STAT 101

Assignment 2

Draft version posted 21 January 2012

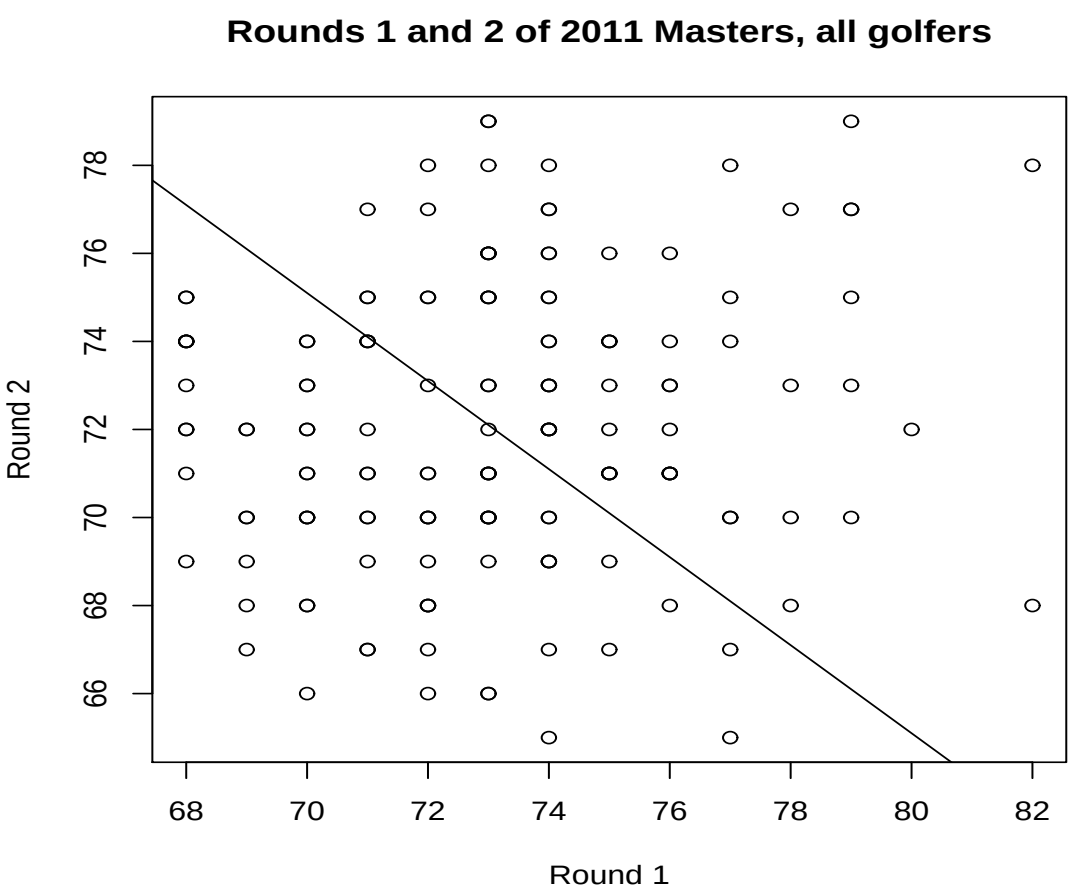
Note: I have used a few questions from the text as is, in spite of having said I would switch over. Mea culpa.

1. This question is based on # 4.24 in the text but uses data in Masters2011.dat. Here is a graph for the scores on the first two rounds for all players in the 2011 Masters golf tournament. The information is drawn from

   www.majorschampionships.com/masters/2011/scoring/index.cfm.

   (a) What was the highest score in the first round? How many golfers had that score? What score did those golfers achieve on the second round?

   (b) Is the correlation in this graph closest to -0.3, 0.1, 0.5 or 0.8?



2. In the previous question imagine that we sent out, along with the professional golfers, a group of amateurs to play the same course twice on two consecutive days. If we put the results of those amateurs together with those of the professionals what would happen to r, the correlation coefficient? Would it be higher than, lower than, or about the same as the correlation in the figure?

3. The graph for the 2011 Masters has a line drawn across it representing the "cut". Golfers whose total score in the first two rounds was 146 or more were cut from the tournament; those whose total was 145 or less were permitted to play in rounds 3 and 4. Look at the group of players who were not cut – the players left of and below the line. Is the correlation coefficient for these players positive or negative? Then answer the same question for the other group of players — those who were cut.

4. Bonus hard question: think about the golfers who made the cut and got to play all 4 rounds. Use the regression effect to predict whether the average scores for these golfers on rounds 3 and 4 will be higher than, lower than or about the same as their average score on rounds 1 and two.

5. From the text: # 4.38 on page 120.

6. From the text: # 5.13 on page 143.

7. From the text: # 5.15 on page 146.

8. From the text: # 5.31 on page 151-152. It is not necessary to make the graph of the line described in b) — just report the results of the prediction.

9. A study is carried out in an elementary school in which there are 7 classes – one class of say 30 students for each grade from 1 to 7. Each student is given two exams: one to measure reading level and one to measure mathematics level. We thus end up with 210 children and each child has an x value for reading and a y value for mathematics.

   (a) Is the correlation between reading level and mathematics level likely to be positive or negative or near 0?

   (b) Suppose that in each class we averaged the reading levels and we also averaged the mathematics levels. Now we have 7 pairs of scores, one pair for each class. Which is most likely to happen: the correlation for these 7 averages is higher than for the 210 individual children, or lower, or about the same? I would like to see you sketch a scatterplot to explain your ideas.

Computing Exercises

The following exercises require you to use either JMP or Excel. They are based on problem 4.44, page 122, and problem 5.30 on page 151 in the text. I will send you the small data set involved by email but it is recorded here because it is small enough to be typed in if need be. The data describe 9 years of data on a small falcon called a merlin. In an isolated area of Sweden researchers counted the number of breeding pairs of this bird each year. They banded the males and measured the percentage of those males who returned the next year.

Breeding Pairs	Percent of males returning
28	82
29	83
29	70
29	61
30	69
32	58
33	43
38	50
38	47

10. Make a scatterplot of Percent Males returning against number of Breeding pairs.

11. Use JMP (or other software) to find the equation of the regression line for predicting Percent Males returning from the Number of Breeding Pairs.

12. Find the correlation between the two variables.

13. Describe the general relation between these two variables in a sentence.

14. Find the residuals, make a plot of the residuals against number of Breeding pairs and comment on whether there seem to be any problems with using linear regression for this data.

15. Use the equation to predict the percent males returning after a season with 30 breeding pairs.

16. What is wrong with doing the same thing for a season with 15 breeding pairs?

To do the calculations in JMP you will:

- Start JMP and use the JMP Starter window to import the data file I send you. It will have two variables.
- Select "Fit Y by X" under the "Analyze" menu.
- Click on "Pairs" then "X, Factor" then click on "Pct" then "Y, Response". Then click "OK".
- A window will open with the required scatterplot in it. It will have "Pct" on the y axis and "Pairs" on the x axis. At the top of the window will be a little red triangle by "Bivariate Fit of Pct By Pairs". Click on the triangle and select "Fit Line" from the menu which pops up.
- The equation for the regression line is there.

- Below the scatterplot is an area called "Linear Fit" with another red triangle at the top. Click on that and select "Save Residuals". This creates a new column in your data set. You can go to "Scatterplot Matrix" under "Graph" to make a plot of Residual on the y axis and Pairs on the x axis. (Under the "Linear Fit" red triangle menu just below "Save Residuals" is "Plot Residuals". The graph called "Residual by Predicted Plot" can be used instead of the plot I am asking you to make the same assessment I am asking you to make but its use is not described in the text.)

- Now under "Analyze" select "Multivariate Methods" and follow the arrow to select 'Multivariate'. A window opens and you need to select both Pairs and Pct and put them in the area "Y, Columns" by highlighting the variables and clicking or by dragging and dropping. Then click "OK".