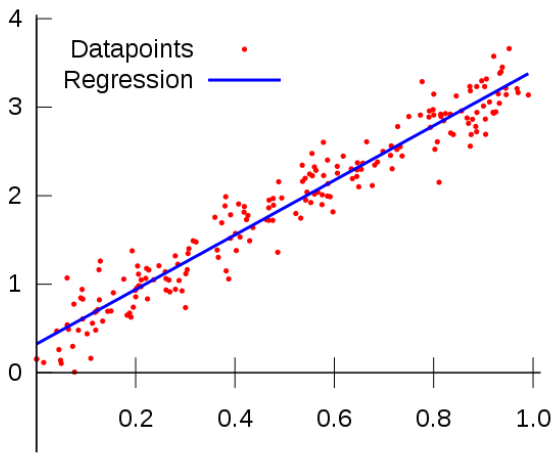
Critical Reasoning 22 - Simple Linear Regression



A scatter plot of data points with a liner regression line superimposed

In Critical Reasoning 19 we looked at some methods for quantifying the relationship between variables, if any, in a data set. We also used that information to test hypotheses concerning these relationships. In this study unit we introduce the method of **linear regression** by which to model the relationship between a dependent variable y, used to represent individual fixed-size data objects, and one or more independent variables, x. The word 'linear' here is used in its adjectival sense to refer to the property of resembling a straight line. Note that the kinds or relationship we are interested in here are statistical rather deterministic. Examples of **deterministic relationships** are the area of a circle and its radius ($A = \pi r^2$) or energy-mass equivalence ($E = mc^2$), where the equation in question *exactly* describes the relationship between the two variables. **Statistical relationships**, on the other hand, simply require one variable's values to increase or decrease when the other variable's values changes, a relationship that is almost never exact or prefect.

Unfortunately there is no consistency in the way that various authors refer to the variables involved. The independent variable(s) may be variously called explanatory variables, exogenous variables, predictor variables, or regressors. Meanwhile the dependent variable may be variously called an outcome variable, a criterion variable, an endogenous variable, or a regressand.

Simple linear regression, to which we shall confine ourselves here, considers only one independent variable at a time. According to the website Statistics Solutions the aim of linear regression is to answer two questions:

1. Does a set of independent (predictor) variables do a good job in predicting a dependent (outcome) variable? and

2. Which variables in particular are significant predictors of the dependent (outcome) variable, and in what way do they… impact the dependent (outcome) variable?

In its simplest form the regression equation for one dependent and one independent variable is given by

$$y = bx + c$$

where y is the estimated dependent variable score, b is the regression coefficient, c is a constant and x is the score of the independent variable. No doubt you will recognize this equation from High School as that of a strait line. If you have been following this series of study units you will also be familiar with scatter plots and the somewhat haphazard method of drawing a line of best fit that may or may not suggest a relationship among the data. Linear regression takes the guesswork out of

this task by providing a statistical method for fitting a single line, given by the above equation, through a scatter plot. However linear regression is not simply about drawing a line through a cloud of points; it is much more powerful. According to Statistics Solutions there are three major uses for Regression Analysis:

> **causal analysis**, which is used to identify the strength of the effect that an independent variable(s) may have on a dependent variable. Typical questions here include "What is the strength of relationship between dose and effect, sales and marketing spending, age and income".
>
> **forecasting an effect** or impact(s) of change, which allows us to understanding how much the dependent variable will change when we change one or more independent variables. A typical question here is, "How much additional Y do I get for one additional unit of X?"
>
> **trend forecasting** which uses regression analysis to predict trends and future values or point estimates. A typical question here might be "What will the price of gold be in 6 months?"

Assumptions of Linear Regression

Like every statistical procedure, linear regression relies on certain assumptions. Although not an assumption *per se*, there is a "rule of thumb" about sample size that recommends that there should be at least 20 cases per independent variable in the analysis. The assumptions are as follows:

> **A linear relationship**: There should be a linear relationship between the independent and dependent variables. The easiest way to assess this visually is to construct a scatter plot. Obvious outliers should be excluded as linear regression is sensitive to outlier effects. (See Critical Reasoning 19) Where some data are not linear but approximate another known function there is sometimes a way of *making them linear* by, for example, plotting them logarithmically. There are other methods; however we shall not consider them here.
>
> **Multivariate normality**: The distribution of two or more variables should be multivariate normal, *i.e.* every linear combination of variables should also have a univariate normal (Gaussian) distribution. This can most easily be assessed visually by constructing histograms. (See Critical Reasoning 13.) However, there are other statistical techniques for checking normality by using a goodness-of-fit test such as the non-parametric **Kolmogorov-Smirnov test**, used to determine whether two distributions differ significantly.
>
> **No or little multicollinearity**: Multicollinearity occurs when the independent variables are too tightly correlated with one another. There a several techniques for checking for multicollinearity, however the one we already know of is to compute a correlation matrix using Pearson's r. The value of the calculated correlation coefficients should be less than 1 but also not too close to 1. (See Critical Reasoning 19)
>
> **No autocorrelation**: Autocorrelation among data occurs when there is correlation between the values of the same variables based on related objects. This typically occurs when sampling data from the same source instead of it being randomly selected. Autocorrelation shows up when, for example, the value of $y(x+1)$ is not independent from that of $y(x)$.

This can often be visualized from scatterplots in which adjacent data points have the same value or which display various kinds of cyclical patterns. Of course, adjacent data points can sometimes have the same value purely by chance; however if this happens repeatedly it is likely that the data were sampled from the same source. The **Durbin Watson test** is one popular test designed to detect the presence of autocorrelation; however we shall not consider it here, except to caution that it has its own set of assumptions that have to be met.

Homoscedasticity: *i.e.* the data should have the same finite variance. This criterion was discussed in Critical Reasoning 19. Recall that graphically, if a scatter plot tends to "splay out" in either direction then the data are almost certainly heteroscedastic. The **Goldfeld-Quandt** test for homoscedasticity divides the data set into two parts or groups and compares the variances of each group to see if they are similar. However we shall not consider that test here. A glance at the scatterplot will be adequate to our purposes.

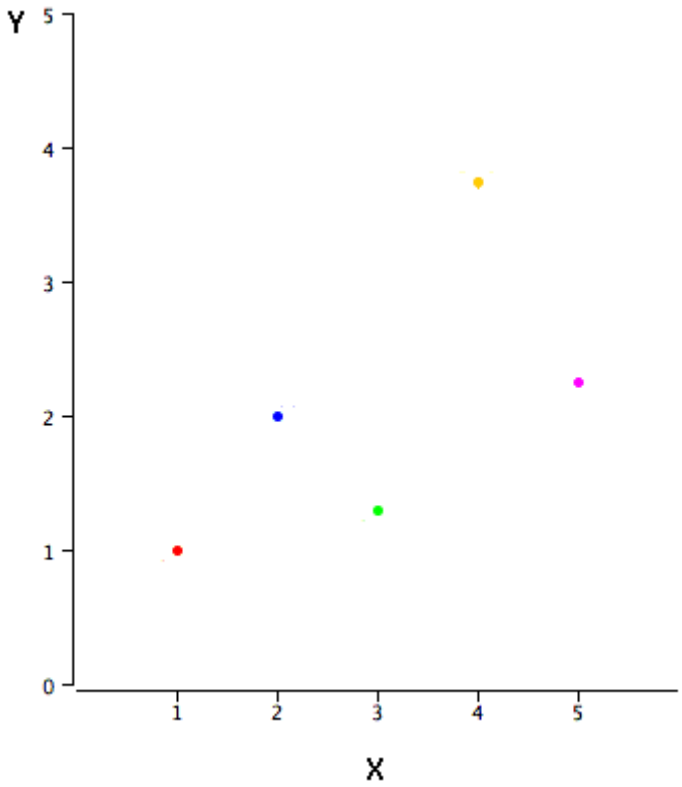
Conducting a Linear Regression

According to Statistics Solutions, Linear Regression Analysis consists of 3 stages: (1) analyzing the correlation and directionality of the data, (2) estimating the model, *i.e.*, fitting the line, and (3) evaluating the validity and usefulness of the model. The example we shall be working through below was found at onlinestatbook.com. All diagrams in this example are reproduced from that website.

E.g. 1 Suppose you are given the data set tabulated below. We don't know what X and Y stand for – they are just for the sake of example. We can see that there appears to be a positive relation between X and Y. If we were to predict the value of Y from its corresponding X, the higher the value of X the higher our prediction of Y would be. Note this example does not comply with our rule of thumb about sample size that recommends that there should be at least 20 cases per independent variable in the analysis; however we can set that aside for now as this is just an introductory example.

X	Y
1.00	1.00
2.00	2.00
3.00	1.30
4.00	3.75
5.00	2.25

The next task is to create a scatter plot of the data so that we can assess it visually. See over page. The positive relation between X and Y is clear to see. It also looks like the relation is a more or less linear one. Although it appears that the points "flare out" somewhat diagonally upwards, there are just too few data points to visually assess whether we are dealing with a heteroscedastic data set. Rather than spending too much time deciding whether every assumption of linear regression has been met, we shall assume that the authors, in the interests of clarity, chose their example in such a way that they were.



Scatter plot of the example data

The next step is to calculate the formula of the regression line. This can easily be computed using statistical software; however we shall calculate it using a spreadsheet program so that we can understand just what the more powerful statistical software is programmed to do. Note that we do not know whether the data represent a population or a sample. However there are so few of them that we shall assume that they represent only a sample. Our calculations should therefore be based on statistics rather than parameters. The following statistical information is required to derive the formula for the regression line: M_X and M_Y being the means of X and Y respectively; s_X and s_Y being the standard deviations of X and Y respectively; and r being Pearson's correlation coefficient between X and Y. The formulae for these statistics should be familiar to you by now, except for one small difference to the formula for the standard deviation of a sample.

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - M)^2}$$

Note that in the case of samples we divide through by n - 1 rather than just n in the case of populations. We also use the sample mean M rather than population mean μ. The following table lists the statistics required for computing the regression line:

M_x	M_Y	s_x	s_y	r
3.00	2.06	1.581	1.072	0.627

Recall that the general formula for a straight line is $y = bx + c$. The slope or gradient b and the y intercept c, together determine a unique straight line. The slope of a linear regression line can be calculated as follows:

$$b = r\frac{s_{\mathrm{Y}}}{s_{\mathrm{X}}}$$

and the intercept as follows:

$$c = M_Y - b\mathrm{M}_{\mathrm{X}}$$

Note that the symbols used differ widely for one text to another; however we have adhered to our existing conventions. All that remains to determine the formula for the regression line is to substitute the tabulated values into the formulae above, thus:

$$b = 0.627\left(\frac{1.072}{1.581}\right) = 0.425$$

and

$$c = 2.06 - (0.425)(3.00) = 0.785$$

Therefore the formula for the regression line is

$$y = 0.425x + 0.785$$

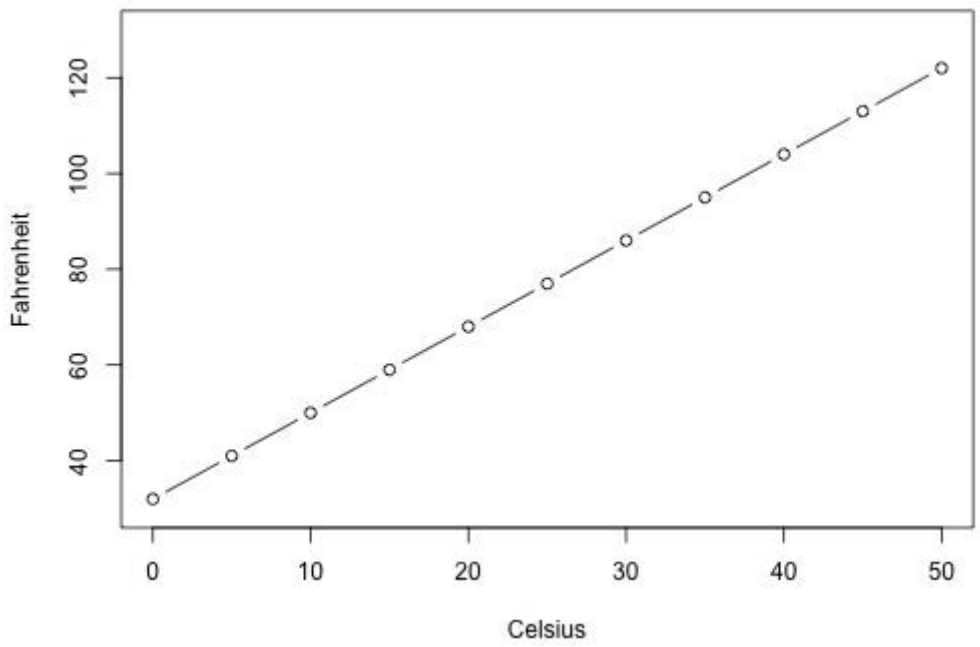
We can see that this line meets with our expectations. The sign of the gradient is positive, so we know that it slopes upwards towards the right. We can also see from the value of the gradient that it is not a very steep line. A value of 1 represents a line at 45° to the horizontal but our line is not so steep, although the somewhat elongated y-axis of the scatter plot above makes this less apparent. Also we can see from the small, positive valve of the y-intercept that it cuts the y-axis above the x-axis fairly close to the origin.

The following two real world examples, including diagrams, were found at Penn State Eberly College of Science's STAT 501 website.

E.g. 2 Assess the strength of the relationship between n = 11 temperatures as given in degrees Celsius *vs*. degrees Fahrenheit. Below, over page, is a scatter plot of the 11 data points together with a 'regression line' passing through them. Pearson's correlation coefficient of degrees Celsius and degrees Fahrenheit was calculated to be r = 1.000. Therefore there is a perfect liner relationship between degrees Celsius and degrees Fahrenheit. In fact we know this relationship to be

$$°\mathrm{F} = 1.8 \times °\mathrm{C} + 32$$

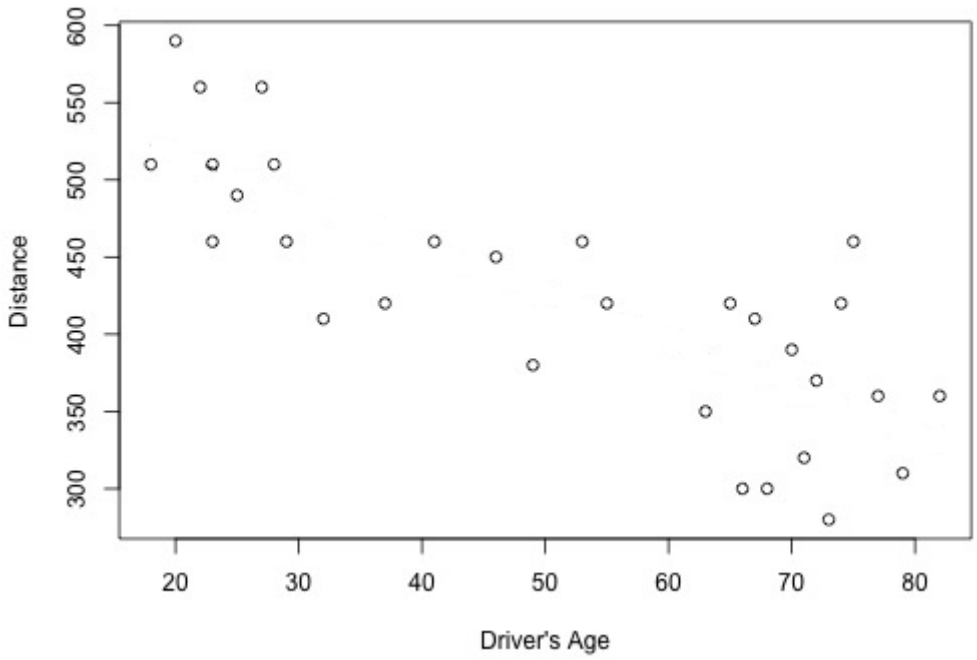
Therefore 100% of the variation in temperatures in Fahrenheit is explained by the temperature in Celsius. Clearly what we are dealing with here is a deterministic relationship between the variables; therefore a statistical approach to this problem would have been uncalled for.



E.g. 3 Asses the strength of the linear relationship between the age of a driver and the distance that that driver can see to the next car. We know that, without the aid of spectacles, the distance that a person can see to a distant object decreases with age; therefore we could hazard a guess that the linear relationship in this example should be negative. Here are the raw data for n = 30 individuals tested, with age given in years and distance in feet. (Converting the distances to meters will not make any difference to this example.)

Age	Distance	Age	Distance	Age	Distance
18	510	37	420	68	300
20	590	41	460	70	390
22	560	46	450	71	320
23	510	49	380	72	370
23	460	53	460	73	280
25	490	55	420	74	420
27	560	63	350	75	460
28	510	65	420	77	360
29	460	66	300	79	310
32	410	67	410	82	360

Our first task is to create a scatter plot. See over page. As we can see, it appears that the data points sweep from the top left corner of the graph down, fairly gently, towards the bottom right. This is consistent with our guess that the linear relation in this example should be negative.



Once we have entered our data into our spreadsheet program as two side-by- side columns, we select the various statistical calculation operations required for performing a linear regression, namely: the arithmetic means and standard deviations (for samples) for both the age and distance columns, followed by Pearson's correlation coefficient r and the square of that, r^2. Next we tabulate our results as below.

M Age	*M* Dist	Std Dev A	Std Dev Dist	r	r^2
51	423.333	21.776	81.720	-0.801	0.642

To determine the equation of our regression line $y = bx + c$, we need to find the gradient b and the y-intercept, c. As in the first example, they are given by:

$$b = r\frac{s_{\mathrm{Y}}}{s_{\mathrm{X}}}$$

and by

$$c = M_Y - b\mathrm{M_X}$$

The rest is simply a matter of substitution, thus

$$b = -0.801\frac{81.720}{21.776} = -3.006$$

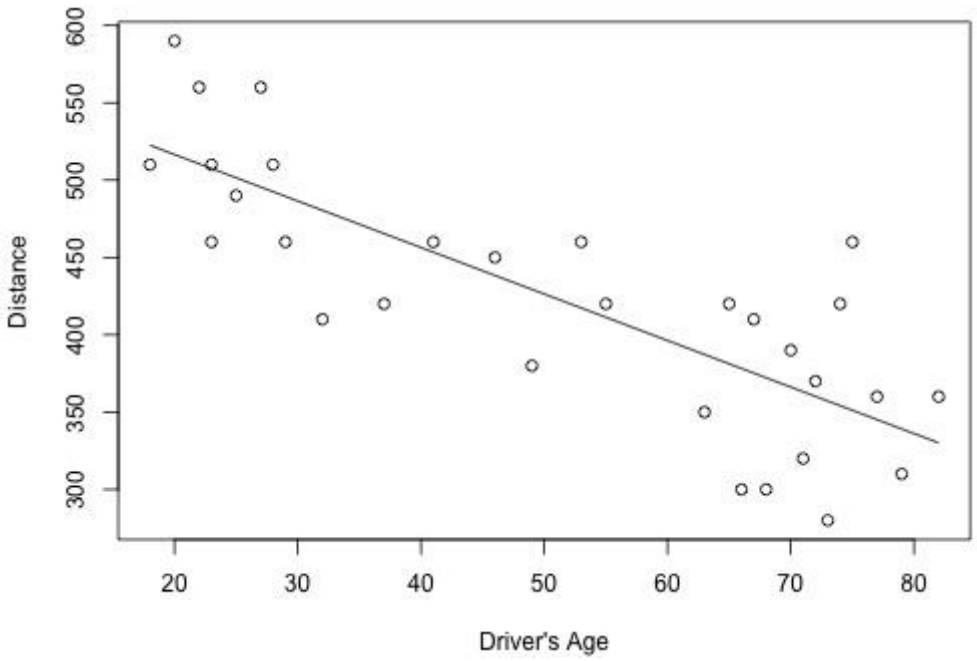
and

$$c = 423.333 - (-3.006)(51) = 576.639$$

Therefore the equation of our regression line is

$$y = -3.006x + 576.639$$

which we can draw in over our scatter plot, as follows:



Let us try to make sense of what this line tells us. Firstly, its gradient is negative, as expected, but rather steep. A gradient of -1 would have been 45° to the horizontal, while a gradient of -3 would lie at a much steeper 72°. The reason for the apparently flatter slope in the plot above is due to the fact that the axes have not been drawn to the same scale. Both axes have been truncated or shortened and the scale of the vertical axis has been considerably compressed relative to that of the horizontal axis. Also, we cannot extrapolate and make literal sense of the y-intercept above, for that would represent a person of age 0 seeing 576.639 feet to the next car, which is clearly nonsense – newborn human babies cannot focus their eyes. Therefore, our relation cannot be totally linear across the entire interval. One way of showing this is not to extend our regression line beyond the first and last data points as in the diagram above. The horizontal axis has been truncated so that the age of a driver is not represented below the legal age for a driver's license, which makes sense.

Note that the absolute value of r of 0.801 which is close to 1, tells us that the linear relationship of obtained is fairly strong but not perfect. If we look at the value of r^2 = 0.642 and multiply it by 100, then this value tells us that 64.2% of the variation in seeing distance to the next car is reduced by taking the age of the driver into account.

r^2 Precautions

Unfortunately the **correlation coefficient r**, which is just Pearson's r, and the **coefficient of determination r^2**, which is a measure of how well differences in one variable can be explained by a difference in a second variable in a statistical model, are frequently misused and misunderstood. Fortunately, the website from which the above example was taken lists seven precautions, including several practice problems, which should be consulted so as not to make the most common mistakes. That page is indexed separately at Penn State Eberly College of Science's STAT501 website. Please consult it. We cannot reproduce it here *verbatim* for copyright reasons.

Linear Extrapolation

Once we have obtained our regression line it is a simple matter to substitute the value of one variable not included in the data set in order predict the other variable. In the example above, no one of age 80 was included in the data set, yet we can substitute that x-value into our linear equation to predict how far we would expect a driver of age 80 to be able to see to the next car based on our sample data. Thus

$$y = -3.006(80) + 576.639 \approx 336 \text{ feet}$$

We can also use the equation of our linear regression line to extrapolate values beyond the interval represented in our dataset so long as the unknown point is not too far off and we have strong reason to believe that the relation is linear throughout, at least up to that point. Thus, if we wanted to estimate the seeing distance (to the next car) of an 85 year old driver based on the existing sample, it would again be a simple matter of substitution.

$$y = -3.006(85) + 576.639 \approx 321 \text{ feet}$$

However using our linear regression line to extrapolate the seeing distance (to the next car) of a centenarian (someone 100 years of age or more) would be ill advised because we cannot assume that the relation would continue to be linear to such an old age - likely it would fall off quite steeply into advanced old age.

Conducting an Hypothesis Test for a Regression Slope

In Critical Reasoning 15 we introduced statistical hypothesis testing as a formal set of procedures by which to decide whether to accept or reject a statistical hypothesis. If we have already calculated a regression line of the form $y = bx + c$ that represents a line of best fit between an independent variable x and a dependent variable y, we may wish to test the significance of the linear relation by focusing on the slope b of the regression line. If the slope is zero, or extremely close to zero, then we can conclude that that there is no linear relation between the independent variable and the dependent variable. If however the slope of the regression line is *significantly* different to zero we are justified in concluding that there is a significant relationship between the independent and dependent variables. The following discussion and example is drawn from the website Stat Trek, Teach yourself statistics: Hypothesis Test for Regression Slope

The test requirements are the same as those for simple linear regression described under the heading "Assumptions of Linear Regression" above. The first step is to state the hypotheses. If there is a significant linear relationship between the independent variable and the dependent variable, the slope will not be zero. Thus:

H_0: $b = 0$

H_1: $b \neq 0$

Here H_0 is the null hypothesis which states that the slope of the regression line is zero while H_1 is the alternative hypothesis which states that it is not zero. Before we begin our test we need to select a level of significance α of say, 0.01, 0.05 or 0.10. The appropriate test here is the linear regression t-

test with n - 2 degrees of freedom for one independent and one dependent variable. The test statistic is given by the following equation:

$$t = {}^{b}/_{SE_b}$$

where b is the slope of the sample regression line and SE_b is the standard error of the slope given by the formula:

$$SE_b = \frac{\sqrt{\frac{1}{n-2}\sum(y_i - \hat{y}_i)^2}}{\sqrt{\sum(x_i - \bar{x})^2}}$$

where:

y_i is the i^{th} observed value of the dependent variable

$\hat{y}_i$ is the i^{th} estimated value of the dependent variable

x_i is the i^{th} observed value of the independent variable

$\bar{x}$ is the mean of the independent variable and n is the number of observations.

Obviously this formula is rather cumbersome; therefore we recommend using a statistical package to calculate the valve of t directly and thence the associated p-value from the t distribution for n - 2 degrees of freedom. Lastly, we compare the calculated p-value with our pre-chosen level of significance α and typically reject the null hypothesis if $p < \alpha$.

E.g. 4 A power company randomly selects 101 of its customers as part of a survey. For each participant of the survey it collects the following information: Annual electricity bill in dollars and the size of the home in square feet. The company wants to know if there is a significant relation at the level of α = 0.05 between the size of a household and its annual electricity bill. Following the procedure in the previous section it determines that the regression equation is:

$$Annual\ bill = 0.55(home\ size) + 15$$

Running a regression analysis using a statistical package produces the following output:

Predictor	Coefficient	SE of Coefficient	t	p
Constant	15	3	5.0	0.00
Home size	0.55	0.24	2.29	0.01

The first step in answering this question is to state the hypotheses:

H_0: The slope of the regression line is zero

H_1: The slope of the regression line is *not* zero

If there is a statistically significant relation between home size and annual electricity bill the slope will not be zero.

Next we use the data provided to conduct a linear regression *t*-test to determine whether the slope of the regression line differs significantly from zero at the level of a = 0.05. Fortunately we have already been provided with the slope (0.55) and the standard error of the slope (0.24). We also know that there are n - 2 degrees of freedom involved which, in this case, is 101 - 2 = 99 *df*. Substituting into the formula above we calculate the *t*-test statistic as:

$$t = {}^{b}/_{SE_b} = \frac{0.55}{0.24} = 2.29$$

Using a spreadsheet or a reliable online *t*-value calculator we establish the associated p-value. Note that our alternative hypothesis is non-directional therefore we must take the "two tailed" p-value calculated or, if not, we must add the associated p-values for both tails, *i.e.*

$$p(t > 2.29) + p(t < 2.29) = 0.0121 \ + \ 0.0121 \ = 0.0242$$

Because the associated p-value of 0.0242 is less than the chosen level of significance of 0.05 we decide to reject the null hypothesis (H_0) and fall back on the alternative hypothesis (H_1). Therefore we are justified in concluding that there is a significant relationship between home size and its annual electricity bill.

Task

Think about two quantities that are easy to measure that you think may have a linear relation. Select a limited random sample of objects or subjects, measure these quantities and from your data use the methods of simple linear regression described in this study unit to determine if indeed there may be such a relation and whether it is significant. (Avoid choosing quantities between which there is a known perfect correlation such as current through and voltage across a resistor or between degrees Fahrenheit and degrees Celsius.) If you wish you may use a statistical package to do the number crunching; however you will have to know which statistics need to be calculated and how to interpret the output.

Feedback

Assuming that you do not have access to laboratory state-of-the-art measuring equipment, we suggest that you use those devices present in most homes *e.g.* a tape measure, a bathroom or kitchen scale, a measuring jug, a home thermometer, a stopwatch or kitchen timer, and so on. For simplicity we chose to measure the height of 15 adult team members using a measuring tape (cm) and their corresponding weight using a bathroom scale (kg).The subjects were all measured in one session and the scale was reset to zero after each measurement. The results are tabulated below.

We designated height as the independent variable and weight as the dependent variable, although, for this task, we could have done so the other way round. Next we state our hypotheses as above:

H_0: $b = 0$

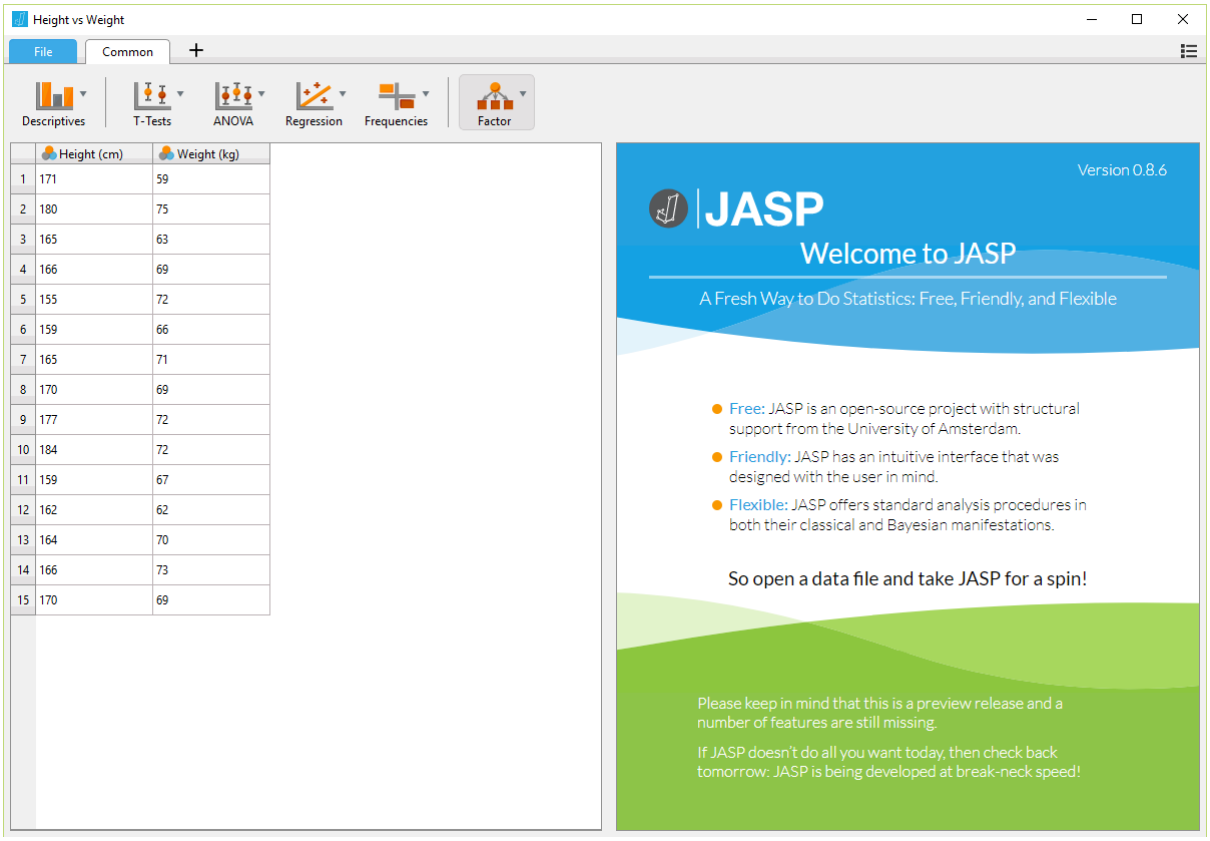
H_1: $b \neq 0$

where b is the slope or coefficient of out linear regression line $y = bx + c$. So as not to cheat, we decide, in advance, to set our level of significance at $\alpha = 0.10$

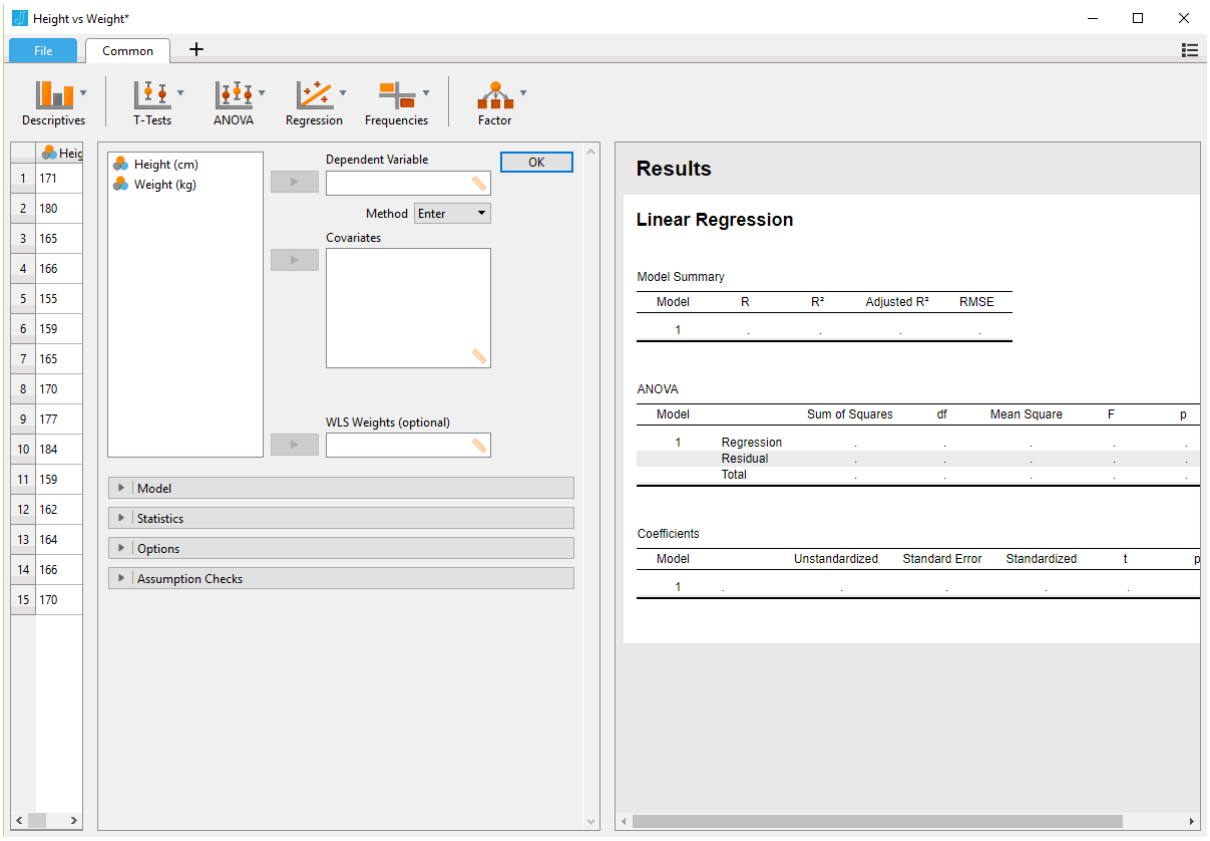
Height (cm)	Weight (kg)
171	59
180	75
165	63
166	69
155	72
159	66
165	71
170	69
177	72
184	72
159	67
162	62
164	70
166	73
170	69

Rather than doing the calculations by hand, we chose to use a statistical package to analyze our data. Unfortunately these programs are very expensive; JASP however is a free open-source graphical program for statistical analysis, provided by the University of Amsterdam. In our opinion JASP is the best and most user friendly statistical package in its class. It has an intuitive graphical interface and is constantly evolving, with new functions being added with each update. If you wish to download this program and install it onto your computer you will be able to follow how we used JASP's inbuilt linear regression function to analyze this data set.

Firstly, you will have to enter the data, exactly as shown in a spreadsheet program such as Excel™ and save it as a .csv file (*e.g.* Table.csv) using the "Save As" option. Once you have saved your data as a .csv file, close the spreadsheet program, open JASP and click the "File" tab. Select "Open" and locate your saved .csv file under the button "Computer" at right. Now open it. You will be presented with the following screen:

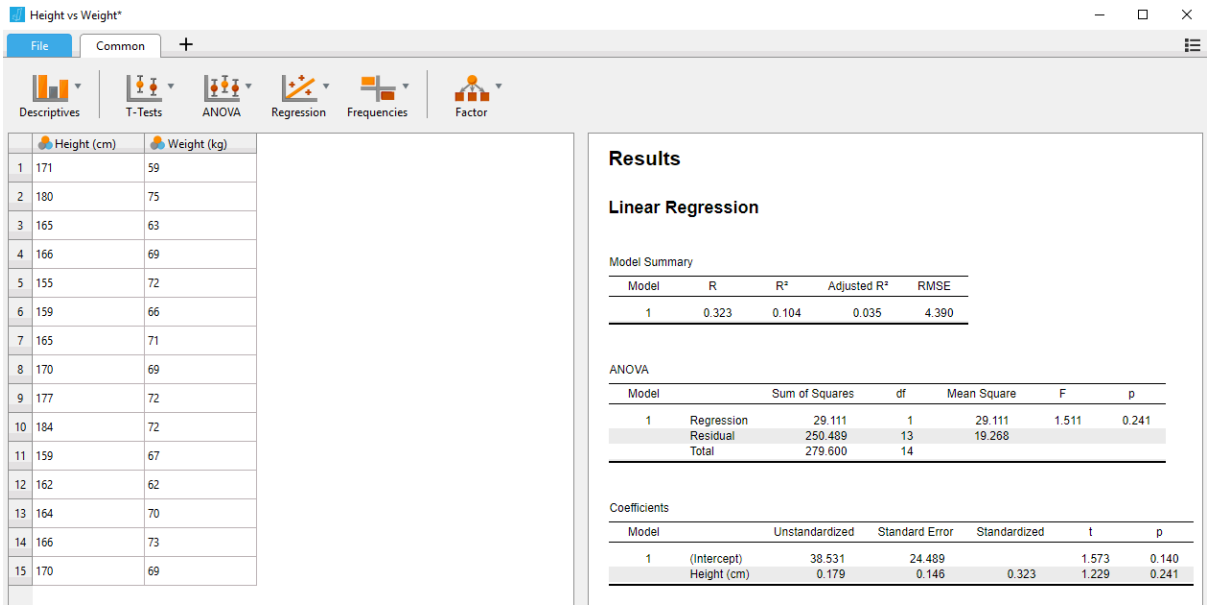


As you can see, the “Common” tab is open and our data are displayed in two columns at left. Next we must decide which tests to perform. Therefore we click the “Regression” button and select “Linear regression” from the pull-down menu. The following screen appears:



Here we need to enter the dependent variable as weight by clicking on it at left and then clicking the grey rectangle with the darker little triangle next to the open box marked “Dependent Variable”. Similarly, we select our independent variable as height in the same way and place it under the open box marked “Covariates”[1]. That done, click OK and the following screen appears:

[1] The term **covariate** is often used loosely to mean *any* continuous predictor variable in a model, including the predictor variable of an hypothesis, but more specifically an independent variable that can influence the outcome of a given statistical trial, but which is not of direct interest. (Oxford Languages)



As can be seen, all the statistics for linear regression have been instantly calculated and tabulated as the output at right. The ones in which we are especially interested for this task are the unstandardized coefficient of 0.179 and its positive sign as well as the p-value of 0.241, displayed at the bottom under the heading “Coefficients”. The sign of the coefficient tells us that there is a positive relation between height and weight but the magnitude of this value shows that the slope of the regression line is very gentle. The p-value not is larger than our α of 0.10; therefore we decide not to reject the null hypothesis. So, for our sample at least, while there may be a positive relation between height and weight, it is not significant. Perhaps for a larger sample of people not drawn from the same team, the results would have been different.

This is just one application offered by JASP. You are encouraged to look at the examples provided with the software and explore other statistical functions with which you are already familiar such as “Descriptives” and “T-Tests”.