**37252 Regression and Linear Models**

**Lab 5: Multiple Linear Regression III**

This lab is marked out of 21.

Please save your file in PDF format with name

**37252\_Lab5\_Surname\_FirstName**

**Due: 12 noon Wednesday 10 April 2024**

In this week’s lab we model the calorie content of breakfast cereals sold in a particular supermarket. The data are nutritional information from 77 products and available in **37252\_Lab5\_data.csv**, which can be downloaded from Canvas.

The variables we consider are summarised in the table below.

|  |  |  |
| --- | --- | --- |
| **Name** | **Role** | **Description** |
| $$calories$$ | response | calorie count |
| $$sugars$$ | predictor | sugar content |
| $$carbo$$ | predictor | carbohydrate content |
| $$fat$$ | predictor | fat content |
| $$shelf$$ | predictor | display shelf (1, 2, 3) |

The variable $shelf$ is a three-state categorical variable which must be recoded into two dummy variables; we can do this either manually or ask R to do it for us by specifying it as a *factor*.

Build a multiple linear regression model with $calories$ as response and $sugars$, $carbo$, $fat$, $shelf1$ (dummy variable for $shelf=1$ ), $shelf2$ (dummy variable for $shelf= 2$), interaction between $shelf$ and $sugars$ as predictors.

>caloriesdat<-read.csv("~/2024\_37252/Labs/Lab5/37252\_Lab5\_data.csv")

> caloriesdat$shelf <- as.factor(caloriesdat$shelf)

> caloriesdat$shelf <- relevel(caloriesdat$shelf, ref = "3")

> mod1 <- lm(calories ~ sugars + carbo + fat + shelf + sugars\*shelf, data = caloriesdat)

> summary(mod1)

1. Write down the estimated regression equation **[1 mark]** and provide interpretations of the estimated beta coefficients for $shelf1$ and $sugars:shelf1$ **[2 marks]**.
2. Use R to calculate $\hat{calories}$ for cereals on shelves 1, 2 and 3 when $sugars=9$, $carbo=16$ and $fat=3$ **[3 marks]**.
3. Determine if there is any statistical evidence of serial correlation which would violate the assumption of independence **[2 marks]**.
4. Determine if there is any statistical evidence of multicollinearity **[3 marks]**.
5. Perform a visual analysis of the residuals for compliance with the normality, independence and constant variance assumptions **[3 marks]**.
6. Identify potentially influential points giving statistical evidence for your answer **[3 marks total, 1 for evidence, 2 for identification of all points]**.

Filter out the 5 points identified in part (f) (hint!) and re-run the regression model again (see Lab 2 for filtering instructions).

From part (e) you probably identified a problem with the assumption of normality (hint!). We’ll now test the normality of the residuals in the filtered dataset model.

1. Using 0.05 significance level, perform a hypothesis test as to the normality of the residuals (see Lab 2 for instructions). Write down the hypotheses **[1 mark]**, the test statistic and p-value **[1 mark]**, the result of the test **[1 mark]** and a conclusion in non-mathematical language **[1 mark]**.