Statistics for the Doctoral School in Biomolecular Sciences
Academic year 2014-2015

Exercises after Lecture 6

Analysis of variance and regression models

EXERCISE 6.1 The authors of a study have studied the expression of gene *for* in bees (*Apis mellifera*. They have compared 'nursing bees' (young bees that feed larvae in the beehive) and 'foraging bees' (older bees that forage for nectar and pollen outside the beehive) out of 3 colonies. The following table presents the mean value of gene expression in the groups

Type of bee	Colony	Gene expression
Nursing	1	0.99
Foraging	1	1.93
Nursing	2	1.00
Foraging	2	2.36
Nursing	3	0.24
Foraging	3	1.96

1. Which model would use to test whether expression of gene *for* differs between foraging and nursing bees

2. Is there something to be gained by considering colony number in the model?

EXERCISE 6.2 What does the statistics R^2 represent in the analysis of variance and in regression?

EXERCISE 6.3 Explain why it can be sometimes useful transforming the data, before performing an analysis of variance on them.

EXERCISE 6.4 The analysis of the dataset on survival of rats after having given them a poison and antidote has provided the following Analysis of Variance Table in which the response variable is 1/survival time,

	Df	Sum Sq	Mean Sq	F value	Pr(¿F)	
poison	2	34.877	17.4386	72.6347	2.310e-13	***
treat	3	20.414	6.8048	28.3431	1.376e-09	***
poison:treat	6	1.571	0.2618	1.0904	0.3867	
Residuals	36	8.643	0.2401			

How do we interpret the numbers in the table? Can we find what is the value of R^2 in this analysis? Explain what is the difference an additive model and a model with and without interaction.

After seeing this table, we proceed with applying and additive model. Why are we justified in doing so?

The results are the following:

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(¿—t—)	
(Intercept)	2.6977	0.1744	15.473	¡ 2e-16	***
poisonII	0.4686	0.1744	2.688	0.01026	*
poisonIII	1.9964	0.1744	11.451	1.69e-14	***
treatB	-1.6574	0.2013	-8.233	2.66e-10	***
treatC	-0.5721	0.2013	-2.842	0.00689	**
treatD	-1.3583	0.2013	-6.747	3.35e-08	***

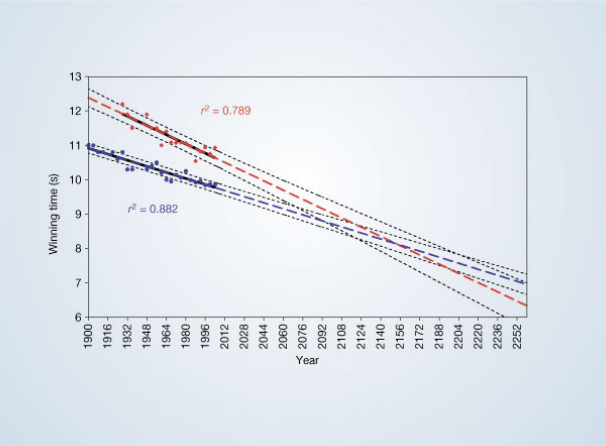
```
   Residual standard error:  0.4931 on 42 degrees of freedom
Multiple R-squared:  0.8441, Adjusted R-squared:  0.8255
F-statistic:  45.47 on 5 and 42 DF, p-value:  6.974e-16
```

Describe clearly which is the final model that is obtained from this analysis. In particular, what is the mean mortality (=1/survival time) predicted for rats subjected to poison II and treatment C?

Describe precisely which are the tests that have been performed (and presented with a p-value); discuss which results appear to be significant, which are the potential problems of this analysis, and how one could proceed.

EXERCISE 6.5 It has been found that many quantities are related to body size according to the so-called allometric relations of the form $y = kx^a$. For instance, such a relationship has been found for metabolic rate, where now y represents metabolic rate, x body size, K and a constants to be estimated. Assuming that we have data on body size and metabolic rate for a large number of animal species, explain how we can use the method of linear regression to estimate a (and K); specify which are the assumptions used in this analysis.

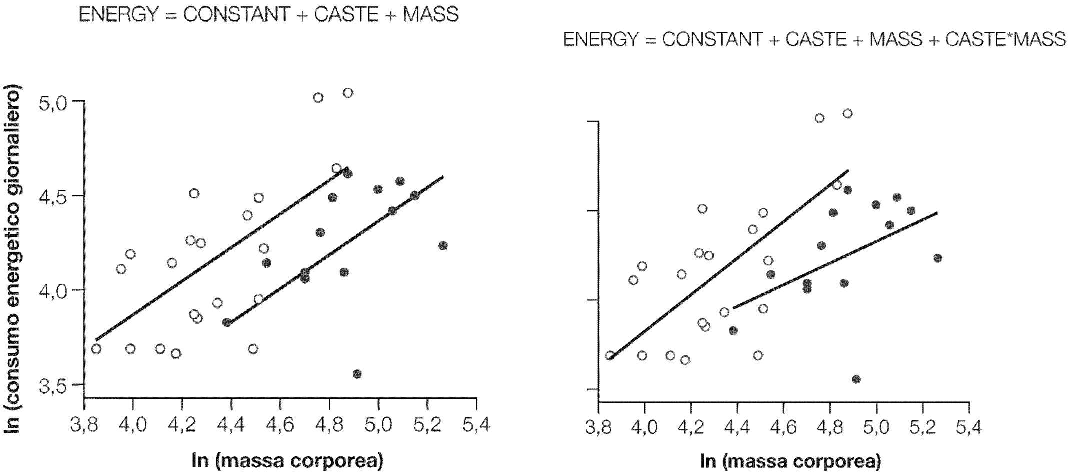
EXERCISE 6.6 A study by Tatem et al. (2004) used the data of the winning time for men and women 100 metres at the Olympics to compute the regression lines, shown in the picture of winning time against year.



From the regression lines, they predict that winning times in the Olympics beyond 2156 will be better for women than for men. What are possible reasons to view cautiously this prediction?

EXERCISE 6.7 An analysis of daily energetic consumption as a function of weight in a species of social insects has shown the role of the caste of the individual. The data collected are shown (in log-log scale) in the two figures, together with two different models fitted to the data.

Describe the mathematical relation (separately for the two castes) used in the two models to describe how daily energetic consumption depends on weight (both in log-log scale, and in the natural scale).



EXERCISE 6.8 On a dataset (present in the standard implementation of R) we have performed the regression of ozone concentration on wind speed (Wind) and month (mese, with values from 5 to 9). Using R, we obtain the following regression table:

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> $\|t\|$)	
(Intercept)	50.748	15.748	3.223	0.00169	**
Wind	-2.368	1.316	-1.799	0.07484	.
mese6	-41.793	31.148	-1.342	0.18253	
mese7	68.296	20.995	3.253	0.00153	**
mese8	82.211	20.314	4.047	9.88e-05	***
mese9	23.439	20.663	1.134	0.25919	
Wind:mese6	4.051	2.490	1.627	0.10680	
Wind:mese7	-4.663	2.026	-2.302	0.02329	*
Wind:mese8	-6.154	1.923	-3.201	0.00181	**
Wind:mese9	-1.874	1.820	-1.029	0.30569	

—

Signif. codes: *** < 0.001 ** < 0.01 $* < 0.05$. < 0.1

```
Residual standard error:  23.12 on 106 degrees of freedom
Multiple R-squared:  0.5473, Adjusted R-squared:  0.5089
F-statistic:  14.24 on 9 and 106 DF, p-value:  7.879e-15
```

Describe clearly which is the final model that is obtained from this analysis. It is advisable writing down separately the model for observations belonging to each month; in other words, write formulae Ozone = ... if month = 5, Ozone = ... if month = 6, ...

Describe precisely which are the tests that have been performed (and presented with a p-value); discuss which results appear to be significant, which are the potential problems of this analysis, and how one could proceed.