STAT 632 HW 5

Brandon Keck

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

Exercise 1

(a)

```
full.model <- lm(hdi_2018 ~ median_age + pctpop65 + pct_internet + pct_labour, data = hdi)
summary(full.model)
```

```
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)

Based on the summary statistics there is evidence of a relationship between hdi_2018 and some of the predictor variables. Those include; median_age, and pct_internet.

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

vs

H_A : at least one $\beta_j \neq 0$

The F-statistic reported is 341.5 with a p-value of 2.2e-16 which is essentially 0. We can conclude that at least one β means is significantly different from the others.

(c)

Only median_age and pct_internet are statistically significant at the 0.05 significance level.

(d)

```r
reduced.model <- lm(hdi_2018 ~ median_age + pct_internet, data = hdi)
anova(reduced.model, full.model)
```

```
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
```

$H_0 : \beta_{median-age} = \beta_{pct-internet}$

vs

H_A : At least one of $\beta_{median-age}, \beta_{pct-internet} \neq 0$

With a p-value of 0.7269, we fail to reject the H_0 and can conclude that there is no significant evidence that pctpop65 and pct_labour improve the model beyond what is already explained by median_age and pct_internet.

(e)

```
s1 <- summary(full.model)
s2 <- summary(reduced.model)

s1$adj.r.squared
```

```
[1] 0.8855708
```

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

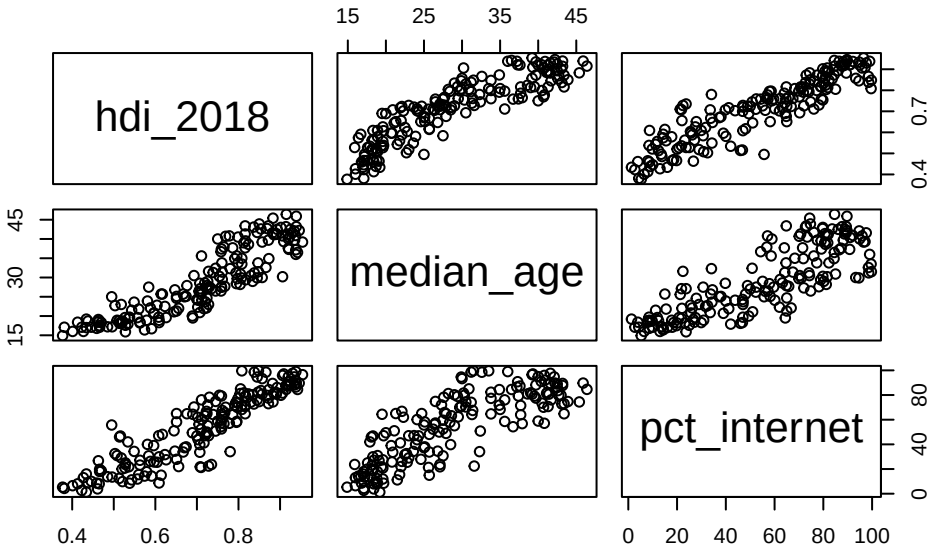
```
[1] 0.8864657
```

According to the adjusted R^2 the reduced model is slightly better than the full model. This supports the result of the partial F-test from part (d), where we found no significant improvement when adding pctpop65 and pct_labour. Therefore, the reduced model is preferable due to model performance.

Exercise 2

(a)

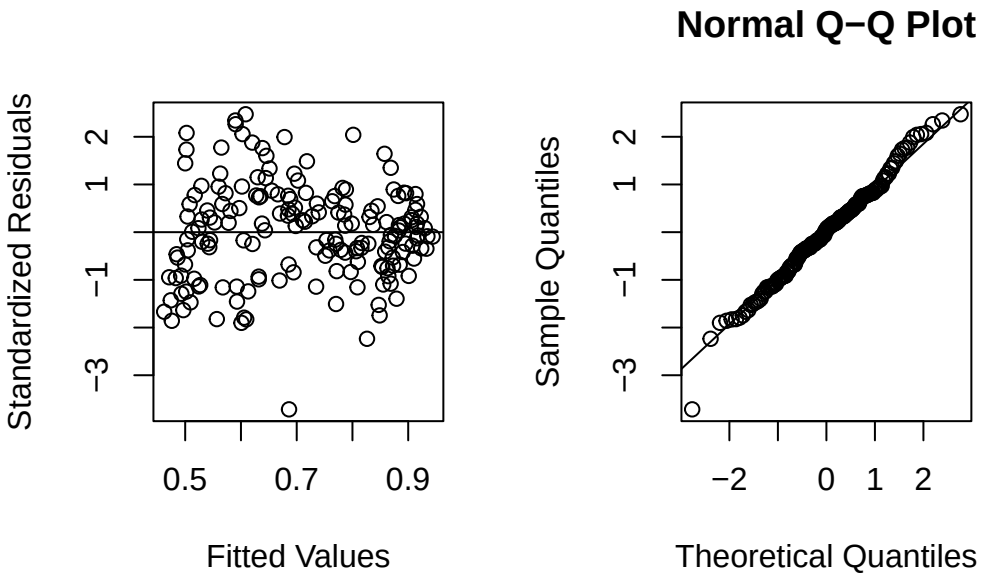
```
pairs(hdi_2018 ~ median_age + pct_internet, data = hdi)
```



The scatterplot matrix shows a strong positive linear relationship between both predictor variables median_age, pct_internet with hdi_2018. Median age and internet usage themselves are positively related, suggesting some correlation between the predictors. These linear trends suggest that a multiple linear regression model may be appropriate for this dataset.

(b)

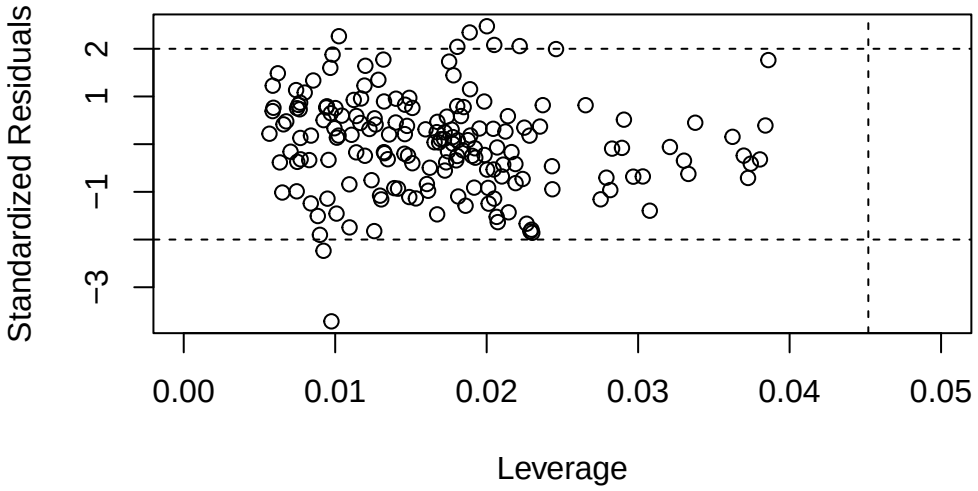
```
par(mfrow = c(1,2))
plot(predict(reduced.model), rstandard(reduced.model),
     xlab = "Fitted Values",
     ylab = "Standardized Residuals")
abline(h = 0)
qqnorm(rstandard(reduced.model))
qqline(rstandard(reduced.model))
```



The residuals vs fitted values plot shows no strong pattern, suggesting the assumptions of linearity and approximately constant variance are met. The QQ plot indicates the residuals are approximately normally distributed, with only slight deviation in the tails. Overall, the model assumptions for multiple linear regression appear adequately satisfied.

(c)

```
p <- 3
n <- nrow(hdi)
plot(hatvalues(reduced.model), rstandard(reduced.model),
     xlab = "Leverage",
     ylab = "Standardized Residuals",
     xlim = c(0, 0.05))
abline(v = 2*(p+1)/n, lty = 2)
abline(h = c(-2,2), lty = 2)
```




```
outlier <- which(abs(rstandard(reduced.model)) > 2)
leverage <- which(hatvalues(reduced.model) > 2*(p+1)/n)
hdi[outlier,]
```

```
                 country hdi_2018 median_age   pctpop65 pct_internet
37                 Congo    0.609       18.9 0.01923077          8.7
46              Djibouti    0.495       25.0 0.00000000         55.7
79                Israel    0.906       30.2 0.11904762         81.6
94                 Libya    0.708       27.1 0.04477612         21.8
106 Moldova (Republic of)    0.711       35.6 0.12195122         76.1
107              Mongolia    0.735       27.1 0.03125000         23.7
136                 Samoa    0.707       20.9 0.00000000         33.6
164          Turkmenistan    0.710       25.6 0.05084746         21.3
    pct_labour
37        69.2
46        63.0
79        64.1
94        52.4
106       42.1
107       59.9
136       31.4
164       65.1
```

```
hdi[leverage,]
```

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

(d)

Based on the scatterplot matrix and diagnostic plots, the assumptions of multiple linear regression appear to be reasonably satisfied. The relationships between predictors and HDI are lienar, the residuals are approximately normally distributed and there is no evidence of strong heteroscedisticity. Additionally, there are no high-leverage or influential observations, although a few data points exhibit large standardized residuals and may warrant closer inspection overall, the model fits well. In order to improve the model we might consider introducing some interactinos terms to determine how that might affect the adjusted R^2.