# 1.1 – What is Simple Linear Regression?

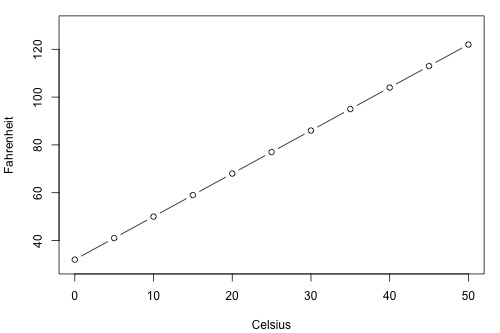
**Simple linear regression** is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables:

* One variable, denoted *x*, is regarded as the **predictor**, **explanatory**, or **independent** variable.
* The other variable, denoted *y*, is regarded as the **response**, **outcome**, or **dependent** variable.

Because the other terms are used less frequently today, we'll use the "**predictor**" and "**response**" terms to refer to the variables encountered in this course. The other terms are mentioned only to make you aware of them should you encounter them in other arenas. Simple linear regression gets its adjective "simple," because it concerns the study of only one predictor variable. In contrast, multiple linear regression, which we study later in this course, gets its adjective "multiple," because it concerns the study of two or more predictor variables.

## Types of relationships

Before proceeding, we must clarify what types of relationships we won't study in this course, namely, **deterministic** (or **functional**) **relationships**. Here is an example of a deterministic relationship.



Note that the observed (*x*, *y*) data points fall directly on a line. As you may remember, the relationship between degrees Fahrenheit and degrees Celsius is known to be:

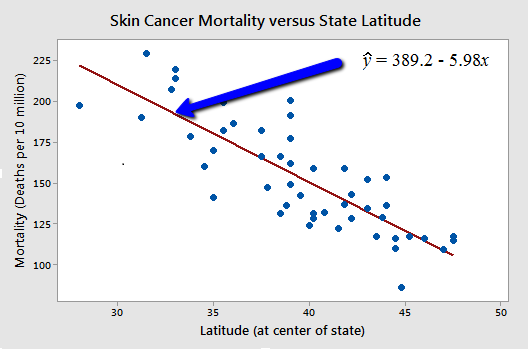
That is, if you know the temperature in degrees Celsius, you can use this equation to determine the temperature in degrees Fahrenheit *exactly*.

Here are some examples of other deterministic relationships that students from previous semesters have shared:

* Circumference = π × diameter
* Hooke's Law: *Y* = α + *βX*, where *Y* = amount of stretch in a spring, and *X* = applied weight.
* Ohm's Law: *I* = *V*/*r*, where *V* = voltage applied, *r* = resistance, and *I* = current.
* Boyle's Law: For a constant temperature, *P* = α/*V*, where *P* = pressure, α = constant for each gas, and *V* = volume of gas.

For each of these deterministic relationships, the equation *exactly* describes the relationship between the two variables. This course does not examine deterministic relationships. Instead, we are interested in **statistical relationships**, in which the relationship between the variables is not perfect.

Here is an example of a statistical relationship. The response variable *y* is the mortality due to skin cancer (number of deaths per 10 million people) and the predictor variable *x* is the latitude (degrees North) at the center of each of 49 states in the U.S. ([skincancer.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/skincancer/index.txt)) (The data were compiled in the 1950s, so Alaska and Hawaii were not yet states. And, Washington, D.C. is included in the data set even though it is not technically a state.)



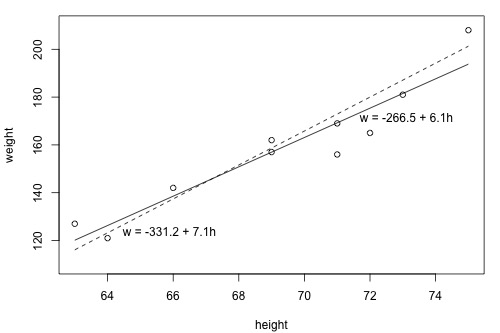
You might anticipate that if you lived in the higher latitudes of the northern U.S., the less exposed you'd be to the harmful rays of the sun, and therefore, the less risk you'd have of death due to skin cancer. The scatter plot supports such a hypothesis. There appears to be a negative linear relationship between latitude and mortality due to skin cancer, but the relationship is not perfect. Indeed, the plot exhibits some "**trend**," but it also exhibits some "**scatter**." Therefore, it is a statistical relationship, not a deterministic one.

Some other examples of statistical relationships might include:

* Height and weight — as height increases, you'd expect weight to increase, but not perfectly.
* Alcohol consumed and blood alcohol content — as alcohol consumption increases, you'd expect one's blood alcohol content to increase, but not perfectly.
* Vital lung capacity and pack-years of smoking — as amount of smoking increases (as quantified by the number of pack-years of smoking), you'd expect lung function (as quantified by vital lung capacity) to decrease, but not perfectly.
* Driving speed and gas mileage — as driving speed increases, you'd expect gas mileage to decrease, but not perfectly.Navigation

# 1.2 - What is the "Best Fitting Line"?

Since we are interested in summarizing the trend between two quantitative variables, the natural question arises — "what is the best fitting line?" At some point in your education, you were probably shown a scatter plot of (*x*, *y*) data and were asked to draw the "most appropriate" line through the data. Even if you weren't, you can try it now on a set of heights (*x*) and weights (*y*) of 10 students, ([student\_height\_weight.txt)](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/student_height_weight/index.txt).   Looking at the plot below, which line — the solid line or the dashed line — do you think best summarizes the trend between height and weight?



Hold on to your answer! To examine which of the two lines is a better fit, we first need to introduce some common notation:

* yiyi denotes the observed response for experimental unit *i*
* xixi denotes the predictor value for experimental unit *i*
* ^yiy^i is the predicted response (or fitted value) for experimental unit *i*

Then, the equation for the best fitting line is:

^yi=b0+b1xiy^i=b0+b1xi

Incidentally, recall that an "**experimental unit**" is the object or person on which the measurement is made. In our height and weight example, the experimental units are students.

Let's try out the notation on our example with the trend summarized by the line *w* = -266.53 + 6.1376 *h*. (Note that this line is just a more precise version of the above solid line, *w* = -266.5 + 6.1 *h*.) The first data point in the list indicates that student 1 is 63 inches tall and weighs 127 pounds. That is, *x*1 = 63 and *y*1 = 127. 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, ^y1y^1 = 120.1. Clearly, our prediction wouldn't be perfectly correct — it has some "**prediction error**" (or "**residual error**"). In fact, the size of its prediction error is 127-120.1 or 6.9 pounds.

You might want to roll your cursor over each of the 10 data points to make sure you understand the notation used to keep track of the predictor values, the observed responses and the predicted responses:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  | | --- | --- | --- | --- | | ***i*** | xixi | yiyi | ^yiy^i | | 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 5, the equation of the line would predict his or her weight to be -266.53 + 6.1376(69) or 157 pounds. The size of the prediction error here is 162-157, or 5 pounds.

In general, when we use ^yi=b0+b1xiy^i=b0+b1xi to predict the actual response *yi*, we make a prediction error (or residual error) of size:

ei=yi−^yiei=yi−y^i

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: ^yi=b0+b1xiy^i=b0+b1xi
* We just need to find the values *b*0 and *b*1 that make the sum of the squared prediction errors the smallest it can be.
* That is, we need to find the values *b*0 and *b*1 that minimize:

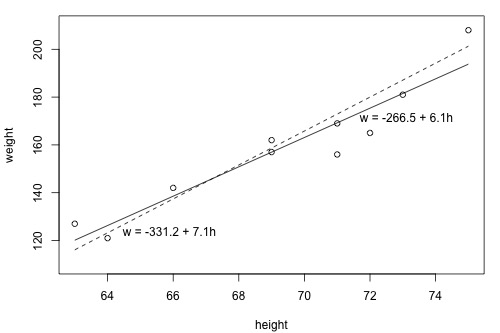
Q=n∑i=1(yi−^yi)2Q=∑i=1n(yi−y^i)2

Here's how you might think about this quantity *Q*:

* The quantity ei=yi−^yiei=yi−y^i is the prediction error for data point *i*.
* The quantity e2i=(yi−^yi)2ei2=(yi−y^i)2 is the squared prediction error for data point *i*.
* And, the symbol ∑ni=1∑i=1n tells us to add up the squared prediction errors for all *n* data points.

Incidentally, if we didn't square the prediction error ei=yi−^yiei=yi−y^i to get e2i=(yi−^yi)2ei2=(yi−y^i)2, 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. Considering the least squares criterion, which line do you now 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 dashed line and the solid line.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | ***w* = -331.2 + 7.1 *h*** (the dashed line) | | | | | | | ***i*** | xixi | yiyi | ^yiy^i | (yi−^yi)(yi−y^i) | (yi−^yi)2(yi−y^i)2 | | 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** | | |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | ***w* = -266.53 + 6.1376 *h*** (the solid line) | | | | | | | ***i*** | xixi | yiyi | ^yiy^i | (yi−^yi) (yi−y^i) | (yi−^yi)2(yi−y^i)2 | | 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 least 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, *w* = -266.53 + 6.1376h, best summarizes the data. But, is this equation guaranteed to be the best fitting line of all the possible lines we didn't even 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'd have our work cut out for us. We'd have to implement the above procedure for an infinite number of possible lines — clearly, an impossible task! Fortunately, somebody has done some dirty work for us by figuring out formulas for the **intercept** *b*0 and the **slope** *b*1 for the equation of the line that minimizes the sum of the squared prediction errors.

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

Q=n∑i=1(yi−(b0+b1xi))2Q=∑i=1n(yi−(b0+b1xi))2

(that is, take the derivative with respect to *b*0 and *b*1, set to 0, and solve for *b*0 and *b*1) and get the "**least squares estimates**" for *b*0 and *b*1:

b0=¯y−b1¯xb0=y¯−b1x¯

and:

 b1=∑ni=1(xi−¯x) (yi−¯y) ∑ni=1(xi−¯x)2b1=∑i=1n(xi−x¯) (yi−y¯) ∑i=1n(xi−x¯)2

Because the formulas for *b*0 and *b*1 are derived using the least squares criterion, the resulting equation — ^yi=b0+b1xiy^i=b0+b1xi — is often referred 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 (¯x, ¯y) (x¯, y¯), since when x=¯xx=x¯, then y=b0+b1¯x=¯y−b1¯x+b1¯x=¯yy=b0+b1x¯=y¯−b1x¯+b1x¯=y¯.

In practice, you won't really need to worry about the formulas for *b*0 and *b*1. Instead, you are being going to let statistical software, such as Minitab, find least squares lines for you. But, we can still learn something from the formulas — for *b*1.

If you study the formula for the slope *b*1:

b1=∑in=1(xi−¯x) (yi−¯y) ∑ni=1(xi−¯x)2b1=∑i=1n(xi−x¯) (yi−y¯) ∑i=1n(xi−x¯)2

you see that the denominator is necessarily positive since it only involves summing positive terms. Therefore, the sign of the slope *b*1 is solely determined by the numerator. The numerator tells us, for each data point, to sum up the product of two distances — 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 the *y*-values. Let's see how this determines the sign of the slope *b*1 by studying the following two plots.

**When is the slope *b*1 > 0?** Do you agree that the trend in the following plot is positive — that is, as *x* increases, *y* tends to increase? If the trend is positive, then the slope *b*1 must be positive. Let's see how!

* Click on 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.
* Now, select clear and then click on the blue data point in the lower left quadrant......... Note that the product of the two distances for this data point is also positive. In fact, the product of the two distances is positive for *any* data point in the lower left quadrant.

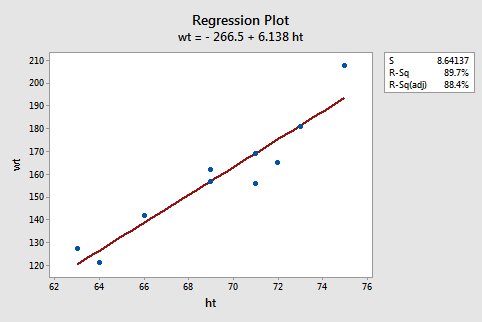
Adding up all these positive products must necessarily yield a positive number, and hence the slope of the line *b*1 will be positive.

**When is the slope *b*1 < 0?** Now, do you 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 *b*1 must be negative. Let's see how!

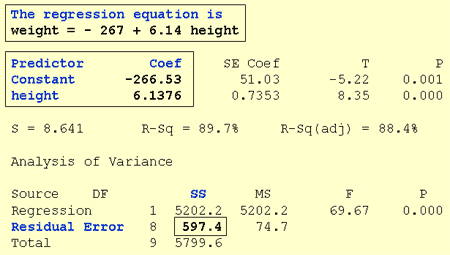
* Click on the blue data point in the upper left quadrant......... Note that the product of the two distances for this data point is negative. In fact, the product of the two distances is negative for *any* data point in the upper left quadrant.
* Now, select clear and then click on the blue data point in the lower right quadrant......... Note that the product of the two distances for this data point is also negative. In fact, the product of the two distances is negative for *any* data point in the lower right quadrant.

Adding up all these negative products must necessarily yield a negative number, and hence the slope of the line *b*1 will be negative.

Now that we finished that investigation, you can just set aside the formulas for *b*0 and *b*1. Again, in practice, you are going to let statistical software, such as Minitab, find 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 (in box) on Minitab's "**fitted line plot**."



The following Minitab output illustrates where you can find the least squares line (in box) in Minitab's "**standard regression analysis**" output.



Note that the estimated values *b*0 and *b*1 also appear in a table under the columns labeled "**Predictor**" (the intercept *b*0 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**" *b*0 and *b*1, 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. | ^yi, wt=−267+6.14xi, hty^i, wt=−267+6.14xi, ht |
| Predict (mean) weight of 66"-inch tall people. | ^yi, wt=−267+6.14(66) =138.24y^i, wt=−267+6.14(66) =138.24 |
| Predict (mean) weight of 67"-inch tall people. | ^yi, wt=−267+6.14(67) =144.38y^i, wt=−267+6.14(67) =144.38 |

**Now, what does *b*0 tell us?** The answer is obvious when you evaluate the estimated regression equation at *x* = 0. 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 *x* = 0. So, here the intercept *b*0 is not meaningful. In general, if the "scope of the model" includes *x* = 0, then *b*0 is the predicted mean response when *x* = 0. Otherwise, *b*0 is not meaningful. There is more information on this [here](https://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-to-interpret-the-constant-y-intercept).

**And, what does *b*1 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 144.38 - 138.24 = 6.14 pounds -- the value of *b*1. Here, it tells us that we predict the mean weight to increase by 6.14 pounds for every additional one-inch increase in height. In general, we can expect the mean response to increase or decrease by *b*1 units for everyone unit increase in *x.*

Navigation