Tutorial topic 8, Simple and Multiple linear regression

Document generated on 2017-04-28

Objectives

- Review calculating (i) ANOVA table of a simple linear regression model (ii) the correlation coefficient;
- Learn to interpret correlation matrix to identify useful predictor variables;
- Learn to calculate regression parameters for a multiple linear regression model.

EXERCISE 1 – REVIEW: FITTING REGRESSION PARAMETERS & ASSESSING SIGNIFICANCE OF A SIMPLE LINEAR REGRESSION MODEL

The number of horses on Canadian farms appeared to decrease after the war:

Year	1944	1945	1946	1947	1948
horses_mult_100.000	28	26	22	20	19

a) Calculate the parameters of the line of best fit (simple linear regression). Use slide 6 - 8 from the lectures in week 8 to assist you

data	Year	Horses	Year.mean	Horses.mean	Product
	1944	28			
	1945	26			
	1946	22			
	1947	20			
	1948	19			
mean					

LINE OF BEST FIT: number of horses = year

b) Complete the following hypotheses and ANOVA:

H_0:
H_1:

Source_of_variation	df	SS	MS	F	p_value
Regression		57.6			
Residual					
Total		60			

c) Is the model significant? Explain why

EXERCISE 2 – CORRELATION MATRIX

A biological survey was performed to examine the nature of the variables thought to influence the breeding of British dippers (thrush-sized birds, living mainly in the upper reaches of rivers; they feed on benthic

invertebrates by probing the river beds with their beaks). Twenty two sites were included in the survey. The predictor variables that were measured are:

- site altitude
- water hardness
- river-bed slope
- the numbers of caddis fly larvae
- the numbers of stonefly larvae
- the numbers of mayfly larvae
- the numbers of all other invertebrates collected
- the breeding density which is the number of breeding pairs of dippers per 10 km of river (this is the response variable)

In the analyses, we transform the four invertebrate variables using a log10(number+1) transformation.

A correlation matrix presents the Pearson correlation coefficients.

variable	Altitude	Hardness	RiverSlope	LogCadd	LogStone	LogMay	LogOther
Hardness	-0.082						
RiverSlope	0.49	0.242					
LogCadd	0.181	0.334	0.431				
LogStone	0.477	0.03	0.575	0.443			
LogMay	-0.128	0.414	-0.096	0.438	-0.167		
LogOther	-0.339	0.554	-0.02	0.352	-0.3	0.388	
Br.Dens	0.406	0.350	0.710	0.613	0.763	0.156	-0.128

a) Why would we have transformed the counts of invertebrates?

b) Why would add 1 to the count before performing the log transformation?

c) What is the best single predictor of breeding density? What is the second best single predictor? What is the worst? Had the correlation between Br Dens and LogStone been -0.763, would any of these conclusions change?

d) If you had to choose two predictors to use in a MLR model, which would you choose?

EXERCISE 3 – CALCULATION OF REGRESSION PARAMETERS

Please look at the lecture slides 29 and 30 from Topic 8 to assist you

The variance-covariance matrix for these variables is:

variable	Altitude	Hardness	RiverSlope	LogCadd	LogStone	LogMay	LogOther	Br.Dens
Altitude	3143.9	-196.8	172.1	9.4	26.7	-13.4	-34.7	33.9
Hardness	-196.8	1845.1	65.1	13.4	1.3	33.2	43.4	22.4
RiverSlope	172.1	65.1	39.2	2.5	3.6	-1.1	-0.2	6.6
LogCadd	9.4	13.4	2.5	0.9	0.4	0.8	0.6	0.8
LogStone	26.7	1.3	3.6	0.4	1.0	-0.3	-0.5	1.1
LogMay	-13.4	33.2	-1.1	0.8	-0.3	3.5	1.3	0.4
LogOther	-34.7	43.4	-0.2	0.6	-0.5	1.3	3.3	-0.3
Br.Dens	33.9	22.4	6.6	0.8	1.1	0.4	-0.3	2.2

The mean values are

(i) breeding density = 4.7
(ii) RiverSlope = 9.9
(iii) log(stonefly numbers +1) = 5.0

a) Write down the equations you would solve to obtain the parameters for the line of best fit for predicting breeding density using the two predictors RiverSlope and log(stonefly numbers +1).

Line of best fit is:

$$breeding_density = b_0 + b_1 RiverSlope + b_2 log(stoneflynumbers + 1).$$

b) **If time permits** solve the simultaneous equations to estimate b_1, b_2 and then b_0.