A Quick Study of Linear Regressions

Aaron Wong

May 2, 2020

This document gives a brief overview of the calculation of two variable linear regression. The derivation will be worked out from first principles and basic ideas, and a step-by-step computational guide will also be presented. (This document is functionally a test case for using R to create documents that are a mix of R and LaTeX.)

An Example Using R

We will begin by just running an example through R. This will generate information that the rest of the document will explain the derivation and interpretation of this information. We will begin by creating the data set.

```
# Data
X = c(3, 5, 8, 10, 17)
Y = c(1, 8, 3, 14, 10)
```

This data set consists of the following ordered pairs:

X	Y
3	1
5	8
8	3
10	14
17	10

We can now have R determine the linear model that best fits this data.

```
# Create the linear model
regression = lm(Y ~ X)
summary(regression)
```

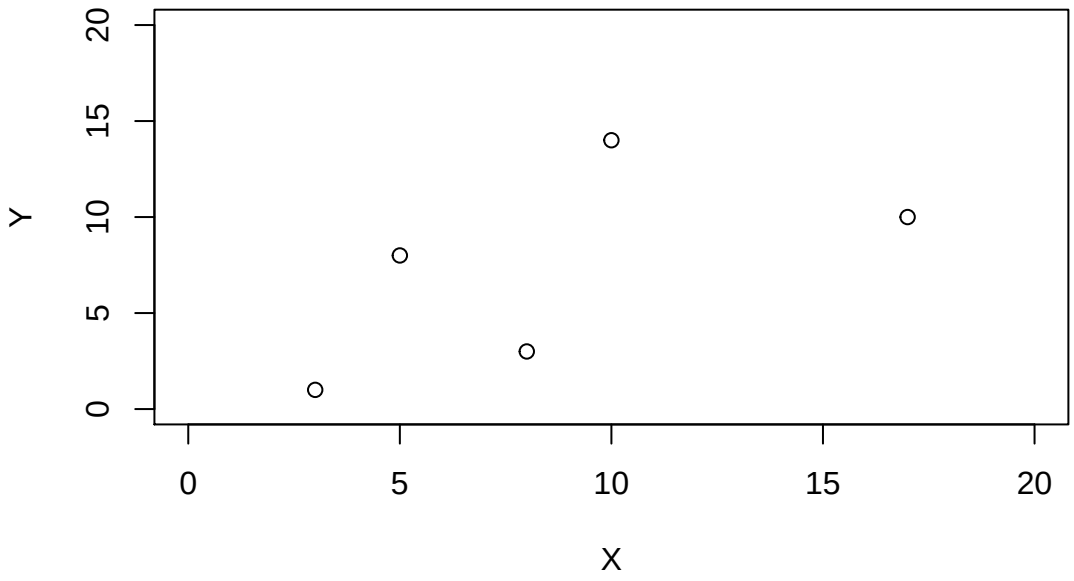
```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##      1      2      3      4      5
## -2.980  2.870 -3.855  5.995 -2.031
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2543     4.4672   0.505    0.649
## X             0.5751     0.4526   1.270    0.293
##
## Residual standard error: 4.9 on 3 degrees of freedom
## Multiple R-squared:  0.3498, Adjusted R-squared:  0.1331
## F-statistic: 1.614 on 1 and 3 DF,  p-value: 0.2935
```

The challenge is that this information requires interpretation. What does all of this information mean? And what other information might be helpful? We're going to explore all of this information.

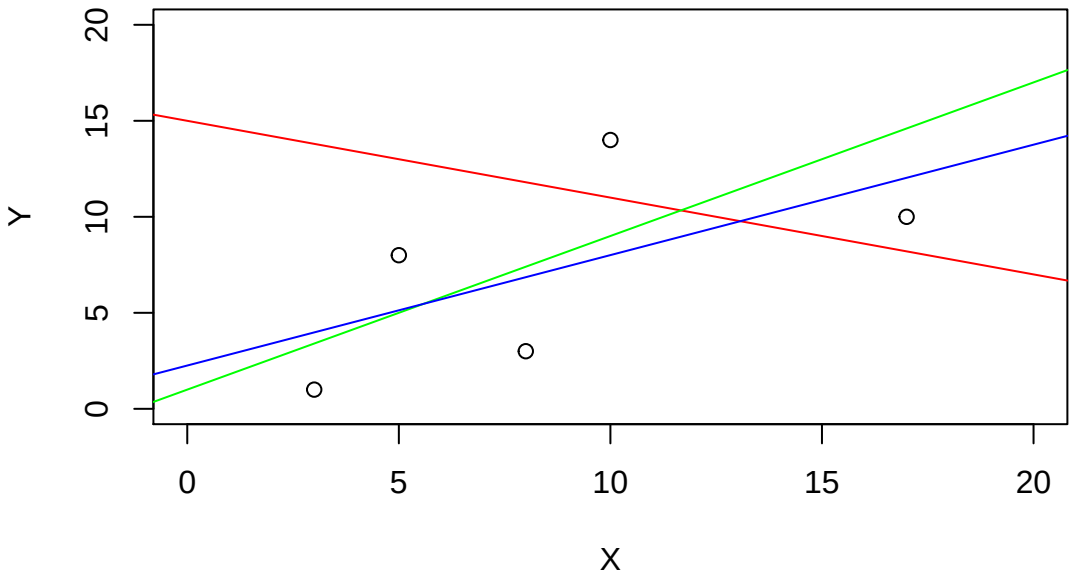
Linear Model

The command `regression = lm(Y ~ X)` creates a linear model called `regression`. The goal of this model is to come up with an equation of the form $Y' = a + bX$ that is the "best fit" for the data stored in the variables X and Y. (Note: Y' is not a derivative in this notation.) Let's start by looking at a graph of the data.

```
# Plot the data with the linear model
plot(X,Y,
     xlim = c(0,20), ylim = c(0,20))
```



We can see from this graph that there is a general lower-left to upper-right behavior of the data. As one variable gets larger, the other variable tends to get larger as well. It is not possible to connect all the dots with a single line, so if we tried to approximate this data with a line, there will always be some type of error. But some lines do a better job than others. In the following graph, the blue and green lines do a better job at matching the behavior as the red line. But between the blue and green line, it's not so obvious which one is better.



Residuals

Before we can even begin to discuss the "best" fit, we need to devise a method for measuring how good a fit actually is. We will do this by establishing an error value for each data point relative to the line. Suppose we have an approximating function $Y' = f(X)$. Then the residual for each point (X_i, Y_i) is defined as the (signed) distance between the point and its approximation (X_i, Y_i'). This graph shows the coordinates of all the points (black), all of the approximations (blue), and all of the residuals (red).

Least Squares

For every line that we can draw (in fact, every function we can imagine), we will be able to compute these residuals. Intuitively, we might imagine that the "total error" we might use would be the sum of the absolute values of the residuals. And while this does give us a value that can be used (referred to as the *absolute deviation*), we will actually use the sum of the squares of the residuals. There are several reasons for this choice, but for our purposes we're going to use the fact that the absolute value function is not differentiable. The reason for this is that our goal will be to find the line that minimizes this error, giving us the *least squares* method.

Deviation Scores

For reasons that will become more clear in the derivation of the least squares solution, it is helpful to define a new set of variables. Let M_X and M_Y be the means of the variables X and Y and let $x_i = X_i - M_X$ and $y_i = Y_i - M_Y$. These are called the *deviation scores* and represent a "re-centering" of the data so that M_x and M_y (the means of x and y) are both zero.

Least Squares Solution

We will now formalize the idea of the least squares method and apply it to the deviation scores. Suppose we have a set of points (x_i, y_i) (for $i = 1, 2, \ldots, N$) and let $\mathcal{F}$ be some set of functions. We define the function $\text{SqE} : \mathcal{F} \to \mathbb{R}$ by $\text{SqE}(f) = \sum_{i=1}^{N}(y_i - y_i')^2$, where $f \in \mathcal{F}$ and $y_i' = f(x_i)$. Also, the only sum that will be in summation notation will be over the index of points, and so we will be lazy and drop that part of the notation. The goal is to identify the function f that minimizes the value of the function SqE.

Since we are doing a linear regression, we will let $\mathcal{F}$ be the set of all linear functions of the form $f(x) = a + bx$. Notice that if we treat a and b as real-valued continuous parameters, we have that

$$\begin{aligned}\frac{\partial}{\partial a}\,\text{SqE} &= \frac{\partial}{\partial a}\left(\sum(y_i - y_i')^2\right) = \sum 2(y_i - y_i') \cdot \frac{\partial}{\partial a}(-y_i') \\ &= -2\sum(y_i - y_i')\frac{\partial}{\partial a}(a + bx_i) = -2\sum(y_i - y_i')\end{aligned}$$

and

$$\begin{aligned}\frac{\partial}{\partial b}\,\text{SqE} &= \frac{\partial}{\partial b}\left(\sum(y_i - y_i')^2\right) = \sum 2(y_i - y_i') \cdot \frac{\partial}{\partial b}(-y_i') \\ &= -2\sum(y_i - y_i')\frac{\partial}{\partial b}(a + bx_i) = -2\sum(y_i - y_i') \cdot x_i\end{aligned}$$

By setting $\frac{\partial}{\partial a}\,\text{SqE} = 0$, we get that $\sum y_i' = \sum y$, which is equivalent to $Na + b\sum x_i = \sum y_i$, or $a + bM_x = M_y$. But we've chose the variables x and y so that their means are zero, and so this implies that $a = 0$.

Next, we will set $\frac{\partial}{\partial b}\,\text{SqE} = 0$. We can see that this leads to $\sum x_i y_i' = \sum x_i y_i$, which can be written as $a\sum x_i + b\sum x_i^2 = \sum x_i y_i$. Since $a = 0$, this implies that $b = \frac{\sum x_i y_i}{\sum x_i^2}$. It turns out to be convenient to write this as $b = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}} \cdot \sqrt{\frac{\sum y_i^2}{\sum x_i^2}} = r\frac{s_y}{s_x}$, where s_x and s_y are the standard deivations of the variables x and y and r is the Pearson correlation coefficient (discussed below). We have implicitly used the population standard deviation here (s instead of σ), but the whole derivation can be done with an entire population with no significant changes.

This shows us that the line of best fit for the deviation scores is given by $y = r\frac{s_y}{s_x}x$. We can substitute back to the orginal variables (using the fact that a translation of variables does not affect the standard deviation) to get $Y - M_Y = r\frac{s_Y}{s_X}(X - M_X)$, which can be rewritten as $Y = \left(r \cdot \frac{s_Y}{s_X}\right)x + (M_Y - rM_X \cdot \frac{s_Y}{s_X})$.

The Pearson Correlation Coefficient

The Pearson correlation ceofficient is a measure of how well (or how poorly) the data is fit by a line. It is defined by $r = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$. (Reminder: We are still using the deviation scores in this section.) Here are some of the properties of the Pearson correlation coefficient:

- $|r| \leq 1$
- The sign of r indicates whether the slope of the line of best fit is positive or negative.
- The magnitude of r indicates how well the line of best fit approximates the data. If $|r| = 1$ then the points lie exactly on a line. If $r = 0$ then the data does not match a line at all (or all the data is on a horizontal line, but that doesn't happen in practice).

A Computational Method

Although all of the calculations are laid out in the derivations give the formulas for calculating the least squares fit of the data, it is helpful to have all of these calculations organized into a computational method that is easy to follow. Here are the charts that perform that organizational role:

	X	Y	x	y	x^2	y^2	xy
1	3	1					
2	5	8					
3	8	3					
4	10	14					
5	17	10					
Σ							
M							

	X	Y	Y'	$Y - Y'$	$(Y - Y')^2$
1	3	1			
2	5	8			
3	8	3			
4	10	14			
5	17	10			
Σ					

Here are the steps for the first chart:

- The initial data is filled in (as shown).
- The Σ row is for the sum of the values in the column.
- The M row is for the mean of the column.
- The formulas for x and y are $x = X - M_X$ and $y = Y - M_Y$. If calculated correctly, the Σ and M values for these columns should be 0.
- The remaining columns are calculated as described at the top of the columns. The Σ row is used to calculate the Pearson correlation coefficient. The means of the x^2 and y^2 give the standard deviations if the data represents a full population, otherwise you would need to divide by $N - 1$. Alternatively, those spaces can simply be left blank and calculate the standard deviation somewhere else.

This is the completed version of the chart:

	X	Y	x	y	x^2	y^2	xy
1	3	1	−5.600	−6.200	31.360	38.440	34.720
2	5	8	−3.600	0.800	12.960	0.640	−2.880
3	8	3	−0.600	−4.200	0.360	17.640	2.520
4	10	14	1.400	6.800	1.960	46.240	9.520
5	17	10	8.400	2.800	70.560	7.840	23.520
Σ	43	36	0.000	−0.000	117.200	110.800	67.400
M	8.6	7.2					

From this chart, we can use the values at the bottom right to compute r:

$$r = \frac{\sum x_i, y_i}{\sqrt{\sum x_i^2 \sum y_i^2}} = \frac{67.4}{\sqrt{(117.2) \cdot (110.8)}} = 0.5914611$$

We can also calculate the sample variance and sample standard deviation for both X and Y:

$$s_X^2 = \frac{\sum x_i^2}{N-1} = \frac{117.2}{5-1} = 29.3 \qquad s_Y^2 = \frac{\sum y_i^2}{N-1} = \frac{110.8}{5-1} = 27.7$$

$$s_X = \sqrt{s_X^2} = 5.4129474 \qquad s_Y = \sqrt{s_Y^2} = 5.2630789$$

From this, we can use the formulas we obtained earlier to calculate the coefficients of the line of best fit:

$$r \cdot \frac{s_Y}{s_X} = 0.5914611 \cdot \frac{5.2630789}{5.4129474} = 0.5750853$$
$$M_Y - rM_X \cdot \frac{s_Y}{S_X} = 7.2 - (0.5914611) \cdot (8.6) \cdot \frac{5.2630789}{5.4129474} = 2.2542662$$

So $Y = \left(r \cdot \frac{s_Y}{s_X}\right) x + (M_Y - rM_X \cdot \frac{s_Y}{s_X}) = 0.5750853X + 2.2542662$. We can use this to proceed to the second chart:

- The Y' column is simply the values obtained by plugging X into the formula. The sum of these values should be the same as the sum of the values in the Y column. The reason for this comes out of hte derivation of the least squares solution when we set the derivative to zero to find the minimum error. This is a good self-check.
- The other two columns are calculations based on the formulas at the top of the chart. The sum at the bottom of the first column should be zero, and this is another self-check moment. (This value should be the difference of columns of the previous two values, which have the same sum.)

Here is the completed version of the second chart:

	X	Y	Y'	$Y - Y'$	$(Y - Y')^2$
1	3	1	3.980	−2.980	8.878
2	5	8	5.130	2.870	8.239
3	8	3	6.855	−3.855	14.861
4	10	14	8.005	5.995	35.939
5	17	10	12.031	−2.031	4.124
Σ	43	36	36.000	0.000	72.039

Comparisons with R's Model

The output from the R summary of the linear model contains a lot of information. This section is just a summary of how that information matches up with the values we've discussed.

- We calculated the `Residuals` in the $Y - Y'$ column of the second chart.
- The `Coefficients` are the coefficients of the line of best fit, which we calculated in preparation to complete the second chart.
- `Multiple R-squared` value is the square of the Pearson correlation coefficient: $0.5914611^2 = 0.3498263$.

There are some other values in the summary table, and those values are related to attempting to determine the statistical significance of the estimates. I may add to this document later to cover those parts.