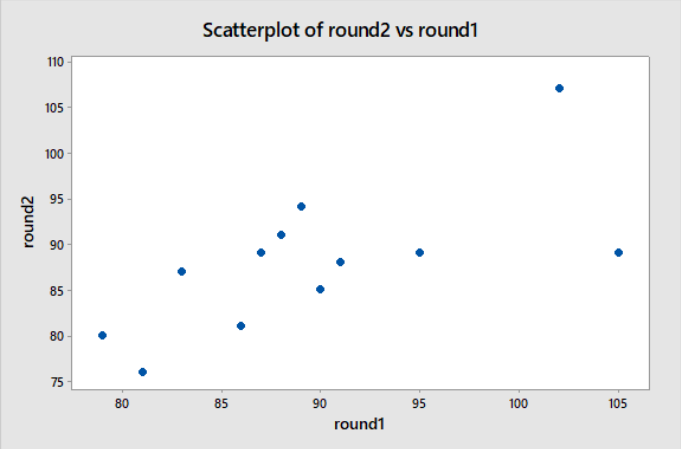
Supplementary exercises 2.2 and 10.7 of IPS7e

Data: Golf scores for 12 members of a college women's golf team in two rounds of tournament play (low scores are better). The scores in round 1 and round 2 are both response variables. Although scores only take integer values, it may be quite reasonable to assume normal distributions because the range of scores can be quite wide.

Exercise 2.2

(a) In the scatterplot, we put round 1 on the x-axis because it is more natural to think of round 2 following round 1 than the other way around. It's debatable whether this in itself makes the first-round score an explanatory variable.



(b) The association is positive; lower scores in round 1 are associated with lower scores in round 2 as well. We would expect a positive association, because the scores reflect the skills of each player. A good player would be expected to have a low score in both rounds, whereas a less skilled player would be expected to have higher scores in both rounds. Of course players do not always perform to their level, but fluctations in scores should be random.

(c) The values for Player 8 fall outside the pattern: she had a very high (poor) score in the first round (105), and an average score (89) in the second round. We have no way of telling which of these two scores represents the "normal" level of this player; thus we cannot say whether round 1 represented a "bad day" or round 2 represented an unusually good round. Note that the scores for Player 7 are both high but quite similar, so this point does not fall outside the pattern.

Exercise 10.7

Minitab commands and output (note the generation of an extra column, in which one value is set to missing):

```
MTB > Correlation  'round1' 'round2'.
MTB > Copy 'round1' c4;
SUBC>   Varnames.
MTB > let c4(8)='*'
MTB > Correlation  'round1_1' 'round2'.
```

Correlation: round1, round2

Correlations

Pearson correlation	0.687
P-value	0.014

Correlation: round1_1, round2

Correlations

Pearson correlation	0.842
P-value	0.001

(a) Already answered in Exercise 2.2.

(b) The sample correlation is 0.687. If we assume a joint normal distribution for the two scores and that the women are an i.i.d. sample from a meaningful population, we can test the hypothesis

$$\mathrm{H}_0 : \rho = 0 \quad \text{versus} \quad \mathrm{H}_a : \rho \neq 0$$

by a t-test. Manual calculation gives: $t = 2.99$ with df $= 10$. The first of the two Minitab listings above gives $P = 0.014$ for this test, and we conclude there is evidence to say that a non-zero correlation exists in the population. The sample correlation shows that the population correlation must be positive, as expected.

(c) Without the outlier we get a sample correlation of 0.842 and strong statistical significance against H_0 ($P = 0.001$); manual calculation gives $t = 4.68$ and df $= 9$. It is not surprising that removal of the outlier has this effect, because without this observation the linear relationship becomes considerably more apparent and convincing.

Extra question:

Although the scores for Player 7 correspond well to the linear pattern, removal of this point affects the correlation as well. Without Player 7 alone the correlation drops down to 0.550 ($P = 0.080$), and is no longer statistically significant. Without both Players 7 and 8 we get a sample correlation of 0.661 ($P = 0.037$), an even stronger impact. We know that extreme points can affect the correlation strongly, and without the scores of Player 8 we would indeed consider the scores for Player 7 as somewhat extreme (in both rounds).

Whether these findings mean that any of the two players should be dropped from the data, is hard to say without additional information about the population these data are supposed to represent. A cautious conclusion is that the association is positive and substantial in any case, and that one would need a larger sample size to decide about whether these points do belong with the others or not. An alternative analytical approach is to use the Spearman rank correlation coefficient which is less sensitive to extreme points (as discussed in Extra exercise 21).

```
MTB > Copy 'round2' c5;
SUBC>   Varnames.
MTB > let c5(7)='*'
MTB > Correlation  'round1' 'round2_1'.
MTB > Correlation  'round1_1' 'round2_1'.
```