June 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) a) B0= B1=B2=B3=0 b) B0= B1=B2=B3 c) B1=B2=B3=0 d) B1=B2=B3=0

- I think this is supposed to be 'what does the null hypothesis in an F-test correspond to' and that would be c or d, I think.
- All Bk in the model equal to 0
- F test refers to the fact that the model as a whole is significant or not right?
- What's the difference between c and d? → I assume one should be without the 0, just all equal
- I think the answer is indeed all Bk zero except for the intercept (B0) I agree!

2) give the predicted value for a female with wage that is equal 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

Exercises (open book on pc, 2 hours, make sure 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 exercise 1) explore the data, are there any potential issues?

2) create lasso and ridge model, discuss performance, 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)

June 2017

Theoretical part: 1 hour, close book, only written, no oral defense
Practical part: 2 hours, open book, script/doc to be uploaded to link provided

Theoretical part

Prof. Dr. Abad

Output of a longitudinal model. 5 test questions: What are the level 2 models? …….

Prof. Dr.Jelier
1.
2. Talk about bagging and boosting, what are they, compare with GAM in terms of flexibility
3. What effects do you see when you cross validate- discuss in terms of bias variance tradeoff. How does it differ in LOOCV or in 10 fold cross validation?

Practical part

Prof. Dr. Abad

Multiple imputation AGAIN. 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

Dataset linear regression

5. Explore the data. What challenges it present?
6. Make a Lasso model
7. How does the coefficient of lcoval in the lasso model compare with its correlation to the response? Does it make sense? explain
8. Do gam. Can you build a better model with generalized additive models? Compare lasso and gam models

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.

1. Explore the data. What challenges it present?
2. 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.
3. Do Lasso. Compare with the previous model. Test performance.
4. Do gam. Can you build a better model with GAM ?Is there evidence for non-linear relationship?

August 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?

June 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?

(August 2014)
1a. How to select linear model or non-linear model, explain bias-variance trade-off and two additional things.
1b. Difference between best subset selection, ridge regression and PCR and when to use which method.
1c. assumption of linear model. correlation variables.

2a. lasso2 package, Prostate data. explore the whole data.
2b. full logistic model, backward selection.
2c. lasso, logistic model. compare with the model of 2b

DOE part
1. concept of factorial design. relationship with linear model. problem of multicollinearity.
2. use a table to show factorial design. if budget insufficient, how to make the design. How will it influence the model

(June/2014)

Two questions from prof. Rob Jelier:
First one didn't require R, it was about splines but can't remember the subquestions. There were 4 parts; but all I remember is "How do you control variance and bias in splines?" They were all "theoretical" questions though (no explicit formulas).
Second question was on a dataset in R. a) explore the dataset and comment on the correlations. b) do a variable selection though backwards subset selection; ridge regression and lasso. The latter two need to be done with cross validation. c) Why do we use cross validation there? (hint: because you said we had to is not a valid answer) d) make the proposed models and check their performance IN A MEANINGFUL WAY (*wink* *nudge*) e) was variable gkl included in the model? Why (not)?

Two questions from Experimental Design:
1) Write a 250 word essay that would explain what the course was about to a classmate who didn't take it. Use the following keywords: optimality, orthogonality, one factor at a time, …
(In another exam it was a similar question only that to explain for someone that is familiar with linear regression and how is related with that)
2) Calculate how many observations we want to do for two different experiments with different variance and different costs. It's in the slides and dead easy.... if you looked at it beforehand, which I didn't. Basic math was too hard.

(June/2014) Second version

the questions of the last part were to give an essay of 250 words about optimal design of experiments with some keywords in it like orthogonality, optimal and a couple more. another question was a simple exercise from the first slides and you could use excel for it

the first part had a number of theoretical questions which were all quite to be expected and an exercise in R.