DATA605 - Assignment 12

Nick Oliver

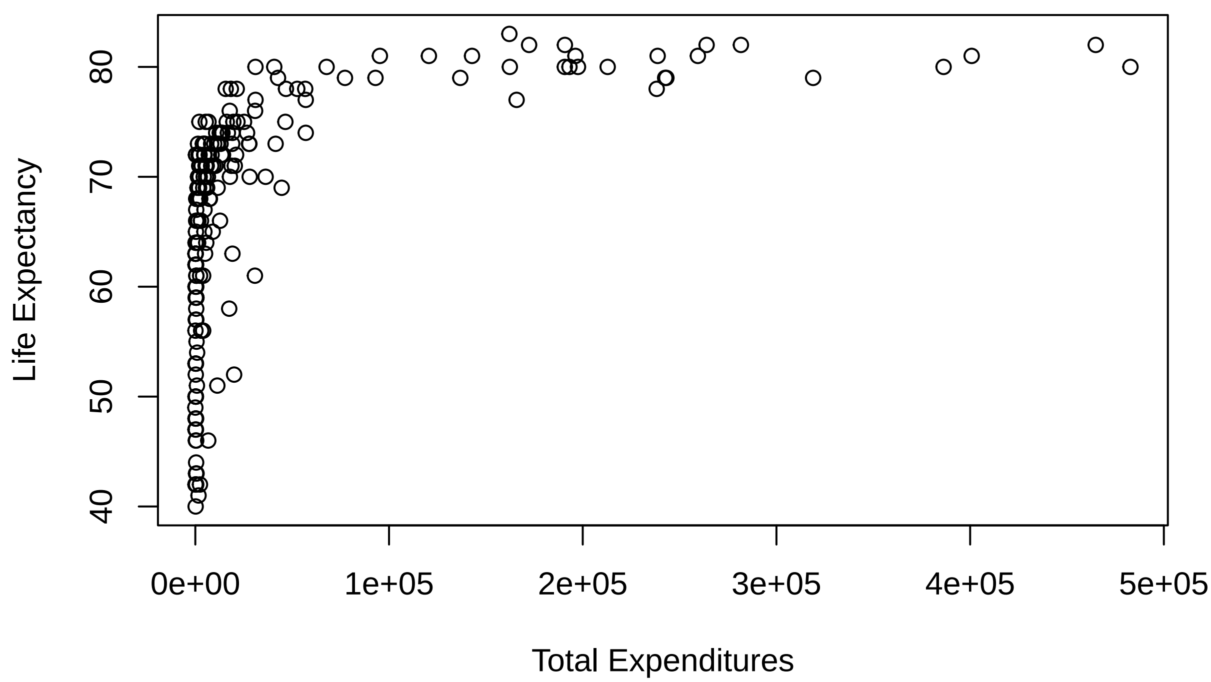
Assignment 12

The attached who.csv dataset contains real-world data from 2008. The variables included follow. Country: name of the country LifeExp: average life expectancy for the country in years InfantSurvival: proportion of those surviving to one year or more Under5Survival: proportion of those surviving to five years or more TBFree: proportion of the population without TB. PropMD: proportion of the population who are MDs PropRN: proportion of the population who are RNs PersExp: mean personal expenditures on healthcare in US dollars at average exchange rate GovtExp: mean government expenditures per capita on healthcare, US dollars at average exchange rate TotExp: sum of personal and government expenditures.

1.

Provide a scatterplot of LifeExp~TotExp, and run simple linear regression. Do not transform the variables. Provide and interpret the F statistics, R^2, standard error,and p-values only. Discuss whether the assumptions of simple linear regression met.

```
df <- read.csv('https://raw.githubusercontent.com/nolivercuny/DATA605/main/week%2012/who.csv')
plot(df$LifeExp ~ df$TotExp,
xlab="Total Expenditures", ylab="Life Expectancy")
```




```
model <- lm(data=df, LifeExp ~ TotExp)
summary(model)
```

```
##
## Call:
## lm(formula = LifeExp ~ TotExp, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.764  -4.778   3.154   7.116  13.292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.475e+01  7.535e-01  85.933  < 2e-16 ***
## TotExp      6.297e-05  7.795e-06   8.079 7.71e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.371 on 188 degrees of freedom
## Multiple R-squared:  0.2577, Adjusted R-squared:  0.2537
## F-statistic: 65.26 on 1 and 188 DF,  p-value: 7.714e-14
```

F-statistic:65.26 is meaningless for single parameter models

R^2:0.2577 indicates a relatively poorly fitting model as it is closer to 0 than 1

Standard Error:7.795e-06 Standard error 6.297e-05 /7.795e-06 = 8.078255. The large ratio indicates little variability in the slope estimate

P-value:7.714e-14 much smaller than 0.05 which indicates statistical significance

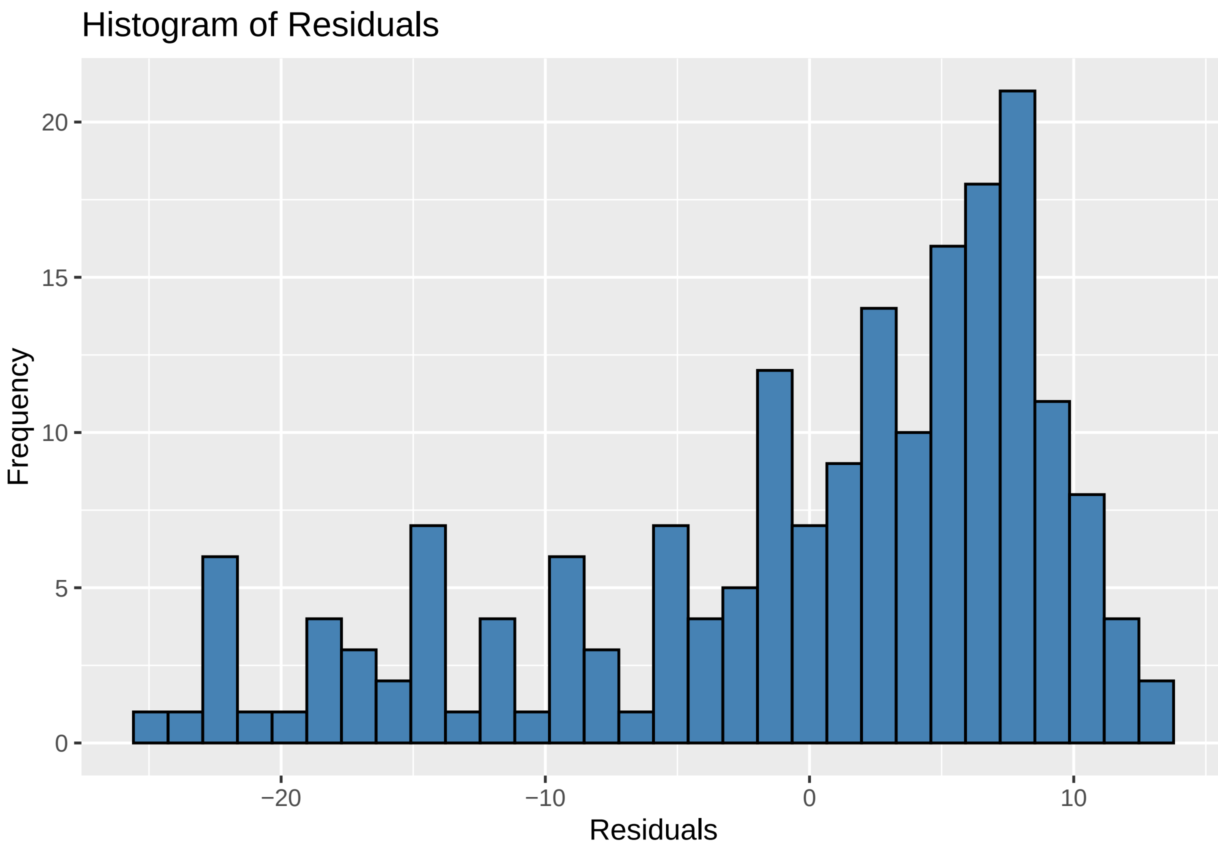
Are the assumptions of a simple linear regression met?

(1) linearity:No model is not linear

(2) nearly normal residuals: Yes residuals look nearly normal based on histogram plot

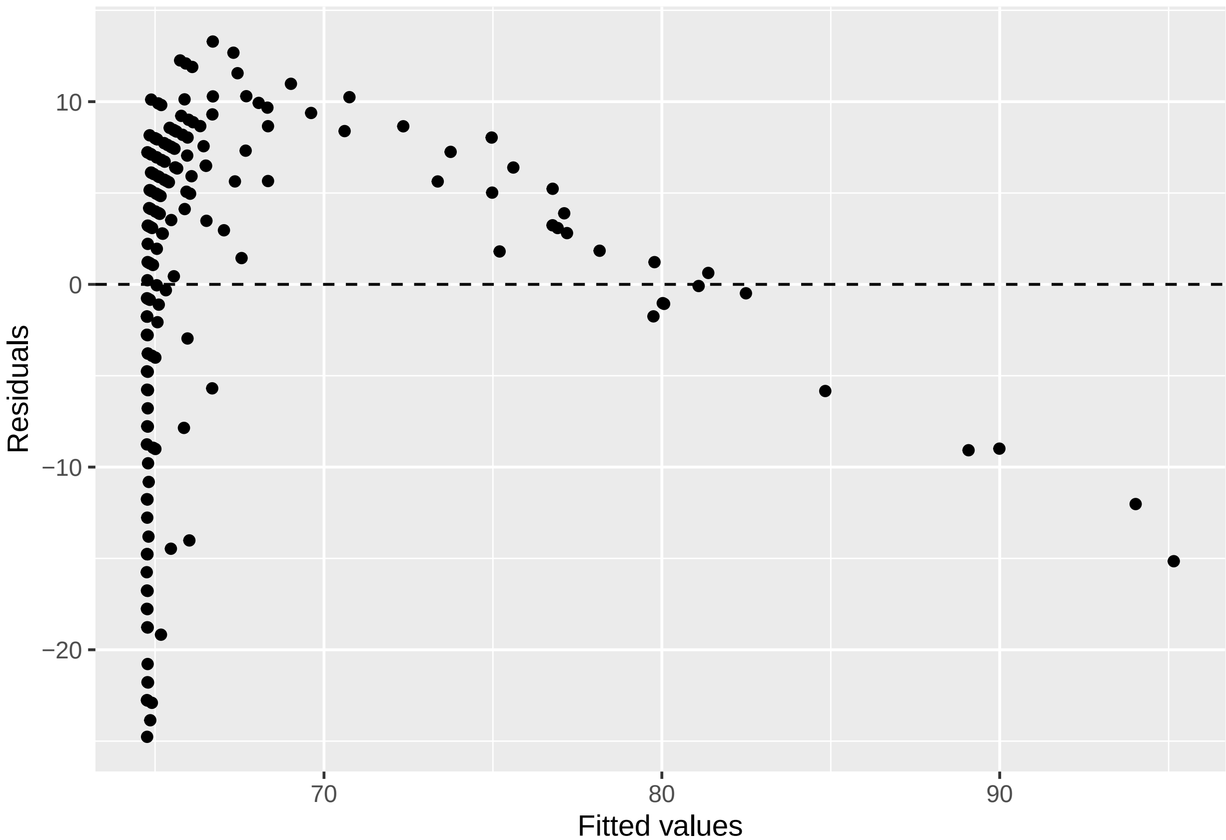
```r
ggplot(data = df, aes(x = model$residuals)) +
    geom_histogram(fill = 'steelblue', color = 'black') +
    labs(title = 'Histogram of Residuals', x = 'Residuals', y = 'Frequency')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



(3) constant variability: No, an obvious pattern in the fitted vs residual plot

```
ggplot(data = model, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")
```



2.

Raise life expectancy to the 4.6 power (i.e., LifeExp^4.6). Raise total expenditures to the 0.06 power (nearly a log transform, TotExp^.06). Plot LifeExp^4.6 as a function of TotExp^.06, and r re-run the simple regression model using the transformed variables. Provide and interpret the F statistics, R^2, standard error, and p-values. Which model is "better?"

```
df <- df %>%
  mutate(LifeExpFourSix = LifeExp^4.6) %>%
  mutate(TotalExpPointSix = TotExp^.06)
model <- lm(data=df, LifeExpFourSix ~ TotalExpPointSix)
summary(model)
```

```
##
## Call:
## lm(formula = LifeExpFourSix ~ TotalExpPointSix, data = df)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -308616089  -53978977   13697187   59139231  211951764
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -736527910   46817945  -15.73   <2e-16 ***
## TotalExpPointSix  620060216   27518940   22.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 90490000 on 188 degrees of freedom
## Multiple R-squared:  0.7298, Adjusted R-squared:  0.7283
## F-statistic: 507.7 on 1 and 188 DF,  p-value: < 2.2e-16
```

F-statistic:507.7 is meaningless for single parameter models

R^2:0.7298 indicates a well fitting model as it is much closer to 1 than 0

Standard Error:27518940 Standard error 620060216 /27518940 = 22.53213. The large ratio indicates little variability in the slope estimate

P-value:2.2e-16 much smaller than 0.05 which indicates statistical significance

The summary statistics all indicate that the second model is a better fitting model.

3.

Using the results from 3, forecast life expectancy when TotExp^.06 =1.5. Then forecast life expectancy when TotExp^.06=2.5.

$LifeExpectancy = -736527910 + 620060216 \times TotExp$

```r
l1 <- -736527910  + (620060216 * 1.5)
l1
```

```
## [1] 193562414
```

```r
l1 ^ (1/4.6)
```

```
## [1] 63.31153
```

```r
l2 <- -736527910  + 620060216 * 2.5
l2
```

```
## [1] 813622630
```

```r
l2 ^ (1/4.6)
```

```
## [1] 86.50645
```

Answer: $\approx$ 63 year life expectancy with 1.5 total expenditures and $\approx$ 87 life expectancy with 2.5 total expendeitures.

4.

Build the following multiple regression model and interpret the F Statistics, R^2, standard error, and p-values. How good is the model? LifeExp = b0+b1 x PropMd + b2 x TotExp +b3 x PropMD x TotExp

```r
multipleModel <- lm(data=df,LifeExp ~ PropMD + TotExp + (PropMD * TotExp))
summary(multipleModel)
```

```
##
## Call:
## lm(formula = LifeExp ~ PropMD + TotExp + (PropMD * TotExp), data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.320  -4.132   2.098   6.540  13.074
##
## Coefficients:
```

```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.277e+01  7.956e-01  78.899  < 2e-16 ***
## PropMD         1.497e+03  2.788e+02   5.371 2.32e-07 ***
## TotExp         7.233e-05  8.982e-06   8.053 9.39e-14 ***
## PropMD:TotExp -6.026e-03  1.472e-03  -4.093 6.35e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.765 on 186 degrees of freedom
## Multiple R-squared:  0.3574, Adjusted R-squared:  0.3471
## F-statistic: 34.49 on 3 and 186 DF,  p-value: < 2.2e-16
```

F-statistic:34.49

R^2:0.3574 indicates a well fitting model as it is much closer to 1 than 0

Standard Error:

2.788e+02 1.497e+03 / 2.788e+02 = 5.36944

8.982e-06 7.233e-05 / 8.982e-06 = 8.052772

1.472e-03 -6.026e-03/ 1.472e-03 =-4.09375

All of the standard error ratios are near 5-10 times smaller than the coefficients

P-value:2.2e-16 much smaller than 0.05 which indicates statistical significance

5.

Forecast LifeExp when PropMD=.03 and TotExp = 14. Does this forecast seem realistic? Why or why not?

$LifeExpectancy = 6.277e+01 + 1.497e+03 \times PropMD + 7.233e-05 \times TotExp - -6.026e-03 \times PropMD \times TotExp$

```r
PropMD <- 0.03
TotExp <- 14
6.277e+01  + (1.497e+03 * PropMD) + (7.233e-05 * TotExp) -(6.026e-03 * PropMD * TotExp)
```

```
## [1] 107.6785
```

Life expectancy of 107.6785 years Seems a little unrealistic as I would assume it is relatively uncommon to live to the age of 107.