ACSC/STAT 3740, Predictive Analytics

WINTER 2025

Toby Kenney

Homework Sheet 1

Due: Friday 24th January: 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.]

Basic Questions

1. A former colleague has produced the code in the file `HW1Q1.R` to process the sports-analytics dataset in the file `HW1Q1.txt`, before leaving the company. The code is intended to remove all rows with `Balls.remaining` equal to zero. However, it does not work. Explain why the code does not work, and how to make it work, and how to restructure it in a better way.

2. A government worker is investigating the effect of various parenting techniques on children's mental health. He has used the code in the file `HW1Q2.R` to process the data in the file `HW1Q2.txt`. Add comments to the code to make it easier to follow.

   The variables in the data set are explained in the following table:

Variable	Meaning
`Living.With`	The child's household status — one of "Both", "Mother", "Father", "Joint custody", "Foster", "Other"
`Family.Income`	The combined annual household income
`Age`	The child's age
`Siblings`	The number of siblings the child has
`Discipline.strict.rules`	The extent to which the caregivers strictly enforce rules
`Discipline.punishment`	The extent to which the caregivers use punishment for misbehaviour
`Parent.attention`	The average number of hours per week that the caregivers spend with the child
`Freedom`	The extent to which the child is allowed to act without supervision.
`Health.index`	An index summarising the child's overall physicall health.
`Programmes`	The average number of hours per week that the child spends in extracurricular programmes.
`Screentime`	The average number of hours per week that the child spends using electronic devices.
`Friends`	The number of friends the child has.
`School.Grades.Mathematics`	The child's average grade in school mathematics.
`School.Grades.English`	The child's average grade in school english.
`Depression.Score`	A summary of various psychological surveys assessing the child's susceptibility to depression.

3. A scientist is studying crop growth. Their research assistant was analysing the data in the file `HW1Q3.txt`, and wrote the code in the file `HW1Q3.R` to process the data, before leaving suddenly. Upon reviewing the code, the scientist discovers that the code does not work. Fix the code.

4. A government researcher is studying the effect of news coverage on elections. Their research assistant was analysing the data in the file `HW1Q4.txt`, and wrote the code in the file `HW1Q4.R` to process the data, before leaving suddenly. Upon reviewing the code, the researcher discovers that the code does not work. Fix the code and improve it to reduce the risk of this type of mistake happenning in future.

5. The code in the file `HW1Q5.R` is a script for processing a reinsurance company's contract records. Improve the code to make it more reusable and less error-prone.

6. A data scientist has produced the code in the file `HW1Q6.R` to process a company's data. Testing the code on a small subset of the data, she finds that it takes 7 hours to process a dataset with 200,000 records, each with 400 predictors.

   (a) Approximately how long would the program be expected to take for the company's whole database of 4,000,000 records with 1,200 variables each?

   (b) Management deems the time required unacceptable. Rewrite the code to run more efficiently for big datasets.

7. The file `HW1Q7.txt` contains data from an entertainment company about electricity usage. The data are not formatted in a very convenient way. Read the data into `R` and reformat into a more convenient way, and use it to create a plot showing electricity used per hour (y-axis) vs number of people (x-axis) with colour showing age group and size showing company size, with a facet grid of type of event versus time of day. Make a list of all corrections made to the data.