ASSIGNMENT 4  
BU ID : U55-32-1699

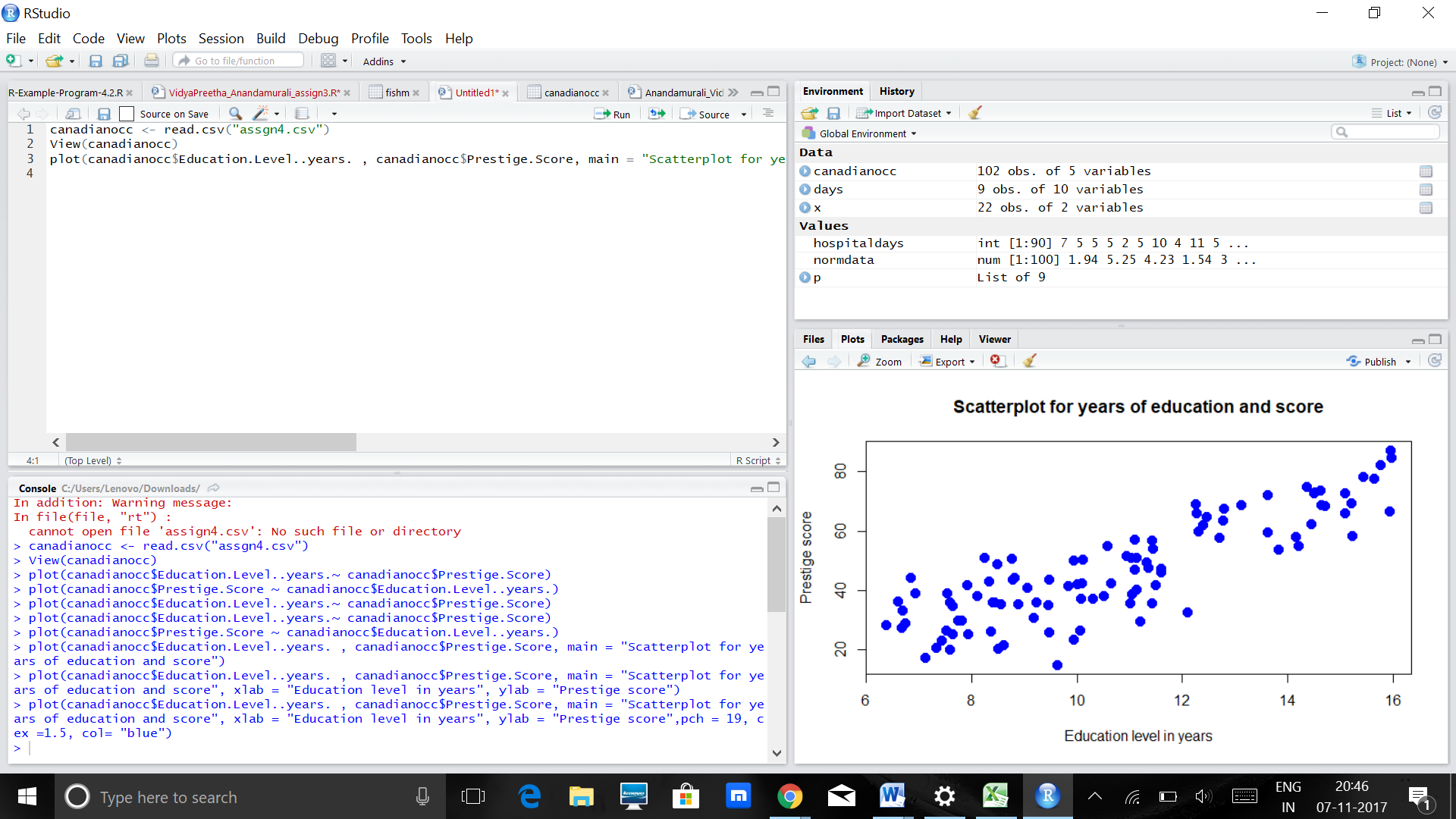
1. Reading the file and setting the directory:

canadianocc <- read.csv("assgn4.csv")

View(canadianocc)

summary(canadianocc)

The summary statistics should be calculated for all the data that is present in the data set so that the initial analysis of the data can be done.

1. Scatterplot for the relationship between the prestige score and the number of educational years put in.  
   

Form: It is evident from the above scatterplot that the relationship is linear. Although the data set does represent a higher spacing in the lower half of the plot, the relationship is linear here.

Direction: The direction of the plot is such that it is positively related to each other, and since they are positively associated, the dependent parameter increases with an increase in value of the independent parameter. In this case, as the number of educational level in years increases, it produces an increase in the prestige score.

Strength: The strength of the plot can be defined by the closeness of the points to each other. The points are present almost near to each other which says that they do have a strong association to each other.  
The correlation of the data: 0.85 – This value is very close to 1 and shows that the dataset is   
highly positively correlated and has a high positive linearity.

|  |
| --- |
| > cor(canadianocc$Education.Level..years. , canadianocc$Prestige.Score)  [1] 0.8501769  Simple linear Regression: > mymodel<- lm(canadianocc$Education.Level..years. ~ canadianocc$Prestige.Score)  > print(mymodel)  Call:  lm(formula = canadianocc$Education.Level..years. ~ canadianocc$Prestige.Score)  Coefficients:  (Intercept) canadianocc$Prestige.Score  4.4236 0.1348 |
|  |
| |  | | --- | |  | |

1. **Residual Plots:**

|  |
| --- |
| > fitted(mymodel)  1 2 3 4 5 6 7 8 9 10 11 12  13.699779 13.740227 12.971703 12.081833 14.333474 14.886272 14.212128 14.953687 14.279543 13.699779 12.782943 12.513285  13 14 15 16 17 18 19 20 21 22 23 24  11.677347 12.809909 14.522234 11.852624 15.519968 12.257110 12.284076 14.239094 15.830074 12.459354 13.335741 16.180629  25 26 27 28 29 30 31 32 33 34 35 36  13.416638 13.645847 13.146981 9.129082 14.144714 13.767193 13.524501 12.135764 12.189696 11.717795 10.625682 10.072884  37 38 39 40 41 42 43 44 45 46 47 48  11.084100 10.126815 10.854891 8.589767 8.832459 9.641432 9.290877 9.439188 9.560534 8.387524 11.313309 9.236945  49 50 51 52 53 54 55 56 57 58 59 60  9.223462 10.018952 9.843675 7.996520 6.419023 7.565068 10.800959 10.773994 11.313309 10.288610 11.380723 8.427972  61 62 63 64 65 66 67 68 69 70 71 72  7.147099 11.825658 7.915623 7.227996 6.756095 7.133616 10.369507 7.322376 9.183014 9.668397 7.821243 9.115599  73 74 75 76 77 78 79 80 81 82 83 84  7.551585 8.913356 8.306626 10.153781 10.382990 9.263911 10.059401 9.263911 10.315576 11.272860 9.439188 8.225729  85 86 87 88 89 90 91 92 93 94 95 96  9.560534 11.205446 8.104383 9.938055 11.191963 11.313309 9.668397 9.304360 8.454938 10.207713 7.996520 13.335741  97 98 99 100 101 102  11.016685 9.263911 7.807760 7.942589 10.113332 9.169531 |
|  |
| > resid(mymodel)  1 2 3 4 5 6 7 8 9 10  -0.58977864 -1.48022728 -0.20170309 -0.66183298 0.28652597 0.75372787 0.87787190 0.48631347 0.24045749 0.94022136  11 12 13 14 15 16 17 18 19 20  -0.39294277 -0.21328515 2.15265344 1.63009147 -0.16223436 2.35737599 0.25003248 1.89288958 2.93592381 0.26090614  21 22 23 24 25 26 27 28 29 30  0.13992623 1.16064637 1.74425913 -0.22062866 2.52336185 1.06415288 -0.68698054 0.32091786 -0.52471370 1.44280695  31 32 33 34 35 36 37 38 39 40  -0.73450120 -1.04576450 0.52030398 -0.27779520 0.96431812 1.41711622 0.23590018 0.51318470 0.50510915 0.58023308  41 42 43 44 45 46 47 48 49 50  3.25754124 1.39856840 -0.07087670 0.63081161 0.94946569 2.81247629 -0.18330879 2.19305482 1.77653770 -0.17895225  51 52 53 54 55 56 57 58 59 60  1.28632519 2.05347983 3.20097685 2.36493201 0.79904068 0.31600644 -0.28330879 -0.81860986 -0.45072319 -0.68797235  61 62 63 64 65 66 67 68 69 70  1.35290130 -1.25565825 1.54437711 0.10200402 0.35390484 0.44638418 -3.52950715 1.27762386 -0.30301366 -2.12839736  71 72 73 74 75 76 77 78 79 80  -0.18124272 -1.47559926 -0.13158511 -2.22335605 -1.56662642 -0.06378106 -1.57299003 -0.86391094 -2.13940089 -0.83391094  81 82 83 84 85 86 87 88 89 90  -1.53557563 -2.51286014 0.85081161 -1.84572914 -1.46053431 -1.10544574 -1.43438322 -0.88805497 -1.26196286 -3.07330879  91 92 93 94 95 96 97 98 99 100  -2.74839736 -2.70435958 -0.64493811 -1.87771258 -0.47652017 -1.06574087 -2.52668541 -1.68391094 0.12224016 0.42741135  101 102  -0.11333242 -0.61953078  Residual Plot between the fitted model and the residual values:  Residual plot with the independent value of the education level in years:  Histogram pertaining to the residuals:  To check if the assumptions are met : a) Linearity – the residual plot generated is linear in nature. b) The observations are independent and have to be taken such that only one value of  The independent data corresponds to one value of the dependent data.  c) The scatter varies almost constantly if we look at it from both the sides of  the graph.  d) From the histogram it is evident that the data is normally distributed.  The points on the extreme ends are outliers and vary greatly from the range of the  Data set provided. As x goes on increasing, the points show an increase in value. This v  Value makes the points closer to the regression line as the x value increases.  The equation for the least-squares regression line for multiple linear regression is an extension of the  equation for simple linear regression: y^=β0^+β1^x1+β2^x2+…+βk^xk  mymodel1 <- lm(canadianocc$Prestige.Score ~ canadianocc$Education.Level..years. +  canadianocc$Income....+canadianocc$Percent.of.Workforce.that.are.Women)  > print(mymodel1)  Call:  lm(formula = canadianocc$Prestige.Score ~ canadianocc$Education.Level..years. +  canadianocc$Income.... + canadianocc$Percent.of.Workforce.that.are.Women)  Coefficients:  (Intercept) canadianocc$Education.Level..years.  -6.794334 4.186637  canadianocc$Income.... canadianocc$Percent.of.Workforce.that.are.Women  0.001314 -0.008905  B0 = -6.794334 + 4.1866 X1 + 0.00131 x2 – 0.008905 x3  Testing with the F test: > anova(mymodel1)  Analysis of Variance Table  Response: canadianocc$Prestige.Score  Df Sum Sq Mean Sq F value Pr(>F)  canadianocc$Education.Level..years. 1 21608.4 21608.4 350.9741 < 2.2e-16 \*\*\*  canadianocc$Income.... 1 2248.1 2248.1 36.5153 2.739e-08 \*\*\*  canadianocc$Percent.of.Workforce.that.are.Women1 5.3 5.3 0.0858 0.7702  Residuals 98 6033.6 61.6  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Anova table to calculate the R squared: Reg SS/ Total SS = 90.55%  This shows that 90.55% of the variability of the data set is shown with the change in  These parameters that are mentioned above. The Pr value as shown above, shows high  Level of significance with respect to the education level in years as well as the  Income of the people. These two factors show high influence on the variability of the  Data.  4.  > summary(mymodel1)  Call:  lm(formula = canadianocc$Prestige.Score ~ canadianocc$Education.Level..years. +  canadianocc$Income.... + canadianocc$Percent.of.Workforce.that.are.Women)  Residuals:  Min 1Q Median 3Q Max  -19.8246 -5.3332 -0.1364 5.1587 17.5045  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -6.7943342 3.2390886 -2.098 0.0385 \*  canadianocc$Education.Level..years. 4.1866373 0.3887013 10.771 < 2e-16 \*\*\*  canadianocc$Income.... 0.0013136 0.0002778 4.729 7.58e-06 \*\*\*  canadianocc$Percent.of.Workforce.that.are.Women -0.0089052 0.0304071 -0.293 0.7702  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 7.846 on 98 degrees of freedom  Multiple R-squared: 0.7982, Adjusted R-squared: 0.792  F-statistic: 129.2 on 3 and 98 DF, p-value: < 2.2e-16  Formal Testing: H0: education level, income and percent of workforce who are women are not predictors of the prestige score.  H1 : atleast one of the slope coefficients are different from 0  F statistic: MS Reg/ MS Res = 350.9741 for education years   = 36.5153 for income  = 0.0858 for the percent of work force   |  | | --- | | > qf(.95, df1=3, df2=98)  [1] 2.697423  From this value of the f statistic it is evident that this value is smaller than  The f value pertaining to the education level in years and the income of the people,  But is greater than the value of the percent of work force and hence, only two factors  Show significance out of the three factors that we did a test on.  Also here, we can boldly reject H0, as the f values of the parameters are bigger than  The F statistic calculated for the model. Therefore, we can say that H1 can be taken  Into consideration.  5.  > confint(mymodel1, level = 0.95)  2.5 % 97.5 %  (Intercept) -1.322220e+01 -0.366468202  canadianocc$Education.Level..years. 3.415272e+00 4.958002277  canadianocc$Income.... 7.623127e-04 0.001864808  canadianocc$Percent.of.Workforce.that.are.Women -6.924697e-02 0.051436660  A plot of the residuals against the fitted values should show no pattern. If a pattern is observed, there may  be "heteroscedasticity" in the errors. That is, the variance of the residuals may not be constant.  The conclusions: The parameters that show high influence to the variation of the prestige score are education level and  income of the work force. These two factors are of great significance.  6. Residual Plot between the fitted model and the residual values:  Outliers do exist and the variation is constant in the graph above. Here, the point  That is close to 3 and -3 have a higher boundary than the other data points.  Although some extreme points do exist in the x direction, they don’t give in too much  Of variation to the data set.  Yes , the model has significant values and is representative of the trends in the data set. The tests made  Above prove that the parameters that have significance will prove to make a change in the value predicted  For the value that depends on them. | |  | | |  | | --- | |  | | |
|  |