HW12

Zachary Palmore

4/20/2021

Directions

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.

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?"

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

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

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

Part 1

First, the data is read and the first few rows are shown for reference. It should match our variables above.

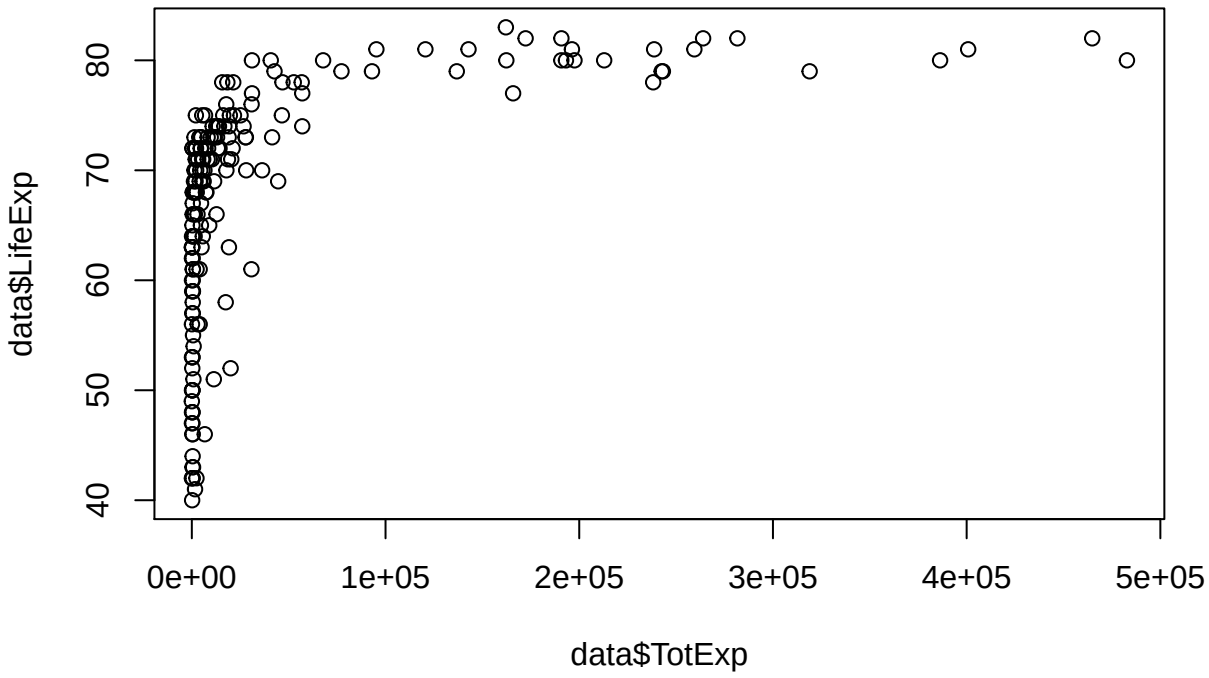
```
data <- read.csv("https://raw.githubusercontent.com/palmorezm/msds/main/605/who.csv")
head(data)
```

```
##               Country LifeExp InfantSurvival Under5Survival  TBFree      PropMD
## 1         Afghanistan      42          0.835          0.743 0.99769 0.000228841
## 2             Albania      71          0.985          0.983 0.99974 0.001143127
```

```
## 3             Algeria      71          0.967          0.962 0.99944 0.001060478
## 4             Andorra      82          0.997          0.996 0.99983 0.003297297
## 5              Angola      41          0.846          0.740 0.99656 0.000070400
## 6 Antigua and Barbuda      73          0.990          0.989 0.99991 0.000142857
##        PropRN PersExp GovtExp TotExp
## 1 0.000572294      20      92    112
## 2 0.004614439     169    3128   3297
## 3 0.002091362     108    5184   5292
## 4 0.003500000    2589  169725 172314
## 5 0.001146162      36    1620   1656
## 6 0.002773810     503   12543  13046
```

Next we create a scatter plot with the sum of personal and government expenditures (or total expenditures) on the x-axis and average life expectancy for each country in years on the y-axis.

```
plot(data$TotExp, data$LifeExp)
```



Our main takeaway is that the points do not follow a straight liner trend. There is a clear pattern that might be exponential or logistic but nothing about the relationship between the two indicates a steady increase or decrease as one changes. Regardless, we run a linear model to assess a few statistics. In it we ensure that life expectancy (LifeExp) is modeled by the total expenditures (TotExp). The steps and summary are provided below.

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

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

This summary shows that the coefficient of determination, or R^2, is 0.2576922 which describes how well the regression line explains the data. In this case, it specifically looks at how much variation in life expectancy is explained by the total personal and government expenditures. As expected, a linear trend on this data without transformation is poor, with the R^2 explaining only about 25% of the model fit. Many would consider this a weakly correlated value given that the range of possible correlation strength extends from 0 to 1 with one indicating a perfect fit for all points.

The p-value for total expenditures is $7.8070708 \times 10^{-153}$ which is statistically significant beyond the alpha level of 0.001. If our null hypothesis is that adding our variable would result in no change in our model, then we could reject the null. In this case, the addition of total expenditures significantly changed the estimates. However, statistical significance alone, does not mean we can make accurate predictions from it. This p-value only shows that the probability of this change occurring by random chance is extremely low.

Results also show that the F statistic is 65.2641982. The F statistic is in an indication of overall significance in the model. It tells us whether the regression plotted has a better fit than an intercept-only model. The further the value is from 1, the better the model. Our critical value is probably less than 2 with 188 degrees of freedom and this value is greater than that critical value. Thus, we could reject the null hypothesis to conclude that the addition of total expenditures is statistically significant and that there is a strong relationship between the variables (stronger than the null hypothesis). However, here again, significance does not equate to a perfect linear model nor its ability to be useful.

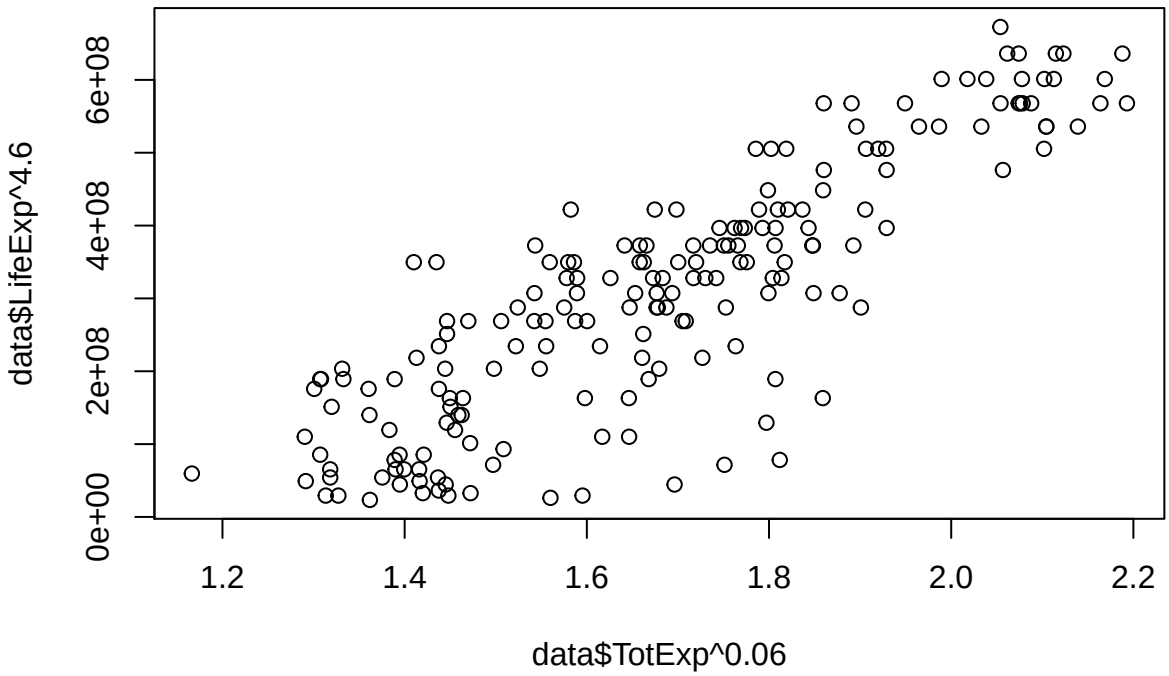
Going further, the standard error is 0.7535366. The closer this value is to 0, the better, since it indicates how far the data is from the model expectations on average. Notice that this value is extremely low. Some consider this an indicator of the quality of the fit in a model. This low value indicates high quality, assuming the other assumptions of linear regression are met. Additionally, the total expenditure coefficient is 64.7533745 +/- 1.4769318 at 95% confidence.

Based on these statistics, the assumptions of simple linear regression are not met. Consider that the initial scatter plot displayed no signs of linearity, an important assumption in linear regression. For us to use linear regression, one must assume that there is a linear relationship in the data. In this case, there is not. Additionally, the normality of a plot such as the one shown, would also exhibit non-normal trends. These two failures in our assumptions would eliminate the option of using a linear regression model to make predictions or interpretations without transformations.

Part 2

To see if the data can appear more linear and normal it is transformed. The total expenditures values are raised to the power of 0.06 and the life expectancy values to the power of 4.6. The resultant scatterplot is shown below.

```
plot(data$TotExp^.06, data$LifeExp^4.6)
```



This shows a much more linear trend with greater potential to fulfill the assumptions of linear regression. We continue with a summary of the linear model to compare those statistics previously mentioned.

```
data$LifeExp4.6 <- data$LifeExp^4.6
data$TotExp.06 <- data$TotExp^.06
lm.data.t <- lm(LifeExp4.6~TotExp.06, data)
summary(lm.data.t)
```

```
##
## Call:
## lm(formula = LifeExp4.6 ~ TotExp.06, data = data)
##
## 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
```

This summary shows that the coefficient of determination, or R^2, is 0.7297673 which is much better than our previous model. With this statistic one might assume that the variation in life expectancy explains about 72% of the expenditures. This is a much strong correlation than our original model.

The p-value for total expenditures is $3.9117741 \times 10^{-36}$ which is statistically significant beyond the alpha level of 0.001. Again, the null hypothesis assumes the addition of total expenditures is not significant and could be due to random chance. In this case, our probability of these changes occurring by chance is again, extremely low, but this p-value also happens to match that of the model without any independent variables.

These results show that the F statistic is 507.6967054 suggesting that this model is almost certainly significantly and we can reject the null hypothesis. This and the p-value means that the addition of total expenditure is statistically significant in our model and that there is a closer relationship between the variables than in our previous model. However, there is something else that must be considered.

The standard error of this transformed data set is 4.6817945×10^7. Recall that, the closer this value is to 0, the better, since it indicates how far the data is from the model expectations on average. This error value is extremely high suggesting the data associated with this model are extremely far away from their fit in the model. As an indicator of model quality, this would be a very poor model. However, based on the information observed and how we were directed, this model is technically better because it allows us to satisfy the necessary assumptions for linear regression and create the model under more appropriate conditions. Although, it is practically useless for prediction due to its high error and these transformations.

Part 3

To forecast life expectancy based on two values of total expenditures, located at 1.5 and 2.5, we create a new data frame using those values and predict with it. Confidence intervals accompany each prediction which is based on the transformed model.

```
newdata <- data.frame(TotExp.06=c(1.5,2.5))
predict(lm.data.t, newdata, interval="predict")^(1/4.6)
```

```
##        fit      lwr      upr
## 1 63.31153 35.93545 73.00793
## 2 86.50645 81.80643 90.43414
```

Based on this output Life expectancy is about 63 years when total expenditures is at a value of about 1.5. Life expectancy then increases with increased expenditures to about 87 when the total expenditures value is at 2.5. The 95% confidence intervals are from approximately 36 - 73 years when total expenditures is at 1.5 and about 82 - 90 years when total expenditures is at 2.5.

Part 4

```r
lm.example <- lm(LifeExp ~ PropMD + TotExp + TotExp:PropMD, data)
summary(lm.example)
```

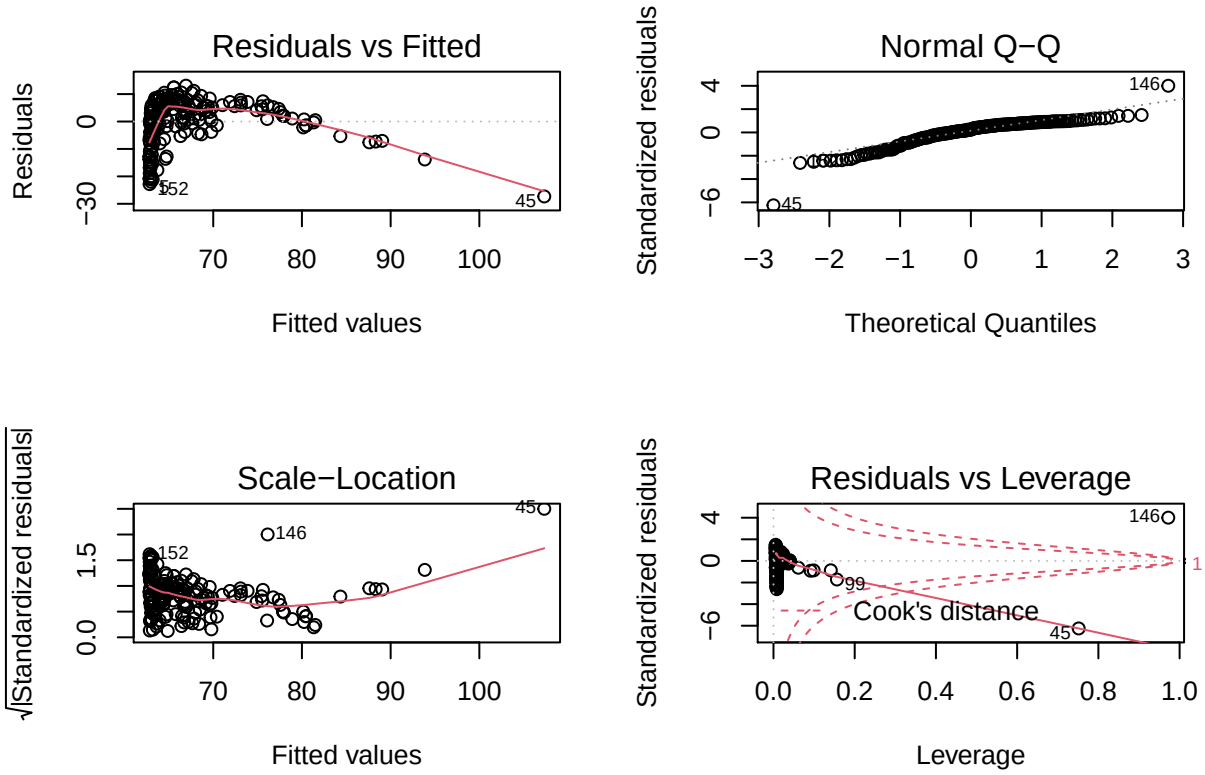
```
##
## Call:
## lm(formula = LifeExp ~ PropMD + TotExp + TotExp:PropMD, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.320  -4.132   2.098   6.540  13.074
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.277e+01  7.956e-01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

Comparing the same statistics as previously mentioned, this model is generally worse than the second model created with the transformations. To start, the R^2 value is much lower, indicating that the model does not explain most of the variation between life expectancy and total expenditures. The F-statistic is also several times smaller than our transformed model. Results are still significant but to a lesser degree. The standard error is also lower, much closer to zero but that may not be enough. If we were to examine more we might create some diagnostic plots to validate the conditional assumptions of linearity as shown:

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

```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



Looking at the diagnostic plots, there is a clear pattern in the residuals and normality is a little off. Thus, our assumptions are already invalid. However, there is also a problem with overly influential values and their leverage on the data, as well as residuals being unequally spread in the Scale-location plot. Nothing about this model indicates a good linear fit for this data.

Part 5

Predicting using this new model and the setting the proportion of doctors to 0.03 with the total expenditures to 14, the results are not realistic.

```
newdata.example <- data.frame(PropMD=0.03, TotExp=14)
predict(lm.example, newdata.example, interval="predict")
```

```
##       fit      lwr      upr
## 1 107.696 84.24791 131.1441
```

Our prediction based on this model is that the average life expectancy would be higher than the age most people could live to be (given current behavior and technology) at about 108 years. From this model we can also be 95% confident that the true life expectancy lies between approximately 84 - 131 years. The highest average life expectancy that is attainable based on this data seems to be about 83 years with a standard deviation of about 10.85 years. The median is 70 years and mean is about 67 years. For this prediction, the average life expectancy is not likely to be three standard deviations higher than the mean as that would indicate increasing the proportion of doctors and total expenditures would boost life expectancy by almost 41 years. This is unrealistic. A full summary of the life expectancy is shown below.

```r
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.0.3
```

```r
describe(data$LifeExp)
```

```
##    vars   n  mean    sd median trimmed   mad min max range skew kurtosis   se
## X1    1 190 67.38 10.85     70   68.47 10.38  40  83    43 -0.8    -0.25 0.79
```

Furthermore, the link between life expectancy and expenditures in this model is far from perfect. There is a weak correlation at best and the data is likely not linear yet we are predicting as if it was linear. Based on the previous models, it is not likely that the life expectancy of humans can continuously increase as expenditures increase. The same applies for the other variables in this model. Increasing the proportion of doctors in the population might have a net positive effect on the average life expectancy for the country but it would require many countries each increasing this proportion simultaneously to see if there truly was as big of an impact as implied by this prediction. However, it is quite unlikely due to the nature of the human life. At some point, there are diminishing returns and this model does not take that into account.