fds <- read.csv(choose.files(),header=TRUE)

NATION HDI LE2013 MEANYRSCH

Afghanistan : 1 Min. :0.3370 Min. :45.60 Min. : 1.300

Albania : 1 1st Qu.:0.5625 1st Qu.:64.95 1st Qu.: 5.450

Algeria : 1 Median :0.7170 Median :73.20 Median : 8.500

Andorra : 1 Mean :0.6856 Mean :70.58 Mean : 7.902

Angola : 1 3rd Qu.:0.8110 3rd Qu.:76.70 3rd Qu.:10.300

Antigua and Barbuda: 1 Max. :0.9440 Max. :83.60 Max. :12.900

(Other) :181

EYRSCH GNI2013 HDI2012 CHINRANK

Min. : 4.10 10,339 : 2 Min. :0.3330 Min. :-5.00000

1st Qu.:11.10 1,011 : 1 1st Qu.:0.5585 1st Qu.: 0.00000

Median :12.90 1,090 : 1 Median :0.7150 Median : 0.00000

Mean :12.81 1,129 : 1 Mean :0.6837 Mean :-0.04278

3rd Qu.:15.10 1,142 : 1 3rd Qu.:0.8100 3rd Qu.: 0.00000

Max. :19.90 1,147 : 1 Max. :0.9430 Max. : 4.00000

(Other):180

DL

high :53

low :43

medium :42

very high:49

summary(fds[,c("MEANYRSCH","EYRSCH")])

MEANYRSCH EYRSCH

Min. : 1.300 Min. : 4.10

1st Qu.: 5.450 1st Qu.:11.10

Median : 8.500 Median :12.90

Mean : 7.902 Mean :12.81

3rd Qu.:10.300 3rd Qu.:15.10

Max. :12.900 Max. :19.90

#In 2012,the MEANYRSCH is skewed left,since the median is larger than the mean,

#Similarly,EYRSCH is also skewed left, yet almost appears normal since the data is

#almost equal.

boxplot(MEANYRSCH,EYRSCH, horizontal=TRUE)



hist(MEANYRSCH)



#the data is left skewed.

hist(EYRSCH)



#The data almost appears normal

qqnorm(MEANYRSCH)



qqline(MEANYRSCH)



#The data almost converges at the midrange

#yet diverge at both sides

qqnorm(EYRSCH)

qqline(EYRSCH)





#The EYRSCH data seems converge quite well

par(mfrow=c(1,2))

hist(EYRSCH)

hist(sqrt(EYRSCH))



#The original histogram of EYRSCH appears more normally distributed

#than the one in the square root,which appears skewed-left

hist(log10(MEANYRSCH))

hist(MEA NYRSCH)

#The graph by log transforming on MEANYRSCH appears more skewed-left than the original one

**e)&&f)**

par(mfrow=c(1,1))

boxplot(HDI,horizontal=TRUE)



hist(HDI)



qqnorm(HDI)

qqline(HDI)



#The histogram of HDI is skewed left

boxplot(HDI ~ DL)



#for high DL,HDL is normally distributed,since its mean seems equal its median

#for Low DL,HDL is left-skewed,since its mean is smaller than its median

#for Medium DL,HDL is almost normal(a bit right skewed),since its mean is a tiny bit larger than its median

#for very high DL,HDL is left-skewed,since its mean is smaller than its median

#To sum up,by categories,they are not all normally distributed.

**Subset part:**

fds <- read.csv(choose.files(),header=TRUE)

fds1<- subset(fds, DL=="very high")

attach(fds1)

par(mfrow=c(1,1))

boxplot(MEANYRSCH,EYRSCH,horizontal=TRUE)



par(mfrow=c(2,1))

hist(MEANYRSCH)

hist(EYRSCH)



#The MEANYRSCH is skewed left while the EYRSCH appears approximately normal

par(mfrow=c(1,1))

qqnorm(MEANYRSCH)

qqline(MEANYRSCH)



qqnorm(EYRSCH)

qqline(EYRSCH)



cor(cbind(MEANYRSCH,EYRSCH,HDI),use= "pairwise.complete.obs")

MEANYRSCH EYRSCH HDI

MEANYRSCH 1.0000000 0.7997989 0.8976834

EYRSCH 0.7997989 1.0000000 0.8950789

HDI 0.8976834 0.8950789 1.0000000

#cor (MEANYRSCH,EYRSCH)is .7997989

#cor (MEANYRSCH,HDI)is .8976834

#cor (EYRSCH,HDI)is .8950789

plot(MEANYRSCH,EYRSCH)

#Yes,it is pretty close.

plot(MEANYRSCH,HDI)

#Since this is better than the plot(MEANYRSCH,EYRSCH),when it

#comes to points convergence,yes,it is close.

plot(EYRSCH,HDI)

#By comparison with plot(MEANYRSCH,HDI),it is not as good while the correlations value is almost the same

**c)&&d)**

lm(HDI~EYRSCH)

Call:

lm(formula = HDI ~ EYRSCH)

Coefficients:

(Intercept) EYRSCH

0.05618 0.04915

model1 =lm(HDI~EYRSCH)

summary(model1)

Call:

lm(formula = HDI ~ EYRSCH)

Residuals:

Min 1Q Median 3Q Max

-0.198922 -0.044069 0.009931 0.043908 0.198798

Coefficients:

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

(Intercept) 0.05618 0.02361 2.379 0.0184 \*

EYRSCH 0.04915 0.00180 27.302 <2e-16 \*\*\*

---

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

Residual standard error: 0.06978 on 185 degrees of freedom

Multiple R-squared: 0.8012, Adjusted R-squared: 0.8001

F-statistic: 745.4 on 1 and 185 DF, p-value: < 2.2e-16

#R-sqr=.8012,80.12% of the variation in HDI is being explained

#by EYRSCH,this is pretty high and good enough

#via t-stat,.0018 is the std.error of .04915

#df=187-(1+1)=185

qt(.975,185)

[1] 1.97287

#since our t-stat + 27.302 is significantly larger than 1.97287

#Ha:Beta is not equal 0,this leads us to believe that EYRSCH is a significant

#predictor of HDI

HDI=.05618+.04915EYRSCH

Error: unexpected input in "HDI=.05618+.04915EY"

# HDI=.05618+.04915EYRSCH

lm(EYRSCH~HDI)

Call:

lm(formula = EYRSCH ~ HDI)

Coefficients:

(Intercept) HDI

1.631 16.302

model2 =lm(EYRSCH~HDI)

summary(model2)

Call:

lm(formula = EYRSCH ~ HDI)

Residuals:

Min 1Q Median 3Q Max

-3.7415 -0.7308 -0.1609 0.7383 3.3208

Coefficients:

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

(Intercept) 1.6307 0.4198 3.885 0.000143 \*\*\*

HDI 16.3015 0.5971 27.302 < 2e-16 \*\*\*

---

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

Residual standard error: 1.271 on 185 degrees of freedom

Multiple R-squared: 0.8012, Adjusted R-squared: 0.8001

F-statistic: 745.4 on 1 and 185 DF, p-value: < 2.2e-16

#R-sqr=.8012,80.12% of the variation in EYRSCH is being explained

#by HDI,which is the same as

#where Y is HDI,when X is EYRSCH

#Next,.5971 is the std.Error for estimate 16.3015

#df is still in 185

#In terms of p-value,both Y by HDI and Y by EYRSCH are the same.

#Yet,the std Error in Y by EYRSCH is smaller,

#Thus,I think Y by HDI works better

#Ho:beta doesn't 0;H1 does 0;

#For EYRSCH=1.6307+16.3015HDI

#df is 185

qt(.975,185)

[1] 1.97287

#Since our t-stat + 27.302 is significantly larger than 1.97287,

#Ho:Beta is not equal 0,this leads us to believe that HDI is a significant

#predictor of EYRSCH

#It is the same the the previous Y by HDI case in the p-value approach

#p-value < 2.2e-16 which is smaller than 0.000143

#Thus,HDI as a predictor of EYRSCH does work fine

**g)**

confint(model2,level=.99)

0.5 % 99.5 %

(Intercept) 0.5381014 2.723232

HDI 14.7475343 17.855483

#The confidence Interval for X=HDI's slope is (14.74753,17.85548)

**h)**

#The confidence Interval for X=HDI's slope is (14.74753,17.85548)

1.6307+16.3015\*5

[1] 83.1382

#Y hat is 83.1382,the residual value = Y - 83.1382

i)

newdata = data.frame(HDI=5)

predict(model2, newdata, interval="confidence", level=.95)

fit lwr upr

1 83.13821 78.05283 88.2236

#The predicted Interval is (78.05283, 88.2236)

#Yet,the predicted interval is not realistic, as even the fit is above the largest value

j)

(16.3015-2)/0.5971

[1] 23.9516

#since 23.9516 is smaller than t-value of HDI:27.302,

#2 is significantly different from the original slope--16.3015