**STAT 35000**

**Introduction to Statistics**

# Project 1

# Due: February 05 (Monday), 2024 Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Task #1 – Refer to the MHW Data – see details below.

# Retrieve the data file MHW.txt from the class site on Canvas and save it on your local work directory. Then set up your working direction.

# Read in the data to an appropriate data frame:

* MHW<- read.table(‘MHW.txt’, header=TRUE, sep=’,’)

1. Print the top 6 lines of the data set

* head(MHW)

1. Print the first 10 rows of this data frame

* MHW[1:10, ]

1. Attach the data frame to be your working data set

* attach(MHW)

# Task #2: Summarize the MHW data:

# Obtain the five-number summary for the data:

# summary(MHW)

# summary(MHW[,3]) or summary(MHW$grain) or summary(grain)

# summary(MHW[,4]) or summary(MHW$straw) or summary(straw)

# Obtain the various ‘statistics’ for the grain yield and do the same for straw

# min(grain)

# max(grain)

# mean(grain)

# median(grain)

# var(grain)

# sd(grain)

# quantile(grain)

# IQR(grain)

# Define and calculate a new variable ‘yield.ratio’ which is the ratio between grain and straw and obtain the summary statistics for it.

# yield.ratio=grain/straw

# Task #3: Visualize the data:

# Obtain a stem and leaf display of the grain and of the straw yields

# stem(MHW$grain) or stem(MHW[ , 3]) or stem(grain)

# Obtain a simple histogram of the grain and of the straw yields

# hist(grain)

# Obtain a fancy histograms of these data

* hist(straw, nclass=30, col = "lightblue", border = "red", main = "The MHW Data", xlab = "Straw yield per plot")
* hist(grain, breaks = seq(2.6, 5.2, by = 0.1), col = "lightblue", border = "red", main = "The MHW Data", xlab = "Grain yield per plot")

# Obtain, and appropriately label, the boxplots of the yields data

# boxplot(grain)

# boxplot(straw)

# boxplot(grain, straw)

# Task #4: Regression Analysis on the data. Two ‘measured’ variables were studied:

* **grain** : Grain yield, lbs per plot
* **straw** : Straw yield, lbs per plot

1. Calculate the mean and SD for the two variables, grain and straw

* apply(cbind(straw, grain), 2, mean)
* apply(cbind(straw, grain), 2, sd)

1. Obtain a clearly labeled scatterplot of straw(Y) against grain (X). What do you see in this plot? Do you think that the simple linear regression model is appropriate here?

* plot(straw ~ grain)

1. Do you think that with and  the simple linear regression model

 is appropriate here?

1. Mark the means of straw (horizontally) and of grain (vertically) on the plot

* abline(h=mean(straw), col=2)
* abline(v=mean(grain), col=2)

1. Considering the plot above, do you think that with and  the simple linear regression model  is appropriate here?
2. Find the covariance , , using the R command below and then calculate the correlation, , between straw and grain using the formula and results in Task#4 Part A):

* cov(grain, straw)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

1. Using the formula given below and results in Task#4 Part A) and F) to calculate the Least Squares Estimates of the regression coefficients and  for the regression of straw on grain as



# 

# and

# Write down the ‘fitted’ regression line of straw (yield) on the grain (yield)

# 

# Fit the regression (LS) line to the plot using the values of and you calculated in part H) above:

* abline(, , col =2, lwd=2)

# Now verify your results by repeating the regression analysis using the lm() function of R

# Yield.fit<-lm(straw~grain)

# Yield.fit

# summary(yield.fit)

# Using this ‘fitted’ regression model, what would you ‘predict’ the straw yield be for a plot with 4.0 lb. of grain yield?

# 

# Below is the information regarding the MHW Data in case you would like to know more about it.

# 

# THE MHW-Data—

# In the early days of scientific agriculture, Mercer and Hall [[1](#_bookmark650)] were trying to determine the optimum plot size for agricultural yield trials: Plots that are too *small* will be too variable; and plots that are too *large* waste resources (land, labor, seed); if the land area is limited, the number of treatments will be unnecessarily small.

So, they performed a very simple experiment: an apparently homogeneous field was selected, prepared as uniformly as possible and planted to the same variety of wheat. They attempted to treat all parts of the field exactly the same in all respects during subsequent farm operations. When the wheat had matured, the field was divided into 500 equally-size plots. Each plot was harvested separately. Both grain and straw were air-dried, then hand-threshed and weighed to a precision of 0.01 lb (= 4.54 g). The reported values are thus air-dry weight, in lb per plot.

The field was a square of 1 acre, which is 0.40469 hectare or 4,046.9 m2, which was divided into a 20 rows by 25 columns, giving 500 plots, each of 1/500 acre, which is about 8.09 m2 (3.30 m long x 2.45 m wide). We can assume that the rows ran W to E, with 25 plots in each row, beginning at 1 on the W and running to 25 at the E, so that columns run N to S with 20 plots in each, running from 1 at the N to 20 at the S. Thus the NW corner (1,1) is plot 1, the NE corner (1, 25) is plot 481, the SE corner (25, 20) is plot 500, and the SW corner (1, 20) is plot 20.

**The CSV Data File:** The data has been prepared as the comma-separated values(“CSV”) file mhw.csv in a plain-text editor. The first line gives the four field names: "**r**","**c**","**grain**","**straw**" standing for:

**r** : Row number in the field

**c** : Column number in the field

**grain** : Grain yield, lbs per plot

**straw** : Straw yield, lbs per plot

Each of the 500 lines in the data file represents a plot; the four data fields are separated by commas. For example, the first data line is: 1,1,3.63,6.37, the second data line is 2, 1, 407, 6.24 and so on…

|  |  |  |  |
| --- | --- | --- | --- |
| r | c | grain | straw |
| 1 | 1 | 3.63 | 6.37 |
| 2 | 1 | 4.07 | 6.24 |
| 3 | 1 | 4.51 | 7.05 |
| 4 | 1 | 3.9 | 6.91 |
| 5 | 1 | 3.63 | 5.93 |
| 6 | 1 | 3.16 | 5.59 |

[1] W B Mercer and A D Hall. The experimental error of field trials. *The Journal of Agricultural Science (Cambridge)*, 4:107–132, 1911.