Ben Marlow

R Assignment No: 1

Part A

The purpose of this analysis is to perform a linear regression to determine the relationship between the expectation of life at birth (ELAB) and three covariates: number of years of education (NOE), GDP per capita (GDPC), and percentage of population economically active (PEA) for different countries.

The linear regression model can be represented as:

$ELAB = \beta0 + \beta1 * NOE + \beta2 * GDPC + \beta3 * PEA + \varepsilon$

where:

ELAB is the expectation of life at birth.
NOE is the number of years of education.
GDPC is the GDP per capita.
PEA is the percentage of population economically active.
β0, β1, β2, and β3 are the coefficients of the linear regression model.
ε is the error term.

After running the linear regression, we'll need to report the estimated coefficients (β0, β1, β2, and β3), the R-squared value, and any other relevant statistics or insights.

β0 (Intercept): 63.0986246
β1 (NOE): 1.6700662
β2 (GDPC): 0.0003303
β3 (PEA): -0.2400004
Adjusted R-squared: 0.6058
F-statistic: 22.51 on 3 and 39 DF, p-value: 1.265e-08

(i) To estimate the change in the mean expectation of life at birth (ELAB) when the percentage of the economically active population (PEA) increases from 57% to 60.8%, we can use the estimated coefficient for PEA from our linear regression model.

The coefficient for PEA in the linear regression model is β3. To estimate the change in ELAB, we can use the following formula:

Change in ELAB = β3 * (New PEA - Old PEA)

Change in ELAB = β3 * (60.8 - 57)

Our estimate for the change in the mean expectation of life at birth for that country is approximately -0.9120017 years.

(ii) To test the hypothesis at a 5% level of significance for the claim that the effect of GDP per capita (GDPC) is nonzero, we will perform a t-test using the t-statistic and p-value for the GDPC coefficient in our linear regression model.

1. State the null and alternative hypotheses:

H0: β2 = 0 (The effect of GDPC is zero), H1: β2 ≠ 0 (The effect of GDPC is nonzero)

2. Determine the test statistic and p-value:

From the output of our linear regression model, we can obtain the t-statistic and p-value for the GDPC coefficient. The t-statistic for GDPC is 2.071884 and the corresponding p-value is 0.04493548.

3. Decision rule:

Using a 5% level of significance (α = 0.05), we will reject the null hypothesis if the p-value is less than α.

4. Decide based on the p-value:

The p-value for GDPC (p_GDPC) is 0.04493548, which is less than α (0.05). Therefore, we reject the null hypothesis (H0: β2 = 0) in favor of the alternative hypothesis (H1: β2 ≠ 0).

In the context of the problem, this means that there is evidence to support the claim that the effect of GDP per capita (GDPC) on the expectation of life at birth (ELAB) is nonzero at a 5% level of significance.

(iii) To find a 95% confidence interval for the difference of the effect of the number of years of education (NOE) and the effect of the percentage of the population economically active (PEA), we can use the standard errors and covariance of the coefficients in our linear regression model.

1. Calculate the standard error for the difference:

Let's assume the standard errors for NOE and PEA coefficients are SE_NOE and SE_PEA, and the covariance between NOE and PEA is Cov_NOE_PEA. We can calculate the standard error for the difference (SE_diff) using the following formula:

SE_diff = sqrt(0.3928538^2 + 0.09644121^2 - 2 * 0.01567128)

SE_NOE: 0.3928538

SE_PEA: 0.09644121

Cov_NOE_PEA: 0.01567128

2. Calculate the 95% confidence interval:

To calculate the confidence interval, we can use the following formula:

CI = (β1 - β3) ± t_critical * SE_diff

Here, β1 is the coefficient for NOE, β3 is the coefficient for PEA, t_critical is the critical t-value for a 95% confidence interval, and SE_diff is the standard error for the difference calculated in step 1.

t_critical: 2.022691

SE_diff: 0.3637203

CI = (1.174373, 2.645760)

(iv) To find the point estimate of the sum of the effect of the number of years of education (NOE) and the effect of GDP per capita (GDPC), we can simply add the coefficients for NOE (β1) and GDPC (β2) in our linear regression model.

Point estimate = β1 + β2 = 1.670396

β1: 1.670066, β2: 0.0003302573

(v) To simultaneously test the hypotheses for the claim that at least one of the effects of GDP per capita (GDPC) and the percentage of the population economically active (PEA) is nonzero, we will perform an F-test using the F-statistic and p-value from the full model and a reduced model. The reduced model will exclude both GDPC and PEA as covariates.

1. State the null and alternative hypotheses:

H0: β2 = β3 = 0 (The effects of GDPC and PEA are both zero)

H1: At least one of β2 or β3 is nonzero (At least one of the effects of GDPC and PEA is nonzero)

2. Determine the F-statistic and p-value:

Fit a reduced model without GDPC and PEA, and compare the full model (with NOE, GDPC, and PEA) with the reduced model (with only NOE) using an F-test. Let's assume the F-statistic is F_value and the corresponding p-value is p_value.

F_value: 5.30718

p_value: 0.009149882

3. Decision rule:

Using a 5% level of significance (α = 0.05), we will reject the null hypothesis if the p-value is less than α.

4. Decide based on the p-value:

The p-value (0.009149882) is less than the chosen significance level of α = 0.05. Therefore, we reject the null hypothesis H0: β2 = β3 = 0. This means that there is evidence to suggest that at least one of the effects of GDP per capita (GDPC) and the percentage of the population economically active (PEA) is nonzero, in the context of the problem.

(vi) To test the overall significance of the regression, we will perform an F-test using the F-statistic and p-value obtained from the full model.

1. State the null and alternative hypotheses:

H0: β1 = β2 = β3 = 0 (All the regression coefficients are zero)

H1: At least one of β1, β2, or β3 is nonzero (At least one of the regression coefficients is nonzero)

2. Determine the F-statistic and p-value:

Use the F-statistic and p-value from the full model. Let's assume the F-statistic is F_model and the corresponding p-value is p_model.

F_model: 22.51314

p_model: 1.264771e-08

3. Decision rule:

Using a 1% level of significance ($\alpha = 0.01$), we will reject the null hypothesis if the p-value is less than α.

4. Decide based on the p-value:

The p-value (1.264771e-08) is less than the chosen significance level of $\alpha = 0.01$. Therefore, we reject the null hypothesis H0: $\beta1 = \beta2 = \beta3 = 0$. This means that there is evidence to suggest that at least one of the regression coefficients (NOE, GDPC, or PEA) is nonzero, indicating that the regression model is significant in explaining the variation in the expectation of life at birth (ELAB).

Part B

```
> # Load required libraries
> library(readr)
>
> # Read the data from the text file
> data_file <- "~/Desktop/STAT-5313/R/Data_Question_01.txt"
> data_raw <- readLines(data_file)
>
> # Process the data
> data_matrix <- do.call(rbind, lapply(data_raw[-1], function(x) {
+   as.numeric(unlist(strsplit(x, "\\s+")))
+ }))
>
> # Assign column names and convert to data.frame
> data <- data.frame(data_matrix)
> colnames(data) <- c("ELAB", "NOE", "GDPC", "PEA")
>
> # Inspect the data
> print(data)
   ELAB   NOE  GDPC   PEA
1  46.05  2.65   165 84.15
2  47.15  4.55   205 90.35
3  66.50  8.80   994 63.10
4  75.30 11.75  4736 55.00
5  72.30 11.60  4014 54.05
6  76.10 11.30  1983 56.90
7  71.00 11.20  1508 62.75
8  66.00  9.80   973 47.50
9  69.50  9.75  1660 60.20
```

```
10 69.45 12.55  2433 62.80
11 47.05  5.05   321 57.50
12 72.60 10.25   343 66.50
13 78.05 13.75  8684 50.40
14 64.50  9.55   726 60.25
15 78.95 12.65 22898 61.90
16 69.15 12.45  4325 52.25
17 65.15  9.95  1019 68.05
18 62.40  8.25 11308 42.15
19 74.60 10.95  1779 68.20
20 72.40 14.50  9736 62.40
21 76.15  9.00 15757 57.20
22 68.40 11.40  1764 60.65
23 40.70 10.45   142 77.95
24 76.85 13.40  8793 50.75
25 46.95  3.40    77 87.35
26 68.20  9.15   464 58.95
27 69.75  9.15  1860 55.90
28 68.40 12.35  2497 42.55
29 68.40 11.00  1093 65.85
30 71.20 13.15  3058 58.60
31 75.35 14.30 10428 58.10
32 72.70 11.10 14013 60.25
33 69.60 11.45  1570 65.05
34 69.30 11.55  1106 58.75
35 71.65  8.70  6583 42.85
36 65.30 13.10  3230 65.00
37 78.00 15.50 14111 50.05
38 60.00 11.15  1389 46.00
39 73.85 10.70  4083 60.20
40 69.10  9.65  2814 53.25
41 75.20 10.05 17690 58.35
42 76.75 15.80 26037 67.10
43 72.85 10.45  3496 61.65
>
> # Check for missing or problematic values
> print(summary(data))
      ELAB            NOE            GDPC            PEA
 Min.   :40.70   Min.   : 2.65   Min.   :   77   Min.   :42.15
 1st Qu.:66.25   1st Qu.: 9.60   1st Qu.: 1056   1st Qu.:54.52
 Median :69.60   Median :11.00   Median : 2433   Median :60.20
 Mean   :68.11   Mean   :10.63   Mean   : 5160   Mean   :60.20
 3rd Qu.:74.22   3rd Qu.:12.40   3rd Qu.: 7634   3rd Qu.:64.05
 Max.   :78.95   Max.   :15.80   Max.   :26037   Max.   :90.35
```

```
>
> # Remove rows with missing or problematic values
> data_clean <- na.omit(data)
>
> # Perform the linear regression
> model <- lm(ELAB ~ NOE + GDPC + PEA, data = data_clean)
>
> # Display the summary of the regression model
> summary(model)

Call:
lm(formula = ELAB ~ NOE + GDPC + PEA, data = data_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-21.1897  -2.1588   0.8958   3.6396   8.9947

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 63.0986246  8.3696575   7.539 3.94e-09 ***
NOE          1.6700662  0.3928538   4.251 0.000128 ***
GDPC         0.0003303  0.0001594   2.072 0.044935 *
PEA         -0.2400004  0.0964412  -2.489 0.017203 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.869 on 39 degrees of freedom
Multiple R-squared:  0.6339,  Adjusted R-squared:  0.6058
F-statistic: 22.51 on 3 and 39 DF,  p-value: 1.265e-08
```

(i)
```
> # Extract the coefficient for PEA (beta3)
> beta3 <- coef(model)["PEA"]
>
> # Calculate the change in ELAB
> old_PEA <- 57
> new_PEA <- 60.8
> change_in_ELAB <- beta3 * (new_PEA - old_PEA)
>
> # Print the change in ELAB
> print(change_in_ELAB)
      PEA
-0.9120017
```

(ii)
```
> # Extract the t-statistic and p-value for the GDPC coefficient
```

```
> t_GDPC <- summary(model)$coefficients["GDPC", "t value"]
> p_GDPC <- summary(model)$coefficients["GDPC", "Pr(>|t|)"]
>
> # Print the t-statistic and p-value for GDPC
> print(t_GDPC)
[1] 2.071884
> print(p_GDPC)
[1] 0.04493548
```

(iii)
```
> # Extract the coefficients, standard errors, and covariance
> beta1 <- coef(model)["NOE"]
> beta3 <- coef(model)["PEA"]
> SE_NOE <- summary(model)$coefficients["NOE", "Std. Error"]
> SE_PEA <- summary(model)$coefficients["PEA", "Std. Error"]
> Cov_NOE_PEA <- vcov(model)["NOE", "PEA"]
>
> # Calculate the standard error for the difference
> SE_diff <- sqrt(SE_NOE^2 + SE_PEA^2 - 2 * Cov_NOE_PEA)
>
> # Calculate the critical t-value for a 95% confidence interval
> df <- nrow(data_clean) - length(coef(model)) # Degrees of freedom
> t_critical <- qt(0.975, df)
>
> # Calculate the 95% confidence interval
> CI_lower <- (beta1 - beta3) - t_critical * SE_diff
> CI_upper <- (beta1 - beta3) + t_critical * SE_diff
>
> # Print the 95% confidence interval
> print(c(CI_lower, CI_upper))
    NOE     NOE
1.174373 2.645760
```

(iv)
```
> # Extract the coefficients for NOE and GDPC
> beta1 <- coef(model)["NOE"]
> beta2 <- coef(model)["GDPC"]
>
> # Calculate the point estimate of the sum of the effects
> point_estimate <- beta1 + beta2
>
> # Print the point estimate
> print(point_estimate)
    NOE
1.670396
>
```

```
> print(beta1)
    NOE
1.670066
> print(beta2)
      GDPC
0.0003302573
```

(v)
```
> # Fit the reduced model without GDPC and PEA
> reduced_model <- lm(ELAB ~ NOE, data = data_clean)
>
> # Calculate the residual sum of squares for the full and reduced models
> rss_full <- sum(resid(model)^2)
> rss_reduced <- sum(resid(reduced_model)^2)
>
> # Calculate the F-statistic
> numerator <- (rss_reduced - rss_full) / 2
> denominator <- rss_full / (nrow(data_clean) - 4)
> F_value <- numerator / denominator
>
> # Calculate the p-value
> p_value <- 1 - pf(F_value, 2, nrow(data_clean) - 4)
>
> # Print the F-statistic and p-value
> cat("F-statistic:", F_value, "\n")
F-statistic: 5.30718
> cat("p-value:", p_value, "\n")
p-value: 0.009149882
```

(vi)
```
> # Extract the F-statistic and p-value from the full model summary
> F_model <- summary(model)$fstatistic[1]
> p_model <- pf(F_model, summary(model)$fstatistic[2],
summary(model)$fstatistic[3], lower.tail = FALSE)
>
> # Print the F-statistic and p-value
> cat("F-statistic:", F_model, "\n")
F-statistic: 22.51314
> cat("p-value:", p_model, "\n")
p-value: 1.264771e-08
```