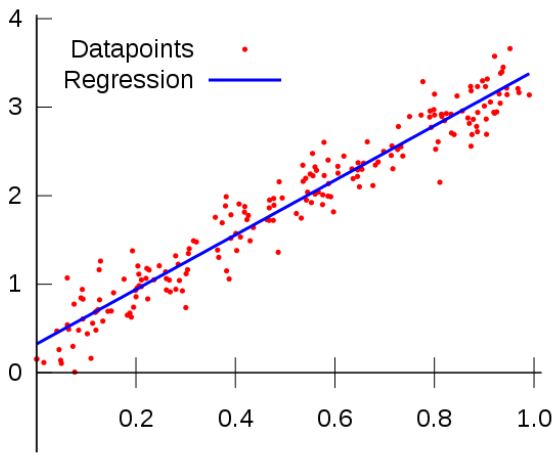
# Critical Reasoning 22 - Simple Linear Regression




*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 and used this 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 scalar dependent variable $y$ and one or more **explanatory**, **predictor** or independent variables, $x$. The word 'linear' here is used in its adjectival form to refer to the property of resembling a 'straight line' or more precisely to describe a polynomial function of one or zero degree. 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 $A = \pi r^2$ or Ohms' $V = IR$ law, where the equation in question *exactly* describes the relationship between the two variables. Statistical relationships, on the other hand, involve a relationship between variables that is not exact or prefect.

**A note on terminology**: Unfortunately there is no consistency in the way that various authors refer to the variables involved. The independent variable(s) may be variously called 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 explanatory 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 predictor variables do a good job in predicting an outcome (dependent) variable? and

2. Which variables in particular are significant predictors of the outcome variable, and in what way do they… impact the 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 on the independent variable. No doubt you will recognize this equation 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 identifying the strength of the effect that the 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** 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**: All variables should be multivariate normal, *i.e.* every linear combination of its components should have a univariate normal distribution. This can be most easily assessed visually by constructing a histogram (See Critical Reasoning 13). However there are other statistical techniques for checking normality by using a goodness of fit test such as the Kolmogorov-Smirnov test, which we shall not consider them here.

- **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 Bivariate Correlation. The value of the calculated correlation coefficients should be less than 1 but also not too close to 1. (See Critical Reasoning 19)

- **No auto-correlation**: 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. Auto-correlation

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 pattern. 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 "spay out" in either direction 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 when 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.

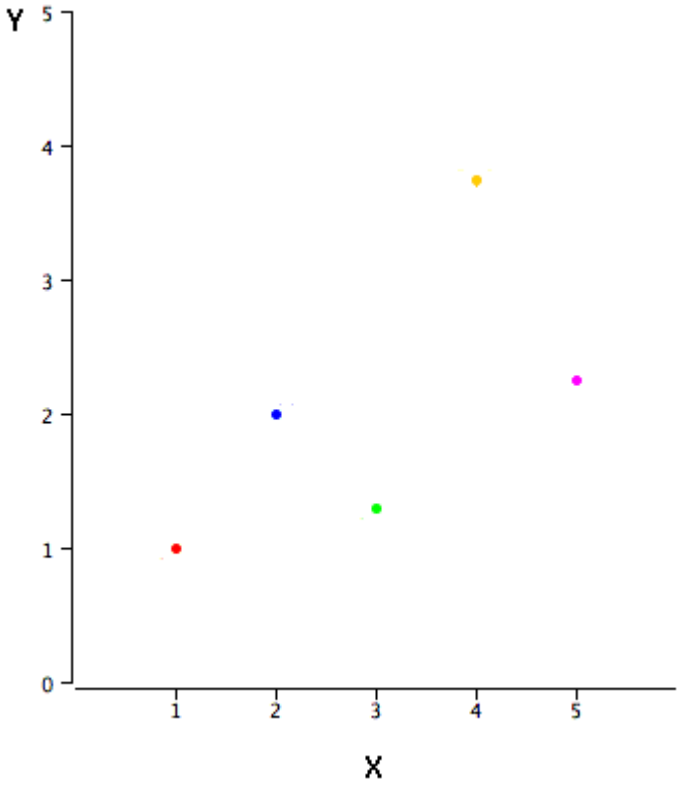
| X | Y |
|---|---|
| 1.00 | 1.00 |
| 2.00 | 2.00 |
| 3.00 | 1.30 |
| 4.00 | 3.75 |
| 5.00 | 2.25 |

*Table of example data*

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.

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 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 represents a population or a sample. However because they are so few of them 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: $\mathrm{M_X}$ and $\mathrm{M_Y}$ being the means of X and Y respectively; $\mathrm{s_X}$ and $\mathrm{s_Y}$ 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 $\mu$. The following table lists the statistics required for computing the regression line:

| $\mathbf{M_X}$ | $\mathbf{M_Y}$ | $\mathbf{s_X}$ | $\mathbf{s_Y}$ | **r** |
|---|---|---|---|---|
| 3 | 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:

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

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

$$c = 2.06 - (0.425)(3) = 0.785$$

Therefore the formula for the regression line is

$$y = 0.425x + 0.785$$

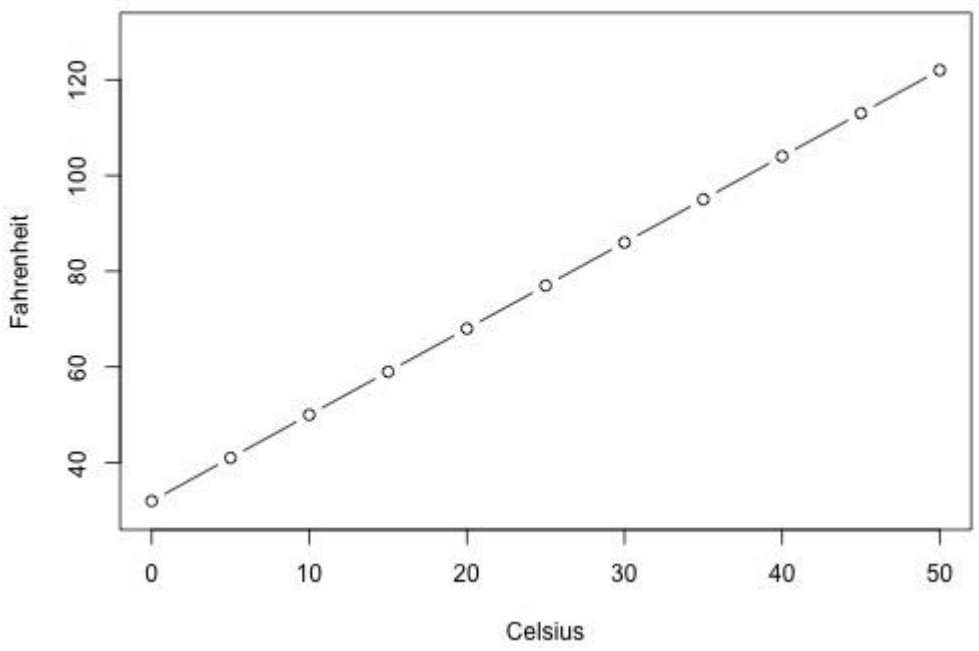
We can see that this line meets with our expectations. The sign of the gradient is positive, so we know that it slopes upwards towards the right. We can also see from the value of the gradient that this is not a very steep line. A value of 1 represents a line at 45° to the horizontal but our line is 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 but 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. Thus there is a perfect liner relationship between degrees Celsius and degrees Fahrenheit. In fact we know this relationship to be

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

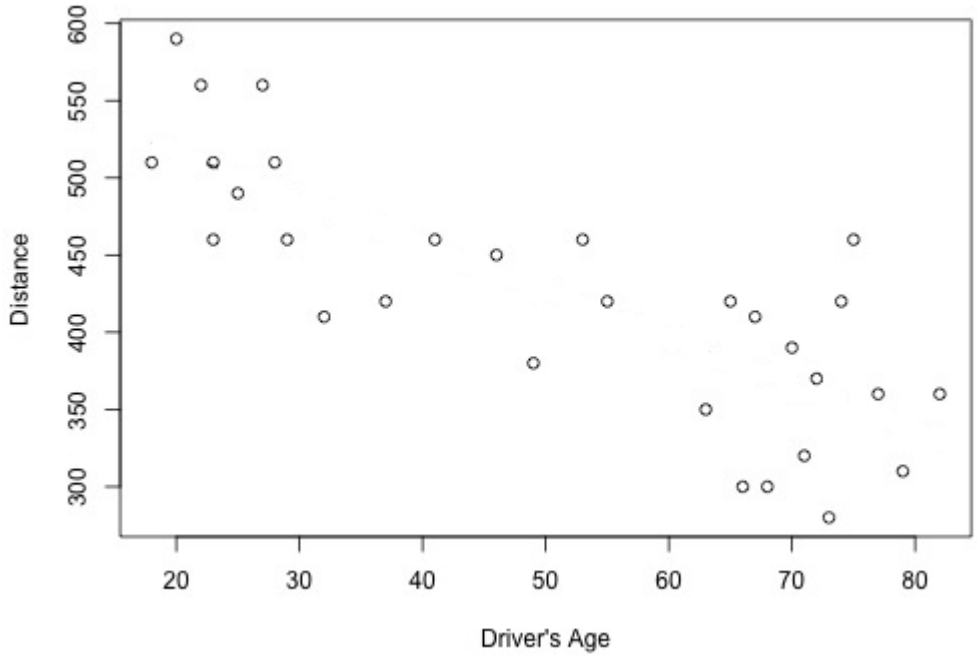
Therefore 100% of the variation in temperatures in Fahrenheit is explained by the temperature in Celsius. Clearly what we are dealing with here is a deterministic relationship between the variables; therefore a statistical approach to this problem would have been uncalled for.




*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.

| 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

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

The rest is simply a matter of substitution, thus

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

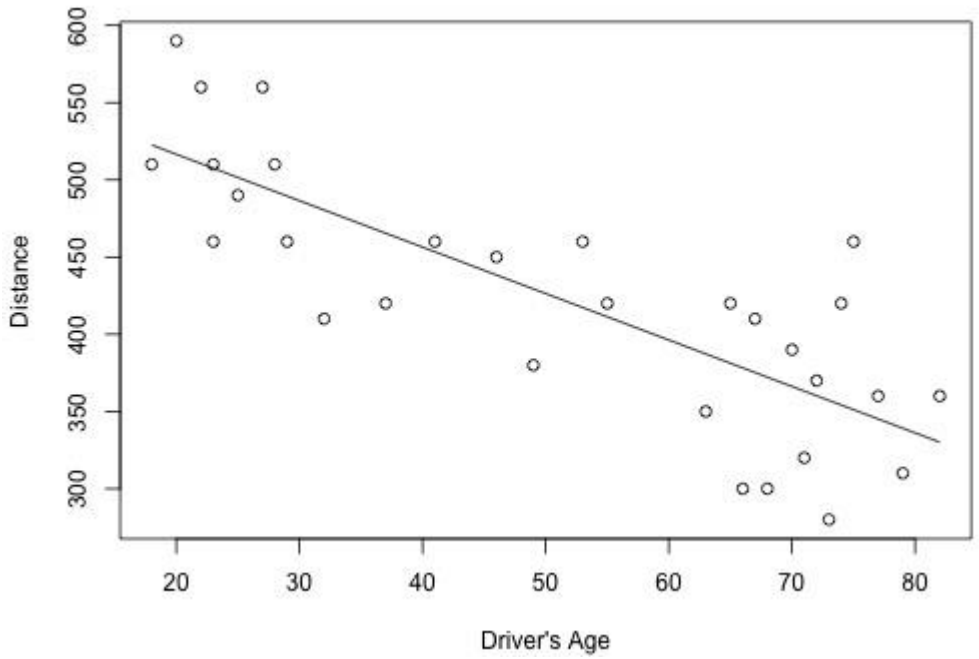
and

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

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. 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. 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 point as in the diagram above. Notice that the axes have also been truncated so that, for example, the age of a driver is not represented below the legal age for a driver's license.

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 is reduced by taking the age of the driver into account.

**$r^2$ Precautions**

Unfortunately the correlation coefficient $r$ and the coefficient of determination $r^2$ are frequently misunderstood and misused. 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 fall victim to the most common mistakes. That page is indexed separately at Penn State Eberly College of Science's STAT 501 website.

**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 that age 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 - quite likely it would fall off quite steeply into advanced old age.

End of Section

There will be an update to this topic posted in the coming weeks.