STAT 101

Assignment 2

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

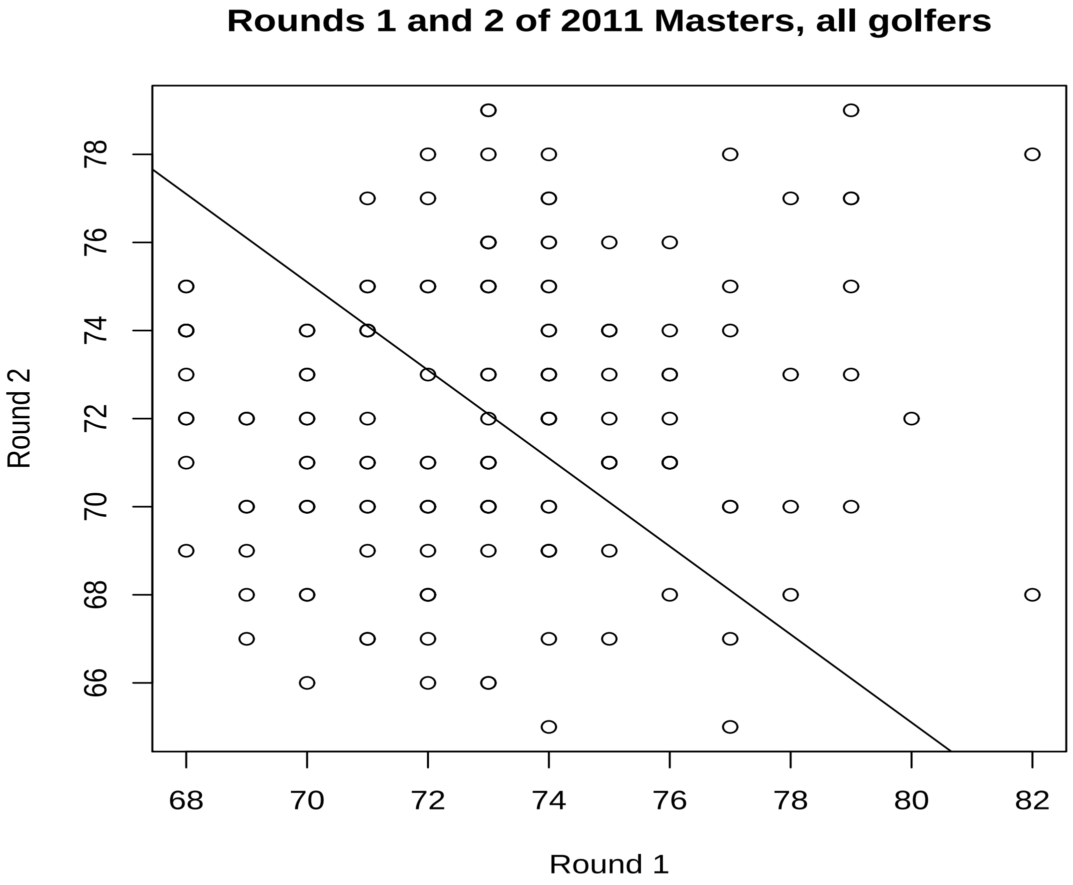
   www.majorschampionships.com/masters/2011/scoring/index.cfm.

   (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?

   This is just to make sure you know how to read the graph. Two golfers shot 82 on round 1. One of those shot 68 on the second round; the other shot 78 on the second round.

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

   The overall trend is very slightly uphill from the lower left to the upper right so the correlation is positive but small — 0.1 is the answer.



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?

   The amateurs would likely score higher than the professionals on average, on both days. Their dots would be up and to the right so the uphill tendency would become more pronounced and r woud be higher than the correlation in the figure shown.

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.

   The dots below the line show a clear trend from upper left to lower right; the correlation in this group of golfers who made the cut is negative. The same is true for those who did not make the cut. The point is important. Splitting a group into subgroups can change the correlation between two variables. It is possible, for instance, for two variables to be positively correlated for men, and positively correlated for women but negatively correlated when sex is not taken into account. If you are looking for evidence that increasing x decreases y for these variables the negative correlation might be misleading.

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.

   Think of the average score on round 1 and two as the x variable and the average score on round 3 and 4 as the y variable. Imagine that the golfers who did not make the cut had also played rounds 3 and 4. Then we would have had a scatter plot of rounds 1 and 2 average for all golfers on the x axis and the rounds 3 and 4 axis on the y axis. The cut picks out those golfers with low values (below average values) of x so the regression method will predict that they will be below average on the rounds 3 and 4 variable but not as much as they were on the round 1 and 2 variable. *That is the regression effect. It is likely that the round 3 and 4 average would be pretty similar to the round 1 and 2 average (a matter of judgment) and so the golfers who made the cut would have a higher average on rounds 3 and 4 than on rounds 1 and 2.*

   In fact in the data the round 1 and 2 average for golfers who made the cut is 70.6 per round. For rounds 3 and 4 those golfers averaged 72.6. The overall average for all golfers for rounds 1 and 2 is also 72.6.

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

   Faculty who are productive researchers are about as good at teaching as those who are not so productive researchers. OR There is no clear relationship between a faculty *member's teaching and his or her research productivity. OR any other statement that says, as the psychologist did, that these two variables show no relationship.*

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

   It might be that the study skills needed to do well in high school math (and therefore take more high school math) are also useful for success in university. It might be that students who come from a less supportive environment are not encouraged to take math and not encouraged to work hard in university. It might be that smarter students are more willing to take math courses and more likely to do well in university. All these variables (work habits, home support, intelligence) are potential confounders or lurking variables.

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

   Possible confounding variables include: family income (children in higher income families can afford to get an education and also have better contacts in the business world to help them get well paying jobs), race or ethnicity (children from cultural groups who are victims of discrimination would be expected to have a harder time getting access to education and a harder time getting a well paying job), intelligence (smarter children get into university and get better paying jobs, perhaps). I am sure there are many others; any one described and well defended should be accepted. The question asks for several so I expect to see at least two to get full marks.

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.

   The slope is

   $$b = rs_y/s_x = 0.5 \times 2.8/2.7 = 0.5185.$$

   You can round off 2 or 3 digits if you wanted (0.52 or 0.518 or 0.519) would be ok by me. The intercept is

   $$a = \bar{y} - b\bar{x} = 69.3 - 0.5185 \times 64 = 36.11.$$

   Just 36.1 is fine, too. Crucial is to recognize that husband's height is y. Regression of Husband's Height on Wife's Height makes Husband's Height the y variable.

   Added note: this correlation between husband's height and wife's height is likely fictitious. It would have to arise from mate choice which is likely a weaker effect than the genetic effect which generates a correlation of 0.5 between the heights of father and son.

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?

It is likely to be positive — quite positive since the older kids will be at higher reading and math levels than the younger ones.

(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.

The correlation will be much higher. As children get older their average math levels go up and so do their reading levels. If you plot clouds of points around the 7 grade averages to represent the behaviour of individual students at each grade level then you will see that the individuals are likely to be more spread out around the upward trend than the averages are.

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.

At the end of these solutions I am including a number of things printed out from JMP; they are what I got when I followed the instructions.

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.

The equation of the line is

$$y = 157.68216 - 2.9934944 * x$$

where x *=Returning Percent and* y *=Breeding Pairs. It is ok to leave all the digits but if you were reporting to other people it would be best to round off to 1 or 2 digits after the decimal point. (Rounding the intercept to one digit is fine but the slope should probably be rounded to 2 It is ok to leave all the digits but if you were reporting to other people it would be best to round off to 1 or 2 digits after the decimal point. (Rounding the intercept to one digit is fine but the slope should probably be rounded to 2.*

12. Find the correlation between the two variables.

    The correlation coefficient is

    $$r = -0.79.$$

    More digits are ok.

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

    In seasons where there are more breeding pairs than average the percent of males returning the next season is below average. Or you could describe it in terms like: Seasons with more breeding pairs are followed by a lower percentage return for the male merlins.

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.

    I don't see compelling problems with the residuals but it does look a bit like the positive residuals are at the edges with the negative residuals in the centre so there may be some curvature to the relationship. The number of data points is not that large so I don't find the evidence terribly compelling one way or the other.

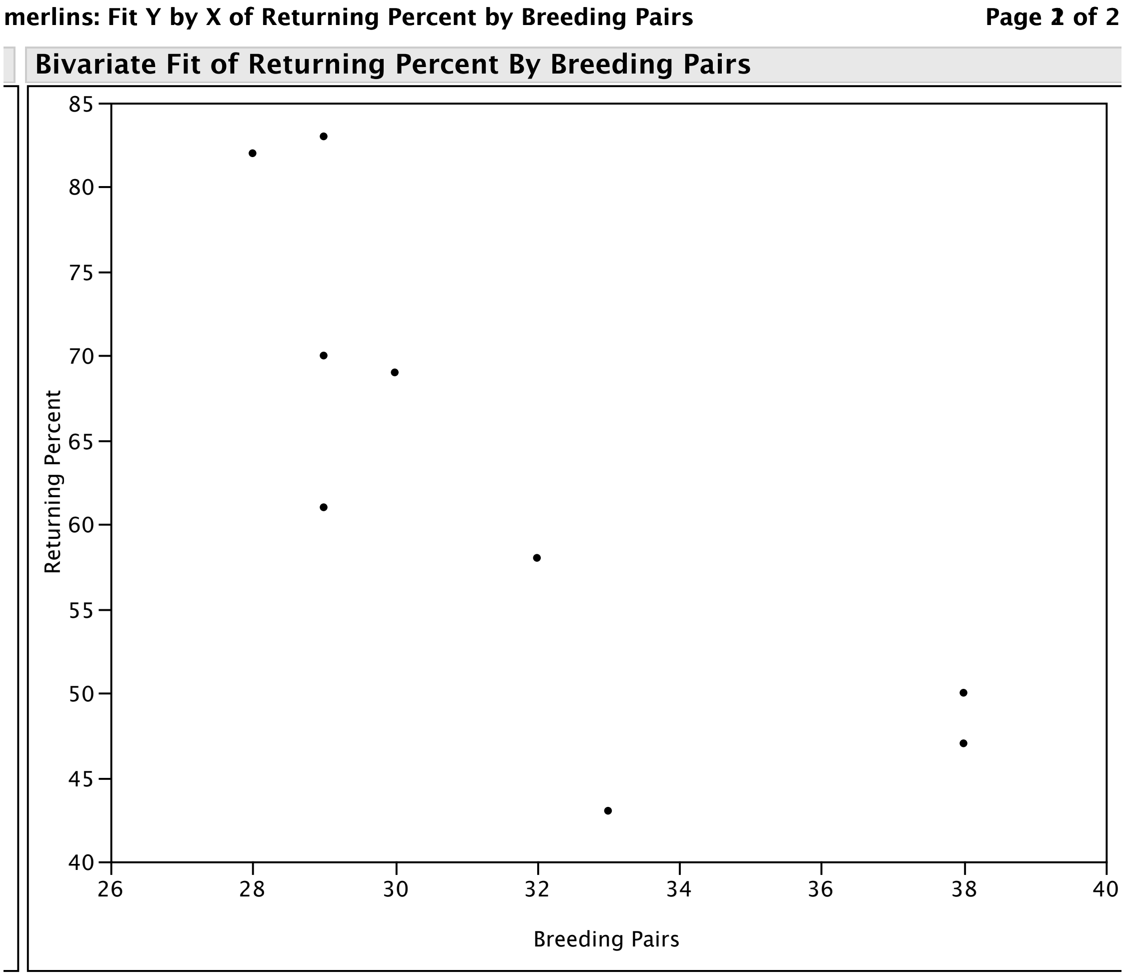
15. Use the equation to predict the percent males returning after a season with 30 breeding pairs.

    Put $x = 30$ *in the regression equation to get the prediction*

    $$157.7 - 2.99 * 30 = 68\% \textit{ males returning.}$$

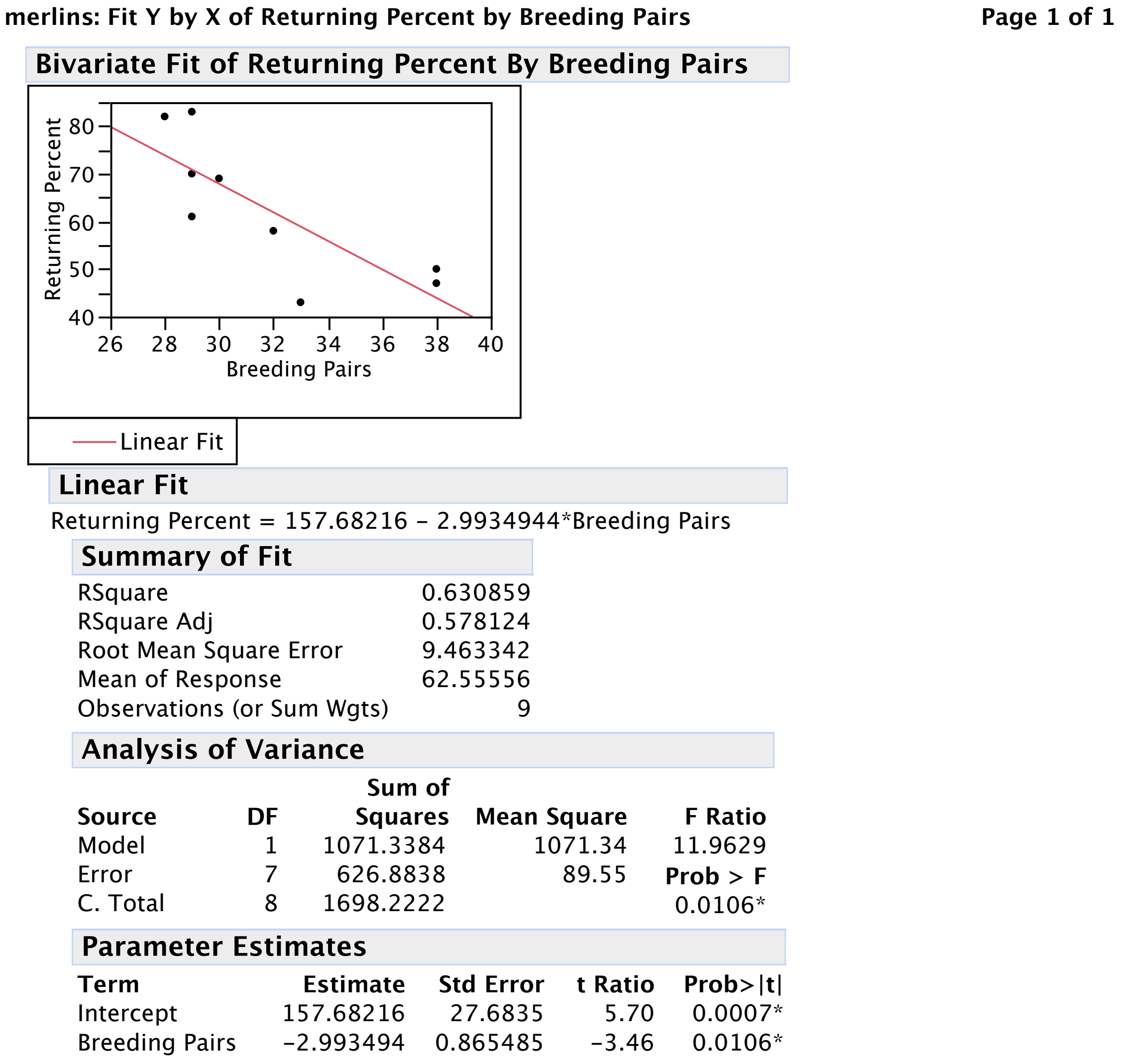
16. What is wrong with doing the same thing for a season with 15 breeding pairs?

    This would be extrapolation well out of the range of the x *values since the smallest* x *value is 28. It is quite likely that the regression line won't give good predictions in such extreme circumstances.*



merlins: Fit Y by X of Returning Percent by Breeding Pairs Page 2 of 2

Bivariate Fit of Returning Percent By Breeding Pairs



merlins: Fit Y by X of Returning Percent by Breeding Pairs **Page 1 of 1**

Bivariate Fit of Returning Percent By Breeding Pairs

Linear Fit

Returning Percent = 157.68216 - 2.9934944*Breeding Pairs

Summary of Fit

RSquare	0.630859
RSquare Adj	0.578124
Root Mean Square Error	9.463342
Mean of Response	62.55556
Observations (or Sum Wgts)	9

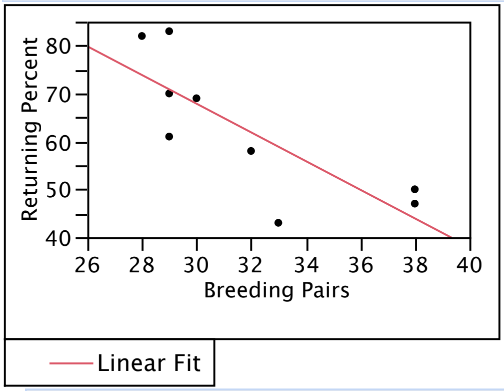
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1071.3384	1071.34	11.9629
Error	7	626.8838	89.55	**Prob > F**
C. Total	8	1698.2222		0.0106*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	157.68216	27.6835	5.70	0.0007*
Breeding Pairs	-2.993494	0.865485	-3.46	0.0106*

Bivariate Fit of Returning Percent By Breeding Pairs

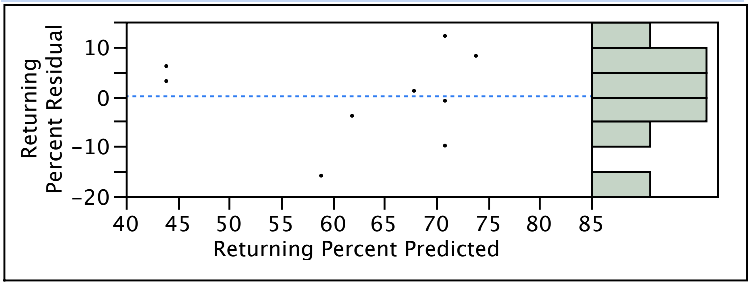


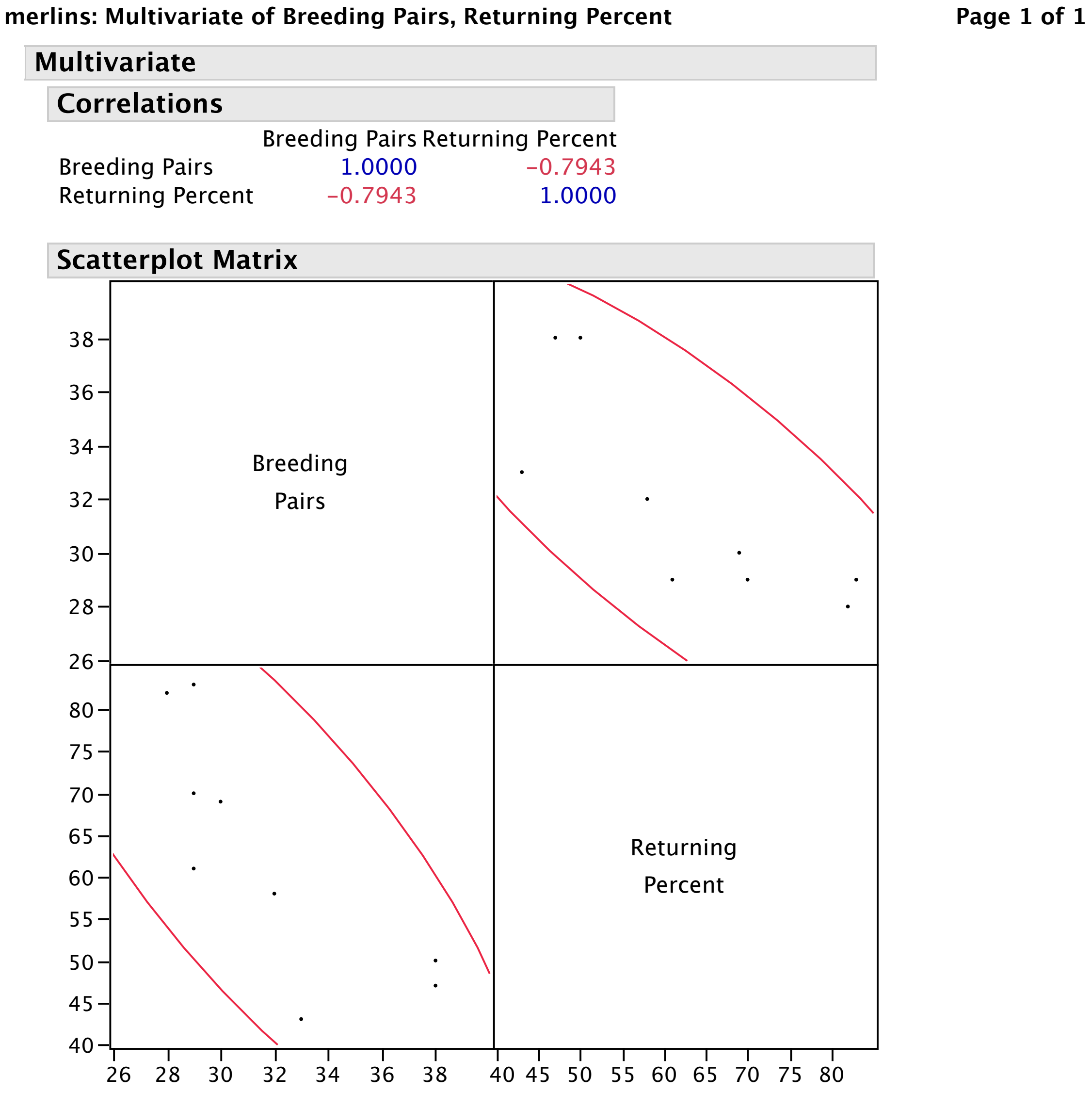
Linear Fit

Returning Percent = 157.68216 - 2.9934944*Breeding Pairs

Diagnostics Plots

Residual by Predicted Plot





merlins: Multivariate of Breeding Pairs, Returning Percent **Page 1 of 1**

Multivariate

Correlations

```
                   Breeding Pairs Returning Percent
Breeding Pairs             1.0000           -0.7943
Returning Percent         -0.7943            1.0000
```

Scatterplot Matrix