Data 605 Homework 12

Jean Jimenez

2023-11-15

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(ggfortify)
```

```
## Warning: package 'ggfortify' was built under R version 4.3.2
```

```
whodata=read.csv(url("https://raw.githubusercontent.com/sleepysloth12/data605_hw12/main/who.csv"))
```

#1

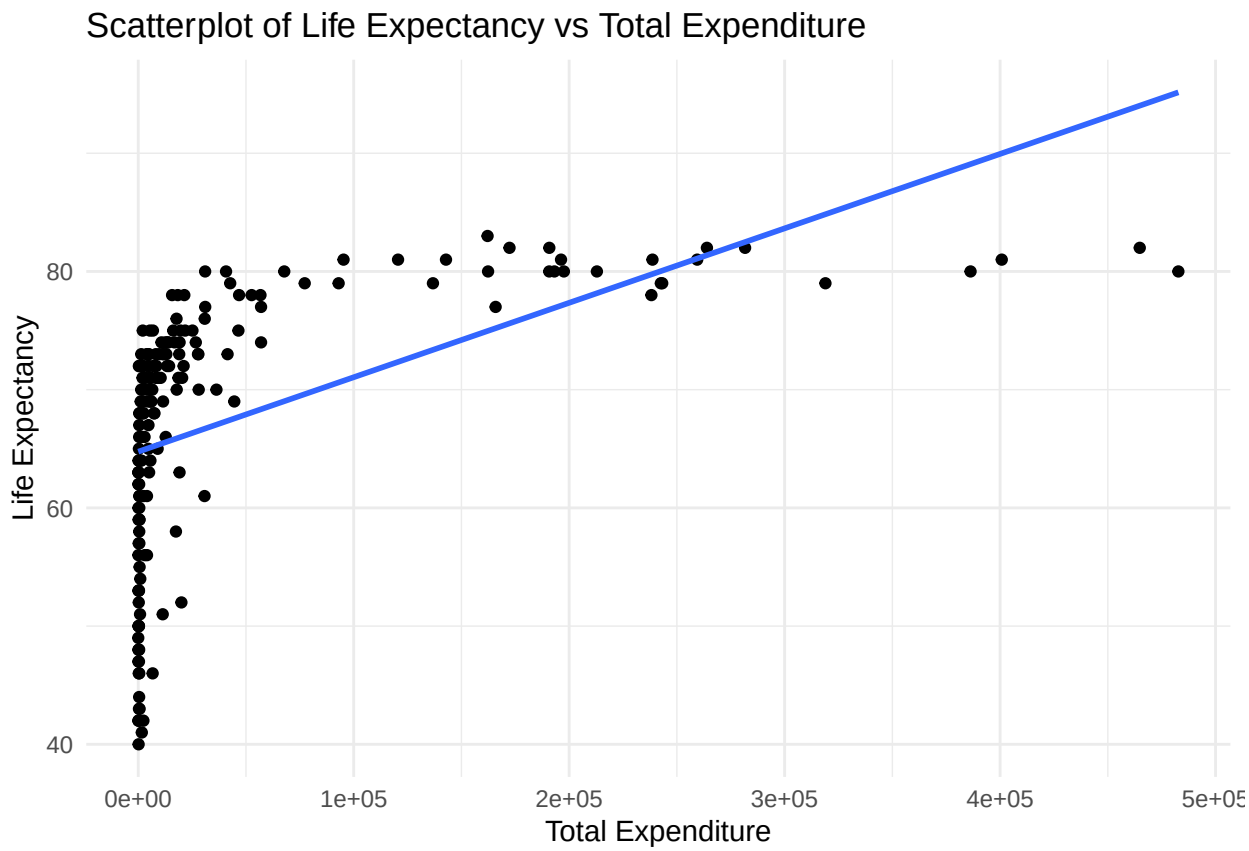
Question

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.

Work and Answer

```
#scatterplot
ggplot(whodata, aes(x = TotExp, y = LifeExp)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_minimal() +
  labs(title = "Scatterplot of Life Expectancy vs Total Expenditure",
       x = "Total Expenditure",
       y = "Life Expectancy")
```

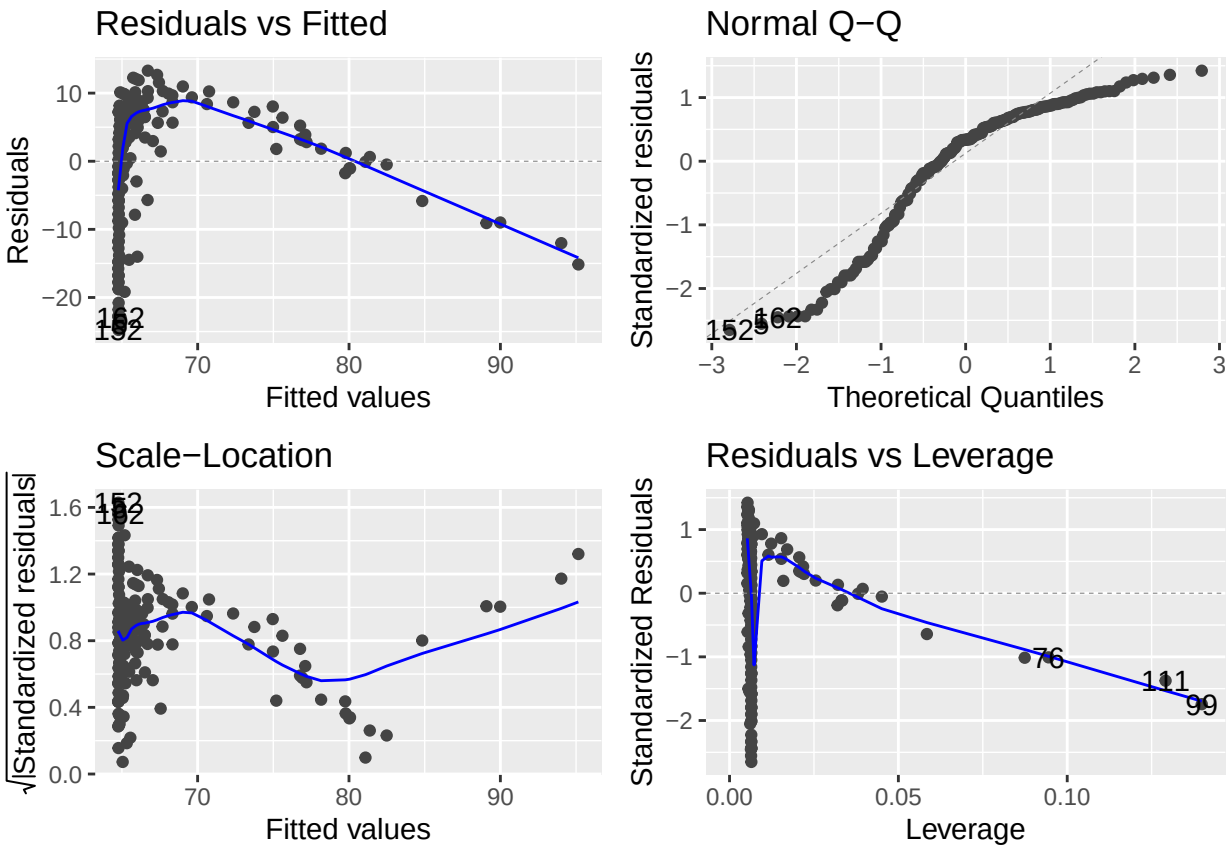
```
## `geom_smooth()` using formula = 'y ~ x'
```




```
lifexp_model = lm(LifeExp ~ TotExp, data = whodata)

summary(lifexp_model)
```

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

```
autoplot(lifexp_model)
```



Results of Linear Models:

The intercept coefficient is ~64.75. There is a standard error of about 0.7535. It is statistically significant with a p-value of <2e-16.

The slope for `TotExp` is ~0.00006297 with a standard error of 0.000007795. The p-vaue is statistically significant (7.71e-14). There is a positive relationship between `TotExp` and `LifeExp`.

The residual standard error is ~9.371 meaning the actual life expectancy values are about 9.371 years away from the predicted value by the model.

The R^2 value is 0.2577, suggesting that approximately 25.77% of the variability in `LifeExp` is explained by the model.

The F-statistic is 65.26 with a p-value of about 7.714e-14. This suggests that the model as a whole is significant, and there is a relationship between `TotalExp` and `LifeExp`.

Assumptions of Regression:

The Residuals vs Fitted plot shows a clear pattern, which mean that the relationship between the variables may not be purely linear, or that there are other variables affecting the relationship that are not included.

The Normal Q-Q plot shows that the residuals deviate from the line at the tails, meaning the residuals may not be normally distributed.

The Scale-Location plot shows that the residuals spread differently along the range of predictors. This is heteroscedasticity and implies that the variance of the residuals is not constant.

The Residuals vs Leverage plot shows that there are no individual points are exerting too much of an influence on the regression model.

Some of the assumptions of linear regression are not fully met. The clear pattern in the residuals and the heteroscedasticity suggest that a linear model may not be the best fit for this data and we have to do something else.

#2

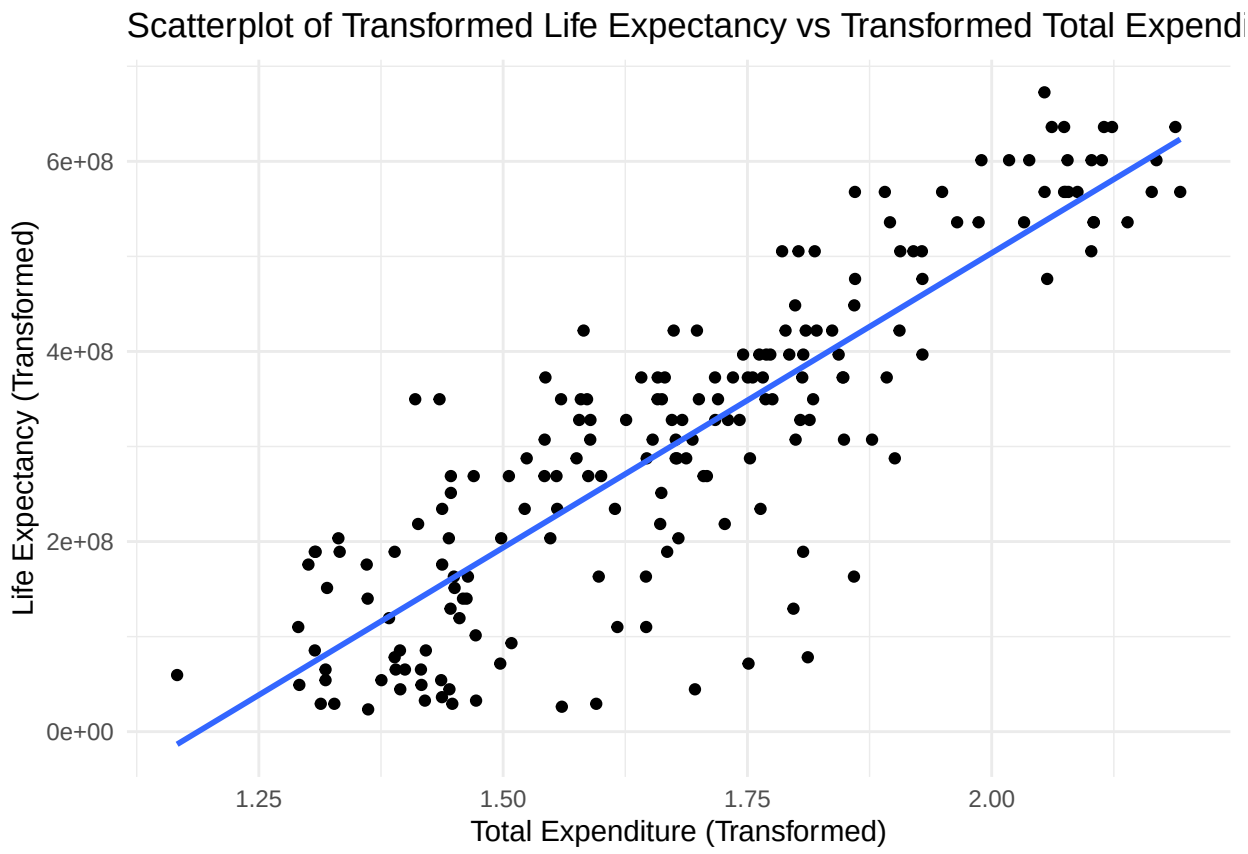
Question

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?"

Work and Answer

```
#transform
whodata$LifeExp_trans = whodata$LifeExp^4.6
whodata$TotExp_trans = whodata$TotExp^0.06

#scatter plo of transformed

ggplot(whodata, aes(x = TotExp_trans, y = LifeExp_trans)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_minimal() +
  labs(title = "Scatterplot of Transformed Life Expectancy vs Transformed Total Expenditure",
       x = "Total Expenditure (Transformed)",
       y = "Life Expectancy (Transformed)")
```

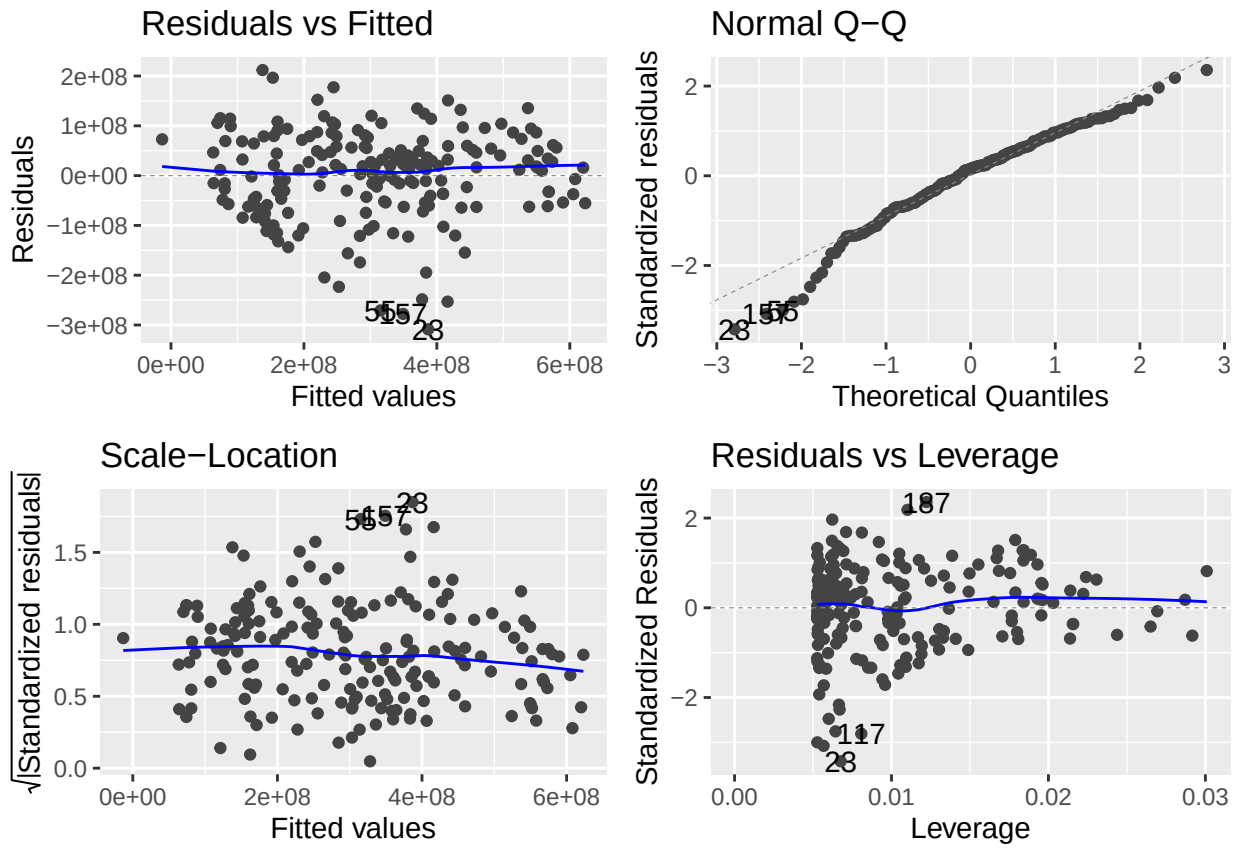
```
## `geom_smooth()` using formula = 'y ~ x'
```




```
#transformed model

trans_model=lm(LifeExp_trans ~ TotExp_trans, data = whodata)

summary(trans_model)
```

```
##
## Call:
## lm(formula = LifeExp_trans ~ TotExp_trans, data = whodata)
##
## 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_trans  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
```

```
autoplot(trans_model)
```




```
#original R squared
original_r_squared = summary(lifexp_model)$r.squared
original_rse = summary(lifexp_model)$sigma

#trans rsquared
trans_r_squared = summary(trans_model)$r.squared
trans_rse = summary(trans_model)$sigma

list(
  original_model = list(
    r_squared = original_r_squared,
    rse = original_rse
  ),
  trans_model = list(
    r_squared = trans_r_squared,
    rse = trans_rse
  )
)
```

```
## $original_model
## $original_model$r_squared
## [1] 0.2576922
##
## $original_model$rse
```

```
## [1] 9.371033
##
##
## $trans_model
## $trans_model$r_squared
## [1] 0.7297673
##
## $trans_model$rse
## [1] 90492393
```

The new intercept coefficient is ~ -736,527,910 with a standard error of ~46,817,945, which is significant with a p-value less than 2e-16.

The slope coefficient for `TotExp_trans` is ~620,060,216 with a standard error of ~27,518,940. This coefficient is also highly significant ($p < 0.001$), indicating a strong positive relationship between the transformed Total Expenditure and transformed Life Expectancy.

The residual standard error is ~90,490,000. This is probably due to the scale of the transformed Life Expectancy variable.

TheR^2 value is 0.7298 meaning ~ 72.98% of the variability in the transformed Life Expectancy is explained by the model.

The F-statistic is 507.7 with a p-value of less than 2.2e-16, which means that the model is highly significant.

When comparing the transformed model to the original, the transformed model has a substantially higher R^2 value (0.7298 vs 0.2577), indicating a better fit to the data as it explains more of the variance in Life Expectancy. Therefore, I think that the transformed model is "better" in terms of the proportion of the variance explained.

#3

Question

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

Work and Answer

```
life_exp_1_5 = predict(trans_model, newdata = data.frame(TotExp_trans = 1.5))

life_exp_2_5 = predict(trans_model, newdata = data.frame(TotExp_trans = 2.5))

forecast_1_5 = life_exp_1_5^(1/4.6)
forecast_2_5 = life_exp_2_5^(1/4.6)

# Return the forecasts
list(
  life_exp_when_TotExp_1_5 = forecast_1_5,
  life_exp_when_TotExp_2_5 = forecast_2_5
)
```

```
## $life_exp_when_TotExp_1_5
##        1
## 63.31153
##
## $life_exp_when_TotExp_2_5
```

```
##        1
## 86.50645
```

The results indicate that:

- When `TotExp^.06` is equal to 1.5, the forecasted life expectancy is approximately 63.31 years.
- When `TotExp^.06` is equal to 2.5, the forecasted life expectancy is approximately 86.51 years.

#4

Question

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

Work and Answer

```
mult_model = lm(LifeExp ~ PropMD + TotExp + PropMD:TotExp, data = whodata)

summary_mult = summary(mult_model)

print(summary_mult)
```

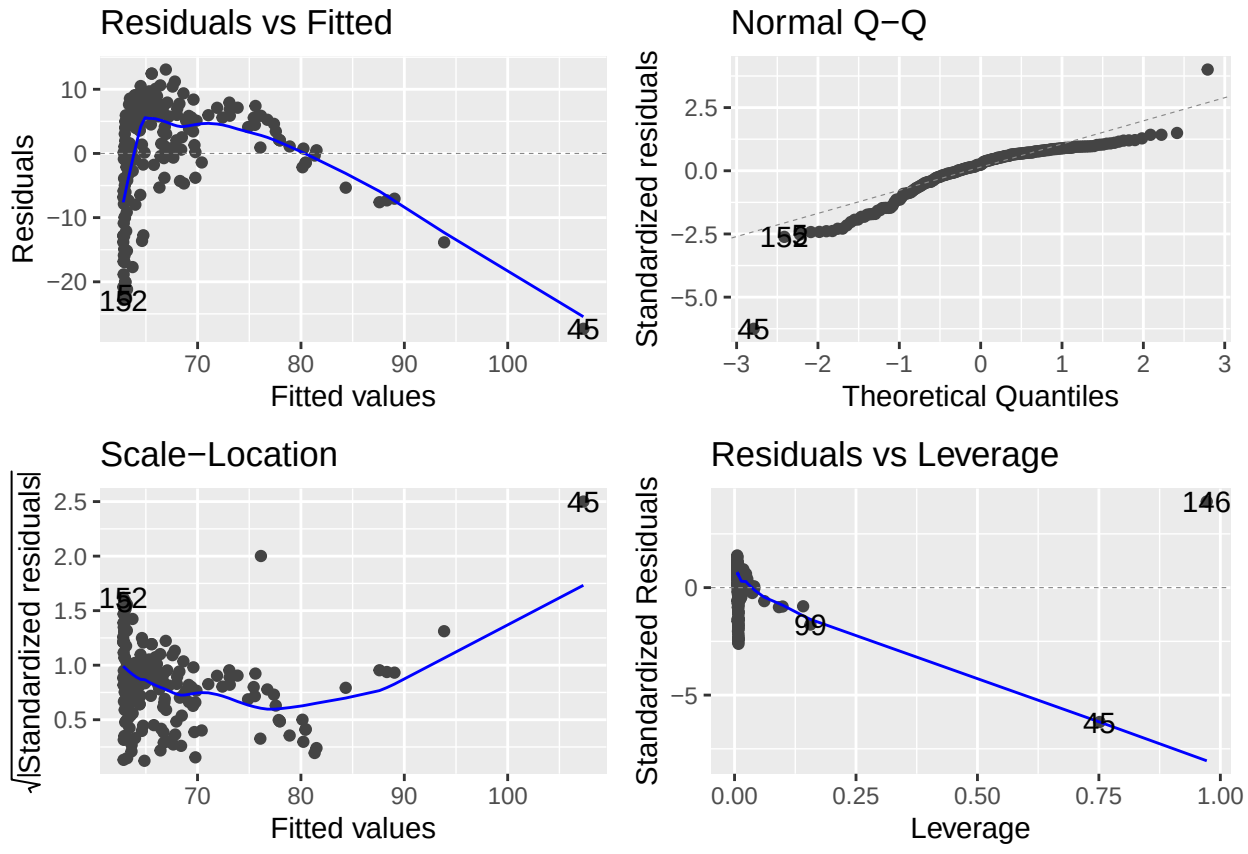
```
##
## Call:
## lm(formula = LifeExp ~ PropMD + TotExp + PropMD:TotExp, data = whodata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.320  -4.132   2.098   6.540  13.074
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.277e+01  7.956e-01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

```
f_statistic_mult = summary_mult$fstatistic
r_squared_mult = summary_mult$r.squared
rse_mult = summary_mult$sigma
coefficients_p_values_mult = summary_mult$coefficients[,4]


model_summary_mult = list(
  f_statistic = f_statistic_mult,
  r_squared = r_squared_mult,
```

```
  residual_standard_error = rse_mult,
  coefficients_p_values = coefficients_p_values_mult
)


model_summary_mult
```

```
## $f_statistic
##     value     numdf     dendf
##  34.48833   3.00000 186.00000
##
## $r_squared
## [1] 0.3574352
##
## $residual_standard_error
## [1] 8.765493
##
## $coefficients_p_values
##   (Intercept)        PropMD        TotExp PropMD:TotExp
## 6.207187e-145  2.320603e-07  9.386290e-14  6.352733e-05
```

```
autoplot(mult_model)
```



The intercept coefficient is ~62.77, with a very small standard error, and it is highly significant with a p-value less than 2e-16.

The coefficient for `PropMD` is ~1497, also with a significant p-value of 2.32e-07, meaning as the proportion of

MDs increases, there is a significant increase in life expectancy.

The coefficient for `TotExp` is ~0.00007233, with a significant p-value of 9.39e-14. This means that as total expenditure increases, life expectancy also increases.

The interaction term `PropMD:TotExp` has a coefficient of ~ -0.006026, with a significant p-value of 6.35e-05. This means that the relationship between `PropMD` and life expectancy changes as `TotExp` changes.

The RSE is about 8.765, which indicates that the typical size of the residuals is around 8.765 years.

The Multiple R^2 value is 0.3574, and the adjusted R^2 is lower at 0.3471. This means that the model explains ~35.74% of the variability in Life Expectancy, which is an OK amount but meaning there are other factors not included in our data that affects this..

The F-statistic is 34.49 with a very low p-value (< 2.2e-16) meaningthat the model is statistically significant.

Because we have a statistically significant interaction term, that the effect of the proportion of MDs on life expectancy is not constant Instead, it depends on the level of total expenditure. This indicates a more complex relationship between the variables than a simple additive model would suggest. While the model is good, there is still variability that the model doesn't explain.

#5

Question

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

Work and Answer

```
new_data = data.frame(PropMD = 0.03, TotExp = 14)
life_expectancy_forecast = predict(mult_model, newdata = new_data)

life_expectancy_forecast
```

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

The forecasted life expectancy is 107.696 years when **`PropMD`** is 0.03 and **`TotExp`** is 14. This forecast is not realistic when compared to current global life expectancy statistics. The highest average life expectancy does not exceed 90 years in any country.