# **Introduction:**

**Simple Linear Regression:**

A simple linear regression uses the presence of a linear relationship to predict the value of a dependent variable based on the value of an independent variable. The dependent variable is also referred to as the outcome, target or criterion variable and the independent variable as the predictor, explanatory or regressor variable. A simple linear regression is also referred to as a bivariate linear regression or simply as a linear regression.

In order to run a simple linear regression, you require the following:

» One independent variable that is **continuous** (e.g., height, exam performance, etc.).  
» One dependent variable that is **continuous** (e.g., height, weight, etc.).

Simple linear regression can be used to answer the following problems:

### 1. Predict new values for the dependent variable given the independent variable

You can use simple linear regression to predict the value of one variable when you know the value of another variable. The value you are predicting is the dependent variable and the value you know is the independent variable. For example, you might have last year’s student's mid-term and final exam results for a biomechanics course. Using this data you construct a linear regression equation. When this year’s class sits the mid-term biomechanics exam, you use the linear regression equation to predict their performance in their final exam based on their mid-term exam results (even though they have not yet sat this final exam).

### 2. Determine how much of the variation in the dependent variable is explained by the independent variable

Often, your goal is not to make predictions, but to determine whether differences in your independent variable can help explain the differences in your dependent variable. This approach is more common in theory building, where you have proposed that your independent variable can help explain some of the variation of your dependent variable. Furthermore, you want to be able to quantify the degree to which your independent variable explains your dependent variable. For example, how much does the amount of time spent exercising influence cholesterol concentration (a fat in the blood linked to heart disease)?

**Multiple Regression**

A standard multiple regression allows you to predict a dependent variable based on multiple independent variables and is an extension to simple linear regression. The dependent variable is also referred to as the outcome, target or criterion variable and the independent variables as predictor, explanatory or regressor variables. This method also allows you to determine the overall fit (variance explained) of the model and the relative contribution of each of the predictors to the total variance explained.

## Y = b0 + b1X1 + b2X2 + e

## Where β0 is the intercept (also known as the constant), β1 is the slope parameter (also known as the slope coefficient) for X1, and so forth, and ε represents the errors. This type of statistical test relies on the initial assumption that there is, in fact, a linear relationship between each independent variable and the dependent variable and a linear relationship between the "composite" of the independent variables and the dependent variable. This assumption can be examined, as you will do in this guide. Confidence intervals can be calculated for the sample intercept and slope parameters to estimate the likely range of values that these parameters might take in the population. Furthermore, predictions can be made based on the regression equation calculated. You will calculate all these statistical measures in this guide.

## What is required

In order to run a multiple regression, you require the following:

1. Two or more independent variables that can be either **continuous** or **categorical** (e.g., height, exam performance, gender, etc.).
2. One dependent variable that is **continuous** (e.g., height, weight, etc.).

Multiple regression can be used to answer the following problems:

### 1. Predict new values for the dependent variable given the independent variables

You can use multiple regression to predict the value of one variable when you know the value of other variables. For example, you might have individuals' heights, weights, age and gender, and you want to predict running performance. Using this data you construct a multiple regression equation, which you then use to predict new individuals' running performance based on their measured physical properties (i.e., their height, weight, age and gender).

### 2. Determine how much of the variation in the dependent variable is explained by the independent variables

Often, your goal is not to make predictions, but to determine how much of the variation in the dependent variable can be explained by all the independent variables. In addition, you can use multiple regression to understand the relative, unique contribution of each independent variable towards this total. For example, you might have individuals' heights, weights, age and gender, and you want to predict running performance. You want to know how much of the variation in running performance can be explained by the predictor variables. Additionally, you want to know the relative contribution of each predictor to the explanation of variance.

**Assumptions of the Regression:**

## Assumption #1: Your two or more independent variables should be continuous or nominal and the dependent variable should be measured at the interval or ratio level (i.e., they are continuous). Examples of variables that meet this criterion include revision time (measured in hours), intelligence (measured using IQ score), exam performance (measured from 0 to 100), weight (measured in kg), and so forth.

## Assumption #2: There needs to be a linear relationship between the two variables. The best way of checking this assumption is to plot a scatterplot and visually inspect the graph.

## Assumption #3: There should be no significant outliers. Outliers are data points within your sample that do not follow a similar pattern to the other data points.

In order to take your analysis to the next stage using prediction, which will be required for most undergraduate and above work, your errors in prediction (residuals) will have to pass the following assumptions:

* Assumption #4: Independence of errors (residuals).
* Assumption #5: Homoscedasticity of residuals (equal error variances).
* Assumption #6: Errors (residuals) are normally distributed.

These extra assumptions will allow you to (1) provide information on the accuracy of your predictions, (2) test how well the regression model fits your data, (3) determine the variation in your dependent variable explained by your independent variable, and (4) test hypotheses on your regression equation.

With real-world data, it is not uncommon for one or more of these six assumptions to be violated. This guide will, therefore, provide explanations of how to implement techniques to overcome these violations and move forward with your analysis, if this is indeed possible with your data. The assumptions will be tackled in the order they have been addressed above. This order has been chosen because it represents an order whereby if a violation is not correctable, you cannot proceed with the analysis (if you want valid results). For example, if you have violated assumption (1), it is pointless testing assumptions (2) through (6) because the regression analysis will already have been rendered invalid by failure of assumption (1).

## **Null and Alternative Hypotheses:**

Sometimes you will be required to explicitly state the null and alternative hypotheses for a simple linear regression, and then state which was accepted or rejected at the end of the experiment. The main null hypothesis for a simple linear regression is:

H0: b1 = 0, the coefficient of the slope equals 0 (zero)

And the alternative hypothesis is:

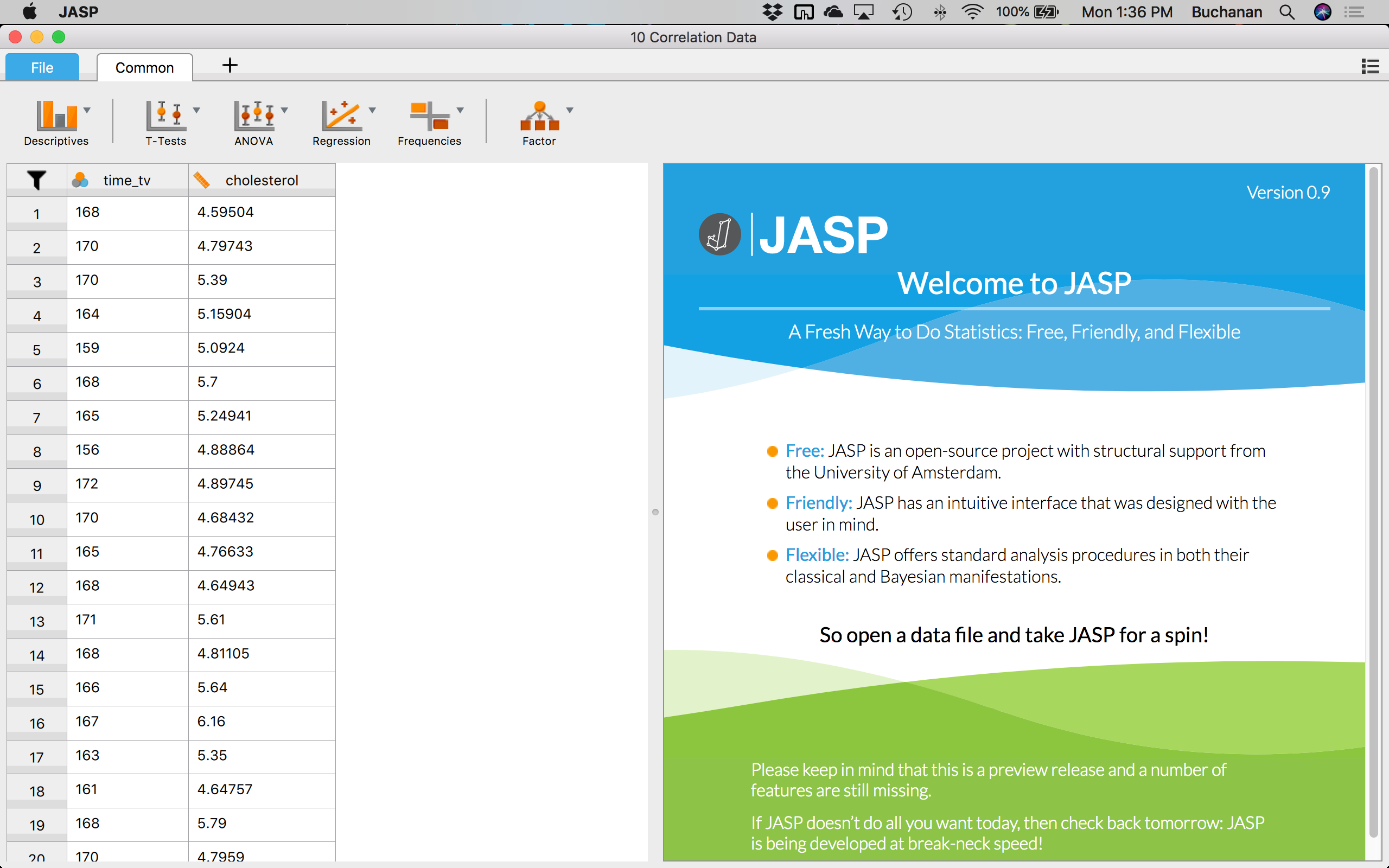
HA: b1 ≠ 0, the coefficient of the slope does not equal 0 (zero)

## **Example Simple Linear Regression:**

Studies show that exercising can help prevent heart disease. Within reasonable limits, the more you exercise, the less risk you have of suffering from heart disease. One way in which exercise reduces your risk is by reducing a fat in your blood called cholesterol. The more you exercise, the lower your cholesterol concentration. It has recently been shown that the amount of time you spend watching TV, an indicator of a sedentary lifestyle, might be a good predictor of heart disease; that is, the more TV you watch, the greater your risk of heart disease. Therefore, a researcher decided to determine if cholesterol concentration was related to time spent watching TV in otherwise healthy 45 to 65 year old men (an at-risk category of people). They believed that there would be a positive relationship: the more time people spent watching TV, the greater their cholesterol concentration. The researcher also wished to be able to predict cholesterol concentration and to know the proportion of cholesterol concentration that time spent watching TV could explain. Daily time spent watching TV was recorded in the variable time\_tv and cholesterol concentration recorded in the variable cholesterol. Expressed in variable terms, the researcher wants to regress cholesterol on time\_tv. (note: this data is fictitious. In addition, they did not decide to predict the direction of the relationship in the statistical analysis.)

To get started, open the dataset for this example in JASP. Remember, you can always use the previous help guides for greater detail in case you do not remember how to do something.

File 🡪 Open 🡪 Computer 🡪 Browse 🡪 Pick the Correlation Data.



## **Check your assumptions:**

**Is the independent variable continuous or nominal and the dependent variable at least scale (ratio or interval)?**

Yes, we are using ratio style data for both variables.

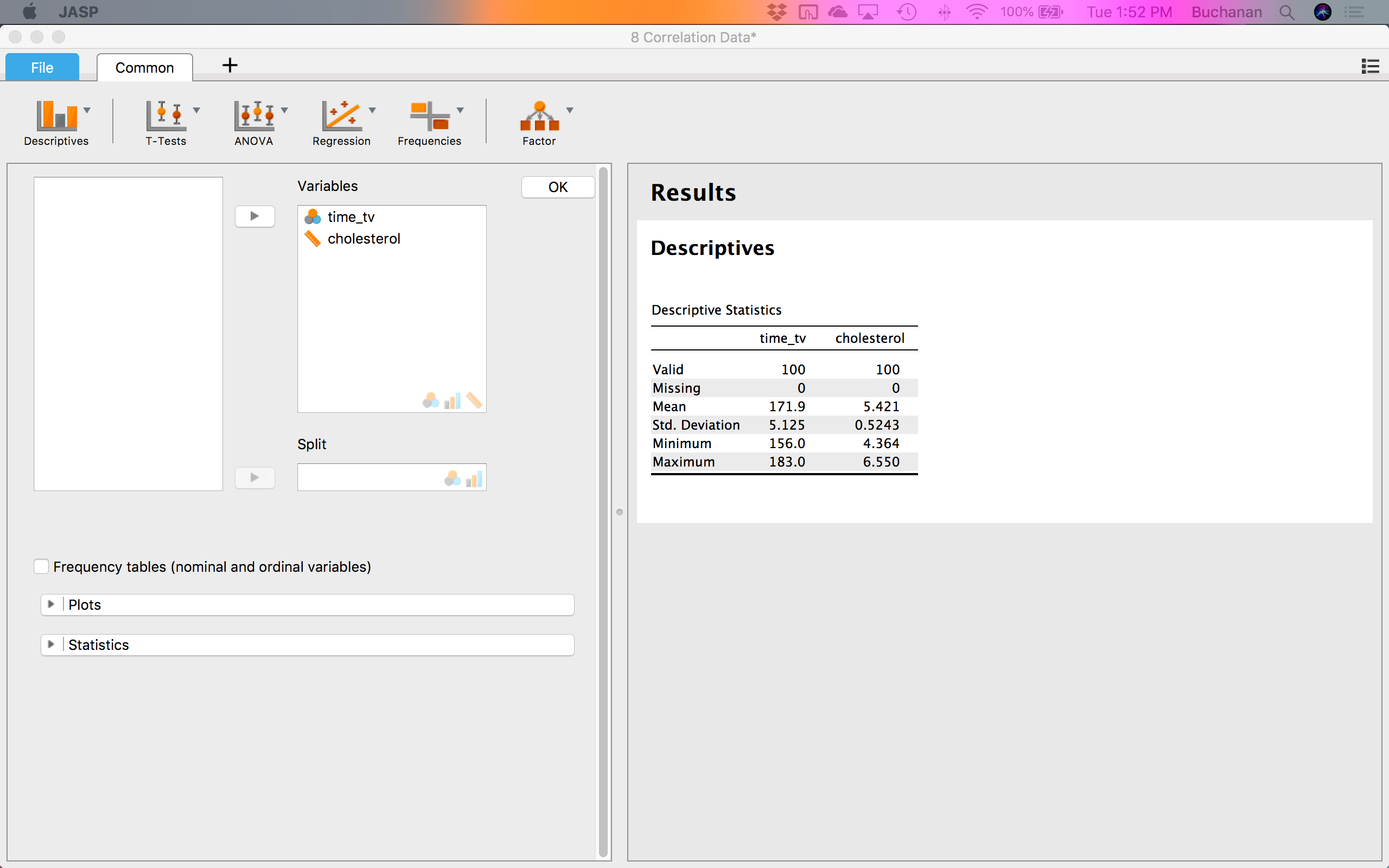
**Is there a linear relationship between the variables?**

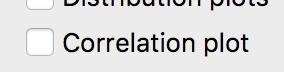
To get a scatterplot, we could use Excel as described in earlier chapters. However, there are graph options in JASP that allow you to answer this question and the outlier question below.

Click Descriptives 🡪 Descriptive Statistics.

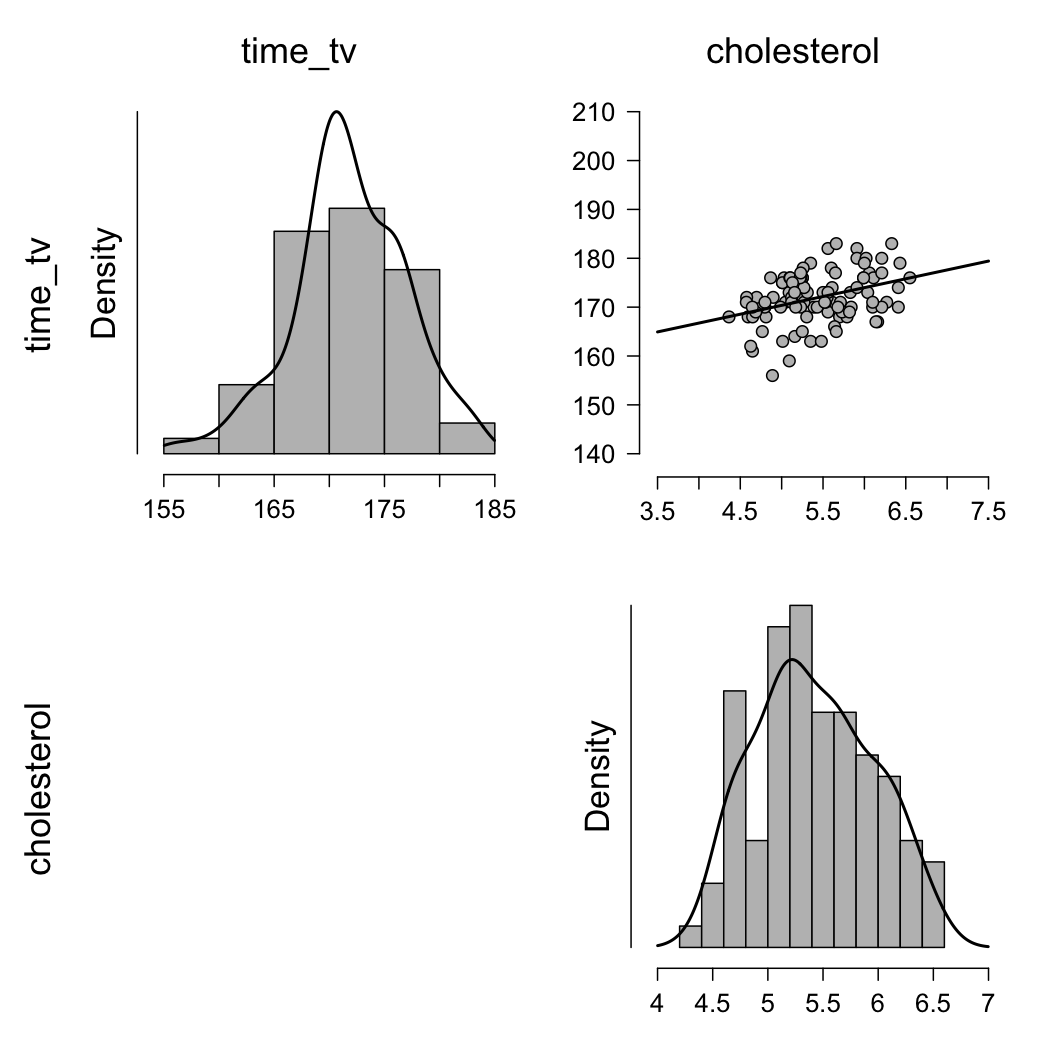


In this window, we want to click on both time\_tv and cholestorol and click the arrow  to move it over to the right hand side under Variables.



Click on the plots options:  to see more available options. Click on correlation plot . If you want to switch the X and Y axis, you can reorder the variables in the Variables window.

#### Correlation plot



The scatterplot is presented in the top right corner. You should inspect the scatterplot above and form an opinion as to whether you believe there is enough evidence to suggest the relationship is linear. The human brain is very good at visualizing straight lines and often you can rely on your own visual inspection to determine whether the relationship is a linear one or not. For this example, you can conclude from visual inspection of the above scatterplot that there is a linear relationship between cholesterol concentration and time spent watching TV. In other situations, the relationship can sometimes be a little bit more tricky to evaluate and more care will have to be taken (especially with setting the correct scales for the x-axis and y-axis).

In this example, the linear relationship between our variables is positive; that is, as the value of time\_tv increases, so does the value of cholesterol. However, when testing your own data, you might discover a negative relationship (i.e., as the value of one variable increases, the value of the other variable decreases). You might also find that your line/relationship might be more steep or more shallow than the line/relationship in this particular example. However, for assessing linearity, all that matters is whether or not the relationship is linear (i.e., a straight line) in order to proceed.

**Are there any outliers in the sample?**

Good news! You already have the graph for outliers. You can look at the scatterplot in the top right corner for outliers. Outliers are data points that do fit the pattern of the rest of the data set. These data points can often be easily identified from the scatterplot you plotted when testing for linearity. Further, you are given the distribution plots for the data, which have used before to determine if there are outliers. These plots are the top left and bottom right plots. From looking at these, we can see that there are not any bars that are far away from the others.

If you did have outliers, your first consideration should be to check whether you have made any data entry errors (i.e., simply keyed in any wrong values into Excel). If any of your outliers are due to data entry errors, you should replace them with the correct values and re-run scatterplot.

If you find that the outliers are not due to data entry errors, you should consider whether they are measurement errors (e.g., equipment malfunction or out-of-range values). Measurement errors usually result in you having to remove those data points from your analysis and this is most likely what you will have to do.

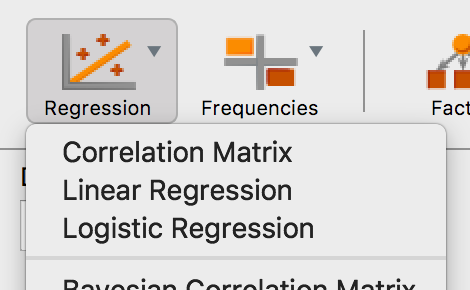
If you have established that an outlier is neither a data entry or measurement error, they most likely represent genuine data points. It is these genuine data points that are the hardest to deal with because although they are not ideal from a statistical perspective (i.e., they upset the linear regression analysis), there is no good reason to reject them as invalid.

**Keep the outlier(s):** If you do not want to remove an outlier, you have three choices: either (1) modify the outlier by replacing the outlier's value with one that is less extreme; (2) transform the dependent variable; or (3) include in the analysis anyway but remember to highlight the outlier in your report. You can also run the linear regression with and without the outlier, and if there is no appreciable difference in the results, keep the outlier. With respect to point two (2), transformation can be an option as it can lead to outliers being disproportionately affected ("reduced in size") so that they are no longer classified as outliers. Remember, however, that transformations can affect homoscedacity and normality so you should check these before transforming your data. Also, if you make any transformations, you will have to re-run all assumptions.

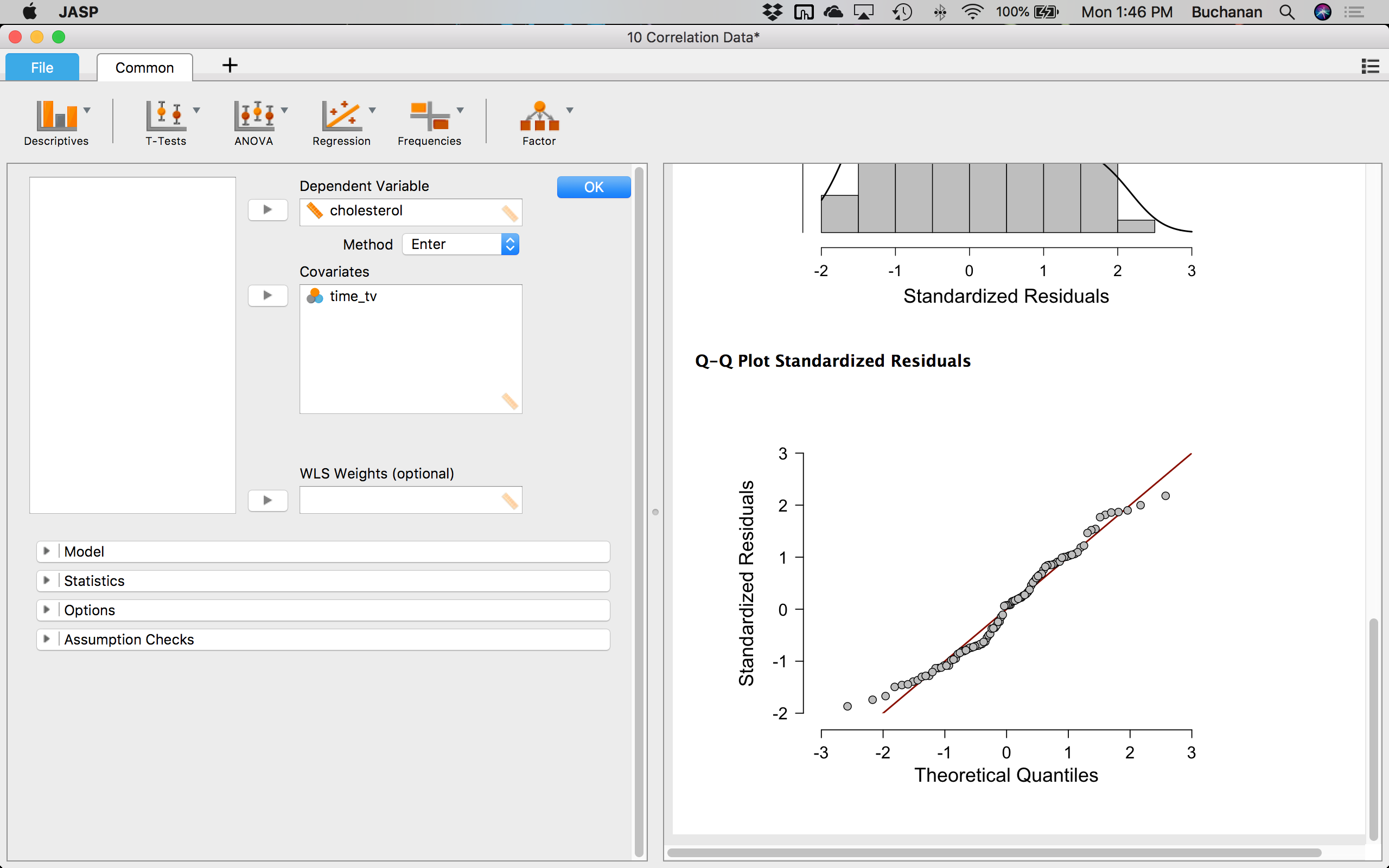
**Remove the outlier(s):** Alternatively, if you are happy to just remove the outlier, you can do so. Remember to note any decisions you take in your written results section. You can justify doing this as you would otherwise be compromising all your data by just one (or a small number) of data points. Your goal, after all, is to generalize your findings to a larger population. For example, if you did remove an outlier, provide information about that data point, so that a reader can make an informed opinion about why you removed it and how it might have affected your results. This can also help dispel any accusations that you might have removed a data point just to make your results look better.

One more question you can ask yourself when you have an outlier, is whether the participant's values highlight that their inclusion in the study should be reconsidered. For example, the outlier in this example has a cholesterol concentration of 7.98 mmol/L, which is a very high concentration, indicating considerable risk of heart disease. Although the study wanted to take a cross-section of individuals, it did not want to study individuals that might have possible underlying clinical complications or be at very high risk of heart disease. With such a high cholesterol concentration, this individual does not represent those that the study aims to generalize to. For this reason, you may justify removing this data point. We are going to remove this data point from our study as the cholesterol concentration is slightly suspect for an otherwise healthy individual, and also because we do not want this single individual having such an undue influence on the generalization of the results.

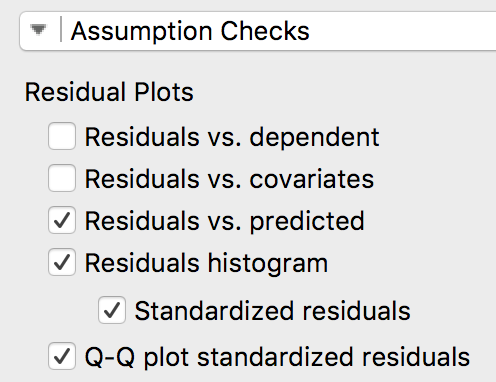
To get the next assumption tests, we need to run part of the regression analysis. Click Regression  🡪 Linear Regression.



Move the dependent variable into the dependent variable box, and put the independent variable in the covariates box.



Click on Assumption Checks  🡪 and check the following boxes: Residuals vs. predicted, Residuals Histogram, Standardized Residuals, and QQ plot standardized residuals.

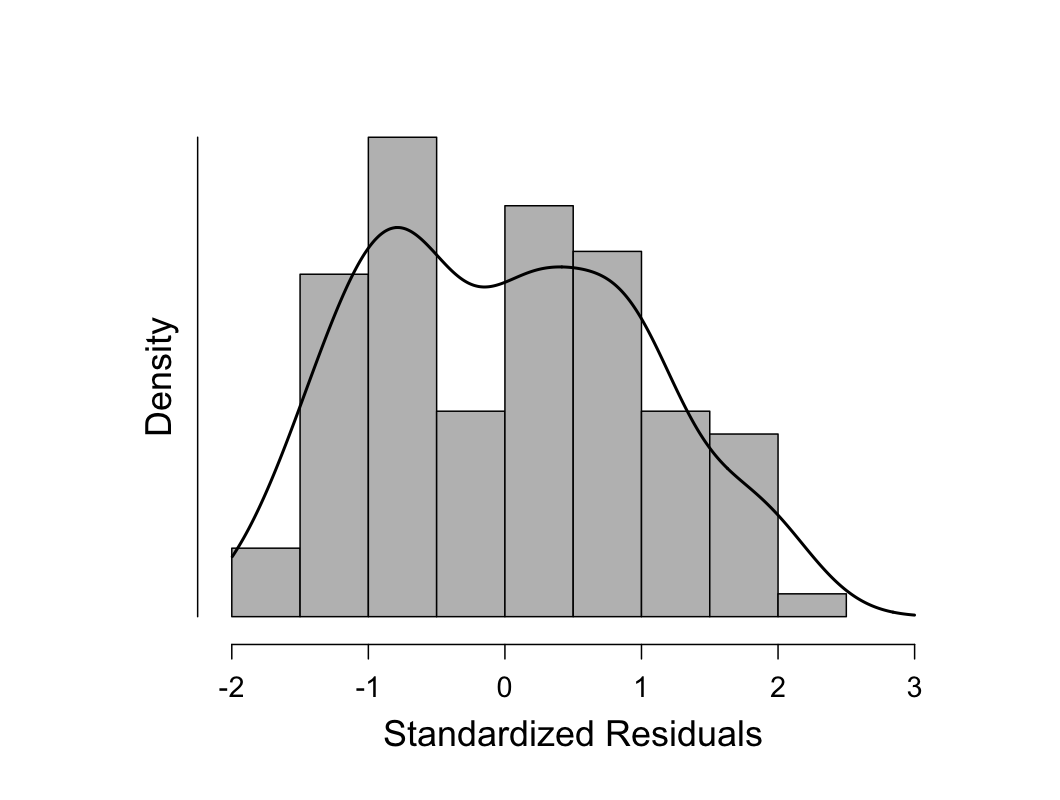


**Are the residuals normally distributed?**

Based on the options that you selected in the linear regression procedure, you will be presented with two methods to ascertain whether the residuals are normally distributed. These are the histogram and the Normal Q-Q Plot, as described below:

### Histogram

### Standardized Residuals Histogram

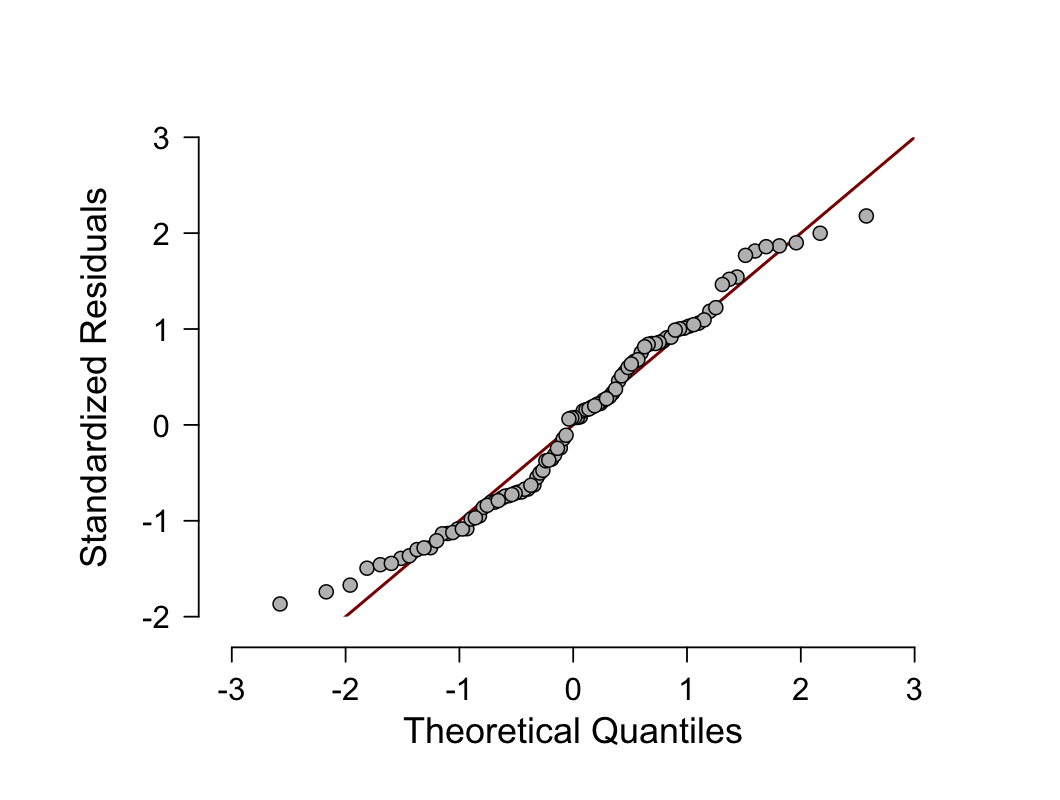


You can see from the above histogram that the standardized residuals appear to be approximately normally distributed. However, histograms can be deceptive as their appearance can be largely dependent on the selection of the correct bin width (column width). You would want to find that the data is centered around zero and looks normal (i.e., even on each side, the bell curve discussed previously).

To confirm your opinion of normality based on the visual inspection of the above histogram, you should also look at the Normal Q-Q Plot that was produced.

### Normal Q-Q Plot

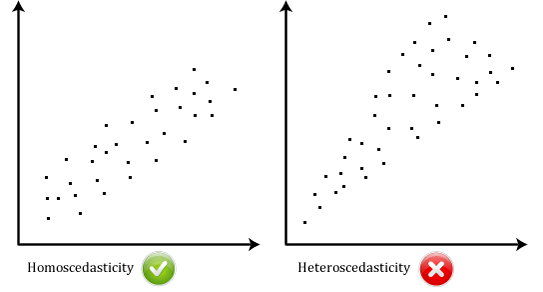
### Q-Q Plot Standardized Residuals



If the residuals are normally distributed, the points will be aligned along the diagonal line. In reality, the points will never be perfectly aligned along the diagonal line. Moreover, you only need the residuals to be approximately normally distributed as the regression analysis is fairly robust to deviations from normality. You can see from the above plot that although the points are not aligned perfectly along the diagonal line, they are close enough to indicate that the residuals are approximately normally distributed. As linear regression analysis is fairly robust against deviations from normality, you can accept this result as meaning that no transformations or otherwise need to take place; you have not violated the assumption of normality.

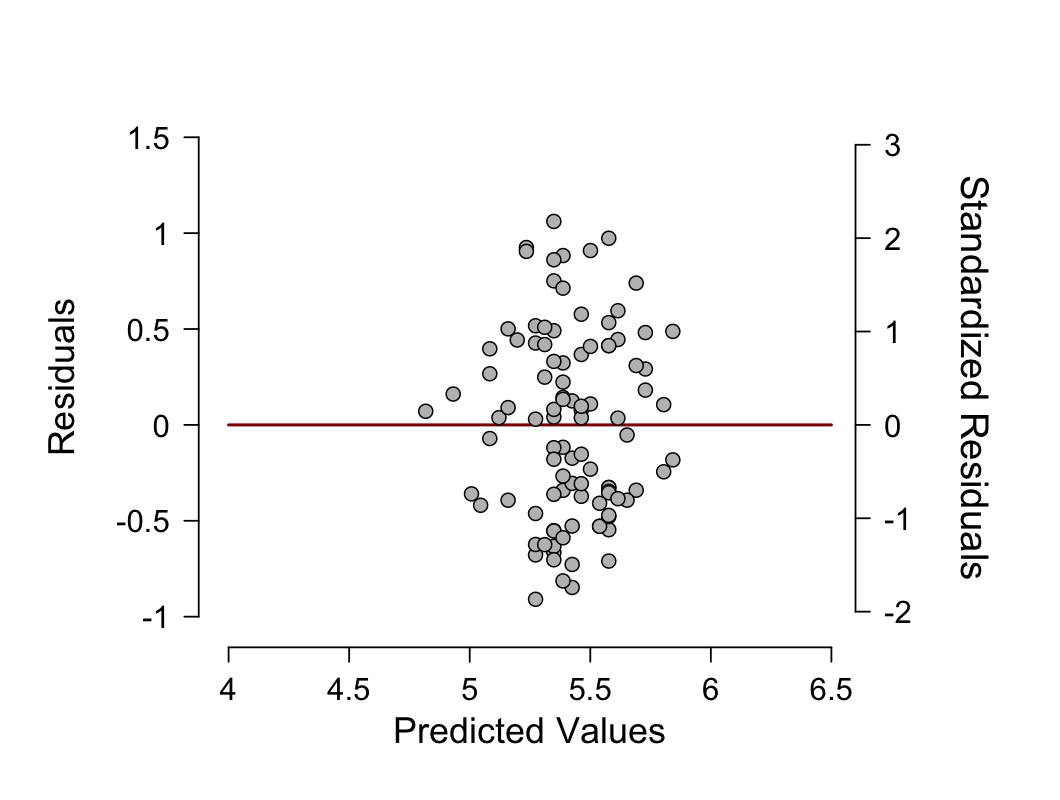
**Do you have homoscedasticity in the residuals?**

An assumption of linear regression is that the variance of the errors is constant across the observations. This assumption means that the variance around the regression line is the same for all values of the predictor variable (X). You can check for this by observing whether the residuals (errors of prediction) are equal across the standardized predicted values.



If there is homoscedasticity, the "Regression Standardized Residuals" scores (y-axis) will remain approximately constantly spread across the "Regression Standardized Predicted Value" (x-axis) scores. If you do not have homoscedasticity, these values will not be evenly spread, but will differ in height (e.g., a funnel shape). You now need to determine how to proceed based on your results:

### Residuals vs. Predicted

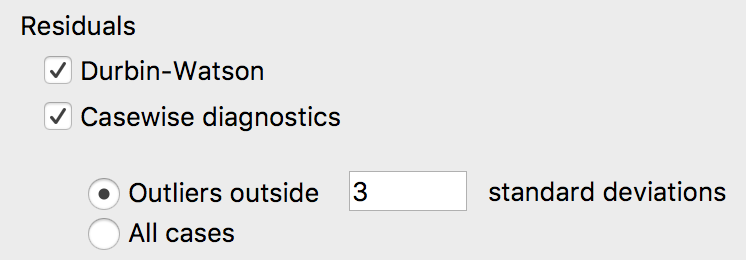


Mainly, you want the dots centered around zero (the line) and not make any triangle or megaphone shapes.

## **The simple linear regression test:**

To finish out the regression analysis, we can check a few more boxes:

Click Statistics  🡪 Click Durbin-Watson, Casewise Diagnostics in the Residuals section.



**Independence of observations**

If you examine the JASP output generated by this test, you will find a table called **Model Summary**, which contains the Durbin-Watson statistic, as highlighted below:

| **Model Summary** | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | | **R** | | **R²** | | **Adjusted R²** | | **RMSE** | | **Durbin-Watson** | |
| 1 |  | 0.371 |  | 0.138 |  | 0.129 |  | 0.489 |  | 1.916 |  |
|  | | | | | | | | | | | |

The Durbin-Watson statistic for our data is 1.957. The Durbin-Watson statistic can range from 0 to 4. You are looking for a value of approximately 2, which indicates that there is no correlation between residuals. You can see that our value is very close to 2 so it can be accepted that there is independence of errors (residuals).

**More on outliers**

Outliers can have a detrimental effect on the regression equation and statistical inferences. An outlier can have a large effect on the variability of residuals leading to problems with normality or homoscedacity, which leads to a reduction in the accuracy of prediction. An outlier can also have a significant effect on the line of best fit (regression line). We can detect outliers visually (as you did earlier) and using casewise diagnostics, which is what you will do here.

The **Casewise Diagnostics** table highlights any cases (i.e., participants, in this example) where that case's standardized residual is greater than ±3 standard deviations. A value of greater than ±3 is a common cut-off criteria used to define whether a particular residual might be representative of an outlier or not.

| **Casewise Diagnostics** | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Case Number** | | **Std. Residual** | | **cholesterol** | | **Predicted Value** | | **Residual** | |
| . |  | . |  | . |  | . |  | . |  |
|  | | | | | | | | | |

In this example, we do not have an outlier, so the table is blank. If the data was an outlier, it might look like:

| **Casewise Diagnostics** | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Case Number** | | **Std. Residual** | | **cholesterol** | | **Predicted Value** | | **Residual** | |
| 10 |  | 4.06 |  | 7.98 |  | 5.79 |  | 2.18 |  |
|  | | | | | | | | | |

In this example, you can see that case number 91 ("**Case Number**" column) has been identified as a potential outlier with a large standardized residual of 4.059 ("**Std. Residual**" column), much greater than the cut-off of 3 standard deviations. The table also informs you that the actual cholesterol concentration value is 7.98 ("**Cholesterol**" column), the predicted cholesterol concentration value is 5.79 ("**Predicted Value**" column) and the difference between these two values is 2.18 ("**Residual**" column). We can use this information, in conjunction with the standardized residual value, to determine whether to remove the outlier or not.

**Understanding the analysis:**

There are two main objectives that you can achieve with the output from a simple linear regression: (1) determine the proportion of the variation in the dependent variable explained by the independent variable; and (2) predict dependent variable values based on new independent variable values. Both of these objectives will be answered in the following sections.

**Determining how well the model fits:**

The **Model Summary** table (shown below) provides the information needed to determine how well the regression model fits the data:

| **Model Summary** | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | | **R** | | **R²** | | **Adjusted R²** | | **RMSE** | | **Durbin-Watson** | |
| 1 |  | 0.371 |  | 0.138 |  | 0.129 |  | 0.489 |  | 1.916 |  |
|  | | | | | | | | | | | |

R is the multiple correlation coefficient ("**R**" column). As there is only one independent variable, R is simply the absolute value of the Pearson correlation between the dependent variable and the independent variable. It simply indicates the strength of the association between the two variables. In this example, R = .371, which indicates a moderate correlation. However, you will not normally have to report this value.

The R2 value represents the proportion of variance in the dependent variable that can be explained by our independent variable (technically it is the proportion of variation accounted for by the regression model above and beyond the mean model). In this example, R2 = .138, which means that the independent variable, time\_tv, explains 13.8% of the variability of the dependent variable, cholesterol. However, R2 is based on the sample and is a positively biased estimate of the proportion of the variance of the dependent variable accounted for by the regression model (i.e., it is too large). JASP also prints out an adjusted R2 value ("**Adjusted R Square**" column), which corrects positive bias to provide a value that would be expected in the population. Adjusted R2 is also an estimate of the effect size, which at 0.129 (12.9%), is indicative of a medium effect size, according to Cohen's (1988) classification.

The **ANOVA** table (shown below) informs you whether the regression model results in a statistically significantly better prediction of the dependent variable, cholesterol, than if you just used the mean value.

| **ANOVA** | | | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | |  | | **Sum of Squares** | | **df** | | **Mean Square** | | **F** | | **p** | |
| 1 |  | Regression |  | 3.745 |  | 1 |  | 3.745 |  | 15.64 |  | < .001 |  |
|  |  | Residual |  | 23.473 |  | 98 |  | 0.240 |  |  |  |  |  |
|  |  | Total |  | 27.218 |  | 99 |  |  |  |  |  |  |  |
|  | | | | | | | | | | | | | |

In this example, the regression model is statistically significant, F(1, 98) = 15.64, p < .001, R2 = .14. The breakdown of the last part is as follows:

|  |  |
| --- | --- |
| **Part** | **Meaning** |
| F | Indicates that we are comparing to an F-distribution (F-test). |
| 1 in (1,98) | Indicates the Regression degrees of freedom ("df") |
| 98 in (1,98) | Indicates the Residual degrees of freedom ("df") |
| 15.64 | Indicates the obtained value of the F-statistic (obtained F-value) |
| p < .001 | Indicates the probability of obtaining the observed F-value if the null hypothesis is correct. |

## 

**Predicting cholesterol concentration:**

The general form of the line to predict cholesterol concentration from time spent watching TV, expressed in SPSS variable form (i.e., cholesterol and time\_tv), is:

cholesterol = b0 + (b1 x time\_tv)

where b0 is the intercept and b1 is the coefficient. By substituting the values for b0 and b1 you will be able to predict cholesterol concentration given time spent watching TV. You can ascertain these value by inspecting the **Coefficients** table:

| **Coefficients** | | | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | |  | | **Unstandardized** | | **Standard Error** | | **Standardized** | | **t** | | **p** | |
| 1 |  | (Intercept) |  | -1.103 |  | 1.651 |  |  |  | -0.668 |  | 0.505 |  |
|  |  | time\_tv |  | 0.038 |  | 0.010 |  | 0.371 |  | 3.954 |  | < .001 |  |
|  | | | | | | | | | | | | | |

Substituting these values into the equation, you have:

cholesterol = -1.10 + (0.038 x time\_tv)

For example, if you wanted to know the mean predicted cholesterol concentration for someone who spent three hours (180 minutes) watching TV per day, then you would calculate this as follows:

predicted cholesterol concentration = -1.10 + (0.038 x 180) = 5.74 mmol/L

For the example we have used throughout this guide, it is doubtful that we would want to use the time spent watching TV as a method to predict cholesterol concentration. More likely, we were trying to understand whether time spent watching TV might be a good surrogate marker for physical inactivity and explain some of the variability in cholesterol concentrations.

## **Reporting All Together:**

A linear regression established that daily time spent watching TV could statistically significantly predict cholesterol concentration, F(1, 98) = 15.64, p < .001, and time spent watching TV accounted for 13.8% of the explained variability in cholesterol concentration. The regression equation was: predicted cholesterol concentration = -1.10 + 0.038 x (time spent watching tv).

## **Example Multiple Linear Regression:**

A health researcher wants to be able to predict maximal aerobic capacity (VO2max), an indicator of fitness and health. Normally, to perform this procedure requires expensive laboratory equipment and necessitates that an individual exercise to their maximum (i.e., until they can longer continue exercising due to physical exhaustion). This can put off those individuals that are not very active/fit and those individuals that might be at higher risk of ill health (e.g., older unfit subjects). For these reasons, it has been desirable to find a way of predicting an individual's VO2max based on more easily and cheaply measured attributes. To this end, the researcher recruits 100 participants to perform a maximum VO2max test, but also records their age, weight, heart rate and gender. Heart rate is the average of the last 5 mins of a 20 mins much easier, lower workload cycling test. The researcher's goal is to be able to predict VO2max based on age, weight, heart rate and gender.

To get started, open the dataset for this example in JASP. Remember, you can always use the previous help guides for greater detail in case you do not remember how to do something.

File 🡪 Open 🡪 Computer 🡪 Browse 🡪 Pick the Correlation Data.

## **Check your assumptions:**

**Is the dependent variable at least scale (ratio or interval)?**

Yes, we are using ratio style data.

**Is there a linear relationship between the variables?**

**Are there any outliers in the sample?**

**Are the residuals normally distributed?**

**Do you have homoscedasticity in the residuals?**

Homoscedasticity occurs when the spread of X is the same all the down Y and vice versa. You do not want X and Y on a scatterplot to be shaped like a megaphone, triangle, or other odd shapes, which would indicate heteroscadasticity (that’s bad).

## **The multiple linear regression test:**

**Effect Size:**

## **Reporting All Together:**