Assignment 2

Start Assignment

Due Wednesday by 23:59   **Points** 10   **Submitting** a file upload   **File types** pdf
Attempts 0   **Allowed attempts** 2   **Available** until 31 Mar at 23:59

- *Throughout this assignment tests should be performed using a confidence level =0.05, unless otherwise specified.*
- *Where appropriate, motivate your answers and check the model assumptions by using relevant diagnostic tools.*

Exercise 1. Trees

The *Amsterdamsche Bos* forestry wishes to estimate the total wood volume of the trees on its domain. To this end the forestry has cut a sample of 59 trees of their most prevalent types beech and oak and collected the data from the cut trees in the file **treeVolume.txt (https://canvas.vu.nl/courses/68030/files/6051344?wrap=1)** ↓ **(https://canvas.vu.nl/courses/68030/files/6051344/download?download_frd=1)** . The volume of these trees alongside with their height and trunk diameter have been measured; these are the columns *volume*, *height* and *diameter*, respectively. Column *type* gives the tree type: *Beech* or *Oak*. The tree type, height and diameter can be measured in the field without sacrificing the tree. The forestry hypothesizes that these are predictive of the wood volume.

a) Investigate whether the tree type influences volume by performing ANOVA, without taking diameter and height into account. Can a t-test be related to the above ANOVA test? Estimate the volumes for the two tree types.

b) Now include *diameter* and *height* as explanatory variables into the analysis. Investigate whether the influence of diameter on volume is similar for the both tree types. Do the same for the influence of height on volume. (Consider at most one (relevant) pairwise interaction per model.) Comment.

c) Using the results from c), investigate how diameter, height and type influence volume. Comment. Using the resulting model, predict the volume for a tree with the (overall) average diameter and height?

d) Propose a transformation of the explanatory variables that possibly yields a better model (verify this). (Hint: think of a natural link between the response and explanatory variables.)

Exercise 2. Expenditure on criminal activities

The data in **expensescrime.txt** (https://canvas.vu.nl/courses/68030/files/6051345?wrap=1) ↓ (https://canvas.vu.nl/courses/68030/files/6051345/download?download_frd=1) were obtained to determine factors related to state expenditures on criminal activities (courts, police, etc.) The variables are: *state* (indicating the state in the USA), *expend* (state expenditures on criminal activities in $1000), *bad* (crime rate per 100000), *crime* (number of persons under criminal supervision), *lawyers* (number of lawyers in the state), *employ* (number of persons employed in the state) and *pop* (population of the state in 1000). In the regression analysis, take *expend* as response variable and *bad*, *crime*, *lawyers*, *employ* and *pop* as explanatory variables.

a) Make some graphical summaries of the data. Investigate the problem of influence points, and the problem of collinearity.

b) Fit a linear regression model to the data. Use the step-up method to find the best model. Comment.

c) Determine a 95% prediction interval for the *expend* using the model you preferred in b) for a (hypothetical) state with *bad*=50, *crime*=5000, *lawyers*=5000, *employ*=5000 and *pop*=5000. Can you improve this interval?

d) Apply the LASSO method to choose the relevant variables (with default parameters as in the lecture and *lambda=lambda.1se*). (You will need to install the R-package *glmnet*, which is not included in the standard distribution of R.) Compare the resulting model with the model obtained in b). (Beware that in general a new run delivers a new model because of a new train set.)

Exercise 3. Titanic

In April 15, 1912, British passenger liner Titanic sank after colliding with an iceberg. There were not enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew. The data file **titanic.txt** (https://canvas.vu.nl/courses/68030/files/6051346?wrap=1) ↓ (https://canvas.vu.nl/courses/68030/files/6051346/download?download_frd=1) gives the survival status of passengers on the Titanic, together with their names, age, sex and passenger class. About half of the ages for the 3rd class passengers are missing, although many of these could be filled in from the original source (so for many models you will need to work only with a part of the data, that is, with the rows for which the age is provided). The columns: *Name* – name of passenger; *PClass* – passenger class (1st, 2nd or 3rd), *Age* – age in years, *Sex* – male or female, *Survived* – survival status (1=Yes or 0=No).

a) Study the data and give a few (>1) summaries (graphics or tables). Fit a logistic regression model (no interactions yet) to investigate the association between the survival status and the predictors *PClass*, *Age* and *Sex*. Interpret the results in terms of odds, comment.

b) Investigate the interaction of predictor *Age* with *PClass*, and the interaction of *Age* with *Sex*. From this and a), choose (and justify) a resulting model. For this model, report the estimate for the probability of survival for each combination of levels of the factors *PClass* and *Sex* for a person of age 55.

c) Propose a method to predict the survival status and a quality measure for your prediction and describe how you would implement that method (you do not need to implement it).

d) Another approach would be to apply a contingency table test and to investigate whether factor passenger class has an effect on the survival status and whether factor gender has an effect on the survival status. Implement the relevant test(s).

e) Is the second approach in d) wrong? Name both an advantage and a disadvantage of the two approaches, relative to each other.

Exercise 4. Military coups

To study the influence of different political and geographical variables on the number of military coups, these data are collected for several countries in the file **coups.txt (https://canvas.vu.nl/courses/68030/files/6051349?wrap=1)** ↓ **(https://canvas.vu.nl/courses/68030/files/6051349/download?download_frd=1)** . The meaning of the different variables:

- *miltcoup* — number of successful military coups from independence to 1989,
- *oligarchy* — number years country ruled by military oligarchy from independence to 1989,
- *pollib* — political liberalization (0 = no civil rights for political expression, 1 = limited civil rights for expression but right to form political parties, 2 = full civil rights),
- *parties* — number of legal political parties in 1993,
- *pctvote* — percent voting in last election,
- *popn* — population in millions in 1989,
- *size* — area in 1000 square km,
- *numelec* — total number of legislative and presidential elections,
- *numregim* — number of regime types.

Remark: do not include interaction terms in models when answering below questions.

a) Perform Poisson regression on the full data set, taking *miltcoup* as response variable. Comment on your findings.

b) Use the step-down approach (using output of the function *summary*) to reduce the number of explanatory variables. Compare the resulting model with your findings in a).

c) Using the model from b), predict the number of coups for a hypothetical country for all the three levels of political liberalization and the (overall) averages of all the other (numerical) characteristics. Comment on your findings.

EDDA: Assignment 2

Criteria	Ratings	Pts
Exer.1 a) 0.5 b) 0.5 c) 0.7 d) 0.6		2.3 pts
Exer.2 a) 0.6 b) 0.7 c) 0.5 d) 0.5		2.3 pts
Exer.3 a) 0.6 b) 0.6 c) 0.5 d) 0.5 e) 0.4		2.6 pts
Exer.4 a) 0.6 b) 0.7 c) 0.5		1.8 pts
General layout		1 pts
		Total points: 10