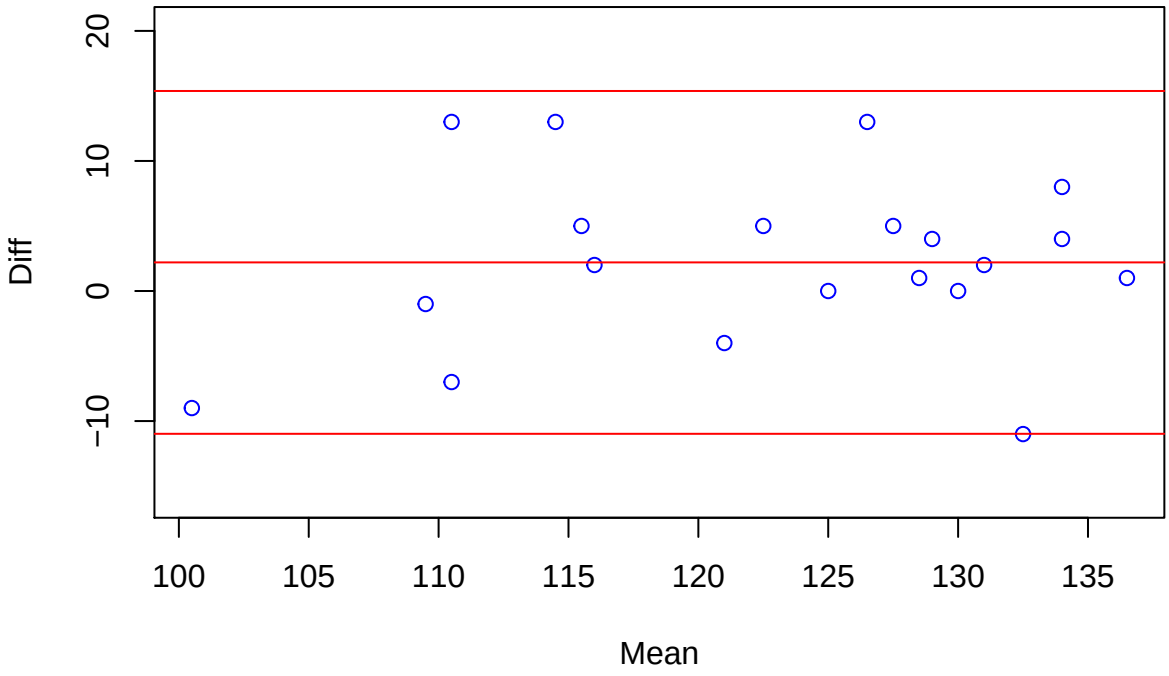
The average difference (the mean of DIFFERENCE), in our sample data, is 2.2. Add a reference line on the Y axis at 2.2.

Repeat the previous step to add reference lines at `2.2+1.96*6.725 = 15.4` (the upper limits of agreement)

Repeat the previous step to add reference lines at `2.2-1.96*6.725 = -11` (the lower limits of agreement)

This will produce the following Bland-Altman Plot for the sample data.

```
plot(Diff ~ Mean, data = BA, type = 'n', ylim = c(lower-5, upper+5))

points(Diff ~ Mean, data = BA, col = 'blue')

abline(h = meanDiff, lty = 1, col = 'red')

abline(h = lower, lty = 1, col = 'red')

abline(h = upper, lty = 1, col = 'red')
```



3.2.4 Interpretation

Bland-Altman plots could be interpreted informally without further analyses.

1. The estimate of bias shows how big the average discrepancy between two methods. This must be interpreted clinically to answer whether the discrepancy is large enough to be important. If there is a consistent bias, it can be adjusted for by subtracting the mean difference from the new method.

2. If the limits of agreement ($mean \pm 1.96 \times SD$) are wide (as defined clinically), the results are ambiguous. If the limits are narrow and the bias is very small, then the two methods are essentially equivalent and may be used interchangeably.

3. The plot reveals if there is a relationship between the differences and the magnitude of measurements. If the difference between methods tend to get larger (or smaller) as the average increases, it could be interpreted as a trend.

4. The plot is also useful to check if the variability is consistent across the range of values. For example, does the scatter around the bias line get larger as the average gets higher?

Chapter 4

Linear Regression

This is the investigation of the relationship between two independent continuous variables, X and Y (several X variables are for multiple linear regression). It can be used to predict Y (*dependent* or *response* variable) from the X (*independent* or *predictor* variable(s)).

4.1 Simple linear regression (SLR)

$$Y = \beta_0 + \beta_1 X + e$$

4.1.1 Assumptions

These have to be checked after conducting the regression since they depend on getting residuals (e in the above equation), however, results should only be interpreted once the assumptions have been checked.

- Linear relationship between X and Y: Check by correlation, scatterplot of X and Y, OR scatter plot of the Normalised Predicted values vs Dependent values
- Normal distribution of RESIDUALS: Check by a Histogram of Residuals, or by inspection of the P-P plots
- Constant variance: Check by a Scatterplot of Residuals vs Normalised Predicted (sausage shaped, random scatter)
- Independent observations: Check by Scatterplot of Normalised Residuals and Predicted - AGAIN (no organisation)

4.1.2 Fit of model

The coefficient of variation, R^2, indicates how much of the variation in Y is explained by the proposed model.

4.1.3 ANOVA

An ANOVA table is generated. This divides the overall variation into the variation explained by the regression model and the residual (left over or not explained) variation and compares them.

H_0: The relationship between X and Y in the population is not informative. The regression coefficients are zero i.e. $H_0 : \beta_0 = \beta_1 = 0$

H_1: There is at least one coefficient that is not zero

If the ANOVA $F > 4$ then H_0 is rejected, and the model is said to be informative.

4.1.4 Tests on individual Coefficients

In simple linear regression, there are two coefficients β_0 and β_1. They have subtly different functions

$H_0^{\beta_0}$: The intercept is zero i.e. $\beta_0 = 0$

$H_0^{\beta_1}$: The effect of X_i is constant $\beta_1 = 0$ i.e. the coefficient is flat

Each coefficient is tested using its estimated value and SE to find t-values.

Any $t >$ *approx* 2, then that coefficient is significant.

The estimated beta coefficients are then used in a prediction model and the impact of X on Y may be investigated.

The predicted Y is given by the following equation:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

4.2 Multiple linear regression

This is a simple extension of simple linear regression, but allows for up to p predictor X variables

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + e$$

4.2.1 Assumptions

Same as simple linear regression discussed in the previous section.

4.2.2 Fit of model

Same as simple linear regression discussed in the previous section.

4.2.3 ANOVA

Same as simple linear regression discussed in the previous section but now it investigates:

H_0: The relationship between all the X's and Y in the population is not informative.

All the regression coefficients are zero, i.e. $\beta_0 = \beta_1 = \beta_2 = ... + \beta_p = 0$

H_1: There is at least one coefficient is not zero

If the ANOVA $F > 4$ then H_0 is rejected, and the model is said to be informative.

4.2.4 Tests on individual coefficients

In simple linear regression, there are two coefficients β_0 and β_1. Now there are $p + 1$

$H_0^{\beta_0}$: The intercept is zero i.e. $\beta_0 = 0$

$H_0^{\beta_j}$: The effect of X_i is constant $\beta_j = 0; j = 1, ..., p$ i.e. the coefficient is flat

As for simple linear regression, each coefficient is tested.

Any t-values that are greater than approximately 2 indicate that coefficient is significant

The prediction model provides the estimates of the impact of the X's on Y.

The prediction model is given by the following equation:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_p X_p$$

It estimates the impact of the X's on Y.

4.2.5 Combinations of variables

The models are additive. These can be a mixture of numerical and categorical variables as main effects (e.g. gender, blood pressure…), interaction effects (e.g. blood pressure for male and female) and even transformations of variables (e.g. BMI^2)

4.2.6 Variable selection

There is conflict between having an informative model and one that overfits the data and will not be generalisable. There are several different views on how to resolve this.

The most parsimonious model comes from a mathematical approach and only keeps significant variables (individually and in combination with others).

However, it may be necessary to adjust for specific variables (e.g. age, gender …) even if not significant.

R provides several options to conduct Forward, Backward and Stepwise model selection.

4.3 Using R

- We will use the data `lowbirthweight.csv`. Please load the data into R as an R object `BW` if you have not loaded it already.
- We will develop a series of regression models and explore model selection to model the birth weight (`bwt`) with other potential predictors.
- After loading the data, change the variables `ethnicity`, `smoke` and `hyper` as factor variables using the function `as.factor`.
- You can load the data using the following code, but make sure that the data are available in your current working directory.

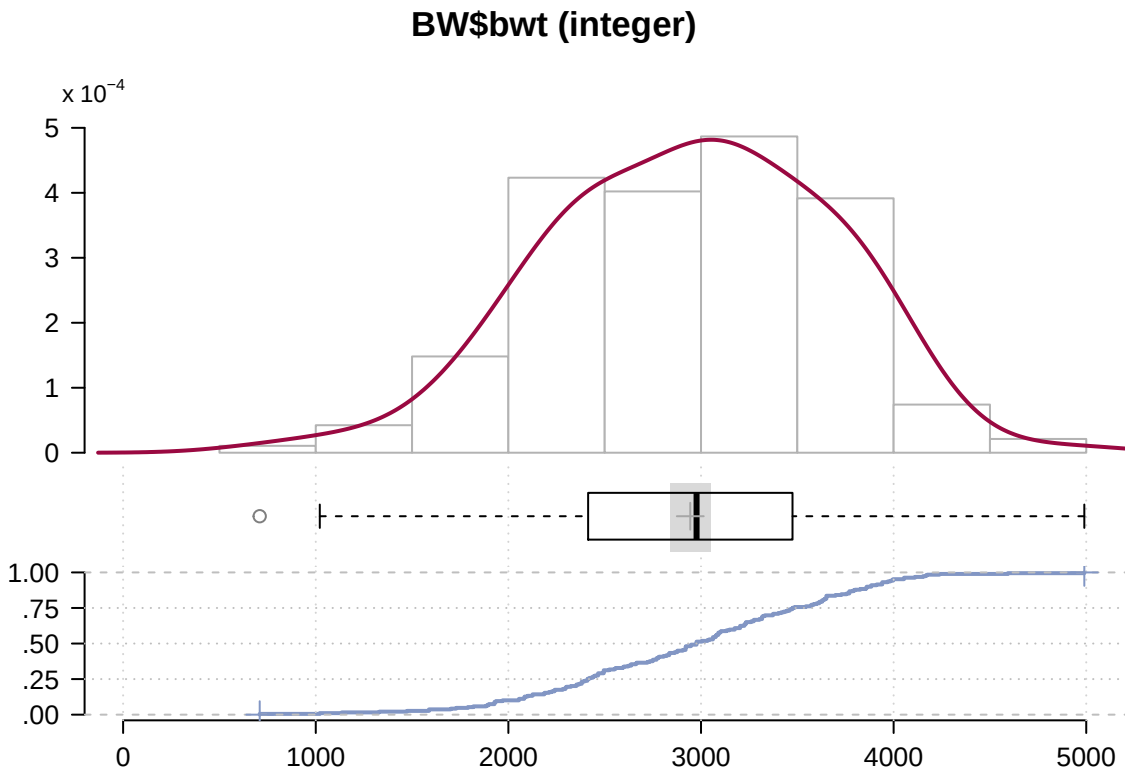
```
BW <- read.csv('lowbirthweight.csv')

BW$ethnicity <- as.factor(BW$ethnicity)
BW$smoke <- as.factor(BW$smoke)
BW$hyper <- as.factor(BW$hyper)
```

4.3.1 Summary Statistic

- We will obtain the summary statistics at first instance.
- Use the `DescTools::Desc` function to obtain the summary statistics and relevant plot.
- Here, we first obtain the summary statistics and histogram for the birth wegith measured in g (`bwt`).

```
library(DescTools)

DescTools::Desc(BW$bwt, plotit = TRUE)
```

```
-------------------------------------------------------------------------------
BW$bwt (integer)

    length         n       NAs    unique        0s      mean    meanCI
       189       189         0       133         0  2'944.29  2'839.68
             100.0%      0.0%               0.0%            3'048.89

      .05       .10       .25    median       .75       .90       .95
  1'801.20  2'038.00  2'414.00  2'977.00  3'475.00  3'864.80  3'997.00

     range        sd     vcoef       mad       IQR      skew      kurt
  4'281.00    729.02      0.25    819.88  1'061.00     -0.21     -0.14

lowest : 709, 1'021, 1'135, 1'330, 1'474
highest: 4'167, 4'174, 4'238, 4'593, 4'990
```

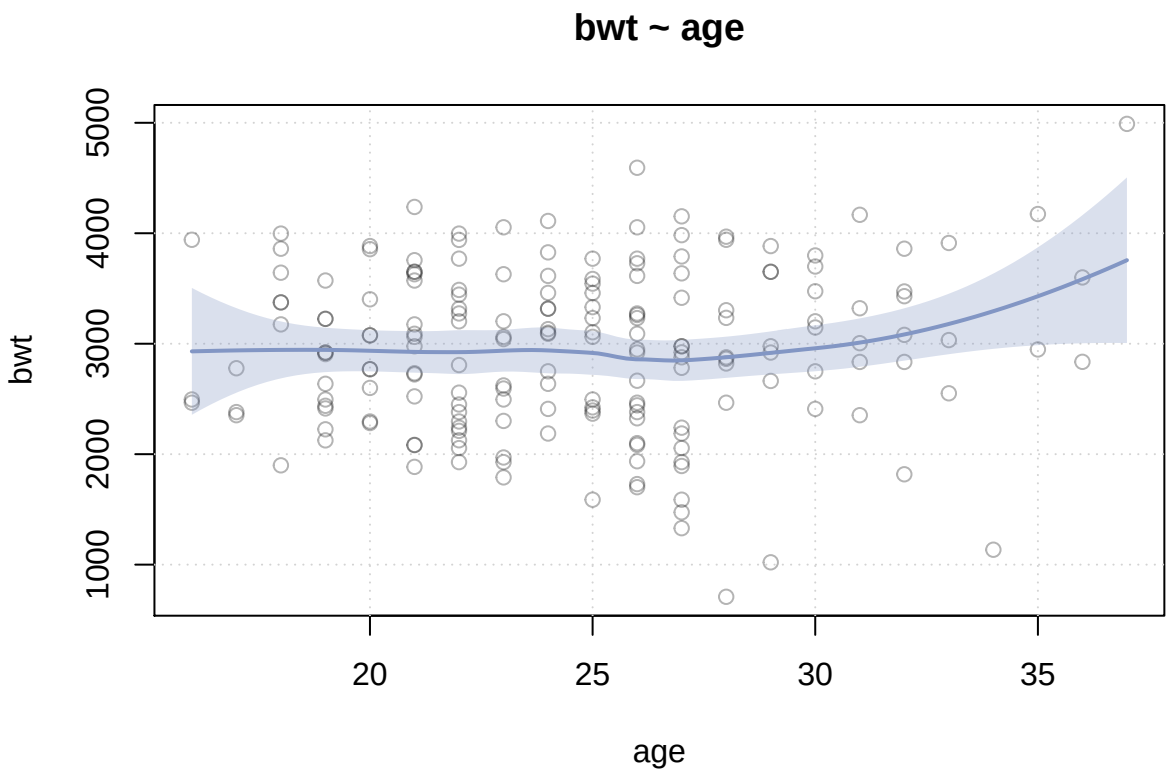


- Next, we explore the relationship between the birth weight (`bwt`) and the following predictor variables:
  - Age of the mother in years (`age`)
  - Mother's ethnic group (`ethnicity`): 1 = Caucasian; 2 = African-Caribbean; 3 = Other
  - Mother smoked during pregnancy (`smoke`): 0 = No; 1 = Yes
  - Mother has a history of hypertension (`hyper`): 0 = No; 1 = Yes
- We can explore this using the formula option in the `Desc` function.

```
library(DescTools)

DescTools::Desc(bwt ~ age + ethnicity + smoke + hyper,
                data = BW, plotit = TRUE)
```

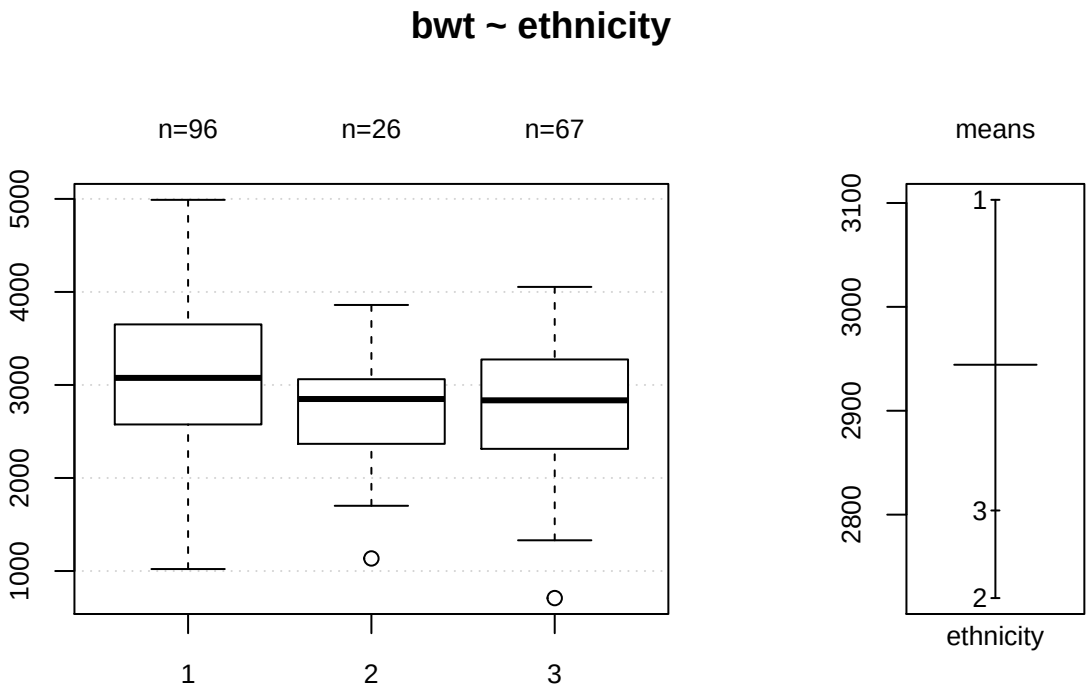
```
------------------------------------------------------------------------------
bwt ~ age

Summary:
n pairs: 189, valid: 189 (100.0%), missings: 0 (0.0%)
```

```
Pearson corr. : 0.071
Spearman corr.: 0.065
Kendall corr. : 0.041
```

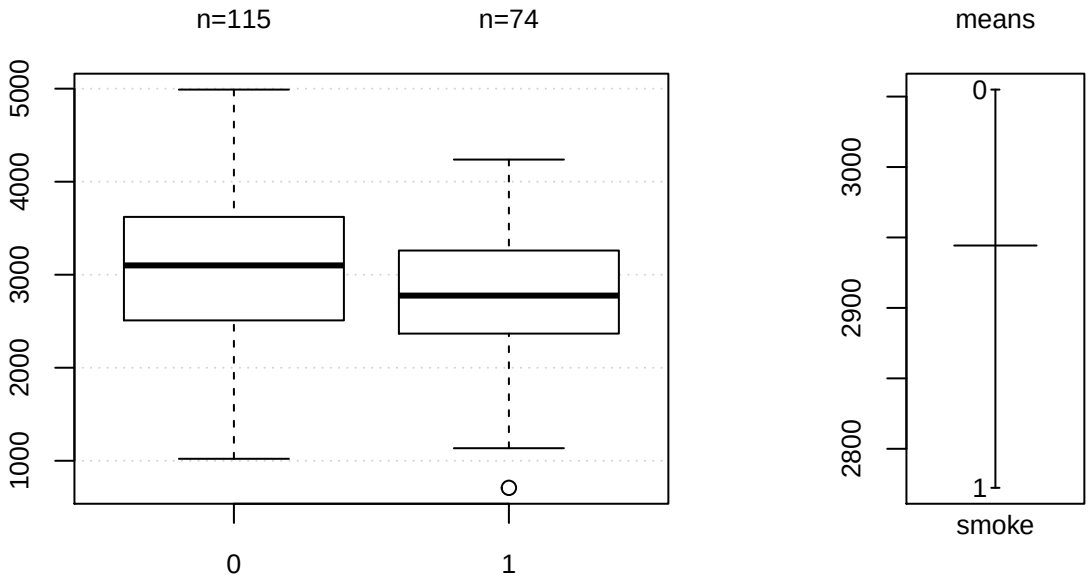


```
bwt ~ ethnicity

Summary:
n pairs: 189, valid: 189 (100.0%), missings: 0 (0.0%), groups: 3


                1          2          3
mean    3'103.010  2'719.692  2'804.015
median  3'076.000  2'849.000  2'835.000
sd        727.872    638.684    721.301
IQR     1'066.250    686.500    961.000
n              96         26         67
np         50.794%    13.757%    35.450%
NAs             0          0          0
0s              0          0          0

Kruskal-Wallis rank sum test:
```

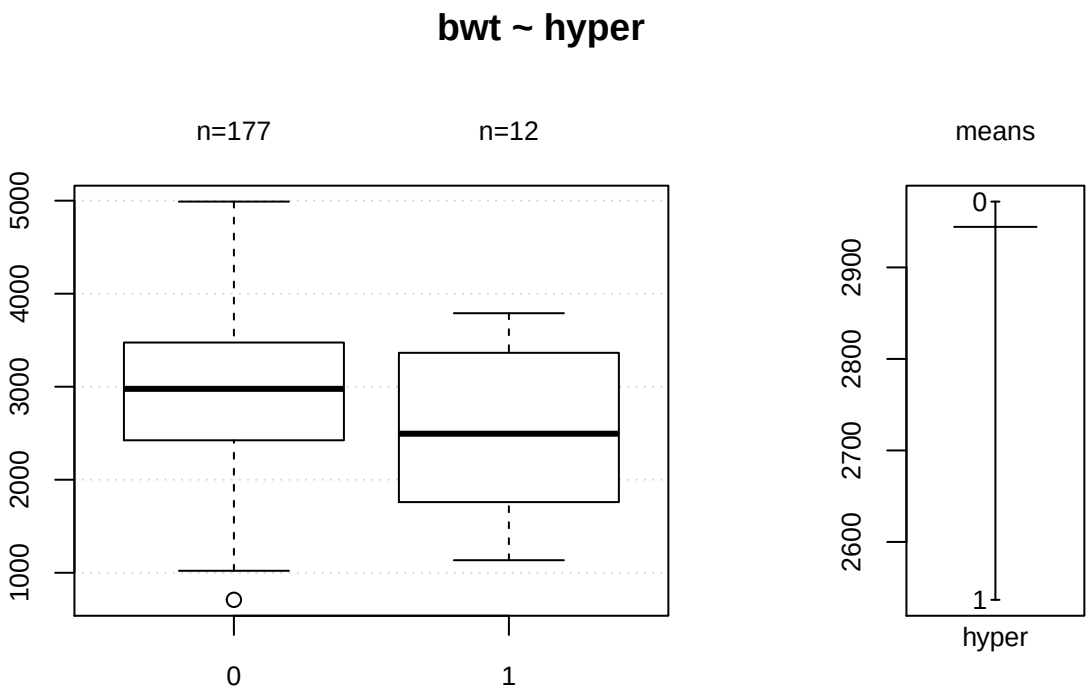
```
  Kruskal-Wallis chi-squared = 8.5525, df = 2, p-value = 0.01389
```




```
--------------------------------------------------------------------------------
bwt ~ smoke

Summary:
n pairs: 189, valid: 189 (100.0%), missings: 0 (0.0%), groups: 2


                0          1
mean    3'054.957  2'772.297
median  3'100.000  2'775.500
sd        752.409    659.807
IQR     1'112.500    875.250
n             115         74
np         60.847%    39.153%
NAs             0          0
0s              0          0

Kruskal-Wallis rank sum test:
  Kruskal-Wallis chi-squared = 7.2819, df = 1, p-value = 0.006965
```




```
--------------------------------------------------------------------------------
bwt ~ hyper

Summary:
n pairs: 189, valid: 189 (100.0%), missings: 0 (0.0%), groups: 2


                0          1
mean    2'971.915  2'536.750
median  2'977.000  2'495.000
sd        709.235    917.341
IQR     1'051.000  1'457.500
n             177         12
np         93.651%     6.349%
NAs             0          0
0s              0          0

Kruskal-Wallis rank sum test:
  Kruskal-Wallis chi-squared = 2.4666, df = 1, p-value = 0.1163
```



4.3.2 Fit simple linear regression model

- The R function `lm` fits linear regression model between x and y; `lm` stands for *linear model*.
- We will use the formula option of R (`y ~ x`) to fit the model
- For this example, y is the continuous variable birth weight (`bwt`) column.
- The x is the continuous variable age (`age`).
- Check the help file for `lm` for further details.

```
fm1 <- lm(bwt ~ age, data = BW)
```

- Use the function `anova` to print the analysis of variance table of the fitted model.

```
anova(fm1, test = 'F')
```

```
Analysis of Variance Table

Response: bwt
           Df   Sum Sq Mean Sq F value Pr(>F)
age         1   503814  503814  0.9477 0.3316
Residuals 187 99411484  531612
```

- The F-statistic and corresponding p-value suggest that the effect of `age` on `bwt` is not statistically significant ($p = 0.332$)
- Use the function `summary(fm)` to print the summary information of the fitted table.

```
summary(fm1)
```

```
Call:
lm(formula = bwt ~ age, data = BW)

Residuals:
     Min       1Q   Median       3Q      Max
-2274.64  -517.64     5.02   523.08  1901.43

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2657.19     299.64   8.868 5.89e-16 ***
age            11.66      11.98   0.974    0.332
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 729.1 on 187 degrees of freedom
Multiple R-squared:  0.005042,  Adjusted R-squared:  -0.0002782
F-statistic: 0.9477 on 1 and 187 DF,  p-value: 0.3316
```

- The summary outputs also confirm the results of the ANOVA table.
- We conclude that the age of the mother has no effect on the birth weight of babies.

4.3.3 Fit multiple linear regression model

- We can fit a multiple regression model including all relevant predictor variables.

```
fm2 <- lm(bwt ~ age + ethnicity + smoke + hyper, data = BW)
```

- The ANOVA of the fitted model is shown below.

```
anova(fm2, test = 'F')
```

```
Analysis of Variance Table

Response: bwt
           Df   Sum Sq Mean Sq F value    Pr(>F)
```

```
age         1   503814  503814  1.0729 0.3016509
ethnicity   2  4608151 2304076  4.9068 0.0083981 **
smoke       1  7240315 7240315 15.4191 0.0001219 ***
hyper       1  1632089 1632089  3.4757 0.0638769 .
Residuals 183 85930929  469568
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The ANOVA outputs show that variables `ethnicity` and `smoke` are statistically significant ($p < 0.05$)
- The summary of the fitted model is shown below.

```
summary(fm2)
```

```
Call:
lm(formula = bwt ~ age + ethnicity + smoke + hyper, data = BW)

Residuals:
     Min       1Q   Median       3Q      Max
-2328.85  -461.64    -4.69   471.60  1647.76

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3377.4384   316.1696  10.682  < 2e-16 ***
age           -0.9514    11.5893  -0.082 0.934665
ethnicity2  -428.0124   155.6755  -2.749 0.006570 **
ethnicity3  -451.7029   118.2967  -3.818 0.000184 ***
smoke1      -425.0583   109.3873  -3.886 0.000142 ***
hyper1      -382.6223   205.2331  -1.864 0.063877 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 685.3 on 183 degrees of freedom
Multiple R-squared:   0.14, Adjusted R-squared:  0.1165
F-statistic: 5.956 on 5 and 183 DF,  p-value: 3.956e-05
```

- The summary outputs confirm the outcomes from the ANOVA.
- We can simplify the model including only the statistically significant ($p < 0.05$) predictors.
- We can also adopt several model selection options provided by R. See the next section.

4.3.4 Model selection: Backward selection

- For the backward selection, we start with the full model.
- Use the function `drop1` to conduct backward model selection.
- The model selection can be based on AIC or F-statistic.
- To conduct the test using the F-statistic, use the argument `test = 'F'`.
- The smaller the AIC, the better the model.
- If the p-value for F-statistic is less than 5% ($p < 0.05$), the predictor is important i.e. it significantly explains the variability of the response variable.

```
drop1(fm2, test = 'F')
```

```
Single term deletions

Model:
bwt ~ age + ethnicity + smoke + hyper
          Df Sum of Sq      RSS    AIC F value    Pr(>F)
<none>                 85930929 2474.2
age        1      3164 85934093 2472.2  0.0067 0.9346647
ethnicity  2   7965685 93896614 2486.9  8.4819 0.0003001 ***
smoke      1   7090245 93021173 2487.2 15.0995 0.0001424 ***
hyper      1   1632089 87563018 2475.7  3.4757 0.0638769 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- From the above outputs, we can conclude that `ethnicity` and `smoke` are the two most important predictors ($p < 0.05$) to explain the variability in `bwt`. Hence the final model should include these variables.

4.3.5 Model selection: Forward selection

- For the forward selection, we start with the simple regression model i.e. `fm1` which was fitted only with one predictor (`age`).
- Use the function `add1` to conduct forward model selection.
- Use the `scope` argument to add other predictors.
- The model selection can be based on AIC or F-statistic.
- The smaller the AIC, the better the model.
- If the p-value for F-statistic is less than 5% ($p < 0.05$), the predictor is important i.e. it significantly explains the variability of the response variable.

```
add1(fm1, scope = ~ age + low + ethnicity + smoke + hyper, test = 'F')
```

```
Single term additions

Model:
bwt ~ age
          Df Sum of Sq      RSS    AIC  F value  Pr(>F)
<none>                 99411484 2493.7
low        1  61101986 38309498 2315.5 296.6619 < 2e-16 ***
ethnicity  2   4608151 94803333 2488.7   4.4962 0.01240 *
smoke      1   3462831 95948653 2489.0   6.7128 0.01033 *
hyper      1   2123167 97288317 2491.6   4.0592 0.04537 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- From the above outputs, we can conclude that `ethnicity` and `smoke` are the two most important predictors ($p < 0.05$) to explain the variability in `bwt`.
- The variable `hyper`, in addition, shows statistical significance ($p < 0.05$) which may need further investigation.

4.3.6 Model selection: Stepwise selection

- For the stepwise selection, we start with the full model.
- Use the function `step` to conduct stepwise model selection.
- Use the `scope` argument to add possible predictors.
- The default direction for `step` is `both`with the `scope` argument.
- The model selection can be based on AIC or $F - statistic$.
- The smaller the AIC, the better the model.
- If the p-value for F-statistic is less than 5% ($p < 0.05$), the predictor is important i.e. it significantly explains the variability of the response variable.

```
step(fm2, scope = ~ age + ethnicity + smoke + hyper,
     test = 'F')
```

```
Start:  AIC=2474.16
bwt ~ age + ethnicity + smoke + hyper

            Df Sum of Sq      RSS    AIC F value    Pr(>F)
- age        1      3164 85934093 2472.2  0.0067 0.9346647
```

```
<none>                   85930929 2474.2
- hyper      1   1632089 87563018 2475.7  3.4757 0.0638769 .
- ethnicity  2   7965685 93896614 2486.9  8.4819 0.0003001 ***
- smoke      1   7090245 93021173 2487.2 15.0995 0.0001424 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Step:  AIC=2472.17
bwt ~ ethnicity + smoke + hyper

            Df Sum of Sq      RSS    AIC F value    Pr(>F)
<none>                   85934093 2472.2
- hyper      1   1634308 87568401 2473.7  3.4993 0.0629811 .
+ age        1      3164 85930929 2474.2  0.0067 0.9346647
- smoke      1   7158219 93092312 2485.3 15.3270 0.0001272 ***
- ethnicity  2   8328793 94262886 2485.7  8.9167 0.0002013 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



Call:
lm(formula = bwt ~ ethnicity + smoke + hyper, data = BW)

Coefficients:
(Intercept)   ethnicity2   ethnicity3       smoke1       hyper1
     3352.6       -425.7       -449.8       -424.0       -382.8
```

- The stepwise model selection selects the final model with three predictors: `ethnicity`, `smoke` and `hyper` considering $p < 0.10$). Sometime it is advisable to keep p-values at higher level to conduct preliminary model selection.
- *Be careful while using the standard model selection approach; always consider the research questions and the underlying biology while conducting any model selection approach or choosing the final model.*

4.3.7 Final Model

- We fit the final model including the variables: `ethnicity`, `smoke` and `hyper` and obtain the relevant outputs.

```
fm <- lm(bwt ~ ethnicity + smoke + hyper, data = BW)

anova(fm, test = 'F')
```

```
Analysis of Variance Table
```

```
Response: bwt
           Df   Sum Sq Mean Sq F value    Pr(>F)
ethnicity   2  5048361 2524181  5.4047 0.0052375 **
smoke       1  7298537 7298537 15.6274 0.0001099 ***
hyper       1  1634308 1634308  3.4993 0.0629811 .
Residuals 184 85934093  467033
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- We note that the variable `hyper` is not significant when the model is adjusted for `ethnicity` and `smoke`.
- Hence, we consider the final model with only two variables: `ethnicity` and `smoke`.

```
fm <- lm(bwt ~ ethnicity + smoke, data = BW)

anova(fm)
```

```
Analysis of Variance Table

Response: bwt
           Df   Sum Sq Mean Sq F value    Pr(>F)
ethnicity   2  5048361 2524181  5.3327 0.0056021 **
smoke       1  7298537 7298537 15.4191 0.0001214 ***
Residuals 185 87568401  473343
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(fm)
```

```

Call:
lm(formula = bwt ~ ethnicity + smoke, data = BW)

Residuals:
     Min       1Q   Median       3Q      Max
-2313.86  -440.83    15.17   492.14  1655.14

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3334.86      91.74  36.350  < 2e-16 ***
ethnicity2   -450.54     153.07  -2.943 0.003662 **
ethnicity3   -454.18     116.44  -3.901 0.000134 ***
smoke1       -428.03     109.00  -3.927 0.000121 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 688 on 185 degrees of freedom
Multiple R-squared:  0.1236,    Adjusted R-squared:  0.1094
F-statistic: 8.695 on 3 and 185 DF,  p-value: 1.995e-05
```

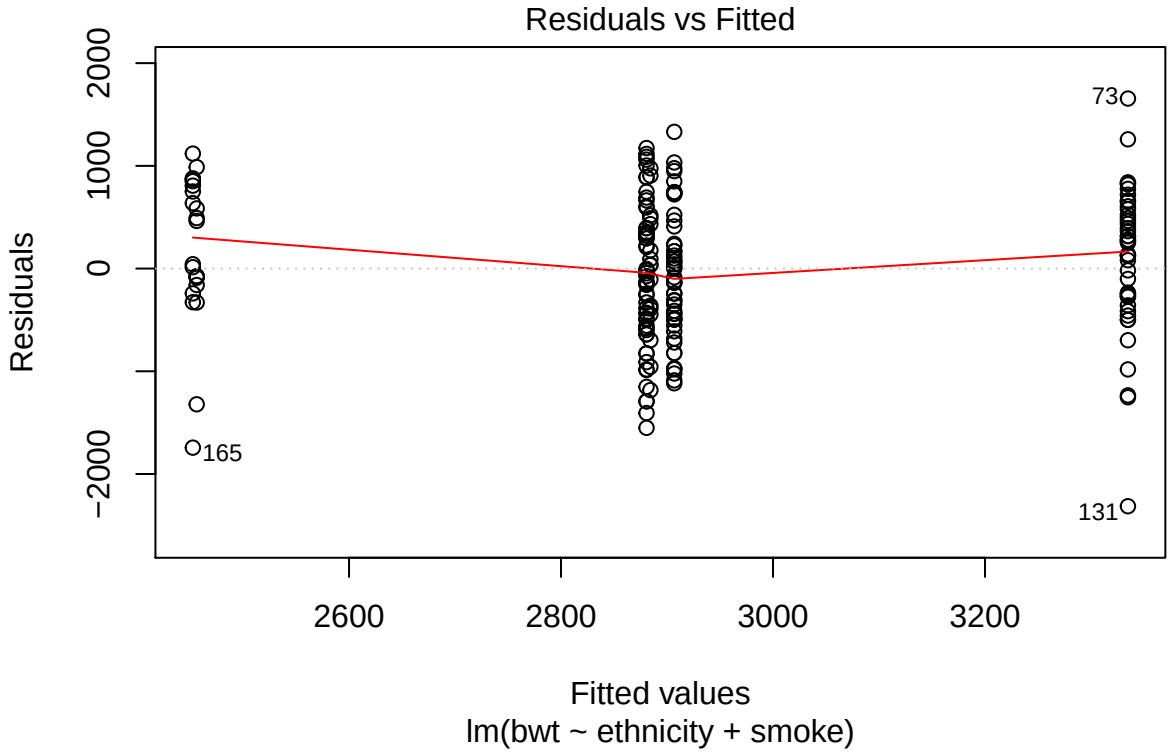
4.3.8 Model Interpretation

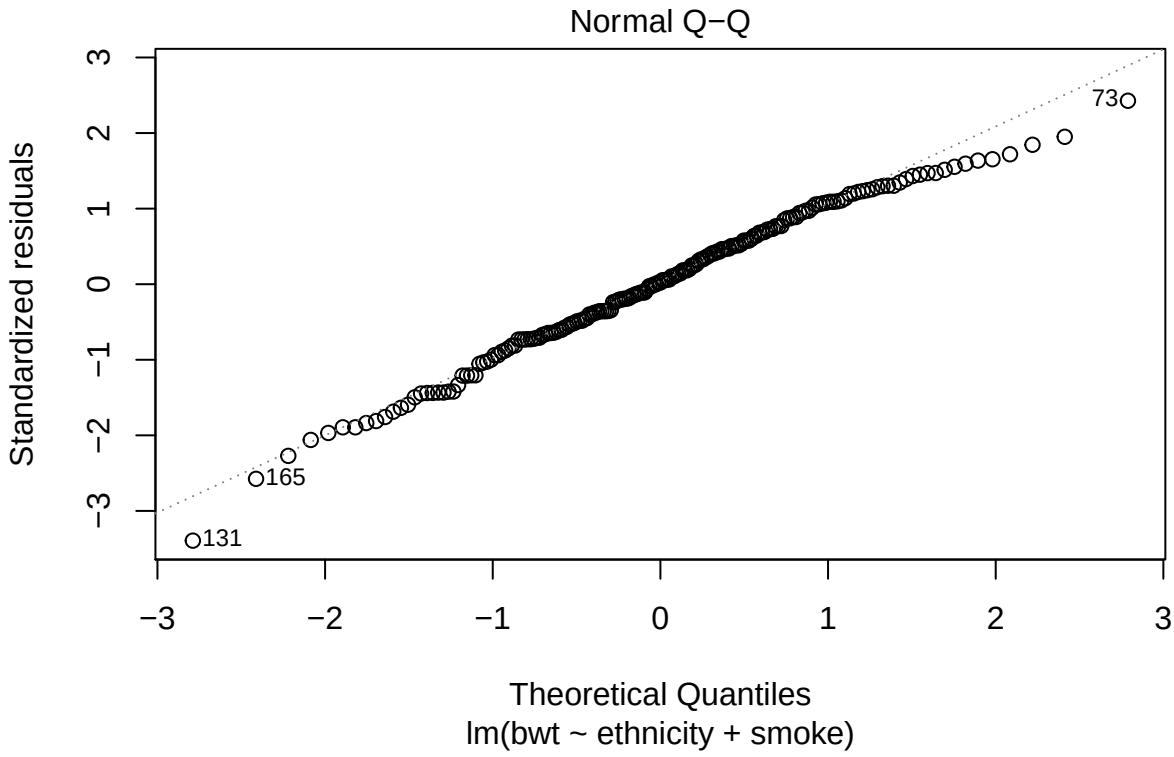
- Overall, the smoking status ($p < 0.001$) and ethnicity ($p = 0.008$) of mothers showed statistically significant association with the mean birth weight of babies.
- The summary outputs show that the birth weight of babies born to mothers with ethnic groups 2 and 3 were significantly lower compared with ethnic group 1.
- The birthweight of babies were significantly lower for mothers who smoked comapred with mother who did not smoke.
- The age and history of hypertension of mothers did not have any effect on the birth weight of babies.

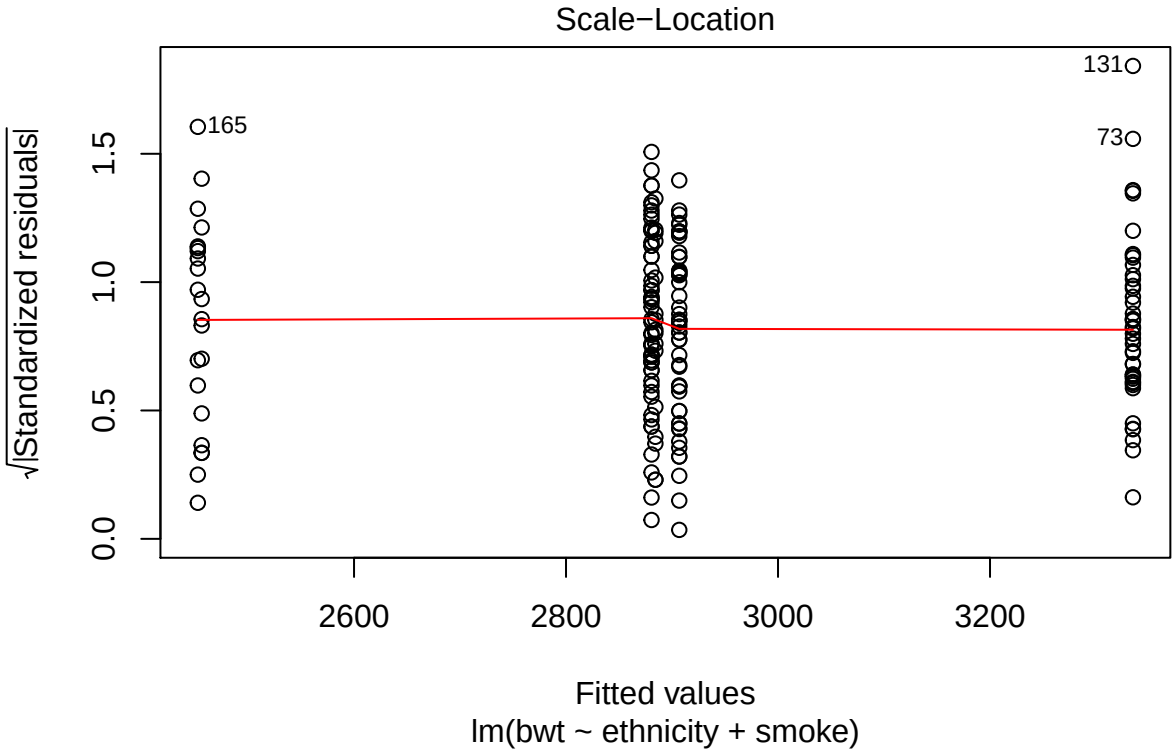
4.3.9 Check Model Assumptions

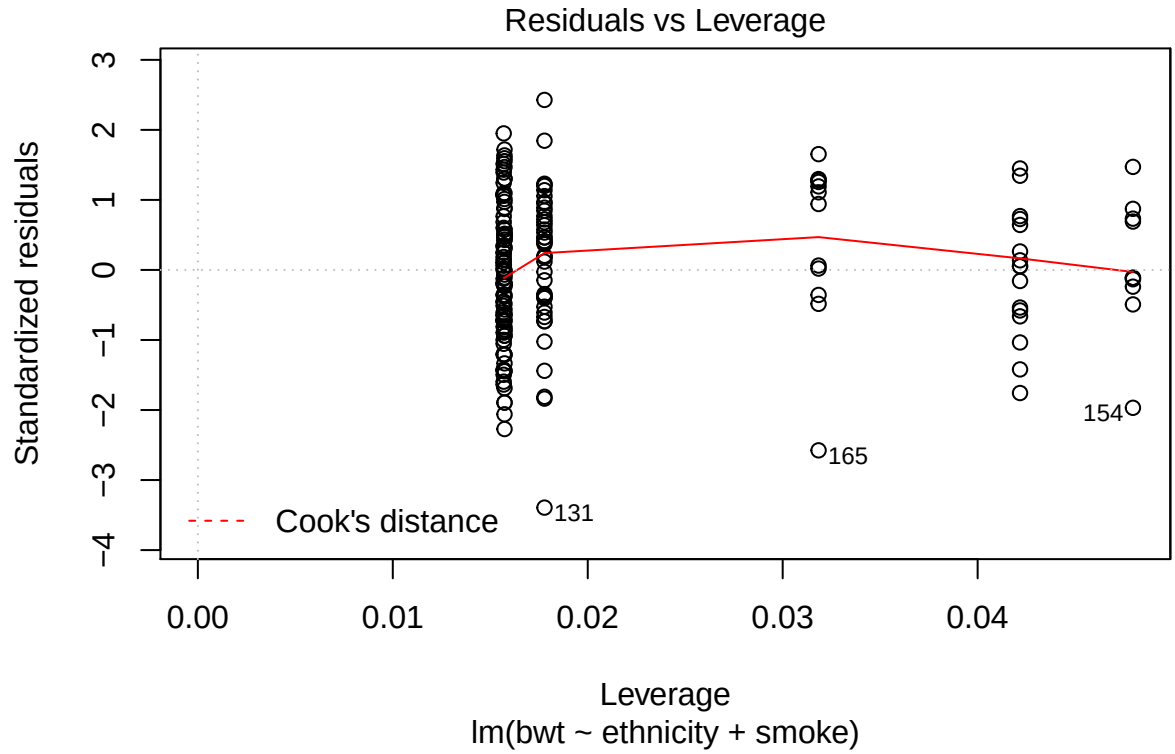
- We can check the assumptions of the model using the `plot` function.
- It produces four plots; you have to press `Enter` or `Return` to move to the next plot.
- RStudio will save all the plots; use the back arrow in the plot tab to view all plots.
- **Residuals vs Fitted plot:** This plot shows if residuals have any non-linear pattern when plotted against the fitted values; any significant non-linear pattern will be reflected by the red line drawn alongside the plot. The plot also checks the assumption of equal variance or homoscedasticity.
- **Normal Q-Q plot:** This plot shows if residuals are normally distributed. If the residuals are normally distributed, all the points should reasonably be on the dotted line along the diagonal.
- **Scale-Location plot:** This plot shows if the standardised residuals are spread equally along the fitted values. This plot checks the assumption of equal variance or homoscedasticity.
- **Residuals vs Leverage:** This plot shows if any particular observation(s) is (are) influential to the model fitting. Depending on the dataset, observations with low and high residuals as well as low and high leverage will be identified in this plot.

```
plot(fm)
```


Residuals vs Fitted
Residuals
2000
1000
0
−2000
73
165
131
2600
2800
3000
3200
Fitted values
lm(bwt ~ ethnicity + smoke)


Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−3
73
165
131
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(bwt ~ ethnicity + smoke)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
165
131
73
2600
2800
3000
3200
Fitted values
lm(bwt ~ ethnicity + smoke)


Residuals vs Leverage
Standardized residuals
−4
−3
−2
−1
0
1
2
3
Cook's distance
131
165
154
0.00
0.01
0.02
0.03
0.04
Leverage
lm(bwt ~ ethnicity + smoke)

Chapter 5

Relative Risk

5.0.1 When is it appropriate?

Relative Risk (RR) is used to compare the risk of an event between two groups. It is used in *cohort studies* and *randomised controlled trials* as a measure of effect.

5.0.2 Example

For this example, we will use the data file `lowbirthweight.csv`.

Please read the data into R environment as a data.frame object `BW` as shown below.

Note that the data file `lowbirthweight.csv` is in the current working directory.

```
BW <- read.csv(file = 'lowbirthweight.csv')
```

Using this dataset, we wish to compare the *risk of having a low birth weight (LBW) baby* ($< 2500g$) in mothers who ***do not smoke*** **with those that** ***do smoke***.

5.0.3 Null hypothesis

H_0 : $RR = 1$, i.e. risk of LBW baby in non-smoking mothers = risk of LBW baby in those smoking mothers

5.0.4 Alternative hypothesis

H_1 : $RR \neq 1$, i.e. risk of LBW baby in non-smoking mothers is not equal to risk of LBW baby in those smoking mothers

5.1 Using R

- The R function `DescTools::Desc` from the `DescTools` library to estimate the relative risk (RR).
- We will use the formula option of R (`y ~ x`) to submit to the function
- The variable x should represent the *event* (here, the variable `low` i.e. high (0) and low (1)); it corresponds to `column`.
- The variable y should represent the *exposure* (here, the variable `smoke` i.e. smoker (0) vs non-smoker (1)); it corresponds to `row`.
- The outputs provide the Chi-squared test as well as other additional information; ignore those outputs.

```
library(DescTools)

DescTools::Desc(smoke ~ low, data = BW, plotit = FALSE)
```

```
------------------------------------------------------------------------------
smoke ~ low


Summary:
n: 189, rows: 2, columns: 2

Pearson's Chi-squared test (cont. adj):
  X-squared = 4.2359, df = 1, p-value = 0.03958
Fisher's exact test p-value = 0.03618
McNemar's chi-squared = 2.6849, df = 1, p-value = 0.1013

                    estimate lwr.ci upr.ci'

odds ratio             2.022  1.081  3.783
rel. risk (col1)       1.258  1.013  1.561
rel. risk (col2)       0.622  0.409  0.945


Phi-Coefficient        0.161
Contingency Coeff.     0.159
Cramer's V             0.161


        low        0      1    Sum
smoke
```

```
0       freq      86     29    115
        perc   45.5%  15.3%  60.8%
        p.row  74.8%  25.2%      .
        p.col  66.2%  49.2%      .

1       freq      44     30     74
        perc   23.3%  15.9%  39.2%
        p.row  59.5%  40.5%      .
        p.col  33.8%  50.8%      .

Sum     freq     130     59    189
        perc   68.8%  31.2% 100.0%
        p.row      .      .      .
        p.col      .      .      .


----------
' 95% conf. level
```

Looking at the outputs on relative risk estimates, the output relevant to us represents `rel. risk (col2)` which represents the value of 1 for the variable `low`.

Hence the estimate of relative risk and corresponding 95% confidence intervals are: $0.622\ (0.409, 0.945)$.

Note:

R has given you two estimates of RR, but you only ever need one. Carefully check the estimate that is relevant for the given objective.

If in doubt, calculate the RR by hand and then cross check with the R output to provide the 95% confidence interval.

5.2 Direct calculation

We can calculate the **RR** directly by hand.

From the output, we can see that 25.2% of non-smoking mothers had low birth weight babies (risk in unexposed group) compared to 40.5% of the smoking mothers (risk in exposed group).

Therefore, the estimated relative risk (RR) is: $0.252/0.405 = 0.622$

This is for **non-smokers relative to smokers**. This is key for interpretation.

5.3 Interpretation

The $RR = 0.622$ with $95\%CI\ (0.409, 0.945)$ indicating that the risk of having a *low birth weight baby* for *non-smokers* is 0.62 times that of *smokers* and in the population it is between 0.41 and 0.95 times lower. This decrease in risk is significant as the 95% CI does not contain one ($p < 0.05$).

Conclusion

We can, therefore, conclude being a non-smoker reduces your risk of having a low birth weight baby.

Note: If you wanted to present the comparison of **smokers compared to non-smokers**, you can do the reciprocal of the RR, i.e. $1/0.622 = 1.607$, so smokers have an increased risk of LBW. It is an alternative interpretation. To calculate the 95% confidence interval, use the reciprocal values of the confidence interval estimated earlier.

Chapter 6

Odds Ratio

6.0.1 When is it appropriate?

If you wish to compare the odds of an exposure between those with the event to those without the event. Often used for *case-control* studies.

6.0.2 Example

For this example, we will use the data file `lowbirthweight.csv`.

Please read the data into R environment as a data.frame object `BW` as shown below.

Note that the data file `lowbirthweight.csv` is in the current working directory.

```
BW <- read.csv(file = 'lowbirthweight.csv')
```

Using the data, we wish to compare the odds of the mother have *hypertension during pregnancy* if she had a baby of *low birth weight* ($< 2500g$) with the odds of the mother having *hypertension during pregnancy* if she had a baby of *normal weight* ($\geq 2500g$). The event is low birth weight (`yes` or `no`) and the exposure of interested is mothers hypertension (`yes` or `no`).

6.0.3 Null hypothesis

$H_0 :$ $OR = 1$, i.e. odds of mother having hypertension during pregnancy is the same for babies born < 2500g compared to babies born $\geq 2500g$.

6.0.4 Alternative hypothesis

$H_1 :$ $OR \neq 1$, i.e. odds of mother having hypertension during pregnancy is different for babies born < 2500g compared to babies born $\geq 2500g$.

6.1 Using R

First we need to calculate a new variable `bwt_lt2500` as shown below.

```
BW$bwt_lt2500 <- as.factor(BW$bwt < 2500)
```

- We will use the R function `DescTools::Desc` from the `DescTools` library to estimate the odds ratio (OR).
- We will use the formula option of R (`y ~ x`) to submit to the function
- The variable x should represent the *event* (here, the variable `hyper`); it corresponds to `column`.
- The variable y should represent the *exposure* (here, the variable `bwt_lt2500`); it corresponds to `row`.
- The outputs provide the Chi-squared test as well as other additional information; ignore those outputs.

```
DescTools::Desc(bwt_lt2500 ~ hyper, data = BW, plotit = FALSE)
```

```
------------------------------------------------------------------------------
bwt_lt2500 ~ hyper


Summary:
n: 189, rows: 2, columns: 2

Pearson's Chi-squared test (cont. adj):
  X-squared = 3.1431, df = 1, p-value = 0.07625
Fisher's exact test p-value = 0.05161
McNemar's chi-squared = 37.123, df = 1, p-value = 1.109e-09

Warning message:
  Exp. counts < 5: Chi-squared approx. may be incorrect!!

                    estimate lwr.ci upr.ci'

odds ratio             3.365  1.021 11.088
rel. risk (col1)       1.091  0.987  1.205
rel. risk (col2)       0.324  0.107  0.979


Phi-Coefficient        0.152
Contingency Coeff.     0.151
```

```
Cramer's V             0.152


             hyper      0      1    Sum
bwt_lt2500

FALSE        freq     125      5    130
             perc   66.1%   2.6%  68.8%
             p.row  96.2%   3.8%      .
             p.col  70.6%  41.7%      .

TRUE         freq      52      7     59
             perc   27.5%   3.7%  31.2%
             p.row  88.1%  11.9%      .
             p.col  29.4%  58.3%      .

Sum          freq     177     12    189
             perc   93.7%   6.3% 100.0%
             p.row      .      .      .
             p.col      .      .      .


----------
' 95% conf. level
```

Looking at the outputs, the output relevant to us represents the odds ratio is given by the first row of the table indicated as `odds ratio`.

Hence the estimate of odds ratio (OR) and corresponding 95% confidence intervals are: $3.365\ (1.021, 11.088)$.

Note:

R has also given you two estimates of RR, but we can ignore this.

If in doubt, calculate the OR by hand and then cross check with the R output to provide the 95% confidence interval.

6.2 Direct calculation

We can calculate the **OR** directly by hand.

Odds of hypertension in *low birth weight babies* ($< 2500g$) $= 7/52 = 0.134$

Odds of hypertension in *normal babies* $\geq 2500g = 5/125 = 0.04$

Therefore, the odds ratio (OR) $= 0.134/0.04 = 3.36$

6.3 Interpretation

The odds ratio is 3.36 with 95%CI (1.02, 11.1); this indicates the odds of a hypertensive mother is over three times as much for a low birth weight baby as it is for a normal baby.

Conclusion

The odds of the mother having *hypertension during pregnancy* with a baby of *low birth weight* ($< 2500g$) is 3.36 times higher compared with a baby of *normal weight* ($\geq 2500g$).