605-Wk12-HW

Jose Mawyin

11/17/2019

DATA 605 Wk12-HW. Regression of Non-Linear Variables

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.

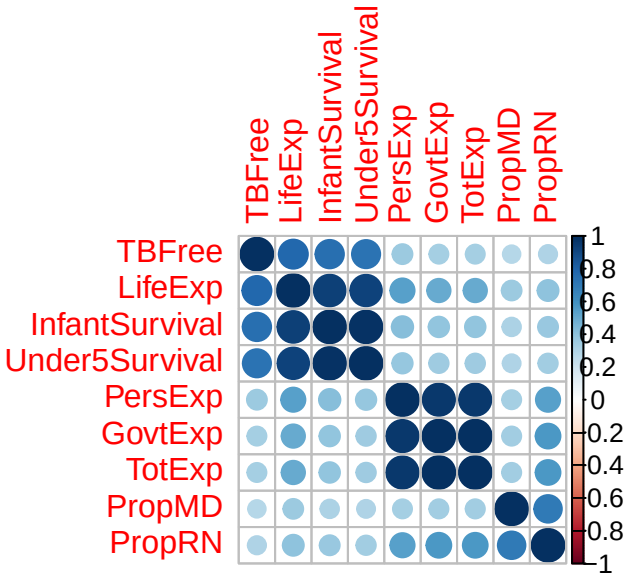

Figure 1: Correlation Plot of all Variables in Dataset

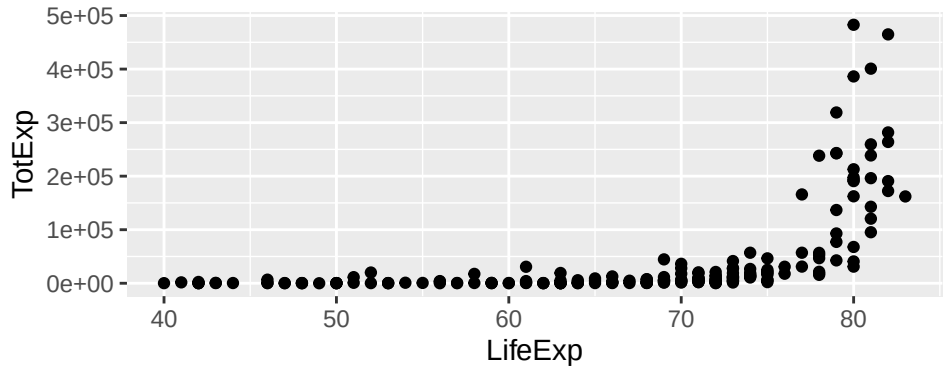

Figure 2: Scatter Plot of LifeExp Vs. TotExp

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.

Looking at the fit parameters we notice that:

*F statistics of 65.26:

*R^2 of 0.2577: The calculated linear fit model of TotExp accounts for only for 25.77% of the variability of the data. Not a good percentage.

*Standard Error of 7.795e-06: In a good model the standard error should five to ten times smaller than the corresponding coefficient. In our case is 8.08 times smaller.

*P-Values of 7.714e-14: Our found p-value is quite low indicating a very low probability that coefficient is not relevant in the model.

In our linear fit graph in Figure. 3 we notice how poorly a linear line fits the data when the relationship LifeExp Vs. TotExp is not linear but a power relationship.

##

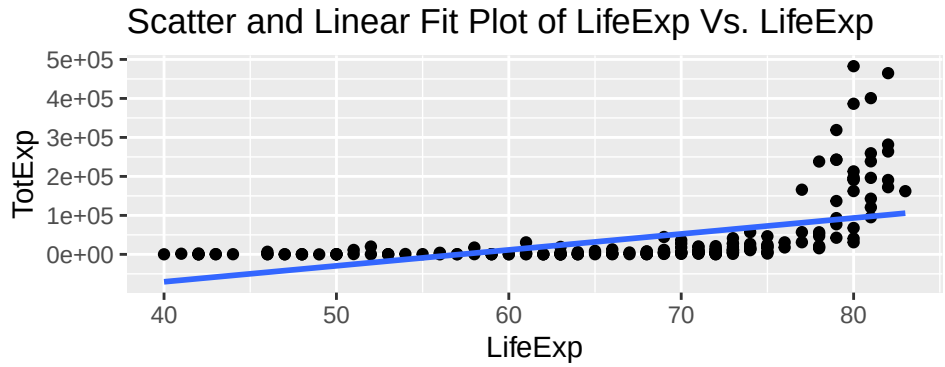

Figure 3: Scatter Plot and Linear Fit of LifeExp Vs. TotExp

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

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?"

*F statistics of 507.7:

*R^2 of 0.7298: The calculated linear fit model of TotExp accounts for 72.98% of the variability of the data. A significant improvement over the 25.77% when using linear variables.

*Standard Error of 27518940: In a good model the standard error should five to ten times smaller than the corresponding coefficient. In our case, by using the using the LifeExp^4.6 Vs. TotExp^.06 variables it has increased from 8.08 to 22.53 times smaller.

*P-Values of 2.2e-16: Our found p-value is quite low indicating a very low probability that coefficient is not relevant in the model.

```
attach(who.data)
who.data.2 <- who.data
who.data.2$LifeExp <- (who.data.2$LifeExp)^4.6
who.data.2$TotExp <- (who.data.2$TotExp)^0.06
```

```
## The following objects are masked from who.data (pos = 3):
##
##     Country, GovtExp, InfantSurvival, LifeExp, PersExp, PropMD,
##     PropRN, TBFree, TotExp, Under5Survival

## The following objects are masked from who.data (pos = 4):
##
##     Country, GovtExp, InfantSurvival, LifeExp, PersExp, PropMD,
##     PropRN, TBFree, TotExp, Under5Survival

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

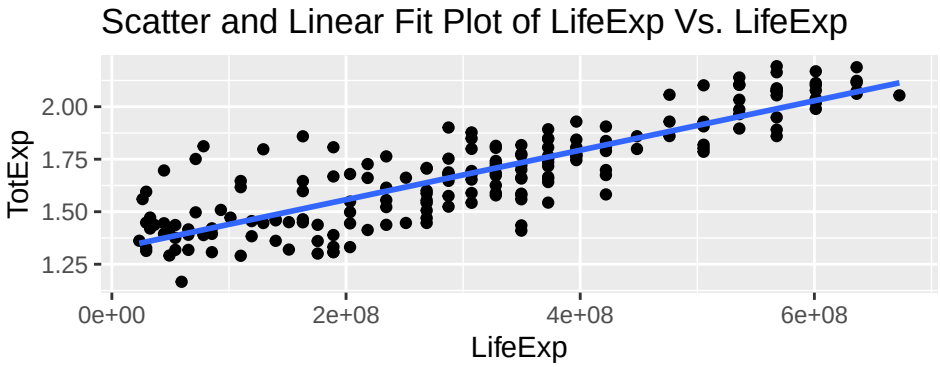

Figure 4: Scatter Plot and Linear Fit of LifeExp^4.6 Vs. TotExp^.06

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

```
who.2.lm
```

```
##
## Call:
## lm(formula = LifeExp ~ TotExp)
##
## Coefficients:
## (Intercept)       TotExp
##  -736527909    620060216
```

```
TotExp_Power <- 1.5
LifeExp_Power<- -736527910 + 620060216*(TotExp_Power)
Life.Exp.at.1.5 <- LifeExp_Power^(1/4.6)
cat("The forecasted life expectancy when TotExp^.06 =1.5 is", (Life.Exp.at.1.5) %>% round(0)," years.")
```

```
## The forecasted life expectancy when TotExp^.06 =1.5 is 63  years.
```

```
who.2.lm
```

```
##
## Call:
## lm(formula = LifeExp ~ TotExp)
##
## Coefficients:
## (Intercept)       TotExp
##  -736527909    620060216
```

```
TotExp_Power <- 2.5
LifeExp_Power<- -736527910 + 620060216*(TotExp_Power)
Life.Exp.at.1.5 <- LifeExp_Power^(1/4.6)
cat("The forecasted life expectancy when TotExp^.06 =2.5 is", (Life.Exp.at.1.5) %>% round(0)," years.")
```

```
## The forecasted life expectancy when TotExp^.06 =2.5 is 87  years.
```

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

```
##
## Call:
## lm(formula = LifeExp ~ PropMD + TotExp + PropMD * TotExp)
##
## 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 statistics of 34.49:

*R^2 of 0.3574: The calculated linear fit model of TotExp accounts for 32.74% of the variability of the data. Worse than the previous model using variables to a power.

- Standard Error: In a good model the standard error should five to ten times smaller than the corresponding coefficient. In our case, we have for:
- PropMD: Standard error 5.37 larger than corresponding coefficient.
- TotExp: Standard error 8.05 larger than corresponding coefficient.
- PropMD:TotExp: Standard error 4.09 larger than corresponding coefficient.

*P-Values of 2.2e-16: Our found p-value is quite low indicating a very low probability that coefficient is not relevant in the model. Also, is the same as in the previous case. Puzzling!

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

```
who.lm.3
```

```
##
## Call:
## lm(formula = LifeExp ~ PropMD + TotExp + PropMD * TotExp)
##
## Coefficients:
##   (Intercept)         PropMD         TotExp  PropMD:TotExp
##     6.277e+01      1.497e+03      7.233e-05     -6.026e-03
```

```
PropMD <- .03
TotExp <- 14
LifeExp<- 6.277e+01 + 1.497e+03*PropMD + 7.233e-05*TotExp + (-6.026e-03)*PropMD*TotExp
cat("The forecasted life expectancy when PropMD=.03 and TotExp = 14 is", (LifeExp) %>% round(0)," years
```

```
## The forecasted life expectancy when PropMD=.03 and TotExp = 14 is 108  years.
```

```
summary(who.data$LifeExp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   40.00   61.25   70.00   67.38   75.00   83.00
```

The forecasted life expectancy of 108 years sounds in the realm of possibility. However, when checking the summary of the life expectancy data given, we see that the highest recorded value of LifeExp is only 83 years. Therefore, we are extrapolating outside the realm of the data used in the regression model:

LifeExp = b0+b1 x PropMd + b2 x TotExp +b3 x PropMD x TotExp