Chapter 8: Regression models

1 Session 1

Objective: simple and multiple linear regression; confidence and prediction intervals, tests on slope, on predicted values.

Exercise 1.1. Upload `bosson.csv`, read data description. Assign columns `aneurysm`, `bmi`, to variables `A`, `B`.

1. Represent the linear regression of `A` onto `B`. Give the value of R-squared. Test significance and conclude. Represent the linear regression.

2. Display typical plots of residuals. Comment on the validation of the model.

3. Give a confidence interval and a prediction interval for the aneurysm size of a patient with bmi equal to 25.

4. Test the pertinence of the regression of `A` onto `B`, for the subpopulations of women, of men, of Vietnamese, of French, of Vietnamese women, men, of French women, men. Can you explain why certain regressions are significant, others are not?

5. Would you use bmi as a predictor of aneurysm size?

Exercise 1.2. Upload `tauber.csv`, read data description. Assign columns `gender`, `age`, `height`, `weight` to variables `G`, `A`, `H`, `W`.

1. Represent the linear regressions of `H` onto `A`, on the global population, on girls, on boys. Compare the 95%-confidence interval on the average height and the 95%-prediction interval for the height of one individual, for a six year old child, a six year old girl, a six year old boy. Why are confidence intervals narrower on the global population? Why do the three prediction intervals have similar amplitudes?

2. Represent the linear regressions of `W` onto `A`, on the global population, on girls, on boys. Compare the 95%-confidence interval on the average weight and the 95%-prediction interval for the weight of one individual, for a six year old child, a six year old girl, a 6 year old boy. Do you observe the same as in the previous question?

3. A six year old boy weighs 17 kg. Should his mother worry? Repeat for a six year old girl.

4. Represent the linear regressions of `W` onto `H`, on the global population, on girls, on boys. Compare the 95%-confidence interval on the average weight and the 95%-prediction interval for the weight of one individual, for a 115 cm child, a 115 cm girl, a 115 cm boy. Are confidence intervals narrower than those of the previous question? Are prediction intervals narrower than those of the previous question?

5. A 115 cm tall boy weighs 17 kg. Should his mother worry? Repeat for a 115 cm tall girl.

6. As a predictor of weight, would you choose age or height?

7. Compute the linear regression of `W` onto `A`, onto `H`, onto `A` and `H`. Display summaries, compare values of adjusted R-squared.

8. Compare with an anova, the linear regression of `W` onto `A`, to the linear regression of `W` onto `A` and `H`. Compare with an anova, the linear regression of `W` onto `H`, to the linear regression of `W` onto `A` and `H`.

9. Which model explains best the response?

2 Session 2

Objective: multiple linear regressions in high dimension; selection of variables.

Exercise 2.1.

1. Upload `LenzT.rds` and `LenzI.rds`. Assign column `follup` to variables `YF`. Sample randomly 500 genes from the `LenzT` matrix and assign to matrix `X`.

2. Apply a linear regression model of `YF` onto `X` with function `lm`. How many coefficients are estimated ? How many are set to `NA` ?

3. Install package `glmnet` with instruction `library(glmnet)`. Run a linear regression model penalised by lasso of `YF` onto `X` with function `glmnet`. Plot the regularisation path, add the labels. Are they any coefficients set to zero only for large regularization parameter λ ?

4. Select the regularization parameter λ by cross-validation using function `cv.glmnet`. Print the non nul coefficients. How sparse is the vector of coefficients ?

5. Run the same procedure with the full matrix `LenzT.rds`. How many coefficients are non nul ? Does the set obtained with the full matrix contain the set of the previous question ? Run again the cross-validation procedure with the full matrix. Is the new set close to the previous one ?

6. Run the cross-validation procedure with a number of sub-samples `nfolds=15` ? Is it slower to run ? Comment the results.

Exercise 2.2.

1. Upload `LenzT.rds` and `LenzI.rds`. Assign column `ecog` to variables `Ye`. `Ye` has missing values. Remove the patients with missing values in `Ye` and assign to variable `Yec`. Sample randomly 500 genes from the `LenzT` matrix and assign to matrix `X` the selected genes for patients without missing values in `Ye`.

2. Apply a linear regression model of `Yec` onto `X` with function `lm`. How many coefficients are estimated ? How many are set to `NA` ?

3. Install package `glmnet` with instruction `library(glmnet)`. Run a linear regression model penalised by lasso of `Yec` onto `X` with function `glmnet`. Plot the regularisation path, add the labels. Are they any coefficients set to zero only for large regularization parameter λ ?

4. Select the regularization parameter λ by cross-validation using function `cv.glmnet`. Print the non nul coefficients. How sparse is the vector of coefficients ?

5. Run the same procedure with the full matrix `LenzT.rds` (removing patients with missing values `ecog`). How many coefficients are non nul ? Does the set obtained with the full matrix contain the set of the previous question ? Run again the cross-validation procedure with the full matrix. Is the new set close to the previous one ?

3 Session 3

Objectives: on your project dataset, linear regression between continuous phenotypes variables and genomics variables.

Steps of the session:

1. Team work: Study the link between
   - each continuous phenotype variable and the other variables with a multiple linear regression model (using function `step` to select the variables)
   - each continuous phenotype variable and the genomics variables on your project dataset. Use lasso penalisation approach to select the most important and influent genes.
   - compare your results with published results
2. Team work: Write the report corresponding to your results
3. Individual work at home: Read the chapter *Regression models*, sections *Logistic regression and survival regression.*

4 Session 4

Objective: multiple logistic regressions in high dimension; selection of variables.

Exercise 4.1.

1. Upload `titanic.csv`. Assign columns `pclass`, `survived`, `gender` to variables `P`, `S`, `G`. Transform `G` in a binary variable equal to 1 for female.
2. Apply a logistic regression model of `S` onto `G` with function `glm`. Compute the odds-ratio for women against men.
3. Apply a logistic regression model of `S` onto `P`. Compute the odds-ratio for third class passengers against second class passengers.
4. Apply a logistic regression model of `S` onto `P` and `G`. Compute the two odds-ratio and compare to the previous ones. Comment.
5. Select the best model with function `step`. Comment the best model.

Exercise 4.2.

1. Upload `LenzI.rds`. Assign columns `status`, `gender`, `diagno`, `ecog`, `age`, `regim`, `stage`, `ldhrat` to variables `S`, `G`, `D`, `E`, `A`, `R`, `St`, `L`. Transform `S` in a binary variable equal to 1 for dead subjects.
2. Apply a logistic regression model of `S` onto `G` with function `glm`. Comment the odds-ratio.
3. Apply a logistic regression model of `S` onto `D` with function `glm`. Comment the odds-ratio.
4. Apply a logistic regression model of `S` onto `E`, onto `E` and `D`, onto `E` and `D` with an interaction between both predictors. Select the best model. Comment the odds-ratio.
5. Assign column `extnod` to variables `Ex`. Remove patients with `Ex` missing values in all the variables.
6. Apply a logistic regression model of `S` onto `E`, `D`, `A`, `R`, `St`, `L`, `Ex`. Select the best model. Comment.

Exercise 4.3.

1. Upload `LenzT.rds` and `LenzI.rds`. Assign column `status` to variable `Ys`. Sample randomly 500 genes from the `LenzT` matrix and assign to matrix `X`.
2. Install package `glmnet` with instruction `library(glmnet)`. Run a logistic regression model penalised by lasso of `Ys` onto `X` with function `glmnet`. Plot the regularisation path, add the labels. Are they any coefficients set to zero only for large regularization parameter λ ?

3. Select the regularization parameter λ by cross-validation using function `cv.glmnet`. Print the non nul coefficients. How sparse is the vector of coefficients ?

4. Run the same procedure with the full matrix `LenzT.rds`. How many coefficients are non nul ? Does the set obtained with the full matrix contain the set of the previous question ?

Exercise 4.4.

1. Upload `LenzT.rds` and `LenzI.rds`. Assign columns `status`, `follup`, `gender` to variables `S`, `Fo`, `G`. Transform `S` in a binary variable equal to 1 for dead subjects.

2. Install package `survival`. Apply a Cox model of `Fo` onto `G` with function `coxph`. Compute the hazard ratio of women with respect to men. Comment.

3. Sample randomly 500 genes from the `LenzT` matrix and assign to matrix `X`. Set nul values of `Fo` to `0.01`. Install package `glmnet` with instruction `library(glmnet)`. Run a Cox model penalised by lasso of `Fo` onto `X` with function `glmnet`. Plot the regularisation path, add the labels. Are they any coefficients set to zero only for large regularization parameter λ ?

4. Select the regularization parameter λ by cross-validation using function `cv.glmnet`. Print the non nul coefficients. How sparse is the vector of coefficients ?

5. Run the same procedure with the full matrix `LenzT.rds`. How many coefficients are non nul ? Does the set obtained with the full matrix contain the set of the previous question ?

5 Session 5

Objectives: on your project dataset, logistic regression and cox models between phenotypes variables and genomics variables.

Steps of the session:

1. Team work: Study the link between
   - each binary phenotype variable and the other variables with a multiple logistic regression model (using function `step` to select the variables)
   - each binary phenotype variable and the genomics variables on your project dataset. Use lasso penalisation logistic model to select the most important and influent genes.
   - a (censored) duration time and the other variables with a multiple cox model (using function `step` to select the variables)
   - a (censored) duration time and the genomics variables on your project dataset. Use lasso penalisation cox model to select the most important and influent genes.
   - compare your results with published results
2. Team work: Write the report corresponding to your results