Chapter 10 Correlation and Regression

In this chapter we will be dealing with paired data – an independent and a dependent random variable. One set of data points must be dependent upon the other set of data points. The data in this section will be quantitative, ratio or interval data. We are trying to establish that a relationship exists between two data sets, and then once the relationship has been established through visual inspection and finally hypothesis testing, the next goal is to describe the relationship with a linear equation and then to test whether that equation is statistically significant.

We start with **correlation**. Correlation is the relationship between two variables. It is quantifiable with the correlation coefficient, ρ (rho, for population) and r for sample.

Assumptions for finding correlation:
1) Paired data from random sample.
2) The pairs come from normally distributed populations (This assumption can be lightly tested by
seeing if x & y are normally distributed. Recall: Normal probability plots.)

Let's start with an example so that we can go through the process of correlation and regression.

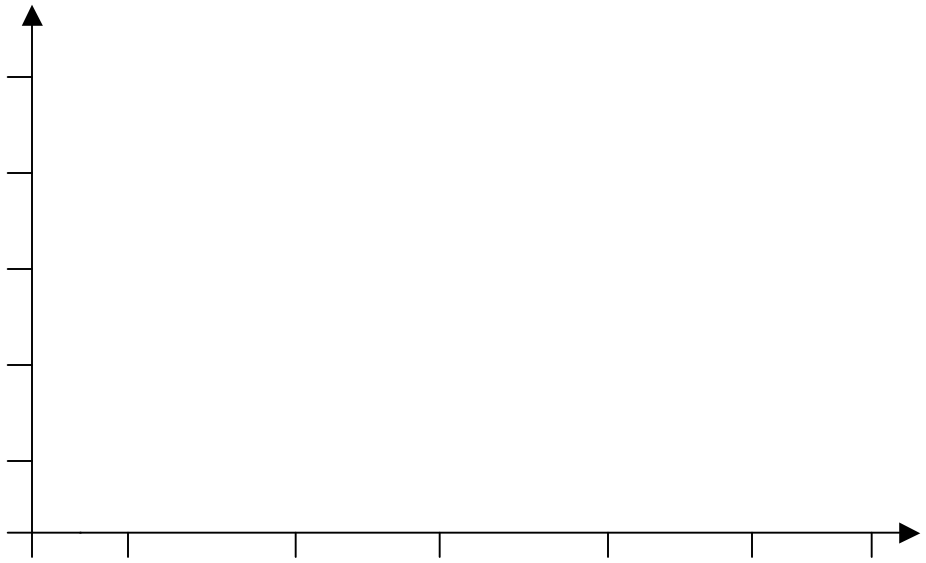
Example: The data points represent the starting salary in thousands of dollars (dependent variable) of a person who scored a certain score on a math reasoning test (independent variable).

X = score	78	85	92	100	85
Y= salary	89	93	99	100	84

The first investigation is a **scatter plot** or **scatter diagram** of the data. This investigation will help us to see if there seems to be a straight line relationship between the independent and dependent variables.

We could see some or strong linear correlation (that correlation could be negative or positive, non-linear correlation, or no correlation what-so-ever. (I'll leave some space so that you can draw the graphs on the board.)

Let's make a scatter plot of our data.



This appears to be strong, positive, linear correlation – of course the data set is very small.

Our **second step** will be to get a measure of the strength of the correlation. We need to compute r to find this mathematical measure, called the **correlation coefficient.**

$$r = \frac{n\,\Sigma xy - \Sigma x\,\Sigma y}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}\ \sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$$

X	Y	X*y	X^2	Y^2
78	89	6942	6084	7921
85	93	7905	7225	8649
92	99	9108	8464	9801
100	100	10000	10000	10000
85	84	7140	7225	7056
440	465	41095	38998	43427

x-bar = 88
y-bar = 93
s_x = 8.336666
s_y = 6.745368782

r = _____

Properties of the Correlation Coefficient

1) $-1 \le r \le 1$
2) R is the same for any scale of measurement
3) R is the same if x & y are reversed
4) R measures the strength of the linear relationship
   - r = 0 no relationship
   - r > 0 positive relationship
   - r < 0 negative relationship
   - r = 1 perfect positive relationship
   - r = -1 perfect negative relationship

The last thing we may wish to find is if the correlation is statistically significant. This will be done with a hypothesis test.

1) Is there any correlation $H_0: \rho = 0$ **vs** $H_a: \rho \neq 0$
2) Is there positive correlation $H_0: \rho \le 0$ **vs** $H_a: \rho > 0$
3) Is there negative correlation $H_0: \rho \ge 0$ **vs** $H_a: \rho < 0$

Test Statistic

$$t = \frac{r}{\sqrt{\frac{(1 - r^2)}{n - 2}}}$$

Critical Value

t with df = n–2

Let's test to see if our correlation is significantly positive at the alpha 0.1 level.

The **last** step is regression. **Regression** is nothing more than fitting a straight line to correlated data so that predictions can be made. Here are the assumptions.

Assumptions

1) There is correlation
2) A prediction about something is desired within the guidelines
   a. Recent data
   b. Prediction is within bounds of current data
   c. Population is assumed to be the same

Regression Equation

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

$a = b_0 = \text{intercept} = (\bar{y}) - b_1(\bar{x})$

$$b = b_1 = \text{slope} = \frac{n\Sigma xy - \Sigma x \Sigma y}{n(\Sigma x^2) - (\Sigma x)^2}$$

The last thing we may want to do is make predictions. (Note: If correlation doesn't exist then y-bar is the best predictor!)

For our example, what salary would you expect to get for a score of 88?

Now we'll go over the example that you had for homework last time.