[Linear Regression](https://en.wikipedia.org/wiki/Linear_regression#Applications_of_linear_regression)

# Overview

In a Linear Regression, the relationships are modeled using linear prediction functions whose unknown model parameters are estimated from the data. Such models are called linear models. Most commonly, the conditional mean of y given the value of X is assumed to be an affine function of X; less commonly, the median or some other quantile of the conditional distribution of y given X is expressed as a linear function of X. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of y given X, rather than on the joint probability distribution of y and X, which is the domain of multivariate analysis.

Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways such as by minimizing the “lack of fit” in some other norm (as with least absolute deviations regression), or by minimizing a penalized version of the least squares cost function as in ridge regression (L2-norm penalty) and lasso (L1-norm penalty). Conversely, the least squares approach can be used to fit models that are not linear models. Thus while the terms “least squares” and “linear model” are closely linked, they are not synonymous.

# Practical applications

Most applications fall into one of two broad categories

1. Prediction, forecasting, or error reduction – Linear regression can be used to fit a predictive model to an observed dataset of y and X values. After developing such a model, if an additional value of X is then given without its accompanying value of y, the fitted model can be used to make a prediction of the value of y.
2. Quantify the relationship between *y* and – Given variable *y* and a number of variables that may be related to *y*, linear regression analysis can be applied to quantify the strength of the relationships between *y* and , to assess which may have no relationship with *y* at all, and to identify which subsets of the contain redundant information about *y*.

## Trend line

A trend line represents a trend, the long-term movement in time series data after other components have been accounted for. It tells whether a particular data set (say GDP, oil prices, or stock prices) have increased or decreased over the period of time. Typically, a trend line’s slope and position is calculated using statistical techniques like a linear regression. Trend lines are usually straight lines, although some variations use higher degree polynomials depending on the degree of curvature desired in the line.

Trend lines are sometimes used in business analytics to show changes in data over time. This has the advantage of being simple. Trend lines are often used to argue that a particular action or even (such as training, or an advertising campaign) caused observed changes at a point in time. This is a simple technique, and does not require a control group, experimental design, or a sophisticated analysis technique. However, it suffers from a lack of scientific validity in cases where other potential changes can affect the data.

## [Economics](https://en.wikipedia.org/wiki/Econometrics)[[1]](#footnote-1)

Linear regression is the predominant empirical tool in Economics. For example, it is used to predict consumption spending, fixed investment spending, inventory investment, purchases of a country’s exports, spending on imports, the demand to hold liquid assets, labor demand, and labor supply.

Econometrics is the application of mathematics, statistical methods, and computer science to economic data and is described as the branch of economics that aims to give empirical content to economic relations. More precisely, it is “the quantitative analysis of actual economic phenomena based on the concurrent development of theory and observation, related by appropriate methods of inference.”

The basic tool for econometrics is the linear regression model. In modern econometrics, other statistical tools are frequently used, but linear regression is still the most frequently used starting point for an analysis.

Econometric theory uses statistical theory to evaluate and develop econometric methods. Econometricians try to find [estimators](https://en.wikipedia.org/wiki/Estimator) that have desirable statistical properties including unbiasedness, efficiency, and consistency.

* [Unbiased estimator](https://en.wikipedia.org/wiki/Bias_of_an_estimator) – the estimated value is equal to the true value of the parameter
* [Consistent estimator](https://en.wikipedia.org/wiki/Consistent_estimator) – the estimated value converges to the true parameter value as sample size gets get larger
* [Efficient estimator](https://en.wikipedia.org/wiki/Efficiency_(statistics)) – the estimator has a lower standard error than other unbiased estimators for a given sample size.

[Ordinary Least Squares (OLS)](https://en.wikipedia.org/wiki/Ordinary_least_squares) is often used for estimation since it provides the Best Linear Unbiased Estimator (BLUE), where “best” means most efficient, unbiased estimator, given the [Gauss-Markov](https://en.wikipedia.org/wiki/Gauss%E2%80%93Markov_theorem) assumptions. When these assumptions are violated or other statistical properties are desired, other estimation techniques such as maximum likelihood estimation, generalized method of moments, or generalized least squares are used. Estimators that incorporate prior beliefs are advocated by those who favor [Bayesian statistics](https://en.wikipedia.org/wiki/Bayesian_statistics) over traditional, classical, or “frequentist” approaches.

[Applied econometrics](https://en.wikipedia.org/wiki/Methodology_of_econometrics)[[2]](#footnote-2) uses theoretical econometrics and real-world data for assessing economic theories, developing [econometric models](https://en.wikipedia.org/wiki/Econometric_model), analyzing economic history, and [forecasting](https://en.wikipedia.org/wiki/Economic_forecasting). Econometrics may us standard statistical models to study economic questions, but most often they are observational data, rather than in controlled experiments. In this, the design of observational studies in economics is similar to the design of the studies in other observational disciplines, such as astronomy, epidemiology, sociology, and political science. Analysis of data from an observational study is guided by the study protocol, although exploratory data analysis may be useful for generating new hypotheses. Economics often analyzes systems of equations and inequalities, such as supply and demand hypothesized to be in equilibrium. Consequently, the field of econometrics has developed methods for identification and estimation of simultaneous-equation models. These methods are analogous to methods used in other areas of science, such as the field of systems identification in systems analysis and control theory. Such methods may allow researchers to estimate models and investigate their empirical consequences, without directly manipulating the system.

# Simple Linear Regression

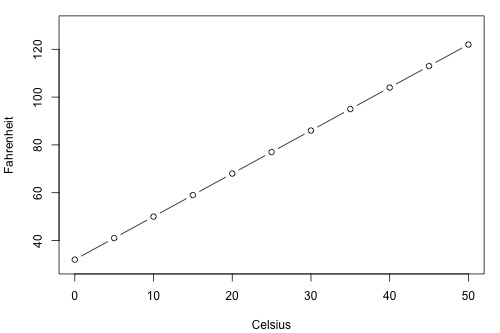
## What is a Simple Linear Regression?

A [simple linear regression](https://en.wikipedia.org/wiki/Simple_linear_regression) is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables. In a simple linear regression there is one explanatory variable (sometimes referred to as the independent, or predictor variable), denoted as *x*, and a second outcome variable (also known as the dependent, or response variable), denoted as *y*. For the purposes of this article, the variables will be known as the “predictor” and “response” variables.

Before proceeding, it is important to make a clear distinction between deterministic (or functional) relationships and statistical relationships. For statistical purposes deterministic relationships are not the subject of regression analysis because for each deterministic relationship, the equation exactly describes the relationship between the two variables.

An example of a deterministic relationship is the conversion of Fahrenheit and Celsius. This conversion equation is:

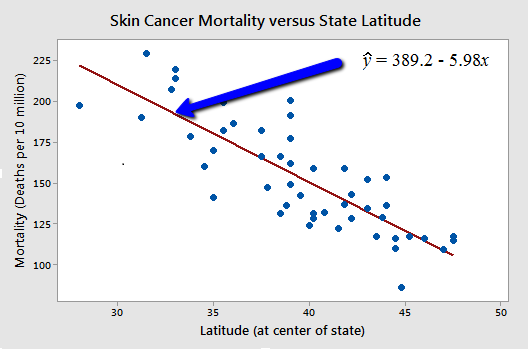
The graph of the above conversion equation is illustrated below:



The point is that if you know the exact value in Celsius then you know the exact value in Fahrenheit, and vice-versa. Regression analysis is instead concerned with statistical relationships in which the relationship between the two variables is not perfect.

Now let’s consider an example of a statistical relationship. Let the response variable *y* be the mortality due to skin cancer (number of deaths per 100 million people) and the predictor variable x be the latitude (degrees North) at the center of each of the listed 49 states in the U.S. The data used was compiled in the 1950’s when neither Alaska nor Hawaii were states. Additionally, Washington D.C. was included as a data point despite not technically being a state. The dataset [skincancer.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/skincancer.txt) can be downloaded utilizing the preceding link.

When analyzing the data, a reasonable hypothesis would be that the instances of deaths due to skin cancer would decrease as latitude decreases. A scatter plot with a trend overlay supports such a hypothesis, as seen below:



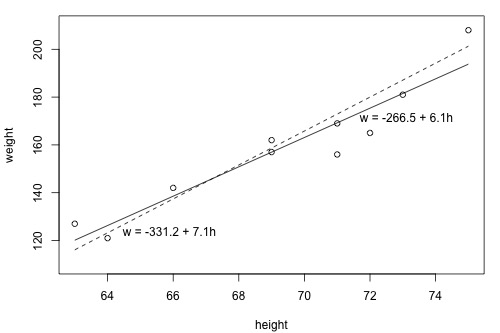
Thus, given the data we have supporting evidence that if you lived in the higher latitudes of the northern U.S., the less exposed you would be to the harmful rays of the sun, and therefore the less risk that you would have of death due to skin cancer. Notice that there is a downward trend but the relationship is not perfect. If the relationship were perfect, then the scatter points would line up perfectly along the trend line and we would instead have a deterministic relationship. As seen from the scatterplot, the relationship is indeed imperfect yet still exhibits a negative linear relationship between latitude and mortality due to skin cancer.

Here are a few more examples of statistical relationships:

* Height and Weight – as height increase, you would expect weight to increase, only imperfectly.
* Alcohol Consumption and Blood Alcohol Content – As alcohol consumption increase then you would expect one’s blood alcohol content to also increase, although imperfectly across subjects.
* Vital Lung Capacity and Pack-Years of Smoking – As the amount of smoking increase (quantified by number of packs-years of smoking) then it would be reasonable to expect lung function (as quantified by vital lung capacity) to decrease, although there is variation as to the degree of degradation across patients.
* Driving Speed and Gas Mileage – As driving speed increase the gas mileage should decrease, but there will be variation between cars and across make and model.

## What is the Best Fitting Line?

Otherwise known as a “trend line”, the best fitting line summarizes the trend between two quantitative variables. This is usually best illustrated by using a scatterplot of *(x,y)* data and drawing a line that best illustrates the plotted points on the resulting graph. For example, take the data found at student\_height\_weight.txt, with weights = *x*, heights = *y*, *obs* = 10. The resulting plot can be found below:



Looking at the above plot which line (dashed or solid) best summarizes the trend between height and weight? The answer to this question can be somewhat difficult as many lines look good but determining best fit using your eyes alone is tricky.

The equation for the best fitting line is

Where

* denotes the observed response for experimental unit
* denotes the predictor value for experimental unit
* is the predicted response (or fitted value) for experimental unit

Remember that the experimental unit is the object or person on which the measurement is made. In this example the heights and weights are measurements of students, which are the experimental unit.

Let’s try out the notation on our example with the trend summarized by the line . Note that this line is just a more precise version of the above solid line, . The first data point in the list indicates that student 1 is 63 inches tall and weighs 127 pounds. That is, and . Do you see this point on the plot? If we know this student’s height but not his or her weight, we could use the equation of the line to predict his or her weight. We’d predict the student’s weight to be -266.53+6.1376(63) or 120.1 pounds. That is . Clearly, our prediction wouldn’t be correct – it has some **prediction** (or **residual**) **error**. In fact, the size of this point’s prediction error is 127-120.1 = 6.9 pounds.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  | | --- | --- | --- | --- | |  |  |  |  | | 1 | 63 | 127 | 120.1 | | 2 | 64 | 121 | 126.3 | | 3 | 66 | 142 | 138.5 | | 4 | 69 | 157 | 157.0 | | 5 | 69 | 162 | 157.0 | | 6 | 71 | 156 | 169.2 | | 7 | 71 | 169 | 169.2 | | 8 | 72 | 165 | 175.4 | | 9 | 73 | 181 | 181.5 | | 10 | 75 | 208 | 193.8 | |  |

As you can see, the size of the prediction error depends on the data point. If we didn’t know the weight of student 4, the equation line would predict their weight to be -266.53+6.1376(69) = 157 pounds. The size of the error of this error can be founds here as 162-157 = 5 pounds.

In general, when we use to predict the actual response , we make a prediction error ( or residual error) of size:

A line that fits the data “best” will be one for which the **n prediction errors** – one for each observed data point – **are as small as possible in some overall sense**. One way to achieve this goal is to invoke the “**least squares criterion**,” which says to “**minimize the sum of the squared prediction errors**.” That is:

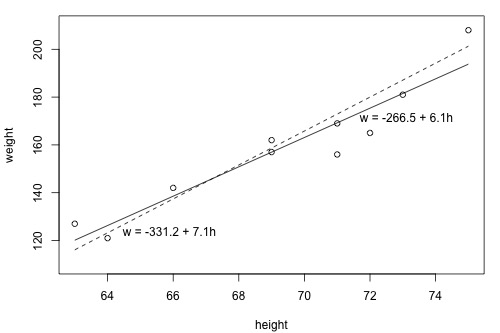
* The equation of the best fitting line is:
* We just need to find the values and that make the sum of the squared prediction errors the smallest it can be.
* That is, we need to find the values and that minimizes:

Here is how you might think about this quantity Q:

* The quantity is the prediction error for data point .
* The quantity is the squared prediction error for data point .
* And, the symbol tells us to add up the squared prediction errors for all data points.

Incidentally, if we didn’t square the prediction error to get , the positive and negative prediction errors would cancel each other out when summed, always yielding 0.

Now being familiar with the least squares criterion, let’s take a fresh look at our plot again. In light of the least squares criterion, which line do you think is the best fitting line?



Let’s see how you did! The following two side-by-side tables illustrate the implementation of the least squares criterion for the two lines up for consideration – the dash line and the solid line.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | (The dashed line) | | | | | | |  |  |  |  |  |  | | 1 | 63 | 127 | 116.1 | 10.9 | 118.81 | | 2 | 64 | 121 | 123.2 | -2.2 | 4.84 | | 3 | 66 | 142 | 137.4 | 4.6 | 21.16 | | 4 | 69 | 157 | 158.7 | -1.7 | 2.89 | | 5 | 69 | 162 | 158.7 | 3.3 | 10.89 | | 6 | 71 | 156 | 172.9 | -16.9 | 285.61 | | 7 | 71 | 169 | 172.9 | -3.9 | 15.21 | | 8 | 72 | 165 | 180.0 | -15.0 | 225.00 | | 9 | 73 | 181 | 187.1 | -6.1 | 37.21 | | 10 | 75 | 208 | 201.3 | 6.7 | 44.89 | |  |  |  |  |  | ------------  **766.5** | | |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | (The solid line) | | | | | | |  |  |  |  |  |  | | 1 | 63 | 127 | 120.139 | 6.8612 | 47.076 | | 2 | 64 | 121 | 126.276 | -5.2764 | 27.840 | | 3 | 66 | 142 | 138.552 | 3.4484 | 11.891 | | 4 | 69 | 157 | 156.964 | 0.0356 | 0.001 | | 5 | 69 | 162 | 156.964 | 5.0356 | 25.357 | | 6 | 71 | 156 | 169.240 | -13.2396 | 175.287 | | 7 | 71 | 169 | 169.240 | -0.2396 | 0.057 | | 8 | 72 | 165 | 175.377 | -10.3772 | 107.686 | | 9 | 73 | 181 | 181.515 | -0.5148 | 0.265 | | 10 | 75 | 208 | 193.790 | 14.2100 | 201.924 | |  |  |  |  |  | ------------  **597.4** | |

Based on the lease squares criterion, which equation best summarizes the data? The sum of the squared prediction errors is 766.5 for the dashed line, while it is only 597.4 for the solid line. Therefore, of the two lines, the solid line, , best summarizes the data. But is this equation to be the best fitting line out of all the possible lines we didn’t consider? Of course not!

If we used the above approach for finding the equation of the line that minimizes the sum of the squared prediction errors, we would be calculating until we die of old age. We’d have to implement the above procedure an infinite number of times – clearly an impossible task. Fortunately, somebody has done some dirty work for us by figuring out formulas for the **intercept** and the **slope** for the equation of the line that minimizes the sum of the squared errors.

The formulas are determined using methods of calculus. We minimize the equation for the sum of the squared prediction errors:

(that is, take the derivative with respect to and , set equal to 0, and solve for and ) and get the “**least squares estimates**” for and :

And:

Because the formulas for and are derived using the least squares criterion, the resulting equation is often refered to as the “**least squares regression line**,” or simply the “**least squares line**.” It is also sometimes called the “**estimated regression equation**.” Incidentally, note that in deriving the above formulas, we made no assumptions about the data other than that they follow some sort of linear trend.

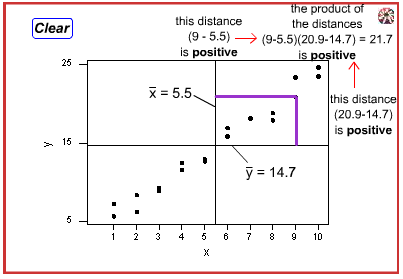
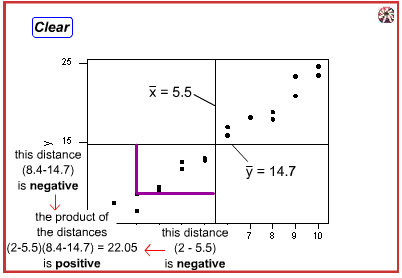
We can see from these formulas that the least squares line passes through the point , since when then .

In practice, you will not really need to worry about the formulas for and . Instead, you are going to let statistical software, such as Minitab or Stata, find the least squares lines for you. But, we can still learn something from the formulas – for in particular.

If you study the formula for the slope of :

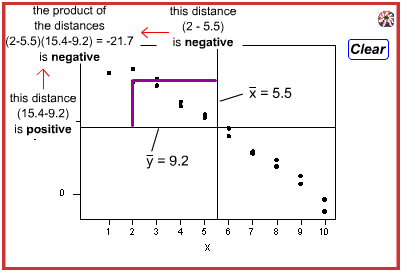
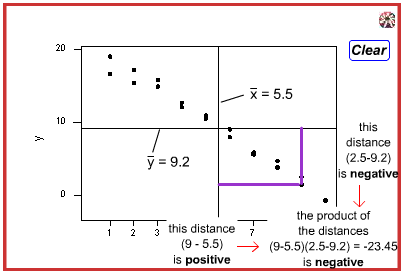
you see that the denominator is necessarily positive since it only involves summing positive terms. Therefore, the sign of the slope is solely determined by the numerator. The numerator tells us to sum up the product of two distances, for each data point – the distance of the x-value from the mean of all the x-values and the distance of the y-value from the mean of all y-values. Let’s see how this determines the sign of the slope of by studying the following two plots.

**When is the slope ?** Do you agree that the trend in the following plot is positive – that is, as increases, tends to increase? If the trend is positive, then the slope must be positive. Let’s see how!

* Look at the blue data point in the upper right quadrant. Note that the product of the two distances for this data point is positive. In fact, the product of the two distances is positive for *any* data point in the upper right quadrant. A positive times a positive will always yield a positive.  
  
* Now view the blue data point in the lower left quadrant. Note that the product of the two distances for this data point will also be positive. In fact, the product of any two distances will be positive for *any* data point in the lower left quadrant. I.e. The product of two negatives will always yield positive result.  
  

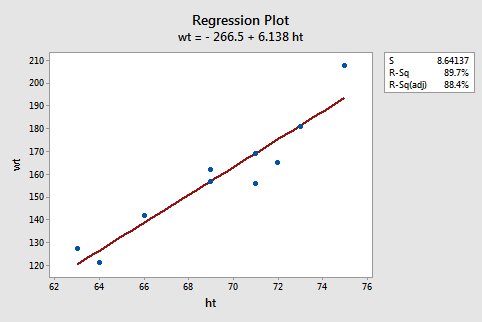
Adding up all of the positive productes must necessarily yield a positive number, and hence the slope of the line will be positive.

**When is the slop of ?** Now, do we agree that the trend in the following plot is negative, that is as x increases, y tends to decrease? If the trend is negative, then the slope if must be negative. Let’s see how!

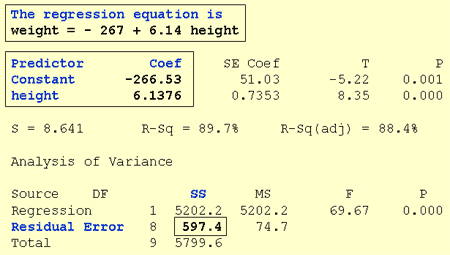
* Observe the blue data point in the upper left quadrant. Note that the product of the two distances for this data point is negative. In fact, the product of the two distances is negative for any data point in the upper left quadrant. The product of a positive and a negative will always be negative.  
  
* Observe the blue data point in the lower right quadrant. Note that the product of the two distances for this data point is negative. In fact, the product of the two distances is negative for any data point in the lower right quadrant. The product of a negative and a positive will always be negative.  
  

Adding up all of these negative products must necessarily yield a negative number, and hence the slope of the line will be negative.

Now that we’ve finished that investigation, you can set aside the formulas for and . Again, in practice, you are going to let statistical software, such as minitab, find the least squares lines for you. We can obtain the estimated regression equation in two different places in minitab. The following plot illustrates where you can find the least squares line on Minitab’s **“fitted line plot.”**



The following Minitab output illustrates where you can find the “least squares line” in Minitab’s **“standard regression analysis”** output.



Note that the estimated values and also appear under the columns labled “**Predictor**” (the intercept is always referred to as the “**Constant**” in Minitab) and “**Coef**” (for “Coefficients”). Also note that the value we obtained by minimizing the sum of the squared prediction errors, 597.4, appears in the “**Analysis of Variance**” table appropriately in a row labeled “Residual Error” and under a column labeled “SS” (for “Sum of Squares”).

Although we’ve learned how to obtain the “estimated regression coefficients” and , we’ve not yet discussed what we learn from them. One thing they allow us to do is to predict future responses – one of the most common uses of an estimated regression line. This use is rather straightforward.

|  |  |
| --- | --- |
| A common use of the estimated regression line |  |
| Predict (mean) weight of 66-inch tall people |  |
| Predict (mean) weight of 67-inch tall people |  |

**Now what does tell us?** The answer is obvious when you evaluate the estimated regression equation at . Here it tells us that a person who is 0 inches tall is predicted to weigh -267 pounds! Clearly this prediction is nonsense. This happened because we “extrapolated” beyond the “scope of the model” (the range of the x values). It is not meaningful to have a height of 0 inches; that is, the scope of the model does not include and the intercept is not meaningful. In general, if the “scope of the model” includes , then is the predicted response when . Otherwise is not meaningful.

And, what does tell us? The answer is obvious when you subtract the predicted weight of 66-inch tall people from the predicted weight of 67-inch tall people. We obtain pounds – the value of . Here it tells us that we predict that the mean weight will increase by 6.14 pounds for every additional 1-inch increase in height. In general, we can expect the mean response to increase or decrease by units for every one unit increase in .

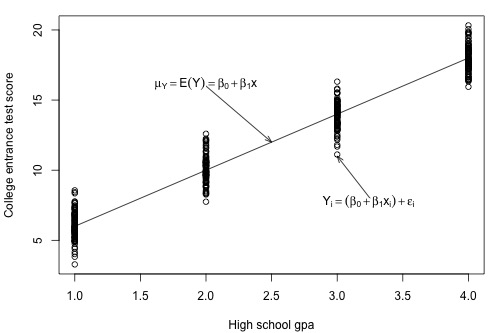
## The Simple Linear Regression Model

We have worked hard to come up with formulas for the intercept and the slope of the least squares regression line. But we haven’t yet discussed what and estimate.

So what do and estimate?

Let’s investigate this question with another example. Below is a plot illustrating a potential relationship between the predictor “high school grade point average (GPA)” and the response “college entrance test score.” Ony five groups (“subpopulations”) of students are considered – those with a GPA of 1.0, 2.0, 3.0, or 4.0.

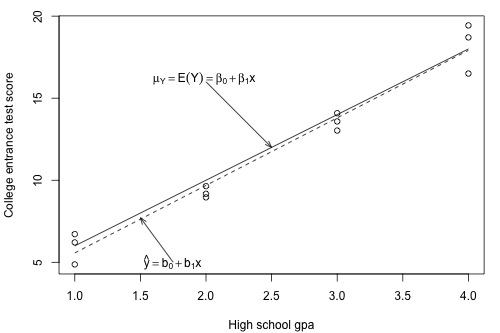
Let’s focus, for now, just on the students with a GPA of 1.0. As you can see there are so many data points – each representing one student – that the data points run together. That is, the data on the entire subpopulation of students with a 1.0 GPA are plotted. Similarly, the data on the entire subpopulations of students with 2.0, 3.0, and 4.0 GPA’s are plotted.



Now take the average college entrance test score for students with a 1.0 GPA and, similarly, take the average college entrance score for students with a 2.0, 3.0, and 4.0 GPA. Connecting the dots, i.e. the averages, you get a line, summarized by the formula . The line, which is called the “population regression line,” summaries the trend in the population between the predictor x and the mean of the responses . We can also express the average college entrance test score for the *i-*th student, . Of course, not every student’s college entrance test score will equal the average . There will be some error. That is, any student’s response will be the linear trend plus some error . So, another way to write the simple linear regression model is .

When looking to summarize the relationship between a predictor *x* and a response *y*, we are interested in knowing the population regression line . However, the only way we could ever know it is to be able to collect data on every person in the population, which is most often an impossible task. We have to rely instead on taking and using a sample of the data from the population in order to estimate the population regression line.

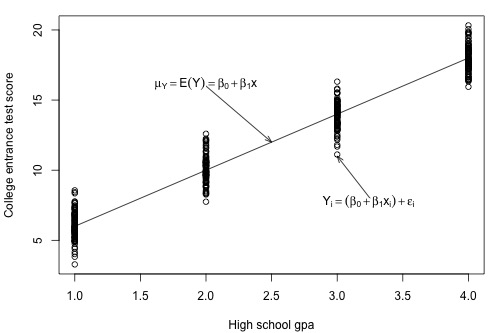
Let’s take a sample of three students from each of the subpopulations, i.e. three students with a 1.0 GPA, three students with a 2.0 GPA, etc…, for a total of 12 students. As the plot below suggests, the least squares regression line through the sample of 12 data points estimates the population regression line . That is, the sample intercept estimates the population intercept and the slope estimates the population slope .



The least squares regression line doesn’t match the population regression line perfectly, but it is a pretty good estimate. And, of course, we’d get a different least squares regression line if we took another (different) sample of 12 such students. Ultimately, we are going to want to use the sample slope to learn about the parameter we care about, the population slope . And we will use the intercept to learn about the population intercept .

In order to draw any conclusions about the population parameters and , we have to make a few more assumptions about the behavior of the data in a regression setting. We can get a pretty good feel for the assumptions by looking at our plot of GPA against college entrance test scores.

First, notice that when we connected the averages of the college entrance test scores for each of the subpopulations, it formed a line. Most often, we will not have the population data at our disposal, as we pretend to have here. If we didn’t, do you think it would be reasonable to assume that the mean college entrance test scores are **linearly related** to high school GPA’s?



Again, let’s focus on just one subpopulation, those students who have a GPA of 1.0. Notice that most of the college entrance scores for these students are clustered near the mean of 6, but few students did much better than the subpopulation’s average scoring around 9, and a few students did a bit worse scoring about a 3. Do you get the picture? Thinking instead about the errors, , most of the errors for these students are clustered near the mean of 0, but a few are as high as 3, and a few are as low as -3. If you could draw a probability curve for the errors above this subpopulation of data, what kind of curve do you think it would be? Does it seem reasonable to assume that the errors for each subpopulation are **normally distributed**?

Looking at the plot again, notice that the spread of the college entrance test scores for the students with a 1.0 GPA is similar to the spread of the college entrance test scores for students with 2.0, 3.0, and 4.0 GPAs. Similarly, the spread of the errors is similar, no matter the GPA. Does it seem reasonable to assume that the errors for each subpopulation have **equal variance**?

Does it also seem reasonable to assume that the error for one’s student’s college entrance test score is independent of the error for another student’s college entrance test score? I’m sure that you can come up with some scenarios – cheating students, for example – for which this assumption would not hold. But if you take a random sample from the population, it should be an assumption that is easily met.

We are now ready to summarize the four conditions that comprise **the simple linear regression model**:

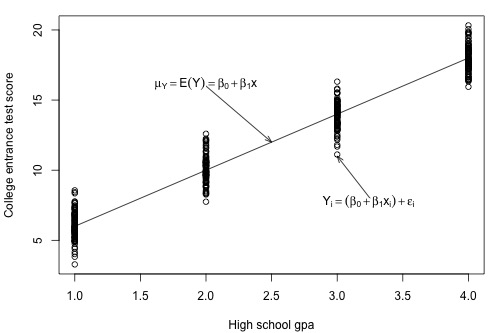
* The mean response, , at each value of the predictor, , is a Linear function of the .
* The errors, , Independent.
* The errors, , at each value of the predictor, , are Normally distributed.
* The errors, , at each value of the predictor, , have Equal variances (denoted ).

Do you notice what the first letters, that are colored in blue, spell? “LINE.” And what are we studying in this course? Lines! Get it?! You might find mnemonic devices a useful way to remember the four conditions that make up what we call the “simple linear regression model.” Whenever you here “simple linear regression model,” think of these four conditions!

An equivalent way to think of the first condition (linearity) is that the mean of the error, , at each value of the predictor, , is **zero**. An alternative way to describe all four assumptions is that the errors, , are independent normal random variables with mean zero and constant variance, .

## What is the Common Error Variance?

The plot of our population of data suggests that the college entrance test scores for each subpopulation have equal variance. We denote the value of this common variance as .

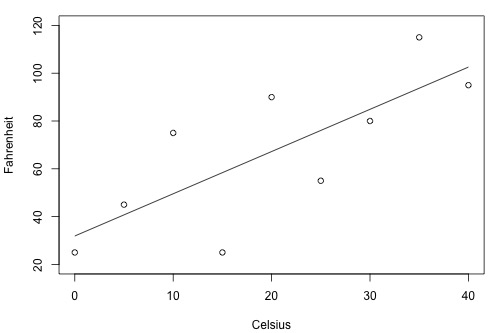


That is, quantifies how much the responses vary around the (unknown) mean population regression line .

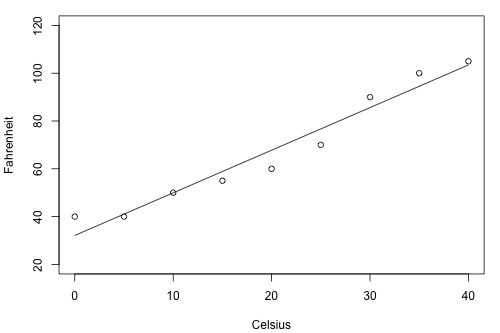
Why should we care about ? The answer to this question pertains to the most common use of an estimated regression line, namely predicting some future response.

Suppose you have two brands (A and B) of thermometers, and each brand offers a Celsius and a Fahrenheit thermometer. You measure the temperature in Celsius and Fahrenheit using each brand of thermometer on ten different days. Based on the resulting data, you obtain two estimated regression lines – one for brand A and one for brand B. You plan to use the estimated regression lines to predict the temperature in Fahrenheit based on the temperature in Celsius.

Will this thermometer brand (A) yield more precise future predictions…



… Or will this one (B)?

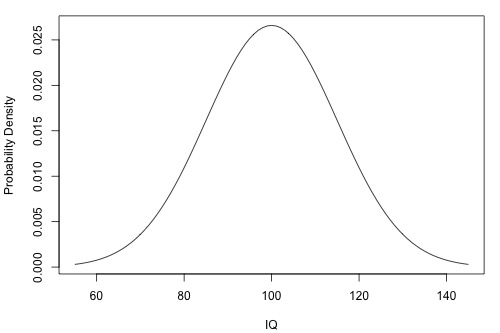


As the two plots illustrate, the Fahrenheit responses for the brand B thermometer don’t deviate as far from the estimated regression equation as they do for the brand A thermometer. If we use the brand B estimated line to predict the Fahrenheit temperature, our prediction should never be too far off from the actual observed Fahrenheit temperature. On the other hand, predictions of the Fahrenheit temperatures using the brand A thermometer can deviate quite a bit from the actual observed Fahrenheit temperature. Therefore, the brand B thermometer should yield more precise future predictions than the brand A thermometer.

To get an idea of how precise future predictions would be, we need to know how much the responses () vary around the (unknown) mean population regression line . As stated earlier, quantifies this variance in the responses. Will we ever know this value ? No! Because is a population parameter, we will rarely know its true value. The best we can do is estimate it!

To understand the formula for the estimate of in the simple linear regression setting, it is helpful to recall the formula for the estimate of the variance of the responses, , when there is only one population.

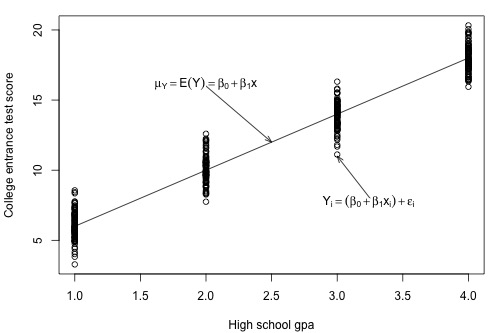
The following is a plot of the (one) population of IQ measurements. As the plot suggests, the average of the IQ measurements in the population is 100. But how much do the IQ measurements vary from the mean? That is, how “spread out” are the IQ values?



The **sample variance**:

Estimates , the variance of the one population. The estimate is really close to being like an average. The numerator adds up how far each response is from the estimated mean in squared units, and the denominator divides the sum by , not like you would expect from an average. What we would really like to is for the numerator to add up, in squared units, how far each response is from the unknown mean . But, we don’t know the population mean , so we estimate it with . Doing so “costs us one degree of freedom.” That is, we have to divide by and not , because we estimate the unknown population mean .

Now let’s expand this thinking to arrive at an estimate for the population variance in the simple linear regression setting. Recall that we assume is the same for each of the subpopulations. For our example on college entrance test scores and GPA’s, how many subpopulations do we have?



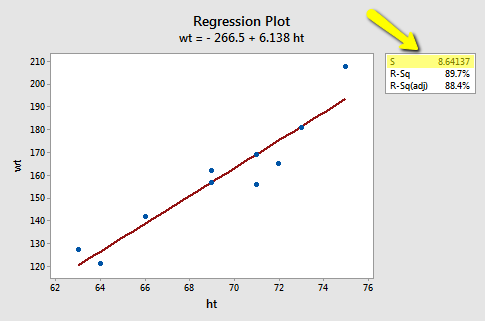
There are four subdivisions depicted in this plot. In general, there are as many subpopulations as there are distinct values in the population. Each subpopulation has its own mean , which depends on through . And each subpopulation mean can be estimated using the estimated regression equation.

The **Mean Squared Error (MSE)**:

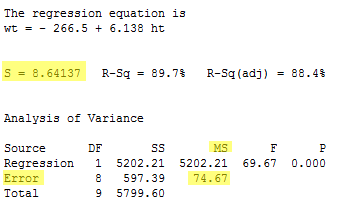
Estimates , the common variance of the many subpopulations.

How does the mean squared error formula differ from the sample variance formula? The similarities are more striking than the differences. The numerator again adds up, in squared units, how far each response is from the estimated mean. In the regression setting, though, the estimated mean is . And, the denominator divides the sum by , not , because in using to estimate , we effectively estimate two parameters – the population intercept and the population slope . That is, we lose two degrees of freedom.

In practice, we will let statistical software, such as Minitab, calculate the mean square error (MSE) for us. The estimate of shows up indirectly on Minitab’s “fitted line plot.” For example, for the student height and weight data ([student\_height\_weight.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/student_height_weight.txt)), the quantity emphasized in the box, , is the squared root of MSE. In general, , which estimates and is known as the *regression standard error* or the *residual standard error*. The fitted plot here indirectly tells us that



The estimate of shows up directly in Minitab’s standard regression analysis output. Again, the quantity is the square root of MSE. In the Analysis of Variance table, the value of MSE, 74.67, appears appropriately under the column labeled **MS** (for **M**ean **S**quared) and in the row labeled **Residual Error** (for Error).



## The Coefficient of Determination, r-squared

Let’s start our investigation of the coefficient of determination, , by looking at two different examples: one example in which the relationship between the response and the predictor is very weak, and a second example in which the relationship between the response and the predictor is very strong. If our measure is going to work well, it should be able to distinguish between these two, very different situations.

Here’s a plot illustrating a very weak relationship between and . There are two lines on the plot, a horizontal line placed at the average response, , and a shallow-sloped estimated regression line, . Note that the slope of the estimated regression line is not very steep, suggesting that as the predictor , increases, there is not much of a change in the average response . Furthermore, note that the data points do not “hug” the estimated regression line:

|  |  |
| --- | --- |
| y vs x plot |  |

The calculation on the right of the plot show contrasting “sums of squares” values:

* SSR is the “regression sum of squares” and quantifies how far the estimated sloped regression line, , is from the horizontal “no relationship line,” the sample mean or .
* SSE is the “error sum of squares” and quantifies how much the data points, , vary around the estimated regression line .
* SSTO is the “total sum of squares” and quantifies how much the data points, , vary around their mean, .

Note that . The sum of squares appears to tell the story quite well. They tell us that most of the variation in the response () is just due to random variation (), not due to the regression of on (). You might notice that divided by is . Do you see where this value appears in Minitab’s fitted line plot?

Contrast the above example with the following one in which the plot illustrates a fairly convincing relationship between and . The slope of the estimated regression line is much steeper, suggesting that as the predictor increase, there is a fairly substantial change (decrease) in the response ; and here, the data points do “hug” the estimated regression line.

|  |  |
| --- | --- |
| y vs x plot |  |

The sum of squares for this data set tell a very different story, namely that most of the variation in the response () is due ot the regression of on () not just due to random error (. Again notice divided by is , which also appears on Minitab’s fitted line plot.

The previous two examples have suggested how we should define the measure formally. In short, the “coefficient of determination” or “r-squared value,” denoted , is the regression sum of squares divided by the total sum of squares. Alternatively, since , the quantity also equals one minus the ratio of the error sum of squares to the total sum of squares:

Here are some basic characteristics of the measure:

* Since is a proportion, it is always between 0 and 1.
* If , all of the data points fall perfectly on the regression line. The predictor accounts for *all* the variation in .
* If , the estimated regression line is perfectly horizontal. The predictor accounts for *none* of the variation in .

We’ve learned the interpretation for the two easy cases – when or – but, how do we interpret , e.g. or ? Here are two similar, yet slightly different ways in which the coefficient of determination can be interpreted. We say either:

“ percent of the variation in is reduced by taking into account predictor ”

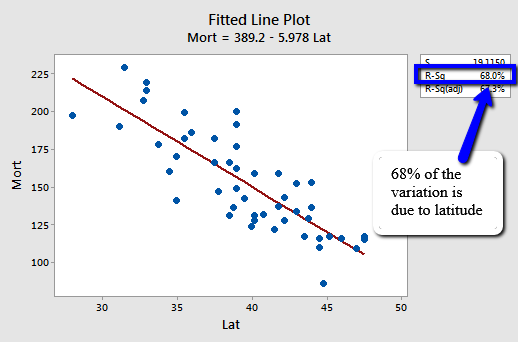
Or

“ percent of the variation in is ‘explained by’ the variation in predictor ”

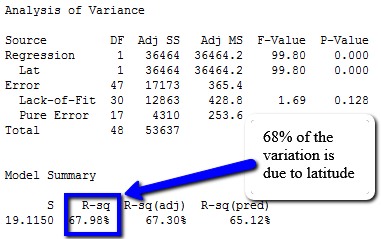
Many statisticians prefer the first interpretation. I tend to prefer the latter. The risk with using the second interpretation – and hence why ‘explained by” spears in quotes – is that it can be misunderstood as suggesting that the predictor causes the change in the response . Association is not causation. That is, just because a data set is characterized by having a large r-squared value, it does not imply that cause the changes in . As long as you keep the correct meaning in mind, it is fine to use the second interpretation. A variation on the second interpretation is to say, “ percent of the variation in is accounted for by the variation in the predictor .”

Student’s often ask: “what’s considered a large r-squared value?” It depends on the research area. Social scientists who are often trying to learn something about the huge variation in uman behavior will tend to find it very hard to get r-squared values much above, say 25% or 30%. Engineers, on the other hand, who tend to study more exact systems would likely find an r-squared value of just 30% merely unacceptable. The moral of the story is to read the literature to learn what typical r-squared values are for your research area.

Let’s revisit the skin cancer mortality example ([skincancer.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/skincancer.txt)). Any statistical software that performs simple linear regression analysis will report the r-squared value for you. It appears in two places in Minitab’s output, namely on the fitted line plot:



And in the standard regression analysis output.



We can say that about 68% of the variation in the skin cancer mortatly rate is reduced by aking into account latitude. Or we can say – with knowledge of what it really means – that about 68% of the variation in skin cancer mortality is “due to” or is “explained by” latitude.

## The Pearson Correlation Coefficient r

The correlation coefficient is directly related to the coefficient of determination in the obvious way. If is represented in decimal form, e.g. 0.39 or 0.87, then all we have to do to obtain is to take the square root of :

The sign of depends on the sign of the estimated slope coefficient .

* If is negative, then takes the negative sign
* If is positive, then takes a positive sign

That is, the estimated slope and the correlation coefficient always share the same sign. Furthermore, because is always a number between 0 and 1, the correlation coefficient is always a number between -1 and 1.

One advantage of is that it is unitless, allowing researchers to make sense of the correlation coefficient calculated on different data sets with different units. The “unitless-ness” of the measure can be seen from an alternative formula for , namely:

If is the height of an individual measured in inches and is the weight of an individual measured in pounds, then the units for the numerator is . Similarly, the units for the denominator is . Because they are the same, the units in the numerator and denominator cancel each other out, yielding a “unitless” measure.

Another formula for that you might see in the regression literature is one that illustrates how the correlation coefficient is a function of the estimated slope coefficient :

We are readily able to see from this version of the formula that:

* The estimated slope of the regression line and the correlation always share the same sign.
* The correlation coefficient is a unitless measure
* If the estimated slope of the regression line is 0, then the correlation coefficient must also be 0.

That’s enough with the formulas! As always, we will let statistical software such as Minitab do the dirty calculations for us. Here’s what Minitab’s output looks like for the skin cancer mortality and latitude example ([skincancer.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/skincancer.txt)):

**Correlation: Mort, Lat**

Pearson correlation of Mort and Lat = - 0.825

The Output tells us that the correlation between skin cancer mortality and latitude is -0.825 for this data set. Note that it doesn’t matter the order in which you specify the variables:

**Correlation: Lat, Mort**

Pearson correlation of Lat and Mort = - 0.825

The output tells us that the correlation between skin cancer mortality and Latitude is still -0.825. What does this correlation coefficient tell us? That is, how do we interpret the Pearson correlation coefficient ? In general, there is no nice practical interpretation for as there is for . You can only use to make a statement about the strength of the linear relationship between and . In general:

* If , then there is a perfect negative linear relationship between and
* If , then there is a perfect positive linear relationship between and
* If , then there is no linear relationship between and

All other values of tell us that the relationship between and is not perfect. The closer is to 0, the weaker the linear relationship. The close is to -1, the stronger the negative linear relationship. And the close is to 1, the stronger the positive relationship. As is true for the value, what is deemed a large correlation coefficient value depends greatly on the research area.

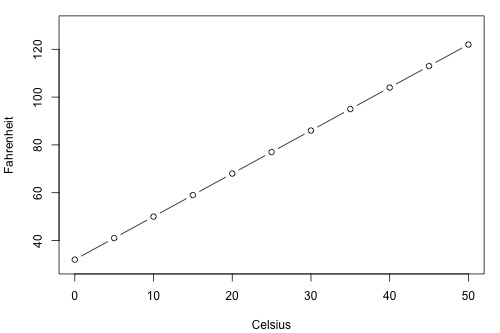
So what does the correlation of -0.825 between skin cancer mortality and latitude tell us? It tells us that:

* The relationship is negative. As the latitude increase, the skin cancer mortality rate decreases (linearly)
* The relationship is quite strong (since the value is pretty close to -1)

## Some Examples concerning and

Let’s take a look at some examples so we can get some practice interpreting the coefficient of determination and the correlation coefficient .

**Example 1:** how strong is the linear relationship between temperatures in Celsius and temperatures in Fahrenheit? Here’s a plot of an estimated regression equation based on data points:

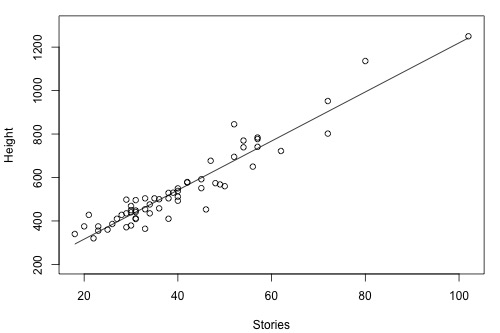


And here is Minitab’s correlation output:

**Pearson correlation coefficient of Celsius and Fahrenheit = 1.00**

It shouldn’t be surprising that Minitab report that and . Both measures tell us that there is a perfect linear relationship between temperature in degrees Celsius and temperature in degrees Fahrenheit. We know that the relationship is perfect, namely that . it should be no surprise then that tells us that 100% of the variation in temperatures in Fahrenheit is explained by the temperature in Celsius.

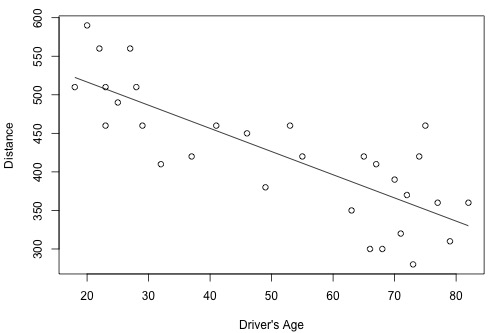
**Example 2:** How strong is the linear relationship between the number of stories a building has and its height? One would think that as the number of stories increases, the height would increase, but not perfectly. Some statisticians compiled data on a set of buildings reported in the 1994 World Almanac ([bldgstories.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/bldgstories.txt)). Minitab’s fitted line plot and correlation output look like:



**Pearson correlation of HEIGHT and STORIES = 0.951**

Minitab reports that and . The positive sign f tells us that the relationship is positive – as the number of stories increase, the height increases – as we expected. Because is close to 1, it tells us that the linear relationship is very strong, but not perfect. The value tells us that 90.4% of the variation in the height of the building is explained by the number of stories in the building.

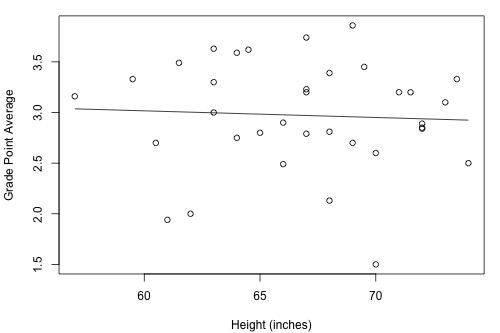
**Example 3:** How strong is the linear relationship between the age of a driver and the distance the driver can see? If we had to guess, we might think that the relationship is negative – as age increases, the distance decreases. A research firm (Last Resources, Inc., Bellefonte, PA) collected data on the sample of drivers ([signdist.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/examples/signdist.txt)). minitab’s fitted line plot and correlation output on the data look like:



**Pearson correlation coefficient of Distance and DrivAge = - 0.801**

Minitab reports that and . the negative sign of tells us that the relationship is negative – as driving age increases, seeing distance decreases – as we expected. Because is fairly close to -1, it tells us that the linear relationship is fairly strong, but not perfect. The value tells us that the 64.2% of the variation in the seeing distance is reduced by taking into account the age of the driver.

**Example 4:** How strong is the linear relationship between the height of the student and his or her GPA? Data were collected on a random sample of students in a statistics course at Penn State University ([heightgpa.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/heightgpa.txt)) and the resulting fitted line plot and correlation output were obtained:



**Pearson correlation of height and gap = - 0.053**

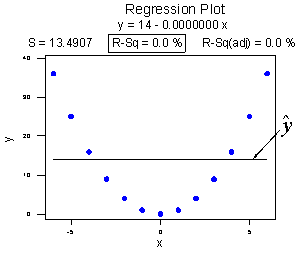
Minitab reports that and . Because is quite close to 0, it suggests – not surprisingly, I hope – that there is next to no linear relationship between height and GPA. Indeed, the value tells us that only of the variation in the GPA of the students in the sample can be explained by their height. In short, we would need to identify another more important variable, such as number of hours studied, if predicting a student’s GPA were important to us.

## Some warnings concerning R-squared

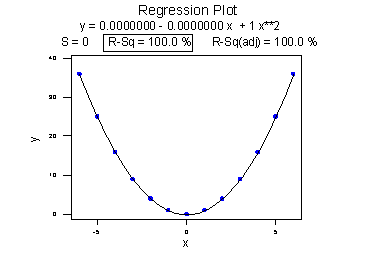
Unfortunately, the coefficient of determination and the correlation coefficient have to be the most often misused and misunderstood measures in the field of statistics. To ensure that you don’t fall victim to the most common mistakes, we review a set of seven different warnings here. Master these and you’ll be a master of measures.

**Warning 1: The coefficient of determination and the correlation coefficient quantify the strength of a linear relationship. It is possible that and , suggesting there is no linear relationship between and , and yet a perfect curved (or “curvilinear” relationship) exists**

Consider the following example. The upper plot illustrates a perfect, although curved, relationship between and , and yet Minitab reports that and . The estimated regression line is perfectly horizontal with slope . If you didn’t understand that and summarize the strength of a linear relationship, you would likely misinterpret the measures, concluding that there is no relationship between and . But, it’s just not true! There is indeed a relationship between and – it’s just not linear.



**Pearson correlation of and = 0.000**



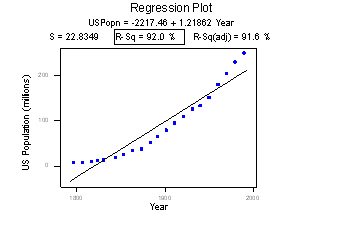
**Pearson correlation of and = 0.000**

The lower plot better reflects the curved relationship between and . Minitab has drawn a quadratic curve through the data, and reports that “R-sq = 100%” and . What is this all about? We’ll learn when we study multiple linear regression later in the course that the coefficient of determination associated with the simple linear regression model for one predictor extends to a “multiple coefficient of determination,” denoted , for the multiple linear regression model with more than one predictor. (The lowercase and the uppercase are used to distinguish between the two situations. Minitab doesn’t distinguish between the two, calling both measures “R-sq.”) The interpretation of is similar to that of , namely “ of the variation in the response is explained by the predictors in the regression model (which may be curvilinear).”

In summary, the value of 100% and the value of 0 tell the story of the second plot perfectly. The multiple coefficient of determination tells us that all of the variation in the response is explained in a curved manner by the predictor . The correlation coefficient tells us that of there is a relationship between and , it is not linear.

**Warning 2: A large value should not be interpreted as meaning that the estimated regression line fits the data well. Another function might better describe the trend in the data.**

Consider the following example in which the relationship between year (1790 to 1990, by decades) and population of the United States (in Millions) is examined:



**Pearson correlation of Year and USPopn = 0.959**

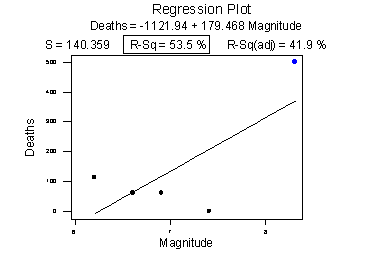
The correlation of 0.959 and the value of 92.0% suggests a strong linear relationship between year and U.S. population. Indeed, only 8% of the variation in U.S population is left to explain after taking into account the year in a linear way! Although, the plot suggests that a curve would describe the relationship even better. That is, the large value of 92% should not be interpreted as meaning that the estimated regression line fits the data well. (Its large value does suggest that taking into account year is better than not doing so. It just doesn’t tell us that we could still do better.)

Again the value doesn’t tell us that the regression model fits the data well. This is the most common misuse of the value! When you are reading the literature in your research area, pay close attention to how others interpret . I am confident that you will find some authors misinterpreting the value in this way. And, when you are analyzing your own data, make sure you plot the data – 99 times out of 100, the plot will tell you more of the story than a simple summary measure like or ever could.

Warning 3: The coefficient of determination and the correlation coefficient can both be greatly affected by just one data point (or a few data points). I.e. sensitive to outliers and extreme values.

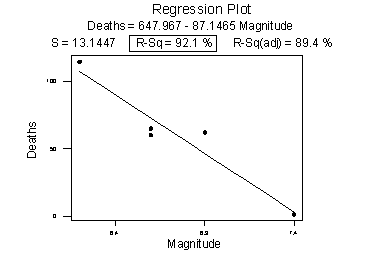
Consider the following example in which the relationship between the number of deaths in an earthquake and its magnitude are examined. Data on earthquakes were recorded, and the fitted line plot on the left was obtained. The slope of line and the correlation of 0.732 suggests that as the magnitude of the earthquake increases, the number of deaths also increases. This is not a surprising result. Therefore, if we hadn’t plotted the data, we wouldn’t notice that one and only one data point (magnitude = 8.3 and deaths = 503) was making the values of the slope and the correlation positive.

**Original Plot**



**Pearson correlation of Deaths and Magnitude = 0.732**

**Plot with unusual point removed**



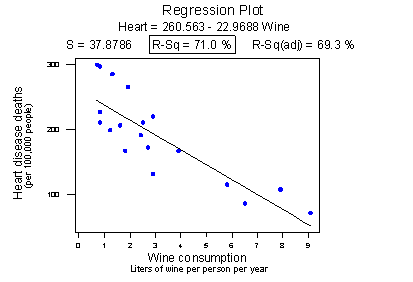
**Pearson correlation of Deaths and Magnitude = -0.960**

The second plot is a plot of the same data, but with the one unusual data point removed. Note that the estimated slope of the line changes from a positive 179.5 to a negative 87.1, just by removing one data point. Also, both measures of the strength of the linear relationship improve dramatically – changes from a positive 0.732 to a negative 0.960, and changes from 53.5% to 92.1%.

What conclusion can we draw from these data? Probably none! The main point of this example was to illustrate the impact of one data point on the and values. One could argue that a secondary point of the example is that a data set can be too small to draw any useful conclusions.

**Warning 4: Correlation (or association) does not imply causation.**

Consider the following example in which the relationship between wine consumption and death due to heart disease is examined. Each data point represents one country. For example, the data point in the lower right corner is France, where consumption averages 9.1 liters of wine per person per year and deaths due to heart disease are 71 per 100,000 people.



**Pearson correlation of Wine and Heart = - 0.843**

Minitab reports that the value is 71.0% and the correlation is -0.843. Based on these summary measures, a person might be tempted to conclude that he or she should drink more wine, since it reduces the risk of heart disease. If only life were that simple! Unfortunately, there may be other differences in the behavior of the people in the various countries that really explain the differences in the heart disease death rates, such as diet, exercise level, stress level, social support structure, and so on.

Let’s push this a little further. Recall the distinction between an experiment and an observational study:

* An **experiment** is a study in which, when collecting the data, the researcher controls the values of the predictor variables.
* An **observational study** is a study in which, when collecting the data, the researcher merely observes and records the values of the predictor variables as they happen.

The primary advantage of conducting experiments is that one can typically conclude that differences in the predictor values is what caused the changes in the response values. This is not the case for observational studies. Unfortunately, most data used in regression analysis arise from observational studies. Therefore, you should be careful not to overstate your conclusions, as well as be cognizant that others may be overstating their conclusions.

**Warning 5: Ecological correlations – correlations that are based on rates or averages – tend to overstate the strength of an association.**

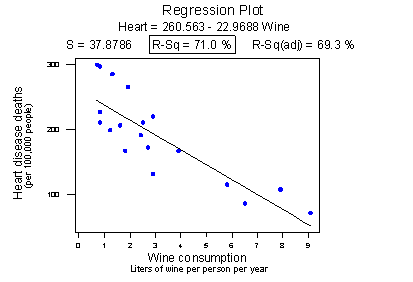
Some statisticians (Freedman, Pisani, Purves, 1997) investigated data from the 1988 Current Population Survey in order to illustrate the inflation that can occur in ecological correlations. Specifically, they consider the relationship between a man’s level of education and his income. They calculated the correlation between education and income in two ways:

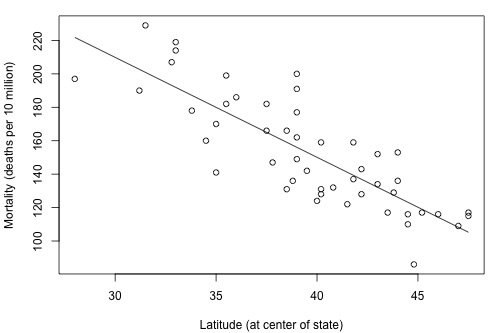
* First, they treated individual men, aged 25-64, as the experimental units. That is, each data point represented a man’s income and education level. Using these data, they determined that the correlation between income and education level for men aged 25-64 was about 0.4, not a convincingly strong relationship.
* The statisticians analyzed the data again, but in the second go-around they treated nine geographical regions as the units. That is, they first computed the average income and average education for men aged 25-64 in each of the nine regions. They determined that the correlation between the average income and average education for the sample of regions was about 0.7, obtaining a much larger correlation than that obtained on the individual data.

Again, ecological correlations, such as the one calculated on the region data, tend to overstate the strength of an association. How do you know what kind of data to use – aggregate data (such as the regional data) or individual data? It depends on the conclusion you’d like to make.

If you want to learn about the strength of the association between an individual’s education level and his income, then by all means you should use individual, not aggregate, data. On the other hand, if you want to learn about the strength of the association between a school’s average salary level and the school’s graduation rate, you should use aggregate data in which the units are the schools.

We hadn’t taken note of it at the time, but you’ve already seen a couple examples in which ecological correlations were calculated in aggregate data:





The correlation between wine consumption and heart disease deaths is 0.71 is an ecological correlation. The units are countries, not individuals. The correlation between skin cancer mortality and state latitude of 0.68 is also an ecological correlation. The units are states, again not individuals. In both cases, we should not use these correlations to try to draw a conclusion about how an individual’s wine consumption or sun tanning behavior will affect their individual risk of dying from heart disease or skin cancer. We shouldn’t try to draw any conclusions anyway, because “association is not causation.”

**Warning 6: A “statistically significant” value does not imply that the slope is meaningfully different from 0.**

This warning is a little strange as we haven’t talked about any hypothesis tests yet. We’ll get to that soon, but before doing so… a number of former students have asked why some article authors can claim that two variables are “statistically associated” with a -value less than 0.01, but their is small, such as 0.09 or 0.16. the answer to that has to do with the mantra that you may recall from your introductory statistics course: “**statistical significance does not imply practical significance.**”

In general, the larger the data set, the easier it is to reject the null hypothesis and claim “statistical significance.” If the data set is very large, it is even possible to reject the null hypothesis and claim that the slope is not 0, even when it is not practically or meaningfully different from 0. That is, it is possible to get a significant -value when is 0.13, a quantity that is likely not to be considered meaningfully different from 0 (of course, it does depend on the situation and the units). Again, the mantra is “statistical significance does not imply practical significance.”

**Warning 7: A large value does not necessarily mean that a useful prediction of the response or estimation of the mean response , can be made. It is still possible to get prediction intervals or confidence intervals that are too wide to be useful.**

We will learn more about such prediction and confidence intervals in Lesson 3.

## Hypothesis Test for the Population Correlation Coefficient

There is one more point we haven’t stressed yet in our discussion about the correlation coefficient and the coefficient of determination – namely, the two measures summarize the strength of a linear relationship *in samples only*. If we obtained a different sample, we would obtain different correlations, different values, and therefore potentially different conclusions. As always, we want to *draw conclusions about populations*, not just samples. To do so, we wither have to conduct a hypothesis test or a calculate a confidence interval. In this section, we learn how to conduct a hypothesis test for the population correlation coefficient (the Greek letter “rho”).

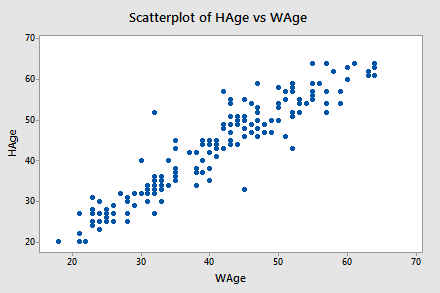
In general, a researcher should use the hypothesis test for the population coefficient to learn of a linear association between two variables, when it isn’t obvious which variable should be regarded as the response. Let’s clarify this point with examples of two different research questions.

Consider evaluating whether or not a linear relationship exists between skin cancer mortality and latitude. We will see in Lesson 2 that we can perform either of the following tests:

* -test for testing :
* ANOVA -test for testing :

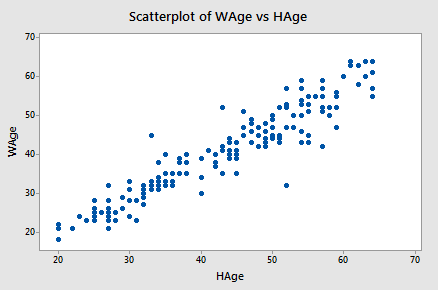
For this example it is fairly obvious that latitude should be treated as the predictor variable and skin cancer mortality as the response.

By contrast suppose we want to evaluate whether or not a linear relationship exists between a husband’s age and his wife’s age ([husbandwife.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/husbandwife.txt))? In this case, one could treat husband’s age as the response



**Pearson’s correlation of HAge and WAge = 0.939**

Or one could treat wife’s age as the response:



**Pearson’s correlation of WAge and HAge = 0.939**

In cases such as these, we answer our research questions concerning the existence of a linear relationship by using the -test for testing the population correlation coefficient : .

Let’s jump right to it! We follow standard hypothesis test proceedures in conducting a hypothese test for the population correalation coefficient . First, we specify the null and alternative hypotheses:

**Null Hypothesis: :**

**Alternative Hypothesis: : , or : or :**

Second, we calculate the value of the test statistic using the following formula:

Test statistic:

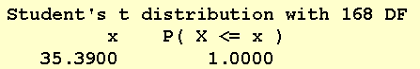
Third, we use the resulting statistic to calculate the -value. As always, the -value is the answer to the question “how likely is it that we’d get a test statistic as extreme as we did if the null hypothesis were true?” The -value is determined by referring to a -distribution with degrees of freedom.

Finally, we make a decision,

* If the -value is smaller than the significance level , we reject the null hypothesis in favor of the alternative. We conclude “there is sufficient evidence at the level to conclude that there is a linear relationship in the population between the predictor and the response .”
* If the -value is larger than the significance level , we fail to reject the null hypothesis. We conclude “there is not enough evidence at the level to conclude that there is a linear relationship in the population between the predictor and the response .”

Let’s perform the hypothesis test on the husband’s age and wife’s age data in which the sample correlation, based on couples, is . To test : against the alternative : , we obtain the following test statistic:

To obtain the -value, we need to compare the test statistic to a -distribution with 168 degrees of freedom (since 170 – 2 = 168). In particular, we need to find the probability that we’d observe a test statistic more extreme than 35.39, and then, since we-re conducting a two-sided test, multiply the probability by 2. Minitab helps us out here:



The output tells us that the probability of getting a test-statistic smaller than 35.39 is greater than 0.999. Therefore, the probability of getting a test-statistic greater than 35.39 is smaller than 0.001. Now we multiply by 2 and determine that the -value is less than 0.002. Since the -value is small – smaller than 0.05, say – we can reject the null hypothesis. There is sufficient statistical evidence at the level to conclude that there is a significant linear relationship between a husband’s age and his wife’s age.

Incidentally, we can let statistical software like Minitab do all of the dirty work for us. In doing so, Minitab reports:

minitab output

One final note… as always, we should clarify when it is okay to use the -test for testing : . The guidelines are a straightforward extension of the “LINE” assumptions made for the simple linear regression model. It’s okay:

* When it is not obvious which variable is the response
* When the pairs are a random sample from a bivariate normal population
  + For each , the ’s are normal with equal variances
  + For each , the ’s are normal with equal variances
  + Either can be considered a linear function of
  + Or can be considered a linear function of
* The pairs are independent

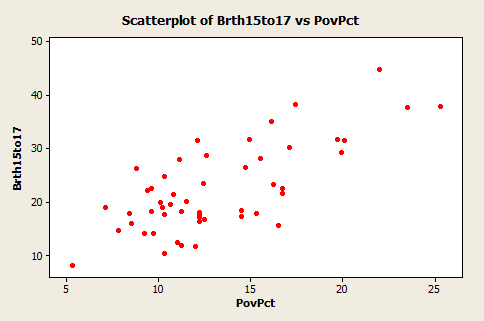
## Further Examples

**Example 1: Teen Birth Rate and Poverty Level Data**

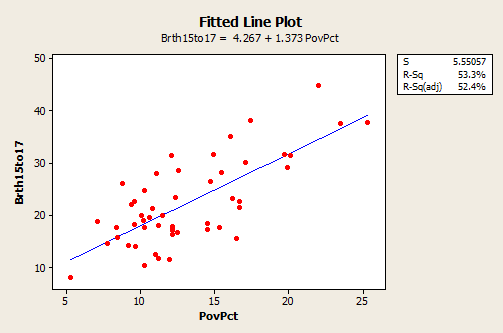
This dataset of size are for the 50 states and the District of Columbia in the United States ([poverty.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/examples/poverty.txt)). The variables are year 2002 birth rate per 1000 females 15 to 17 years old and poverty rate, which is the percent of the state’s population living in households with incomes below the federally defined poverty level. (Data source: Mind on Statistics, 3rd Edition, Utts and Heckard).



The plot of the data below (birth rate on the vertical) shows a generally linear relationship, on average, with a positive slope. As the poverty level increases, the birth rate for 15 to 17-year-old females tends to increase as well.



The figure below, created in Minitab using *Stat >> Regression >> Fitted Line Plot*, shows a regression line superimposed on the data. The equation is given near the top of the plot. Minitab should have written that the equation is for the “average” birth rate (or “Predicted” birth rate would be okay too) because a regression equation describes the average value of as a function of one or more -variables. In statistical notation, the equation could be written .



* The interpretation of the slope (value = 1.373) is that the 15 to 17-year-old birth rate increase 1.373 units, on average, for each one unit (one percent) increase in the poverty rate.
* The interpretation of the intercept (value = 4.267) is that if there were states with a poverty rate = 0, the predicted average for 15 to 17-year-old birth rate would be 4.267 for those states. Since there are no states with a poverty rate = 0, this interpretation of the intercept is not practically meaningful for this example.

In the graph with the regression line present, we also see the information that and .

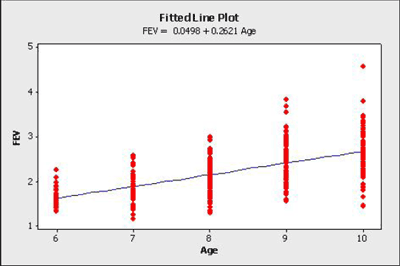
* The value of tells us roughly the average difference between the -values of individual observations and predictions of based on the regression line.
* The value of can be interpreted to mean that poverty rates “explain” 53.3% of the observed variation in the 15 to 17-year-old average birth rates of the states

The (adj) value (52.4%) is an adjustment to based on the number of -variables in the model (only one here) and the sample size. With only one -variable, the adjusted is not that important.

**Example 2: Lung Function in 6 to 10 Year Old Children**

The data are from children between 6 and 10 years old. The variables aare = Forced Exhalation Volume (FEV), a measure of how much air somebody can forcibly exhale from their lungs, and = age in years. (Data source: The data here are a part of a dataset given in Kahn, Michael (2005). “[An Exhalent Problem for Teaching Statistics](http://www.amstat.org/publications/jse/v13n2/datasets.kahn.html)”, *The Journal of Statistical Education*, 13(2).

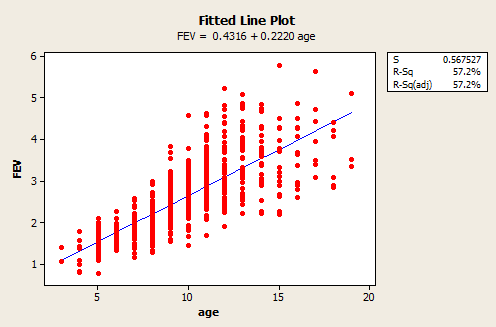
Below is a plot of the data with a simple linear regression line superimposed. The plot was done in Minitab and as pointed out earlier, the word “average” should come before the -variable name.



* The estimated regression equation is that average FEV = 0.0498 + 0.2621 \* age. For instance, for an 8-year-old we can use the equation to estimate that the average FEV = 0.0498 + 0.2621 \* (8) = 2.1466.
* The interpretation of the slope is that the average FEV increases 0.2621 for each one year increase in age (in the observed age range)

An interesting and possibly important feature of these data is that the variance of individual -values from the regression line incrases as age increases. This feature of data Is called non-constant variance. For example, the FEV values of 10 year olds are more variable than FEX value of 6 year olds. This is seen by looking at the vertical ranges of the data in the plot. This maul lead to problems using a simple linear regression model for these datam which is an issue we’ll explore in more detail in lesson 4.

Above, we only analyzed a subset of the entire dataset. The full dataset ([fev\_dat.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/examples/fev_dat.txt)) is shown in the plot below:



As we can see, the range of ages now spans 3 to 19-years-old and the estimated regression equation is FEV = 0.4316 + 0.2220 \* age. Both the slope and intercept have noticeably changed, but the variance still appears to be non-constant. This illustrates that it is important to be aware of how you are analyzing your data. If you only use a subset of your data that spans a shorter range of predictor values, then you could obtain noticeably different results than if you had used the full dataset.

Quick reference: , ,

1. See Econometrics [↑](#footnote-ref-1)
2. See Methodology of econometrics [↑](#footnote-ref-2)