Lecture Notes in Quantitative Biology

Correlation - Model Revision

Chapter 20.2 Revised 25 Nov 1997, 19 Nov 2005

ReCap
Correlation: Model Revision

Handout: SRBX15_7.wpd

```
Model revision: added in 1995,
used in 1996,
not used afterward.
```

ReCap--Correlation

Correlation measures the association between two variables.
It is based on a model with two response variables and an explanatory variable that is estimated from the data, rather than measured directly.

The correlation coefficient relates the two response variables to the fabricated explanatory variable. The squared correlation coefficient measures the variation held in common. It does not measure the variance in one quantity, explained by the other.

Correlation is based on a linear model so we use many of the same techniques we learned for the General Linear model. State variables, state model (optional), state H_o/H_A pair, and check residuals.

Most packages do not calculate residuals for correlation analysis. A good data analyst will want to look at scattergram of the data itself to check residuals. Visual inspection of the residuals is used to determine whether a line (new variable X) is a good measure of association.

Rank-based correlation methods are never questioned. They should be because they greatly reduced the power of the test. The type I error estimates were p = 0.004 for randomization or parametric analysis with the t-distribution. The type I error estimate for the rank-based test was twice as great, p = 0.009 for rank-based test.

ReCap--Correlation measures association between two variables.
The variables being correlated are both response variables.
(the explanatory variable is constructed from the data).
The statistic is r, the correlation coefficient.
This is an estimate of the true correlation, ρ (greek rho)

A t-distribution was used to calculate probabilities.
We could have used randomization methods to calculate probabilities, if necessary.
Correlation assumes a monotonic relation between variables.

Today: Model revision with correlation

Wrap-up: Minor deviations from monotonic relation do not matter.
Bowl or arch shaped relation of variable to another are not detected by correlation.

Introduction

The formula for the correlation coefficient ρ:

$$\rho = \frac{\sum (Y_1 - \mu_1)(Y_2 - \mu_2)}{\sigma_1 \sigma_2}$$

In words, the correlation coefficient ρ is the product of the deviations of each quantity from its true mean; this product is then scaled to (divided by) σ_1 and σ_2, the root mean squared deviations of each quantity from its true mean.

This correlation coefficient will be in the range from –1 to +1.

$$-1 \leq \rho \leq +1$$

Model Revision

Work through example from Box 15.7, Sokal and Rohlf 1995.

We will use the generic recipe for hypothesis testing with the GLM.

1. Construct model

Response variables. There are two.
ltot, length of stem mothers
lthor, thorax width of parthenogenetic offspring

Explanatory variable
X = best fitting straight line through the cloud of points.
(not ltot or lthor)
All three quantities (ltot lthor X) are on a ratio type of measurement scale

Write model. This is not usually needed for correlation analysis. It will be done here just to emphasize that correlation is based on a linear model.

$$[\ \text{ltot lthor}\ ] = [\ X\ ]\ [\ \sin^{-1}\rho\ ] + \epsilon$$

2. Execute model

r = 0.65 (see handout)

Calculate and plot residuals.

3. Evaluate Model. Bowls or arches ?

This is not computed for us by Minitab, but it can be done effectively by eye. Draw the best fitting line through the data shown on handout. A line drawn by eye will usually minimize the deviations perpendicular to the line. This is how correlation analysis forms the explanatory variable X. Correlation minimizes the deviations perpendicular to the line, rather than minimizing the deviations perpendicular to the x-axis, as in regression.

We cannot easily plot residuals, but we can accomplish much the same thing by looking at the data points relative to our line. How good is the model ? Does the data show bows or arcs relative to the line ?

It does. The relation between ltot and lthorax appear to be curvilinear. This not a surprise. Morphological data often shows a curvilinear relation of parts with total length (See Lab manual, equations lab). Doubling the length of an aphid evidently does not double the width of the thorax.

According to the recipe, we go back to step 1.

1. State the model

If we want a better model, we can use an exploratory approach: rescale ltot, lthor, or both, then plotting the rescaled quantities. Here are a series of rescalings.

log(ltot)	log(lthor)
ltot	log(lthor)
log(ltot)	lthor
ltot	$lthor^{0.5}$
ltot	$lthor^{2}$
ltot	$lthor^{3}$

2. Execute

3. Evaluate

In each case Minitab was used to compute then plot the rescaled quantities. The plot of ltot versus $lthor^3$ offers a slight improvement. Because the relation between ltot and $lthor^3$ looks more linear than that between ltot and lthor, we expect a higher correlation coefficient. The coefficient for ltot and $lthor^3$ is 0.663, not much greater that 0.65 for ltot and lthor.

1. State the model (continued)

Let's try a completely different model: a monotonic relation between the two variables. As one variable increases the other increase, but not necessarily in a proportional manner. This can be expressed in terms of ranks. If we move from the lowest ranking to the next ranking value of ltot, we expect to move from the lowest to the next lowest ranking value of lthor.

2. Execute

To examine this model, we rescale the values from ratio scale to rank type of measurement scale. Then plot the data, expressed as ranks (see Handout).
The correlation coefficient based on ranked data is r = 0.649, essentially the same as the unranked data. (This is called the Spearman rank correlation coefficient).
The Spearman rank coefficient as based on a model of a linear relation between ranks.

3. Evaluate

This is still not much of an improvement. The departure from a linear relation is still noticeable, even when the data are reduced to ranks. Based on examination of the residuals, the monotonic model (ranked data leading to Spearman coefficient) is no better than original linear model (original data leading to Pearson coefficient).

$$[\ \text{ltot lthor}\ ] = [\ X\ ]\ [\ \sin^{-1}\rho\ ] + \epsilon$$

A series of models were tried. None were linear. It seems that any thorax length is possible at low total lengths (lengths below 7 micrometer units).

1. State Model

Let's examine the relation between variables when ltot > 7 units. (Graph in Handout).

2. Execute

The correlation is r = 0.663741 (n=12), about the same as with n = 15.

3. Evaluate
The relation is now a straight line (n = 12, not 15).
Model is now good. Coefficient with this model is about the same as with all the data, so we will continue with all the data, secure in knowledge that non-linear relation does not affect results.

$r_{Pearson} = 0.65$

$r_{Spearman} = 0.649$ (monotonic relation, based on ranks)

$r_{chopped} = .6637$ (straight line, for lthor > 7 units)

4. State the population
All possible measurements on a small number of aphid stem mothers and their parthenogenetic offspring. There is not enough information to determine whether these aphids were representative of the entire aphid population. Hence this is a statistical population, not a biological population.

5. Mode of inference. Hypothesis testing, we wish to declare whether association is present.

6. State H_A about parameters
H_A: $\rho \neq 0$
H_o: $\rho = 0$
$\alpha = 5\%$

7. Calculate the variance explained by the model
The variance explained by the model is $r^2 = 0.650^2 = 42.25\%$
This is the variance held in common between the quantities. It is not the proportion of variance in one quantity explained by the other.

There are two response variables. So instead of partitioning the variance in the response variable, we will calculate a t-statistic directly. The formula for the t-statistic for r is

$$t = (r - \rho) / s_r$$

$$s_r^2 = (1 - r^2) / df = (1 - r^2) / (n - 2)$$

$$s_r^2 = (1 - 42.25\%) / (15 - 2) = 0.5775 / 13 = 0.04442$$

s_r = sqrt(0.04442) = 0.21076

t = 0.65 / 0.21076 = 3.084

Calculate Type I error.
p = 1 – 0.9956 = 0.0044
This is one tail. For two tailed test
p = 0.0044 + 0.0044 = 0.0088

```
MTB > cdf 3.084;
SUBC> t 13.
      3.084  0.9956
```

8. Recompute p-value if necessary.
We have assumed normal deviations, independent, identically distributed, same as any other application of F or t distribution.
Checking this will be difficult, because most packages do not compute the residuals in a correlation analysis.
n is small
p value far from α
Hence no need to recompute p-value, even if residuals not normal

If we had to defend this p-value, we could always compute a randomized p-value.

9. Declare decision
0.0043 = p < α = 0.05 Reject H_o

10. Report and interpret parameters of biological interest, with measure of uncertainty .
Total length and offspring thorax length are related (r = 0.065 n = 15 p = 0.0043).