ACSC/STAT 3740, Predictive Analytics

WINTER 2023

Toby Kenney

Homework Sheet 2

Due: Thursday 9th February: 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 in this homework are simulated.]

Standard Questions

1. The file `HW2Q1.txt` contains the following data

Variable	Meaning
population	The population of the district.
average.income	The average annual income of full-time workers aged 18–65
income.inequality	A measure of inequality in income in the region with 0 representing perfect equality and 100 representing maximal inequality.
percent.unemployed	The percentage of people aged 18–65 unemployed in the region.
education.level	The average number of years spent in full-time education
police.officers	Number of police officers per 100,000 inhabitants
government.spending	Amount of government spending on the district per inhabitant.
crime.rate	Number of crimes committed per 1,000,000 inhabitants.

These data were collected from government data in a variety of countries. Government spending, police officers and crime rate data were from government reports. Income data were from government tax data. Unemployment data were from government records of individuals claiming unemployment benefits. Population data is from the government census.

Perform data exploration on this data set, and summarise (with tables and plots to support where appropriate) your initial conclusions about data issues and appropriate models. You should take into account any concerns with data collection and processing.

2. The file `HW2Q2.txt` contains the following data from a university's international development office about trends in overseas student applications

Variable	Meaning
`country.of.residence`	The country of the applicant's residence
`year.of.application`	The year the application was made
`application.gpa`	The GPA at time of application
`applied.major`	The major to which the student applied
`time.to.outcome`	The time the student spent at the university
`outcome`	The result of the student's studies
`gpa`	The student's final GPA
`final.major`	The students declared major at the time they left the university

Perform data exploration on this data set, and summarise (with tables and plots to support where appropriate) your initial conclusions about data issues and appropriate models. You should take into account any concerns with data collection and processing.

3. The file `HW2Q3.txt` contains the following data from an electricity company, who are trying to forecast electricity demand in each city.

Variable	Meaning
city	The city being supplied.
year	The year.
population	The population at the time.
average.temp	The average daytime high temperature over the year.
rainfall	The total annual rainfall.
price.adj	CPI-adjusted price per KWh of electricity
consumption	Total electricity consumption

Electricity consumption is from the company's meters. Population is from government census data. Weather data is from government historical weather records. Pricing data is from the company's records, and the CPI adjustment is using government economic data.

Perform data exploration on this data set, and summarise (with tables and plots to support where appropriate) your initial conclusions about data issues and appropriate models. You should take into account any concerns with data collection and processing.

4. An pensions company is modelling improvements in mortality. It collects the following data on its policyholders:

Variable	meaning
init.age	The age of the policyholder at the time of plan initiation.
init.year	The year of plan initiation
death.age	The age of the policyholder at death (0 if policyholder is still alive)
death.year	The year of the policyholder's death
sex	The sex of the policyholder
race	The race of the policyholder
income	The policyholder's income (adjusted for inflation) at time of initiation.
smoking	Whether and how much the policyholder smokes at time of initiation.
health	A measure of the policyholder's overall health at time of initiation, with 100 representing perfect health and 0 bein

The data are in the file `HW2Q4.txt`.

The data are from the pension company's records.

Perform data exploration on this data set, and summarise (with tables and plots to support where appropriate) your initial conclusions about data issues and appropriate models.

5. A scientist is studying the effect of social habits on microbial communities in the guts of animals. He collects the following data

Variable name	Meaning
species	The species of animal
social.type	The type of social behaviour of the animal
diet	Carnivore, herbivore, omnivore
age	The age of the animal at sample collection
wild	Whether the animal is wild or captive
Bacteroides	The percentage of the gut community consisting of the phylum Bacteroides
Firmicutes	The percentage of the gut community consisting of the phylum Firmicutes

The data are in the file `HW2Q5.txt`. For captive animals, the species, age and social behaviour are identified by careful examination. For wild animals, they are determined by video surveillance of the habitat, with animals observed fewer than 3 times removed from the sample. The percentages of Bacteroides and Firmicutes are determined by sequencing the bacterial community in a faecal sample. For captive animals, this sample is collected within two hours, and sequenced the following day. For wild animals, the sample is collected at the following site visit, which can be up to a week after the sample is produced. The sample is then sequenced upon return to the laboratory, which may be another week.

Perform data exploration on this data set, and summarise (with tables and plots to support where appropriate) your initial conclusions about data issues and appropriate models.