# WHO dataset Regression Analysis

Christian Thieme

11/11/2020

# Forecasting Life Expectancy

## The Data

Using the provided 2008 dataset from the World Health Organization, we'll look at factors that affect life expectancy as well as determine if we can build a model to predict life expectancy. Our dataset has the following columns:

- 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.

We'll load the dataset using the `readr` package and then look at the first five rows of data:

```
who <- readr::read_csv('C:/Users/chris/OneDrive/Master Of Data Science - CUNY/Fall 2020

head(who)
```

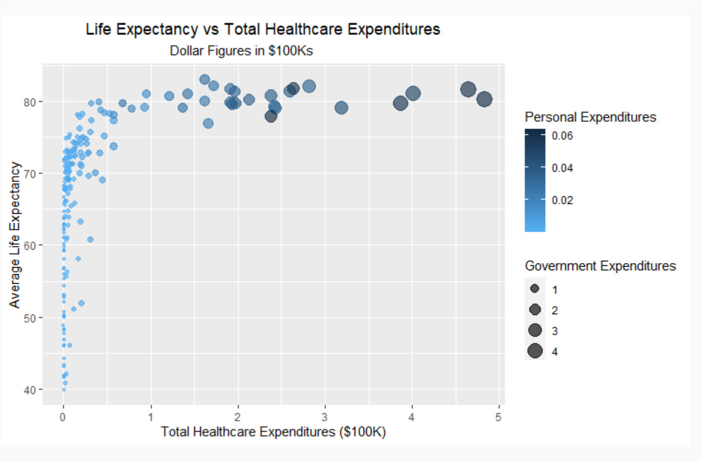
```
## # A tibble: 6 x 10
##   Country LifeExp InfantSurvival Under5Survival TBFree  PropMD  PropRN PersExp
##   <chr>     <dbl>          <dbl>          <dbl>  <dbl>   <dbl>   <dbl>   <dbl>
## 1 Afghan~      42          0.835          0.743  0.998 2.29e-4 5.72e-4      20
## 2 Albania      71          0.985          0.983  1.00  1.14e-3 4.61e-3     169
## 3 Algeria      71          0.967          0.962  0.999 1.06e-3 2.09e-3     108
## 4 Andorra      82          0.997          0.996  1.00  3.30e-3 3.50e-3    2589
## 5 Angola       41          0.846          0.74   0.997 7.04e-5 1.15e-3      36
## 6 Antigu~      73          0.99           0.989  1.00  1.43e-4 2.77e-3     503
## # ... with 2 more variables: GovtExp <dbl>, TotExp <dbl>
```

## Analysis

**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 are met.**

We'll build a scatterplot with our dependant variable, Life Expectancy, on the y-axis and our independent variable, Total Healthcare Expenditures, on the x-axis. Additionally, we'll add Personal Expenditures as a color gradient variable and Government Expenditures as a size variable. This will help us to determine if either Personal or Healthcare expenditures are more pronounced in Life Expectancy.

```
ggplot(who) +
  aes(x = TotExp/100000, y = LifeExp, size = GovtExp/100000, color = PersExp/100000) +
  geom_jitter(alpha = 0.65) +
  #geom_smooth(method = lm, se = FALSE) +
  labs(title = "Life Expectancy vs Total Healthcare Expenditures", subtitle = "Dollar F
  ylab("Average Life Expectancy") +
  xlab("Total Healthcare Expenditures ($100K)") +
  theme(
    plot.title = element_text(hjust = 0.40),
    plot.subtitle = element_text(hjust = 0.40)
    ) +
  scale_color_continuous(high ="#132B43" , low = "#56B1F7")
```




Looking at the above plot we can see several things:

- The relationship between Life Expectancy and Total Healthcare Expenditures is not linear
- Many countried live long healthy lives up to between age 70-80 with limited government expenditures.
- Those countries where more money is spent on healthcare (both government and personal) have longer life expectancies
  - There appears to be a threshold for life expectancy at 80 that can't consistently be crossed without government expenditures

**Simple Linear Regression**

We'll now run a simple regression model using these two variables:

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

Looking at the above model diagnostics, we can see that the F-statistic is 65.26. What does the F-statistic tell us? The F-test for overall significance has two hypotheses:

- The null hypothesis says that a model with no independent variables fits the data as well as your model
- The alternative hypothesis says that your model fits the data better than an intercept-only model.

We can use the degrees of freedom and the F-statistic to get the probability that a intecept-only model is better than our current model. In our case, that value is 7.714e-14, which is an extremely small number (basically zero). This gives us extreme confidence that our model fits the data better than an intercept-only model.

Next, we'll take a look at the R^2 value. Our Adjusted R-Squared value is 0.2537. This means that our model currently accounts for about 25.37% of the variability within life expectancy. This tells us that we don't have a great model and that there is a lot of variability that we aren't capturing with our current variables.

The standard error is 9.371. The standard error here tells us the typical (average) distance that data points are falling from the regression line in **units of the dependent variable**. So if a least-squares regression line is drawn, typically our data points are falling ~9 years away from that line. The standard error tells you how precise your model is using the units of your dependent variable.

We discussed the overall p-value above when we discussed the F-statistic. The p-values of the variables used in the model tell us that probability that they are **NOT** significant to the model. Here we can see both the intercept and the TotExp have incredibly small p-values, indicating that they are indeed significant to the model.

**Assumptions For Linear Regression**

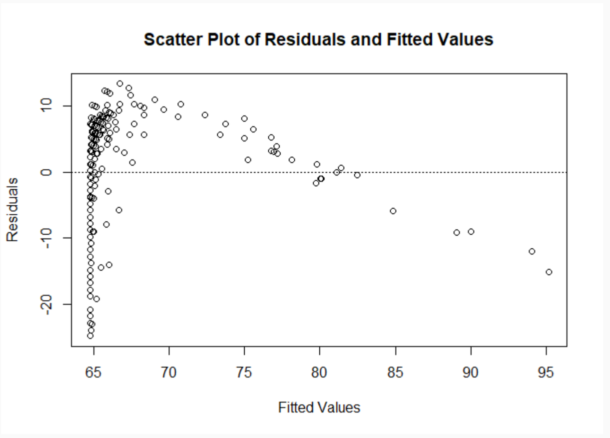
The assumptions for linear regression are:

- Linearity: The relationship between X and the mean of Y is linear
- Homoscedasticity: The variance of residuals is the same for any value of X.
- Independence: Observations are independent of each other
- Normality: For any fixed value of X, Y is normally distributed

Looking at our scatter plot above, we know that the relationship here is not linear. So we fail the first assumption for using a linear model.

Next, let's check to see if the variability in our residuals in nearly constant.

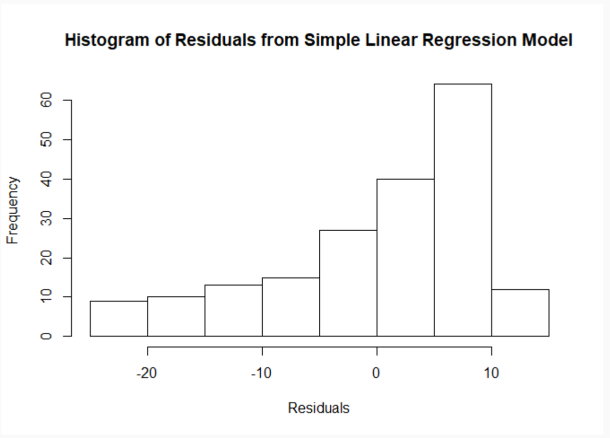
```
plot(model.lm$residuals ~ model.lm$fitted.values,
     main = "Scatter Plot of Residuals and Fitted Values",
     xlab = "Fitted Values", ylab = "Residuals")
abline(h = 0, lty = 3)
```




The variability in our residuals is not nearly constant and we can see that there is a shape/pattern to the residuals.

We can say that our observations are independent of each other. Lastly, lets check to see if our residuals are normally distributed:

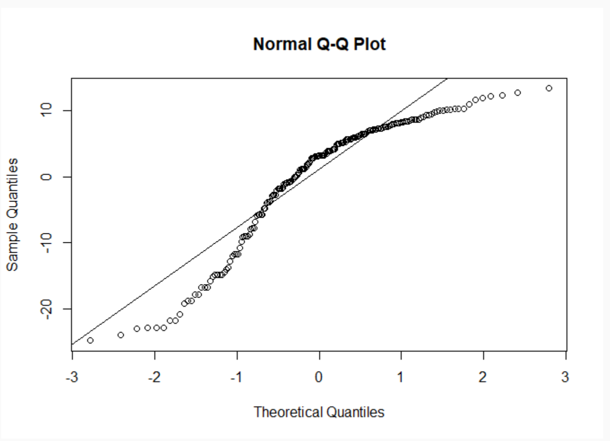
```
hist(model.lm$residuals, main = "Histogram of Residuals from Simple Linear Regression M
     xlab = "Residuals")
```




In looking at the histogram of the residuals, we can see that they are not quite *nearly* normal. The histogram is strongly left-skewed.

Let's also look at the *quantile-versus-quantile* plot (Q-Q plot).

```
qqnorm(resid(model.lm))
qqline(resid(model.lm))
```

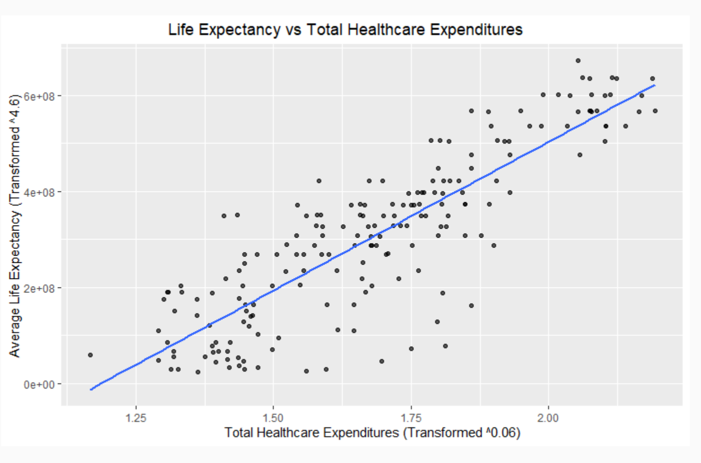



We can see again here, that our residuals are not nearly normal. In evaluating the above tests, we can say that we have not met the assumptions to use linear regression and we should use extreme caution when using our current model for any type of predictions/analysis.

**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 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?"**

```
who2 <- who %>%
  mutate(LifeExp = LifeExp^4.6) %>%
  mutate(TotExp = TotExp^0.06)
```

```
ggplot(who2) +
  aes(x = TotExp, y = LifeExp) +
  geom_jitter(alpha = 0.65) +
  geom_smooth(method = lm, se = FALSE) +
  #geom_smooth(method = lm, se = FALSE) +
  labs(title = "Life Expectancy vs Total Healthcare Expenditures",  color = "Personal E
  ylab("Average Life Expectancy (Transformed ^4.6)") +
  xlab("Total Healthcare Expenditures (Transformed ^0.06)") +
  theme(
    plot.title = element_text(hjust = 0.40),
    plot.subtitle = element_text(hjust = 0.40)
    ) +
  scale_color_continuous(high ="#132B43" , low = "#56B1F7")
```




We can see that after our transformations, the relationship between Life Expectancy and Total Healthcare Expenditures is fairly linear.

Let's now re-run our simple linear regression model with our transformed variables:

```
model.lm2 <- lm(LifeExp ~ TotExp, data = who2)
summary(model.lm2)
```

```
##
## Call:
## lm(formula = LifeExp ~ TotExp, data = who2)
##
## 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
```

Looking at the above model diagnostics, we can see that the F-Statistic is 507.7. We can use the degrees of freedom and the F-statistic to get the probability that a intercept-only model is better than our current model. In our case, that value is 2.2e-16, which is an extremely small number (basically zero). This gives us extreme confidence that our model fits the data better than an intercept-only model.

Next, we'll take a look at the R^2 value. Our Adjusted R-Squared value is 0.7283. This means that our model currently accounts for about 72.83% of the variability within life expectancy. This tells us that our model is fitting the data better than our previous model.

The standard error is 90,490,000. This is a really large number, but we raised the LifeExp column to the 4.6 power. The standard error here tells us the typical (average) distance that data points are falling from the regression line in **units of the dependent variable**. Since we performed some transformations on the original data, this value isn't super meaningful to us. To make it meaningful, we could take the 4.6 root of all our predicted values and see what the error is, however, in this case we can rely on the R^2 value to tell us about our model fit.

Here, again, we can see both the intercept and the TotExp have incredibly small p-values, indicating that they are indeed significant to the model.

Our biggest indicator that this model is performing better than the previous model is the Adjusted R-Squared value that has almost tripled.

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

Using the output from our model above, we can see that the function to estimate life expectancy is: $y = 620060216x - 736527910$. To solve the above question, we can plug 1.5 in for x in our function:

$$y = 620060216(1.5) - 736527910 = 193562414$$

These units are to the 4.6 power, so we can take the 4.6 root and get:

```
193562414 ^ (1/4.6)
```

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

Now for the next part where we plug in 2.5:

$$y = 620060216(2.5) - 736527910 = 813622630$$

Again, these units are to the 4.6 power, so we can take the 4.6 root and get:

```
813622630 ^ (1/4.6)
```

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

**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**

```
who3 <- who %>%
  mutate(PropMDxTotExp = PropMD * TotExp)

multiple_model.lm <- lm(LifeExp ~ PropMD + TotExp + PropMDxTotExp, data = who3)
summary(multiple_model.lm)
```

```
##
## Call:
## lm(formula = LifeExp ~ PropMD + TotExp + PropMDxTotExp, data = who3)
##
## 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 ***
## PropMDxTotExp -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 at the above model diagnostics, we can see that the F-statistic is 34.49. We can use the degrees of freedom and the F-statistic to get the probability that a intercept-only model is better than our current model. In our case, that value is 2.2e-16, which is an extremely small number (basically zero). This gives us extreme confidence that our model fits the data better than an intercept-only model.

Next, we'll take a look at the R^2 value. Our Adjusted R-Squared value is 0.3471. This means that our model currently accounts for about 34.71% of the variability within life expectancy. While this is not a great adjusted r-squared value, it is an improvement over our simple linear regression model where we had a value of 0.2537.

The standard error is 8.765. The standard error here tells us the average distance that data points are falling from the regression line in years.

The p-values for all of our variables as well as the intercept are extremely small (almost 0) which indicates that they are indeed significant to the model.

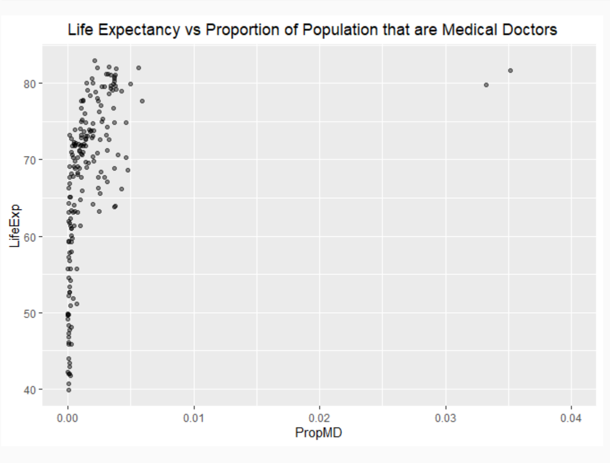
**5. Forecast LifeExp when PropMD=.03 and TotExp = 14. Does this forecast seem realistic? Why or why not?**

First we'll need to get the coefficients from the model:

```
summary(multiple_model.lm)$'coefficients'
```

```
## $coefficients
[illegible]
```

[illegible]



[illegible]

## Sources

- [illegible]
- Standard Error
- Regression Assumptions
- Model Output in R