Biostats VHM 802 Course
Winter 2016, Atlantic Veterinary College, PEI
Henrik Stryhn

Exercises for Session 2: Linear and multiple linear regression

Exercise 2.1

Simple linear regression and model check
In the middle of the 19th century, the Scottish physicist James D. Forbes studied the relation between pressure and boiling points. The aim of his work was (among other things) to estimate altitude above sea level by measuring the boiling point of water (barometers were fragile and difficult to transport at that time). His own measurements from the Alps and Scotland were the following (from Weisberg (1985): *Applied Linear Regression*; data from Forbes, J. D. (1857): Further experiments and remarks on the measurement of heights by the boiling point of water, *Trans. R. Soc. Edinburgh* **21**, 135–143.):

Location	1	2	3	4	5	6	7	8
Boiling point (°F)	194.5	194.3	197.9	198.4	199.4	199.9	200.9	201.1
Pressure (inches Hg)	20.79	20.79	22.40	22.67	23.15	23.35	23.89	23.99

9	10	11	12	13	14	15	16	17
201.4	201.3	203.6	204.6	209.5	208.6	210.7	211.9	212.2
24.02	24.01	25.14	26.57	28.49	27.76	29.04	29.88	30.06

Analyse these data, answering to the following points.

1) Formulate a statistical model (choice of dependent and independent variables).

2) Check the model using residual plots and regression diagnostics. Are there any outlying observations? Is a linear relation appropriate? If not so, make the necessary adjustments to achieve a good model.

3) Estimate the parameters of the model you chose in 2), and carry out the necessary diagnostics to convince yourself that the model assumptions are met to a reasonable degree.

4) On theoretical grounds, Forbes chose to analyse the dependence of the logarithm of the pressure on the boiling point. Conduct this analysis as well, and compare the two analyses.

Exercise 2.2

Data error
The table below shows data from a study of 20 patients with chronic congestive heart failure. Two measurements are shown — ejection fraction x (in percent), which is a measure of left ventricular dysfunction, and pulmonary arterial wedge pressure y (in *mm* Hg): (from Altman (1991): *Practical statistics for medical research*; data from Caruana *et al.* (1988): Efects of chronic congestive heart failure secondary to coronary artery disease on the circadian rhythm of blood pressure and heart rate, *Am. J. Cardiol.* **62**, 755–759.)

	Patient																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
x	28	26	42	29	16	21	25	35	30	36	37	41	20	26	38	26	10	18	10	31
y	15	14	15	12	37	30	7	14	28	13	5	13	24	8	13	17	27	29	8	5

One value has been mistranscribed from the paper. Estimate a linear regression of y on x, and use residuals and/or regression diagnostics to determine which patient's data is most likely to be wrong.

Exercise 2.3

Simple linear regression and transformation
Consider the data below on colon cancer mortality in different age groups (given as interval midpoints). The mortality is in parts per million per year. (Data from Cairns, J. (1975): The cancer problem, *Scientific American* **233**, 64–78.)

age	22.5	27.5	32.5	37.5	42.5	47.5	52.5	57.5	62.5	67.5	72.5	77.5	82.5
mortality	0.58	1.42	3.41	6.63	13.0	24.9	45.7	82.2	128	187	283	369	462

1) Investigate whether the data can be described by a regression of a) y on x, b) $\ln y$ on x, or c) $\ln y$ on $\ln x$ (where y = mortality and x = age). Estimate the parameters in the best model. Write down the equation for predicting y from x. Compute prediction intervals for (some of) the age values in the data, and compare with prediction intervals from the second best model.

2) Under some assumptions (to be explained below), the relation between y and x should approximately take the form

$$\ln y = const + (n-1)\ln x - \lambda x\,,$$

where λ is a parameter of a so-called Poisson process. Is this equation in agreement with the data? If so, estimate the parameters n and λ. (The assumptions are: *i)* every cell has n protective genes, *ii)* a cell can develop into a cancer cell only when all these n genes have mutated, *iii)* genes mutate independently, and *iv)* the probability of mutation in a small time interval $(t, t+h)$ equals λh, corresponding to a Poisson process.)

Exercise 2.4

Linear and polynomial regression
The data below are measurements of height h (in m) and diameter d (in cm) of 18 Corsican Pines; from Jeffers, J. N. R. (1959): *Experimental Design and Analysis in Forest Research*, Almquist & Wiksell, Stockholm.

Tree	1	2	3	4	5	6	7	8	9
Diameter (cm)	32	31	30	29	29	28	25	23	20
Height (m)	22.7	22.7	22.6	22.6	21.9	21.9	21.8	21.0	20.4
Tree	10	11	12	13	14	15	16	17	18
Diameter (cm)	18	17	17	16	16	15	13	11	11
Height (m)	18.6	19.2	18.9	18.5	18.1	17.7	17.2	16.5	15.5

1) It is of interest to describe the tree height as a function of the diameter. Explore the relationship between the variables using linear and polynomial regression. What polynomial order seems necessary to describe the data? Compute 95% prediction intervals for diameters 11, 20 and 30 cm.

2) Alternatively, and supposedly better, the data can be analysed using the relation

$$h = \alpha\, d^{\beta} \quad \text{or} \quad \ln h = \ln \alpha + \beta \ln d\,.$$

Formulate a corresponding statistical model and analyse the data. Compute similar 95% prediction intervals as in 1), and compare.

Exercise 2.5

Box-Cox transformation
In a study of reproductive performance in dairy cattle in Reunion Island, the calving to first service interval was measured (among other variables). We consider here only data for 124 cows in one of the study herds, and as our single explanatory variable the origin of the cows: imported, from a central farm on the island, or from the local producer. In the datafile, the origins are denoted by levels 1–3, corresponding to their order above. (Data from Dohoo *et al.* (2001), *Preventive Veterinary Medicine* **50**, 127–144.)

Determine whether there seems to be a need for transformation for the (1-way ANOVA) analysis, and try the standard transformations. Determine also approximately the optimal Box-Cox transformation of the data. (For the entire data set, an optimal λ-value of -0.576 was reported.)

Exercise 2.6

Stata analysis equivalents for linear regression in another statistics package
Carry out the analyses corresponding to Exercises 1-3 for linear regression in VHM 812 (VER 14.1-3) based on the dataset on times between episodes of bovine tuberculosis in cattle herds in Ireland. Use either Minitab or your preferred statistical software (which can fit linear models). Some of the analyses from Stata may not be available in the software you're using. Specifically, for Minitab (version 16 or later) you may skip (or replace as indicated) the following items in Exercise 3:

- Cook-Weisberg test for heteroscedasticity,
- Shapiro-Wilk test for normality (the Ryan-Joiner test is "similar"),
- dbeta statistics for influence.

Exercise 2.7

Multiple linear regression for small dataset
Prater's data on gasoline comprises 32 measurements on chemical variables involved in the production of gasoline, reproduced here from Atkinson (1985), *Plots, Transformations and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis.* The variables are: y, the percentage of gaoline obtained from crude oil; x_1, the crude oil gravity (°API); x_2, the crude oil vapor pressure (lbs/in^2); x_3, the temperature at which 10% of the crude oil is vaporized (°F); x_4, the temperature at which all of the crude oil is vaporized (°F). The table below shows a subset of the data.

obs. i	y_i	x_{1i}	x_{2i}	x_{3i}	x_{4i}
1	6.9	38.4	6.1	220	235
2	14.4	40.3	4.8	231	307
3	7.4	40.0	6.1	217	212
4	8.5	31.8	0.2	316	365
...	...	...	...	...	...
32	45.7	50.8	8.6	190	407

1) Use the data to determine a good model for predicting gasoline yield from the four other variables.

2) Evaluate the model assumptions carefully using residuals and diagnostics. Do you note any problems with the model or any pecularities with the data?

3) Explore how model selection procedures perform for these data. Do they lead to the same final model as your analysis?

Exercise 2.8

Multiple linear regression for intermediate size dataset
We consider here an excerpt from the Los Angeles Heart Study supervised by J. M. Chapman, reproduced from Dixon & Massey (1983), *Introduction to Statistical Analysis*, 4th ed. The following measurements were obtained for 60 men: y, weight in pounds; x_1, age in years; x_2, systolic blood pressure (mm Hg); x_3, diastolic blood pressure (mm Hg); x_4, cholesterol (mg/dl); x_5 height in inches. The table below shows a subset of the data.

obs. i	y_i	x_{1i}	x_{2i}	x_{3i}	x_{4i}	x_{5i}
1	44	124	80	254	70	190
2	35	110	70	240	73	216
3	41	114	80	279	68	178
4	31	100	80	284	68	149
...	...	...	...	...	...	...
60	68	110	80	268	62	138

1) Use the data to determine a good model for predicting weight from the five other variables.

2) Evaluate the model assumptions carefully using residuals and diagnostics. Pay special attention to whether any of the observations has strong influence on the final model.

3) Explore how model selection procedures perform for these data. Do they lead to the same final model as your analysis?

Exercise 2.9

Multiple linear regression for large dataset
The `bodyfat` dataset from the Statlib data server gives estimates of the percentage of body fat determined by underwater weighing and various body circumference measurements for 252 men. It is of interest to develop an equation to estimate body fat from the circumference measurements (see the description of the dataset and the biological context at

`ftp://rcom.univie.ac.at/mirrors/lib.stat.cmu.edu/datasets/.index.html`

for details; the Statlib server seems to have been discontinued, but this mirror site is still accessible). The outcome variable is y, percent bodyfat estimated from Siri's equation (%), and the predictors of interest are: age (years), weight (pounds), height (inches), and circumferences (all in *cm*) of neck, chest, abdomen, hip, thigh, knee, ankle, biceps, forearm, and wrist.

1) Use the data to determine a good model for predicting percent bodyfat from the other variables (except body density!).

2) Evaluate the model assumptions carefully using residuals and diagnostics. Make decisions about whether any observations should be omitted from analysis.

3) Explore how model selection procedures perform for these data. Do they lead to the same final model as your analysis?

Exercise 2.10

Stata analysis equivalents for linear model building in another statistics package
Carry out the analyses corresponding to the VER Chapter 15 Model building exercise based on the dataset on average daily gain of pigs in 15 farms in Prince Edward Island. Use either Minitab or your preferred statistical software (which can fit linear models). Focus specifically on Questions 6, 8 and 11; some Minitab hints:

Q6 Use the Minitab menus to carry out the descriptive analyses requested.

Q8 Variable selection by stepwise forwards and backwards procedures are in Minitab 17 built directly into the comprehensive regression menus (for linear, logistic and Poisson regressions). The models may contain both continuous and categorical predictors, and the latter are treated as combined variables (including all indicator variables needed to represent the categories). Moreover, when an interaction or a quadratic term is included, the default is a "hierarchical model", by which main effects are always included with an interaction and a linear term is always included with an interaction.
In contrast, the Best Subsets menu does not have these new features, and therefore categorical predictors must be represented by indicator variables, and derived variables such as interactions and quadratic terms must be calculated and included manually. It is however still possible to force variables into models, so one option is to force in all terms for combined effects.

Q11 The comprehensive regression menu by default displays the predictive R^2, and the PRESS statistics is also displayed in the expanded tables. Therefore, there is little interest in manually carrying out a split-sample analysis with estimation and validation datasets.

Exercise 2.11

Several regression lines

An experiment about the accumulation of salts was conducted by measuring the concentration of rubidium and bromide ions in potato slices after the potatoes had been immersed in a solution containing these ions for several hours. The concentrations are given in the table below, in mg per 1000 g of water in the tissue. (Data from Steward, F. C. & Harrison, J. A. (1939): The absorption and accumuluation of salts by living plant cells, *Annals of Botany*, New Series **3**, 427–453.)

Duration of immersion hours	Rubidium conc. $mg/1000\ g$	Bromide conc. $mg/1000\ g$
21.7	7.2	0.7
46.0	11.4	6.4
67.0	14.2	9.9
90.2	19.1	12.8
95.5	20.0	15.8

Biochemical considerations might indicate the absorptions of bromide and rubidium ions to follow similar patterns.

1) Estimate first two separate linear regression models, one for each of the ions.

2) It is of interest to analyze the two ions in a combined statistical model. Formulate a combined model with separate regression equations for the two ions, and estimate the parameters. Which additional assumptions are made about the data compared to the analyses in 1) ?

3) Use a statistical test to assess whether the data show evidence against equal rates of absorption for the two ions. If not, estimate the parameters of a model assuming equal rates of absorption.

4) Finally, examine whether the regression equations for bromide and rubidium ions are identical. Summarize the statistical analysis and give the parameter estimates of the best model you found, with associated standard errors.

Exercise 2.12

Regression and analysis of covariance

The table below gives nave height and total height, both in feet, for medieval English cathedrals. In addition, the cathedrals can be classified according to their architectural style, either Romanesque or Gothic. Some cathedrals have both a Gothic and Romanesque part, each of different height; these cathedrals are included twice. (Data fra Weisberg (1985): *Applied Linear Regression.*)

Romanesque cathedral	height feet	length feet	*Gothic* cathedral	height feet	length feet
Durham	75	502	York	100	519
Canterbury	80	522	Bath	75	225
Gloucester	68	425	Bristol	52	300
Hereford	64	344	Chichester	62	418
Norwich	83	407	Exeter	68	409
Peterborough	80	451	Gloucester	86	425
St. Albans	70	551	Lichfield	57	370
Winchester	76	530	Lincoln	82	506
Ely	74	547	Norwich	72	407
			Ripon	88	295
			Southwark	55	273
			Wells	67	415
			St. Asaph	45	182
			Winchester	103	530
			Old St. Paul	103	611
			Salisbury	84	473

Our aim is to investigate the relation between height and length, and whether this relationship depends on architectural style. Analyse the logarithm of the length as a function of the logarithm of the height, and eliminate the cathedrals of Bath and Ripon from the analysis. (Here we take these choices for granted, but you are invited to analyse the data from scratch yourself...)

1) Formulate a statistical model where the (logarithmic) height enters as a regression variable and the architectural style as a factor. Based on this model, do the relations between height and length seem to differ for the two architectural styles?

2) Extend the model with a quadratic term — in ln(height), and repeat question 1). Compare the answers.

3) Examine the two models more closely using regression diagnostics and model checks, and possibly by analysing additional models. Do you understand better now the results obtained in 1) and 2)?