**STAT 35000**

**Introduction to Statistics**

# Project 1

# Due: February 05 (Monday), 2024 Name: Raja Allmdar Tariq Ali

# 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.

# setwd("~/Desktop/STAT-350/Project\_1")

# 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)

**> # C) Printing the top 6 lines of the data set**

**> head(MHW)**

**r c grain straw**

**1 1 1 3.63 6.37**

**2 2 1 4.07 6.24**

**3 3 1 4.51 7.05**

**4 4 1 3.90 6.91**

**5 5 1 3.63 5.93**

**6 6 1 3.16 5.59**

1. Print the first 10 rows of this data frame

* MHW[1:10, ]

**> # D) Printing the first 10 rows of this data frame**

**> MHW[1:10,]**

**r c grain straw**

**1 1 1 3.63 6.37**

**2 2 1 4.07 6.24**

**3 3 1 4.51 7.05**

**4 4 1 3.90 6.91**

**5 5 1 3.63 5.93**

**6 6 1 3.16 5.59**

**7 7 1 3.18 5.32**

**8 8 1 3.42 5.52**

**9 9 1 3.97 6.03**

**10 10 1 3.40 5.66**

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

* attach(MHW)

**> # E) Attach the data frame to be your working data set**

**> attach(MHW)**

**The following objects are masked from MHW (pos = 4):**

**c, grain, r, straw**

# Task #2: Summarize the MHW data:

# Obtain the five-number summary for the data:

# summary(MHW)

# 

# > # A) Obtaining the five number summary of the data (including individual columns)

# > print("This is the five number summary of the whole data set")

# [1] "This is the five number summary of the whole data set"

# > summary(MHW)

# r c grain straw

# Min. : 1.00 Min. : 1 Min. :2.730 Min. :4.100

# 1st Qu.: 5.75 1st Qu.: 7 1st Qu.:3.638 1st Qu.:5.878

# Median :10.50 Median :13 Median :3.940 Median :6.360

# Mean :10.50 Mean :13 Mean :3.949 Mean :6.515

# 3rd Qu.:15.25 3rd Qu.:19 3rd Qu.:4.270 3rd Qu.:7.170

# Max. :20.00 Max. :25 Max. :5.160 Max. :8.850

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

# print("This is the five number summary of the grain column")

# [1] "This is the five number summary of the grain column"

# > summary(MHW$grain)

# Min. 1st Qu. Median Mean 3rd Qu. Max.

# 2.730 3.638 3.940 3.949 4.270 5.160

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

# > print("This is the five number summary of the straw column")

# [1] "This is the five number summary of the straw column"

# > summary(MHW$straw)

# Min. 1st Qu. Median Mean 3rd Qu. Max.

# 4.100 5.878 6.360 6.515 7.170 8.850

# 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)

# > # Grain Yield Statistics

# > print("Statistics for Grain Yield:")

# [1] "Statistics for Grain Yield:"

# > print(paste("Minimum:", min(grain)))

# [1] "Minimum: 2.73"

# > print(paste("Maximum:", max(grain)))

# [1] "Maximum: 5.16"

# > print(paste("Mean:", mean(grain)))

# [1] "Mean: 3.94864"

# > print(paste("Median:", median(grain)))

# [1] "Median: 3.94"

# > print(paste("Variance:", var(grain)))

# [1] "Variance: 0.210020190781563"

# > print(paste("Standard Deviation:", sd(grain)))

# [1] "Standard Deviation: 0.458279598914858"

# > print("Quantiles:")

# [1] "Quantiles:"

# > print(quantile(grain))

# 0% 25% 50% 75% 100%

# 2.7300 3.6375 3.9400 4.2700 5.1600

# > print(paste("Interquartile Range:", IQR(grain)))

# [1] "Interquartile Range: 0.632499999999999"

# > # Straw Yield Statistics

# > print("Statistics for Straw Yield:")

# [1] "Statistics for Straw Yield:"

# > print(paste("Minimum:", min(straw)))

# [1] "Minimum: 4.1"

# > print(paste("Maximum:", max(straw)))

# [1] "Maximum: 8.85"

# > print(paste("Mean:", mean(straw)))

# [1] "Mean: 6.5148"

# > print(paste("Median:", median(straw)))

# [1] "Median: 6.36"

# > print(paste("Variance:", var(straw)))

# [1] "Variance: 0.806955270541083"

# > print(paste("Standard Deviation:", sd(straw)))

# [1] "Standard Deviation: 0.89830689106846"

# > print("Quantiles:")

# [1] "Quantiles:"

# > print(quantile(straw))

# 0% 25% 50% 75% 100%

# 4.1000 5.8775 6.3600 7.1700 8.8500

# > print(paste("Interquartile Range:", IQR(straw)))

# [1] "Interquartile Range: 1.2925"

# 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

# > # Define and Calculate Yield Ratio

# > yield.ratio <- grain / straw

# > # Obtaining summary statistics for yield.ratio

# > print("Summary statistics for yield.ratio:")

# [1] "Summary statistics for yield.ratio:"

# > summary(yield.ratio)

# Min. 1st Qu. Median Mean 3rd Qu. Max.

# 0.3910 0.5737 0.6044 0.6105 0.6420 0.8500

# 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)

# > # For Grain Yield

# > print("Stem and Leaf Display for Grain Yield:")

# [1] "Stem and Leaf Display for Grain Yield:"

# > stem(MHW$grain)

# The decimal point is 1 digit(s) to the left of the |

# 27 | 38

# 28 | 45

# 29 | 279

# 30 | 144555557899

# 31 | 4446678999

# 32 | 2345589999

# 33 | 002455666677789999

# 34 | 00112233444444566777777888999

# 35 | 01112334444555666677789999

# 36 | 0001111133333444445666666777778889999

# 37 | 00011111122222233344444555556666667777899999

# 38 | 0011222223334444455566667777999999

# 39 | 0111111112222233333444444555666666777777777999

# 40 | 011122333344555666666677777778888899999999

# 41 | 0001111122333445555777779999

# 42 | 00001111111222333344444466677777788999999

# 43 | 0111223333566666777778888999999

# 44 | 0011111222234445566667777899

# 45 | 0112222234445667888899

# 46 | 1344446678899

# 47 | 3356677

# 48 | 466

# 49 | 12349

# 50 | 279

# 51 | 3336

# > # For Straw Yield

# > print("Stem and Leaf Display for Straw Yield:")

# [1] "Stem and Leaf Display for Straw Yield:"

# > stem(MHW$straw)

# The decimal point is 1 digit(s) to the left of the |

# 40 | 0

# 42 | 8

# 44 | 367

# 46 | 226167

# 48 | 1155667

# 50 | 00557911228

# 52 | 0014452338

# 54 | 01111235567778889022266788888899

# 56 | 00115566688889900001112568888899

# 58 | 00012222234444456677778990112223334555666677889999

# 60 | 112333333445566778888999000111123333444444455555556777889

# 62 | 0001113344556778899999012333333444555556667777889

# 64 | 2334455667990344455566779

# 66 | 002334446789000122455678899

# 68 | 01223556677789901113335556678888899

# 70 | 0023334455556678889001134667777899

# 72 | 000233456668888990011122333445667899

# 74 | 1136790001112233457789

# 76 | 0144577912233333359

# 78 | 024466789123689

# 80 | 256675578

# 82 | 333317

# 84 | 5388

# 86 | 1342458

# 88 | 55

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

# hist(grain)

# 

# hist(straw)

# A graph of straw Description automatically generated

# Obtain a fancy histograms of these data

* hist(straw, nclass=30, col = "lightblue", border = "red", main = "The MHW Data", xlab = "Straw yield per plot")

A graph of straw yield

Description automatically generated

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

# A graph of a grain yield Description automatically generated

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

# boxplot(grain)

# 

# # Boxplot for Grain Yield

# > boxplot(grain, main = "Boxplot of Grain Yield", ylab = "Grain Yield")

# A diagram of a grain yield Description automatically generated

# boxplot(straw)

# # Boxplot for Straw Yield

# > boxplot(straw, main = "Boxplot of Straw Yield", ylab = "Straw Yield") A diagram of a straw yield Description automatically generated

# boxplot(grain, straw)

# > # Combined Boxplot for Grain and Straw Yields

# > boxplot(grain, straw, names = c("Grain", "Straw"), main = "Combined Boxplot of Grain and Straw Yields", ylab = "Yield")A diagram of a grain and straw Description automatically generated

# 

# 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)

**> print("Mean of Grain and Straw:")**

**[1] "Mean of Grain and Straw:"**

**> print(apply(cbind(grain, straw), 2, mean))**

**grain straw**

**3.94864 6.51480**

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

**> print("Standard Deviation of Grain and Straw:")**

**[1] "Standard Deviation of Grain and Straw:"**

**> print(apply(cbind(grain, straw), 2, sd))**

**grain straw**

**0.4582796 0.8983069**

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)

**> plot(straw ~ grain, main = "Scatterplot of Straw Yield against Grain Yield", xlab = "Grain Yield", ylab = "Straw Yield")**

A graph of grain yield

Description automatically generated

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

 is appropriate here?

Yes, and these are the observations to support my claim:  
1. There is a clear positive linear relationship between grain yield and straw yield. As the grain yield increases, the straw yield also tends to increase.

2. There is a strong linear correlation between the two variables

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

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

A graph of grain yield

Description automatically generated

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

A graph of straw yield

Description automatically generated

1. Considering the plot above, do you think that with and  the simple linear regression model  is appropriate here?

Yes, and these are the observations to support my claim:  
1. There is a clear positive linear relationship between grain yield and straw yield. As the grain yield increases, the straw yield also tends to increase.

2. There is a strong linear correlation between the two variables

3. There is also an observation that the spread of data points increases with an increase in grain yield (possible heteroscedasticity) where the variance of the dependent variable changes across different levels of the independent variable.

1. 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)

**> # F) Calculate Covariance and Correlation between Straw and Grain**

**> covariance <- cov(grain, straw)**

**> print(paste("Covariance between Grain and Straw:", covariance))**

**[1] "Covariance between Grain and Straw: 0.300433394789579"**

**> correlation <- covariance / (sd(grain) \* sd(straw))**

**> print(paste("Correlation between Grain and Straw:", correlation))**

**[1] "Correlation between Grain and Straw: 0.729781667762202"**

0.300433394789579 0.729781667762202.

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



# 

# 

# > # G) Calculate Least Squares Estimates of the regression coefficients for the regression of Straw on Grain

# > fit <- lm(straw ~ grain)

# > coefficients <- coef(fit)

# > print("Coefficients of the regression line:")

# [1] "Coefficients of the regression line:"

# > print(coefficients)

# (Intercept) grain

# 0.8662797 1.4304977

# > # Direct calculation using the formula

# > beta <- cor(straw, grain) \* (sd(straw) / sd(grain))

# > alpha <- mean(straw) - (beta \* mean(grain))

# > print(paste("Calculated Intercept (alpha):", alpha))

# [1] "Calculated Intercept (alpha): 0.866279657326044"

# > print(paste("Calculated Slope (beta):", beta))

# [1] "Calculated Slope (beta): 1.4304976758261"

# and

# 0.866279657326044 and 1.4304976758261

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

# > # H) Fitted regression line

# > print("Fitted regression line: Straw Yield = a + b \* Grain Yield")

# [1] "Fitted regression line: Straw Yield = a + b \* Grain Yield"

# > cat("Straw Yield =", coefficients[1], "+", coefficients[2], "\* Grain Yield\n")

# Straw Yield = 0.8662797 + 1.430498 \* Grain Yield

# > print("Fitted regression line: Straw Yield = a + b \* Grain Yield")

# [1] "Fitted regression line: Straw Yield = a + b \* Grain Yield"

# > cat("Straw Yield =", alpha, "+", beta, "\* Grain Yield\n")

# Straw Yield = 0.8662797 + 1.430498 \* 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)

**> # I) Fit the regression line to the plot**

**> abline(alpha, beta, col = 2, lwd = 2)**

**A graph of straw yield

Description automatically generated**

# 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)

# 

# > # J) Verify results using the lm() function

# > yield.fit <- lm(straw ~ grain)

# > print("Linear Model Fit:")

# [1] "Linear Model Fit:"

# > print(summary(yield.fit))

# Call:

# lm(formula = straw ~ grain)

# Residuals:

# Min 1Q Median 3Q Max

# -2.02226 -0.35289 0.01039 0.37339 3.03420

# Coefficients:

# Estimate Std. Error t value Pr(>|t|)

# (Intercept) 0.86628 0.23872 3.629 0.000314 \*\*\*

# grain 1.43050 0.06005 23.821 < 2e-16 \*\*\*

# ---

# Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# Residual standard error: 0.6148 on 498 degrees of freedom

# Multiple R-squared: 0.5326, Adjusted R-squared: 0.5316

# F-statistic: 567.4 on 1 and 498 DF, p-value: < 2.2e-16

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

# 

# > # K) Predicting straw yield for a plot with 4.0 lb of grain yield

# > predicted\_straw\_yield <- predict(yield.fit, newdata = data.frame(grain = 4.0))

# > print(paste("Predicted Straw Yield for 4.0 lb of Grain Yield:", predicted\_straw\_yield))

# [1] "Predicted Straw Yield for 4.0 lb of Grain Yield: 6.58827036063044"

# > Manual\_Predicted\_Straw\_Yield <- alpha + beta \* 4

# > print(paste("Predicted Straw Yield for 4.0 lb of Grain Yield:", Manual\_Predicted\_Straw\_Yield))

# [1] "Predicted Straw Yield for 4.0 lb of Grain Yield: 6.58827036063043"

# 

# Predicted straw yield = 6.5883 (4 Decimal Places)

# 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.