ACSC/STAT 3740, Predictive Analytics

WINTER 2025

Toby Kenney

Homework Sheet 3

Due: Friday 14th March: 13:00

Note: This homework assignment is only valid for WINTER 2025. If you find this homework in a different term, please contact me to find the correct homework sheet.

Standard Questions

1. An insurance company has collected the following data on life expectancy in the file HW3Q1.

Variable	Meaning
current.age	The individual's current age.
sex	The individual's current sex.
BMI	The individual's BMI
cigarettes.per.day	The average number of cigarettes the individual smokes each day
daily.exercise	The average number of minutes per day spent doing physical exercise
health.index	An index measuring overall health
survival.five.year	Whether the individual survives 5 years

   Fit a generalised linear model, with a binomial response variable (and a logistic link function), to predict the probability of dying within 5 years. Use this model to predict the probability of dying for the individuals in the file HW3Q1test .

2. A company is analysing data on the effect of maintainance on productivity in the file HW3Q2.

Variable	Meaning
machine.age	The age of the machine.
machine.operators	The number of workers operating the machine.
machine.preemptive.maintainance	The amount spent on pre-emptive maintainance of the machine over the past year.
machine.corrective.maintainance	The amount spent on corrective maintainance of the machine over the past year.
machine.power	The power consumed by the machine.
machine.output	The number of parts produced by the machine.
machine.defect.rate	The proportion of part output by the machine that are defective.

Fit a random forest to predict the machine defect rate from the other predictors. Use this model to predict defect rates for the machines in the file HW3Q2test .

3. The file `HW3Q3.txt` contains measurements of the total annual rainfall in a certain city over the last century

   (a) Fit a quadratic model to estimate log annual rainfall as a function of time.

   (b) Use AIC to fit the best ARMA model to the residuals of the quadratic model.

   (c) Fit a GARCH model to model the variance.

   (d) Based on this model, what is the probability that average annual rainfall will exceed 2500 in the decade from 2090 to 2099? [You can use the `ugarchboot` function to run a simulation to estimate this.]

4. The file `HW3Q4.txt` contains the following data about school performances in standardised tests for Grade 8:

Variable	Meaning
no.students	The number of students in Grade 8 attending the school.
teacher.student.ratio	The average number of students per teacher in a class at the school.
funding	The schools source of funding — government, independent or private.
specialist.teacher	Whether the school employs teachers with specialist knowledge for each subject.
teacher.5.years	The percentage of teachers at the school with at least 5 years of experience.
parent.employment	The percentage of parents of children at the school who are employed.
median.parent.salary	The median salary of parents of children at the school
mean.parent.education	The average number of years of full-time education of parents of children at the school.
average.score.mathematics	The average score of children in Grade 8 at the school on the standardised mathematics test.
average.score.english	The average score of children in Grade 8 at the school on the standardised English test.

Fit generalised additive models with Gaussian response and identity link function to predict `average.score.mathematics` and `average.score.english` from the other predictors.

5. A company has collected the following data on employee training effectiveness in the file HW3Q5.

Variable	Meaning
training.type	The type of training.
compulsory	Whether the training was compulsory for the employee.
employee.experience	The number of years of experience of the employee.
employee.salary	The employee's annual salary.
employee.gender	The employee's gender.
work.type	The type of work.
training.time	The amount of time spent on the training.
productivity.before	The employee's productivity rating before the training.
productivity.after	The employee's productivity rating after the training.

Fit a linear model, using LASSO for variable selection and regularisation to predict sales from the other predictors. Use this model to predict sales for the scenarios in the file HW3Q5test .