Biomedical Data Science & AI

Exercise sheet 3 - Introduction

SoSe2021

Prof. Dr. Holger Fröhlich, Mohamed Aborageh, Vinay Bharadhwaj, Yasamin Salimi

Due date: May 11th

Questions

Exercise 1 - Probability (7 points)

1. The amount of wine bottles sold in a shop follows a Poisson distribution with 180 bottles per week (6 days). If C = the random variable for bottles per day, how is:

    a. The probability that the shop will only sell 20 bottles per day? **(2 points)**
    b. The probability that the demand is more than average for a particular day? **(2 points)**
    c. The expected number of units per day E[C]? **(1 point)**
    d. What is Var[C]? **(1 point)**
    e. The standard deviation of C? **(1 point)**

Exercise 2 - Hypothesis testing (8 points)

This exercise illustrates a gene expression data set with its normally distributed values. Consider the gene expression data of the *Golub* dataset. Load the file “golub.csv”. It contains gene expression data of 3051 genes from 38 tumor mRNA samples. The expression data is organized in a matrix where rows correspond to genes and columns to samples. The tumor class of the columns is given in the file “golub.cl”. The names of the genes (rows) are given in “golub.gnames”.

1. Calculate the sample mean of all genes $\hat{\beta}$ in the pooled expression matrix. Use these means to determine the overall mean β_0 by just taking the average. **(1 point)**

2. Based on the t-statistic defined as follows:

$$t_{\hat{\beta}} = \frac{\hat{\beta} - \beta_0}{\text{s.e.}(\hat{\beta})}$$

    obtain the 100 most significant genes. [Hint : $\hat{\beta}$ is the sample mean of a particular gene] **(1 point)**

3. Perform two sampled student t-tests for all genes comparing the distributions for ALL and AML. **(1 point)**
4. Based on the p-values obtained in 3., obtain the top 10 genes with the lowest p-values? **(1 point)**
5. Shapiro-Wilk test is used to test if a random variable follows a Normal distribution (Null-hypothesis). Using this test identify the top 100 genes which deviate significantly from normal. **(1 point)**
6. Out of the 100 genes obtained in g), use an appropriate statistical test to obtain the 10 most differentiating genes between ALL and AML classes. **(1 point)**
7. Inform yourself about the multiple testing problem. Apply one appropriate method to deal with it and explain how it works. **(2 points)**

Exercise 3 - Linear regression (10 points)

1. Using the *fish.csv* dataset, generate a multiple linear regression model with weight as the response variable and length1, length2, length3, height, and width as the predictors. Answer the following questions based on the regression model.
    a. How large is the coefficient of the predictors? **(1 point)**
    b. What is the value of the adjusted R-squared? What does this tell us? **(1 point)**
    c. Which predictors lead to the increase in the weight of the fish, and which have a negative effect? **(2 points)**
    d. Using a simple regression model, predict the weight of a fish with length1 of 15? **(2 points)**
    e. Predict the weight of a fish with length1, height, and width of 20.0, 7.3, and 5.3 using a multiple linear regression model. **(2 points)**
    f. What are the associated 95% confidence intervals for the predictors, as well as for intercept? **(2 points)**