**06/2024**

**Alonso part**[[edit](https://lbk.be/examenwiki/index.php?title=Statistical_Methods_for_Bioinformatics&action=edit&section=6)]

Educational study: In educational research study a new and standard arithmetic training programs were compared. The objective of the study was to evaluate if as an average the new program can produce a faster increase in arithmetic ability over time than the standard program for third year primary school children. To that end a total of 200 children were randomized to follow the new or standard program for 5 months and their arithmetic ability was assessed using a validated test (SAT test) with larger values indicating higher arithmetic ability. All the students were assessed at the beginning of the study (after randomization but just before starting the training) and then monthly during the duration of the study. No missing values were present in the data. The final data set had the following variables:

* id: a number identifying the student
* y: response variable containing the scores of the SAT test
* time: evaluation moments 0 (before starting the training) 1 2 3 4 and 5 months after starting the training
* prog: indicator variable for the program taking values 1 for the new training program and 0 for the standard one

**Fit the following hierarchical model to the data using maximum likelihood:**

*Y ij = π0i + π1i timej + ϵij*

where i denotes the student j the measurement and

*ϵij ∼ N (0 σ^2\_ϵ)*

and

*π0i = γ00 + γ01PROGi + b0i*

*π1i = γ10 + γ11PROGi*

and

*b0i ~ N (0 σ^2ϵ)*

Based on the point estimates obtained from the previous model which of the following statements is correct?

* The variability between individuals is larger than the variability within individuals.
* The variability between individuals is smaller than the variability within individuals.
* The variability between individuals is equal to the variability within individuals.
* The previous model does not allow to answer this question.

Based on the previous model which of the following statements is correct?

* As an average the new program is significantly better than the standard program.
* As an average the new program is significantly worse than the standard program.
* As an average based on the point estimates the new program seems to be worse than the standard program but the result is not significant.
* As an average based on the point estimates the new program seems to be better than the standard program but the result is not significant.

Based on the previous model which of the following statements is true?

* The point estimate for the main effect of program is negative and significant.
* The point estimate for the main effect of program is negative and not significant.
* The point estimate for the main effect of program is positive and significant.
* The point estimate for the main effect of program is positive and not significant.

Based on the previous model what is the average SAT score for children in the standard program one month after starting the experiment?

* 6.32
* 5.65
* 0.48
* 6.8

Based on the previous model which of the following statements is correct?

* The p-value associated with γ01 is smaller than 5%
* The p-value associated with γ01 is between 5% and 80%
* The p-value associated with γ01 is larger than 90%
* None of the above

**Rob part**[[edit](https://lbk.be/examenwiki/index.php?title=Statistical_Methods_for_Bioinformatics&action=edit&section=7)]

**Theory Exercise:**

* Q.1: Rank (lasso, gam, linear model, random forest ) by sensitivity to outliers
* Q.2: Rank (Wald test on mixed effects model, validation set, cross validation, bootstrap ) by how they can estimate bias of model
* Q.3: Open considerations you should consider in high dimension data
* Q.4.a: When would you use gam instead of glm.
* Q.4.b: What are the formulations of gam in comparison to glm (really a strange question, it was not formulated so precisely)

**Practical Exercise:**

The exercise is open book you can use additional information sources (but not your laptop or other hardware). Any form of communication is forbidden and communication by the PCs is monitored; this includes pages such as Google docs that facilitate discussions; any violation will lead to immediate exclusion from the exam. The use of AI such as ChatGPT is not permitted during the exam. You will have an additional two hours for the practical part (total limit three hours). Prepare a file with your answers: explain and motivate your answers and provide your code. When you have finished upload the file as assignment to Toledo and signal the examiner to confirm your submission. The data file can be found attached to the Toledo assignment.

Data: Cortex2.rdata

The dataset contains the expression levels of 71 proteins measured in the cerebral cortex of 8 classes of control and Down syndrome mice exposed to context fear conditioning a task used to assess associative learning. The eight classes of mice are described based on genotype (control or trisomic) behavior (stimulated to learn (context-shock) or not (shock-context)) and treatment (saline or drug memantine). The original experiment aimed to test the effect of the drug memantine in recovering the ability to learn in trisomic mice.

1.Study and describe the data. Do you see indications of potential issues when statistically modeling the data? Explain.

2. Train and compare a GAM model and a boosting model to separate the Memantine from Saline treatment (the Treatment variable) samples based on protein expression. Interpret the results of the optimizations; at least discuss:

* Do correlations between variables influence the results?
* Is over-learning a problem for finding the optimal model?
* Is there evidence for non-linear effects?
* Is there evidence for important interactions between variables?

3.Evaluate if Memantine treatment induces any specific changes in protein levels of trisomic mice who were stimulated to learn.

Exam questions (2020)[[edit](https://lbk.be/examenwiki/index.php?title=Statistical_Methods_for_Bioinformatics&action=edit&section=8)]

**12/06/2020** *Theory* (closed book, 1 hour, only written, no oral defense)

Prof. Rob Jelier

1) WHat is bootstrapping, how does it work, when to use it instead of CV.

2) What is boosting explain (with deviations)

3) Discuss Bias-variance trade off, how does it differ between LOOCV and 10-fold?

Prof. Alonso Abad

Output of LM model, asking multiple choice questions (really need to understand everything of the output):

1) What does the F-test refer to? a) B0= B1=B2=B3=0 b) B0= B1=B2=B3 c) B1=B2=B3=0 d) B1=B2=B3=0

2) give the predicted value for a female with wage that is equals to the average plus two time standard deviation ,...

3) To which model is the third-line of the anova model refering to? female + wage <-> female + wage + female:Wage <-> ...

4) Aikake weights, scores, deviation, ... given: Is there certainty that there is one correct model

5...

*Exercises* (open book on pc, 2 hours, make shure to refresh your memory on basic R code commands, you might have to do some table modifications)

Prof. Alonso Abad

Make a multilevel model, fill in table of output + p-values, calculate bootstrapping, discuss: does the variable have a significant impact? Is it positive? is it negative?

Prof. Dr.Jelier

Was similar to the vanDeVijver execrise 1) explore the data, are there any potential issues?

2) create lasso and ridge model, discuss preformance, is overlearning a problem, can a subset of variables have good predictive power, ...

3) (given the categorical variable X): does X influence the protein expression levels when it comes to predictive performance for the genotype and treatment? (was an open question with many possibilities to solve, he said)

Exam questions (2016)[[edit](https://lbk.be/examenwiki/index.php?title=Statistical_Methods_for_Bioinformatics&action=edit&section=9)]

**15/06/2016**

*Theory* (closed book, 1 hour, only written, no oral defense)

Prof. Alonso Abad

Output of a longitudinal model about the effect of medication. Three test questions: What are the level 2 models? Which fixed effect we would expect to be 0? Does the new medication have a better influence on the progress of the patient compared to the old one?

Prof. Rob Jelier

Define smoothing and cubic spline. How to control their degrees of freedom? Give 3 considerations when analyzing a dataset with many predictors. What are bootstrap and bagging methods? Why does random forests outperform both?

*Exercises* (open book on pc, 2 hours)

Prof. Alonso Abad

Inverse probability weighting. Very similar to the titanic exercise in the slides. You don’t have to send the script, but complete the exam (it has tables and you have to write the numbers there).

Prof. Dr.Jelier

Cancer dataset Explore the data (correlation, histograms, boxplots,...). What challenges does it present? Do Lasso and Ridge. Test performance. Do gam. Is it better than the above models?

**07/06/2016**

*Theory* (closed book, 1 hour, only written, no oral defense)

Prof. Alonso Abad

Output of a longitudinal model. Three test questions: What are the level 2 models? Which fixed effect we would expect to be 0? Comparing between kids in the old and in the new program, the kids in the new program have less, same or bigger slope?

Prof. Rob Jelier

Describe three ways of deciding between linear or non-linear model. For one of them, discuss the bias-variance trade-off What are regression splines? How do they relate to cubic splines? Discuss Lasso, Ridge and PCR. When to use each and what are the differences between them?

*Exercises* (open book on pc, 2 hours)

Prof. Alonso Abad

Multiple imputation. Very similar to the titanic exercise in the slides. You don’t have to send the script, but complete the exam (it has tables and you have to write the numbers there). Remember to set the seed he says in the exam, otherwise you’ll get different results!

Prof. Dr.Jelier

Prostate dataset, with gleasonBin as response. Explore the data. What challenges it present? Do forward selection. (Careful here, it is logistic regression so you cannot use regsubsets. Write it yourself or use bestglm or step functions). Test performance. Do Lasso. Compare with the previous model. Test performance. Do gam. Is there evidence for non-linear relationship?

Exam questions (2015)[[edit](https://lbk.be/examenwiki/index.php?title=Statistical_Methods_for_Bioinformatics&action=edit&section=10)]

**08/2015**

*Theory* (closed book, 1 hour, only written, no oral defense)

Prof. Peter Goos

1. What are the three components of GLM? Give three examples of applications of GLM and explain what are the three components in your examples.
2. How do you make a regression on qualitative data? Give two examples for a qualitative variable with 3 levels that are ready to use in a regression

Prof. Rob Jelier

1. When would you use a linear or nonlinear model ( at least 2 considerations) and explain bias-variance tradeoff and how you would consider it in the above question.
2. Explain the differences between ridge, lasso, and PCR. Explain when you use each.
3. What is the bagging and boosting? Why does random forest outperform these two methods?

*Exercises* (open book on pc, 2 hours)

Prof. Peter Goos

Data set given has variables: day, flow, screw speed, moisture, inflation index. User is interested in model with interaction effect and quadratic effect.

1. Build a model with the data set using inflation index as response variable.
2. What is the flow and moisture to achieve inflation index of 12 when the screw speed is fixed at 200? (Use profiler)
3. What is the flow and moisture to achieve inflation index of 12 when the screw speed is fixed at 400? (Use profiler)
4. Interpret the significant interaction effect of the model

Prof. Rob Jelier

Given a microarray data set with 184 observations and 4849. Interested to use gene expression to predict distant metastasis (DM) or no distant metastasis (NODM) in cancer patients.

1. Explore the data, what could be the challenge to model this data set?
2. ...
3. Is there correlation within the data set? How will this affect the model?
4. Use LASSO to build a model using column 1(two outcomes: DM or NODM) as response variable. How many genes does your final model uses to predict the response?
5. How well does your model perform?

**09/06/2015**

*Theory* (closed book, 1 hour)

Prof. Peter Goos

1. What are the three components of a GLM? Give three examples of applications of GLM and what the components would be in your examples.
2. How do you make a regression on qualitative data? Give two examples for a qualitative variable with 3 levels that are ready to use in a regression

Prof. Rob Jelier

1. When would you use a linear or non linear model ( at least 2 considerations) and explain bias-variance trade off and how you would consider it in the above question.
2. Explain the differences between ridge, lasso, and PCR. Explain when you use each.
3. What is bagging and does it overlearn.

*Exercises* (open book on pc, 2 hours)

Prof. Peter Goos

1. Make a model of this data set
2. What is the are the effect variables when the response is 12 (use profiler).
3. What is the significance of the interaction effect in this model.

Prof. Rob Jelier

1. Dataset about influence of age and various SNPs on the change in muscle mass after sport (CH). [data set has more than 100 variables] Is there evidence for the non-random influence of age on CH? Is there evidence for a non-linear influence of age on CH?
2. What would pose difficulties of making a model that predicts the response in this data set?
3. Do forward, ridge, and lasso. Compare results, what is the best?
4. Make a correction for the effect of the age variable on the response. What changes? Does this have a significant effect on the model?