MATH 3070 Introduction to Probability and Statistics
Lecture notes
Relationships: Simple Regression

Objectives:

1. Learn the equation for simple regression
2. Compute the regression equation for a given data set

Simple Linear Regression

Once we've established there is a relationship, we'd like to make use of this knowledge. Specifically, we'd like to be able to use this relationship to predict behavior of the response variable. Having established there is a relationship between square footage and sale price of a house, it would be useful to a prospective buyer to know what a house might list for given its size.

To do this we use a technique known as **linear regression**. The name gives a hint as to what type of relationship must exist (linear) for this to work. We will use the equation of a line which we fit to the data as a predictive equation. Once we know what the line is we can "plug in" new values for the explanatory (independent) variables, turn the crank, and get the predicted output. Simple. But how do we identify the line?

The first step in the process is to produce a scatterplot of the variables. (**Note:** This will only illustrate the simple case where one explanatory variable is assumed to adequately explain all the behavior of a single response variable.) Once we have the data points plotted we can fit a line to them. For any given set of data points there are an infinite number of lines that can be fit. The trick is to find the one that fits the best. To illustrate this we will use a small dataset and fit two lines to it[1]:

x	y
4	6
9	10
1	2
6	2

First let us plot a horizontal line ($y = 5$). This gives a predicted value of 5 for all values of y regardless of input values for x. To differentiate the real values of y from the predicted values, we will use the notation of "y-hat" ($\hat{y}$) to indicate the predicted values. Given this, we see there are errors in the prediction. We can measure the error by subtracting the predicted value from the actual value, like so:

x	y	$y - \hat{y}$
4	6	$6 - 5 = 1$
9	10	$10 - 5 = 5$
1	2	$2 - 5 = -3$
6	2	$2 - 5 = -3$

[1] From *Modern Elementary Statistics, 9th ed*, Freund and Simon, Prentice Hall, 1997

If we sum these errors we get zero, which isn't very interesting (as with the variance and standard deviation). So we will square these values to remove the sign

x	y	$y - \hat{y}$	$(y - \hat{y})^2$
4	6	$6 - 5 = 1$	1
9	10	$10 - 5 = 5$	25
1	2	$2 - 5 = -3$	9
6	2	$2 - 5 = -3$	9

Now the sum of our errors is 44, a much more meaningful number. This doesn't look very good, though. Let's try to fit a better line to the data such as $y = 1 + x$. This actually fits two of the points exactly $[(1, 2), (9, 10)]$ which reduces our error greatly. Following the same procedure as before (subtracting the predicted value for y from the real value for y, squaring the result) we get the following:

		$y = 5$		$y = 1 + x$	
x	y	$y - \hat{y}$	$(y - \hat{y})^2$	$y - \hat{y}$	$(y - \hat{y})^2$
4	6	$6 - 5 = 1$	1	$6 - 5 = 1$	1
9	10	$10 - 5 = 5$	25	$10 - 10 = 0$	0
1	2	$2 - 5 = -3$	9	$2 - 2 = 0$	0
6	2	$2 - 5 = -3$	9	$2 - 7 = -5$	25

The sum of our squared terms is now 26, which is much better than the 44 we got with the horizontal line. Comparing the two lines we would conclude that the second line is a better fit.

This line is called the **least-squares line** because its sum of squared errors is less than the other line's. The procedure we will use is called the method of least-squares precisely because of this. We want the line that minimizes the errors between the actual values for y and the predicted values for y ($\hat{y}$). We accept that there will be error since only a perfect line has all the data points exactly aligned and that would be rare (or manufactured).

Calculating the Regression Line

The following discussion is based upon the symbols and formulas used in Johnson and Kuby.

The equation for a regression line is the same as we learned before, only we use some slightly different symbols. The equation is written

$$\hat{y} = b_0 + b_1 x$$

We compute the value for b_1 first since we actually use that value to calculate b_0. The formula for b_1 is

$$b_1 = \frac{SS_{xy}}{SS_{xx}}$$

where SS_{xy} is the "sum of squares" for each pair of observations x and y and SS_{xx} is the "sum of squares" for each x observation. The values for these are computed by the following formulas:

$$SS_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

$$SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

To compute b_0 we use the following formula:

$$b_0 = \frac{\sum y - (b_1 \times \sum x)}{n}$$

So let's see what this looks like when we apply the formulas to our previous example. First we need to square our x values for the SS_x formula:

x	y	x^2
4	6	16
9	10	81
1	2	1
6	2	36

Next we will calculate the product of each x, y pair of observations. This is for the SS_{xy} formula.

x	y	x^2	$x \times y$
4	6	16	24
9	10	81	90
1	2	1	2
6	2	36	12

Now we will sum the columns to get totals for x, y, x^2, and $x \times y$.

x	y	x^2	$x \times y$
4	6	16	24
9	10	81	90
1	2	1	2
6	2	36	12
20	20	134	128

Now we can calculate the values for $SS_x x$ and SS_{xy}, like so:

$$SS_x x = \sum x^2 - \left(\frac{(\sum x)^2}{n}\right) = 134 - \frac{(20)^2}{4} = 34$$

$$SS_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} = 128 - \frac{(20 \times 20)}{4} = 28$$

The slope of the line (b_1) is then calculated as

$$b_1 = \frac{SS_{xy}}{SS_x x} = \frac{28}{34} = 0.8235$$

We use that value to calculate the y-intercept of the line (b_0)

$$b_0 = \frac{\sum y - (b_1 \times \sum x)}{n} = \frac{20 - (0.8235 \times 20)}{4} = 0.8824$$

So the equation of our line of "best fit" is

$$\hat{y} = 0.8824 + 0.8235 \times x$$

To quickly check to see if we've done all the math correctly, plug in the mean value for x in the equation. You should get the mean value of y as the predicted value. In this case, the mean of x ($\bar{x}$) is 5 ($\frac{20}{4}$). So using that value we get

$$\hat{y} = 0.8824 + 0.8235 \times 5 = 4.999$$

This value is almost exactly our mean of y, the difference is 0.0001 and due to rounding error. So our equation seems to be correctly derived. Now, if we use our equation to calculate our predicted values, we get the following results:

x	y	$\hat{y} = 0.8824 + 0.8235 \times x$
4	6	4.1764
9	10	8.2939
1	2	1.7059
6	2	5.8234

And our residuals ($y - \hat{y}$) are

x	y	$\hat{y} = 0.8824 + 0.8235 \times x$	$y - \hat{y}$
4	6	4.1764	1.8236
9	10	8.2939	1.7061
1	2	1.7059	0.2941
6	2	5.8234	-3.8234

Now, square the residuals to get the *Sum of Squared Error*, or SSE, which we can then compare to our two previous lines.

x	y	$\hat{y} = 0.8824 + 0.8235 \times x$	$y - \hat{y}$	$(y - \hat{y})^2$
4	6	4.1764	1.8236	3.3255
9	10	8.2939	1.7061	2.9108
1	2	1.7059	0.2941	0.0865
6	2	5.8234	-3.8234	14.6184

The sum of our squared errors (residuals) is 20.9412. Our previous lines had sum of squared errors of 44 ($y = 5$) and 26 ($y = 1 + x$). This value is better than either of those, so this line is a better fit to the data.

Interpreting b_0 and b_1

What are these mysterious variables b_0 and b_1? In other terms we can think of b_0 as the y-intercept and b_1 as the slope of the line.

We have to be careful with the regression line, though, because the values produced are not always rational. Yes the line crosses the y-axis at b_0 when we plug in zero for x,

but the value at zero may not make sense. For example, a regression line that predicts food expenditure (in hundreds of dollars) based upon income has the equation

$$\hat{y} = 1.1414 + 0.2642x$$

so when we plug in zero for x we get \$1.1414 which means the family spends \$114.14 when they have zero income. This is a problem known as **extrapolation**. The regression line is calculated using a given data set and the line is valid for all data points within the limits (minimum and maximum) of that data set. Outside those limits the regression line may not be valid. There is a demonstrated relationship between age and height in children, and height can reasonably be estimated from age with a sufficiently large data set, but growth stops eventually and age continues. Predicting height for someone at age 21 would probably yield a value that would be meaningless.