CUNY 605 Wk 12 Regression Analysis

Jen Abinette

2023-04-23

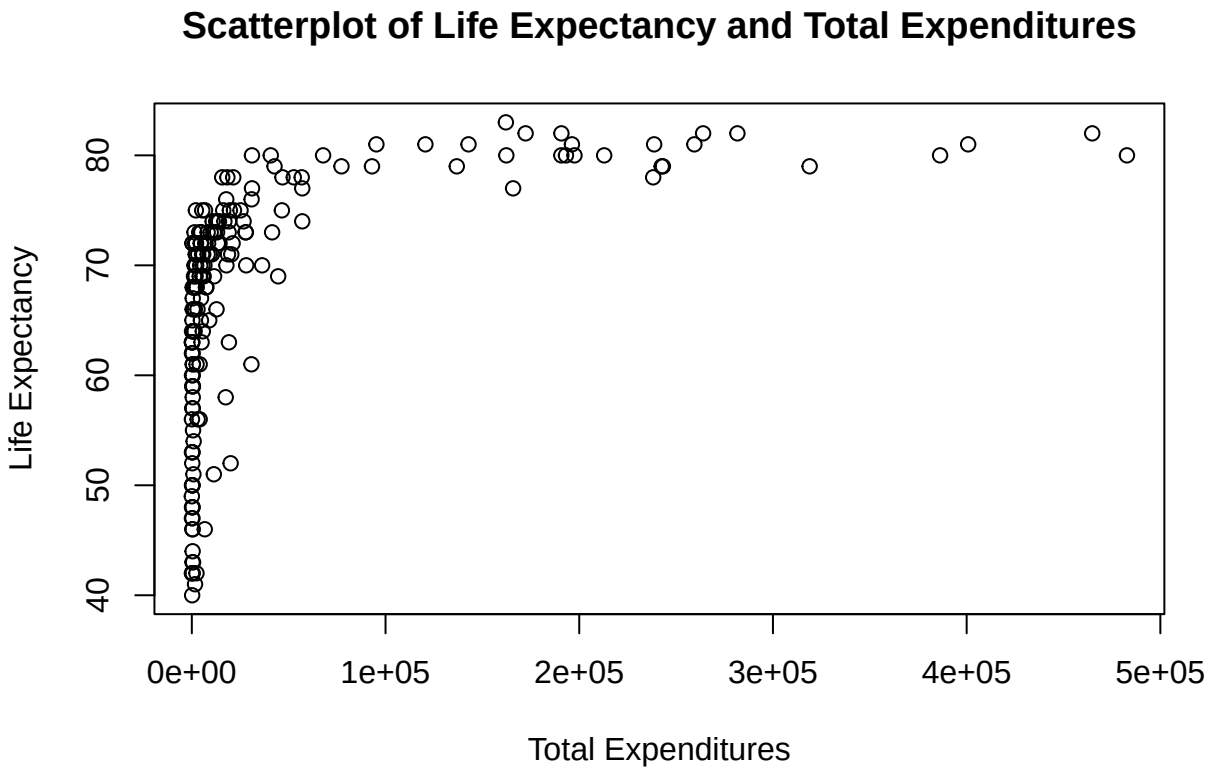
Load csv file from Github

```
who <- read.csv('https://raw.githubusercontent.com/JAbinette/CUNY-605-Computational-Math/main/12%20-%20
```

Simple Linear Regression Model

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.

```
plot(who$TotExp,who$LifeExp,xlab="Total Expenditures", ylab="Life Expectancy", main="Scatterplot of Life
```



```r
who.lm <- lm(LifeExp~TotExp, data=who)
summary(who.lm)
```

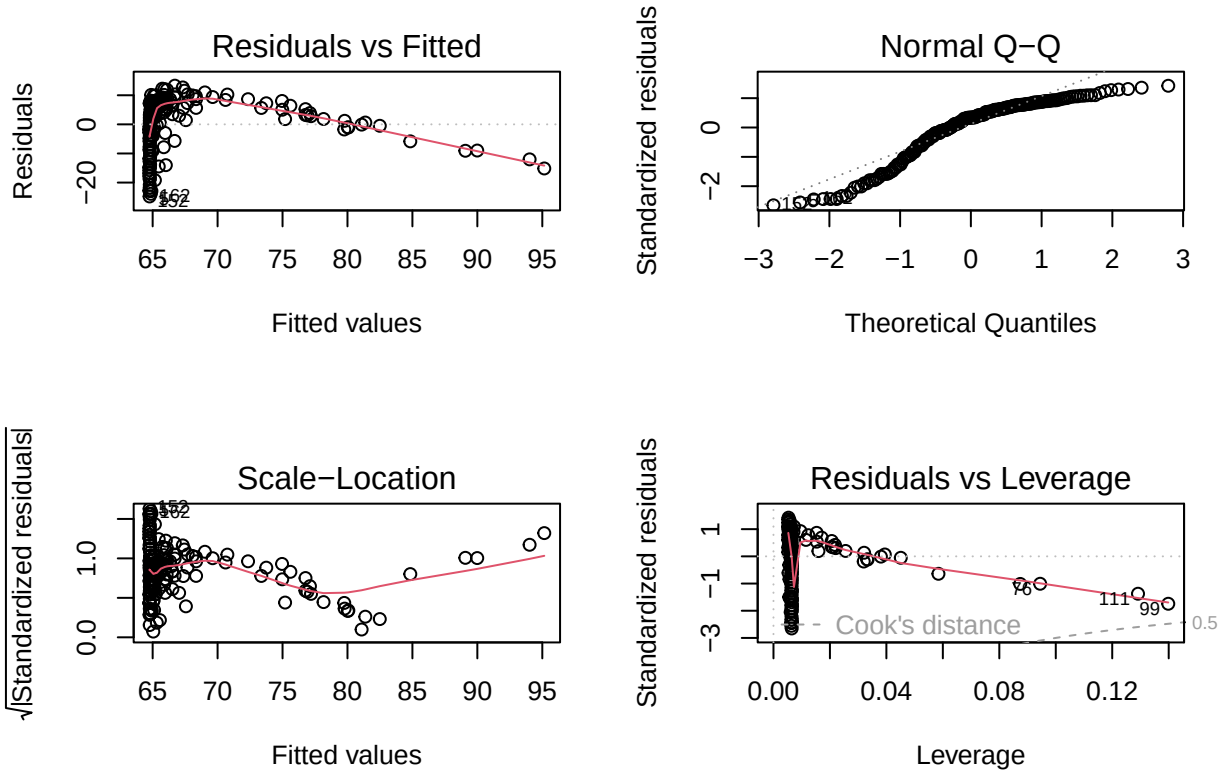
```
##
## Call:
## lm(formula = LifeExp ~ TotExp, data = who)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.764  -4.778   3.154   7.116  13.292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.475e+01  7.535e-01  85.933  < 2e-16 ***
## TotExp      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
```

The standard errors for the intercept and Total Expenditure coefficients in the model are both more than five times smaller than the coefficients and significant with p-values less than .001 supporting this is a good model and there's little variability in the slope estimates.

The effect size of this model, R-squared, indicates that 26% of the variability in Life Expectancy is explained by the variation in Total Expenditures.

Lastly, the F Statistic supports that this model using Total Expenditures as a predictor is significantly better than a Intercept-only model.

```r
par(mfrow=c(2,2))
plot(who.lm)
```

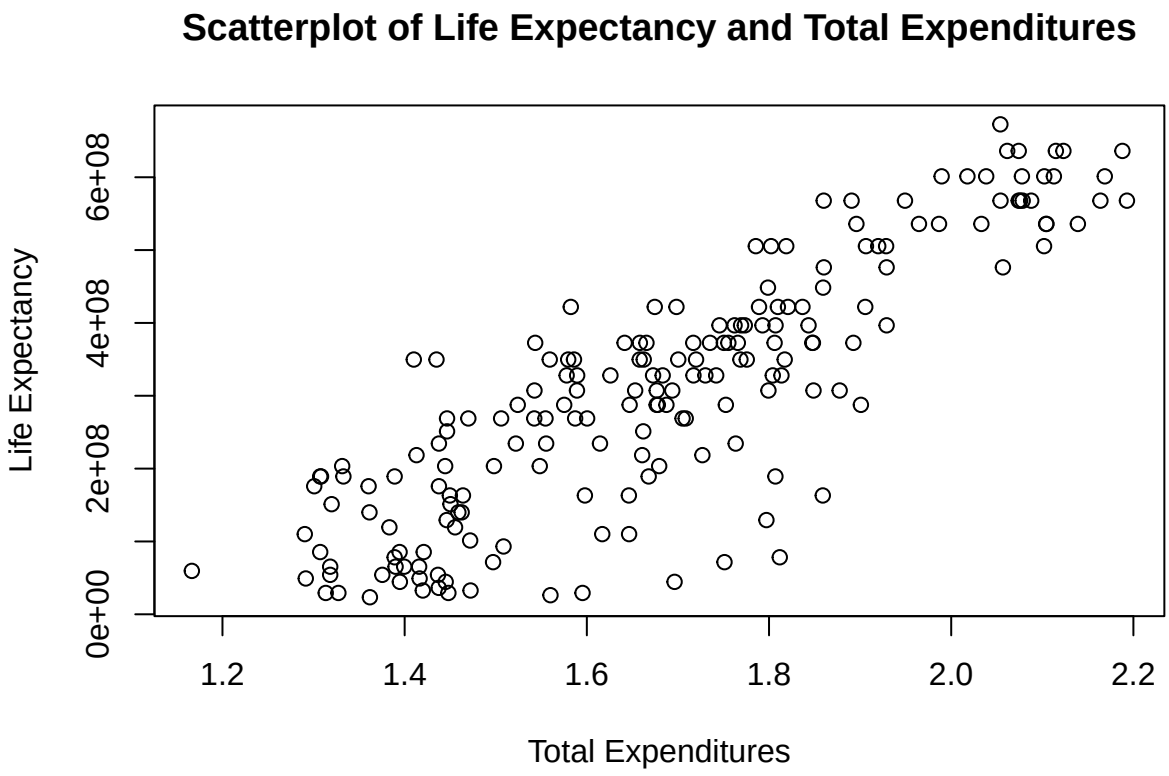


Although the model looked promising based on the summary statistics, the residual plots indicate an issue with our model assumptions as the residuals are not nearly normally distributed and there is evidence of an exponential pattern. Based on these plots, we should explore transforming variables to create better fitting model.

Model with Transformed Variables

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?"

```
who$LifeExp_4.6 <- who$LifeExp^4.6
who$TotExp_06 <- who$TotExp^.06
plot(who$TotExp_06,who$LifeExp_4.6,xlab="Total Expenditures", ylab="Life Expectancy", main="Scatterplot
```




```
transform.lm <- lm(LifeExp_4.6~TotExp_06, data=who)
summary(transform.lm)
```

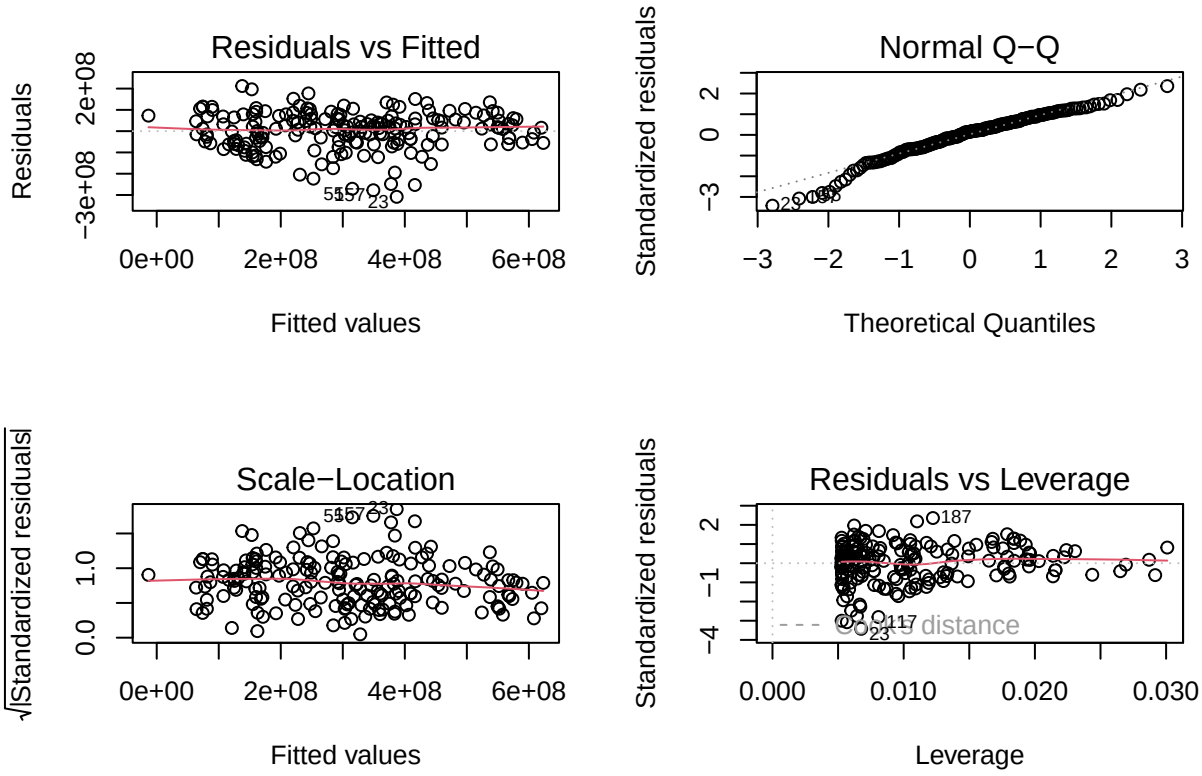
```
##
## Call:
## lm(formula = LifeExp_4.6 ~ TotExp_06, data = who)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -308616089  -53978977   13697187   59139231  211951764
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -736527910   46817945  -15.73   <2e-16 ***
## TotExp_06    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
```

The standard errors for the intercept and Total Expenditure coefficients in this model are both more than ten times smaller than the coefficients and significant with p-values less than .001 supporting this is a good model and there's little variability in the slope estimates.

The effect size of this model, R-squared, indicates that 73% of the variability in Life Expectancy is explained by the variation in Total Expenditures.

Lastly, the F Statistic supports that this model using Total Expenditures as a predictor is significantly better than a Intercept-only model.

```
par(mfrow=c(2,2))
plot(transform.lm)
```



Which Model is Better?

Our second model using transformed variables is a better model as it has a greater effect size and the model assumptions are not violated. This is evidenced by the second model explaining an additional 47% of the variability in Life Expectancy as the first model predicted 26% of the variability as compared to 73% in the transformed model. Additionally, the residuals plots support the greater reliability of this model given the linearity, nearly normal residuals, and constant variability.

Forecast Life Expectancy

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

```
intercept = -736527910
TotExp_06_Coeff = 620060216
y1 = intercept+TotExp_06_Coeff*1.5
y2 = intercept+TotExp_06_Coeff*2.5

paste("Using our transformed variable model, we predict life expectancy to be approximately ",round(y1^
```

```
## [1] "Using our transformed variable model, we predict life expectancy to be approximately  63  and  8
```

Multiple Regression Model

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

```
mreg_lm <- lm(LifeExp~TotExp+PropMD+(PropMD*TotExp),data=who)
summary(mreg_lm)
```

```
##
## Call:
## lm(formula = LifeExp ~ TotExp + PropMD + (PropMD * TotExp), data = who)
##
## 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 ***
## TotExp         7.233e-05  8.982e-06   8.053 9.39e-14 ***
## PropMD         1.497e+03  2.788e+02   5.371 2.32e-07 ***
## TotExp:PropMD -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
```

Similar to the first Linear Model, the standard errors for the intercept and coefficients in this model are both more than five times smaller than the coefficients with the intercept having the least variability. Additionally, all predictors are significant with p-values less than .001 supporting this is a good model and there's little variability in the slope estimates.

The effect size of this model, R-squared, indicates that 36% of the variability in Life Expectancy is explained by the model, which is greater than the effect size in our first model of 26%, but less than the 73% explained by our second model using transformed variables.

Lastly, the F Statistic supports that this model is significantly better than a Intercept-only model.

Forecast Life Expectancy using Multiple Regression Model

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

```
MR_intercept = 6.277e+01
MR_TotExp_Coeff = 7.233e-05
MR_PropMD_Coeff = 1.497e+03
MR_Interaction_Coeff = -6.026e-03

MR_y1 = MR_intercept + MR_TotExp_Coeff*14 + MR_PropMD_Coeff*.03 + MR_Interaction_Coeff*14*.03

paste("Using our Multiple Regression model, we predict life expectancy to be approximately ",round(MR_y
```

```
## [1] "Using our Multiple Regression model, we predict life expectancy to be approximately  108  years
```