Simon Chaisouang

Assignment 1:

Part A

(i)

The least-square equation of the regression is given by:

Where is approximately equal to 1.67. If the NOE for a country change from 6.4 to 8.15, the mean expectation of life at birth (ELAB) is expected to change by times the difference between 8.4 and 6.4, under the circumstance that all other covariates remain unchanged. The estimated change in estimated mean response y can is calculated below:

Y is expected to increase by 2.9225 when NOE for a country change from 6.4 to 8.15 and all other covariates remain unchanged.

(ii)

Null Hypothesis (H0): **βGDPC = 0**

Alternative Hypothesis (Ha): **βGDPC ≠ 0**

Computing variance:

The variance is needed in order to compute the test statistic. Sigma hat square was calculated to be approximately 34.43 and a was set to [0,0,1,0] to test the effect of GDPC being nonzero.

Computing the test statistics:

Computing p-value:

The distribution of the test statistic under H0 follows a t-distribution with n-p-1=39 degrees of freedom. The p-value was found to be about 0.04 which is less than 0.05. This means at 5% level, there is enough statistical evidence to prove that the effect of GDPC is significant and nonzero. Therefore, the null hypothesis is rejected.

(iii)

Using the R code from Part B, the 95% upper confidence interval of the difference of effect of NOE and effect of PEA is given below:

(iv)

The point estimate of sum of effect of NOE and effect of GDPC is calculated using the formula below in R:

The point estimate is approximately 1.67. a was set to [0,1,1,0] to test for the sum of effect of NOW and effect of GDPC.

(v)

Null Hypothesis (H0): =

Alternative Hypothesis (Ha): **≠** , At least one of the effects is nonzero.

Computing Test Statistics:

The formula above was used to calculate the test statistics which is approximately 5.30718. A was set to to show that we’re testing for the effect of GDPC or PEA, one being at least nonzero. q is set to 2 since two claims are being tested. The test statistic can then be taken to calculate as shown below:

The distribution of the test statistic under H0 follows an F-distribution with 2 and n-p-1=39 degrees of freedom. The p-value was calculated to be approximately 0.009. Since the p-value is less than 0.005, at level 5, there is enough statistical evidence to prove that at least one of the effects of GDPC and effect of PEA is nonzero. Therefore, the null hypothesis is rejected.

(vi)

Null Hypothesis (H0): **βNOE = βGDPC = βPEA = 0**

Alternative Hypothesis (Ha): **βNOE** or **βGDPC** or **βPEA ≠ 0**, at least one of the effects is nonzero.

Computing the Test Statistics:

The F-statistic was calculated to be approximately 22.51. The test statistic for the significance of regression was calculated by taking the mean square due to regression (MSR) divided by the mean squared error (MSE) in R. The test statistics can then be taken to calculate the p-value using the formula below. q in this case is set 3, to test the 3 effects of the regression.

The distribution of the test statistic under H0 follows an F-distribution with 3 and n-p-1=39 degrees of freedom. The p-value was calculated to be approximately 1.264771e-08 which significantly less than 0.01. This means at level 1%, there is enough statistical evidence to prove that at least one of the covariates (NOE, GDPC, PEA) have a significant effect on the response variable y. Therefore, we reject the null hypothesis under significance level 1%.

(vii)

To find out what percentage of the variation in ELAB is linearly explained by the three covariates can be found from computing R2.

The R2 is approximately 0.634. This means that about 63.4% of the total variation in ELAB can be linearly explained by the three covariates.

Part B

#### Forming the Data

regression\_data = read.table("Data\_Question\_1.txt",header=TRUE)

print(colnames(regression\_data))

y = matrix(regression\_data[ , 1],ncol=1)

n = length(y)

X = matrix(1,n,1)

colnames(X) = "Intercept"

X = cbind(X,regression\_data[, appropriate\_column\_numbers])

X = as.matrix(X)

X = cbind(X,regression\_data[, 2:4])

X = as.matrix(X)

#### Regression Code

p = ncol(X) - 1

output\_information = lm(y~0+X)

beta\_hat = matrix(output\_information$coefficients,ncol=1)

y\_hat = X%\*%beta\_hat

e = matrix(y-y\_hat,ncol=1)

sigmahat\_square = sum(e^2)/(n-p-1)

#### Test of hypothesis for effect of GDPU at 5% Level of Significance

a = matrix(c(0,0,1,0),ncol=1)

alpha = 0.05

variance\_compute = sigmahat\_square\*(t(a)%\*%solve(t(X)%\*%X)%\*%a)

test\_stat = t(a)%\*%beta\_hat/sqrt(variance\_compute)

p\_value = 2\*(1-pt(abs(test\_stat), df = (n-p-1)))

#### Finding 95% Upper Confidence Interval for difference of effect of NOE and effect of PEA

a = matrix(c(0,1,0,-1),ncol=1)

point\_estimate = t(a)%\*%beta\_hat

alpha = 0.05

variance\_compute = sigmahat\_square\* ( t(a)%\*%solve(t(X)%\*%X)%\*%a )

conf\_int\_L = -Inf

conf\_int\_U = point\_estimate +

qt(alpha,df=n-p-1,lower.tail=F)\*sqrt(variance\_compute)

cat("95% Upper Confidence Interval: [", conf\_int\_L, ",", conf\_int\_U, "]\n")

#### Finding the point estimate of sum of effect of NOE and effect of GDPC

a = matrix(c(0,1,1,0),ncol=1)

point\_estimate = t(a)%\*%beta\_hat

#### Simultaneous test of hypotheses for the claim: At least one of the effects of GDPC and effect of PEA is nonzero.

q =2

A = matrix(0,q,p+1)

A[1,] = c(0,0,1,0)

A[2,] = c(0,0,0,1)

Dispersion = A%\*%solve(t(X)%\*%X)%\*%t(A)

test\_stat = t(A%\*%beta\_hat)%\*%solve(Dispersion)%\*%(A%\*%beta\_hat)/(q\*sigmahat\_square)

alpha = 0.05

p\_value = 1 - pf(test\_stat, df1 = q, df2 = n-p-1)

####ToH for significance of regerssion code (from ANOVA)

SSE = t(e)%\*%e

SSR = t(y\_hat)%\*%y\_hat - n\*(mean(y)^2.0)

SST = SSR + SSE

MSR = SSR/p

MSE = sigmahat\_square

Fstatistic = MSR/MSE

q=3

p\_value = 1 - pf(Fstatistic, df1=(q) , df2=(n-p-1))

####Computing R square

R\_square = SSR/SST