PShaji_Assignment12

Priya Shaji

11/15/2019

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?

Load the dataset

```
who_df <- read.csv("https://raw.githubusercontent.com/PriyaShaji/Data605/master/week%2012/who.csv")
head(who_df)
```

	Country	LifeExp	InfantSurvival	Under5Survival	TBFree	PropMD	PropRN
	<fctr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	Afghanistan	42	0.835	0.743	0.99769	0.000228841	0.000572294
2	Albania	71	0.985	0.983	0.99974	0.001143127	0.004614439
3	Algeria	71	0.967	0.962	0.99944	0.001060478	0.002091362
4	Andorra	82	0.997	0.996	0.99983	0.003297297	0.003500000
5	Angola	41	0.846	0.740	0.99656	0.000070400	0.001146162
6	Antigua and Barbuda	73	0.990	0.989	0.99991	0.000142857	0.002773810

6 rows | 1-8 of 11 columns

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.

Answer

```
attach(who_df)
```

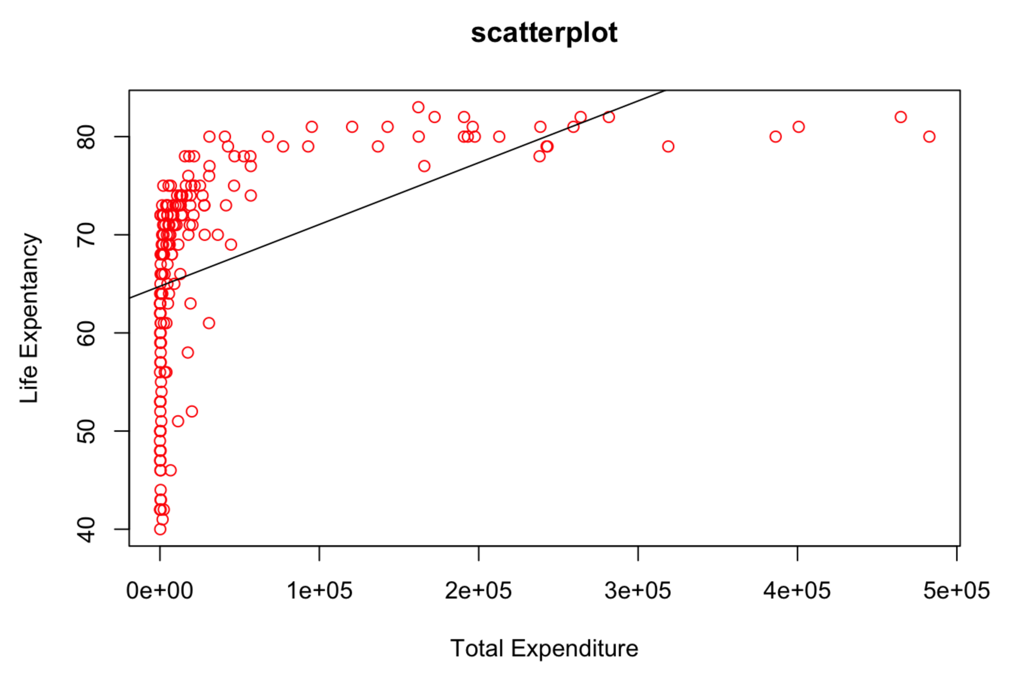
Check for correlation between 2 variables

```
cor(LifeExp,TotExp)
```

```
## [1] 0.5076339
```

Plot a scatterplot

```
plot(TotExp,  LifeExp, main='scatterplot', ylab='Life Expentancy', xlab = 'Total Expenditure', col=2)
abline(lm(LifeExp~TotExp), col=1)
```



Run a simple linear regression

```
linear_regression = lm(LifeExp~TotExp)
linear_regression
```

```
##
## Call:
## lm(formula = LifeExp ~ TotExp)
##
## Coefficients:
## (Intercept)       TotExp
##   6.475e+01    6.297e-05
```

Summarise the coefficients

```
summary(linear_regression)
```

```
##
## 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
```

Linear Regression Model Summary:

Life Expentancy = 64.75 + .000063 * Total Expenditure

The model above shows a negative y intercept (Total Expenditure on healtcare). Which means the model would give negative Total Expenditure if life expendtacy is less than ~65. The model at the onset, is not realistic. The model reflects the very small amount low Life Expentancy countries spend on healthcare when compared to the Total Expentiture by high Life Expentancy countries.

Multiple R-squared: 0.2577, Adjusted R-squared: 0.2537 - The low R-squared value tells us that our model only explains around 25% of the response variable (Life expentancy in response to Total Expenditure) around the mean.

F-statistic: 65.26 on 1 and 188 DF, p-value: 7.714e-14 - the p-value of the model is really low which means we can confindetly reject the null hypothesis (that Total Expenditure DOES NOT contribute to a country’s Life Expentancy). We can say that the variable does contribute to the model, its ony a minor contributor.

Residual standard error: 9.371 on 188 degrees of freedom - 9.371 Residual standard error also tells us the SE is somewhat high (about 10 man years). This means that some the sample data points are significantly off the fitted line. This means that countries who contribute significantly less in healthcare expenditure than what the model would predict, have nonetheless sustain a life expectancy that is significantly higher than expected.

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?”

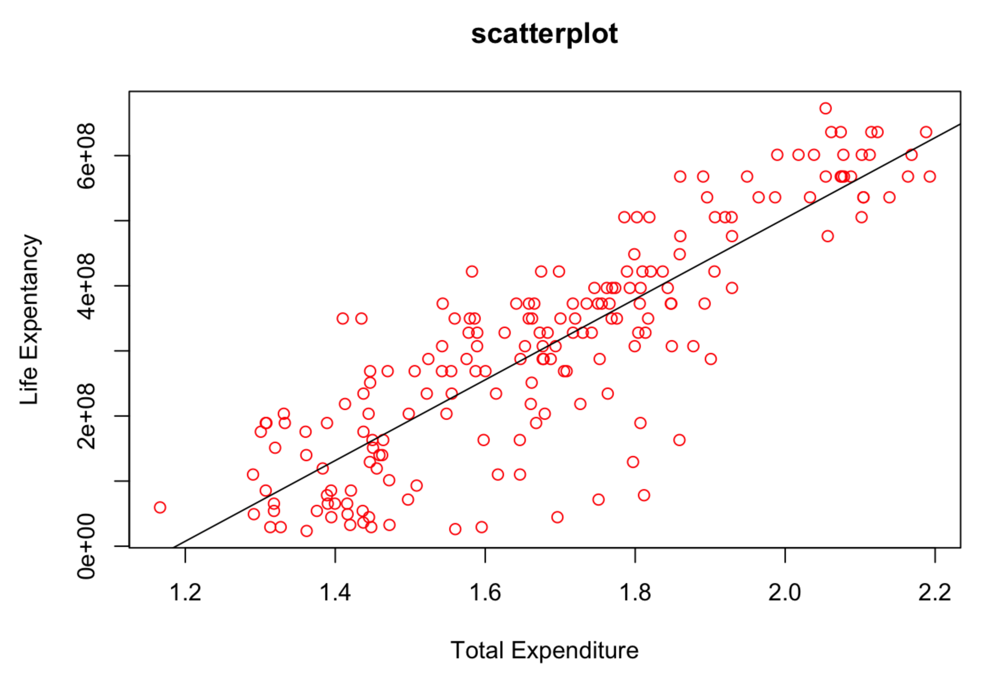
Answer

Check for correlation between two variables

```
TotExp2 = TotExp^0.06
LifeExp2 = LifeExp^4.6
cor(LifeExp2,TotExp2)
```

```
## [1] 0.8542642
```

```
plot(TotExp2, LifeExp2, main='scatterplot', ylab='Life Expentancy', xlab = 'Total Expenditure', col=2)
abline(lm(LifeExp2~TotExp2), col=1)
```




```
linear_regression_second = lm(LifeExp2~TotExp2)
linear_regression_second
```

```
##
## Call:
## lm(formula = LifeExp2 ~ TotExp2)
##
## Coefficients:
## (Intercept)      TotExp2
##  -736527909    620060216
```

```
summary(linear_regression_second)
```

```
##
## Call:
## lm(formula = LifeExp2 ~ TotExp2)
##
## 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 ***
## TotExp2      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
```

Linear Regression Model Summary:

Life Expectancy^4.6 = -736527909 + 620060216 * Total Expenditure^0.06

By looking at the regression line for this transformed model and comparing it against the previous model, I can say that the transformed model is the better model since the data points are more closely clustered around the regression line of the model.

Multiple R-squared: 0.7298, Adjusted R-squared: 0.7283 R-squared value of close to 73% is much better than the ~26% R-squared value for the first model. This means that the response variable (life expentancy^4.6) explains the model’s variability around the mean 75% of the time.

F-statistic: 507.7 on 1 and 188 DF, p-value: < 2.2e-16 - the p-value of the model is really low which means we can confindetly reject the null hypothesis (that Total Expenditure^0.06 DOES NOT contribute to a country’s Life Expentancy^4.6). We can say that the variable does contribute to the model, in a greater way than the orignial model.

Residual standard error: 90,490,000 on 188 degrees of freedom Suprising high Residual SE even when we consider that life Expentancy was increased expontially by 4.6. This contradicts the R-squared and F-statistics finding but since the original scatterplot does show that countries with low life expectancy have even much lower Total Expenditures. Since we increase these values exponentially, the SE should would also increase exponentially.

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

Answer

Linear Regression Model Summary:

Life Expectancy^4.6 = 64.75 + 620060216 * Total Expenditure^0.06

```
LifeExp_46 = -736527909 +  620060216 * (1.5)
LifeExp_15 = exp(log(LifeExp_46)/4.6)
LifeExp_15
```

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

```
LifeExp_46 = -736527909 +  620060216 * (2.5)
LifeExp_25 = exp(log(LifeExp_46)/4.6)
LifeExp_25
```

```
## [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

Answer

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

```
multiple_regression = lm(LifeExp~TotExp + PropMD + PropMD * TotExp)
multiple_regression
```

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

```
summary(multiple_regression)
```

```
##
## Call:
## lm(formula = LifeExp ~ TotExp + PropMD + 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 ***
## 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
```

Life Expentancy MR = 62.8 + .000072 Total Expenditure + 1,497 PropMD + .006 * Total Expenditrure * PropMD

Multiple R-squared: 0.3574, Adjusted R-squared: 0.3471 - with an adjusted R-squared value of only ~35%, this is not a good model. This means that the response variables in this model account for only ~35% of the variability of the model.

F-statistic: 34.49 on 3 and 186 DF, p-value: < 2.2e-16 the F-statistic shows that the p-value is really low (close to zero), which means we can reject the null hypothesis and state with confidence that the response variables do contribute to the true value of the dependent variable.

Residual standard error: 8.765 on 186 degrees of freedom - The residual SE is significant at 8.765. Which means that datapoints on the average are off by 8.765 from what the model would have predicted. By this measure, I would have to say the model is not a good fit to its corresponding data points.

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

Answer

```
Life_Expectancy = 62.8 + .000072 * 14 + 1497 * 0.03 + .006 * 14 * 0.03
Life_Expectancy
```

```
## [1] 107.7135
```

The above forecast is not realistic. It summarises that if we increse the proportion of doctors in the population and drastically reduce spending, we can dramatically increase life expectancy from ~80s (high life expectancy countries) to 107. Since proportion of Doctors is not independent of Total Expenditure in healthcare. Huge amount of money is spent to train good doctors and good doctors also expect to be well compensated. Thus, it is not realistic to have a drastic increase in doctors in a population and at the same time have a drastic decrease in healthcare spending. The Total Expenditure came to be as 14, which is too low a number for Total Expenditure even for countries that have a very expensive and inefficient health care systems. The US, for example, spends more for healthcare per capita than any other country at around 7,000$ per capita. To drastically reduce this to 14$ per capita and expect to have a surge in medical doctors (x1,000 to x10,000) would not make sense.