

Math107: Linear Regression

Dr. Richard Mikula

Spring 2010

Regression Analysis:

Suppose that one is studying an experiment where there are two random variables X and Y under study, and one wishes to determine if there is a relationship between the ordered pairs $(X(\text{outcome}), Y(\text{outcome}))$.

Supposing that we are allowed to freely choose x , we wish to determine if we can express y

$$Y(\textit{outcome}) = y = \alpha x + \beta + \epsilon(\textit{outcome})$$

where $X(\textit{outcome}) = x$, α, β are constants and ϵ is a random variable* with mean value 0 and standard deviation σ . Moreover, the error terms ϵ are independent, and each value of x determines a population of y -values. Thus the expected value is given by

$$E(Y) = \alpha x + \beta$$

for any fixed x value.

*Which we usually assume is Normally Distributed – this we shall discuss later.

Least Squares Regression:

To determine the α and the β for the two random variables X and Y , we often select a sample of data values for n observations from our population

$$O_1, O_2, O_3, \dots, O_n$$

and let $X(O_i) = x_i$ and $Y(O_i) = y_i$. This gives rise to n ordered pairs

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n).$$

For simplicity, we suppose that we have a set of points

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

with all the x'_i s distinct values. We wish to find the so-called **line of best-fit**, or the **trend line**, and let

$$y = mx + b$$

be this line. The plotted points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ in the xy -plane are called the **scatter plot** of the data set.

This line will be the line that is as close to all the points

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

on a whole, as possible.

Without loss of generality, we assume

$$x_1 < x_2 < \cdots < x_n.$$

Let

$$SSE = \sum_{i=1}^n (mx_i + b - y_i)^2.$$

Then SSE , which is always greater than or equal to zero, measures the error in using $\hat{y}_i = mx_i + b$ in stead of y_i . If $SSE = 0$ then we know $y_i = mx_i + b = \hat{y}_i$ for all i . In all practicality, for more than two points, $SSE > 0$ will almost certainly be the case, so we seek the line which makes SSE as small as possible.

It turns out that to determine the m, b so that the quantity SSE given by

$$SSE = \sum_{i=1}^n (\hat{y}_i - y_i)^2,$$

is as small as possible we need multi-variable calculus. It turns out that m is given by

$$m = \frac{n\bar{x}\bar{y} - \sum_{i=1}^n x_i y_i}{n\bar{x}^2 - \sum_{i=1}^n x_i^2},$$

and b is given by

$$b = \bar{y} - m\bar{x}.$$

Note that the values m, b should approximate the parameters α, β discussed earlier. Moreover, the unbiased estimator of σ , the standard deviation of the error term ϵ is given by

$$s_e = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2}}$$

or simply

$$s_e = \sqrt{\frac{SSE}{n - 2}}.$$

Before we do any explicit examples, it turns out that we will rewrite the formulae given for m, b . We seek formulae that may seem a bit easier to remember.

Let

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}),$$

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$$

and

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2.$$

One can actually show that

$$m = \frac{S_{xy}}{S_{xx}}, \quad b = \bar{y} - m\bar{x},$$

$$SSE = S_{yy} - \frac{S_{xy}^2}{S_{xx}}.$$

Example: For the points

$$(1, 3), (2, 5), (3, 5), (6, 7)$$

find the line of best fit.

Solution:

Step 1: Draw the so-called **scatter plot**, which is the data points given plotted in the Cartesian plane (x, y -plane).

Step 2: We find \bar{x} , the average of the x -coordinates of the 4 points, and \bar{y} , the average of the y -coordinates of the 4 points.

$$\bar{x} = \frac{1 + 2 + 3 + 6}{4} = 3,$$

and

$$\bar{y} = \frac{3 + 5 + 5 + 7}{4} = 5.$$

Step 3: We fill out the following table:

x	y	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$
1	3	$(1 - 3)(3 - 5)$	$(1 - 3)^2$
2	5	$(2 - 3)(5 - 5)$	$(2 - 3)^2$
3	5	$(3 - 3)(5 - 5)$	$(3 - 3)^2$
6	7	$(6 - 3)(7 - 5)$	$(6 - 3)^2$

Performing the calculations in this table yields:

x	y	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$
1	3	4	4
2	5	0	1
3	5	0	0
6	7	6	9

Step 4: Next you sum the right two columns – that is the one under the $(x - \bar{x})(y - \bar{y})$ and $(x - \bar{x})^2$ – to get S_{xy} and S_{xx} respectively, that is:

$$S_{xy} = 4 + 0 + 0 + 6 = 10$$

and

$$S_{xx} = 4 + 1 + 0 + 9 = 14.$$

Step 5: Thus we compute $m = \frac{S_{xy}}{S_{xx}}$ and then $b = \bar{y} - m\bar{x}$. Then use these in $y = mx + b$ and graph this line on the drawing with the data points (the scatter plot):

Thus

$$m = \frac{10}{14} = \frac{5}{7}$$

and

$$b = 5 - \frac{5}{7} \cdot 3 = \frac{20}{7}.$$

Therefore, the line of best-fit is given by

$$y = \frac{5}{7}x + \frac{20}{7}.$$

Finally we graph this line.

Home Work: Find the line of best fit for the points

$$(1, 1), (2, 0), (3, 3), (5, 8).$$

Answer: $y = \frac{68}{35}x - \frac{82}{35}$.

Example: Suppose we wanted to see if there was a connection between hours of study for an exam and how well a student has done on an exam. Suppose that in a class of ten students, we found out the following:

Student	Hours of Study	Exam Score
1	3	57
2	6	87
3	4	50
4	5	77
5	0	45
6	1	48
7	7	95
8	3.5	63
9	8	100
10	2.5	70

To begin with, we need to pick our variable. Let x be the hours studied and y the exam score. Then we have the following data set:

$$(0, 45), (1, 48), (2.5, 70), (3, 57), (3.5, 63), \\ (4, 50), (5, 77), (6, 87), (7, 95), (8, 100)$$

In this case

$$\bar{x} = \frac{0 + 1 + 2.5 + 3 + 3.5 + 4 + 5 + 6 + 7 + 8}{10} \\ = 4,$$

$$\bar{y} = \frac{1}{10} \cdot \{45 + 48 + 70 + 57 + 63 \\ + 50 + 77 + 87 + 95 + 100\} \\ = 69.2.$$

x	y	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$
3	57	$(3 - 4)(57 - 69.2)$	$(3 - 4)^2$
6	87	$(6 - 4)(87 - 69.2)$	$(6 - 4)^2$
4	50	$(4 - 4)(50 - 69.2)$	$(4 - 4)^2$
5	77	$(5 - 4)(77 - 69.2)$	$(5 - 4)^2$
0	45	$(0 - 4)(45 - 69.2)$	$(0 - 4)^2$
1	48	$(1 - 4)(48 - 69.2)$	$(1 - 4)^2$
7	95	$(7 - 4)(95 - 69.2)$	$(7 - 4)^2$
3.5	63	$(3.5 - 4)(63 - 69.2)$	$(3.5 - 4)^2$
8	100	$(8 - 4)(100 - 69.2)$	$(8 - 4)^2$
2.5	70	$(2.5 - 4)(70 - 69.2)$	$(2.5 - 4)^2$

x	y	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$
3	57	12.2	1
6	87	35.6	4
4	50	0	0
5	77	7.8	1
0	45	96.8	16
1	48	63.6	9
7	95	77.4	9
3.5	63	3.1	0.25
8	100	123.2	16
2.5	70	-1.2	2.25
		$S_{xy} = 418.5$	$S_{xx} = 58.5$

Therefore,

$$m = \frac{418.5}{58.5} \approx 7.154,$$

and

$$b = \bar{y} - m\bar{x} \approx 69.2 - 7.154 \cdot 4 = 40.584.$$

Thus,

$$y = 7.154x + 40.584.$$

Thus, if you wanted to estimate what score a student may get if he or she studied 5.5 hours, you would say that he or she might get a

$$\begin{aligned}y &= 7.154 \cdot 5.5 + 40.584 \\ &= 79.731 \approx 80\end{aligned}$$

on the exam.

Correlation:

Given a data set $\{(x_i, y_i)\}_{i=1}^n$, we may observe that if S_x, S_y are the standard deviations of the x_i 's and y_i 's respectively, then we may show

$$S_{xx} = S_x^2(n - 1), \quad S_{yy} = S_y^2(n - 1).$$

We define the **Pearson Coefficient of Correlation** r by

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}.$$

One may show that

$$r = m \frac{S_x}{S_y}.$$

Note that if $y_i = mx_i + b$, then $\bar{y} = m\bar{x} + b$ and $S_y^2 = m^2 S_x^2$. Thus $m^2 \frac{S_x^2}{S_y^2} = 1$. Note as well that the sign of m determines the sign of r .

Consider again the example with test scores and study time. We had seen that

$$S_{xy} = 418.5, \quad S_{xx} = 58.5.$$

Recall the table as well:

x	y	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$
3	57	12.2	1
6	87	35.6	4
4	50	0	0
5	77	7.8	1
0	45	96.8	16
1	48	63.6	9
7	95	77.4	9
3.5	63	3.1	0.25
8	100	123.2	16
2.5	70	-1.2	2.25
		$S_{xy} = 418.5$	$S_{xx} = 58.5$

We simply add one more column to this to get S_{yy} , namely:

x	y	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$	$(y - \bar{y})^2$
3	57	12.2	1	$(57 - 69.2)^2$
6	87	35.6	4	$(87 - 69.2)^2$
4	50	0	0	$(50 - 69.2)^2$
5	77	7.8	1	$(77 - 69.2)^2$
0	45	96.8	16	$(45 - 69.2)^2$
1	48	63.6	9	$(48 - 69.2)^2$
7	95	77.4	9	$(95 - 69.2)^2$
3.5	63	3.1	0.25	$(63 - 69.2)^2$
8	100	123.2	16	$(100 - 69.2)^2$
2.5	70	-1.2	2.25	$(70 - 69.2)^2$
		$S_{xy} = 418.5$	$S_{xx} = 58.5$	

Simplifying, then summing the rightmost column we get S_{yy} :

x	y	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$	$(y - \bar{y})^2$
3	57	12.2	1	148.84
6	87	35.6	4	316.84
4	50	0	0	368.64
5	77	7.8	1	60.84
0	45	96.8	16	585.64
1	48	63.6	9	449.44
7	95	77.4	9	665.64
3.5	63	3.1	0.25	38.44
8	100	123.2	16	948.64
2.5	70	-1.2	2.25	0.64
		$S_{xy} = 418.5$	$S_{xx} = 58.5$	$S_{yy} = 3583.6$

Thus, we see that

$$\begin{aligned} r &= \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \\ &= \frac{418.5}{\sqrt{58.5 \cdot 3583.6}} \\ &= \frac{418.5}{\sqrt{209,640.6}} \\ &\approx \frac{418.5}{457.8653} \\ &\approx 0.9129. \end{aligned}$$

Hence, there appears to be a strong correlation between study time and exam score.

Unfortunately, not all data sets are strongly linearly correlated, even though they may be strongly correlated. Take for instance

$$(1, 20), (2, 9), (3, 4), (4, 5), (5, 12), (6, 25).$$

Here $r = 0.22$, however,

$$y_i = 3x_i^2 - 20x_i + 37.$$

Homework: For the previous homework exercise, compute the Pearson Coefficient of Correlation. **Answer:** $r = 0.9323$.

It turns out that the correlation coefficient is a measure of how close the data points in the scatter plot (the data points plotted in the plane) are to the line of best fit.

We observed above that if the data points all lie on the line of best fit, then the correlation coefficient r is 1 if the slope of the line of best fit is positive, and the correlation coefficient r is -1 if the slope of the line of best fit is negative.

Correlation Versus Causality:

Here we note that the Pearson Correlation coefficient r is simply a measure of how far away the data points in the scatter plot are to the line of best fit. Do not read more into it than this. With the example above with points on the parabola $y = 3x^2 - 20x + 37$ we saw that the two variables may be strongly correlated, but not necessarily in a linear way. Moreover, it could also be that r is close to one or minus one, but it doesn't mean that there is a cause and effect relationship between the two variables. There could be a third **lurking variable** that is determining the values of the two original variables x and y .

Example: Lurking Variables

A study by Johns Hopkins University, reported in November of 1985, suggested that coffee drinkers that consume five or more cups of coffee per day had three times the heart disease risk of non-coffee drinkers.

However, a later study, published by Dr. Peter W.F. Wilson, suggested that the original study done by Johns Hopkins University failed to take into account that many of the coffee drinkers were also smokers. It was the smoking that actually increased the chances of heart disease, not the coffee consumption. In fact, the study carried out by Wilson concluded that there is no increase in heart disease linked to drinking coffee.