Correlation & Linear Regression

Exam Questions

Download the files “foetus.csv”, “cucumbers.csv” and “performance.csv” and set the download location as your working directory in RStudio.

1. You must refer to the file “foetus.csv” when answering the following questions. Read the file into R and attach the variables.

This data set contains the data on the age and length of 84 foetuses measured from ultrasounds scans.

	1. Produce descriptive statistics to summarise the variables age and length and comment on the distribution of each.

|  |
| --- |
|  |

* 1. Produce a scatter plot with age on the x-axis and length on the y-axis.

Edit the graph to have a suitable title and axis labels and comment on the relationship between the age and length of typically developing foetuses.

|  |
| --- |
|  |

* 1. Compute the correlation coefficient between age and length.

What can be concluded about the relationship between age and length?

|  |
| --- |
|  |

* 1. Compute the least squares linear regression line which would model the length of a foetus in terms of the age.

|  |
| --- |
|  |

* 1. Compute the coefficient of determination for the regression model.
	How can this be interpreted in terms of the fitted model?

|  |
| --- |
|  |

* 1. Use the fitted model to predict the length of a foetus at 85 days and the length of a foetus at 120 days.
	For each prediction, comment on two factors which would indicate the predictions are accurate.

|  |
| --- |
|  |

Detach the variables in “foetus.csv”.

1. You must refer to the file “cucumbers.csv” when answering the following questions. Read the file into R and attach the variables.

This data set contains randomly collected data on growing season precipitation (mm) and cucumber yield (kg/m2).

	1. Produce a scatter plot with Precipitation on the x-axis and Yield on the y-axis.

|  |
| --- |
|  |

* 1. Compute the correlation between these two variables and comment on the result.

|  |
| --- |
|  |

Detach the variables in “cucumbers.csv”.

1. You must refer to the file “performance.csv” when answering the following questions. Read the file into R and attach the variables.

This data set shows the results of a small study done on a class of high school students who were asked to record the average number of hours sleep per night they got over a one week period.

	1. Produce a scatter plot of amount of sleep and exam performance on the appropriate axes.

|  |
| --- |
|  |

* 1. For the scatter plot, comment on the average number of hours sleep that the students have recorded, explaining why some seem unusual.
	Suggest how this might be measured.

|  |
| --- |
|  |

* 1. Compute the correlation between amount of sleep and exam performance and write a sentence to interpret the result.

|  |
| --- |
|  |

* 1. Compute the least squares linear regression line to predict exam performance from amount of sleep. Interpret the model coefficients in the context of the problem.

|  |
| --- |
|  |

* 1. Predict the performance of an individual who has 10 hours of sleep and comment on the accuracy of this prediction.

|  |
| --- |
|  |

**Useful R commands**

**Entering data to R Studio**

To read in data from an Excel csv file called *excel\_data.csv* to R Studio and name it *mydata*, first use the drop down menus in R Studio **Session > Set Working Directory > Choose Directory** to indicate the location of *excel\_data.csv* on your computer. The following code will then read the data in to R Studio:

mydata<-read.csv("excel\_data.csv")

attach(mydata) — this adds the variable names

At the end of the analysis remember to use detach(mydata) to disassociate the variable names.

**(a) Graphics**

hist(X,col="yellow",main="Histogram of X (units)") — this produces a histogram of the variable named ‘X’

plot(X,Y,xlab="x-axis label",ylab="y-axis label", main="Scatterplot of Y on X",pch=21,bg="black") — produces a scatterplot of X vs Y with black dots of the size specified by ‘pch’

pie(table(X), main="Title") — this gives a simple pie chart of the categories in variable X with the specified title

barplot(table(X), main="title", xlab="x-axis label",col="orange") — this gives a bar chart of the categories in the variable X with the required title, axis labels and colour

boxplot(Y) — produces a boxplot of the numerical variable Y

**(b) Descriptive Statistics**

mean(X) — computes the mean of X

sd(X) — computes the standard deviation of X

summary(X) — computes the mean, median, minimum, maximum and upper and lower quartiles

table(X) — computes the number of observations in each level of the categorical variable X

prop.table(table(X)) — returns the proportion of observations in each level of the categorical variable X

prop.table(table(X))\*100 — returns the percentage of observations in each level of the categorical variable X

table(X,Y) — produces a cross-tabulation between the two categorical variables X and Y

**(c) Correlation and Regression**

cor.test(X,Y) — computes the correlation between X and Y and performs a test of the null hypothesis of zero correlation

lm(Y~X) — fits a linear regression line to the data (lm command stands for linear model)

abline(lm(Y~X)) — produces a scatterplot with the least squares linear regression line superimposed on the data

summary(lm(Y~X)) — displays the coefficient of determination (r-squared)

predict(lm(Y ~ X), newdata=data.frame(X=C),interval = "pred") — computes the predicted value of Y when X=C along with a 95% prediction interval