ACSC/STAT 3740, Predictive Analytics

WINTER 2025

Toby Kenney

Homework Sheet 2

Due: Wednesday 12th February: 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.

[Note: all data in this homework are simulated.]

Standard Questions

1. The file `HW2Q1.txt` contains the following data from an insurance company's records on investment returns.

Variable	Meaning
Term	The length of time the investment was to be held
Liquidity	A measure of the liquidity of the investment
Risk.level	A measure of the relative risk of the investment
Index.Return	The return on a comparable market index.
Return	The percentage return on the investment.

Construct a plot or plots to show this data for the purpose of data exploration.

2. The file `HW2Q2.txt` contains the following data from an experiment on the effect of climate on fertility of wolves.

Variable	Meaning
Ave.winter.temp	The average daily maximum temperature in the period Dec–Mar
Ave.summer.temp	The average daily maximum temperature in the period Jun–Aug
Precipitation	The total annual precipitation
Ave.wind	The average wind speed during the year.
Population	The total adult population of the pack.
Pregnancies	The number of pregnancies.
Live.births	The number of live births in the pack.

The climate data are average readings from a nearby weather station over the previous 10-year period. The population, pregnancies an live births data are estimated using a capture-recapture experiment, where wolves are marked, and then set loose, and the proportion of observed individuals marked is used to estimate the total population.

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. A government has collected the following data about the effect of educational grants on social mobility in the file `HW2Q3.txt`.

Variable	Meaning
GDP.growth	The annual growth in GDP.
Gini.coefficient	A measure of income inequality in the country.
Political.system	The system of government in the country.
Percent.Tech	The percentage of GDP attributable to the technology industry.
Percent.Service	The percentage of GDP attributable to the service industry.
Percent.Manufacture	The percentage of GDP attributable to the manufacture industry.
Percent.Agriculture	The percentage of GDP attributable to the agriculture industry.
Percent.Resources	The percentage of GDP attributable to the resources (e.g. mining) industry.
Unemployment	The percentage of individuals seeking employment who are unable to find it.
Education.years	The average number of years spent in full-time education.
Percent.University	The percent of individuals who attend university.
Education.Grants	The per-capita amount spent on educational grants.
Social.Mobility	An index measuring social mobility.

The economic data are from the government websites for each country. Political data are from the classification of government systems in an academic paper. Data on education systems, university attendance are from a website giving international survey results on education. Education grant data are obtained from government websites. Social mobility data are from an international website that provides survey results about social mobility and various other lifestyle factors.

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. The file `HW2Q4.txt` contains the following data from a company's human resources department.

Variable	Meaning
Age	The employee's age.
Gender	The employee's gender.
Education	The number of years of post-secondary education that the employee has.
Job.type	The type of job the employee has.
Salary	The employee's annual salary.
Total.hours	The employee's average number of weekly hours.
Remote.hours	The employee's average number of hours working remotely.
5-year retention	Whether the employee remains at the company for 5 years.

5. An advertising company is studying internet search terms. It collects the following data :

Variable	meaning
Part.of.speech	The grammatical type of word that the search term is.
Number.of.searches	The number of searches involving this search term.
Average.search.length	The average length in words of a search involving this term.
Click.rate	The proportion of searches involving this term that result in a click on an advertisement.
Term.coverage	The proportion of searches involving this term that also involve one of the 100 most common search terr

The data are in the file `HW2Q5.txt`.

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.

6. The file HW2Q6.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 proportion of the patient's diet that consists of fats.
percent.fibre	The proportion 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.

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.