Name: Raj Shah

PSID: 1499521

**Home Work 3**

**Prob 1:**

Code:

*library(readr)*

*my\_two\_test<-function(succ,n,type="z"){*

*su1<-succ[1]*

*su2<-succ[2]*

*p\_hat<-sum(succ)/sum(n)*

*n1<-n[1]*

*n2<-n[2]*

*p1\_hat<-su1/n1*

*p2\_hat<-su2/n2*

*ts<-(p1\_hat-p2\_hat)/sqrt(p\_hat\*(1-p\_hat)\*((1/n1)+(1/n2)))*

*p\_val<-2\*(1-pnorm(abs(ts)))*

*if(type == "z"){*

*result<-matrix(c(ts,type,p\_val,p1\_hat,p2\_hat),ncol=5)*

*colnames(result)<-c("TS","Type","P-Value","Sample Estimate","Sample Estimate")*

*result<-as.table(result)*

*return(result)*

*}*

*if(type == "chi"){*

*result<-matrix(c(ts^2,type,p\_val,p1\_hat,p2\_hat),ncol=5)*

*colnames(result)<-c("TS","Type","P-Value","Sample Estimate","Sample Estimate")*

*result<-as.table(result)*

*return(result)*

*}*

*}*

my\_two\_test(c(15,20),c(30,50)) #Default type = “Z”

TS Type P-Value Sample Estimate

A 0.872871560943969 z 0.382733088885226 0.5

Sample Estimate

A 0.4

> my\_two\_test(c(15,20),c(30,50),type="chi")

TS Type P-Value Sample Estimate

A 0.761904761904762 chi 0.382733088885226 0.5

Sample Estimate

A 0.4

prop.test(c(15,20),c(30,50),correct = FALSE)

2-sample test for equality of proportions without continuity

correction

data: c(15, 20) out of c(30, 50)

X-squared = 0.7619, df = 1, p-value = 0.3827

alternative hypothesis: two.sided

95 percent confidence interval:

-0.1246134 0.3246134

sample estimates:

prop 1 prop 2

0.5 0.4

*p.set.1<-c(0.80,0.28)*

*p.set.2<-c(0.47,0.54)*

*p.set.3<-c(0.25,0.25)*

*for (n in c(10^2,10^4,10^6)){*

*b1<-rbinom(1,n,p.set.1[1])*

*b2<-rbinom(1,n,p.set.1[2])*

*vec\_set1<-c(b1,b2)*

*my\_two\_test(vec\_set1,c(n,n))*

*b3<-rbinom(1,n,p.set.2[1])*

*b4<-rbinom(1,n,p.set.2[2])*

*vec\_set2<-c(b3,b4)*

*my\_two\_test(vec\_set2,c(n,n))*

*b5<-rbinom(1,n,p.set.3[1])*

*b6<-rbinom(1,n,p.set.3[2])*

*vec\_set3<-c(b5,b6)*

*my\_two\_test(vec\_set3,c(n,n))*

*}*

Output

[1] 4.440892e-16

[1] 0.0891085

[1] 0.3811502

[1] 0

[1] 0

[1] 0.7066422

[1] 0

[1] 0

[1] 0.7575314

|  |  |  |  |
| --- | --- | --- | --- |
| Table of P-Values | N=10^2 | N=10^4 | N=10^6 |
| P1=0.8, P2=0.28 | 4.440892e-16 | 0 | 0 |
| P1=0.47, P2=0.54 | 0.0891085 | 0 | 0 |
| P1=0.25, P2=0.25 | 0.3811502 | 0.7066422 | 0.7575314 |

For the case 1) The p-value for n=10^2,10^4,10^6, is very low, so we can certainly reject the null hypothesis.

For the case 2) When the # of samples are less, the p-value is close to alpha. So sometimes we may or may not reject the null hypothesis, but when you increase the no. of samples, we can reject the null hypothesis (p1 not equal to p2).

For the case 3) The p-value is greater than alpha (0.05), so for all the # of samples we don’t reject the null hypothesis.

**Prob 2**

Here null hypothesis is H0, which is p1=p2. The value of p1 = 23/30 and p2= 18/31 for drug 1 and 2 respectively.

The alternate hypothesis(Ha) is given by:

p1>p2 or p1<p2.

Fishers Test is more efficient for smaller sample sizes as it calculates exact values. As the no. of samples increases, since Fisher’s Test is more accurate, it becomes computationally difficult

*prop.test(matrix(c(23,18,7,13),2))*

*fisher.test(matrix(c(23,18,7,13),2))*

Output

prop.test(matrix(c(23,18,7,13),2),correct = FALSE)

2-sample test for equality of proportions

without continuity correction

data: matrix(c(23, 18, 7, 13), 2)

X-squared = 2.394, df = 1, p-value = 0.1218

alternative hypothesis: two.sided

95 percent confidence interval:

-0.04436961 0.41641262

sample estimates:

prop 1 prop 2

0.7666667 0.5806452

> fisher.test(matrix(c(23,18,7,13),2))

Fisher's Exact Test for Count Data

data: matrix(c(23, 18, 7, 13), 2)

p-value = 0.1737

alternative hypothesis: true odds ratio is not equal to 1

95 percent confidence interval:

0.6936416 8.4948588

sample estimates:

odds ratio

2.339104

The p-value for fisher’s Exact test is more than X-squared test, which signifies that the presence of hypothesized mean in Confidence interval is indicated more by Fishers Exact test.

The 95% confidence interval includes the odds ratio of 1, it is insufficient evidence to conclude that groups are statistically sufficiently different. The CI in the fisher’s test describes the odds ratio. It’s the ratio of success probability to failure probability.

**Prob 3**

As I had taken a time series dataset for first HW1(I didn’t know about it then), I am taking a new dataset for this example. Here I am trying to see the relationship between the Lung capacity with respect to age of the person.

Code

*library(readxl)*

*LungCapData <- read\_excel("C:/Users/Raj Shah/Desktop/Studies UH/MATH 6359 Stastical Computing/Homework/Answers/HW3/LungCapData.xls")*

*attach(LungCapData)*

*Lungcap.lm<-lm(LungCap~Age,data=LungCapData)*

*summary(Lungcap.lm)*

Output

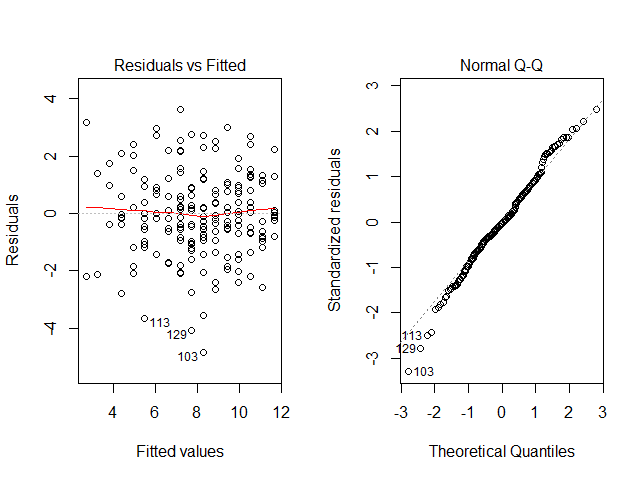
Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.01671 0.36835 2.76 0.00635 \*\*

Age 0.55950 0.02788 20.07 < 2e-16 \*\*\*

*par(mfrow=c(1,2))*

*plot(Lungcap.lm)*



*Lungcap.lm<-lm(LungCap~Age,data=LungCapData,subset=-103) #Removing an outlier 103*

*summary(Lungcap.lm)*

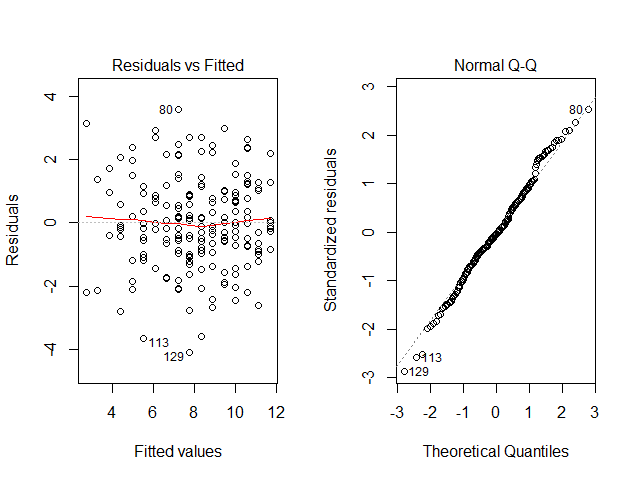
Output

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

(Intercept) 1.03461 0.35845 2.886 0.00436 \*\*

Age 0.56012 0.02712 20.650 < 2e-16 \*\*\*

*plot(Lungcap.lm)*



Summary

There is significant relationship between the Lung capacity and Age of the person, this is evident from the extremely small p-values. It is a linear relationship between two variables. If Age(x) increases by 1 unit, then the Lung capacity(y) increases by 2.4 times. By observing the diagnostic plots, we see there is an outlier. On removing that, doesn’t making any significant difference to overall result, i.e. both the variables have significant relationship. (p-value = 2e-16)

**Prob 4**

*library(readr)*

*library(datasets)*

*data(mtcars)*

*attach(mtcars)*

*pairs(mtcars, gap=0, cex.labels=0.9)*

*cars.lm<-lm(mpg~cyl+disp+hp+drat+wt+qsec+vs+am+gear+carb,data=mtcars)*

*summary(cars.lm)*

*step(cars.lm)*

Output

Call:

lm(formula = mpg ~ cyl + disp + hp + drat + wt + qsec + vs +

am + gear + carb, data = mtcars)

Residuals:

Min 1Q Median 3Q Max

-3.4506 -1.6044 -0.1196 1.2193 4.6271

Coefficients:

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

(Intercept) 12.30337 18.71788 0.657 0.5181

cyl -0.11144 1.04502 -0.107 0.9161

disp 0.01334 0.01786 0.747 0.4635

hp -0.02148 0.02177 -0.987 0.3350

drat 0.78711 1.63537 0.481 0.6353

wt -3.71530 1.89441 -1.961 0.0633 .

qsec 0.82104 0.73084 1.123 0.2739

vs 0.31776 2.10451 0.151 0.8814

am 2.52023 2.05665 1.225 0.2340

gear 0.65541 1.49326 0.439 0.6652

carb -0.19942 0.82875 -0.241 0.8122

---

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

Residual standard error: 2.65 on 21 degrees of freedom

Multiple R-squared: 0.869, Adjusted R-squared: 0.8066

F-statistic: 13.93 on 10 and 21 DF, p-value: 3.793e-07

Step: AIC=61.31

mpg ~ wt + qsec + am

Df Sum of Sq RSS AIC

<none> 169.29 61.307

- am 1 26.178 195.46 63.908

- qsec 1 109.034 278.32 75.217

- wt 1 183.347 352.63 82.790

Call:

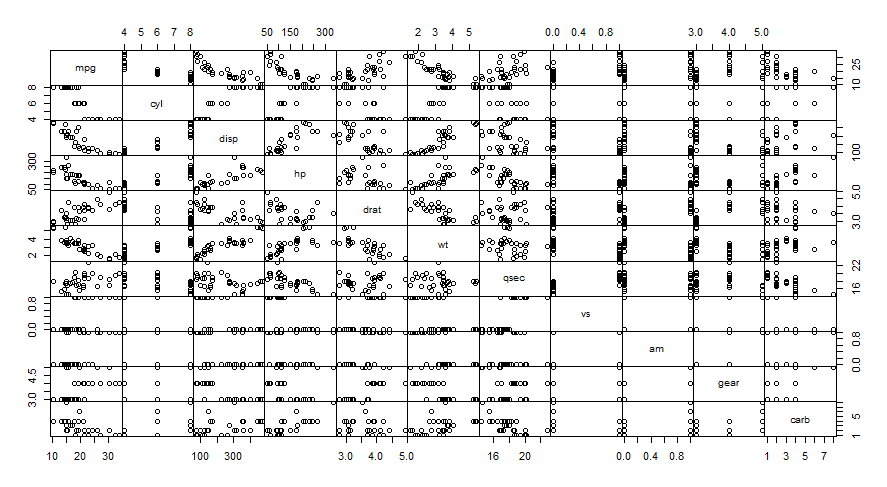
lm(formula = mpg ~ wt + qsec + am, data = mtcars)

Coefficients:

(Intercept) wt qsec am

9.618 -3.917 1.226 2.936

Plot:



Summary:

Performing linear regression on mileage of car on other variables. We see some collinearity between disp and hp.

The linear regression didn’t claim any dominant explanatory variables, but the overall model is significant because of the low p-value(3.973e-07). The step-wise variable selection procedure revealed wt,qsec and am as the optimal subset to describe mpg.

Variable Selection by dropping terms as per domain knowledge

summary(lm(mpg~cyl+disp+hp+drat+wt+qsec+vs+am+gear+carb,data=mtcars))

There were no dominant explanatory variable, so I removed the variable with high p-value.

summary(lm(mpg~cyl+disp+hp+drat+wt+qsec+vs+am+gear,data=mtcars))

Call:

lm(formula = mpg ~ cyl + disp + hp + drat + wt + qsec + vs +

am + gear, data = mtcars)

Residuals:

Min 1Q Median 3Q Max

-3.3038 -1.6964 -0.1796 1.1802 4.7245

Coefficients:

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

(Intercept) 12.83084 18.18671 0.706 0.48790

cyl -0.16881 0.99544 -0.170 0.86689

disp 0.01623 0.01290 1.259 0.22137

hp -0.02424 0.01811 -1.339 0.19428

drat 0.70590 1.56553 0.451 0.65647

wt -4.03214 1.33252 -3.026 0.00621 \*\*

qsec 0.86829 0.68874 1.261 0.22063

vs 0.36470 2.05009 0.178 0.86043

am 2.55093 2.00826 1.270 0.21728

gear 0.50294 1.32287 0.380 0.70745

---

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

Residual standard error: 2.593 on 22 degrees of freedom

Multiple R-squared: 0.8687, Adjusted R-squared: 0.8149

F-statistic: 16.17 on 9 and 22 DF, p-value: 9.244e-08

The wt parameter showed some significance after removing carb. After that I kept selecting variables with higher p-value to be removed in step

summary(lm(mpg~cyl+disp+hp+drat+wt+qsec+vs+am,data=mtcars))

summary(lm(mpg~cyl+disp+hp+drat+wt+qsec+vs,data=mtcars))

summary(lm(mpg~cyl+disp+hp+drat+wt+qsec,data=mtcars))

summary(lm(mpg~cyl+disp+hp+drat+wt,data=mtcars))

summary(lm(mpg~cyl+disp+hp+wt,data=mtcars))

summary(lm(mpg~cyl+hp+wt,data=mtcars))

summary(lm(mpg~cyl+wt,data=mtcars))

Call:

lm(formula = mpg ~ cyl + wt, data = mtcars)

Residuals:

Min 1Q Median 3Q Max

-4.2893 -1.5512 -0.4684 1.5743 6.1004

Coefficients:

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

(Intercept) 39.6863 1.7150 23.141 < 2e-16 \*\*\*

cyl -1.5078 0.4147 -3.636 0.001064 \*\*

wt -3.1910 0.7569 -4.216 0.000222 \*\*\*

---

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

Residual standard error: 2.568 on 29 degrees of freedom

Multiple R-squared: 0.8302, Adjusted R-squared: 0.8185

F-statistic: 70.91 on 2 and 29 DF, p-value: 6.809e-12

After I removed hp, there was significance of wt as well as cyl. Wt was more significant so I removed cyl in the last step.

> summary(lm(mpg~wt,data=mtcars))

Call:

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

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 \*\*\*

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

The elimination method ending up with just one variable wt is coincidence. There is probably connection between disp and hp.