DATA7202 Lab 1 Semester 1, 2018

"The only useful function of a statistician is to make predictions, and thus to provide a basis for action." W. Edwards Deming

Objectives

On completion of this laboratory session you should be able to:

1. Produce scatterplots showing the relationship between two continuous scale measurements,
2. Correctly interpret the correlation between two continuous scale measurements,
3. Correctly interpret a regression line summarizing the relationship between two continuous scale measurements.

Exercise 1

Data 1		Data 2		Data 3		Data 4	
x1	y1	x2	y2	x3	y3	x4	y4
4	4.26	4	3.1	4	5.39	8	6.58
5	5.69	5	4.74	5	5.73	8	5.76
6	7.24	6	6.13	6	6.08	8	7.71
7	4.82	7	7.26	7	6.42	8	8.84
8	6.95	8	8.14	8	6.76	8	8.47
9	8.81	9	8.77	9	7.11	8	7.04
10	8.04	10	9.14	10	7.46	8	5.25
11	8.33	11	9.26	11	7.81	8	5.56
12	10.84	12	9.13	12	8.15	8	7.91
13	7.58	13	8.74	13	12.74	8	6.89
14	9.96	14	8.1	14	8.84	19	12.5

The table above shows four sets of data prepared by the US statistician Frank Anscombe to illustrate the dangers of calculating correlation coefficients and regression lines without first plotting the data. (Frank J. Anscombe, "Graphs in statistical analysis," *The American Statistician*, 27 (1973), pp. 17-21).

Usually, we would start our analysis with exploratory data analysis, including producing numerical summaries (e.g. tables) and various plots before moving onto formal inference (e.g. hypothesis tests, confidence intervals, regression, ANOVA). Just for today, we will do the exploratory data analysis after the formal inference and see if and where we are misled.

This data is stored in the file **Anscombe.xls** on the Blackboard site (under Datasets). Find it and open it in Excel. Open R Studio and copy all the data in. Refer to the R Summary sheet for instructions.

Analysis 1

1. Obtain and record the correlation coefficient for data set 1 and its significance.

You should produce a table of estimated correlation coefficients, with the associate p-values. As you might guess, x1 is perfectly correlated with itself – the same is true for any variable, so the diagonal running from top left to bottom right is of no interest. Also since corr(x1, y1) = corr(y1, x1), i.e. the order of the two variables in the pair doesn't matter. The p-value is the result of a hypothesis test in which the null hypothesis is that the correlation of the two variables is 0; the alternative hypothesis is that the correlation is non-zero (could be negative or positive).

2. Obtain the regression equation for the relationship between the x1 and y1 values, treating **x1** as the **explanatory (independent)** variable and **y1** as the **response (dependent)** variable.

The response variable is the one we want to predict or to know more about. We are hoping that the explanatory variable can help explain at least some of what happens with the response variable. Once we have constructed a linear regression line, we can observe an instance of the explanatory variable (e.g. measurement of an attribute of a new subject) and use it to try to predict the response variable (another attribute of the same subject). The regression equation is of the type y1 = intercept + slope * x1, where intercept and slope are real numbers. Given a dataset such as this one, R finds the best possible choice of intercept and slope values for the prediction of y1 from x1.

The results here tell you that the straight line described by the equation: y1 = 3 + 0.5 x1 fits the data best (at least when rounded to three significant figures). This can be interpreted as saying that if x1 = 0 then the best prediction for y1 is 3 and that for every increase of one unit in x1, the prediction for y1 will increase by half a unit. A predicted y1 value is also the predicted mean of all members of the population with a given x1.

Of the many things reported for a linear regression, there are a few worth noting.

(i): a p-value and confidence interval for the slope (quite important): listed for x1. The null hypothesis is that the slope is 0, the alternative (used here anyway) is that the slope is non-zero.
(ii): a p-value and confidence interval for the intercept (less important). Null and alternative hypotheses as per slope.
(iii) a p-value for the overall regression equation: listed under Prob>F . Null hypothesis is that the regression line explains none of the variance in y1; alternative hypothesis is that it does explain some of the variance.
(iv) R-squared or Adj (adjusted) R-squared. The latter is generally better since it adjusts for the number of explanatory variables, which can be important in multiple regression, but is less important here with just one explanatory variable - “simple linear regression”. The R^2 value is the proportion of variation in y1 values which is explained by the regression line.

3. Repeat analyses in steps 1 and 2 for the other three data sets (i.e. use x2 and y2, etc.) What do you notice?

4. Now produce scatterplots for each data set.
We could do with seeing the best regression line (line of best fit) on the same graph, so l create another plot based on the first one so that they appear together.
Copy the graph somewhere (e.g. Word) for later comparison and then repeat everything in step 4 for each of the other three pairs of x & y variables. Note that this is easiest if you just edit the first two plots (after getting back to the plot list via the menus).

What do you notice about the plots ? Why do the correlations and regressions turn out similarly while the scatterplots look so different ? What can you conclude about correlation and regression results for these datasets and in general ?

The message from this exercise is that you should always plot the data first so that you can correctly interpret any summary statistics about the relationships between two continuous variables. In 1973 when Anscombe published his article, plotting the data was less common due to the relative lack of access to computers. Now exploratory data analysis including various graphs is what statisticians almost always do first, and you should too.

Exercise 2

For this exercise, suppose you wanted to estimate the weight of a person using only a tape measure. The data set below might help (at least if the person you were measuring was a young male).

For his MS305 data project, Michael Larner weighed and measured 22 male subjects aged 16 – 30. Subjects were randomly chosen volunteers, all in reasonably good health. A range of bodily measurements were made. Subjects were requested to slightly tense any muscles in the area being measured to ensure measurement consistency. Apart from Mass, all measurements are in cm.

Variable	**Description**
Mass	Weight in kg
Fore	Maximum circumference of forearm
Bicep	Maximum circumference of bicep
Chest	Distance around chest directly under the armpits
Neck	Distance around neck, approximately halfway up
Shoulders	Distance around shoulders, measured around the peak of the shoulder blades
Waist	Distance around waist, approximately trouser line
Height	Height from top to toe
Calf	Maximum circumference of calf
Thigh	Circumference of thigh, measured halfway between the knee and the top of the leg
Head	Circumference of the head

The data are shown on the next page. You can copy the data into R-Studio from the file **Mass.xls** from the Blackboard site. You may want to clear the data from the first analysis first.

Analysis 2

1. Obtain scatterplots for all the variables.

Make sure you can see all the details clearly. Which variables look most promising for predicting weight (mass)? A promising variable would be fairly highly correlated with mass, i.e. fit some non-horizontal straight line pretty closely. Which variables look least promising?

2. Obtain the bivariate (pair-wise) correlations among the variables and check the significance of each. Which variables look most promising for predicting weight? Are all these correlations significant ? Which look least promising? Are any of these

correlations not significant ? Are the lists of best and worst variables (for predicting mass) the same as those you inferred from the scatterplot matrix ?

3. Find the regression equation for the relationship between mass and the variable most highly correlated to it.

Now we will conduct the formal linear regression analysis with mass as the response variable and your chosen explanatory variable, which mainly aims to find the optimal slope and intercept and whether or not either are significantly non-zero.

What is the estimated increase in mass for a one cm increase in your chosen predictor measurement?

Data:

Mass	Fore	Bicep	Chest	Neck	Shoulder	Waist	Height	Calf	Thigh	Head
77.0	28.5	33.5	100.0	38.5	114.0	85.0	178.0	37.5	53.0	58.0
85.5	29.5	36.5	107.0	39.0	119.0	90.5	187.0	40.0	52.0	59.0
63.0	25.0	31.0	94.0	36.5	102.0	80.5	175.0	33.0	49.0	57.0
80.5	28.5	34.0	104.0	39.0	114.0	91.5	183.0	38.0	50.0	60.0
79.5	28.5	36.5	107.0	39.0	114.0	92.0	174.0	40.0	53.0	59.0
94.0	30.5	38.0	112.0	39.0	121.0	101.0	180.0	39.5	57.5	59.0
66.0	26.5	29.0	93.0	35.0	105.0	76.0	177.5	38.5	50.0	58.5
69.0	27.0	31.0	95.0	37.0	108.0	84.0	182.5	36.0	49.0	60.0
65.0	26.5	29.0	93.0	35.0	112.0	74.0	178.5	34.0	47.0	55.5
58.0	26.5	31.0	96.0	35.0	103.0	76.0	168.5	35.0	46.0	58.0
69.5	28.5	37.0	109.5	39.0	118.0	80.0	170.0	38.0	50.0	58.5
73.0	27.5	33.0	102.0	38.5	113.0	86.0	180.0	36.0	49.0	59.0
74.0	29.5	36.0	101.0	38.5	115.5	82.0	186.5	38.0	49.0	60.0
68.0	25.0	30.0	98.5	37.0	108.0	82.0	188.0	37.0	49.5	57.0
80.0	29.5	36.0	103.0	40.0	117.0	95.5	173.0	37.0	52.5	58.0
66.0	26.5	32.5	89.0	35.0	104.5	81.0	171.0	38.0	48.0	56.5
54.5	24.0	30.0	92.5	35.5	102.0	76.0	169.0	32.0	42.0	57.0
64.0	25.5	28.5	87.5	35.0	109.0	84.0	181.0	35.5	42.0	58.0
84.0	30.0	34.5	99.0	40.5	119.0	88.0	188.0	39.0	50.5	56.0
73.0	28.0	34.5	97.0	37.0	104.0	82.0	173.0	38.0	49.0	58.0
89.0	29.0	35.5	106.0	39.0	118.0	96.0	179.0	39.5	51.0	58.5
94.0	31.0	33.5	106.0	39.0	120.0	99.5	184.0	42.0	55.0	57.0

Source: Larner, M. (1996). Mass and its Relationship to Physical Measurements. MS305 Data Project, Department of Mathematics, University of Queensland. From the Oz-DASL web page at http://www.statsci.org/