(1) To get a sense of the data, generate a scatterplot to examine the association between prestige score and years of education. Briefly describe the form, direction, and strength of the association between the variables. Calculate the correlation. (3 points )

Chart, scatter chart

Description automatically generated

Based on the graph, this is a linear association. And there is a strong positive correlation between prestige score and years of education. As the years of education increase, the prestige score is also increase.

# Calculate the correlation

cor(canada$Prestige.Score, canada$Education.Level..years.)

#0.8501769

(2) Perform a simple linear regression

lg = lm(canada$Prestige.Score ~ canada$Education.Level..years.)

summary(lm(canada$Prestige.Score ~ canada$Education.Level..years.))Graphical user interface, text, application

Description automatically generated

Text, letter

Description automatically generated

Generate a residual plot

Diagram, schematic

Description automatically generated

Assess whether the model assumptions are met?

For Q-Q plot, the data points do fellow the line in the Q-Q plot. Therefore, it is normal.

For “Scale-Location”, there do have a constant variance.

For “Residuals vs Leverage”, there are observations that are being influential.

Thus, the assumptions are met.

Are there any outliers or influence points? If so, identify them by ID and comment on the effect of each on the regression. (3 points)

Chart, histogram

Description automatically generated

The outliers are 24, 53, 67. The 24 and 53 are close to the dataset, they have very little effect. However, 67 is very far away from the dataset. It will have big influence with mean, correlation, and regression. 67 will make error rate higher. And base on “residuals vs leverage” graph, 67 and 24 are above the regression line.

(3) Calculate the least squares regression equation that predicts prestige from education, income, and percentage of women. Formally test whether the set of these predictors are associated with prestige at the = 0.05 level. (6 points)



(1)

predicts prestige = -6.794 + 4.1867Education + 0.001314Income + (-0.008905) percentage of women

(2)

- samples must be independent (not influencing each other) and randomly selected from the two distinct populations of interest

- the variable of interest must be measured in the same way in each of the populations

- the parameter of interest should be normally distributed (or at least have similar shapes and without outliers)

HO: the set of these predictors are not associated with prestige

H1: the set of these predictors are associated with prestige

F – STATIC = 129.2

p-value: < 2.2e-16

Since F – STATIC > P, therefore, we reject the the null hypothesis and conclude that the set of these predictors are associated with prestige.

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text

Description automatically generated

(4) If the overall model was significant, summarize the information about the contribution of each variable separately at the same significance level as used for the overall model (no need to do a formal 5-step procedure for each one, just comment on the results of the tests). Provide interpretations for any estimates that were significant. Calculate 95% confidence intervals where appropriate. (5 points)

Graphical user interface, text

Description automatically generated

Since the t-value for Education level is 10.771, it is greater than P, then, Education level is significance contribution to Prestige Score.

Since the t-value for Income is 4.729, it is greater than P, then, income is significance contribution to Prestige Score.

Since the t-value for Percent of Workforce that are Women is -0.293, it is not greater than P(0.7702), then, Percent of Workforce that are Women is not significance contribution to Prestige Score.

For education level: by increasing 1 year of education, the Prestige Score will increase 4.1867.

For income: By increasing 1 dollars of income, the Prestige Score will increase 0.0013136.

# Calculate 95% confidence intervals

Text

Description automatically generated with medium confidence

1. 95% confidence intervals for education level is(3.4152, 4.958)
2. 95% confidence intervals for income is (0.00076, 0.00186)

(5) Generate a residual plot showing the fitted values from the regression against the residuals. Is the fit of the model reasonable? Are there any outliers or influence points? (3 points)

Chart, scatter chart

Description automatically generated

the model is not reasonable, because there do not have an equally spread residuals around a horizontal line without distinct patterns.

There are outliers in the data.

R code:

library(UsingR)

setwd("/Users/estherji")

#Part 1

canada <- read.csv("Canadian1970.csv", header=TRUE)

head(canada)

#(1)

plot(canada$Education.Level..years.,canada$Prestige.Score)

cor(canada$Prestige.Score, canada$Education.Level..years.)

#0.8501769

#(2)

lg = lm(canada$Prestige.Score ~ canada$Education.Level..years.)

summary(lm(canada$Prestige.Score ~ canada$Education.Level..years.))

#residual plot

par(mfrow = c(2,2))

plot(lm(canada$Prestige.Score ~ canada$Education.Level..years.))

par(mfrow = c(1,1))

#outlier

plot(lg, 4)

#3

rg = lm(canada$Prestige.Score~canada$Education.Level..years.+canada$Income....+canada$Percent.of.Workforce.that.are.Women)

summary(rg)

#4

summary(rg)

confint (rg)

#5

#residual plot

par(mfrow = c(2,2))

plot(lm(canada$Prestige.Score ~ canada$Education.Level..years.))

par(mfrow = c(1,1))