Predictive Anayltics Assignment 1 (worth 10% of your final grade)

Deadline for completion is the 2nd of November 2020

Mastitis occurs when bacteria gets into a cows udder which causes an infection. On average mastitis costs farmers €60 per cow per year. Mastitis accounts for a loss of 20% of the total agricultural revenue in Ireland. Healthy udders are: economically profitable, lead to a better quality product, and better cow welfare.

Somatic cell count (**scc**) is the total number of cells per millilitre in milk. Primarily, **scc** is composed of leukocytes, or white blood cells, that are produced by the cow's immune system to fight a mastitis infection. Since leukocytes in the udder increase as the condition worsens, **scc** provides an indication of the degree of mastitis in an individual cow.

Automated (robotic) milking systems are becoming more popular in Ireland and provide information on various measures of the composition of milk. Casein and whey protein are the major proteins in milk. Casein constitutes approximately 80% (29.5 g/L) of the total protein in bovine milk, and whey protein accounts for about 20% (6.3 g/L). The objective of this project is to analyze the relationship between the somatic cell count **scc** with the protein levels recorded by the automated (robotic) milking system, which are **protein** and **casein**. We also consider the percentage concentrate feed (supplements) in the cows' diet **conc_fed**.

This data set contains

- **protein** the recorded protein in the milk for cow i,
- **casein** the casein in the milk for cow i,
- **scc** the somatic cell count in the milk for cow i,
- **conc_fed** the percenatge concentrate feed (supplements) in the cows diet for cow i,

for $i = 1, \ldots, N$, where N is the number of cows recorded in the data set. The observations relate to individual cows on four farms in Ireland.

Read the data set available on Brightspace contained in a .csv file into R.

Exploratory Data Analysis (35 marks):

For each question in the EDA section please provide the lines of R code required to produce your results and the tables and figures produced by R.

1. Using a boxplot, histogram and the descriptive statistics (mean, min, max, median, and quantiles). Describe the distribution of the somatic cell count **scc**. (5 marks)
2. Using a boxplot, histogram and the descriptive statistics (mean, min, max, median, and quantiles). Describe the distribution of the log of the somatic cell counts **scc**. (5 marks)
3. Using a boxplot, histogram and the descriptive statistics (mean, min, max, median, and quantiles). Describe the distribution of the protein levels **protein**. (5 marks)
4. Using a boxplot, histogram and the descriptive statistics (mean, min, max, median, and quantiles). Describe the distribution of the casein levels **casein**. (5 marks)

5. Convert the categorical variable **conc_fed** to a factor. Describe and illustrate the frequency and proportions of the categorical variable concentrate feed **conc_fed** (5 marks)

6. Using the descriptive statistics (mean, standard deviation, median, mad: median absolute deviation (from the median), minimum, maximum, skew and standard error) and a boxplot describe how the log of somatic cell counts **scc** varies with respect to the variable concentrate feed **conc_fed** (5 marks)

7. Using the correlation and scatter plots discuss the relationship between **log(scc)** and each of the variables **protein** and **casein**. (3 marks)

8. Based on the results from Q 7, which variable **protein** or **casein** would provide a better predictor variable in your regression model with **log(scc)** as the response. Provide a justification for your selection. (2 marks)

Regression Model (65 marks):

1. Using R fit a simple linear regression model to the data with **log(scc)** as the response variable and the variable chosen in Q8 of the exploratory analysis section as the predictor variable. Define and describe the mathematical equation for the model. (Also provide you R code) (4 marks)

2. Interpret the estimate of the intercept term. (2 marks)

3. Interpret the estimate of the slope term. (2 marks)

4. Calculate the variance of the estimate of the intercept and slope term. (2 marks)

5. Calculate and interpret the confidence intervals for β_0 (Provide you R code) (5 marks)

6. Calculate and interpret the confidence intervals for β_1 (Provide you R code) (5 marks)

7. Compute and interpret the hypothesis test $H_0 : \beta_0 = 0$ vs $H_a : \beta_0 \neq 0$. State the test statistic. Compare the test statistic to the correct distribution value and state your conclusion. Also, report the p-value and the conclusion in the context of the problem. (8 marks)

8. Compute and interpret the hypothesis test $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$. State the test statistic. Compare the test statistic to the correct distribution value and state your conclusion. Also, report the p-value and the conclusion in the context of the problem. (8 marks)

9. Interpret the F-statistic in the output in the summary of the regression model. Hint: State the hypothesis being tested, the test statistic and p-value and the conclusion in the context of the problem. (6 marks).

10. Interpret the R-squared value. (2 marks)

11. Interpret the residual standard error of the simple linear regression model. (2 marks)

12. Calculate, plot and comment on the shape of the confidence intervals for the estimated values of Y (Provide you R code) (4 marks)

13. List the assumptions of the linear regression model required for small sample inference (5 marks)

14. Examine the residuals of the regression model and comment on whether you think the residuals satisfy the assumptions of the linear model. Provide the rationale for your answer (10 marks).