Regression, Part 2

Sam Reeves

Who.csv contains real-world data from 2008:

$Country$	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

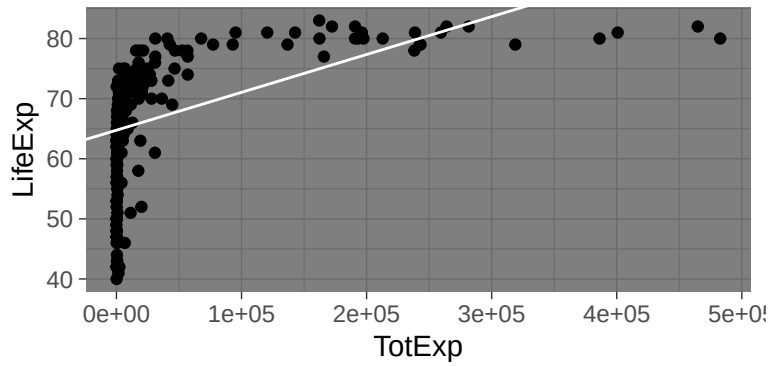
---

1. Provide a scatterplot of LifeExp~TotExp, and run a 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 are met.

```
library(ggplot2)
library(dplyr)

f <- read.csv("who.csv")
lm1 <- lm(LifeExp ~ TotExp, f)

ggplot(f, aes(x = TotExp, y = LifeExp)) +
  theme_dark() +
  geom_point() +
  geom_abline(intercept = lm1$coefficients[1],
              slope = lm1$coefficients[2],
              color = "white")
```



The scatterplot doesn't visually suggest a linear relationship. I have a very strong urge to take the log of the total expenditures, however, we've been instructed not to transform the variables at all.

```
summary(lm1)
```

```
##
## Call:
## lm(formula = LifeExp ~ TotExp, data = f)
##
## 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
```

```
anova <- aov(lm1)
summary(anova)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## TotExp        1   5731    5731   65.26 7.71e-14 ***
## Residuals   188  16509      88
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking first at the p-value, which is really small, we can reject the null hypothesis... That means that the relationship here is not easily explained by regular variance alone.

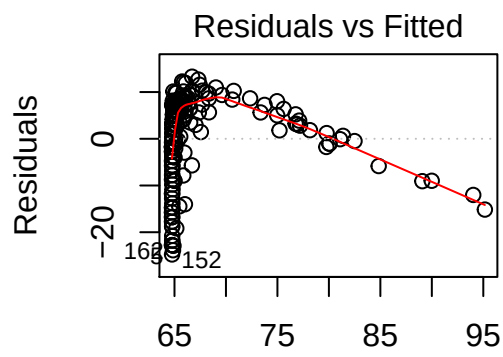
We don't need to consider the F-statistic. This model has only one independent variable.

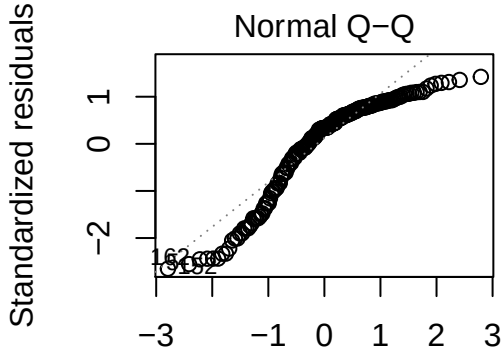
R^2 is only about 26%. That means that our independent variable only accounts for about a quarter of the variability in the y values. Our model has very little predictive power.

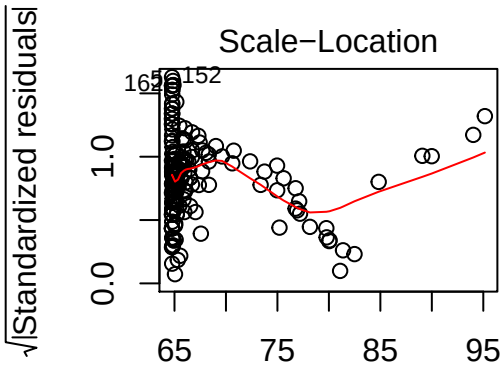
The standard error seems high. The range of values is from 40 to 80, and with a SE of nearly 10, I would assume that this is not a viable linear model.

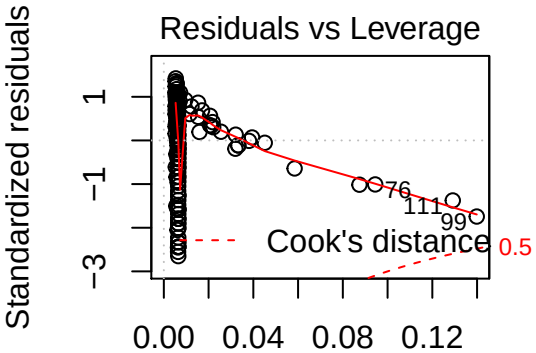
We do not pass the tests for linearity.

```
plot(lm1)
```





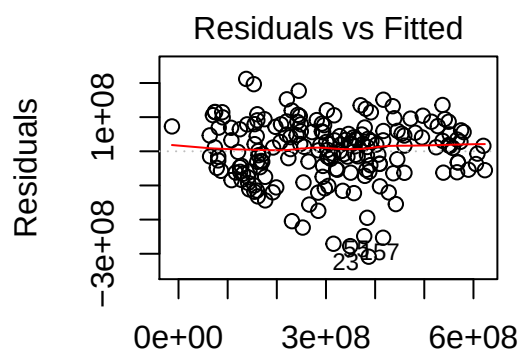


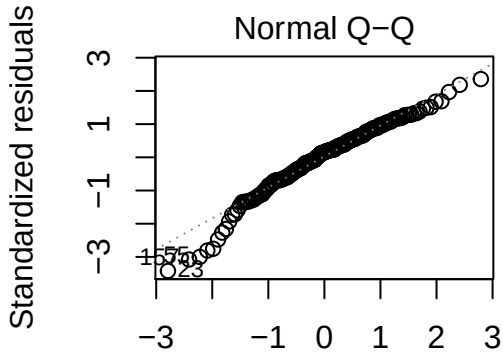


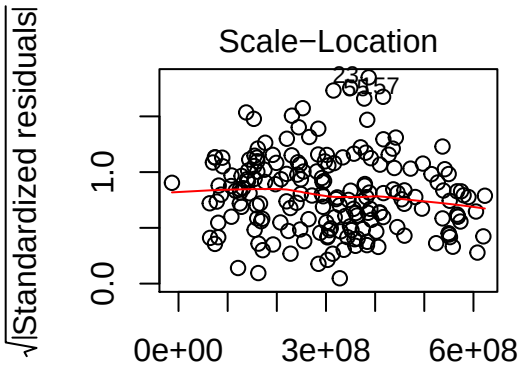
2. Raise life expectancy to the 4.6 power (i.e., LifeExp^4.6). Raise total expenditures to the 0.06 power (nearly a log transform, TotExp^.06). Plot LifeExp^4.6 as a function of TotExp^.06, and r re-run the simple regression model using the transformed variables. Provide and interpret the F statistics,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

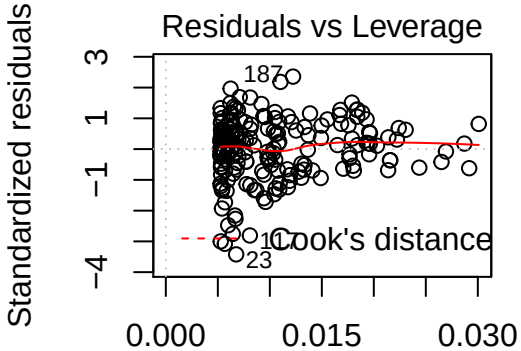
```
g <- f %>%
  mutate(LifeExp = LifeExp^4.6,
         TotExp = TotExp^0.06)

lm2 <- lm(LifeExp ~ TotExp, g)
plot(lm2)
```









Holy cow! Our linear model is suddenly behaving very nicely!

```
summary(lm2)
```

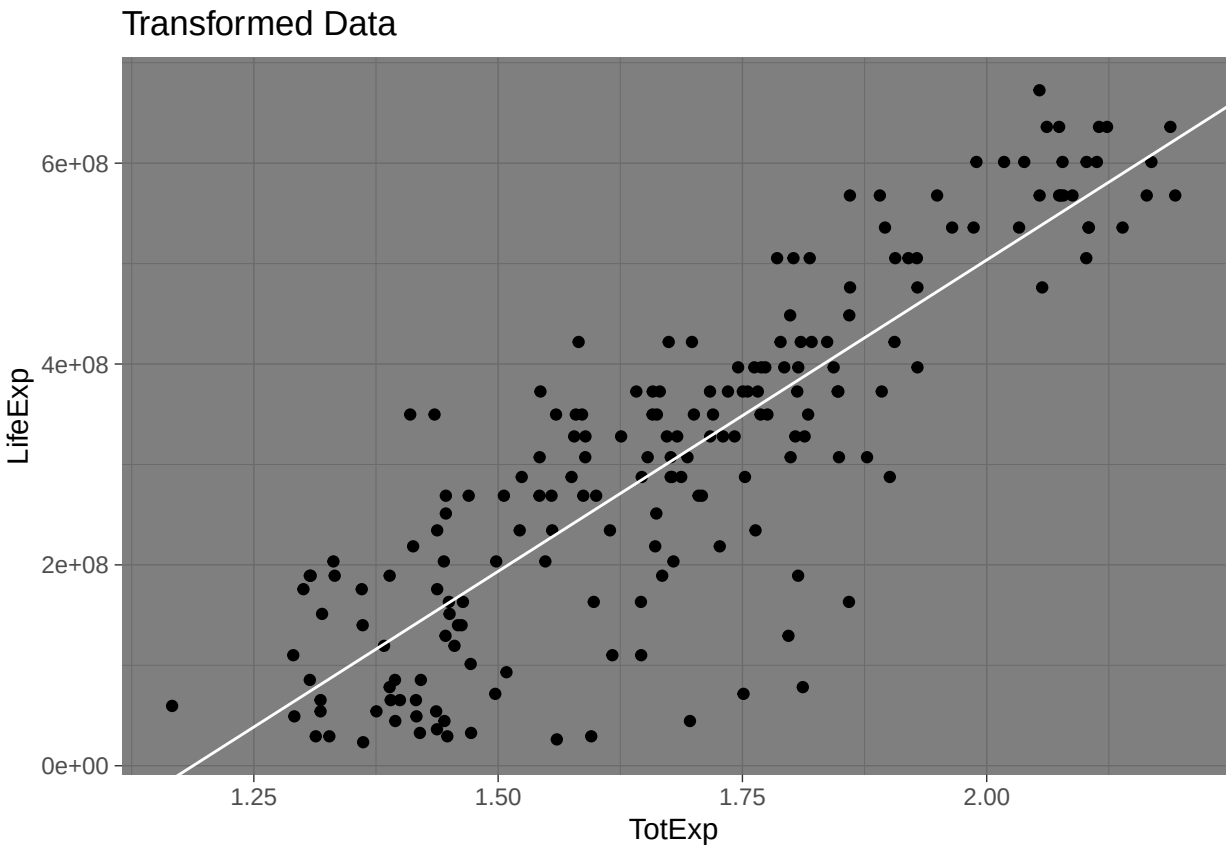
```
##
## Call:
## lm(formula = LifeExp ~ TotExp, data = g)
##
## Residuals:
##        Min          1Q      Median          3Q         Max
## -308616089  -53978977   13697187   59139231  211951764
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -736527910   46817945  -15.73   <2e-16 ***
## TotExp       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
```

```
anova2 <- aov(lm2)
summary(anova2)
```

```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## TotExp        1 4.157e+18 4.157e+18   507.7 <2e-16 ***
## Residuals   188 1.540e+18 8.189e+15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ggplot(g, aes(x = TotExp, y = LifeExp)) +
  theme_dark() +
  geom_point() +
  geom_abline(intercept = lm2$coefficients[1],
              slope = lm2$coefficients[2],
              color = "white") +
  labs(title = "Transformed Data")
```



3. Using the results from 2, forecast life expectancy when $TotExp^{.}06 = 1.5$. Then forecast life expectancy when TotExp^.06=2.5.

```
forecast <- function(x, lm) {
  intercept <- lm$coefficients[1]
  slope <- lm$coefficients[2]

  y <- (slope * x) + intercept

  names(y) <- c()
  LE <- y^(1/4.6)

  return(LE)
}

forecast(1.5, lm2)
```

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

```
forecast(2.5, lm2)
```

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

I suppose this is reasonable?

4. Build the following multiple regression model and interpret the F Statistics, R^2, standard error, and p-values. How good is the model?

$$LifeExp = b_0 + (b_1 \times PropMD) + (b_2 \times TotExp) + (b_3 \times PropMD \times TotExp)$$

```
lm3 <- lm(LifeExp ~ PropMD + TotExp + (PropMD * TotExp), f)
summary(lm3)
```

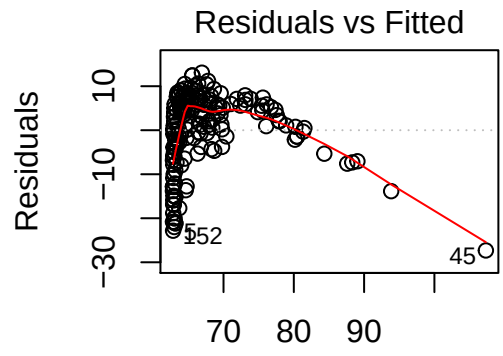
```
##
## Call:
## lm(formula = LifeExp ~ PropMD + TotExp + (PropMD * TotExp), data = f)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.320  -4.132   2.098   6.540  13.074
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.277e+01  7.956e-01  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
```

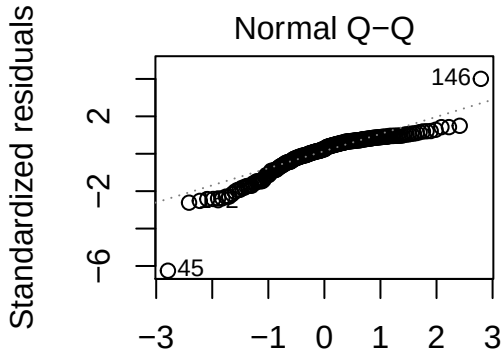
Looking again at the p-values first, all are very near zero. The F-statistic and its p-value suggest that the model is not overly complex, that the combination of independent variables together compose a model for which we can reject the null hypothesis.

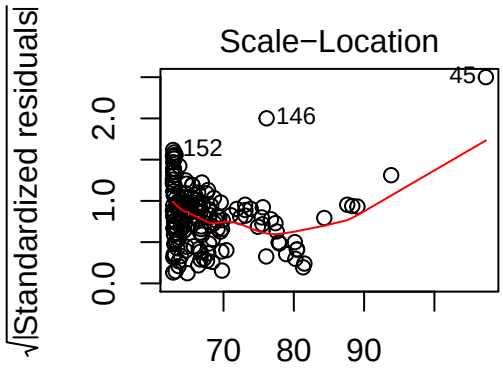
```
anova3 <- aov(lm3)
summary(anova3)
```

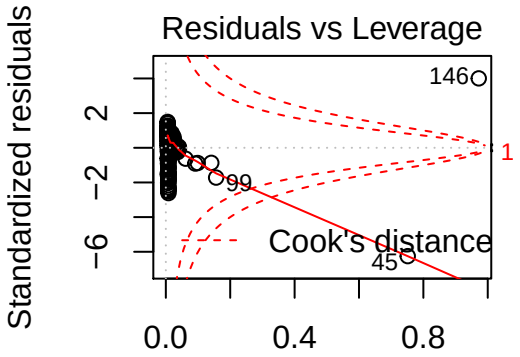
```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## PropMD          1   2967    2967   38.61 3.31e-09 ***
## TotExp          1   3696    3696   48.11 6.47e-11 ***
## PropMD:TotExp   1   1287    1287   16.75 6.35e-05 ***
## Residuals     186  14291      77
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(lm3)
```









These plots visually confirm that we have an appropriate model.

5. Forecast $LifeExp$ when $PropMD = 0.03$ and $TotExp = 14$. Does this forecast seem realistic? Why or why not?

```
forecast2 <- function(x1, x2, lm) {
  intercept <- lm$coefficients[1]
  m1 <- lm$coefficients[2]
  m2 <- lm$coefficients[3]

  y <- intercept + (m1 * x1) + (m2 * x2)

  names(y) <- c()

  return(y)
}

forecast2(0.03, 14, lm3)
```

```
## [1] 107.6985
```

Well, unfortunately, this doesn't seem like a reasonable answer.

This is a ridiculously high life expectancy for a couple of reasons. It falls way outside of the range of ages in the initial dataset, so we can't assume that the relationship we have described in this model goes out to infinity. It's a really advanced age. It's much safer to assume that there are diminishing returns on these expenditures past a certain point... Where to cap the expected values? I would say that maybe it's possible to arrive at a life expectancy a bit higher than the highest available in the data, but certainly not by more than a few years.