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

attach()

summary(cbind(X, Y))

summary(fds[,c("A","B","C")])

#summary directly on the variables u require

#a boxplot default by vertical

boxplot(A,horizontal=TRUE)

########boxplot(Y~X,horizontal=TRUE)

########boxplot(Y~X)

######i.e. boxplot(HDI~DL)

Make boxplots of your m th variable by a categorical variable.

text-mark:

plot(X,Y)

text(X, Y, "Sebastian")

hist(A)

qqnorm(A)

qqline(A)

#Make a side by side (or straight down) plot or hists

par(mfrow=c(1,2))

hist(EYRSCH)

hist(sqrt(EYRSCH))

sd(claims)—standard dev

#Take a subset

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

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

attach(fds1)

# correlations Table

cor(cbind(Y,X,),use="pairwise.complete.obs")

###weak, strong, positive, negative relation

Single variable data fitting

i.e. model1 = lm(Y ~ X)

Result: Y = kX + b

## Interpret the slope

**for every 1 more unit added to the X, Y would increase by 3.281e-05 in index, holding the other X’s variables constant.**

### Interpret R-SQR

i.e.

R-sqr=.8012,80.12% of the variation in Y is being explained by X ,this is pretty high and good enough.

S is the standard error of coefficient estimate .04915

R-sqr adjusted is the proportion of variation explained by the regression, having taken in account for the penalty on the N explanatory variables in the model.

## Confidence Interval VS Predicted Interval

newdata=data.frame(X=num)

predict(model1, newdata, interval="confidence", level=.95)

predict(model1, newdata, interval="prediction", level=.95)

i.e.expected result:

fit lwr upr

1 83.13821 77.46831 88.80811

## CI= if X has num as its value, we have 95% confidence that predicted HDI would have an average of score at the range of () .

#Focus on the average of a group

##PI=If an individual has num as its value in X, we have 95% confidence that the predicted Y for that individual would have a value at () .

#Focus on the individual’s interval

Confident Intervals

i.e.

confint(model2, level=.99)

**A scatter plot matrix of all explanatory variables and Y variable**

varb = data.frame(x, , , , , Y)

pairs(varb,upper.panel=NULL)

**coincide correlation table with a scatter plot matrix?**

Yes, they are coincide with each other, since the correlations of each pair of variables

could be reflected by a scatter plot. we could see how each plots appears

to fit in a line.

**A significant explanatory variable?**

HDI2012 is definitely the most significant, as it is marked by 3 stars and almost 0 in p-value and is significantly smaller than alph= .05.

**Partial correlation:**

model1 =lm(Y ~ X + M + E + G + H)

model2 =lm(Y~ M + E + G + H)

model3 =lm(X ~ M + E + G + H)

e2 =residuals(model2)

e3 =residuals(model3)

cor(e3,e2)

plot(e3,e2) == plot(X,Y)

Graph matches with corr table:

This graph matches the partial correlation of 0.03861117 because there is hardly any positive correlation. We see a horizontal-like line at best instead of a line with a steep slope. This is extremely different from the scatterplot of the two variables in the matrix. This is because without all the other explanatory variable, these two variables are not much correlated, meaning that other variables has play a big part at predicting Y hat.

**plot label:**

plot(X, Y)

plot(X, Y, xlim=c(0,10), ylim=c(2,9), xlab="X axis", ylab="Y axis")

# Removal of observations

model5 = lm(Y~X, subset=-c(20,21))

Removals of all columns associated with 20 and 21 Observations

model6 = lm(Y~X, subset=-c(20,21,22))

**subset removal on the categorical variables**

term1 = subset(term, subset=FACE>0)

###### term1 = subset(term, subset=FACE!=0)

fds1= subset(fds, subset=DL==1)

attach(fds1)

**Hypothesis Test:**

**i.e.**

qt(.95, DF)

1.653132

Ho: B = 2

H1: B ≠ 2

(16.3015 – 2)/0.5971

[1] 23.9516

T-score

Since **T-score** 23.9516 **> critical t-value** of HDI:1.653132 ,we **reject Ho**:B =2; 2 is significantly different from the original slope of 16.3015.