PHCM9795 Foundations of Biostatistics

Course notes

Term 2, 2023

Contents

Course introduction

Welcome to PHCM9795 Foundations of Biostatistics.
This is an introductory course in biostatistics for postgraduate students. The course aims to provide students with core biostatistical skills to analyse and present quantitative data from different study types. These are essential skills required in your degree and throughout your career.
We hope you enjoy the course and will value your feedback and comment throughout the course.

Course information

Biostatistics is a foundational discipline needed for the analysis and interpretation of quantitative information and its application to population health policy and practice.
This course is central to becoming a population health practitioner as the concepts and techniques developed in the course are fundamental to your studies and practice in population health. In this course you will develop an understanding of, and skills in, the core concepts of biostatistics that are necessary for analysis and interpretation of population health data and health literature.
In designing this course, we provide a learning sequence that will allow you to obtain the required graduate capabilities identified for your program. This course is taught with an emphasis on formulating a hypothesis and quantifying the evidence in relation to a specific research question. You will have the opportunity to analyse data from different study types commonly seen in population health research.
The course will allow those of you who have covered some of this material in your undergraduate and other professional education to consolidate your knowledge and skills. Students exposed to biostatistics for the first time may find the course challenging at times. Based on student feedback, the key to success in this course is to devote time to it every week. We recommend that you spend an average of 10-15 hours per week on the course, including the time spent reading the course notes and readings, listening to lectures, and working through learning activities and completing your assessments. Please use the resources provided to assist you, including online support.

Units of credit

This course is a core course of the Master of Public Health, Master of Global Health and Master of Infectious Diseases Intelligence programs and associated dual degrees, comprising 6 units of credit towards the total required for completion of the study program. A value of 6 UOC requires a minimum of 150 hours work for the average student across the term.

Course aim

This course aims to provide students with the core biostatistical skills to apply appropriate statistical techniques to analyse and present population health data.

Learning outcomes

On successful completion of this course, you will be able to:

1. Summarise and visualise data using statistical software.

2. Demonstrate an understanding of statistical inference by interpreting p-values and confidence intervals.
3. Apply appropriate statistical tests for different types of variables given a research question, and interpret computer output of these tests appropriately.
4. Determine the appropriate sample size when planning a research study.
5. Present and interpret statistical findings appropriate for a population health audience.

Changelog

Module 1

Correlation and linear regression

Learning objectives

By the end of this module you will be able to:

- Explore the association between two continuous variables using a scatter plot;
- Estimate and interpret correlation coefficients;
- Estimate and interpret parameters from a simple linear regression;
- Decide whether a regression model is valid;
- Test a hypothesis using regression coefficients;
- Outline the concept of multiple regression and its role in investigative epidemiology.

Readings

Kirkwood and Sterne [2001]; Chapter 10. [UNSW Library Link]
Bland [2015]; Chapter 11. [UNSW Library Link]
Acock [2010]; Chapter 8.

1.1 Introduction

In Module 5, we saw how to test whether the means from two groups are equal - in other words, whether a continuous variable is related to a categorical variable. We often want to know how closely two continuous variables are related. For example, we may want to know how closely blood cholesterol levels are related to dietary fat intake in adult men. To measure the strength of association between two continuously distributed variables, a correlation coefficient is used.
We may also want to know how well one continuous measurement predicts the value of another continuous measurement. For example, we may want to know how well height predicts values of lung capacity in a community of adults. A regression model allows us to use one measurement to predict another measurement.
Although both correlation coefficients and regression models can be used to describe the degree of association between two continuous variables, the two methods provide very different statistical information. It is important to note that both methods only measures the strengths of an association between variables and does not imply a causal relationship.

1.2 Correlation

We use correlation to measure the strength of a linear relationship between two variables. Before calculating a correlation coefficient, a scatter plot should first be obtained to give an understanding of the nature of the relationship between the two variables.

1.2.1 Worked Example

The file `Example_8.1.csv` has information about height and lung function collected from a sample of 120 adults. A random sample of adults was approached to take part in the research study, but the response rate was low at 45%. Information was collected on height (cm) and lung function, which was measured as forced vital capacity (FVC). We can obtain a *scatter-plot* shown in Figure 1.1). This shows that as height increases, lung function also increases, which is as expected. One or two of the data points are separated from the rest of the data but are not so far away as to be considered outliers because they do not seem to stand out of other observations.

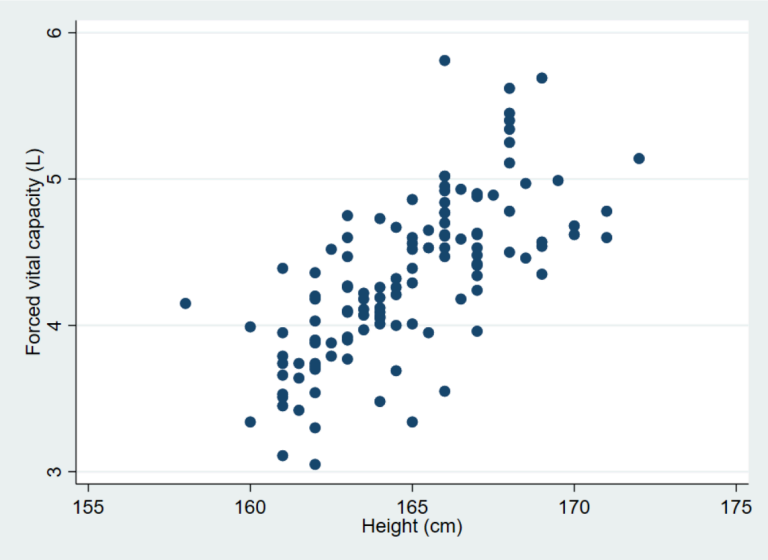


Figure 1.1: Association between height and lung function in 120 adults

1.2.2 Correlation coefficients

A correlation coefficient (r) describes how closely the variables are related, that is the strength of linear association between two continuous variables. The range of the coefficient is from +1 to −1 where +1 is a perfect positive association, 0 is no association and −1 is a perfect inverse association. In general, an absolute (disregarding the sign) r value below 0.3 indicates a weak association, 0.3 to < 0.6 is fair association, 0.6 to < 0.8 is a moderate association, and $\geq$ 0.8 indicates a strong association.

The coefficient is positive when large values of one variable tend to occur with large values of the other, and small values of one variable (y) tend to occur with small values of the other (x) (Figure 1.2 (a and b)). For example, height and weight in healthy children or age and blood pressure.

The coefficient is negative when large values of one variable tend to occur with small values of the other, and small values of one variable tend to occur with large values of the other (Figure 1.2 (c and d)). For example, percentage immunised against infectious diseases and under-five mortality rate.

The P value associated with an r value is an estimate of whether the correlation coefficient is significantly different from zero. However, a correlation coefficient that does not have a significant P value does not imply that there is no relationship because the correlation coefficient only tests for a linear association and there may be a non-linear relationship such as a curved or irregular relationship.

The assumptions for using a Pearson's correlation coefficient are that:

- observations are independent;
- both variables are continuous variables;
- the relationship between the two variables is linear.

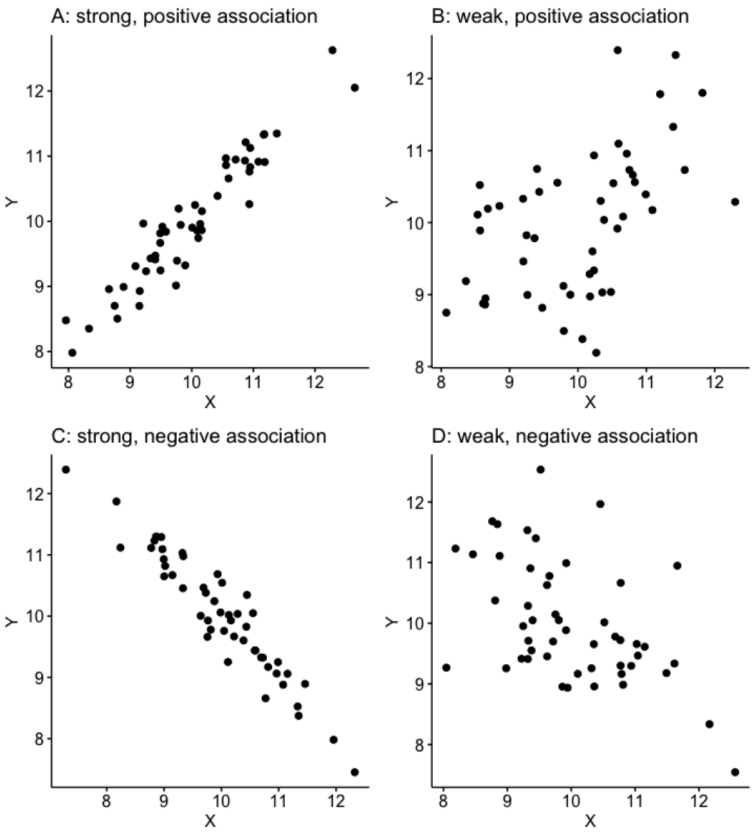


Figure 1.2: Scatter plots demonstrating strong and weak, positive and negative associations

There is a further assumption that the data follow a bivariate normal distribution. This assumes: *y* follows a normal distribution for given values of *x*; and *x* follows a normal distribution for given values of *y*. This is quite a technical assumption that we do not discuss further.
There are two types of correlation coefficients– the correct one to use is determined by the nature of the variables as shown in Table 1.1).

Table 1.1: Correlation coefficients and their application

Correlation coefficient	**Application**
Pearson's correlation coefficient: r	Both variables are continuous and a bivariate normal distribution can be assumed
Spearman's rank correlation: rho	Bivariate normality cannot be assumed. Also useful when at least one of the variables is ordinal

Spearman's ρ is calculated using the ranks of the data, rather than the actual values of the data. We will see further examples of such methods in Module 9, when we consider non-parametric tests, which are often based on ranks.
Correlation coefficients are often presented in the form of a *correlation matrix* which can display the correlation between a number of variables in a single table (Table 1.2).

Table 1.2: Correlation matrix for Height and FVC

	Height	FVC
Height	1	0.70 P < 0.0001
FVC	0.70 P < 0.0001	1

This correlation matrix shows that the Pearson's correlation coefficient between height and lung function is 0.70 with P<0.0001 indicating very strong evidence of a linear association between height and FVC. A correlation matrix sometimes includes correlations between the same variable, indicated as a correlation coefficient of 1. For example, $Height$ is perfectly correlated with itself (i.e. has a correlation coefficient of 1). Similarly, FVC is perfectly correlated with itself.
This r value was calculated for the full data set of 120 adults who had heights ranging from 160 to 172cms. If the r value is calculated for the 60 adults with a height less than 165cms, it is much lower at 0.433 although significant at P=0.001. In general, r values are higher for a wider range of values on the x axis even though the relationship between the two variables remains the same.
Correlation coefficients are rarely used as important statistics in their own right because they do not fully explain the relationship between the two variables and the range of the data has an important influence on the size of the coefficient. In addition, the statistical significance of the correlation coefficient is often over interpreted because a small correlation which is of no clinical importance can become statistically significant even with a relatively small sample size. For example, a poor correlation of 0.3 will be statistically significant if the sample size is large enough.

1.3 Linear regression

The nature of a relationship between two variables is more fully described using regression. There are two principal purposes for building a regression model. The most common is to build a predictive model, for example in situations in which age and gender are used to predict normal values of characteristics such as lung size or body mass index. Normal values are the range of values that occur naturally in the general population.
The second purpose for using a regression model is for testing the hypothesis that there is a linear relationship between one or more explanatory variables and an outcome variable. For example, a regression model can be used to test the extent to which age predicts BMI or to test the hypothesis that two groups with a different dietary regime have significantly different BMI values after adjusting for age differences.
From Worked Example 8.1, we can be also plot a regression line through the scatter. Figure 1.3 shows the data overlayed with the fitted regression line.
The line through the plot is called the line of 'best fit' because the size of the deviations between the data points and the line is minimised in the calculation. The distance between each data point and the regression line is called a 'residual'.

1.3.1 Regression equations

The mathematical equation for the line explains the relationship between the two variables. The equation of the regression line is as follows:

$$y = \beta_0 + \beta_1 x$$

This line is shown in Figure 1.4 using the notation shown in Table 1.3.
The intercept is the point at which the regression line intersects with the y-axis when the value of 'x' is zero. In most cases, the intercept does not have a biologically meaningful interpretation as the explanatory variable cannot take a value of zero. In our working example, the intercept is not meaningful as it is not possible for an adult to have a height of 0cm.

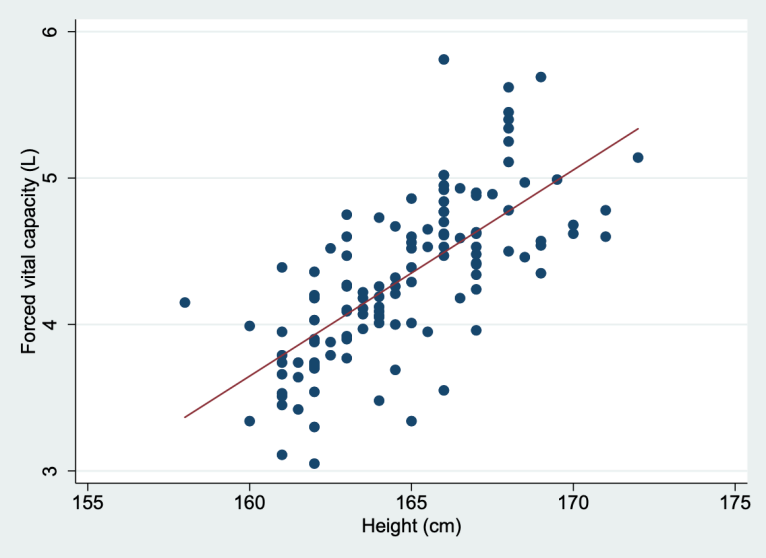


Figure 1.3: Association between height and lung function in 120 adults

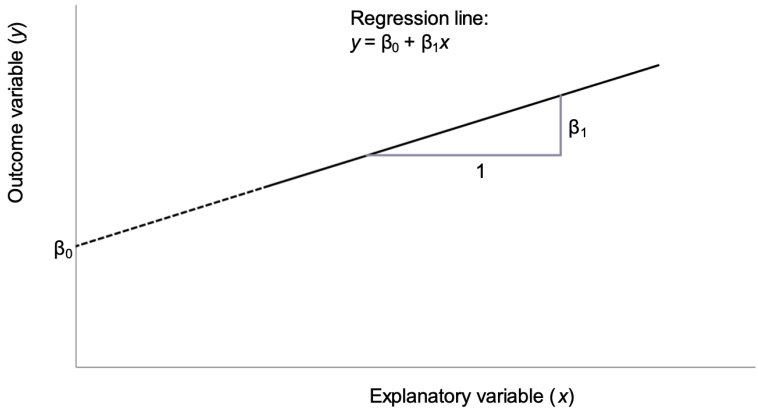


Figure 1.4: Coefficients of a linear regression equation

The slope of the line is the predicted change in the outcome variable 'y' as the explanatory explanatory variable 'x' increases by 1 unit.
An important concept is that regression predicts an expected value of 'y' given an observed value of 'x': any error around the explanatory variable is not taken into account. For this reason, measurements that can be taken accurately, such as age and height, make good explanatory variables.

1.4 Regression coefficients: estimation

Software is always used to estimate the regression equation for a set of data, using the *method of least squares*. This method estimates the intercept and the slope, and also their variability (i.e. standard errors).

Table 1.3: Notation for linear regression equation

Symbol	Interpretation
y	Observed value of the outcome variable
x	Observed value of the explanatory variable
β_0	Intercept of the regression line
β_1	Slope of the regression line

1.5 Regression coefficients: inference

We can use the estimated regression coefficients and their variability to calculate 95% confidence intervals. Here, a t-value from a t-distribution with

$$n - 2$$

degrees of freedom is used:
b0 plus/minus t_n-2 * SE(b0) *b1 plus/minus t_n-2* SE(b1)
Note that as the constant (b0) is not often biologically plausible, the 95% confidence interval for the constant is often not reported.
The significance of the estimated slope (and less commonly, intercept) can be tested using a t-test. The null hypotheses and the alternative hypothesis for testing the slope of a simple linear regression model are:
H0: beta_1 = 0 H1: beta_1 NE 0
To test the null hypothesis for the regression coefficient beta_1, the following t-test is used:
t = beta_1 /SE(beta_1)
This will give a t statistic with an underlying t distribution with n − 2 degrees of freedom, with a corresponding P-value.
Table X.X shows the estimated regression coefficients for our working example.

Table 1.4: Estimated regression coefficients

Term	Estimate	Standard error	t value	P value	95% Confidence interval
Intercept	-18.9	2.19	t=-8.60, 118df	$<$0.001	-23.22 to -14.53
Height	0.14	0.013	t=10.58, 118df	$<$0.001	0.11 to 0.17

From this output, we see that the slope is estimated as 0.141 with an estimated intercept of -18.873. Therefore, the regression equation is estimated as:
FVC (L) = − 18.873 + (0.141 $\times$ Height in cm)
This equation can be used to predict FVC for a person of a given height. For example, the predicted FVC for a person 165 cm tall is estimated as:
FVC = − 18.87347 + (0.1407567 $\times$ 165.0) = 4.40 L.
Note that for the purpose of prediction we have kept all the decimal places in the coefficients to avoid rounding error in the intermediate calculation.
From this model, there is very strong evidence of a linear association between FVC and height in cm (P < 0.001).

1.5.1 Fit of a linear regression model

After fitting a linear regression model, it is important to know how well the model fits the observed data. One way of assessing the model fit is to compute a statistic called coefficient of determination, denoted by R^2. It is the square of the Pearson correlation coefficient r : $r^2 = R^2$. Since the range of r is from −1 to 1, R^2 must lie between 0 and 1.
R^2 can be interpreted as the proportion of variability in y that can be explained by variability in x. Hence, the following conditions may arise:
If $R^2 = 1$, then all variation in y can be explained by variation of x and all data points fall on the regression line.
If $R^2 = 0$, then none of the variation in y is related to x at all, and the variable x explains none of the variability in y.
If $0 < R^2 < 1$, then the variability of y can be partially explained by the variability in x. The larger the R^2 value, the better is the fit of the regression model.

1.5.2 Assumptions for linear regression

Regression is robust to moderate degrees of non-normality in the variables, provided that the sample size is large enough and that there are no influential outliers. Also, the regression equation describes the relationship between the variables and this is not influenced as much by the spread of the data as the correlation coefficient is.
The assumptions that must be met when using linear regression are as follows:

- observations are independent;
- the relationship between the explanatory and the outcome variable is linear;
- the residuals are normally distributed.

A residual is defined as the difference between the observed and predicted outcome from the regression model. If the predicted value of the outcome variable is denoted by $\widehat{y}$ then:

$$\text{Residual} = \text{observed} - \text{predicted} = y - \widehat{y}$$

It is important for regression modelling that the data are collected in a period when the relationship remains constant. For example, in building a model to predict normal values for lung function the data must be collected when the participants have been resting and not exercising and people taking bronchodilator medications that influence lung capacity should be excluded. In regression, it is not so important that the variables themselves are normally distributed, but it is important that the residuals are. Scatter plots and specific diagnostic tests can be used to check the regression assumptions. Some of these will not be covered in this introductory course but will be discussed in detail in the **Advanced Biostatistics** course.
The distribution of the residuals should always be checked. Large residuals can indicate unusual points or points that may exert undue influence on the estimated regression slope.
The histogram of residuals from the model is shown in Figure 8.5. They are normally distributed and indicate that there are no influential outliers so that the assumptions for regression are met.

1.5.3 Critical appraisal

When reading the literature, it is important to be critical about how correlation coefficients are interpreted. It is a good idea to check if a scatter plot is shown to help interpret the relationship and to indicate if there are any influential outliers. Also, question whether the correlation coefficient has been calculated from a random sample and if not, what selected samples the value can be generalised to.
When regression is reported it is essential that the axes are correctly presented so that the equation is predictive. Thus, the explanatory variable must be presented on the x axis and the outcome on the y axis. It is also a good idea to check that all the assumptions are met. Outliers which result in a non-normal distribution of the residuals can severely bias the regression coefficients.

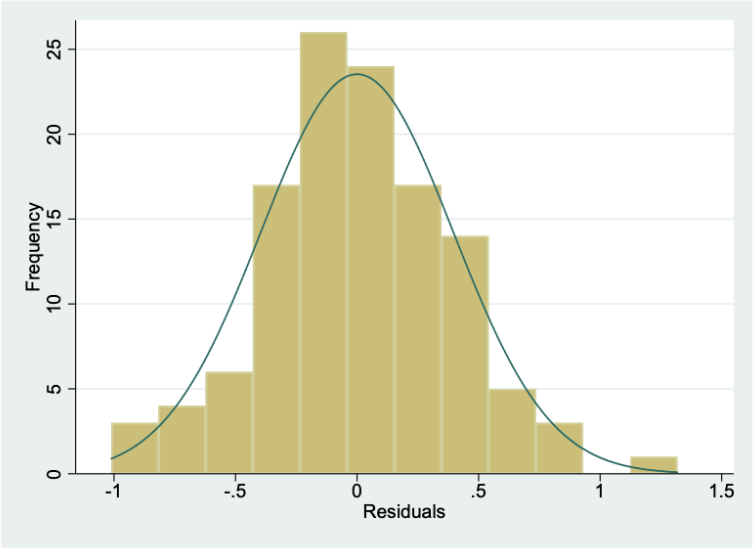


Figure 1.5: Histogram of regression residuals

1.6 Multiple linear regression

In the above example, we have only used a simple linear regression model of two continuous variables. Other more complex models can be built from this e.g. if we wanted to look at the effect of gender (male vs. female) as binary indicator in the model while adjusting for the effect of height. In that case we would include both the variables in the model as explanatory variables. In the same way we can include any number of explanatory variables (both continuous and categorical) in the model: this is called a multivariable model. Multivariable models are often used for building predictive equations, for example by using age, height, gender and smoking history to predict lung function, or to adjust for confounding and detect effect modification to investigate the association between an exposure and an outcome factor.

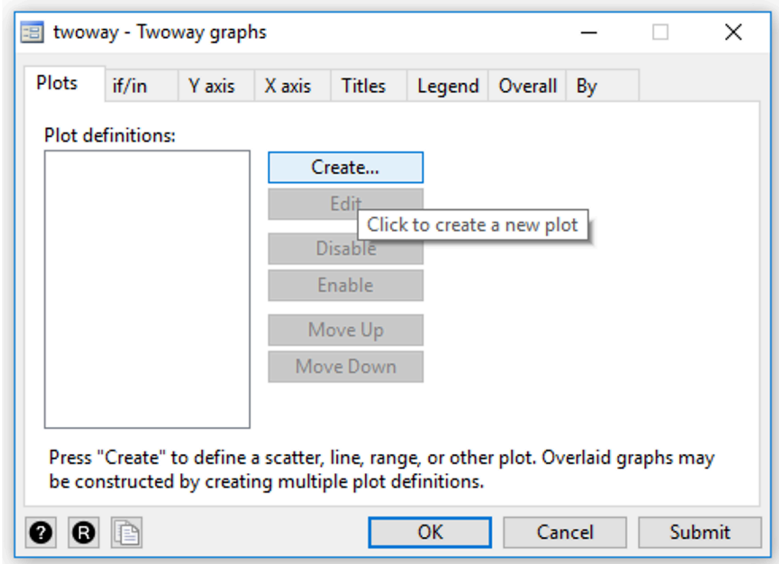
Multiple regression has an important role in investigating causality in epidemiology. The exposure variable under investigation must stay in the model and the effects of other variables which can be confounders or effect-modifiers are tested. The biological, psychological or social meaning of the variables in the model and their interactions are of great importance for interpreting theories of causality.

Other multivariable models include binary logistic regression for use with a binary outcome variable, or Cox regression for survival analyses. These models, together with multiple regression, will be taught in **PHCM9517: Regression Methods in Biostatistics**.

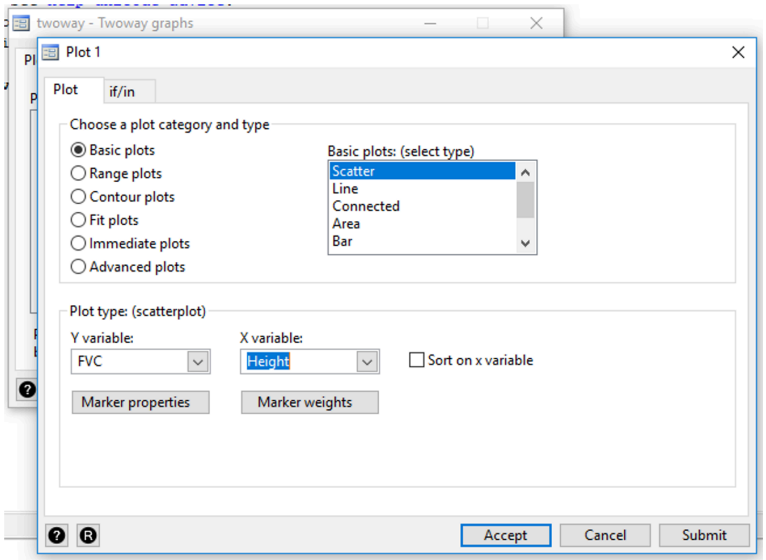
8 Stata notes

1.7 Creating a scatter plot

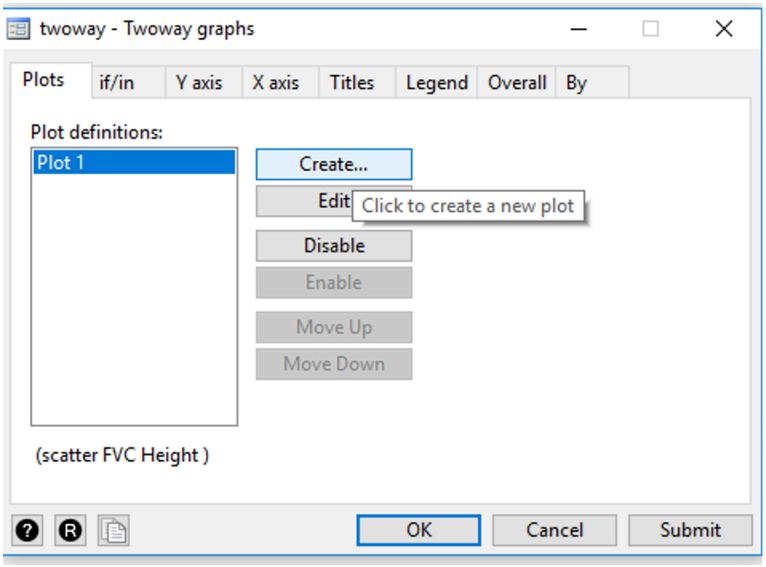
We will demonstrate using Stata for correlation and simple linear regression using the dataset `Example_8.1.dta`.
To create a scatter plot to explore the association between height and FVC click: **Graphics > Twoway graph (scatter, line, etc.)**. In the `twoway` dialog box, click **Create…**



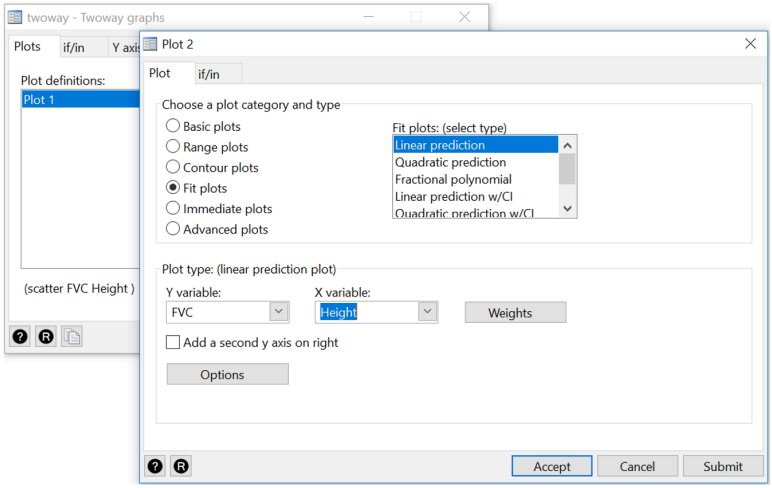
A new dialog box will open. Select the **Basic plots** radio button and highlight **Scatter** under **Basic plots: (select type)**. Choose **FVC** for the **Y variable** and **Height** for the **X variable**.



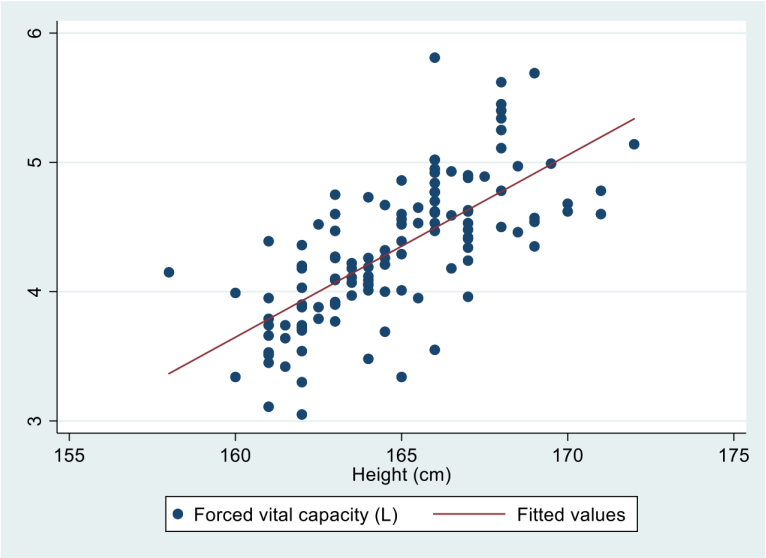
Click the **Accept** button in the **Plot 1** dialog box to return to the **twoway** dialog box, then click the **OK** or **Submit** button to produce the scatter plot shown in **Figure 8.1**.
[Command: `twoway (scatter FVC Height)`]
To add a fitted line, go back to the `twoway` dialog box. If you clicked the **OK** button, you can go to **Graphics > Twoway graph (scatter, line, etc.)** to bring it back again.



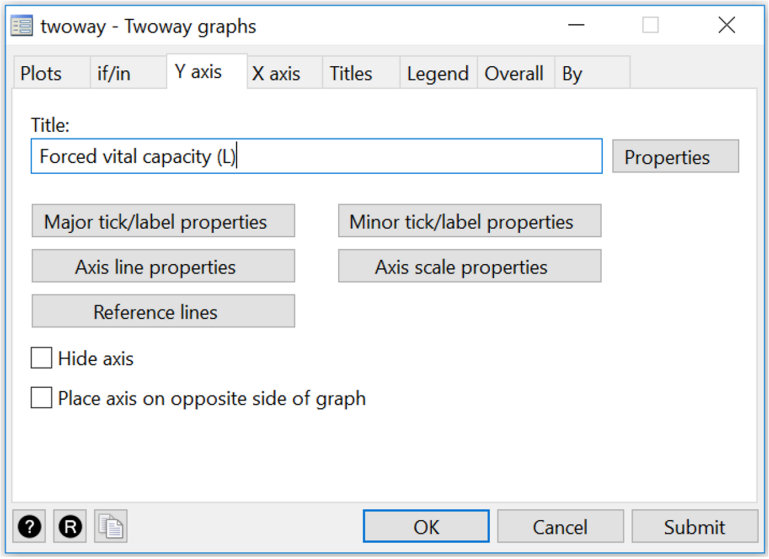
Click **Create…**, then select the **Fit plots** radio button and **Linear prediction** under **Fit plots: (select type)**. Choose **FVC** for the **Y variable** and **Height** for the **X variable**.



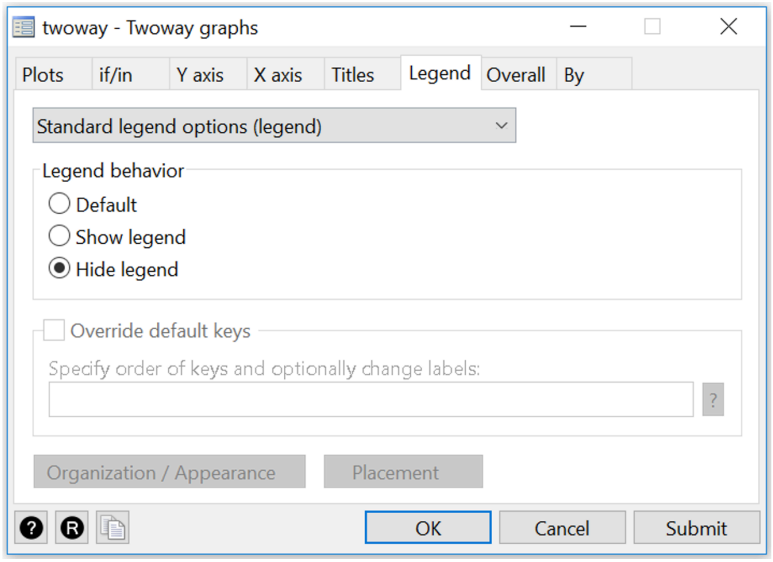
Click the **Accept** button, then the **OK** or **Submit** button to produce the scatterplot below.
[Command: `twoway (scatter FVC Height) (lfit FVC Height)`]



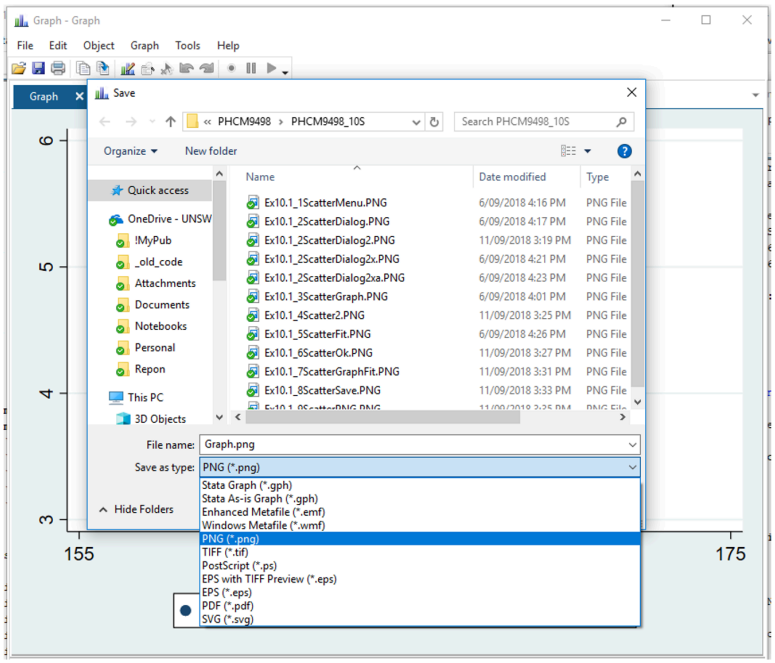
Notice that a legend now appears, and the y-axis title is missing. To add a y-axis title, go to the **Y axis** tab in the **twoway** dialog box to enter your title as shown below.



You can click the **Submit** button to check how the scatter plot looks like. Next go the **Legend** tab and select the **Hide legend** radio button.

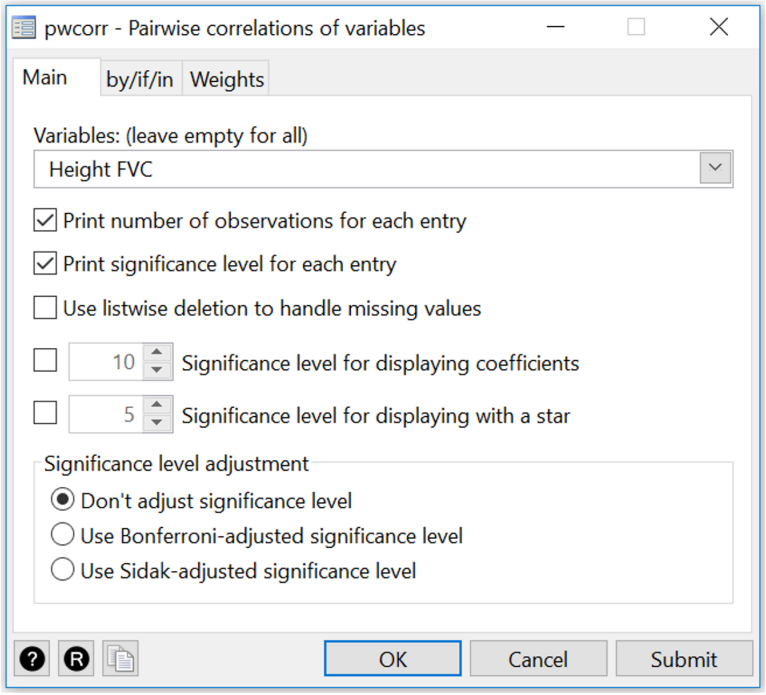


Click the **OK** or **Submit** button when you are finished to produce **Figure 8.3**.
[Command: `twoway (scatter FVC Height) (lfit FVC Height), ytitle(Forced vital capacity (L)) legend(off)`]
To save your graph, go to **File > Save** in the **Graph** window, and be sure to save your file as a PNG file:



1.8 Calculating a correlation coefficient

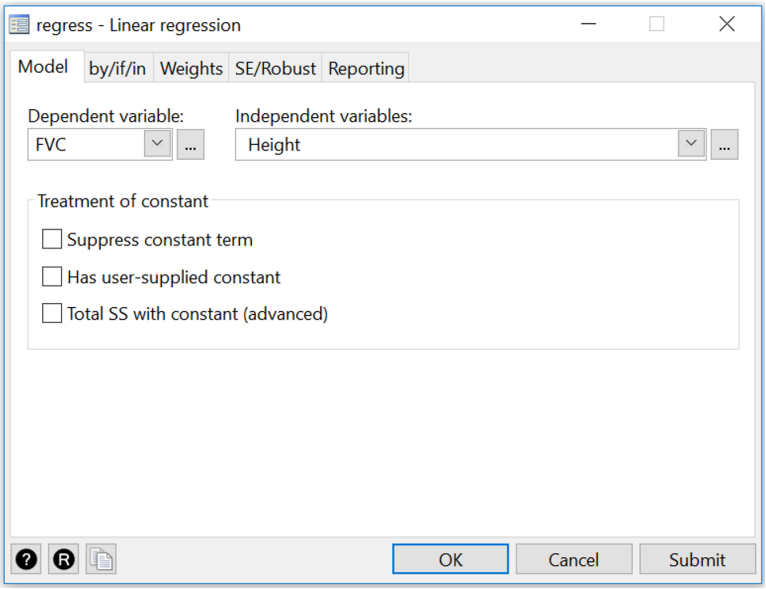
To calculate the Pearson's correlation using the dataset `Example_8.1.dta` go to: **Statistics > Summaries, tables, and tests > Summary and descriptive statistics > Pairwise correlations**
Select the two variables, **FVC** and **Height** in the **Variables** box. You can click the **Submit** button to check the output. Next, tick the box for **Print significance level for each entry** to obtain the P-value and the box for **Print number of observations for each entry** to obtain the number of observations used as shown below.



Click the **OK** or the **Submit** button when you are done to produce **Output 8.1**,
[Command: `pwcorr Height FVC, obs sig`]

1.9 Fitting a simple linear regression model

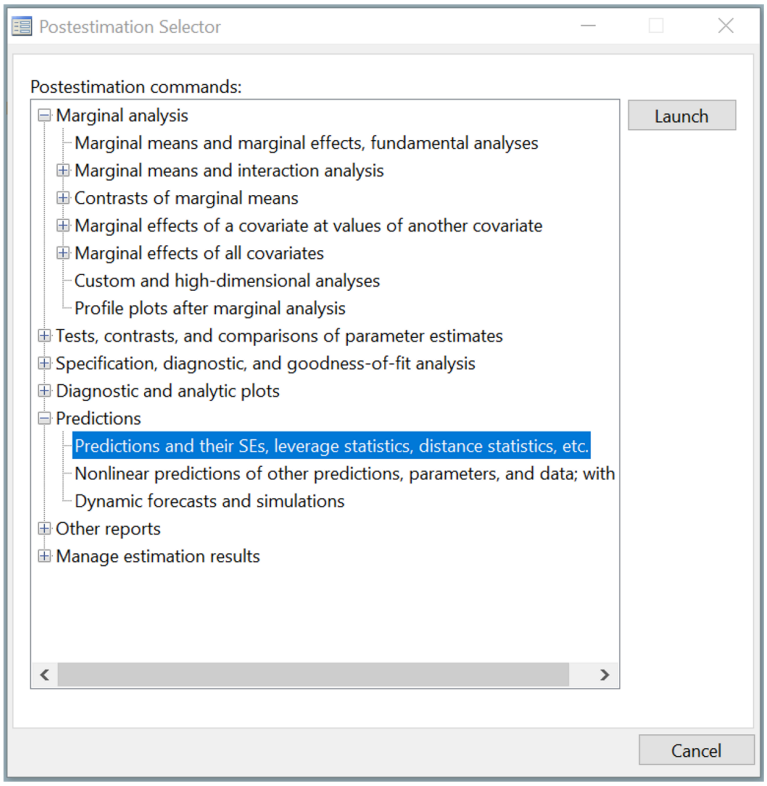
We will fit a simple linear regression model with `Example_8.1.dta` to quantify the relationship between FVC and height.
Choose **Statistics > Linear models and related > Linear regression**
In the `regress` dialog box, select `FVC` as the **Dependent variable**, and `Height` as the **Independent variable**.



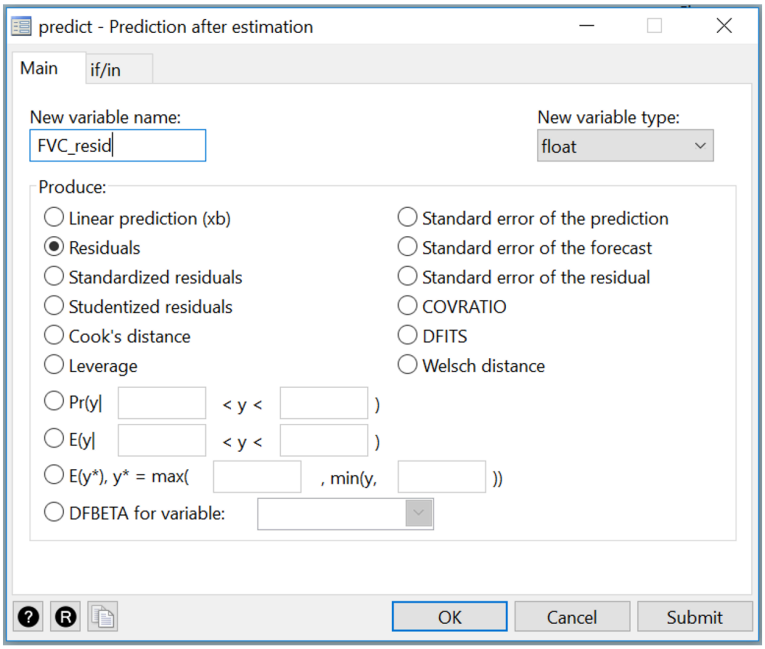
Click the **OK** or the **Submit** button when you are done to produce **Outputs 8.2 and 8.3**.
[Command: `reg FVC Height`]

1.10 Plotting residuals from a simple linear regression

To obtain the residuals, go to **Statistics > Post estimation** after running the regress command.
In the `Postestimation Selector` dialog box, select **Predictions and their SEs, leverage statistics, distance statistics, etc.** in the list under Predictions as shown below.

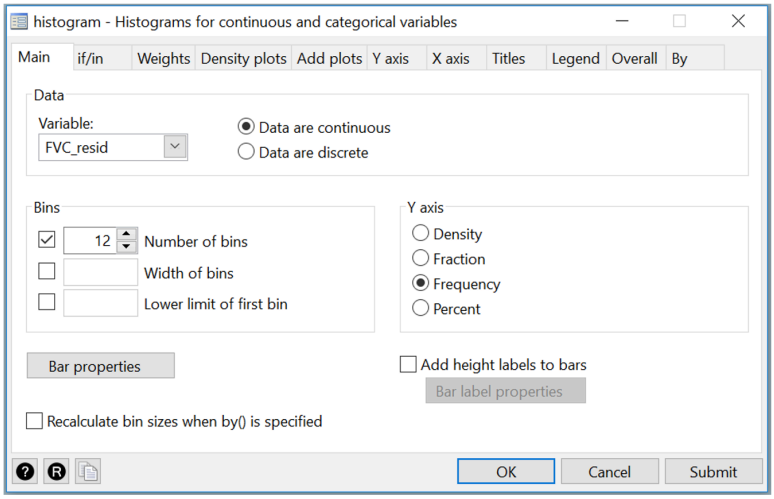


In the `predict` dialog box, choose the **Residuals** button and enter a New variable name (e.g. `FVC_resid`) for the residuals from the regression model.



Click **OK** button when you are done.
[Command: `predict FVC_resid, residuals`]

You can now check the assumption that the residuals are normally distributed by creating a histogram with the normal curve using **Graphics > Histogram** as shown in **Stata Notes** section for **Module 2**. Below is the **histogram** dialog box used to produce the graph in **Figure 8.5**.



[Command: `histogram FVC_resid, bin(12) frequency normal`]

Module 2

Correlation and simple linear regression

```
library(ggformula)
```

```
## Loading required package: ggstance
```

```
##
## Attaching package: 'ggstance'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     geom_errorbarh, GeomErrorbarh
```

```
## Loading required package: scales
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:huxtable':
##
##     number_format
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
## Loading required package: ggridges
```

```
##
## New to ggformula?  Try the tutorials:
##  learnr::run_tutorial("introduction", package = "ggformula")
##  learnr::run_tutorial("refining", package = "ggformula")
```

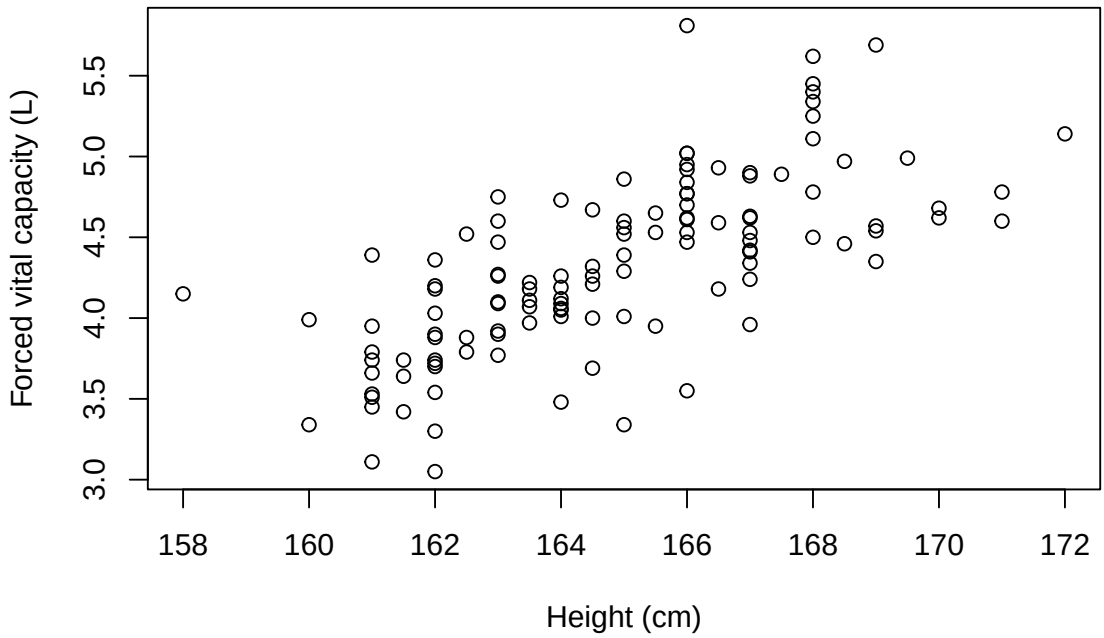
We will demonstrate using R for correlation and simple linear regression using the dataset `Example_8.1.csv`.

```
lung <- read.csv("data/examples/Example_8.1.csv")
```

2.1 Creating a scatter plot

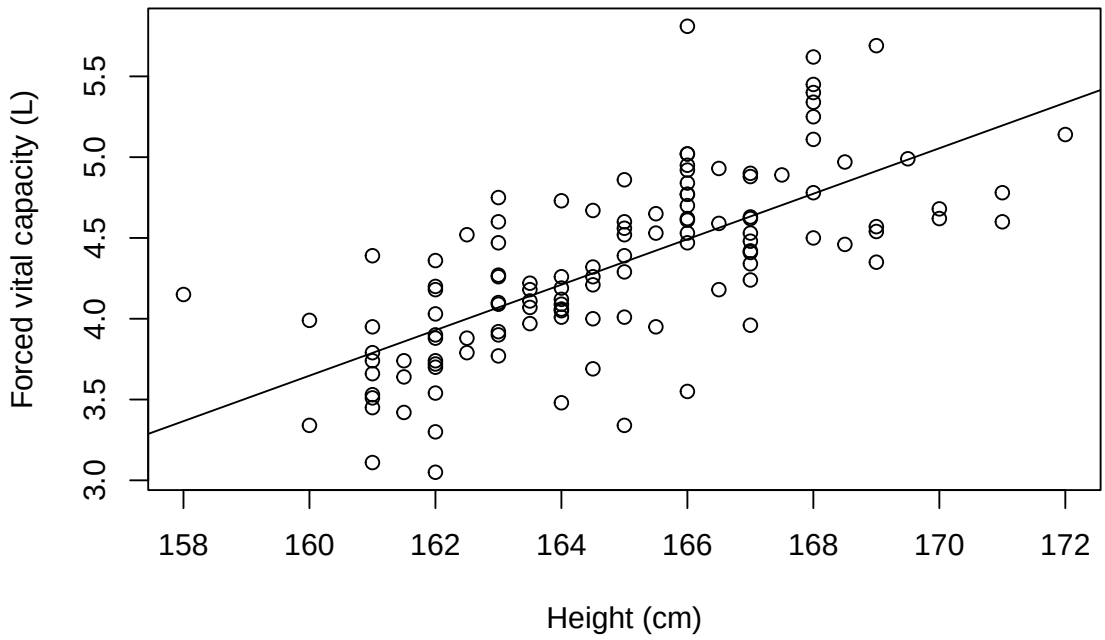
We can use the `plot` function to create a scatter plot to explore the association between height and FVC, assigning meaningful labels with the `xlab` and `ylab` commands:

```
plot(x=lung$Height, y=lung$FVC,
     xlab="Height (cm)",
     ylab="Forced vital capacity (L)")
```



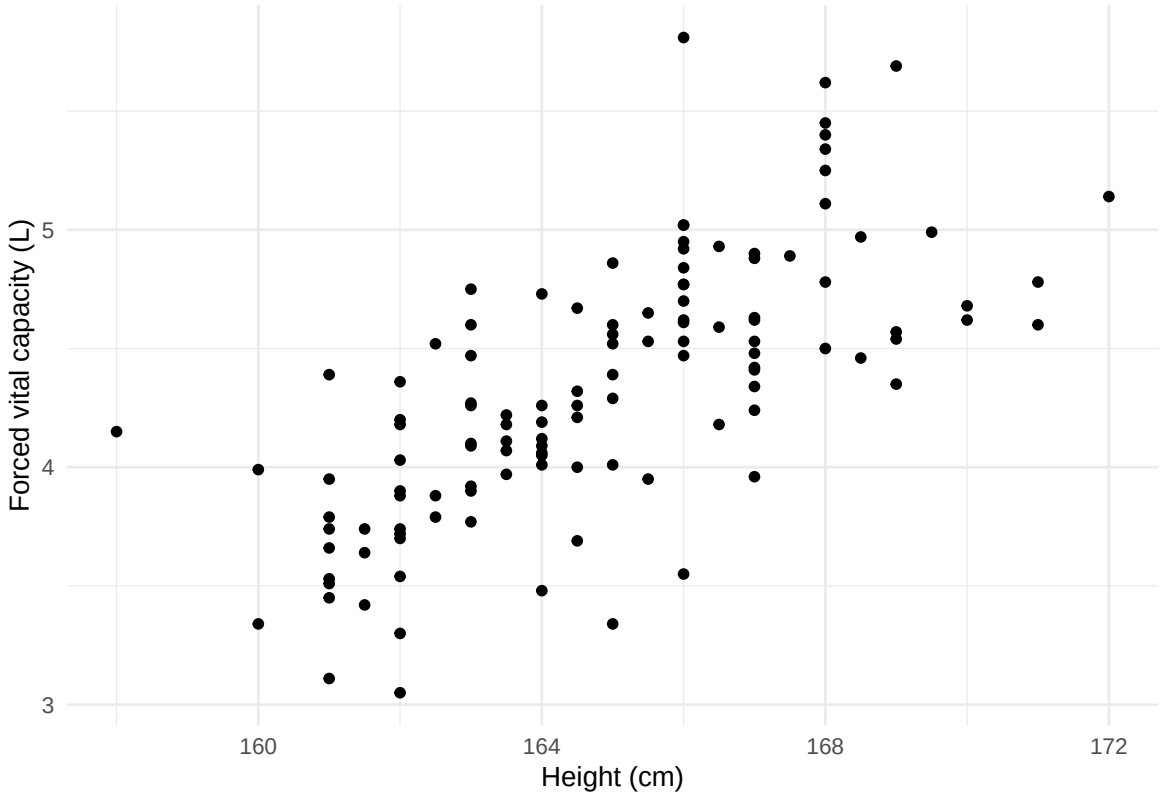
To add a fitted line, we can use the `abline()` function which adds a straight line to the plot. The equation of this straight line will be determined from the estimated regression line, which we specify with the `lm()` function, which fits a *linear model*.
The basic syntax of the `lm()` function is: `lm(y ~ x)` where `y` represents the *outcome* variable, and `x` represents the *explanatory* variable. Putting this all together:

```
plot(x=lung$Height, y=lung$FVC,
     xlab="Height (cm)",
     ylab="Forced vital capacity (L)")

abline(lm(lung$FVC ~ lung$Height))
```



Or using the `ggformula` package, we form the basic plot using the following:

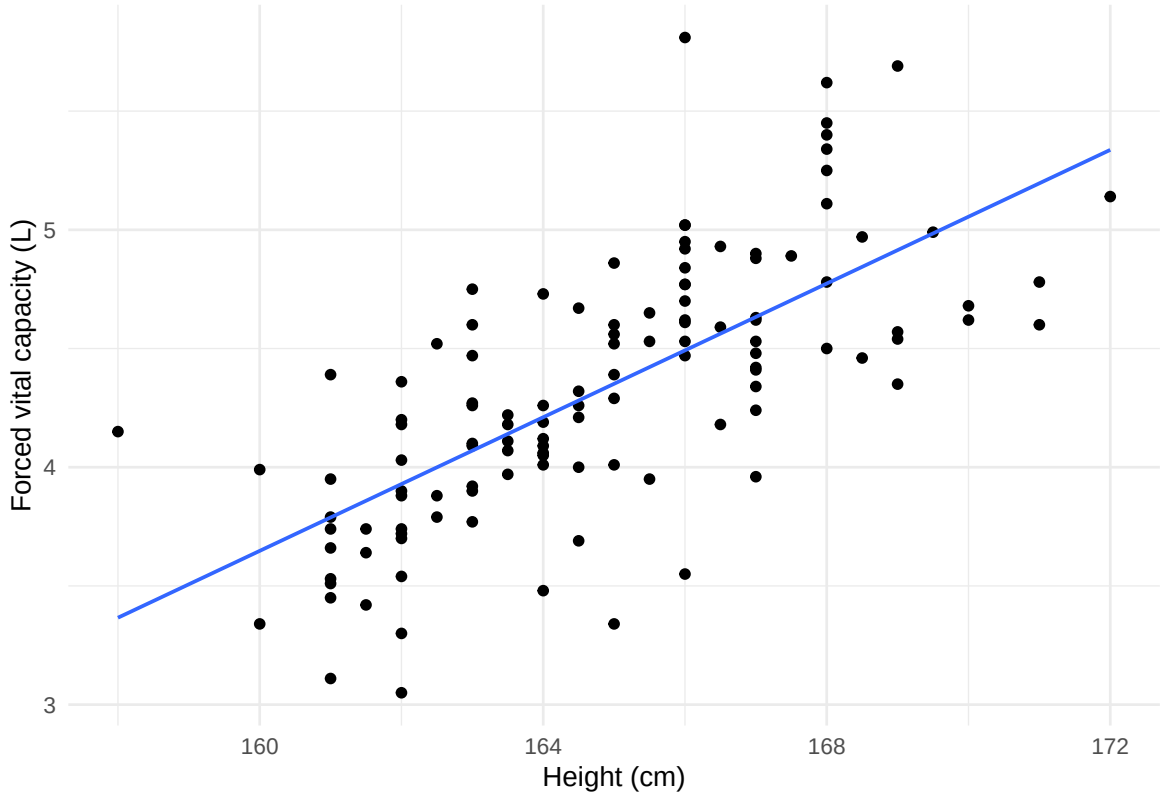
```
gf_point(FVC ~ Height, data=lung,
     xlab="Height (cm)",
     ylab="Forced vital capacity (L)") |>
  gf_theme(theme = theme_minimal())
```



We can add an estimated linear regression line by piping the command `gf_lm()`:

```
gf_point(FVC ~ Height, data=lung,
     xlab="Height (cm)",
     ylab="Forced vital capacity (L)") |>
  gf_lm() |>
  gf_theme(theme = theme_minimal())
```

```
## Warning: Using the `size` aesthietic with geom_line was deprecated in ggplot2 3.4.0.
## i Please use the `linewidth` aesthetic instead.
```



Calculating a correlation coefficient

We can use the `cor.test(x, y)` function to calculate a Pearson's correlation coefficient:

```
cor.test(lung$Height, lung$FVC)
```

```
##
##  Pearson's product-moment correlation
##
## data:  lung$Height and lung$FVC
## t = 10.577, df = 118, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5924715 0.7794090
## sample estimates:
##       cor
## 0.6976279
```

2.2 Fitting a simple linear regression model

We can use the `lm` function to fit a simple linear regression model, specifying the model as `y ~ x` where `y` represents the *outcome* variable, and `x` represents the *explanatory* variable. Using `Example_8.1.rds`, we can quantify the relationship between FVC and height:

```
lm(lung$FVC ~ lung$Height)
```

```
##
## Call:
```

```
## lm(formula = lung$FVC ~ lung$Height)
##
## Coefficients:
## (Intercept)  lung$Height
##    -18.8735       0.1408
```

The default output from the `lm` function is rather sparse. We can obtain much more useful information by defining the linear regression model as an object, then using the `summary()` function:

```
model <- lm(lung$FVC ~ lung$Height)
summary(model)
```

```
##
## Call:
## lm(formula = lung$FVC ~ lung$Height)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.01139 -0.23643 -0.02082  0.24918  1.31786
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -18.87347    2.19365  -8.604 3.89e-14 ***
## lung$Height   0.14076    0.01331  10.577  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3965 on 118 degrees of freedom
## Multiple R-squared:  0.4867, Adjusted R-squared:  0.4823
## F-statistic: 111.9 on 1 and 118 DF,  p-value: < 2.2e-16
```

Finally, we can obtain 95% confidence intervals for the regression coefficients using the `confint` function:

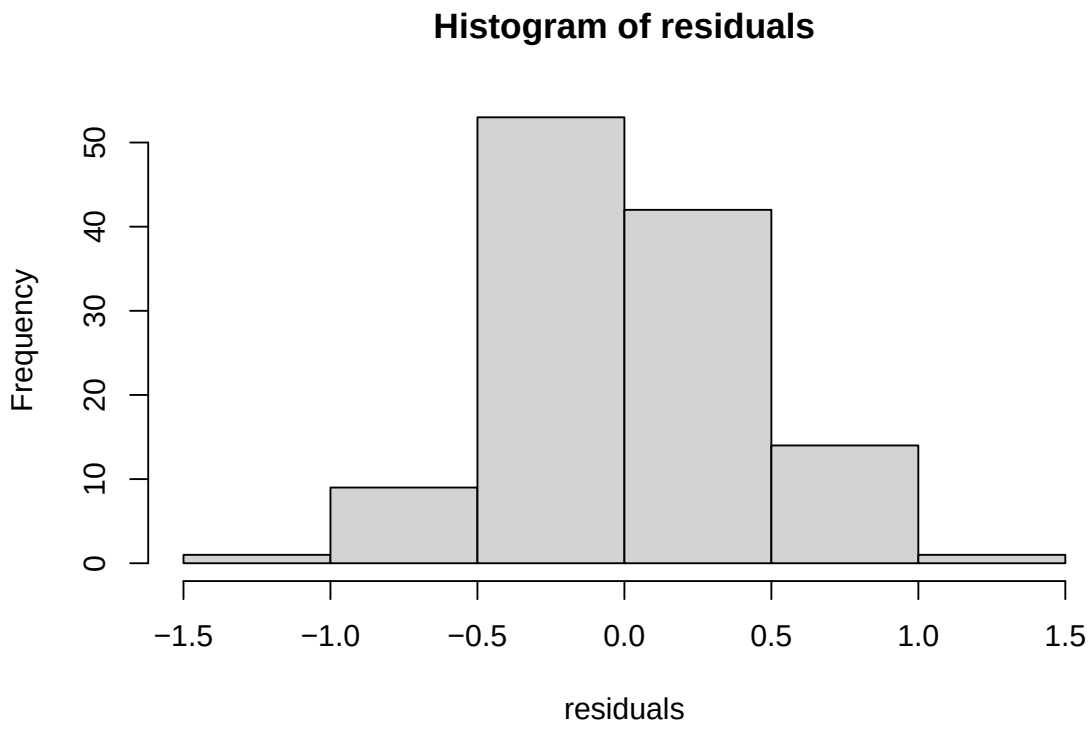
```
confint(model)
```

```
##                   2.5 %      97.5 %
## (Intercept) -23.2174967 -14.5294441
## lung$Height   0.1144042   0.1671092
```

2.3 Plotting residuals from a simple linear regression

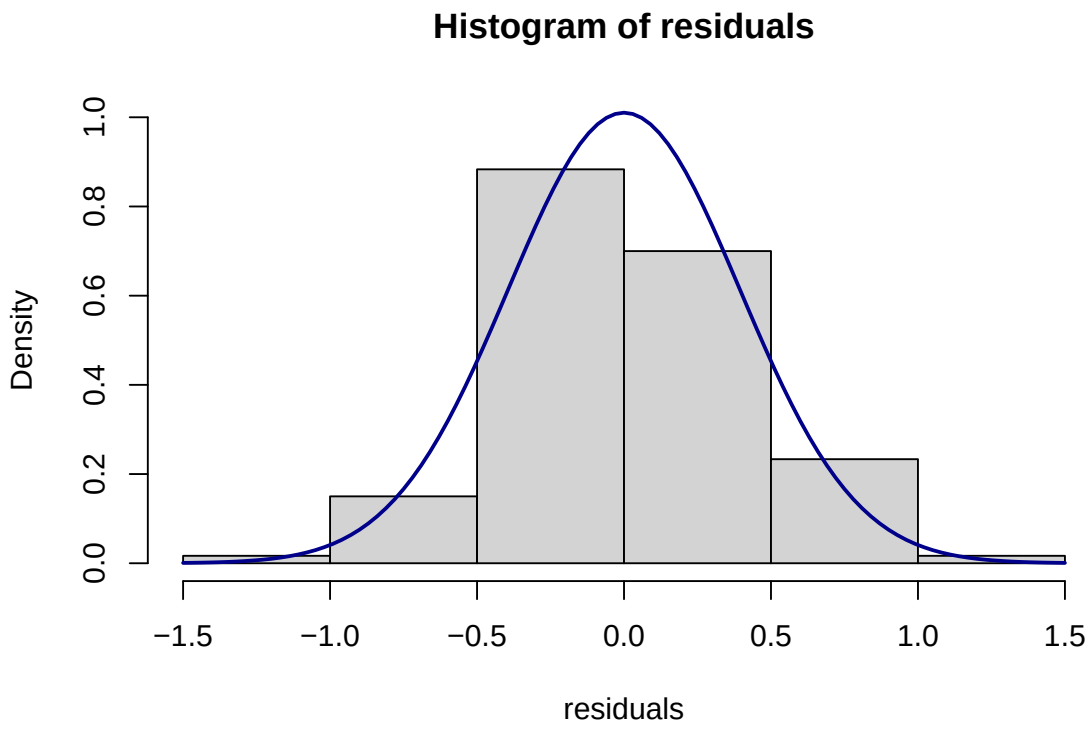
We can use the `resid` function to obtain the residuals from a saved model. These residuals can then be plotted using a histogram in the usual way:

```
residuals <- resid(model)
hist(residuals)
```



A Normal curve can be overlaid if we plot the residuals using a probability scale.

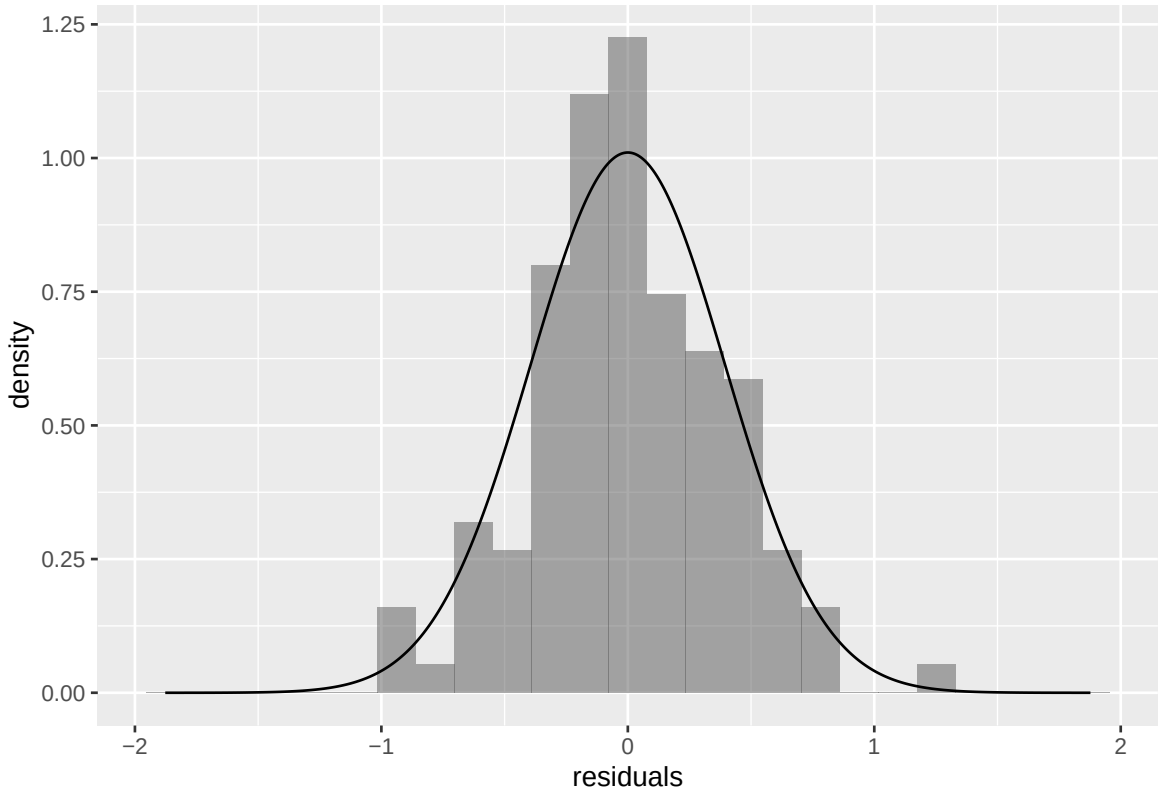
```
hist(residuals, probability = TRUE,
     ylim = c(0, 1))

curve(dnorm(x, mean=mean(residuals), sd=sd(residuals)),
      col="darkblue", lwd=2, add=TRUE)
```



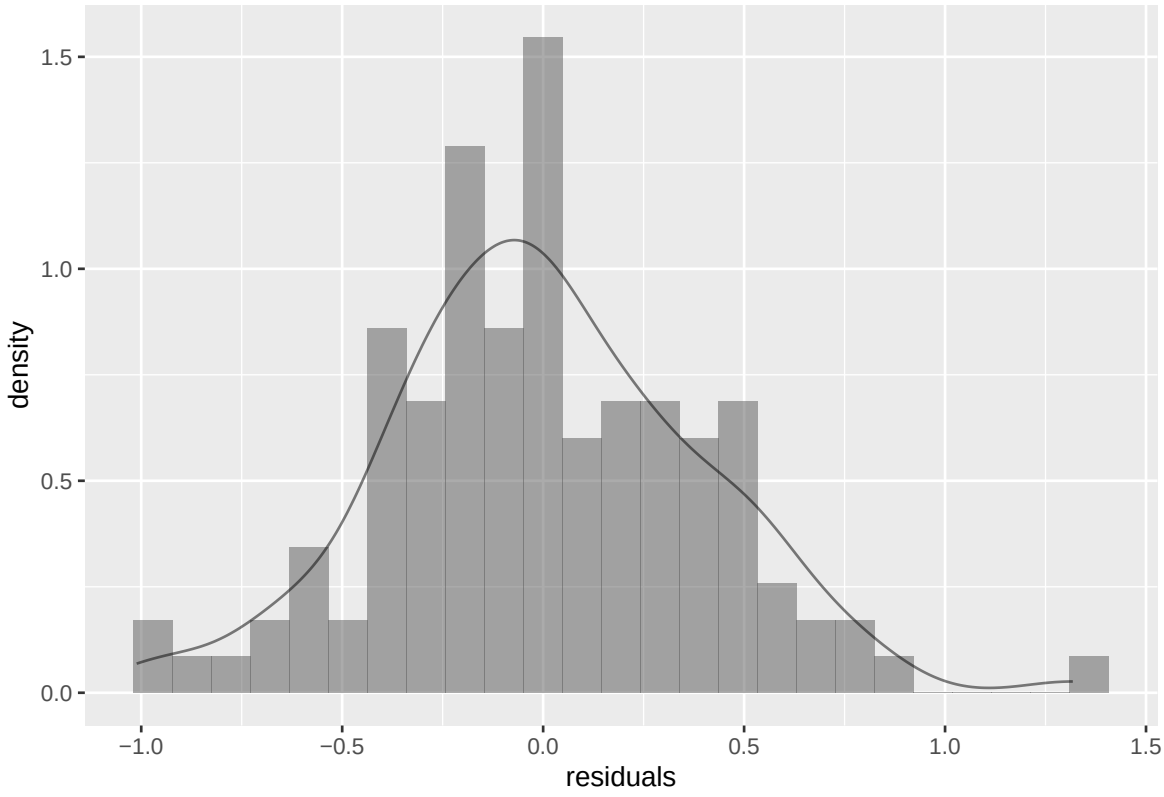
Using `ggformula`, we can plot the residuals as a histogram:

```
gf_dhistogram(~ residuals, data=model) |>
  gf_dist("norm",
          params=list(mean=mean(model$residuals),
                      sd=sd(model$residuals)))
```

```
## Warning: `stat(density)` was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
```




```
gf_dhistogram(~ residuals, data=model) |>
  gf_dens()
```


1.5
1.0
0.5
0.0
density
−1.0
−0.5
0.0
0.5
1.0
1.5
residuals

8 Learning Activities

Activity 8.1

To investigate how body weight (kg) effects blood plasma volume (mL), data were collected from 30 participants and a simple linear regression analysis was conducted. The slope of the regression was 68 (95% confidence interval 52 to 84) and the intercept was −1570 (95% confidence interval −2655 to −492).
[*You do not need Stata for this Activity*]

a) What is the outcome variable and explanatory (exposure) variable?
b) Interpret the regression slope and its 95% CI
c) Write the regression equation
d) If we randomly sampled a person from the population and found that their weight is 80kg, what would be the predicted value of plasma volume for this person?

Activity 8.2

To examine whether age predicts IQ, data were collected on 104 people. Use the data in the Stata file `Activity_8.2.dta` to answer the following questions.

a) What are the outcome variable and the explanatory variable?
b) Create a scatter plot with the two variables. What can you infer from the scatter plot?
c) Using Stata, obtain the correlation coefficient between age and IQ and interpret it.
d) Conduct a simple linear regression using Stata and report the relationship between the two variables including the interpretation of the R2 value. Are the assumptions for linear regression met in this model?
e) What could you infer about the association between age and IQ in the population, based on the results of the regression analysis in this sample?

Activity 8.3

Which of the following correlation coefficients indicates the weakest linear relationship and why?

a) r = 0.72
b) r = 0.41
c) r = 0.13
d) r = −0.33
e) r = −0.84

Activity 8.4

Are the following statements true or false?

a) If a correlation coefficient is closer to 1.00 than to 0.00, this indicates that the outcome is caused by the exposure.
b) If a researcher has data on two variables, there will be a higher correlation if the two means are close together and a lower correlation if the two means are far apart.

Bibliography

Alan C. Acock. *A Gentle Introduction to Stata*. Stata Press, College Station, Tex, 3rd edition, August 2010. ISBN 978-1-59718-075-7.

Martin Bland. *An Introduction to Medical Statistics*. Oxford University Press, Oxford, New York, 4th edition, July 2015. ISBN 978-0-19-958992-0.

Betty Kirkwood and Jonathan Sterne. *Essentials of Medical Statistics*. Wiley-Blackwell, Malden, Mass, 2nd edition, April 2001. ISBN 978-0-86542-871-3.