ACSC/STAT 3740, Predictive Analytics

WINTER 2023

Toby Kenney

Homework Sheet 4

Due: Thursday 23rd March: 11:30

Note: This homework assignment is only valid for WINTER 2023. If you find this homework in a different term, please contact me to find the correct homework sheet.

Note: All data sets in this homework are simulated.

Standard Questions

1. The file `HW4Q1.txt` contains data on the relation between economic policy and child poverty rates. The data set contains the following variables:

Variable	Meaning
base.tax	The lowest rate of income tax
top.tax	The highest marginal rate of income tax
gdp	The per.capita gdp
free.health	Whether the country has government-provided healthcare
free.school.years	Number of years of government-funded education
free.higher.edu	Whether the government funds higher education.
child.poverty	The percentage of children living in poverty

A data analyst uses the following code to fit a linear regression model to the data.

```
HW4Q1<-read.table("HW4Q1.txt")
HW4Q1_linear<-lm(child.poverty~.,data=HW4Q1)
```

Use appropriate diagnostics to assess how appropriate the assumptions of the linear regression model are. What changes would you suggest making to the model to better model the data?

2. A data scientist at a car manufacturing company is analysing data about engine efficiency in the file `HW4Q2.txt`.

Variable	Meaning
cylinder.number	The number of cylinders
fuel.type	Regular, premium, diesel or electric
vehicle.weight	The weight of the vehicle.
vehicle.speed	The speed at which the vehicle is being driven
vehicle.make	The manufacturer of the vehicle
mpg	The vehicles miles per gallon

He has fitted a linear model to predict mpg, using the code in the file `HW4Q2_linear.R`. Perform diagnostics to test which of the assumptions of this model are reasonable. What changes would you suggest making to the model to better model the data?

3. A scientist is reviewing data about the relation between the strength of a material and the production technique, in the file `HW4Q3.txt`.

Variable	Meaning
carbon.proportion	The proportion of carbon in the mixture
titanium.proportion	The proportion of titanium in the mixture
production.temp	The temperature used to produce the material
production.pressure	The pressure used to produce the material
cooling.time	The time period over which the mixture is allowed to cool
tensile.strength	The strength of the eventual material

She has fitted a generalised additive model, a random forest model and a generalised linear model including a number of interaction terms and polynomial terms, to predict the total damage, using the code in the file `HW4Q3_models.R`. Assess which of these models is better at predicting the data. [You may need to modify the code provided to do this.]

4. The file `HW4Q4.txt` contains data from an insurance company about the probability that a settlement offer is accepted. The data set contains the following variables:

Variable	Meaning
accident.year	The year of the accident
number.affected	The number of individuals affected by the accident
property.damage	The estimated amount of property damage.
injury.loss	The direct loss due to injury.
injured.sex	The sex of the injured party.
injured.age	The age of the injured party.
injured.salary	The salary of the injured individual.
settlement.amount	The amount of settlement offered.
settlement.accepted	Whether the settlement was accepted.

A data analyst uses the following code to fit a decision tree to the data:

```
Reaction_data<-read.table("HW4Q4.txt")

library(rpart)

Reaction_dt<-rpart(formula=reaction.time~.,
                   data=Reaction_data,
                   control=rpart.control(minbucket=10, # smallest size of node
                                         maxdepth=10)) # largest depth of tree.
```

and uses the following code to select variables using stepwise regression with AIC:

```
Reaction_Null_model<-lm(reaction.time~1,data=Reaction_data)

Reaction_Full_model<-lm(reaction.time~.,data=Reaction_data)


library(MASS)
Reaction_Forward<-stepAIC(Reaction_Null_model,
                          direction="forward",
                          scope=list(lower=Reaction_Null_model,
                                     upper=Reaction_Full_model))
```

The code is in the files `HW4_Q4_Decision_tree.R` and `HW4_Q4_Stepwise_AIC.R` respectively.

Based on the results of these analyses, how could he try to adjust the models to better fit the data?