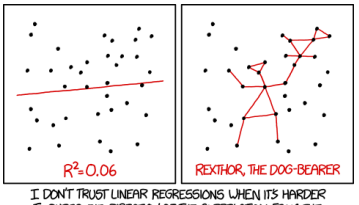
Multiple regression



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

Lecture 4
Biological Statistics III
Ayco Tack

https://xkcd.com/1725/

Outline

- *Multiple regression*
  - ❖ *Linear regression with two or more x-variables*
- *The problem of collinearity*
  - ❖ *Effects of correlation between x-variables*
- *Advice*
  - ❖ *Multiple regression and checking assumptions*

Multiple regression

Situation

- We have measured several quantitative variables on each unit
- It is meaningful to regard one variable, *y*, as possibly depending on, or being predicted by, the other variables, x_1, x_2, ...
- The expected value of *y* is (approximately) a linear function of x_1, x_2, ...

The case with two variables: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$
, where the residual ε is normal with mean zero and standard deviation σ.

We can now test the **null hypothesis** that the true regression coefficients $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ both are zero (but this is rather uncommon).

More typically, we want to test separately the two hypotheses that $\boldsymbol{\beta_1}$ is zero and $\boldsymbol{\beta_2}$ is zero.

Partitioning the sum of squares

Assuming a regression line $\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$ we can calculate the total sum of squares as:

$$SS_{total} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

and the regression sums of squares as:

$$SS_{regression} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

and the residual sums of squares as:

$$SS_{residual} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

(Note that this is similar to simple regression)

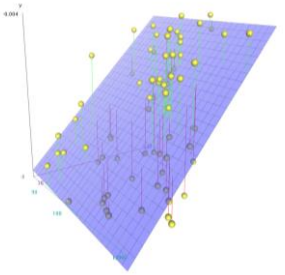
The (least squares) regression is found by making $SS_{residual}$ as small as possible, which determines $\boldsymbol{b_0}$, $\boldsymbol{b_1}$ and $\boldsymbol{b_2}$. For these parameter estimates, we have the partitioning:

$$SS_{total} = SS_{regression} + SS_{residual}$$

With the following degrees of freedom:

$$df_{total} = n - 1,\ df_{regression} = 2, \text{ and } df_{residual} = n - 3$$

Example 1: Plant types and geography

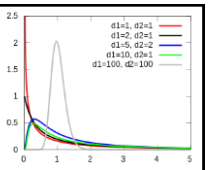


Linear regression of $y = \log_{10}(\text{proportion C3 grasses} + 0.1)$ on Latitude (x_1) and Longitude (x_2)

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$
$$= -1.8497 + 0.0424 x_1 - 0.0037 x_2$$

(see Box 6.1 and Figure 6.1 in Quinn & Keough)

Plant types and geography



We can test the null hypothesis that both regression coefficients are zero using the sums of squares:

ANOVA table

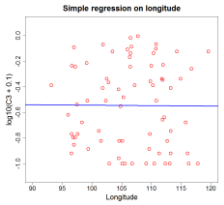
Source	df	SS	MS	F	P
Regression	2	3.6116	1.8058	30.68	<0.000001
Residual	70	4.1200	0.0589		
Total	72	7.7315			

Or we can make a separate test for the two slopes:

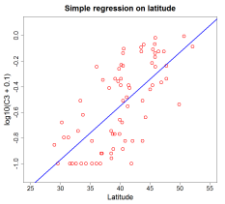
ANOVA table (type II sums of squares)

Source	df	SS	MS	F	P
Latitude	1	3.6112	3.6112	61.36	<0.000001
Longitude	1	0.0412	0.0412	0.70	0.41
Residual	70	4.1200	0.0589		
Total	72	7.7315			

Plant types and geography



The significance test for longitude can be interpreted as a test for a significant improvement in model fit when longitude is taken into account, over and above the other *x*-variables (i.e., latitude)



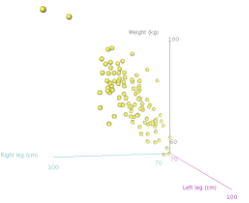
Similarly for latitude: The model fit (measured as residual SS) with only longitude as x-variable is compared to the fit when both longitude and latitude are x-variables.

This is an example of a general method of hypothesis testing in relation to model fitting: *test whether the improvement in model fit when including an x-variable is greater than would be expected from the null hypothesis of no effect*

7

Example 2: The problem of collinearity

A tricky example:
y = weight (kg)
x_1 = length (cm) of the left leg
x_2 = length (cm) of the right leg
n = 100 male students



The regression of weight on left and right leg lengths is:

$$\hat{y} = b_0 + b_1x_1 + b_2x_2$$
$$= -2.01 - 8.15x_1 + 9.14x_2$$

The negative coefficient for the left leg seems strange. Also, the tests of significance for the two coefficients (i.e. if they are different from zero) says that neither is significant.

What is happening??

8

The problem of collinearity

We test for the significance of the left and right leg length:

ANOVA table (type II sums of squares)

Source	df	SS	MS	F	P
Left leg	1	66.09	66.09	2.32	0.13
Right leg	1	83.50	83.5	2.93	0.09
Residual	97	2760.90	28.46		
Total	99	6454.8			

In this example, the x-variables are highly correlated. So, if one already knows the length of (say) the left leg one gets very little extra information about weight by also knowing the length of the right leg, and vice versa. (The type II sums of squares only account for this extra information; they do not sum to the total). Thus, for correlated x-variables one must decide which other x-variables to take into account when one considers the influence of a given x-variable.

Considering only the left leg, the regression is:

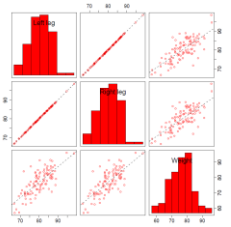
$\hat{y} = -3.74 + 1.01x_1$

, with a highly significant ($p < 10^{-6}$) effect of leg length on weight. Note also that the regression coefficient for x_1 is quite different from the one obtained when x_2 is included in the regression

9

The problem of collinearity

Pairwise correlations:



There is an extremely tight relationship between the two leg lengths: the variance inflation factor (VIF) of either leg length is 3563, which is very large. The VIF for X_j is simply $\frac{1}{1-R_j^2}$ from the regression of X_j against the other predictor variables [see Q&K p. 128]. Either one of them predict weight rather accurately. (Textbooks often state a maximum VIF of 10, but recommendations of 5 or 4 are also found.)

Should we get rid of one of the legs?

10

What should we do about the legs?

We can test the null hypothesis that both regression coefficients are zero using the sums of squares:

ANOVA table (type II sums of squares)

Source	df	SS	MS	F	P
Regression	2	3693.9	1847.00	64.89	<0.000001
Residual	97	2760.7	28.46		
Total	99	6454.8			

Another possibility is to decide in which order to take the x-variables into account. For instance, take the left leg first:

ANOVA table (type I sums of squares)

Source	df	SS	MS	F	P
Left leg	1	3610.4	3610.4	126.8	<0.000001
Right leg	1	83.5	83.5	2.93	0.09
Residual	97	2760.9	28.5		
Total	99	6454.8			

We can conclude that the length of the left leg has a very significant predictive effect on weight, but once the left leg is known, the right leg does not provide a significant amount of additional information.

11

Model comparison

When we have multiple x-variables, we can compare the fit of different models to decide the best-fitting model

- *We can use* anova() *to compare nested models, like:*
  - $y = x1 + x2 + x3$ and $y = x1 + x2$
- *We could use a formal criterion like the Akaike Information Criterion, AIC (The R Book pp. 415-416)*
  - *The best model has the lowest AIC (note that it can become negative)*
  - *There is a penalty for additional parameters*
  - If AIC_{min} is the model with the minimum AIC, then model i is $e^{(AIC_{min} - AIC_i)/2}$ times as probable as the minimum model (to minimize the information loss)
  - *If AIC_{min} and AIC_i differ by less than 2, both models receive strong support and both should receive consideration when making inferences [thus, if the difference is ~2, then the second model is $e^{-2/2} = 0.368$ as probable as the first-ranked model]*
  - *If two models differ by 4-7, the second model has considerable less support*
  - *If two models differ by >10, the second model has (nearly) no support*

12

How should we analyze a multiple regression?

- It is sometimes difficult to know which variables should be *x* and which should be *y* in a multiple regression. For correlated *x*-variables it can also be difficult to determine if there is a statistically significant effect of an individual *x*-variable.
  - *If x-variables are correlated we cannot fully separate the influence of the different x-variables*
    - One possibility is to test for the added effect of an x-variable when all other x-variables are present in the model (type II sums of squares)
    - Another possibility is to test for the joint effect of several x-variables
    - We might want to drop some of the x-variables from the model (using lack of statistical significance, or a criterion like the Akaike Information Criterion, AIC, to determine an acceptable simpler model)
  - *For multiple regression, a good way to think is that the x-variables should predict the value of y*

13

Summary of criticism for multiple regression

- *Checking the assumptions (model criticism) for multiple regression is often done graphically. We should inspect:*
  - *the regression plot for:*
    - *non-linearity*
    - *normal distribution of residuals*
  - *The plot of residuals versus fitted values*
  - *the distribution of residuals for*
    - *outliers*
    - *non-normality*
  - *outliers with extreme x-values are particularly problematic*
- *If there are problems we should try to fix them by transforming the variables*

14

Related reading and information

- **Quinn & Keough:** Chapter 6.1
- **Crawley:** Section 10.13

Some note on (over)fitting elephants:
https://fermatslibrary.com/s/drawing-an-elephant-with-four-complex-parameters#email-newsletter
https://aip.scitation.org/doi/10.1063/1.5031956

Distribution of the Test exercises & Tips for Test Exercises

15