ACSC/STAT 3740, Predictive Analytics

WINTER 2025

Toby Kenney

Homework Sheet 4

Due: Friday 21st March: 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 sets in this homework are simulated.

Standard Questions

1. A home insurance company has collected the following data about fire damage in the file `HW4Q1`.

Variable	Meaning
material	The main material used to build the house.
living.area	The living area of the house.
recent.rain	The amount of rainfall in the week preceding the fire.
fire.alarm	Whether the home is equipped with fire alarms.
sprinkler	Whether the home is equipped with sprinklers.
occupied	Whether the home was occupied at the time of the fire.
fire.station.distance	The distance from the home to the nearest fire station.
damage	The total cost to repair the house.

They have used the code in the the file `HW4Q1_code` to fit a linear regression model to predict the treatment outcome for each patient. 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?

2. A company has collected the following data about customer service in the file `HW4Q2`.

Variable	Meaning
age	The age of the customer.
gender	The gender of the customer.
previous.customer	Whether the customer bought anything within the past 12 months.
amount.spent	The amount the customer spent.
agent.experience	The number of years of experience the customer service agent had.
agent.gender	The agent's gender.
time	The time spent with the customer (minutes)
rating	The customer's rating of the agent.

They have used the code in the file `HW4Q2_code.R` to predict the rating given in each case. Assess the model assumptions and predictive performance of the model. How might the model be improved?

3. A health researcher is studying the effect of access to a family doctor on long-term health outcomes. She has collected the following data in the file `HW4Q3`.

Variable	Meaning
population	The population of the region
family.doctors	Number family doctors in the region.
ave.travel	The average time an inhabitant must travel to attend an appointment.
over.sixty	The proportion of the population over 60.
ave.income	The average income in the region.
cancer.deaths	The number of cancer deaths.
heart.deaths	The number of deaths caused by heart problems.

They have used the code in the file `HW4Q3code` to fit two models to predict heart death rates. The first model is a GAM. The second is a random forest model. Determine which model is better for predicting the heart death rates. [You may need to modify the code for this.]

4. A doctor is studying the effect of antibiotic usage on obesity. He has collected the following data in the file HW4Q4.

Variable	Meaning
patient.BMI.before	The patient's BMI before the start of treatment.
patient.age	The patient's age
patient.sex	The patient's sex.
antibiotic.dosage	The total dosage of antibiotics prescribed.
patient.BMI.after	The patient's BMI 6 months after the start of treatment.

He has used the code in the file HW4Q4_code to fit two GAM models to predict patient BMI afterwards, with different choices for nonlinear terms. Determine which model is better for predicting patient BMI afterwards.