

ACSC/STAT 3740, Predictive Analytics

WINTER 2023

Toby Kenney

Homework Sheet 3

Due: Thursday 9th 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.

Standard Questions

1. A music streaming company is building a recommendation system to suggest songs to its readers. It has collected the following data in the file `HW3Q1.txt`.

Variable	Meaning
<code>genre</code>	The genre (type of music) of the song
<code>artist</code>	The identifier of the artist.
<code>rating</code>	The songs average user rating (scale 1–5)
<code>same.artist</code>	A measure of how much the user listens to songs by the artist. (scale 0–5)
<code>same.genre</code>	A measure of how much the user listens to songs from this genre (scale 0–5)
<code>friend.listen</code>	The number of the users “friends” that listen to the song
<code>friend.recommend</code>	The average of the recommendation scores for the song give by the user’s friends
<code>listen</code>	Whether the user listens to the recommended song.

- (a) Fit a logistic regression model to predict whether the user will listen to the recommended song.
 - (b) The predictor `friend.listen` is skewed and heavy tailed. Try a log transformation and a square root transformation of this variable. Fit models including all combinations of these transformations.
2. The file `HW3Q2.txt` contains data from a study on the effect of exercise on the risk of heart disease in men. The variables included are

Variable	Meaning
age	The age of the patient
ave.weekly.exercise	The number of hours per week spent exercising.
weekly.cals	The number of calories consumed weekly.
percent.fat	The percentage of the patient's diet that consists of fats.
percent.fibre	The percentage of the patient's diet that consists of fibre.
fam.hist	Whether the patient has family history of heart disease.
BMI	The patient's BMI.
SBP	The patients systolic blood pressure.
heart.5.year	Whether the patient develops heart disease within the following 5 years.

Fit a decision tree to predict whether an individual will develop heart disease in the next 5 years.

3. The file `HW3Q3.txt` contains daily new influenza infections counts in a particular country.
 - (a) log-transform the counts and fit a seasonal trend using the function $\sin(2\pi t)$ and $\cos(2\pi t)$ where t is the time in years.
 - (b) After subtracting the seasonal trend, fit an ARMA model to the residuals, using AIC to determine the best choices for p and q .
 - (c) Fit a GARCH model to model the variance.
 - (d) Based on this model, what is the probability that there are fewer than 15000 flu cases in the first four months of 2023? [You can use the `ugarchboot` function to run a simulation to estimate this.]

4. A reinsurance company has collected the following data on earthquakes in the file `HW3Q4.txt`.

Variable	Meaning
magnitude	The magnitude of the earthquake on the Richter scale
population	The population of the affected city or region
distance	The distance of the epicentre from the affected area
depth	The depth of the epicentre
year	The year of the earthquake
years.since.5	The number of years since a magnitude 5 earthquake hit the same region
country.gdp	The annual per-capita gdp of the affected country
damage	The total damage caused by the earthquake

Fit generalised linear models to predict the probability that an earthquake will cause damage, and for an earthquake which does cause damage, to predict the total damage, using a gamma response variable and a log-link function.

Use these models to predict the total damage for the earthquakes in the file `HW3Q4_test.txt`.

5. A scientist has collected the following data on the effect of organic farming on butterfly populations. The data are in the file `HW3Q5.txt`.

Variable	Meaning
total.agriculture	The proportion of the habitat that is used for agriculture.
main.crop	The most grown crop in the region.
percent.organic	The proportion of agricultural land that uses organic farming methods.
ave.summer.temp	The average temperature during the summer months ($^{\circ}C$).
ave.winter.temp	The average temperature during the winter months ($^{\circ}C$).
rainfall	The average total annual rainfall.
year	The year.
butterflies	The number of butterflies caught in the region.

(a) Fit a decision tree to predict number of butterflies from the other variables. Choose an appropriate transformation for the response variable, and make any necessary adjustments to the data.

(b) Fit a random forest model to predict number of butterflies from the other variables. Test this model on the dataset in the file `HW3Q4_test.txt`.