hm5stat632

Jessica Grover

4/1/2022

hdi = read.csv("hdi2018.csv")  
head(hdi)

## country hdi\_2018 median\_age pctpop65 pct\_internet pct\_labour  
## 1 Afghanistan 0.496 17.2 0.02688172 13.5 66.0  
## 2 Albania 0.791 34.9 0.13793103 71.8 56.1  
## 3 Algeria 0.759 27.5 0.06398104 59.6 41.3  
## 4 Angola 0.574 16.4 0.02272727 14.3 77.7  
## 5 Argentina 0.830 30.5 0.11036036 74.3 60.5  
## 6 Armenia 0.760 33.8 0.10000000 64.7 58.8

#Problem 1 #a. Fit a multiple linear regression model with hdi 2018 as the response, and the other four variables as predictors.

df<-hdi  
model1<-lm(hdi\_2018 ~median\_age+pctpop65+pct\_internet+pct\_labour,data=df)  
summary(model1)

##   
## Call:  
## lm(formula = hdi\_2018 ~ median\_age + pctpop65 + pct\_internet +   
## pct\_labour, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.194838 -0.034699 0.003272 0.031096 0.122529   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.3374494 0.0319098 10.575 < 2e-16 \*\*\*  
## median\_age 0.0080796 0.0011337 7.127 2.7e-11 \*\*\*  
## pctpop65 -0.0697020 0.1022759 -0.682 0.496   
## pct\_internet 0.0028967 0.0002451 11.817 < 2e-16 \*\*\*  
## pct\_labour -0.0001738 0.0003809 -0.456 0.649   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.05193 on 172 degrees of freedom  
## Multiple R-squared: 0.8882, Adjusted R-squared: 0.8856   
## F-statistic: 341.5 on 4 and 172 DF, p-value: < 2.2e-16

#b. Using the model fit in (a), is there evidence of a relationship between hdi 2018 and at least one of the predictor variables? Write the null and alternative hypotheses, report the F-test statistic and p-value, and state your conclusion. From the above model fit, we can see that the p-value of hdi\_2018 is same for pct\_internet predictor variable. H0 : β1 = β2 = β3 = β4 = 0 HA : at least one βj != 0 We are rejecting predictor variable pct\_labor because its p-value is 0.649 which is too high and is not significant.

model2<-lm(hdi\_2018 ~median\_age+pctpop65+pct\_internet,data=df)  
summary(model2)

##   
## Call:  
## lm(formula = hdi\_2018 ~ median\_age + pctpop65 + pct\_internet,   
## data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.194667 -0.033487 0.003334 0.031132 0.123101   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.3258064 0.0191190 17.041 < 2e-16 \*\*\*  
## median\_age 0.0081012 0.0011301 7.169 2.11e-11 \*\*\*  
## pctpop65 -0.0670332 0.1018747 -0.658 0.511   
## pct\_internet 0.0028964 0.0002446 11.842 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.05181 on 173 degrees of freedom  
## Multiple R-squared: 0.888, Adjusted R-squared: 0.8861   
## F-statistic: 457.4 on 3 and 173 DF, p-value: < 2.2e-16

#c. Using the model fit in (a), which predictor variables are statistically significant according to the individual t-tests?

median\_age and pct\_internet are statiscally significant to the individual t-test.

lm3<-lm(hdi\_2018 ~median\_age+pct\_internet,data=df)  
summary(lm3)

##   
## Call:  
## lm(formula = hdi\_2018 ~ median\_age + pct\_internet, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.191236 -0.034675 0.002006 0.030777 0.126611   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.3341527 0.0142820 23.397 <2e-16 \*\*\*  
## median\_age 0.0075581 0.0007706 9.807 <2e-16 \*\*\*  
## pct\_internet 0.0029287 0.0002392 12.244 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.05172 on 174 degrees of freedom  
## Multiple R-squared: 0.8878, Adjusted R-squared: 0.8865   
## F-statistic: 688.1 on 2 and 174 DF, p-value: < 2.2e-16

#d. Fit a reduced model with median age and pct internet as predictors. Use the anova() function to conduct a partial F-test that compares this reduced model with the full model specified in (a). Make sure to write the null and alternative hypotheses, report the p-value, and state your conclusion.

Reduced model is:

model3<-lm(hdi\_2018 ~median\_age+pct\_internet,data=df)  
summary(model3)

##   
## Call:  
## lm(formula = hdi\_2018 ~ median\_age + pct\_internet, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.191236 -0.034675 0.002006 0.030777 0.126611   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.3341527 0.0142820 23.397 <2e-16 \*\*\*  
## median\_age 0.0075581 0.0007706 9.807 <2e-16 \*\*\*  
## pct\_internet 0.0029287 0.0002392 12.244 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.05172 on 174 degrees of freedom  
## Multiple R-squared: 0.8878, Adjusted R-squared: 0.8865   
## F-statistic: 688.1 on 2 and 174 DF, p-value: < 2.2e-16

Hypothesis test: H0 : β1 = β2 = β3 = β4 = 0 HA : at least one βj != 0

anova(model1,model3)

## Analysis of Variance Table  
##   
## Model 1: hdi\_2018 ~ median\_age + pctpop65 + pct\_internet + pct\_labour  
## Model 2: hdi\_2018 ~ median\_age + pct\_internet  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 172 0.46380   
## 2 174 0.46552 -2 -0.0017236 0.3196 0.7269

P-value is 2.2e-16 which is small so we are rejecting null hypothesis.

#e. According to the adjusted-R2, how does the full model in (a) compare with the reduced model in (d)? Is this consistent with your conclusion for the partial F-test?

summary(model1)$adj.r.squared

## [1] 0.8855708

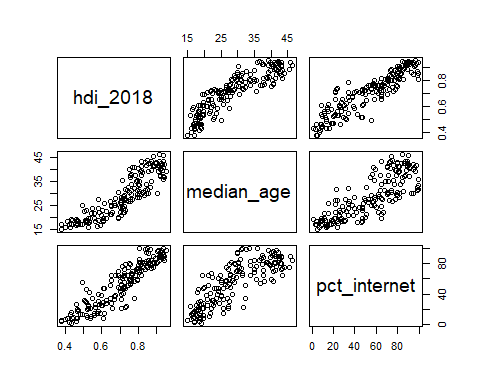
summary(model3)$adj.r.squared

## [1] 0.8864657

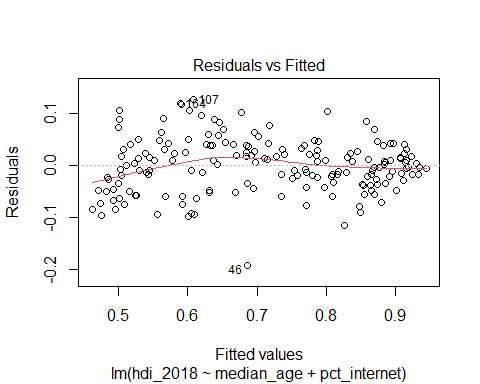
The adjusted R squared of both reduced model and full model is about the same but the reduced model is little higher. This agrees with the conclusion of F-test that the decision of removing pct\_labour is correct.

#Problem 2 #a. Make a scatterplot matrix for the three variables. Describe the associations between the variables in the scatterplot matrix.

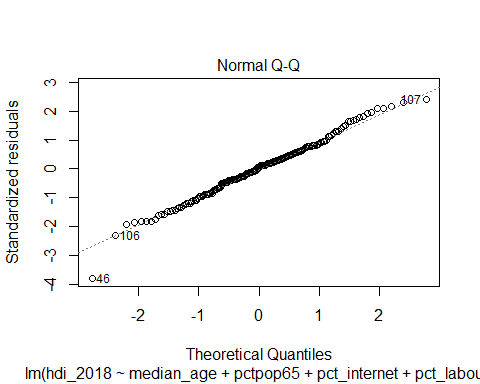
pairs(hdi\_2018 ~ median\_age+ pct\_internet , data=hdi)

 #b. Make a plot of the residuals versus fitted values, and a QQ plot of the standardized residuals. Residual vs fitted plot

plot(model3,which = 1)

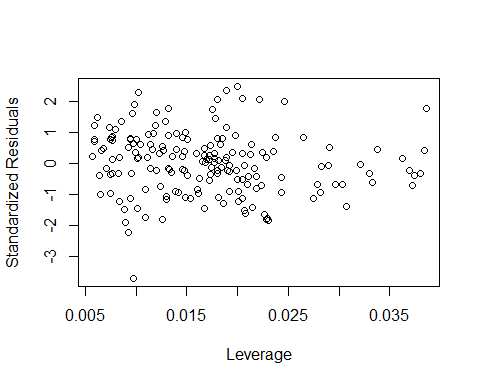
 Standardized residual plot

plot(model1,which = 2)



#c. Make a plot with the leverage values (hi) on the x-axis, and standardized residuals (ri) on the y-axis. Identify any points (countries) that have high standardized residuals or leverage.

p <- 3  
 n <- nrow(hdi)  
plot(hatvalues(model3), rstandard(model3),  
xlab='Leverage', ylab='Standardized Residuals')  
 abline(v = 2\*(p+1)/n, lty=2)

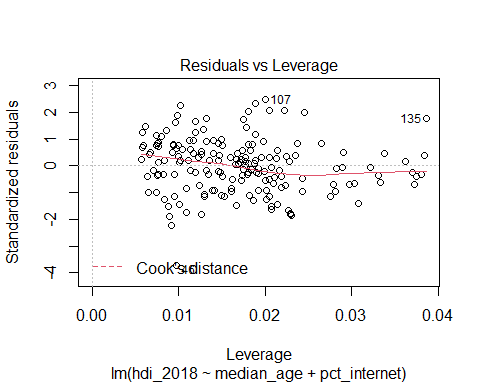
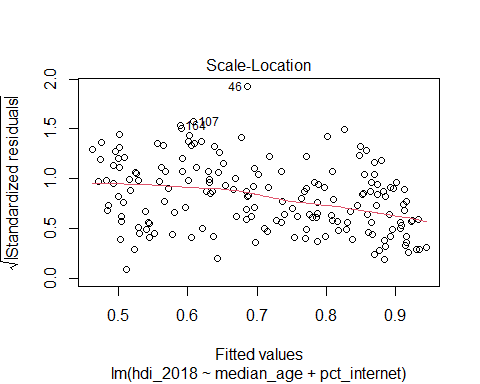
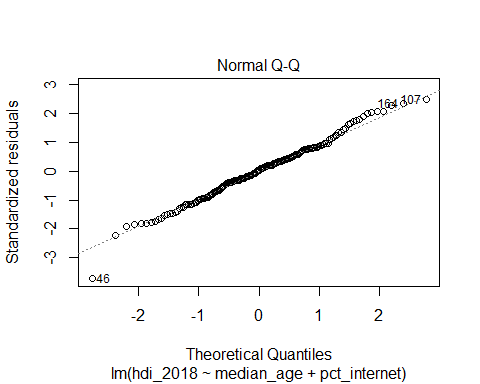
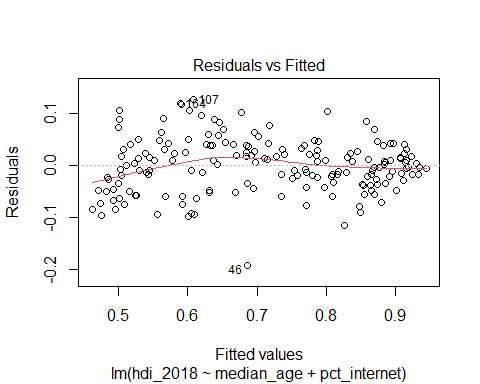


ind <- which(hatvalues(model3) > 4/n)  
hdi[ind, ]

## country hdi\_2018 median\_age pctpop65  
## 11 Bahrain 0.838 31.2 0.00000000  
## 12 Bangladesh 0.614 25.7 0.05142503  
## 23 Brunei Darussalam 0.845 29.9 0.00000000  
## 24 Bulgaria 0.816 43.4 0.21126761  
## 26 Burundi 0.423 17.1 0.02678571  
## 31 Central African Republic 0.381 17.1 0.02127660  
## 40 Croatia 0.837 43.1 0.19047619  
## 41 Cuba 0.778 40.8 0.15044248  
## 44 Cote d'Ivoire 0.516 18.3 0.02788845  
## 52 Eritrea 0.434 19.2 0.05714286  
## 62 Germany 0.939 45.9 0.21419976  
## 64 Greece 0.872 43.4 0.21904762  
## 65 Guatemala 0.651 21.3 0.04651163  
## 67 Guinea-Bissau 0.461 18.4 0.05263158  
## 73 Iceland 0.938 36.0 0.33333333  
## 80 Italy 0.883 45.4 0.22772277  
## 82 Japan 0.915 46.4 0.27594340  
## 83 Jordan 0.723 22.1 0.04000000  
## 87 Kuwait 0.808 33.5 0.02439024  
## 99 Malaysia 0.804 28.2 0.06666667  
## 116 Niger 0.377 14.9 0.02678571  
## 122 Palestine, State of 0.690 19.5 0.04081633  
## 129 Portugal 0.850 43.9 0.22330097  
## 130 Qatar 0.848 31.5 0.00000000  
## 135 Saint Vincent and the Grenadines 0.728 31.6 0.00000000  
## 138 Saudi Arabia 0.857 30.0 0.03264095  
## 139 Senegal 0.514 18.1 0.03144654  
## 144 Slovenia 0.902 43.0 0.19047619  
## 149 Sri Lanka 0.780 32.3 0.10377358  
## 166 Ukraine 0.750 40.0 0.16515837  
## 167 United Arab Emirates 0.866 32.6 0.01041667  
## pct\_internet pct\_labour  
## 11 98.6 72.8  
## 12 15.0 58.7  
## 23 94.6 65.2  
## 24 64.8 55.4  
## 26 2.7 79.0  
## 31 4.3 72.1  
## 40 72.7 51.7  
## 41 57.1 53.6  
## 44 46.8 57.3  
## 52 1.3 80.6  
## 62 89.7 60.6  
## 64 73.0 52.8  
## 65 65.0 62.3  
## 67 3.9 73.0  
## 73 99.0 76.3  
## 80 74.4 48.9  
## 82 84.6 60.7  
## 83 66.8 39.3  
## 87 99.6 73.9  
## 99 81.2 64.6  
## 116 5.3 78.7  
## 122 64.4 45.4  
## 129 74.7 58.7  
## 130 99.7 86.9  
## 135 22.4 68.3  
## 138 93.3 56.5  
## 139 46.0 46.4  
## 144 79.8 58.0  
## 149 34.1 52.5  
## 166 58.9 54.0  
## 167 98.5 83.0

#d. Based on the scatterplot matrix and model diagnostics, do the assumptions for MLR appear adequately satisfied? Can you think of any ways in which the model might be improved to better fit the data?

plot(model3)

 From the above plots, all the assumptions of linearty are satisfied.