**INFERENTIAL STATISTICS**

**ARVIND PAWAR**

**DHRUVIN KOTHARI**

**NORTHEASTERN UNIVERSITY**

**ALY6015 INTERMEDIATE ANALYTICS**

**INSTRUCTOR NAME- TENGLONG LI**

**DATE- 03/05/2019**



**1. Please install and load the library “MASS” into your R Studio so that we may use sample data available within this package. To accomplish this task, you may use the following code:**

install.packages(“MASS”)

library(MASS)

**Using the X dataset, conduct the appropriate t-tests in R (using the “t.test” function) to answer the following questions.**

**one-sample t-test. Use the “chem” dataset to answer the question: “Is the flour production company producing wholemeal flour with greater than 1 part per million copper in it?” (Hint: chem is a vector of values representing the parts per million of copper in the flour).**

Solution:

> Chem\_Dataset<-MASS::chem

> mean(Chem\_Dataset)

**[1] 4.280417**

> t.test(Chem\_Dataset,mu=1,alternative = "greater")

**One Sample t-test**

**data: Chem\_Dataset**

**t = 3.0337, df = 23, p-value = 0.002952**

**alternative hypothesis: true mean is greater than 1**

**95 percent confidence interval:**

**2.427162 Inf**

**sample estimates:**

**mean of x**

**4.280417**

Here we are using one sample T test to check whether our hypothesis is true or false.

Ho: flour production company producing whole meal flour less than equal to 1 ppm copper in it

With a confidence interval of 95%

H1: flour production company producing whole meal flour greater than 1 ppm copper in it

We stored all the data present in “chem” dataset in variable ‘Chem\_Dataset’. We used t.test function and applied the parameters to check hypothesis, we used alternative as greater, as we are checking for value to be greater than 1.

We shall also check values manually if there are any values present in data to be less than one, by using ‘for and if condition’. By both we could conclude that the values are greater than 1 and can say that our claim is true and can reject null hypothesis.

In above calculations, p-value is also low which indicates less support for the null hypothesis.

> i=c(1:24)

> print (i)

> for(Chem\_Dataset in i)+

if(sd[Chem\_Dataset]>1)+

{print("True")}

[1] "True"

[1] "True"

[1] "True"

[1] "True"

[1] "True"

[1] "True"

[1] "True"

[1] "True"

[1] "True"

[1] "True"

[1] "True"

[1] "True"

[1] "True"

[1] "True"

[1] "True"

[1] "True"

[1] "True"

[1] "True"

[1] "True"

[1] "True"

[1] "True"

[1] "True"

[1] "True"

[1] "True"

In the above code the “True” value represents values that are greater than one. From the above output we can see that all the values are greater than 1.

**two-sample t-test with unequal variance. Using the “cats” dataset, answer the question: do male and female cat samples have the same bodyweight(“Bwt”)? [Hint: one way to get separate vectors for male and female cat bodyweights is to use the subset function as follows: “male <- subset(cats, subset=(cats$Sex=="M"))”].**

**Solution:**

> Cats\_Dataset<-MASS::cats

> male <- subset(cats,cats$Sex=="M")

> male

Sex Bwt Hwt

48 M 2.0 6.5

49 M 2.0 6.5

50 M 2.1 10.1

51 M 2.2 7.2

52 M 2.2 7.6

53 M 2.2 7.9

54 M 2.2 8.5

> sum(male$Bwt)

[1] 281.3

> mean(male$Bwt)

[1] 2.9

> Female<- subset(cats,cats$Sex=="F")

> sum(Female$Bwt)

[1] 110.9

> mean(Female$Bwt)

[1] 2.359574

> t.test(male$Bwt,Female$Bwt,var.equal=FALSE)

**Welch Two Sample t-test**

**data: male$Bwt and Female$Bwt**

**t = 8.7095, df = 136.84, p-value = 8.831e-15**

**alternative hypothesis: true difference in means is not equal to 0**

**95 percent confidence interval:**

**0.4177242 0.6631268**

**sample estimates:**

**mean of x mean of y**

**2.900000 2.359574**

Ho: the weights of male and females have same weight.

H1: the weights of male and females have different weight.

Here we are storing the values present in “cats” dataset in ‘sa’ then filtering the data using gender and storing the values into ‘male’ and ‘female’ then calculating the sum and average of their weights to see if they are equal, we can see that their sum and average is not equal by which we can say that their weights are not equal, we would check this with the help of two sample t test now by using“> t.test(male$Bwt,female$Bwt, var.equal=FALSE)” here we took variance as False as it is mentioned in the question that they have unequal variance and the other two argument passed to the t.test function is male body weight, female body weight.

In the result we can observe, p value is less than level of significance (0.05) and p value is also

Very small, we can reject null hypothesis and say that the weights of male and female is

different.

**2. Use the R function “lm” and “summary” to fit a linear regression model to data from the**

**“mtcars” dataset in order to answer the following questions:**

1. **Create a corrplot and indicate the attributes that are positively and negatively correlated to mpg. (Hint: Gear is positively correlated and wt is negatively correlated)**

**Solution:**

install.packages("corrplot")

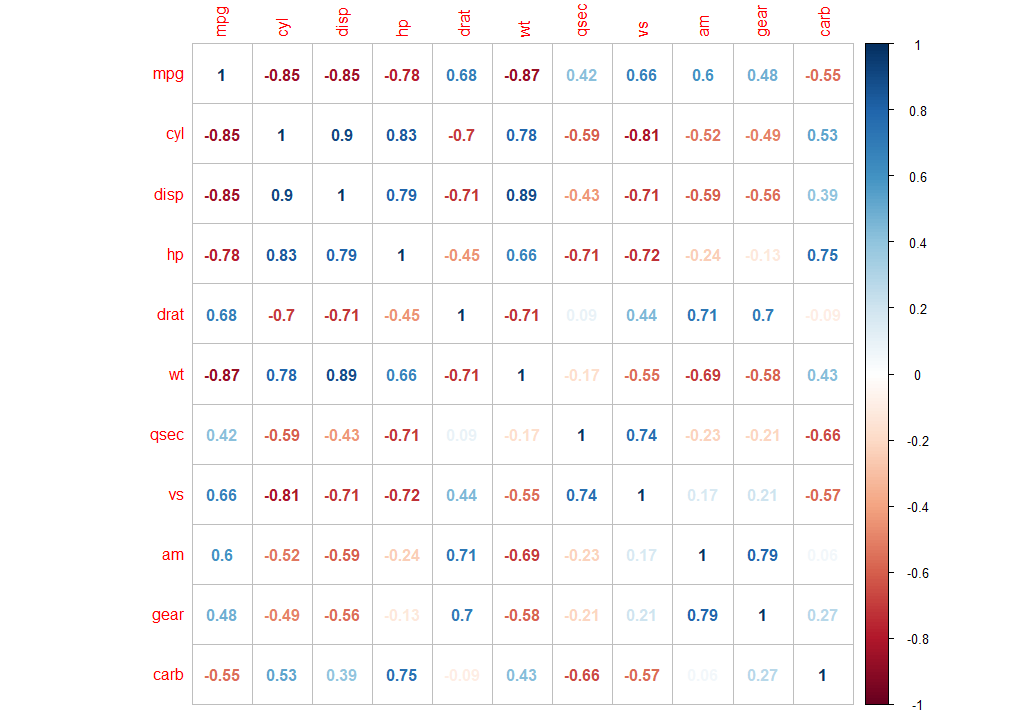
library(corrplot)

Car\_dataset<-cor(mtcars)

corrplot(Car\_dataset, method = "number") ##postive are in blue and negative are in red

data<-mtcars

cor.test(x=data$wt,y=data$mpg) ##test for correlation



1. **“Is car weight (wt) linearly correlated with car gas mileage (mpg), and if so, what is the extent (slope) of this relationship, and how much of the total variance in gas mileage is explained by car weight?”**

**Solution:**

attach(mtcars)

plot(wt, mpg, main="Scatterplot Example",

xlab="Car Weight ", ylab="Miles Per Gallon ", pch=10)

abline(lm(mpg~wt), col="red") # regression line (y~x)

lines(lowess(wt,mpg), col="blue") # lowess line (x,y)

A<-lm(formula = data$mpg ~ data$wt, data = data) ###Summarize your main analytic findings (e.g., parameter estimate, SE, t-test, and p-value, R-squared, the F-statistic, and residual standard error

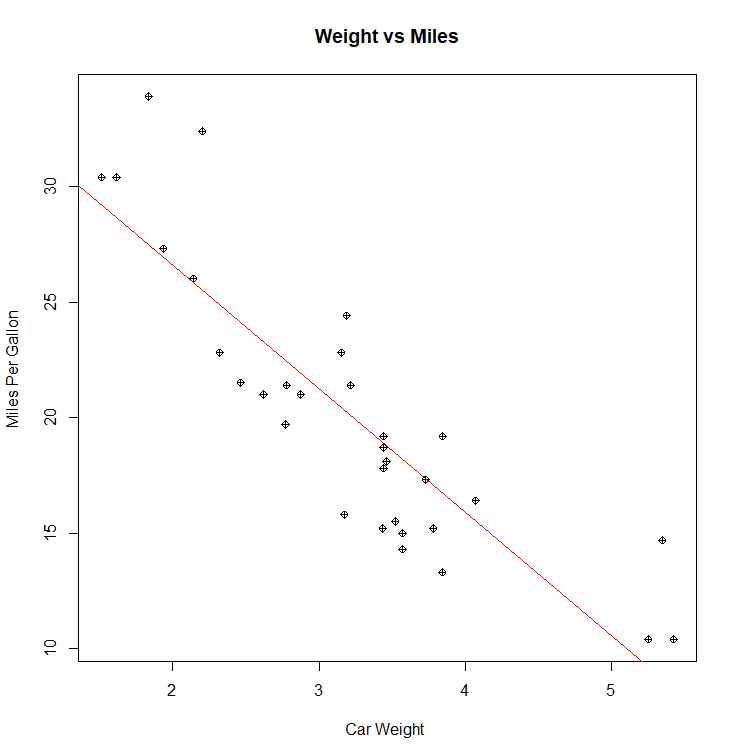
summary(A)

> attach(mtcars)

> plot(wt, mpg, main="Scatterplot Example",

+ xlab="Car Weight ", ylab="Miles Per Gallon ", pch=10)

> abline(lm(mpg~wt), col="red") # regression line (y~x)



> A<-lm(formula = data$mpg ~ data$wt, data = data) ###Summarize your main analytic findings (e.g., parameter estimate, SE, t-test, and p-value, R-squared, the F-statistic, and residual standard error

> summary(A)

Call:

lm(formula = data$mpg ~ data$wt, data = data)

**Residuals:**

**Min 1Q Median 3Q Max**

**-4.5432 -2.3647 -0.1252 1.4096 6.8727**

**Coefficients:**

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

**(Intercept) 37.2851 1.8776 19.858 < 2e-16 \*\*\***

**data$wt -5.3445 0.5591 -9.559 1.29e-10 \*\*\***

**---**

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

**Residual standard error: 3.046 on 30 degrees of freedom**

**Multiple R-squared: 0.7528, Adjusted R-squared: 0.7446**

**F-statistic: 91.38 on 1 and 30 DF, p-value: 1.294e-10**

In the above step we checked, if mpg and weight are linearly correlated or not, by looking at the summary we can say that, they both are negatively correlated i.e. if weight increases mpg decreases. we can also observe that clearly in the scatter plot above. The slope of the relationship between dependent and independent variable is -5.3445, this means as the weight increases by 1-unit mpg decreases by 5.3445.

74.46% of the variation in miles per gallon would be the explained by weight.

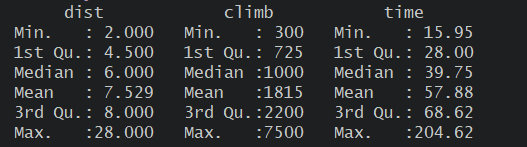
**3. Using a real-world data set that interests you, conduct one of the tests you learned this week (or fit a linear regression model). Make sure to document all steps in the hypothesis testing process, including stating your hypotheses, your code, your output and your findings along with interpretation. You may use data from the MASS library if you wish or load external data.**

**Solution:**

> ##part 3

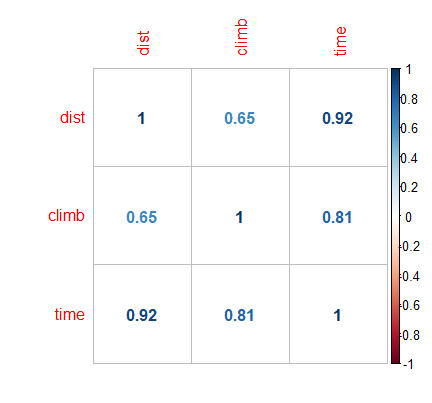
> h <- MASS::hills

> summary(h)



> c<-cor(hills)

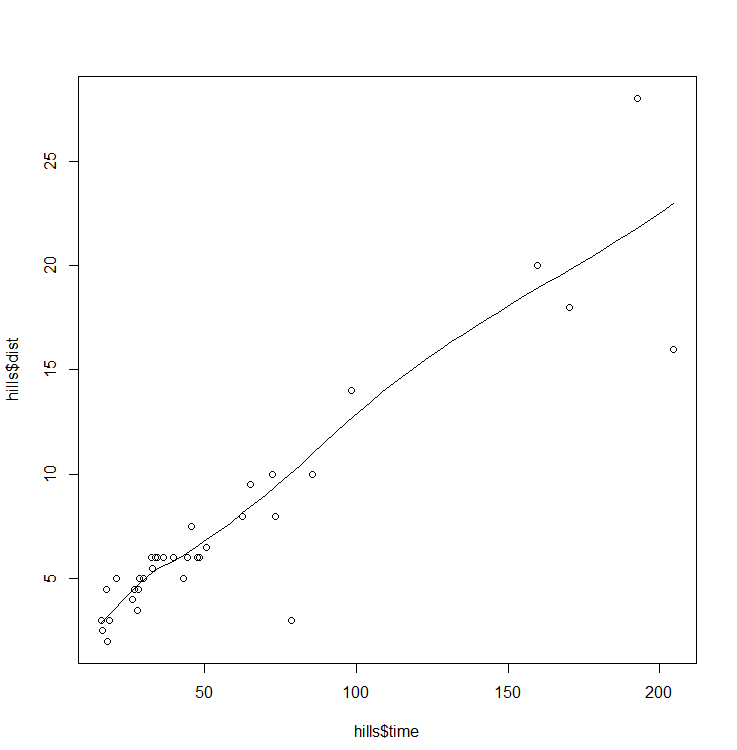
> corrplot(c, method="number")



> m<-lm(dist~time, data = hills)

> scatter.smooth(hills$dist~hills$time)

**Time vs. Distance**



We created a correlation plot that show’s positively correlated data with respect to time and distance for the dataset ‘hills’.

> summary(m)

**Call:**

**lm(formula = dist ~ time, data = hills)**

**Residuals:**

**Min 1Q Median 3Q Max**

**-6.6374 -0.5558 0.0257 0.9714 6.7884**

**Coefficients:**

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

**(Intercept) 1.65347 0.57408 2.88 0.00693 \*\***

**time 0.10151 0.00755 13.45 6.08e-15 \*\*\***

**---**

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

**Residual standard error: 2.203 on 33 degrees of freedom**

**Multiple R-squared: 0.8456, Adjusted R-squared: 0.841**

**F-statistic: 180.8 on 1 and 33 DF, p-value: 6.084e-15**

**> t.test(hills$dist, alternative = "greater", mu=2)**

**One Sample t-test**

**data: hills$dist**

**t = 5.921, df = 34, p-value = 5.452e-07**

**alternative hypothesis: true mean is greater than 2**

**95 percent confidence interval:**

**5.949729 Inf**

**sample estimates:**

**mean of x**

**7.528571**

Ho: distance is less than 2

H1: distance is greater than 2

In t test we are checking if our distance is greater than 2 or not, our null hypothesis is “distance is less than 2” and Alternative hypothesis is “distance is greater than 2”. by the results we can see that the p value is very small so, we can reject null hypothesis that concludes that our distance is greater than 2.

By Analyzing the linear regression between distance and time, we can clearly see that they both have positive regression. by which we can say that, as the distance increases, time increases. We also plotted a corrplot to see their relationship with each other parameter. We can see that time is positively correlated to distance and climbs.

References:

* Spector, P. (2019). Using t-tests in R | Department of Statistics. Statistics.berkeley.edu. Retrieved 6 March 2019, from <https://statistics.berkeley.edu/computing/r-t-tests>
* t.test function | R Documentation. (2019). Rdocumentation.org. Retrieved 6 March 2019, from <https://www.rdocumentation.org/packages/stats/versions/3.5.2/topics/t.test>
* Visualize correlation matrix using correlogram - Easy Guides - Wiki - STHDA. (2019). Sthda.com. Retrieved 6 March 2019, from <http://www.sthda.com/english/wiki/visualize-correlation-matrix-using-correlogram>
* An Introduction to corrplot Package. (2019). Cran.r-project.org. Retrieved 6 March 2019, from <https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html>
* Linear Regression With R. (2019). R-statistics.co. Retrieved 6 March 2019, from <http://r-statistics.co/Linear-Regression.html>
* (2019). Cran.r-project.org. Retrieved 6 March 2019, from <https://cran.r-project.org/web/packages/MASS/MASS.pdf>