605 HW 11

Tenzin

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.

Load Data

```
data <- read.csv("https://raw.githubusercontent.com/tenzinda97/Discussion-11/main/who.csv")

summary(data)
```

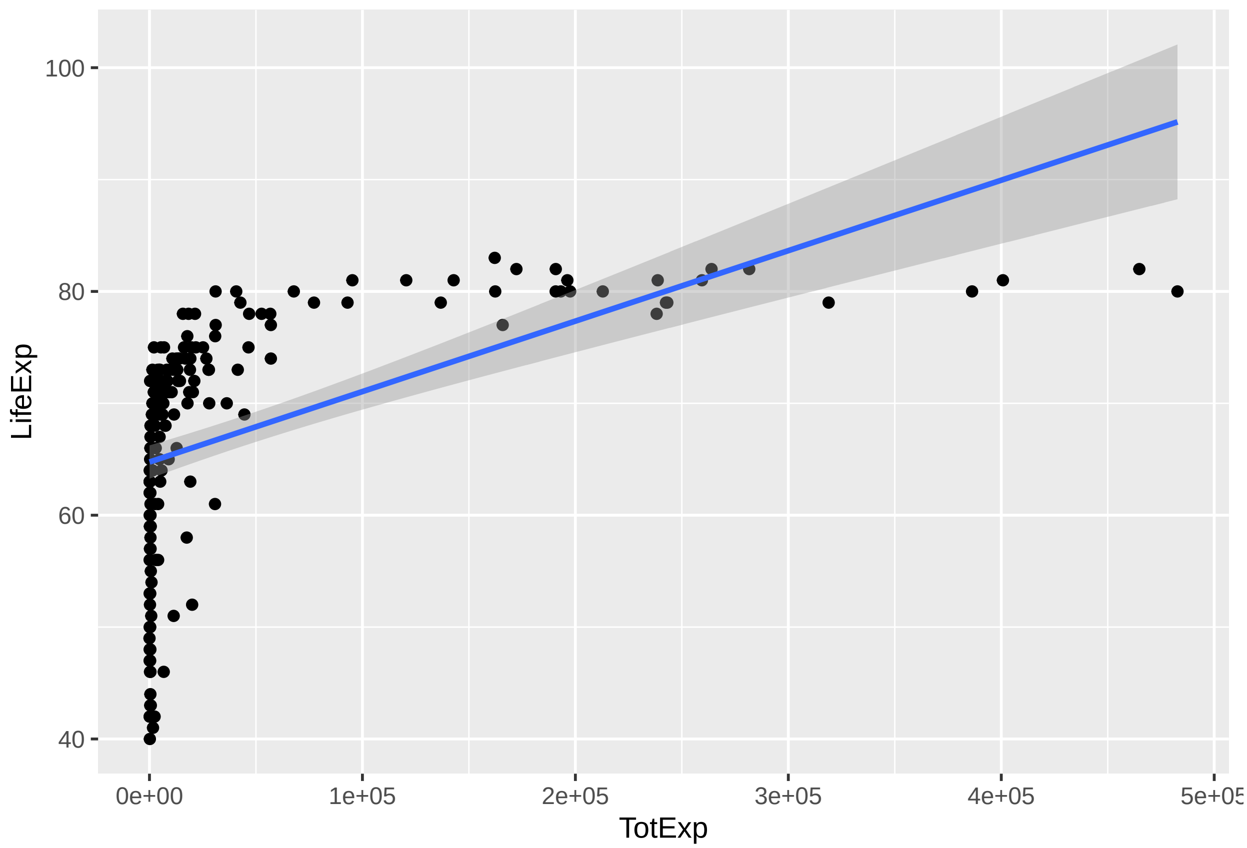
```
##    Country             LifeExp      InfantSurvival   Under5Survival
##  Length:190         Min.   :40.00   Min.   :0.8350   Min.   :0.7310
##  Class :character   1st Qu.:61.25   1st Qu.:0.9433   1st Qu.:0.9253
##  Mode  :character   Median :70.00   Median :0.9785   Median :0.9745
##                     Mean   :67.38   Mean   :0.9624   Mean   :0.9459
##                     3rd Qu.:75.00   3rd Qu.:0.9910   3rd Qu.:0.9900
##                     Max.   :83.00   Max.   :0.9980   Max.   :0.9970
##      TBFree           PropMD              PropRN             PersExp
##  Min.   :0.9870   Min.   :0.0000196   Min.   :0.0000883   Min.   :   3.00
##  1st Qu.:0.9969   1st Qu.:0.0002444   1st Qu.:0.0008455   1st Qu.:  36.25
##  Median :0.9992   Median :0.0010474   Median :0.0027584   Median : 199.50
##  Mean   :0.9980   Mean   :0.0017954   Mean   :0.0041336   Mean   : 742.00
##  3rd Qu.:0.9998   3rd Qu.:0.0024584   3rd Qu.:0.0057164   3rd Qu.: 515.25
##  Max.   :1.0000   Max.   :0.0351290   Max.   :0.0708387   Max.   :6350.00
##     GovtExp             TotExp
##  Min.   :    10.0   Min.   :    13
##  1st Qu.:   559.5   1st Qu.:   584
##  Median :  5385.0   Median :  5541
##  Mean   : 40953.5   Mean   : 41696
##  3rd Qu.: 25680.2   3rd Qu.: 26331
##  Max.   :476420.0   Max.   :482750
```

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.

```
data |> ggplot(aes(x = TotExp, y = LifeExp)) +
  geom_point() +
  geom_smooth(method = "lm")
```

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




```
model1 <- lm(LifeExp ~ TotExp, data = data)
summary(model1)
```

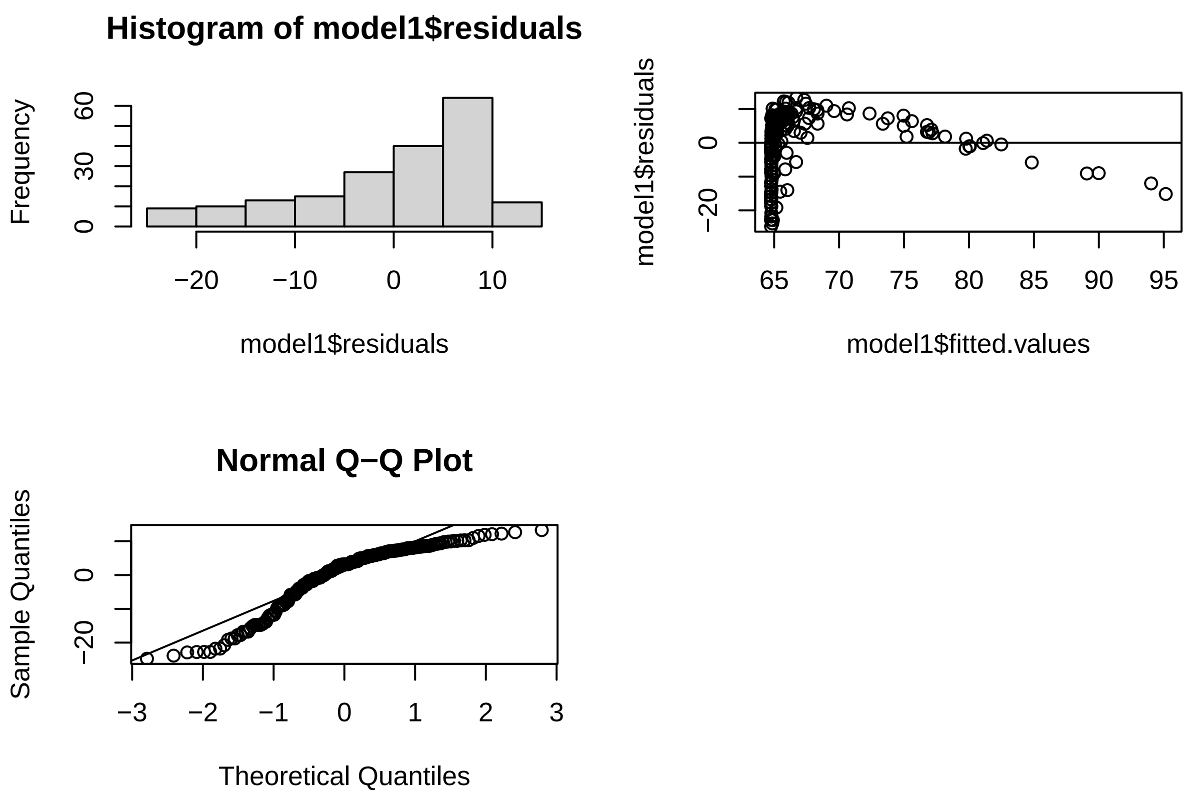
```
##
## Call:
## lm(formula = LifeExp ~ TotExp, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -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
```

The linear regression model has an F-statistic of 65.26, and a p-value of 7.714e-14, which means that the model is statistically significant and we would reject the null hypthesis that there isn't a relationship. The R-Squared value of 0.2577 indicates that 25.77% of the variance in life expectancy can be explained by total expenditures. The adjusted R-Squared value of 0.2537 is similar to the R-Squared value, indicating that the model is not overfitting the data, this is due to only plotting one variable against another. The scatterplot shows that while there is a clear positive relationship between life expectancy and total expenditures, the relationship is not linear. This also explains the discrepancy between the F-statistic and p-values showing a clear relationship and the R-Squared value being relatively low at around 25%. Additionally, the residual standard error of 9.371 seems a bit high for a variable that has a range of 40-83 with a mean of 67.38 but we would have to see if we can improve on that later.

The assumptions of linear regression are linearity, independence, homoscedasticity, and normality. We will check these assumptions below.

```r
par(mfrow=c(2,2))
hist(model1$residuals)
plot(model1$fitted.values, model1$residuals)
abline(h=0)
qqnorm(model1$residuals)
qqline(model1$residuals)
```



The histogram of the residuals and the Q-Q plot show that the residuals are not normally distributed, thus violating the assumption of normality. The residuals vs fitted values plot shows that the residuals are not homoscedastic, violating the assumption of homoscedasticity. We have previosuly shown that the relationship between life expectancy and total expenditures is not linear, thus violating the assumption of linearity. The residuals vs fitted values plot also shows that the residuals are not independent, violating the assumption of independence. We can concluse that the linear regression model is not a good fit for the data.

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, TotExpˆ.06). Plot LifeExp^4.6 as a function of TotExpˆ.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?”

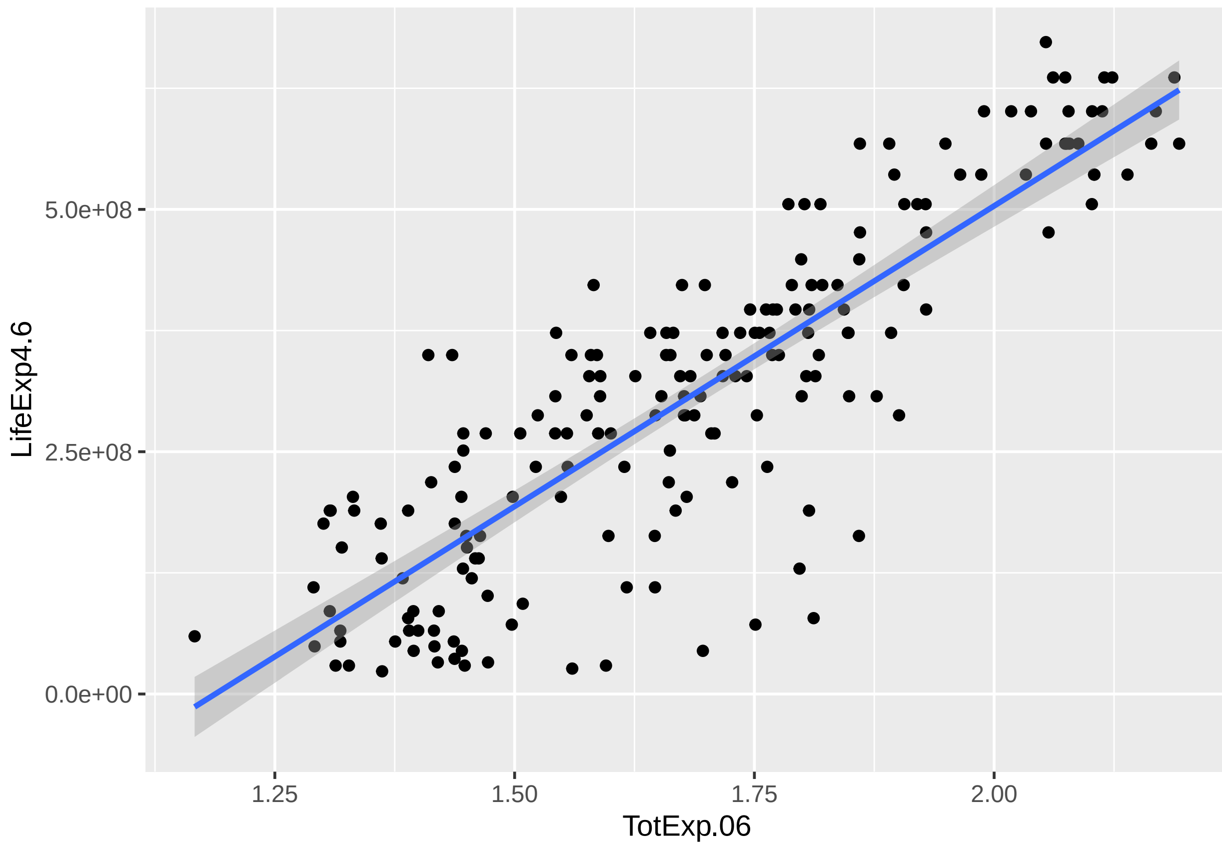
```
data <- data |>
  mutate(LifeExp4.6 = LifeExp^4.6,
         TotExp.06 = TotExp^0.06)

summary(data$LifeExp4.6)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##  23414019 166291095 307221061 307957212 421970229 672603658
```

```
data |> ggplot(aes(x = TotExp.06, y = LifeExp4.6)) +
  geom_point() +
  geom_smooth(method = "lm")
```

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




```
model2 <- lm(LifeExp4.6 ~ TotExp.06, data = data)
summary(model2)
```

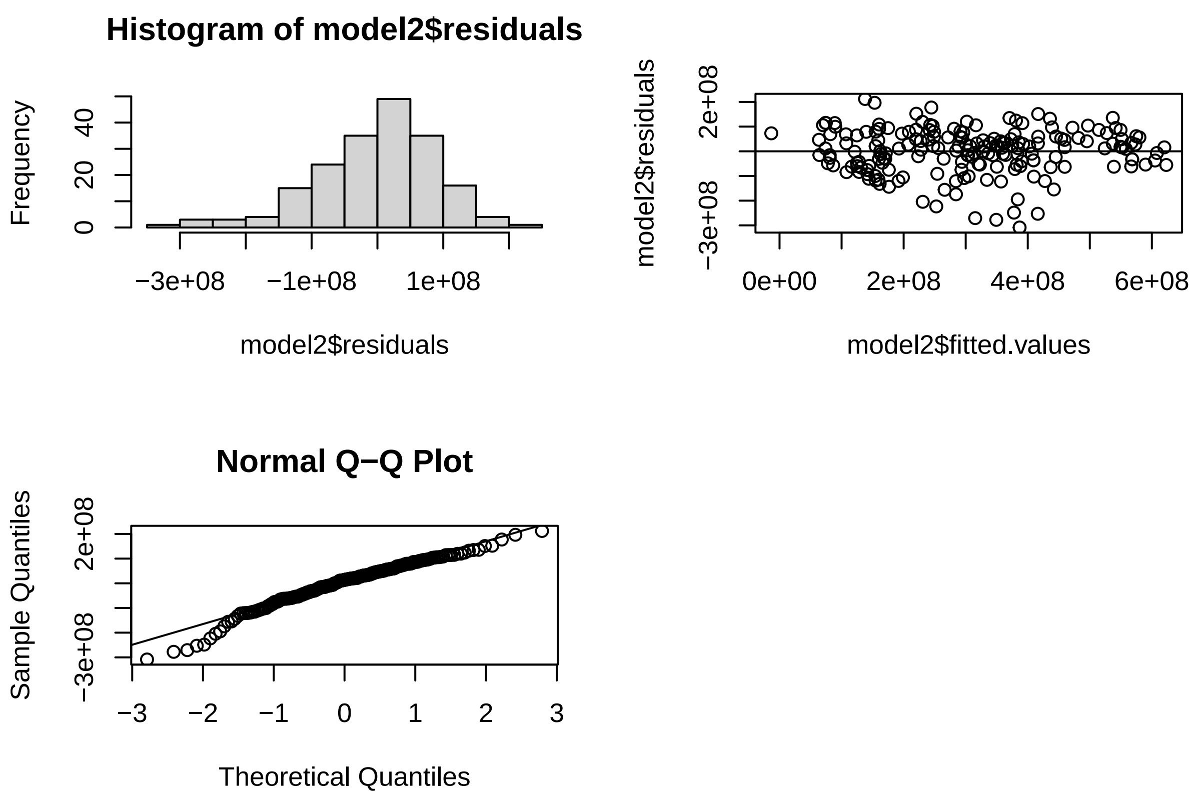
```
##
## Call:
## lm(formula = LifeExp4.6 ~ TotExp.06, data = data)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -308616089  -53978977   13697187   59139231  211951764
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -736527910   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
```

The linear regression model has an F-statistic of 507.7, and a p-value of < 2.2e-16, both much better than the previous model, implying strong statistical significance. The R-Squared value of 0.7298 indicates that 72.98% of the variance in life expectancy can be explained by total expenditures. The adjusted R-Squared value of 0.7283 is similar to the R-Squared value, indicating that the model is not overfitting the data. This is a huge improvement over the previous model, which only explained 25.77% of the variance in life expectancy. The scatterplot shows that there is a clear positive relationship between life expectancy and total expenditures, and the relationship is much more linear than before. The minimum/maximum of the transformed Life Expectancy variable are 23414019/672603658 and the mean is 307221061. Based off these values, the residual standard error of 90490000 seems reasonable.

We will now check the assumptions of the linear regression model.

```r
par(mfrow=c(2,2))
hist(model2$residuals)
plot(model2$fitted.values, model2$residuals)
abline(h=0)
qqnorm(model2$residuals)
qqline(model2$residuals)
```



The histogram and the Q-Q plot of the residuals show that the residuals are pretty normally distributed, albeit with a bit of a skew. The residuals vs fitted values plot shows that the residuals are homoscedastic, and the residuals are independent. The assumptions of linearity, independence, homoscedasticity, and normality are met. We can conclude that the transformed linear regression model is a much better fit for the data. However, the transformation of the variables makes it difficult to interpret the coefficients of the model and how to make predictions on new data.

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

```
new_data <- data.frame(TotExp.06 = c(1.5, 2.5))
predict_LifeExp4.6 <- predict(model2, newdata = new_data)

#transform the predictions back to the original scale
predict_LifeExp <- predict_LifeExp4.6^(1/4.6)
predict_LifeExp
```

```
##        1        2
## 63.31153 86.50645
```

My predictions for the given values are a life expectancy of 63.31153 for a TotExp^.06 of 1.5 and 86.50645 for a TotExp^.06 of 2.5. It is very important to note that the second case an extrapolation as the highest value in the dataset achieved a TotExp^.06 of 2.193 based off the highest LifeExp being 80. The model may not be accurate for values outside of the range of the data.

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

```
model3 <- lm(LifeExp ~ PropMD + TotExp + PropMD*TotExp, data = data)
summary(model3)
```

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

The multiple linear regression model has an F-statistic of 34.49, and a p-value of < 2.2e-16, which means that the model is statistically significant and we would reject the null hypthesis that there isn't a relationship. The R-Squared value of 0.3574 indicates that 35.74% of the variance in life expectancy can be explained by the predictors. The adjusted R-Squared value of 0.3471 is similar to the R-Squared value, indicating that the model is not overfitting the data. The residual standard error of 8.765 seems reasonable for a variable that has a range of 40-83 with a mean of 67.38. The p-values for the coefficients of the model are all less than 0.05, indicating that the predictors are statistically significant. The model is an improvement over the simple linear regression model, which only explained 25.77% of the variance in life expectancy. However, the model is not as good as the transformed linear regression model, which explained 72.98% of the variance in life expectancy. However, since it doesn't rely on transformed variables, it is easier to interpret the coefficients and make predictions on new data. The decision on which model to use would depend on the specific use case and the importance of interpretability vs accuracy. The simple non-transformed linear model is the easiest to read and interpret, but the least accurate. The multiple linear regression model is almost as simple with a significant increase in accuracy. The transformed linear regression model is by far the most accurate, but the most difficult to interpret.

5.

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

```r
new_data2 <- data.frame(PropMD = 0.03, TotExp = 14)
predict_LifeExp <- predict(model3, newdata = new_data2)
predict_LifeExp
```

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

The predicted life expectancy for the given data points based on the model from part 4 is 107.696. This does not seem like a realistic prediction. The highest life expectancy in the dataset is 83. While both values given as predictors are within the bounds of our training data, they are both on the extreme ends of the boundaries, where the model may not be as accurate. As noted in part 3, predicting values outside of the range of the data can lead to inaccurate predictions. It is also important to remember that this specific model is not nearly as accurate as some other models we can build including the model from part 2.