KSooklall Homework 12 DATA 605

Kenan Sooklall

4/19/2021

who.csv dataset 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.

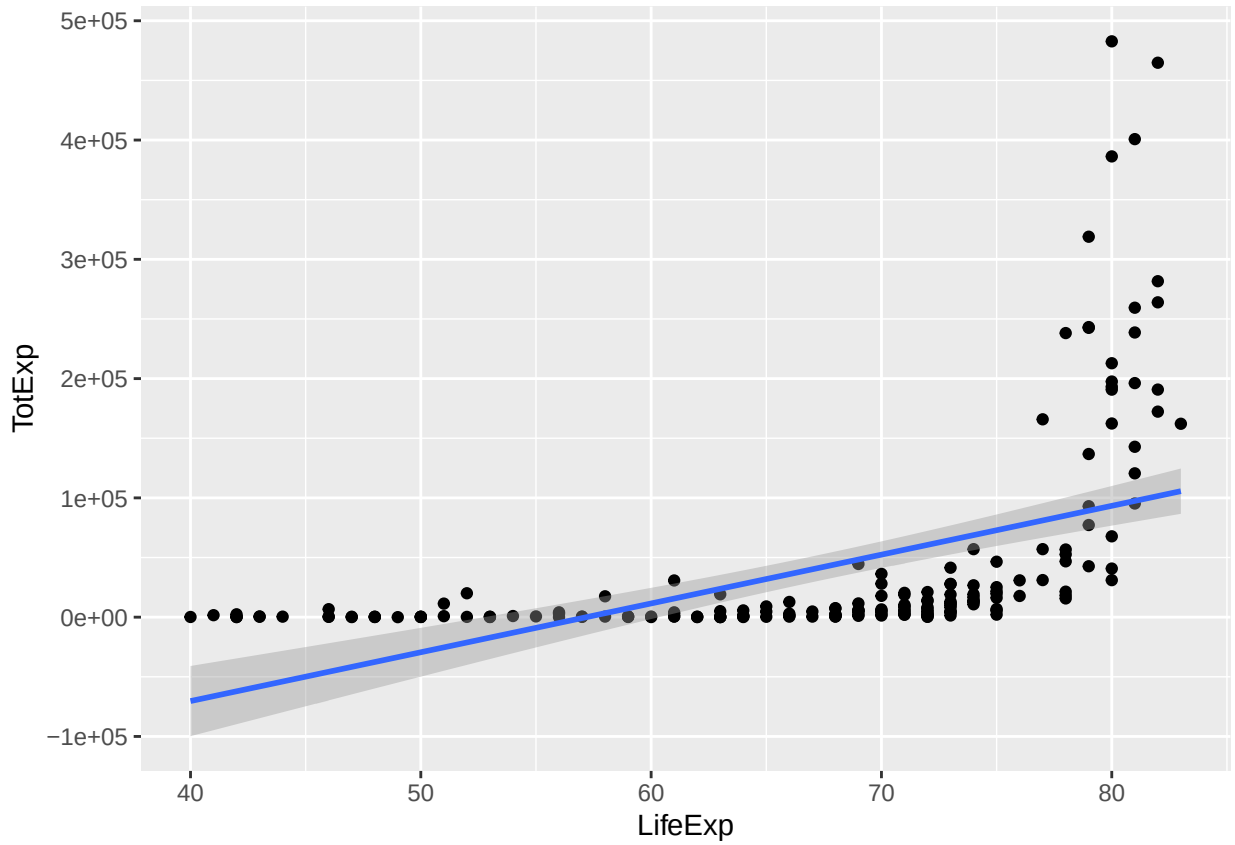
```
path = '/home/kenan/Documents/learning/masters/CUNY-SPS-Masters-DS/DATA_605/homeworks/homework12/'
df = read.csv(paste0(path, 'who.csv'))
```

Exercise 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.

```
df %>% ggplot(aes(x=LifeExp, y=TotExp)) + geom_point() +   stat_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




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

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

An R^2 of 0.2577 is very low and states there isn't a strong correlation between LifeExp and TotExp. The standard error is 7.795e-06 and the p value is very small 7.71e-14. F-statistic = 65.2 which is the ratio of two variances (SSR/SSE), the variance explained by the parameters in the model (sum of squares of regression, SSR) and the residual or unexplained variance (sum of squares of error, SSE).

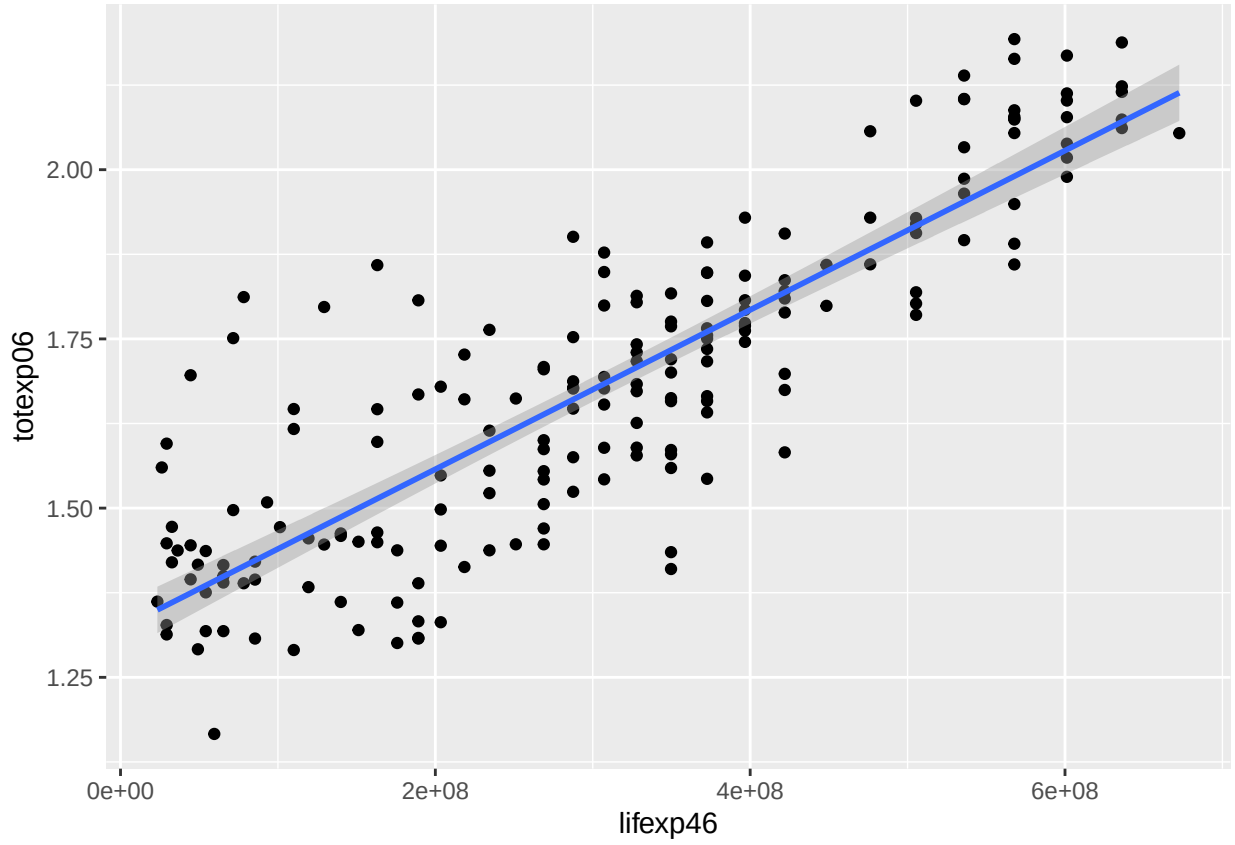
Exercise 2

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?"

```
df$lifexp46 <- df$LifeExp ^ (4.6)
df$totexp06 <- df$TotExp ^ (0.06)

df %>% ggplot(aes(x=lifexp46, y=totexp06)) + geom_point() +   stat_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
model <- lm(lifexp46~totexp06, data=df)
summary(model)
```

```
##
## Call:
## lm(formula = lifexp46 ~ totexp06, data = df)
##
## 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 ***
## totexp06     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
```

Very surprising results. The R^2 tripled making the relationship look more linear. The F-statistic and the std error also went up as well with the P-vaues about the same ~0. This model is "better" than the last.

Exercise 3

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

```
totalexp06 = c(1.5, 2.5)
yhat1 = predict(model, data.frame(totexp06=totalexp06[1]))
print(yhat1)
```

```
##         1
## 193562414
```

```
yhat2 = predict(model, data.frame(totexp06=totalexp06[2]))
print(yhat2)
```

```
##         1
## 813622630
```

Exercise 4

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

$LifeExp = b0 + b1xPropMd + b2xTotExp + b3xPropMDxTotExp$

```
model = lm(LifeExp ~ PropMD + TotExp + (PropMD * TotExp), data=df)
summary(model)
```

```
##
## Call:
## lm(formula = LifeExp ~ PropMD + TotExp + (PropMD * TotExp), data = df)
##
## 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
```

This model is more complex since there are more variables. The statistic values are better than the model from exercise 2 but worst than exercise 3.

Exercise 5

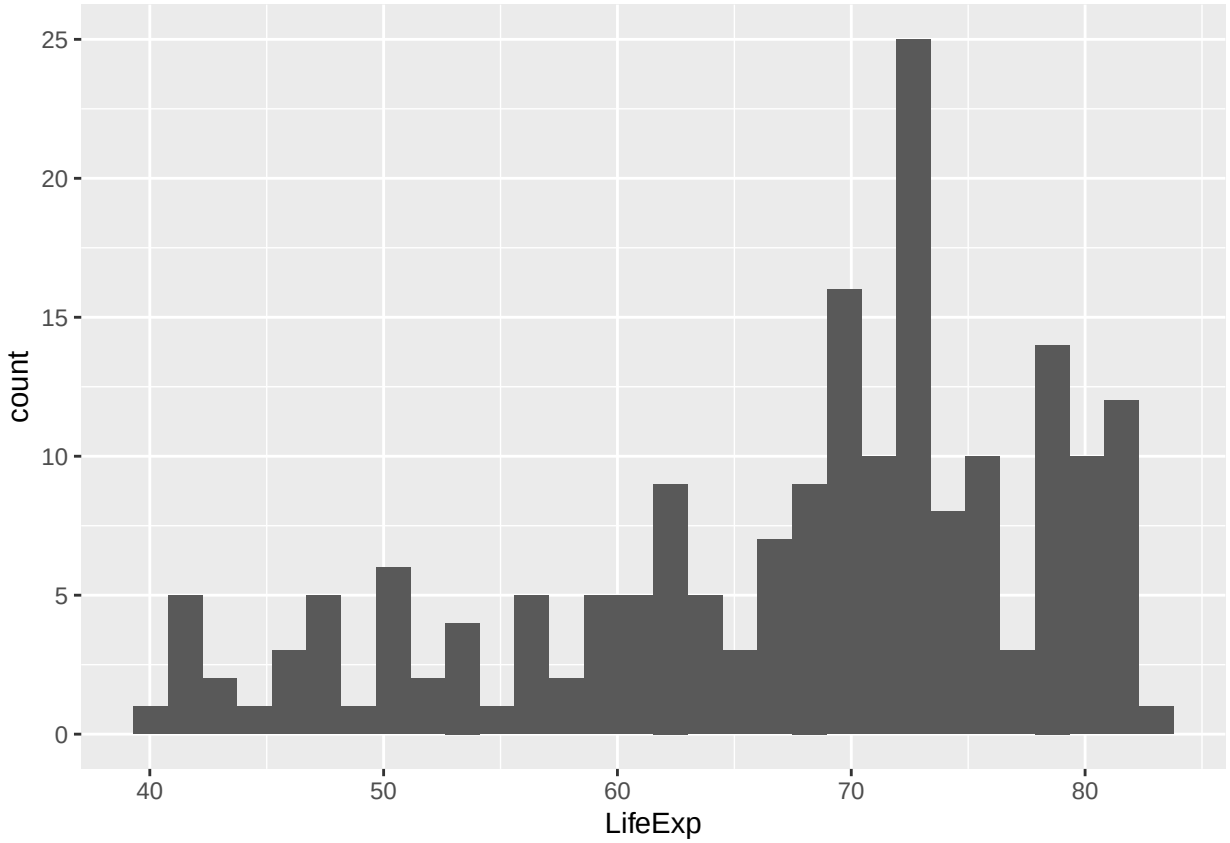
Forecast LifeExp when PropMD=.03 and TotExp = 14. Does this forecast seem realistic? Why or why not?

```
print(predict(model, data.frame(PropMD=0.03, TotExp=14)))
```

```
##       1
## 107.696
```

```
df %>% ggplot(aes(LifeExp)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The forecast is realistic, many have lived to 108 years old; however that is very far above the average in the real world and would be an extrapolation value in this data set.