STAT632 Hmwk 5

Winnie Lu

2022-03-21

```
hdi <- read.csv("hdi2018.csv")
head(hdi)
```

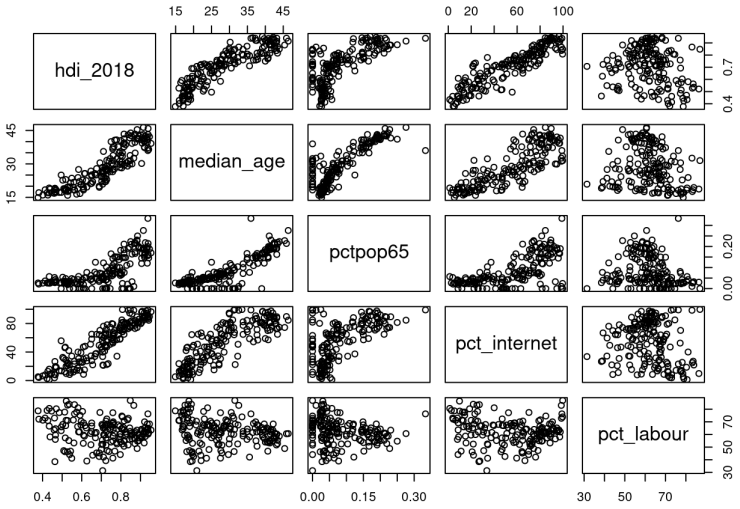
```
##       country hdi_2018 median_age   pctpop65 pct_internet pct_labour
## 1 Afghanistan    0.496       17.2 0.02688172         13.5       66.0
## 2     Albania    0.791       34.9 0.13793103         71.8       56.1
## 3     Algeria    0.759       27.5 0.06398104         59.6       41.3
## 4      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
```

Exercise 1

a

- full model: $\widehat{HDI2018} = 0.3374 + 0.008$ median age -0.0697 pctpop65 $+0.003$ pct_internet -0.0002 pct_labour
- null model: $\widehat{HDI2018} = 0.7113 + e$

```
pairs(hdi_2018 ~ median_age + pctpop65 + pct_internet + pct_labour, data=hdi)
```




```
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

- $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
  - There is no relationship between the response (HDI) and predictor variables
- H_1 at least one $\beta_j \neq 0$
  - There is a relationship between the response (HDI) and at least one of the predictor variables
- According to F-statistic above (F-stat: 341.5 on 4 and 172 DF), and the p-value, (p-value<2.2e-16) < a=.05, we reject the H_0 assumption; therefore, at least one of the predictor variables are significant to the model.

c

- According the fit in part a, the predictor variables that are significant based on individual t-test are: median_age and pct_internet given that their p-values are < $\alpha = 0.05$.

d

To test if our predictors, pctpop65 and pct_labour, are significant to the model, we test these subset of predictors via the partial F-test. ($\beta_2 =$ pctpop65, $\beta_4 =$pct_labour) - $H_0 : \beta_2 = \beta_4 = 0$ - $H_1 : \beta_2 \neq 0$ or $\beta_4 \neq 0$ - Given that the p-value = 0.7269 is large, so we fail to reject the H_0, so we can drop both predictors, pctpop65 and pct_labour, from our model.

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

e

- The full model adjusted $R^2 = 0.8855708$, while the reduced model adjusted $R^2 = 0.8864657$. The adjusted R^2 agrees with our partial F-test result since the variability of the reduced model is higher than of our full model. When we run the reduced model again after dropping the previous two predictors, we can see that the remaining predictors, median_age and pct_internet, are highly significant given their individual t-test p-value values are 2e-16. Therefore, there is no motivation for further simplifying our current reduced model.

```
summary(lm_full)$adj.r.squared
```

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

```
summary(lm_reduced)$adj.r.squared
```

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

```
summary(lm_reduced)
```

```
##
## 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
```

Exercise 2

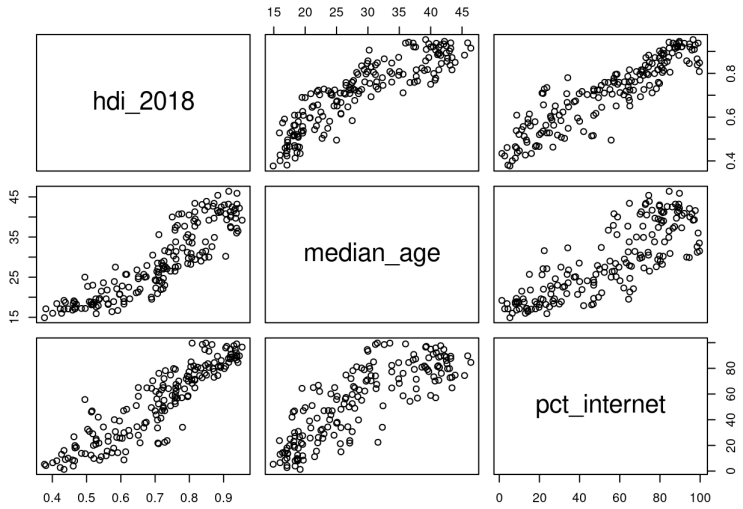
a

- Both predictors, median_age and pct_internet, have a positive, moderately linear relationship with the response variable, hdi_2018. There is also a slight correlation between the predictors, median_age and pct_internet, indicating possible co linearity.

```
lm1 <-lm(hdi_2018 ~ median_age + pct_internet, data=hdi)
summary(lm1)
```

```
##
## 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
```

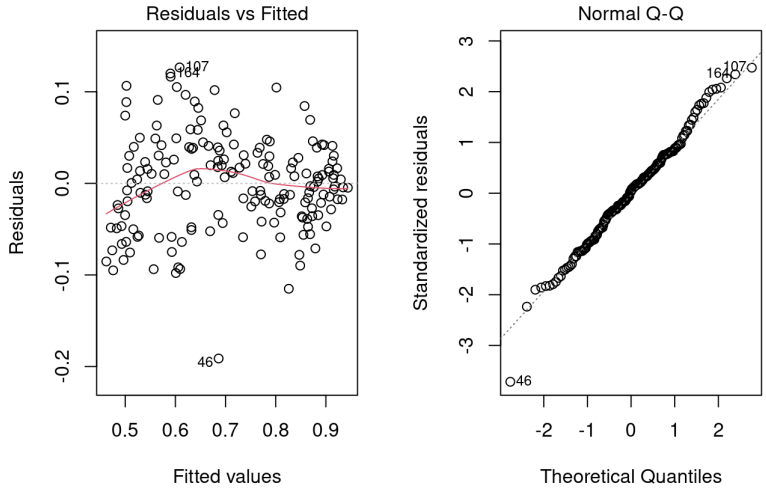
```
pairs(hdi_2018 ~ median_age + pct_internet, data=hdi)
```



b

- From the std residuals vs fitted plot, there appears to be no discernible pattern and the points are randomly scattered around 0. In addition, a majority of the data points in our QQ plot fall onto the line. Notably, there are a few points on both the left and right tail ends that deviate from the line slightly, indicating a slight deviation from the normality assumption. But otherwise, our data upholds the assumption of normality and equal variance.

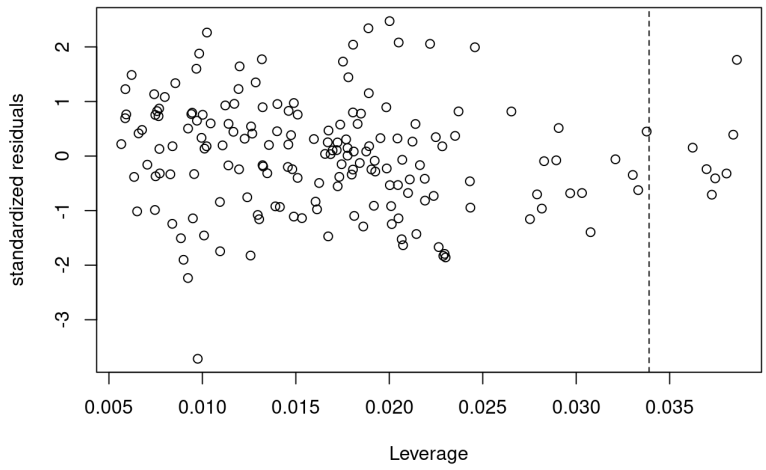
```
par(mfrow=c(1,2))
plot(lm1, 1:2)
```



c

- The countries with high leverage points are Bahrain, Brunei Darussalam, Bulgaria, Italy, State of Palestine, Qatar, Saint Vincent and the Grenadines.

```
p <-2
n <-nrow(hdi)
plot(hatvalues(lm1), rstandard(lm1), xlab="Leverage", ylab="standardized residuals")
abline(v = 2*(p+1)/n, lty=2)
```




```
ind <-which(hatvalues(lm1) > .035)
hdi[ind,]
```

```
##                             country hdi_2018 median_age   pctpop65
## 11                          Bahrain    0.838       31.2 0.00000000
## 23                Brunei Darussalam    0.845       29.9 0.00000000
## 24                         Bulgaria    0.816       43.4 0.21126761
## 80                            Italy    0.883       45.4 0.22772277
## 122              Palestine, State of    0.690       19.5 0.04081633
## 130                           Qatar    0.848       31.5 0.00000000
## 135 Saint Vincent and the Grenadines    0.728       31.6 0.00000000
##     pct_internet pct_labour
## 11          98.6       72.8
## 23          94.6       65.2
## 24          64.8       55.4
## 80          74.4       48.9
## 122         64.4       45.4
## 130         99.7       86.9
## 135         22.4       68.3
```

d

- Using the box-cox method, the estimated value of the parameter is $\hat{\lambda} = 1.8$. We can round and use Y^2 to transform our response. Thus, the regression model with the transformed response would be $Y^2 = -0.002 + 0.011\text{median\_age} + 0.004\text{pct\_internet}$. With the transformation, the F-stat allows us to conclude that at least one of the predictors are significant to our model. In the individual t-test, we see that all predictors are significant to our model. Our $R^2value = 0.8984$, and our $R^2_{adj} = 0.8973$, so compared to the original model($R^2 = 0.8877558$, $R^2_{adj} = 0.8864657$) , the transformed model is slightly better.
- To verify this, we run the diagnostics again and see that it definitely looks slightly better compared to the first model: both equal variance and normality assumptions are held.

```
library(car)
```

```
## Loading required package: carData
```

```
summary(powerTransform(lm1))
```

```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1    1.8521           2       1.4749       2.2294
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                             LRT df       pval
## LR test, lambda = (0) 86.74687  1 < 2.22e-16
##
## Likelihood ratio test that no transformation is needed
##                            LRT df      pval
## LR test, lambda = (1) 19.6975  1 9.072e-06
```

```
lm2 <-lm((hdi_2018)^2 ~ median_age + pct_internet, data=hdi)
summary(lm2)
```

```
##
## Call:
## lm(formula = (hdi_2018)^2 ~ median_age + pct_internet, data = hdi)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.246775 -0.043990  0.000893  0.039620  0.170744
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.0024412  0.0187635   -0.13    0.897
## median_age    0.0111327  0.0010125   11.00   <2e-16 ***
## pct_internet  0.0038765  0.0003143   12.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06796 on 174 degrees of freedom
## Multiple R-squared:  0.8984, Adjusted R-squared:  0.8973
## F-statistic: 769.6 on 2 and 174 DF,  p-value: < 2.2e-16
```

```
summary(lm1)$r.squared
```

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

```
summary(lm1)$adj.r.squared
```

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

```
par(mfrow=c(1,2))
plot(lm2, 1:2)
```