MET CS555: Homework 4

Patrick Ryan



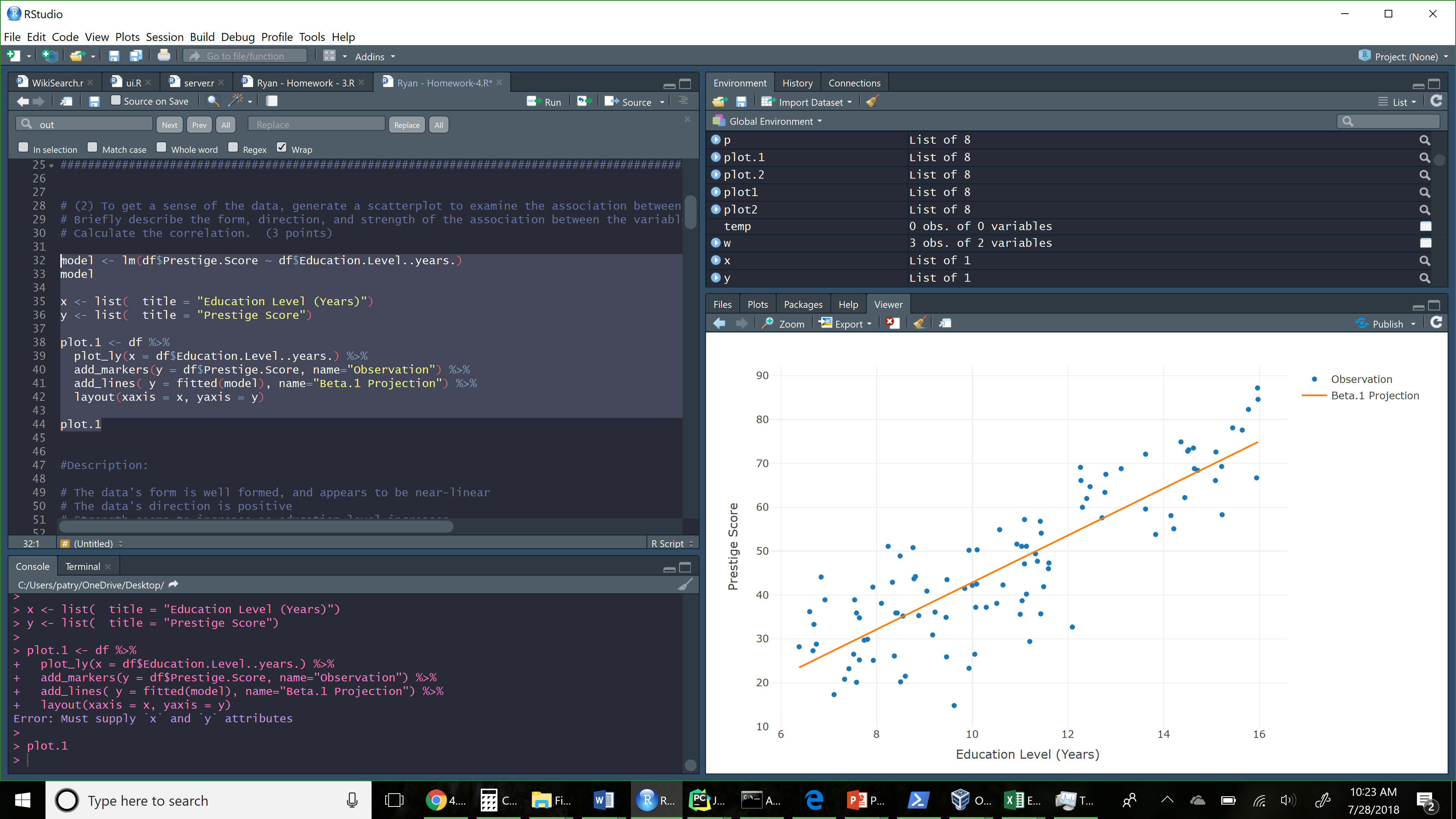
**(1) Save the data to excel or CSV file and read into R for analysis. (1 point)**

|  |
| --- |
| setwd("C:/Users/patry/OneDrive/Desktop")  # Assign the .csv information to a dataframe  df = read.csv("homework4.csv", header=T, sep=",", fileEncoding="UTF-8-BOM", stringsAsFactors = F)  df  library(plotly) |

**(2) 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)**

|  |
| --- |
| model <- lm(df$Prestige.Score ~ df$Education.Level..years.)  model  x <- list( title = "Education Level (Years)")  y <- list( title = "Prestige Score")  plot.1 <- df %>%  plot\_ly(x = df$Education.Level..years.) %>%  add\_markers(y = df$Prestige.Score, name="Observation") %>%  add\_lines( y = fitted(model), name="Beta.1 Projection") %>%  layout(xaxis = x, yaxis = y)  plot.1 |

[Figure 1: Two-Dimensional (Education vs Prestige) Scatter Plot with Beta.1 Trend Line]



Description:

* The data's form is well formed, and appears to be relatively linear
* The data's direction is positive (and thus y increases as x increases)
* Strength seems to increase as education level increases, yielding lower variance, (although this could be also be due to the relative decreasing number of points for education years > 12)

(3) Perform a simple linear regression. Generate a residual plot. Assess whether the model 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. (4 points)

Discussion of Model Assumptions:

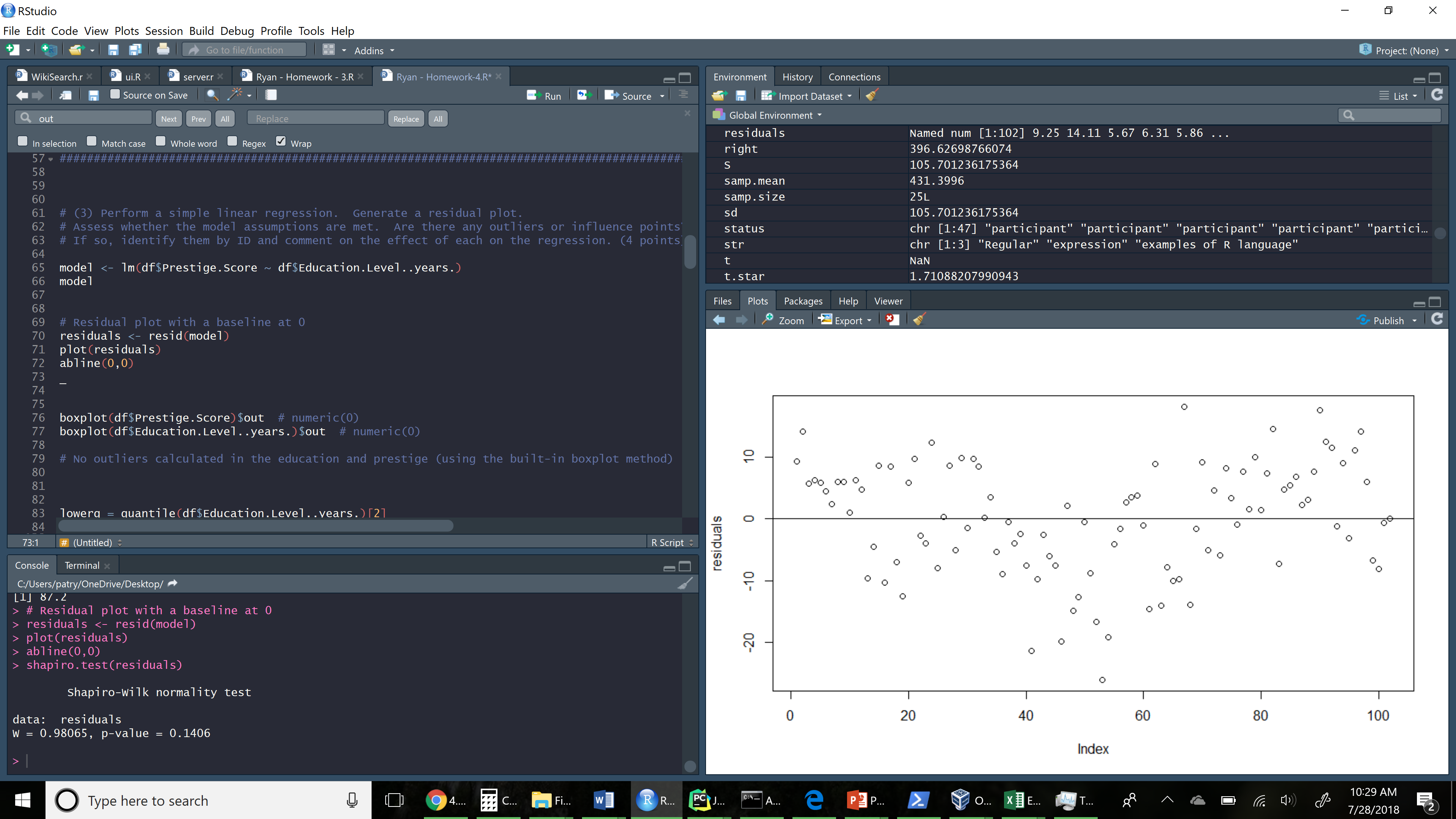
1) Linearity and additivity of the relationship: Mostly (negative )

2) The observations are independent: Yes

3) The variation of the response variable around the regression line is constant: Mostly

4) Normality of the error distribution: Yes, see Shapiro test results. (w=0.98065)

[Figure 2: Two-Dimensional Residual Plot (Residual value vs. Observation)]

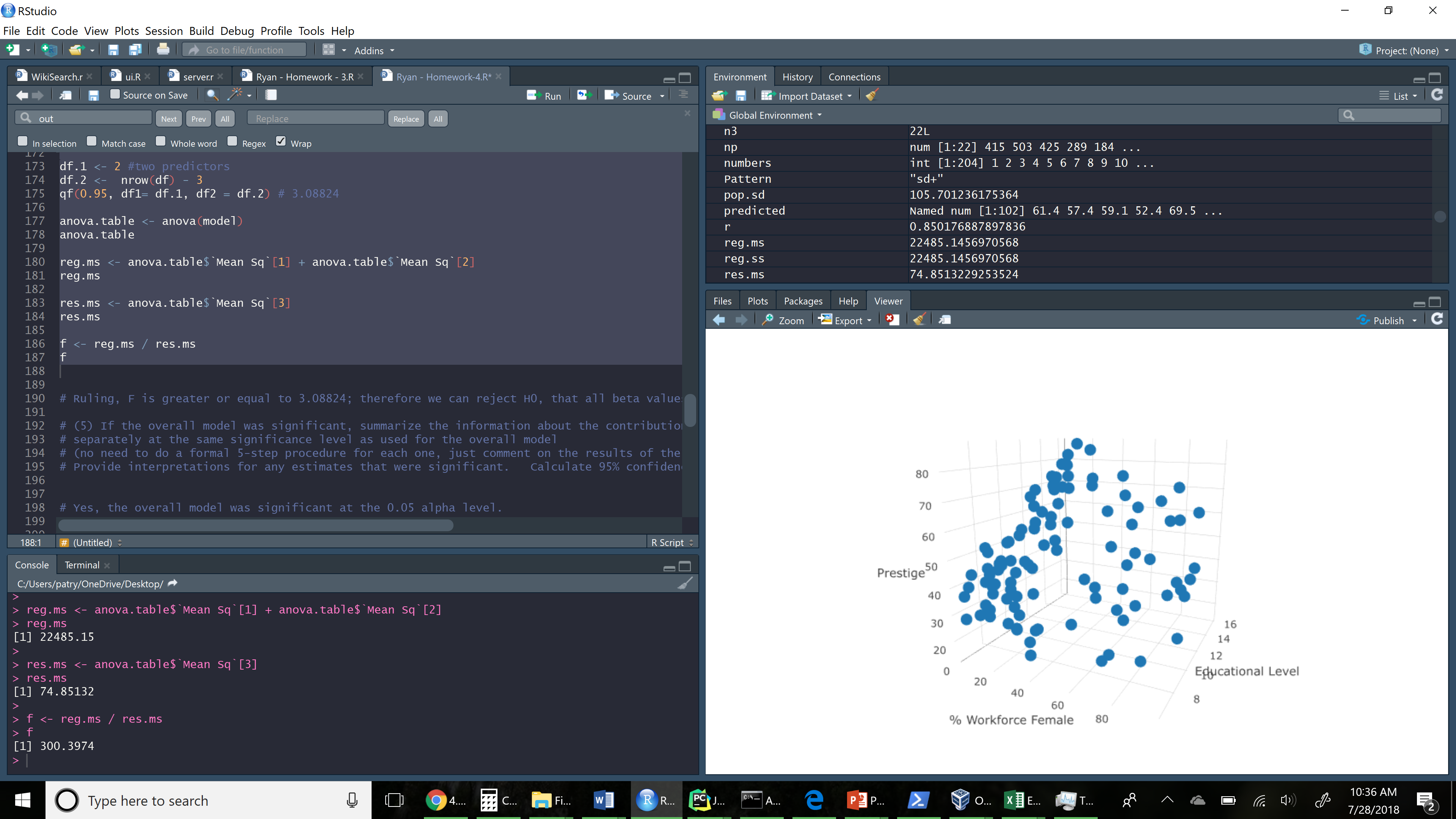


|  |
| --- |
| model <- lm(df$Prestige.Score ~ df$Education.Level..years.)  model  # Residual plot with a baseline at 0  residuals <- resid(model)  plot(residuals)  abline(0,0)  boxplot(df$Prestige.Score)$out # numeric(0)  boxplot(df$Education.Level..years.)$out # numeric(0)  # No outliers calculated in the education and prestige (using the built-in boxplot method)  lowerq = quantile(df$Education.Level..years.)[2]  upperq = quantile(df$Education.Level..years.)[4]  iqr <- IQR(df$Education.Level..years.)  threshold.upper = (iqr \* 1.5) + upperq  threshold.lower = lowerq - (iqr \* 1.5)  threshold.lower # 2.14125  threshold.upper # 18.95125  subset(df$Education.Level..years., df$Education.Level..years. < threshold.lower)  subset(df$Education.Level..years., df$Education.Level..years. > threshold.upper)  lowerq = quantile(df$Prestige.Score)[2]  upperq = quantile(df$Prestige.Score)[4]  iqr <- IQR(df$Prestige.Score)  threshold.upper = (iqr \* 1.5) + upperq  threshold.lower = lowerq - (iqr \* 1.5)  threshold.lower # -0.85  threshold.upper # 95.35  subset(df$Prestige.Score, df$Prestige.Score < threshold.lower)  subset(df$Prestige.Score, df$Prestige.Score > threshold.upper)  # After manually calculating we can see that there are still no outliers (defined as IQR \* 1.5 in both directions)  # This can be further confirmed by visualing comparing the scatterplot, or by identifying the max and min for each variable.  min(df$Prestige.Score) # 14.8  max(df$Prestige.Score) # 87.2  shapiro.test(residuals)  # Shapiro-Wilk normality test  #  # data: residuals  # W = 0.98065, p-value = 0.1406 |

**(4) 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. (4 points)**

|  |
| --- |
| model <- lm(df$Prestige.Score ~ df$Education.Level..years. + df$Percent.of.Workforce.that.are.Women)  model  # Two different summary tables for the model  summary(model)  anova(model)  # 3d rendering of the plot  plot.2 <- plot\_ly(df, z = ~df$Prestige.Score, y = ~df$Education.Level..years., x = ~df$Percent.of.Workforce.that.are.Women) %>%  add\_markers() %>%  layout(scene = list(zaxis = list(title = 'Prestige'),  yaxis = list(title = 'Educational Level'),  xaxis = list(title = '% Workforce Female')))  plot.2  #1) Set up hypothesis and select the alpha level:  # H0: B(education.level) and B(%.female) == 0  # H1: B(education.level) != 0 and/or B(%.female) != 0  #2) Select the appropriate test statistics:  # F = reg.ms/res.ms  #3) State the decision rule:  # Reject H0 if F >= 3.08824; Otherwise, do not reject H0  df.1 <- 2 #two predictors  df.2 <- nrow(df) - 3  qf(0.95, df1= df.1, df2 = df.2) # 3.08824  anova.table <- anova(model)  anova.table  reg.ms <- anova.table$`Mean Sq`[1] + anova.table$`Mean Sq`[2]  reg.ms  res.ms <- anova.table$`Mean Sq`[3]  res.ms  f <- reg.ms / res.ms  f  # Ruling, F is greater or equal to 3.08824; therefore we can reject H0, that all beta values are 0. |

[Figure 3: Three-Dimensional Scatterplot (Education Level, % Female Workforce and Prestige)]



**(5) 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. (4 points)**

|  |
| --- |
| # Yes, the overall model was significant at the 0.05 alpha level.  anova(model)  # Analysis of Variance Table  #  # Response: df$Prestige.Score  # Df Sum Sq Mean Sq F value Pr(>F)  # df$Education.Level..years. 1 21608.4 21608.4 288.685 < 2.2e-16 \*\*\*  # df$Percent.of.Workforce.that.are.Women 1 876.7 876.7 11.713 0.0009039 \*\*\*  # Residuals 99 7410.3 74.9  # ---  # Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  summary(model)  # Education p-value: < 2e-16  # Female Worlforce % p-value: 0.000904  # Call:  # lm(formula = df$Prestige.Score ~ df$Education.Level..years. +  # df$Percent.of.Workforce.that.are.Women)  #  # Residuals:  # Min 1Q Median 3Q Max  # -28.010 -4.069 1.050 5.027 18.942  #  # Coefficients:  # Estimate Std. Error t value Pr(>|t|)  # (Intercept) -8.75416 3.54213 -2.471 0.015164 \*  # df$Education.Level..years. 5.42780 0.31612 17.170 < 2e-16 \*\*\*  # df$Percent.of.Workforce.that.are.Women -0.09305 0.02719 -3.422 0.000904 \*\*\*  # ---  # Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  #  # Residual standard error: 8.652 on 99 degrees of freedom  # Multiple R-squared: 0.7521, Adjusted R-squared: 0.7471  # F-statistic: 150.2 on 2 and 99 DF, p-value: < 2.2e-16  # Education Confidence Interval:  qt(0.025, df=nrow(df), lower.tail=F)  education.lower.ci <- 5.42780 - ( qt(0.025, df=97, lower.tail= F) \* 0.31612 )  education.lower.ci # 4.800389  education.upper.ci <- 5.42780 + ( qt(0.025, df=97, lower.tail= F) \* 0.31612 )  education.upper.ci # 6.055211  # Female Workforce Confidence Interval:  female.workforce.lower.ci <- -0.09305 - ( qt(0.025, df=97, lower.tail= F) \* 0.02719 )  female.workforce.lower.ci # -0.1470146  female.workforce.upper.ci <- -0.09305 + ( qt(0.025, df=97, lower.tail= F) \* 0.02719 )  female.workforce.upper.ci # -0.03908538 |

Significance of findings:

* Education Level has a significant influence prestige, while the percent of female workers in a given sector had a much lesser influence on prestige.
  1. Education Beta.1: 5.42780
  2. Education v.s. Prestige r-value: 0.8501
  3. % Female Beta.1: -0.09305
  4. % Female r-value: -0.1183
* Both independent variables have very low p-values, indicating that the observed value has a very low likelihood of having arisen by chance:

1. Education p-value: < 2e-16
2. Female Workforce % p-value: 0.000904

**(6) Generate a residual plot showing the fitted values from the regression against the residuals. Is the fit of the model reasonable? (2 points):**

Yes, this is a moderately reasonable fit, with relatively few outliers.

|  |
| --- |
| numbers <- c(1:102,1:102)  numbers  values <- c(resid(model), fitted(model))  tags <- c(rep("residual", 102), rep("fitted", 102))  df <- data.frame(residuals=values, tag = tags, numbers=numbers)  df  p <- plot\_ly(y = ~df$residuals, x = numbers, color = df$tag, type="scatter" )  p |

**(7) Are there any outliers or influence points? (2 points)**

Yes, there are three outliers in the residual set of points, ( -18.258, -28.009 and -21.500).

There are no outliers in the fitted set of points.

|  |
| --- |
| boxplot(resid(model))$out  boxplot(fitted(model))$out  # Residual outliers  # 46 53 54  # -18.25840 -28.00991 -21.50051 |