MATH 1070 Introductory Statistics
Lecture notes
Relationships: Correlation and Simple Regression

Objectives:

1. Learn the concepts of independent and dependent variables
2. Learn the concept of a scatterplot
3. Learn the interpretation of Pearson's correlation coefficient
4. Learn the equation for simple regression
5. Compute the regression equation for a given data set

Explanatory and Response Variables

We use the relationship between variables to explain observed phenomena. Does smoking cause cancer, for example. Or the effect of lower temperatures on bacterial growth. We use this information to explain the world around us.

When investigating the relationship between variables we have to differentiate between which of the variables is causative and which is reactive. We usually think of one variable as causing some effect on the other. The variable that we think is causing the effect we refer to as an **explanatory variable**, also known as an **independent variable**. The variable being affected we refer to as a **response variable** or **dependent variable**. Here again we like to use symbols to represent concepts, and we use the letter 'x' to represent the independent variable(s) and the letter 'y' to represent the dependent variable(s).

Scatterplots

As we saw previously, graphs of various types present information in the least painful way possible. There is a special type of graph for investigating relationships, the **scatterplot**. This is also known as an $x - y$ plot or a dot plot.

To create the scatterplot, put the explanatory (independent) variable along the x-axis and the response (dependent) variable along the y-axis. Place a dot (or other character) at the coordinates given by the values for the two variables.

What are we looking for in this graph?

- Overall pattern
- Striking deviations from the pattern
- Form
- Direction
- Strength

The direction of the dots is important as it tells us something about the relationship between the variables. When above-average values in one variable tend to accompany above-average variables in the other we say the variables have a **positive association**. When above-average values accompany below-average values we say the variables have a **negative association**. This is evidence by whether the points trend upward or downward.

The strength of the relationship is shown by how closely the points on the plot follow a given form. The closer to a linear form the stronger the relationship.

Correlation

While a scatterplot is a nice graphical representation, we would prefer something more compact and easier to use. The scatterplot is also open to too much interpretation as to strength and form of relation.

The statistic we use to measure the relationship between two variables was developed by Karl Pearson, and English scholar and mathematician, in the late 1800's. The statistic is named **Pearson's r** in his honor.

The computation of r involves the use of standardized values for both the x and y variables. The sum of the products for each pair of variables is divided by the total number of observations minus one (see below). It is **very important** that you do not sort your data! Sorting the values for the two variables destroys the relationship.

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

This value has some interesting and useful characteristics. It is always between -1.0 and 1.0, never outside this range. Values close to zero indicate weak, or nonexistent, relationships. Values close to the extreme values (-1.0, 1.0) indicate strong relationships. If the value is actually 1.0 (or -1.0) we say it is a perfect linear relationship. Positive values for r indicate a positive relationship. Negative values indicate a negative relationship.

Correlation has no units. Since it is based on standardized variables the units have been removed. Correlation also only works for linear relationships between quantitative variables. If the scatterplot does not show some linearity, correlation will not work. Also, we cannot compute a relationship between qualitative, or categorical, variables.

Strength of the Relationship

r value	Interpretation
0.0 - .20	Slight; almost negligible relationship
.20 - .40	Low correlation; definite but small relationship
.40 - .70	Moderate correlation; substantial relationship
.70 - .90	High correlation; marked relationship
.90 - 1.00	Very high correlation; very dependable relationship

From J.P. Guilford as cited in *Basic Statistical Analysis*, Sprinthall

Caution - Association is NOT Causation

Beware the lurking variable!

It is tempting to interpret the correlation of two variables as implying a causative relationship when there is not one. In one example, a study found a correlation between number of televisions in a country and an increase in the life expectancy. This does not mean that more televisions leads to longer life, does it? There could be some other explanation that would include both phenomena.

In this case there probably is - economic standard of living. A higher standard of living allows people to own television sets and probably have electricity to use as well. A higher standard of living also provides for better health care, which leads to higher life expectancy.

The standard of living is the true explanatory variable in this case, not the number of televisions owned. This is what we call a **lurking variable**, one that is not studied but has an effect on the relationship.

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	$2 - 5 = -3$

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

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

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

$$\hat{y} = a + bx$$

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

$$b = \frac{S_{xy}}{S_{xx}}$$

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

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

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

If we have the correlation coefficient (r) we can use that with the standard deviations for x and y to calculate a value for b using the following formula:

$$b = r \times \frac{s_y}{s_x}$$

To compute a we use a relatively simple formula:

$$a = \bar{y} - b\bar{x}$$

Interpreting a and b

What are these mysterious variables a and b? In the simplest terms we can think of a as the y-intercept and b 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 a 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.

The value for y gives "the change in y due to a change of one unit in x" (Mann, p. 637).