Homework 5, STAT 632

Chandnee das

3/21/2022

```
hdi<-read.csv('hdi2018.csv')

# To fit multiple linear regression
lm_full<-lm (hdi_2018 ~ median_age + pctpop65 + pct_internet + pct_labour ,
data=hdi)
summary(lm_full)

##
## Call:
## lm(formula = hdi_2018 ~ median_age + pctpop65 + pct_internet +
##     pct_labour, data = hdi)
##
## 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) Since the p-value < 0.001, we can reject null hypothesis and we conclude that at least one predictor has a relationship between hdi_2018.

Null hypothesis: H0 : /beta_2 = /beta_4 = 0. Alternative hypothesis: HA :/beta_1 != or /beta_2 != 0 or /beta_3 !=0 or /beta_4!=0

(c)For predictor variable 1 and 3 , p value is < .05, so we can reject null hypothesis.Thus , we conclude that predictor variables median_age and PCI_internet are statistically significant according to individual T test.

(d)

```
lm_fulll<-lm (hdi_2018 ~ median_age + pctpop65 + pct_internet + pct_labour ,
data=hdi)
lm_red<-lm(hdi_2018 ~ median_age + pct_internet , data =hdi)
anova(lm_red,lm_full)
```

```
## Analysis of Variance Table
##
## Model 1: hdi_2018 ~ median_age + pct_internet
## Model 2: hdi_2018 ~ median_age + pctpop65 + pct_internet + pct_labour
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1    174 0.46552
## 2    172 0.46380  2 0.0017236 0.3196 0.7269
```

The p-value = 0.73 is large, so we do not reject the null hypothesis that

Null hypothesis: H0 : beta_2 = beta_4 = 0. So we can remove both predictors, pctpop65 and pct_labour from the model. Alternative hypothesis : HA: beta_1 != 0 , beta_3!=0.

(e)

```
# The R2 for the full and reduced models
s1<-summary(lm_full)
s2<-summary(lm_red)
s1$r.squared
```

```
## [1] 0.8881714
```

```
s2$r.squared
```

```
## [1] 0.8877558
```

```
s1$adj.r.squared
```

```
## [1] 0.8855708
```

```
s2$adj.r.squared
```

```
## [1] 0.8864657
```

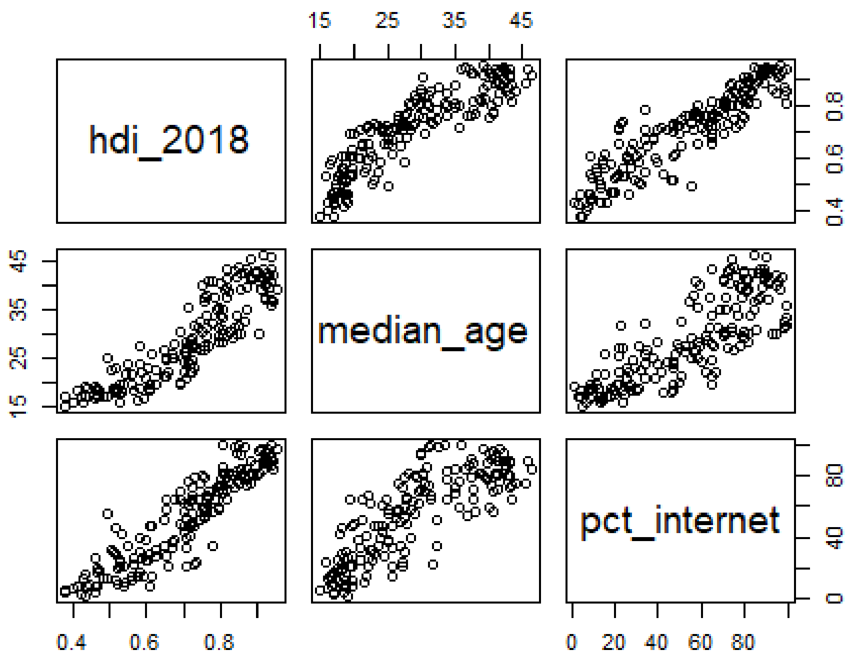
The R2 for the full and reduced models are about the same, R2 for full model is little bit higher than reduced model and theadjusted R2 for the reduced model is a little higher.

This agrees with the conclusion of the F-test. So the adjusted-R2 also indicates that we can remove pctpop65 and pct_labour.

Exercise 2

(a)

```
pairs(hdi_2018 ~ median_age + pct_internet, data=hdi)
```



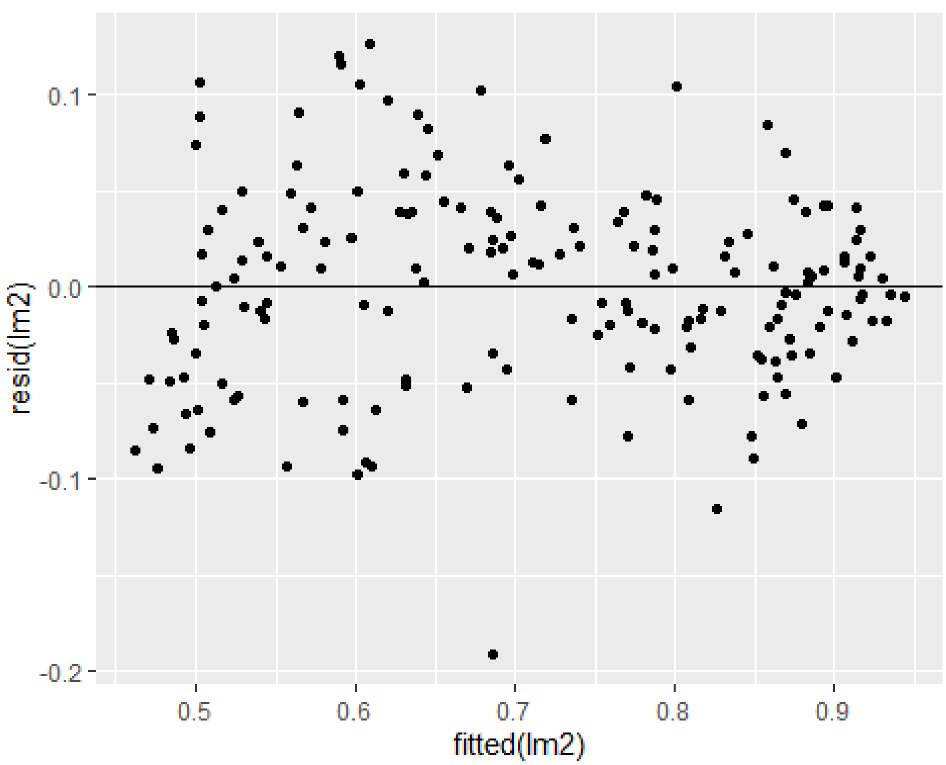
There is a linear relationship between variables in the scatterplot matrix. Also there seems to be some colinearity between predictor variables median_age and pct_internet.

```
lm2<-lm(hdi_2018~median_age+pct_internet, data=hdi)
summary(lm2)

##
## Call:
## lm(formula = hdi_2018 ~ median_age + pct_internet, data = hdi)
##
## Residuals:
```

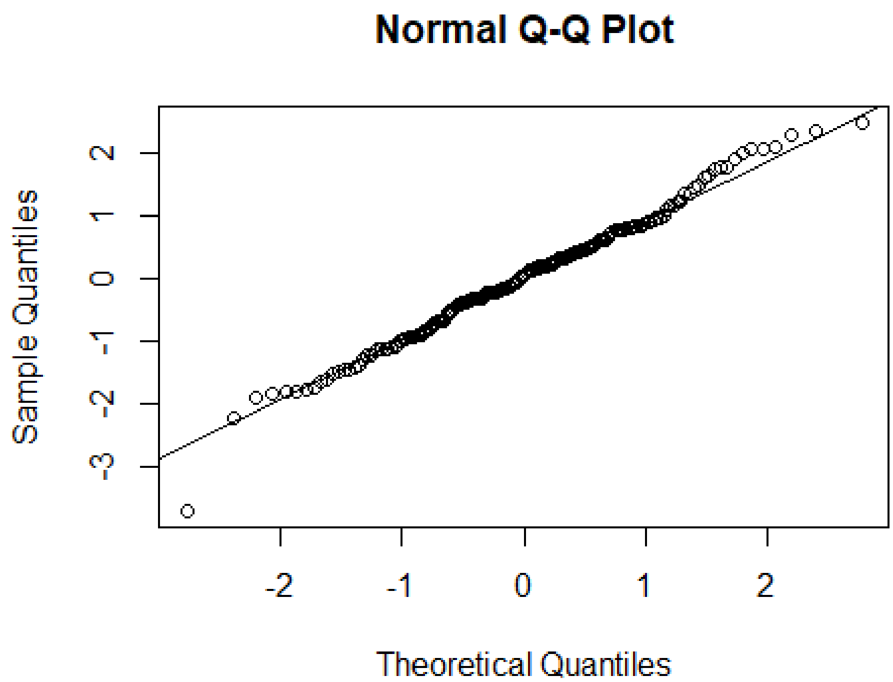
```
##        Min         1Q     Median         3Q        Max
## -0.191236 -0.034675  0.002006  0.030777  0.126611
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.3341527  0.0142820  23.397   <2e-16 ***
## median_age   0.0075581  0.0007706   9.807   <2e-16 ***
## 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
```

(b)

```
library(ggplot2)
ggplot(lm2, aes(fitted(lm2), resid(lm2) )) +
  geom_point() +
  geom_hline(yintercept=0)
```



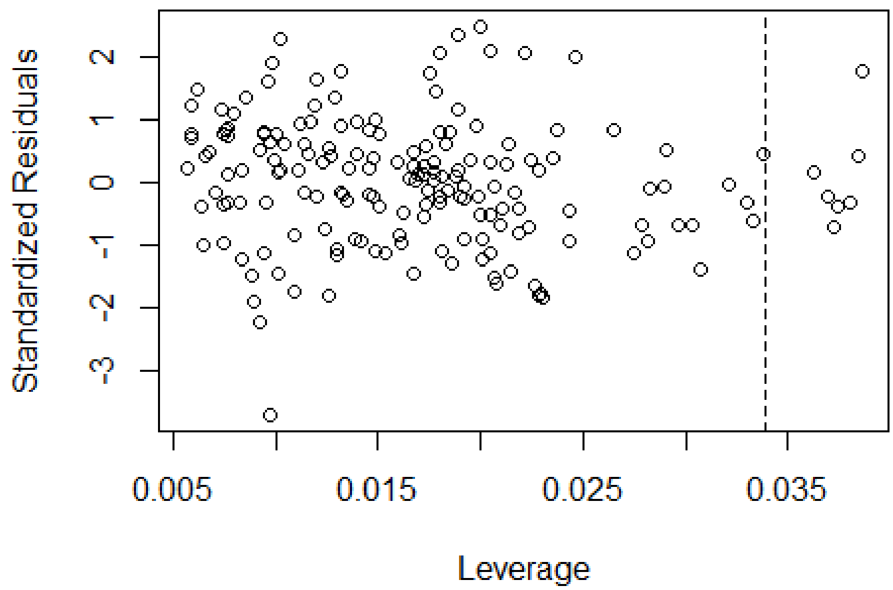

```
qqnorm(rstandard(lm2))
qqline(rstandard(lm2))
```



From residual vs fitted plot shows like fan pattern and qq plot indicates approximately normal with some deviation from normality near the tail.

(c)

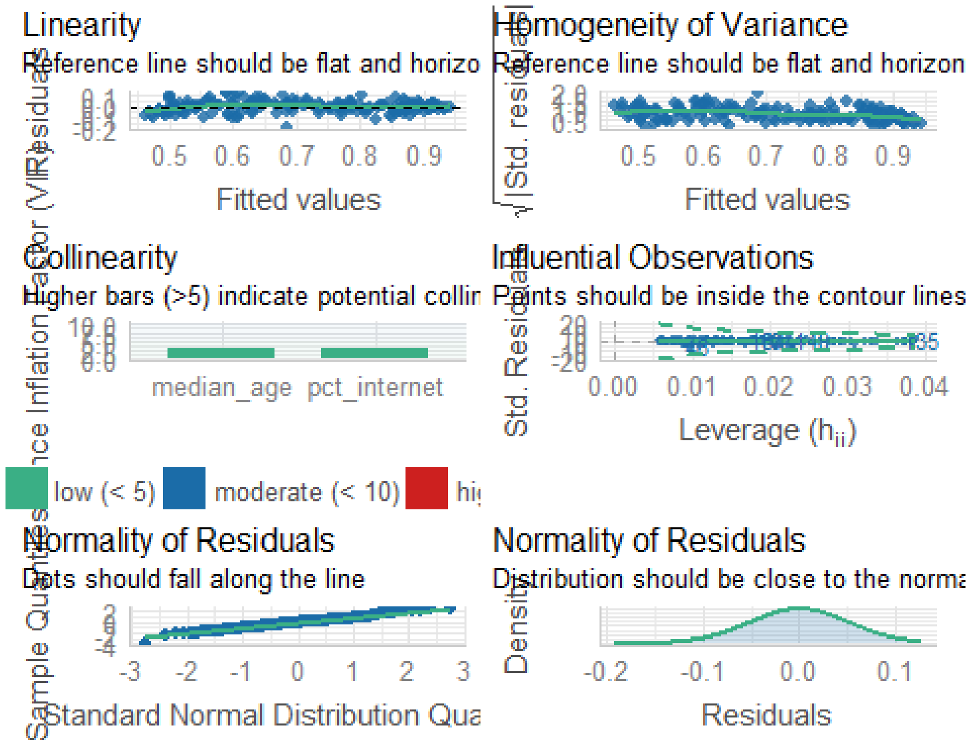
```
plot(hatvalues(lm2), rstandard(lm2),xlab='Leverage', ylab='Standardized
Residuals')
p <- 2
n <- nrow(hdi)
abline(v = 2*(p+1)/n, lty=2)
```




```
ind <- which(hatvalues(lm2) > 0.1)
hdi[ind, ]

## [1] country      hdi_2018     median_age   pctpop65     pct_internet
## [6] pct_labour
## <0 rows> (or 0-length row.names)
```

Nigeria and United Kingdom countries has high leverage point.

```
library('performance')

## Warning: package 'performance' was built under R version 4.1.3

library('see')

## Warning: package 'see' was built under R version 4.1.3

library('patchwork')

## Warning: package 'patchwork' was built under R version 4.1.3

performance::check_model(lm2)
```



(d)

From scatter plot matrix and model diagnostic,we can say that the assumptions of multiple linear regression apparently satisfied.To better fit the model, we can remove outliers and we can do transformation to make the variance constant in the plot.