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.).

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)

## Problems solved using a simple linear regression

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 years 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)?

## What you will calculate

1. A linear regression equation.  
2. The statistical significance of β1 (null hypothesis significance testing).  
3. A measure of effect size.  
4. Confidence and prediction intervals.

## Assumptions & order of testing

For a simple linear regression to be a valid test to use (e.g., provide valid predictions), the following two assumptions must hold:

1. A linear relationship between the two variables (or transformed to linearity).  
2. No significant outliers or influential points.

If these two assumptions are not violated, you can make predictions about the value of your dependent variable given your knowledge of the value of an independent variable. If they are violated, you need to make corrections and re-test these assumptions. If they still do not pass, you must find alternative statistical tests.

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

3. Independence of errors (residuals).  
4. Homoscedasticity of residuals (equal error variances).  
5. 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 five 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 as 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 (3) then it is pointless testing assumptions (4) and (5), as the regression analysis will already have been rendered invalid by failure of assumption (3).

## EXAMPLE

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.)

## TESTING OF ASSUMPTIONS

**Establishing if a linear relationship exists**

Linear regression is only appropriate when there is a linear relationship between your two variables.

1. Click **Graphs > Chart Builder...** on the main menu.

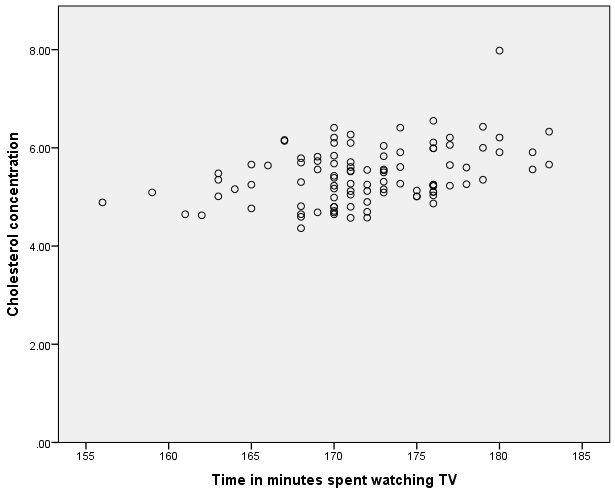
2. Select "Scatter/Dot" from the Choose from: box in the bottom-left-hand corner of the **Chart Builder** dialogue box.

3. Selecting "Scatter/Dot" will present eight different scatter/dot options in the lower-middle section of the **Chart Builder** dialogue box (as shown above and below). Drag-and-drop the top left-hand option (you will see it labeled as "Simple Scatter" if you hover your mouse over the box) into the main chart preview pane.

4. You will be presented with the screen below, which shows a simple scatterplot in the main chart preview pane with boxes for the y-axis ("Y-Axis?") and x-axis ("X-Axis?") for you to populate with the appropriate variables.

5. Drag-and-drop time\_tv from the Variables: box into the "X-axis?" box in the main chart preview screen and do the same for cholesterol but into the "Y-axis?" box.

6. Click on "Y-Axis1 (Point1)" in the Edit Properties of: box located in the **Element Properties** dialogue box (the box on the right-hand side of the main **Chart Builder** dialogue box).

7. Uncheck the Minimum option in the -Scale Range- area so that the Custom box is highlighted and has a value of 0 (zero).

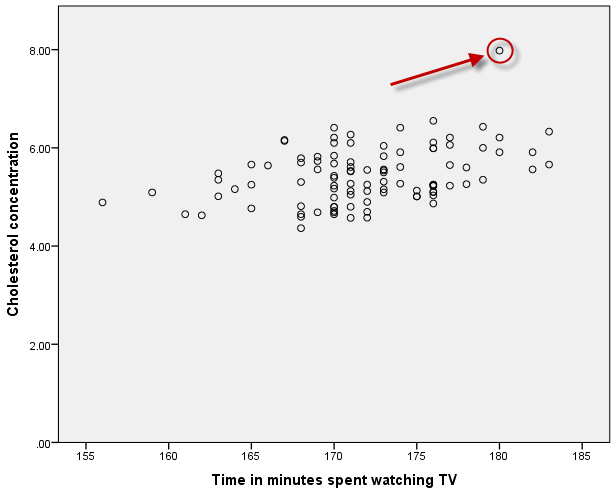
8. Click the https://statistics.laerd.com/premium/pc/img/apply-button.pngbutton to confirm these changes.

## Interpretation of linearity

If you have been following this example, you will have generated the following scatterplot:

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.

## Testing for outliers

When conducting a Pearson's correlation analysis, 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.

The data plotted in this particular scatterplot would appear to show a linear relationship between the two variables, although there appears to be an outlier (shown to the left).

Your first consideration should be to check whether you have made any data entry errors (i.e., simply keyed in any wrong values into SPSS). 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.

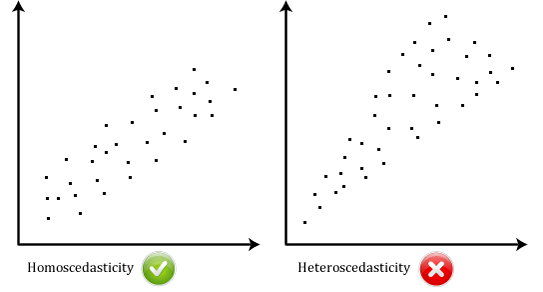
**Keep the outlier(s):** If you do not want to remove an outlier, you have three choices: either (1) modify the outlier by replacing the outlier's value with one that is less extreme; (2) transform the dependent variable; or (3) include in the analysis anyway but remember to highlight the outlier in 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.

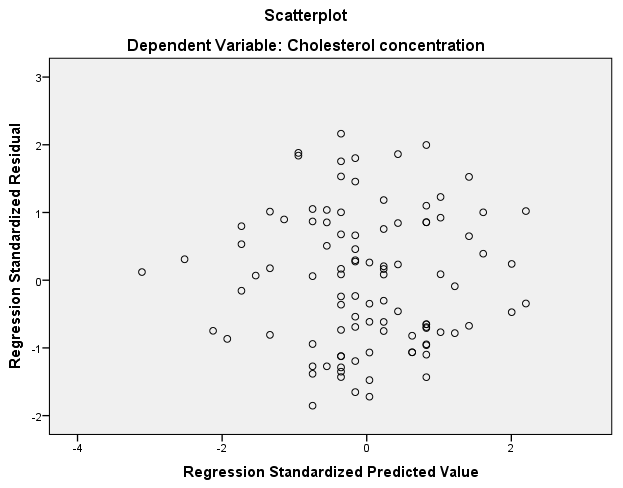
One more question you can ask yourself when you have an outlier, is whether the participant's values highlight that their inclusion in the study should be reconsidered. For example, the outlier in this example has a cholesterol concentration of 7.98 mmol/L, which is a very high concentration, indicating 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.

## Testing for heteroscedasticity

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.



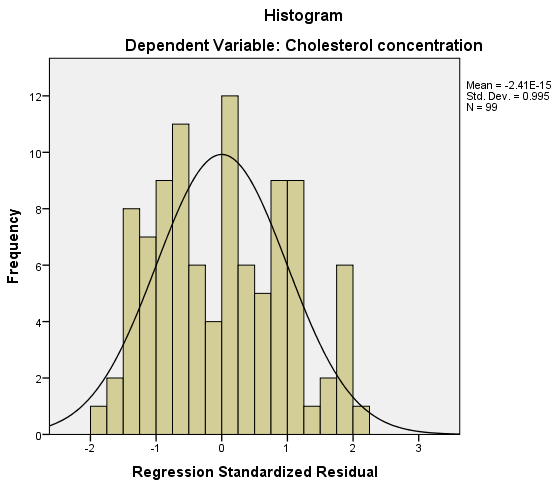
If there is homoscedasticity, the "Regression Standardized Residuals" scores (y-axis) will remain approximately constantly spread across the "Regression Standardized Predicted Value" (x-axis) scores. If you do not have homoscedasticity, these values will not be evenly spread, but will differ in height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



## Checking for normality of residuals (errors)

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 P-P Plot, as described below:

### Histogram

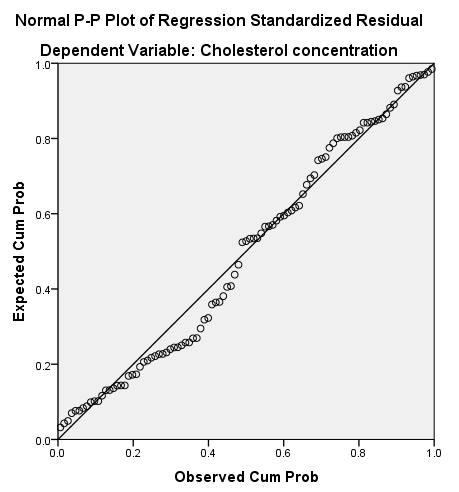
The following histogram (with superimposed normal curve) is produced automatically from the options you selected in the **Linear Regression: Plots** dialogue box:

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).

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

### Normal P-P Plot

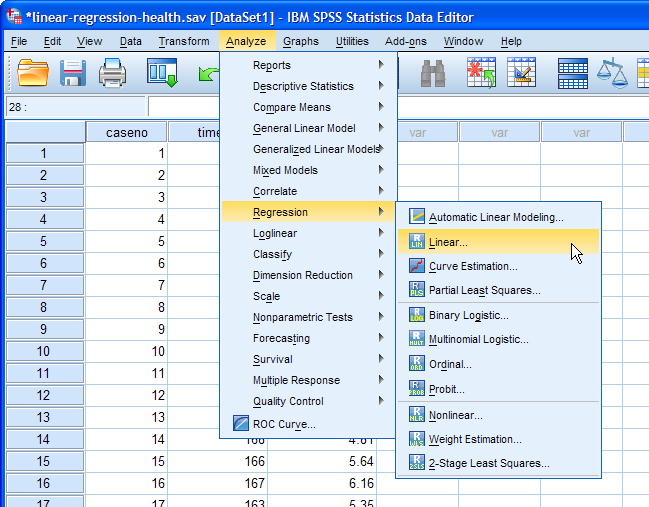
The following Normal P-P Plot is produced automatically from the options you selected in the **Linear Regression: Plots** dialogue box:

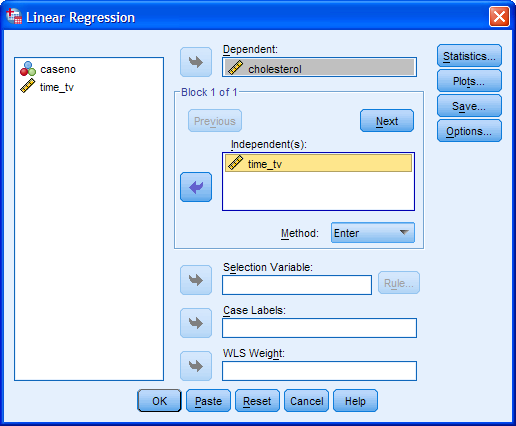


If the residuals are normally distributed, the points will be aligned along the diagonal line. In reality, the points will never be perfectly aligned along the diagonal line. Moreover, you only need the residuals to be approximately normally distributed as the regression analysis is fairly robust to deviations from normality. You can see from the above Normal P-P 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.

**Regression Procedure**

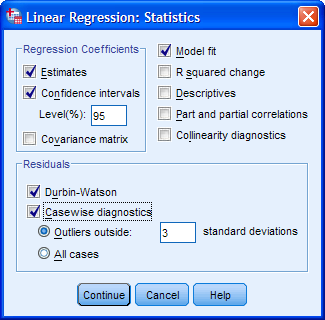
1. Click **Analyze > Regression > Linear...** on the main menu, as shown below:



2. Transfer the dependent variable, cholesterol, into the Dependent: box and the independent variable, time\_tv, into the Independent(s): box, using the https://statistics.laerd.com/premium/lr/img/right-arrow-button.pngbuttons, as shown below. All other boxes can be ignored.

3. Click the https://statistics.laerd.com/premium/lr/img/statistics-button.pngbutton. You will be presented with the **Linear Regression: Statistics** dialogue box.

4. In addition to the options that are already selected, check Confidence intervals from the -Regression Coefficients- area and leave the Level(%): at 95 and check Casewise diagnostics from the -Residuals- area and leave the option value at 3 standard deviations. Also check Durbin-Watson from the -Residuals- area. You will end up with the following screen:

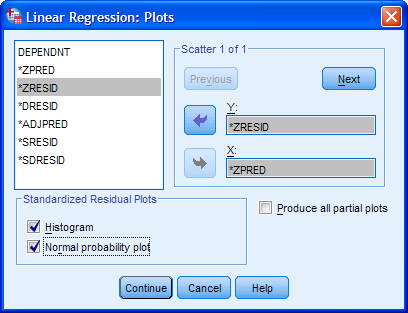


5. Click the https://statistics.laerd.com/premium/lr/img/continue-button.pngbutton. You will be returned to the **Linear Regression** dialogue box.

6. Click the https://statistics.laerd.com/premium/lr/img/plots-button.pngbutton and you will be presented with the **Linear Regression: Plots** dialogue box.

7. Transfer "**\*ZRESID**" into the Y: box and "**\*ZPRED**" into the X: box, using the appropriate https://statistics.laerd.com/premium/lr/img/right-arrow-button.pngbuttons.

8. Select Histogram and Normal probability plot from the -Standardized Residual Plots- area, as shown below:

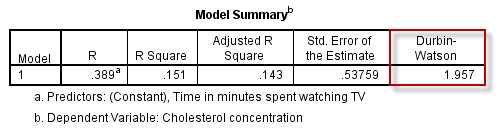


9. Click the https://statistics.laerd.com/premium/lr/img/continue-button.pngbutton. You will be returned to the **Linear Regression** dialogue box.

10. Click the https://statistics.laerd.com/premium/lr/img/ok-button.pngbutton. This will generate the output.

**Independence of observations**

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



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).

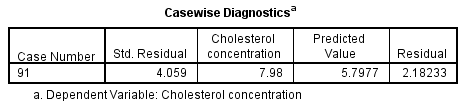
## DEALING WITH UNUSUAL POINTS

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.

### Casewise diagnostics

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, which you have instructed SPSS to treat as an outlier. 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. You will have a table called **Casewise Diagnostics** that contains the relevant information, as shown below:

Note: If all your cases have standardized residuals less than ±3, this table will not be produced as part of the SPSS output. If you successfully remove an outlier, the **Casewise Diagnostics** table produced the first time you run the analysis will probably not then be produced the second time around.



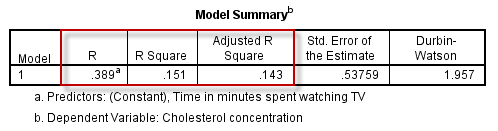
## In this example, you can see that case number 91 ("****Case Number****" column) has been identified as a potential outlier with a large standardized residual of 4.059 ("****Std. Residual****" column), much greater than the cut-off of 3 standard deviations. The table also informs you that the actual cholesterol concentration value is 7.98 ("****Cholesterol concentration****" column), the predicted cholesterol concentration value is 5.7977 ("****Predicted Value****" column) and the difference between these two values is 2.18233 ("****Residual****" column). We can use this information, in conjunction with the standardized residual value, to determine whether to remove the outlier or not. In this example, the outlier will be removed.

## INTERPRETING AND REPORTING THE OUTPUT

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

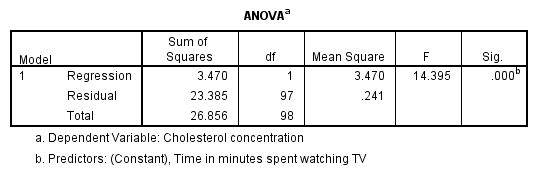
The **Model Summary** table (shown below) provides the information needed to determine how well the regression model fits the data:



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 = 0.389, which indicates a moderate correlation. However, you will not normally have to report this value.

The R2 value ("**R Square**" column) 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 = 0.151, which means that the independent variable, time\_tv, explains 15.1% 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). SPSS 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 0.143 (14.3%), 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.



In this example, the regression model is statistically significant, F(1, 97) = 14.395, p < .001. The breakdown of the last part is as follows:

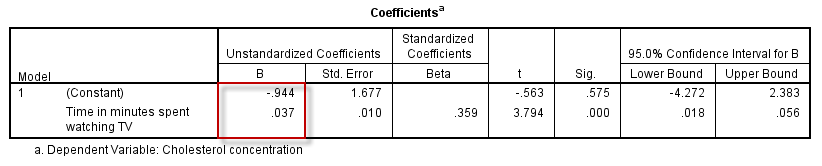
|  |  |
| --- | --- |
| **Part** | **Meaning** |
| F | Indicates that we are comparing to an F-distribution (F-test). |
| 1 in (1,97) | Indicates the Regression degrees of freedom ("df") |
| 97 in (1,97) | Indicates the Residual degrees of freedom ("df") |
| 14.395 | Indicates the obtained value of the F-statistic (obtained F-value) |
| p < .0005 | 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 (highlighted in red below):



Substituting these values into the equation, you have:

cholesterol = -0.944 + (0.037 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 = -0.944 + (0.037 x 180) = 5.72 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.

## Graphing the output

An appropriate graph you can use to present the results of a simple linear regression is a scatterplot.

**Reporting Results in APA:** A linear regression established that daily time spent watching TV could statistically significantly predict cholesterol concentration, F(1, 97) = 14.395, p < .0005 and time spent watching TV accounted for 14.3% of the explained variability in cholesterol concentration. The regression equation was: predicted cholesterol concentration = -0.944 + 0.037 x (time spent watching tv).