DATA605 Homework 12

```
library(tidyverse)
library(cowplot)
library(car)
```

Read in Data

```
who <- read_csv('data/who.csv')
```

```
## Rows: 190 Columns: 10
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (1): Country
## dbl (9): LifeExp, InfantSurvival, Under5Survival, TBFree, PropMD, PropRN, Pe...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
glimpse(who)
```

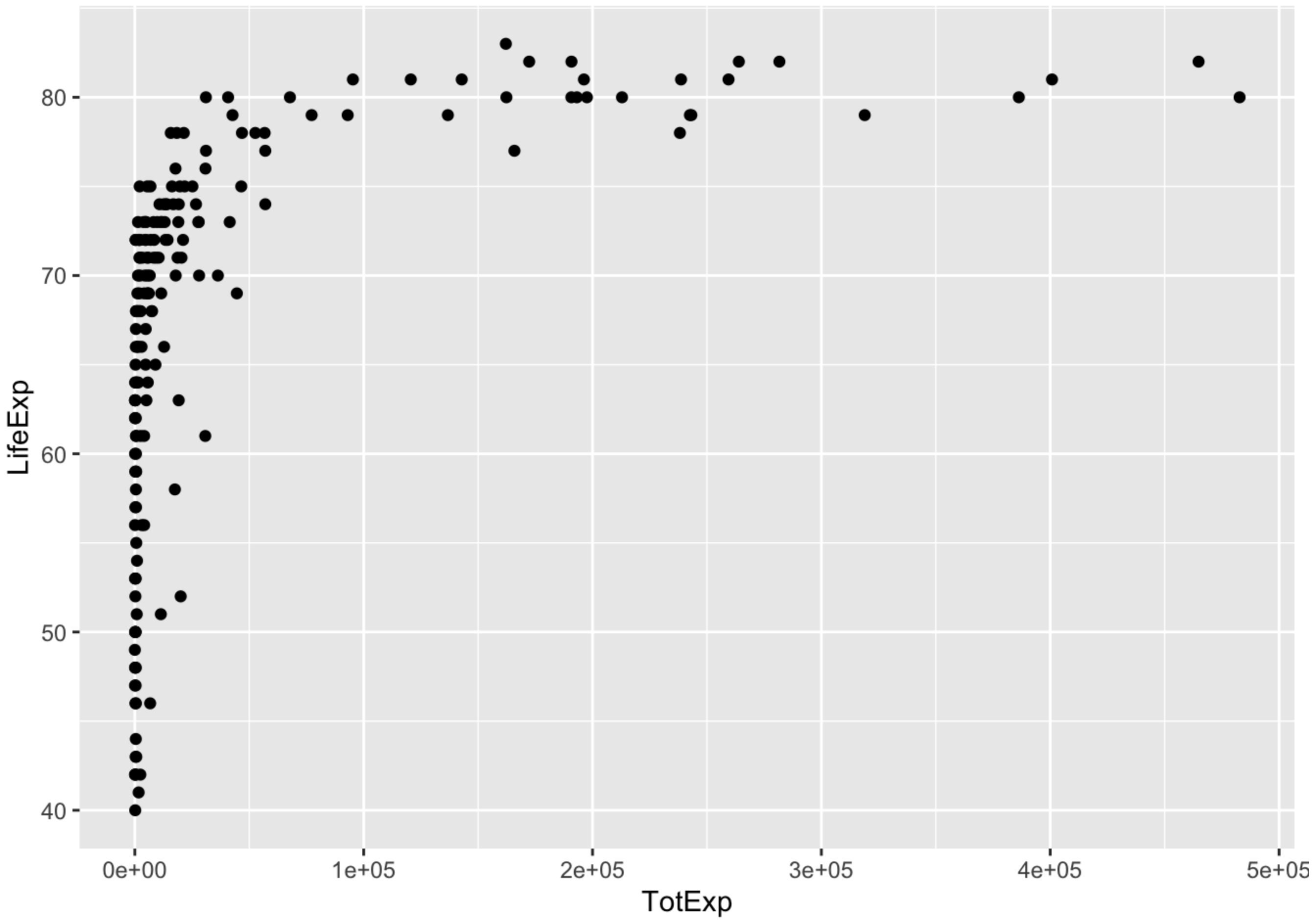
```
## Rows: 190
## Columns: 10
## $ Country        <chr> "Afghanistan", "Albania", "Algeria", "Andorra", "Angola…
## $ LifeExp        <dbl> 42, 71, 71, 82, 41, 73, 75, 69, 82, 80, 64, 74, 75, 63,…
## $ InfantSurvival <dbl> 0.835, 0.985, 0.967, 0.997, 0.846, 0.990, 0.986, 0.979,…
## $ Under5Survival <dbl> 0.743, 0.983, 0.962, 0.996, 0.740, 0.989, 0.983, 0.976,…
## $ TBFree         <dbl> 0.99769, 0.99974, 0.99944, 0.99983, 0.99656, 0.99991, 0…
## $ PropMD         <dbl> 0.000228841, 0.001143127, 0.001060478, 0.003297297, 0.0…
## $ PropRN         <dbl> 0.000572294, 0.004614439, 0.002091362, 0.003500000, 0.0…
## $ PersExp        <dbl> 20, 169, 108, 2589, 36, 503, 484, 88, 3181, 3788, 62, 1…
## $ GovtExp        <dbl> 92, 3128, 5184, 169725, 1620, 12543, 19170, 1856, 18761…
## $ TotExp         <dbl> 112, 3297, 5292, 172314, 1656, 13046, 19654, 1944, 1907…
```

Task 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.

Scatterplot

```
who %>%
  ggplot(aes(TotExp, LifeExp)) +
  geom_point()
```



This scatter plot does not indicate a clear linear relationship. There does seem to be *some* kind of relationship, however, as the high values of Total Expenditures appear correlated with high values of Life Expectancy.

Model

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

```
##
## 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
```

Based on this summary alone, the model appears half-way decent. The F-stat indicates the model is significant, and the sole predictor appears significant as well, given that the p-values of both approach zero. The R^2 is quite low, however, indicating that Total Expenditures only explains ~25% of the variance in Life Expectancy. The standard error is relatively low, too, at about 1/10th the value of the β coefficient.

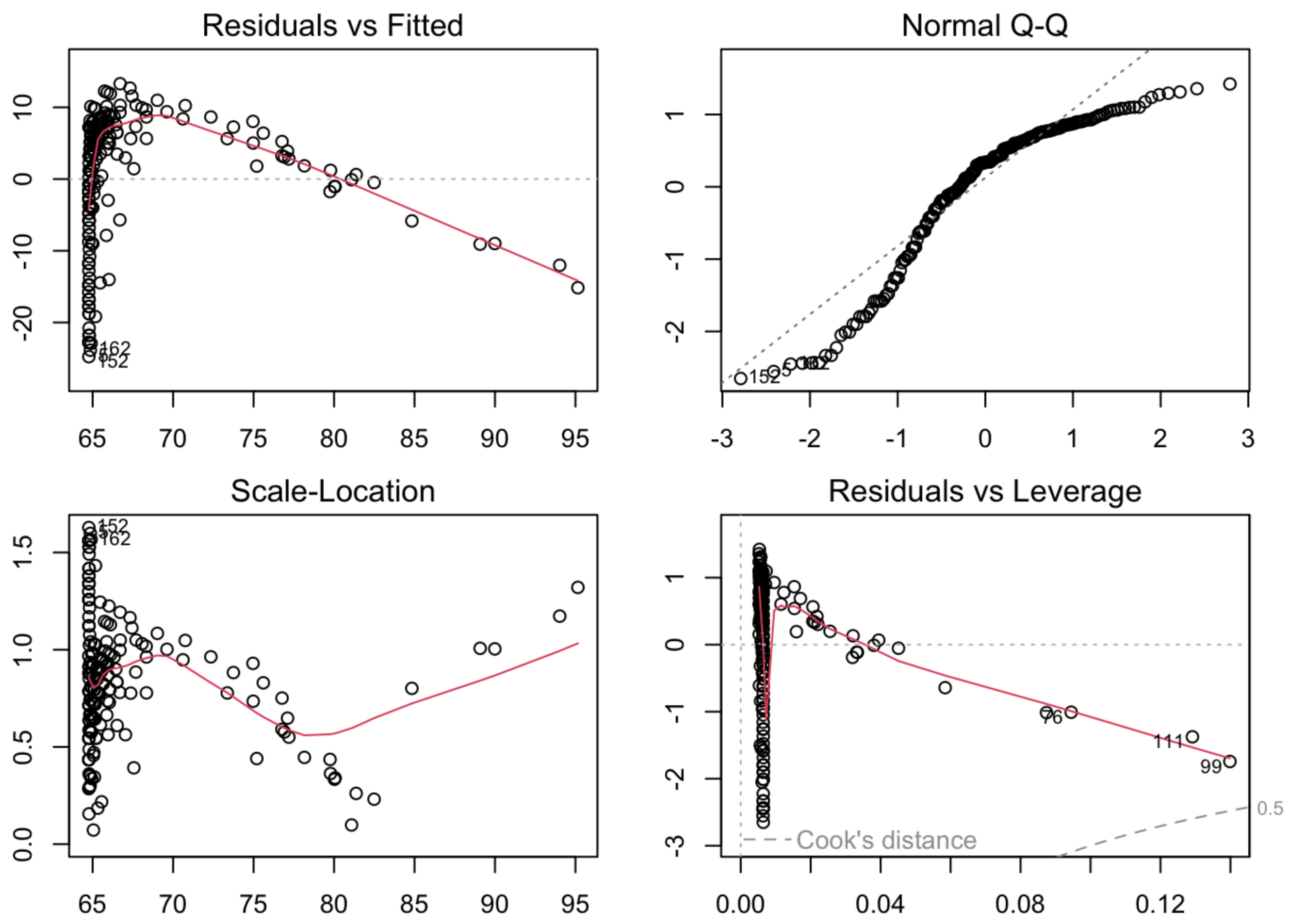
```
results <- summary(model)
results$coefficients[2,2] / results$coefficients[2,1]
```

```
## [1] 0.1237834
```

Residuals and Assumptions

When we look at the residuals, however, the issues resurface. They display non-constant variance and appear non-normally distributed. These indicate the our predictor and response variables do not share a linear relationship. The assumptions of OLS appear violated.

```
par(mfrow = c(2, 2), mar = c(2,2,2,2))
plot(model)
```

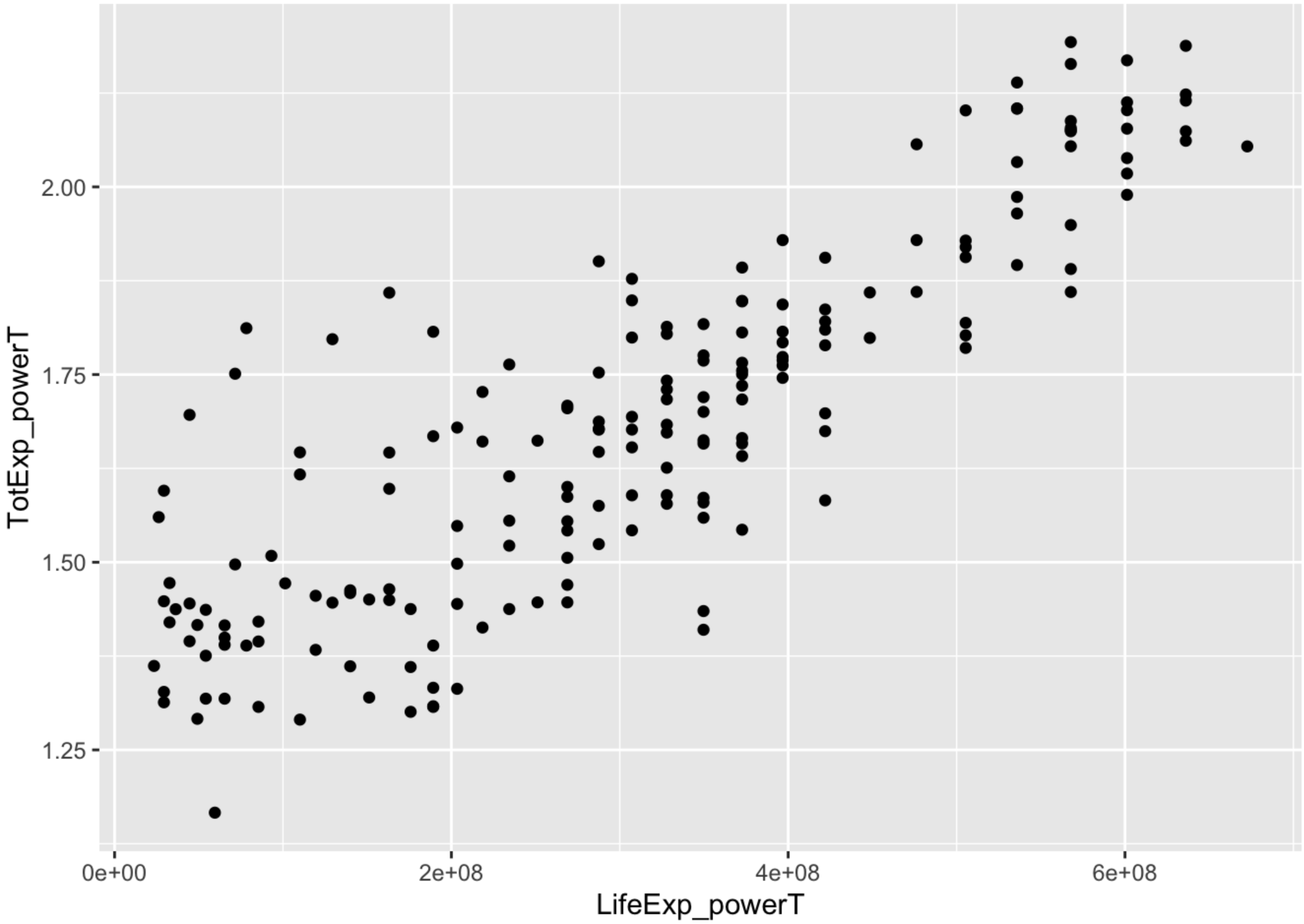


Task 2

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?"

Plot

```
who <- who %>%
  mutate(
    LifeExp_powerT = LifeExp^4.6,
    TotExp_powerT = TotExp^0.06
  )

who %>%
  ggplot(aes(LifeExp_powerT, TotExp_powerT)) +
  geom_point()
```



Already, things look better. There is now a clear linear relationship!

Model

```
model <- lm(LifeExp_powerT ~ TotExp_powerT, data = who)
summary(model)
```

```
## 
## Call:
## lm(formula = LifeExp_powerT ~ TotExp_powerT, data = who)
## 
## 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_powerT  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 overall fit and the TotExp variable appear significant, with very low p-values. This time, however, our R^2 is much igher at ~72%, indicating our transformed predictor explains much more of the transformed response's variance.

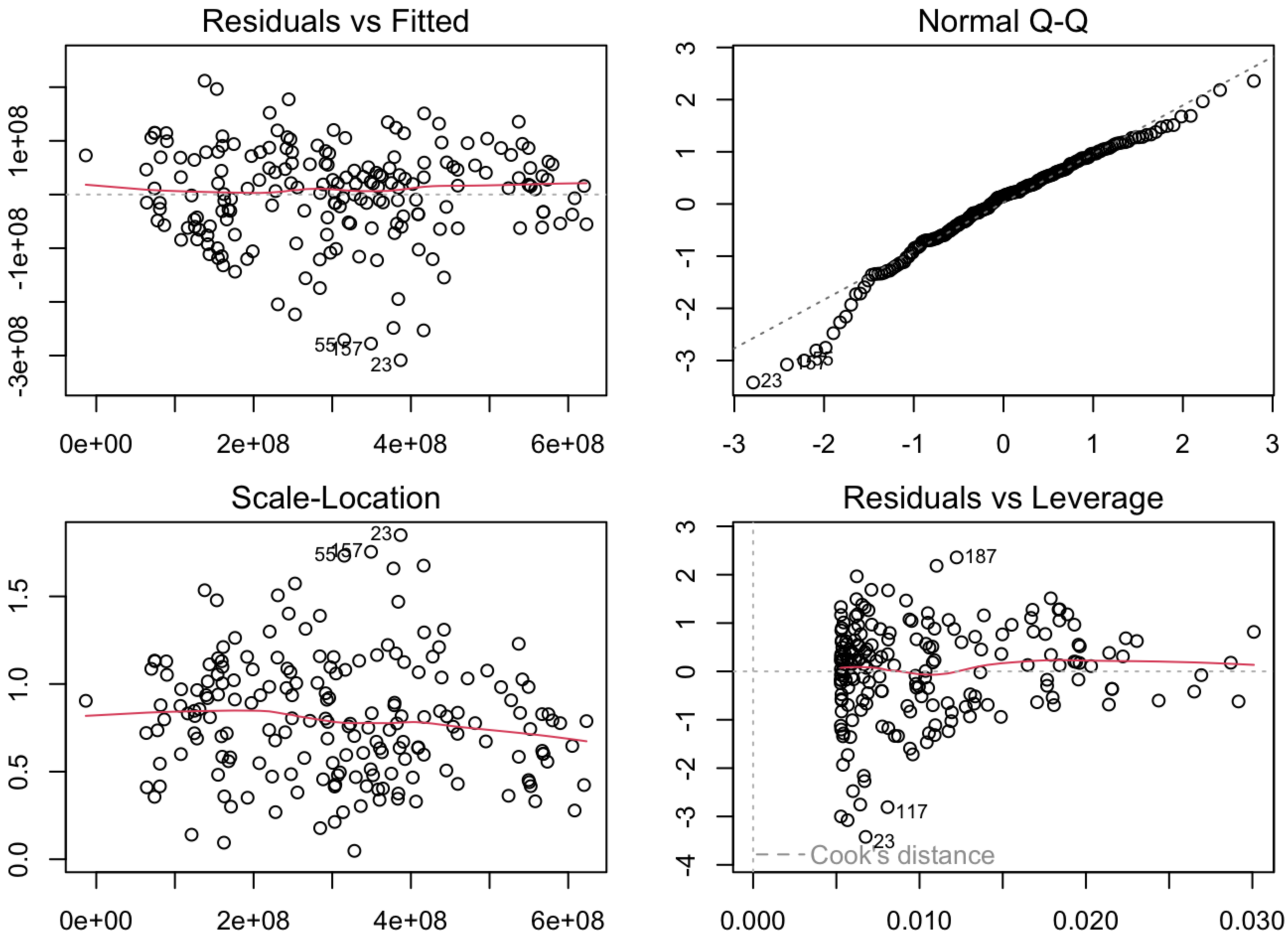
```
results <- summary(model)
results$coefficients[2,2] / results$coefficients[2,1]
```

```
## [1] 0.04438108
```

Our standard error is also much smaller.

Residuals and Assumptions

```
par(mfrow = c(2, 2), mar = c(2,2,2,2))
plot(model)
```



Our residuals are also different. They appear much more uniformly scattered and normally distributed (despite some left tail outliers). This model most certainly provides a better fit.

```
powerTransform(LifeExp ~ TotExp, data = who)
```

```
## Estimated transformation parameter
##     Y1
## 4.7328
```

Task 3

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

```
prediction1 <- predict(model, newdata = data.frame(TotExp_powerT = 1.5))^(1/4.6)
prediction2 <- predict(model, newdata = data.frame(TotExp_powerT = 2.5))^(1/4.6)

cat(
  'Prediction with 1.5: ',
  scales::comma(prediction1),
  '\nPrediction with 2.5: ',
  scales::comma(prediction2),
  sep = ''
)
```

```
## Prediction with 1.5: 63
## Prediction with 2.5: 87
```

We must undo the transformation we applied to the response variable when forming our predictions. With that, we get reasonable predictions.

Task 4

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

$$LifeExp = \beta_0 + \beta_1 PropMd + \beta_1 TotExp + \beta_1 PropMD \times TotExp$$

```
model <- lm(LifeExp ~ PropMD + TotExp + PropMD:TotExp, data = who)
results <- summary(model)
print(results)
```

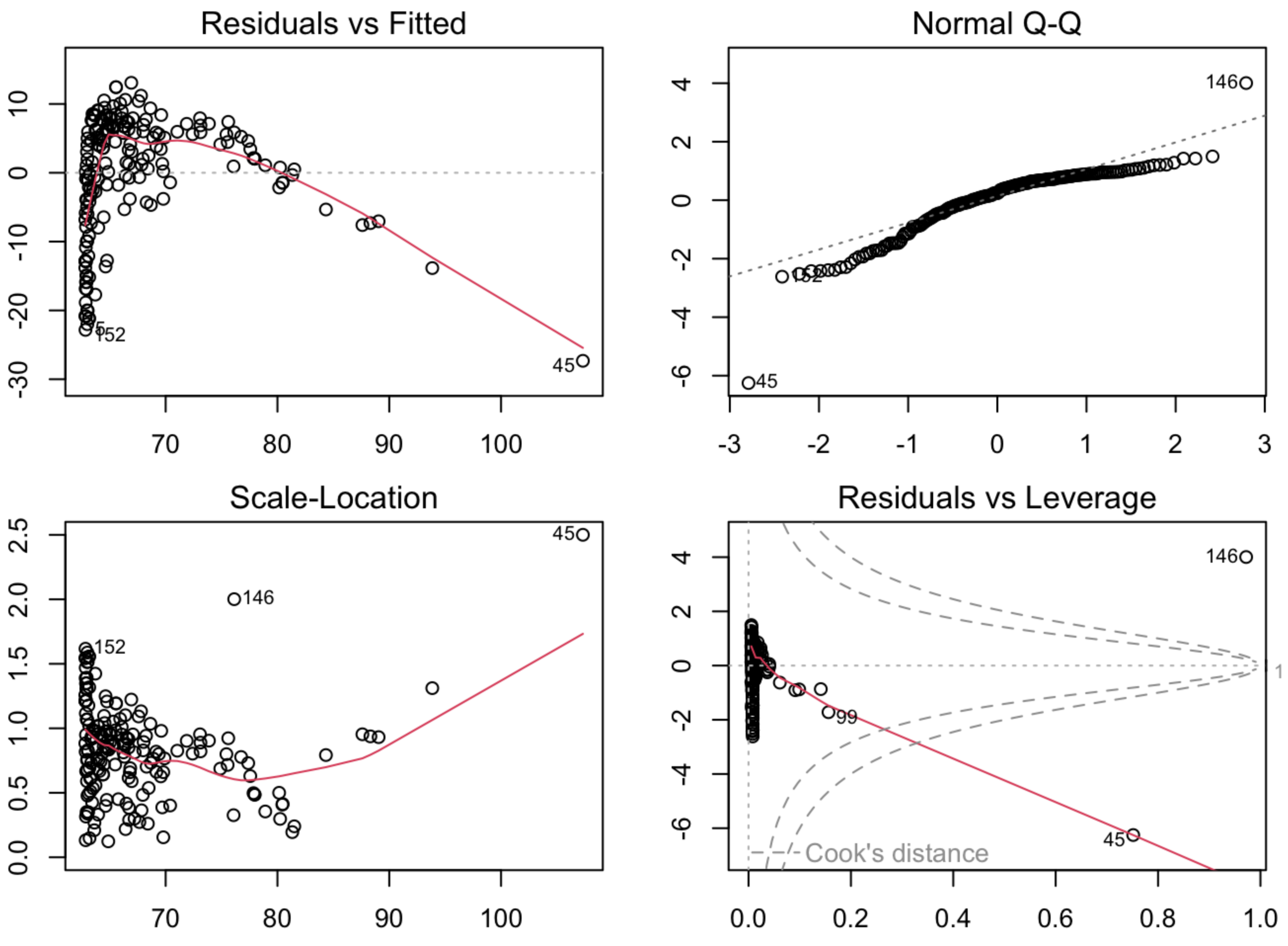
```
##
## Call:
## lm(formula = LifeExp ~ PropMD + TotExp + PropMD:TotExp, data = who)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.320  -4.132   2.098   6.540  13.074
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.277e+01  7.956e-01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

```
cat(
  '--Standard Error to Coefficient Ratios--\n',
  'PropMD: ',abs(results$coefficients[2,2] / results$coefficients[2,1]),'\n',
  'TotExp: ',abs(results$coefficients[3,2] / results$coefficients[3,1]),'\n',
  'Interaction: ',abs(results$coefficients[4,2] / results$coefficients[4,1]),'\n',
  '--------\n',
  sep=''
)
```

```
## --Standard Error to Coefficient Ratios--
## PropMD: 0.186189
## TotExp: 0.1241742
## Interaction: 0.2443468
## --------
```

```
par(mfrow = c(2,2), mar = c(2,2,2,2))
plot(model)
```

```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



We get similar results as we saw in our initial model fit. The F-test and predictor t-test p-values indicate the overall model and all predictors are significant. Our R^2 is relatively low, explaining only ~35% of the response variance. Standard errors are relatively higher, all above 10% of the relevant coefficient. And finally, residuals again appear non-normal and heterskedastic.

Task 5

Forecast LifeExp when PropMD = 0.03 and TotExp = 14. Does this forecast seem realistic? Why or why not?

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

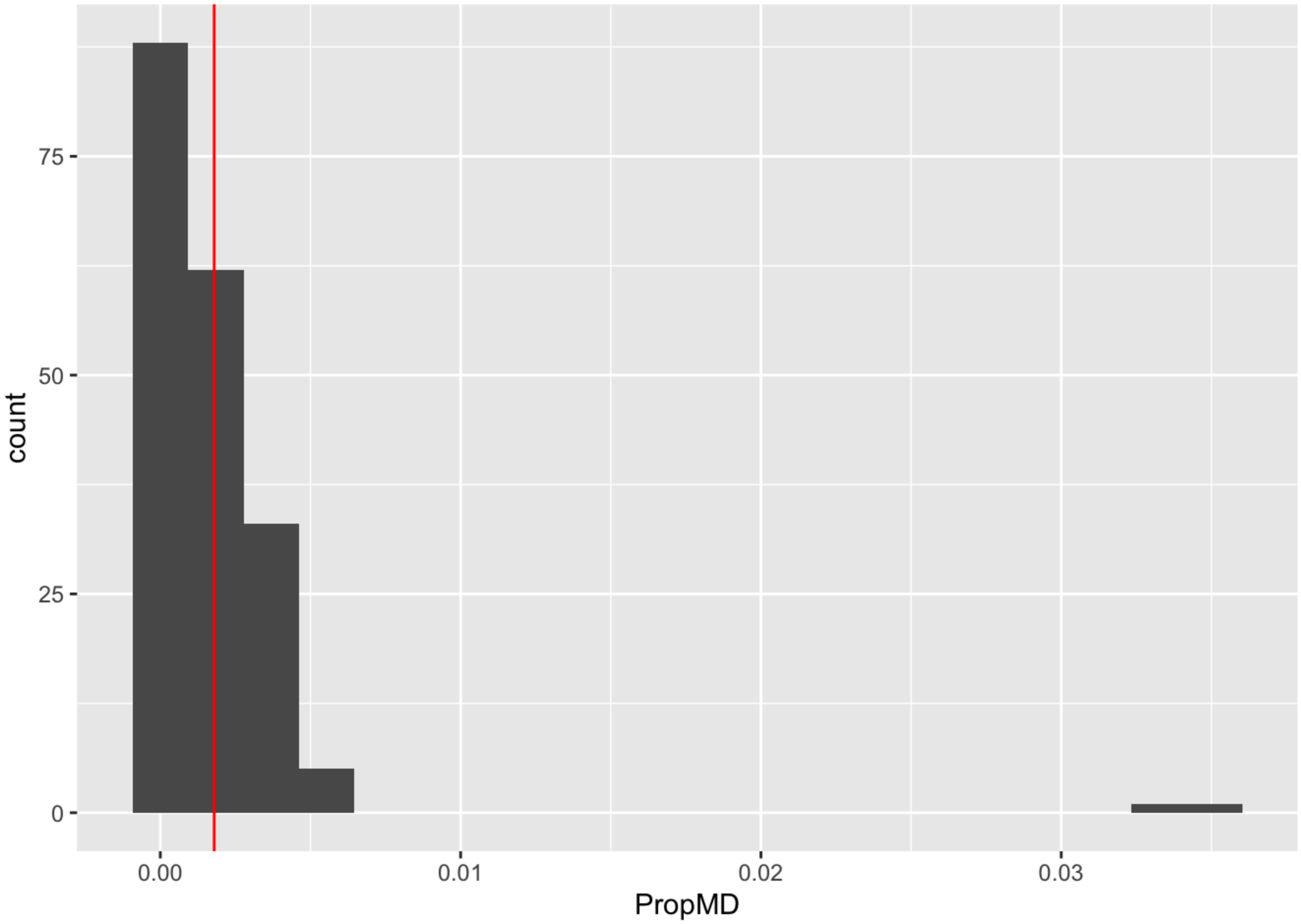
```
##       1
## 107.696
```

A predicted Life Expectancy of ~107.7 years does not seem reasonable. First, the maximum life expectancy in the whole dataset is 83. Second, a Total Expenditure of 14 is quite low, as seen in the plot below, so a prediction of very high life expectancy defies expectations. A Proportion of MDs of 0.3 is, on the other hand, quite high, but not high enough to warrant a life expectancy of 107 years. The country with the highest PropMD in the dataset (San Marino with ~0.35) only has a Life Expectancy of 82 years.

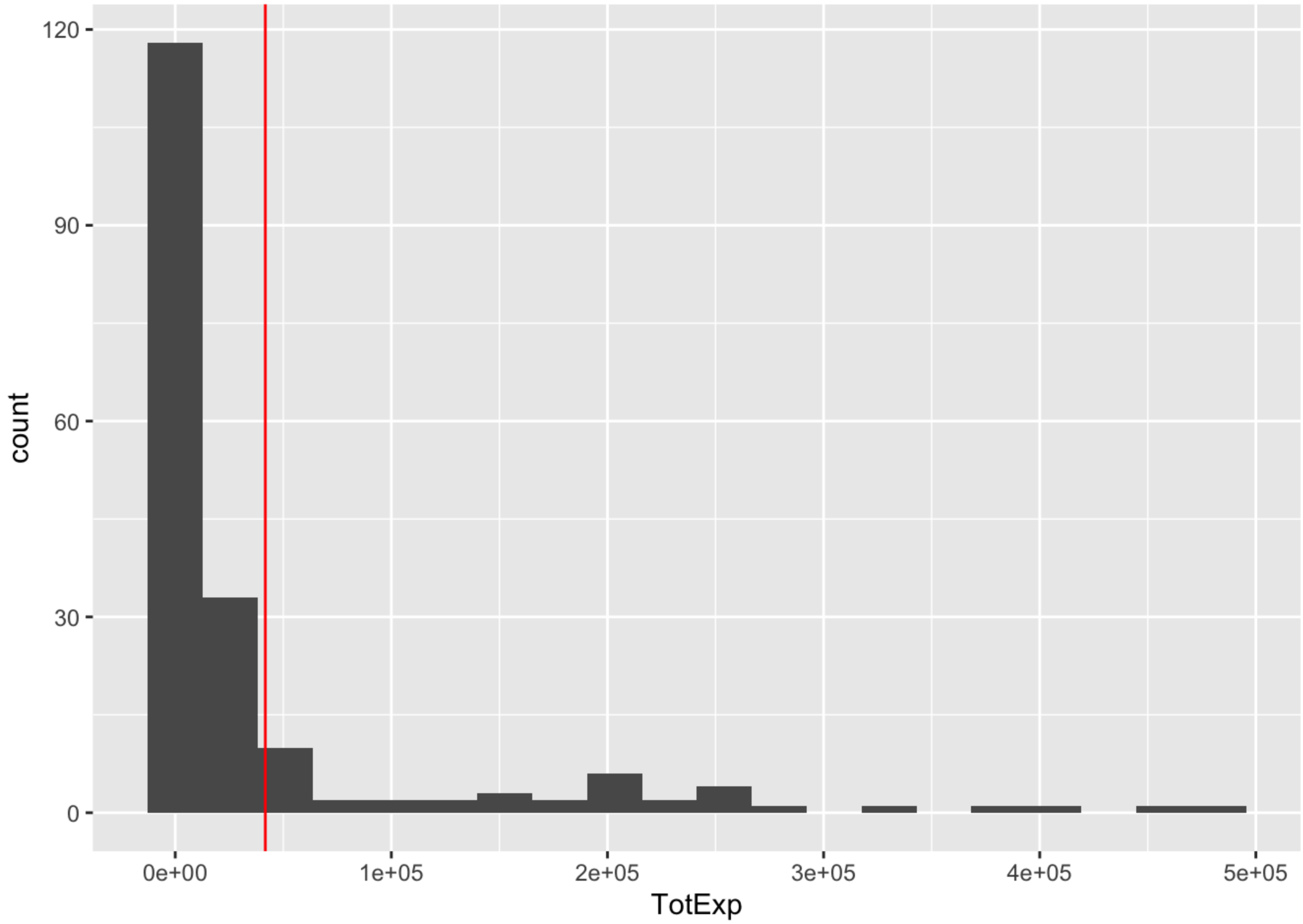
```
max(who$LifeExp)
```

```
## [1] 83
```

```
ggplot(who, aes(PropMD)) +
  geom_histogram(bins = 20) +
  geom_vline(xintercept = mean(who$PropMD), color = 'red')
```



```
ggplot(who, aes(TotExp)) +
  geom_histogram(bins = 20) +
  geom_vline(xintercept = mean(who$TotExp), color = 'red')
```



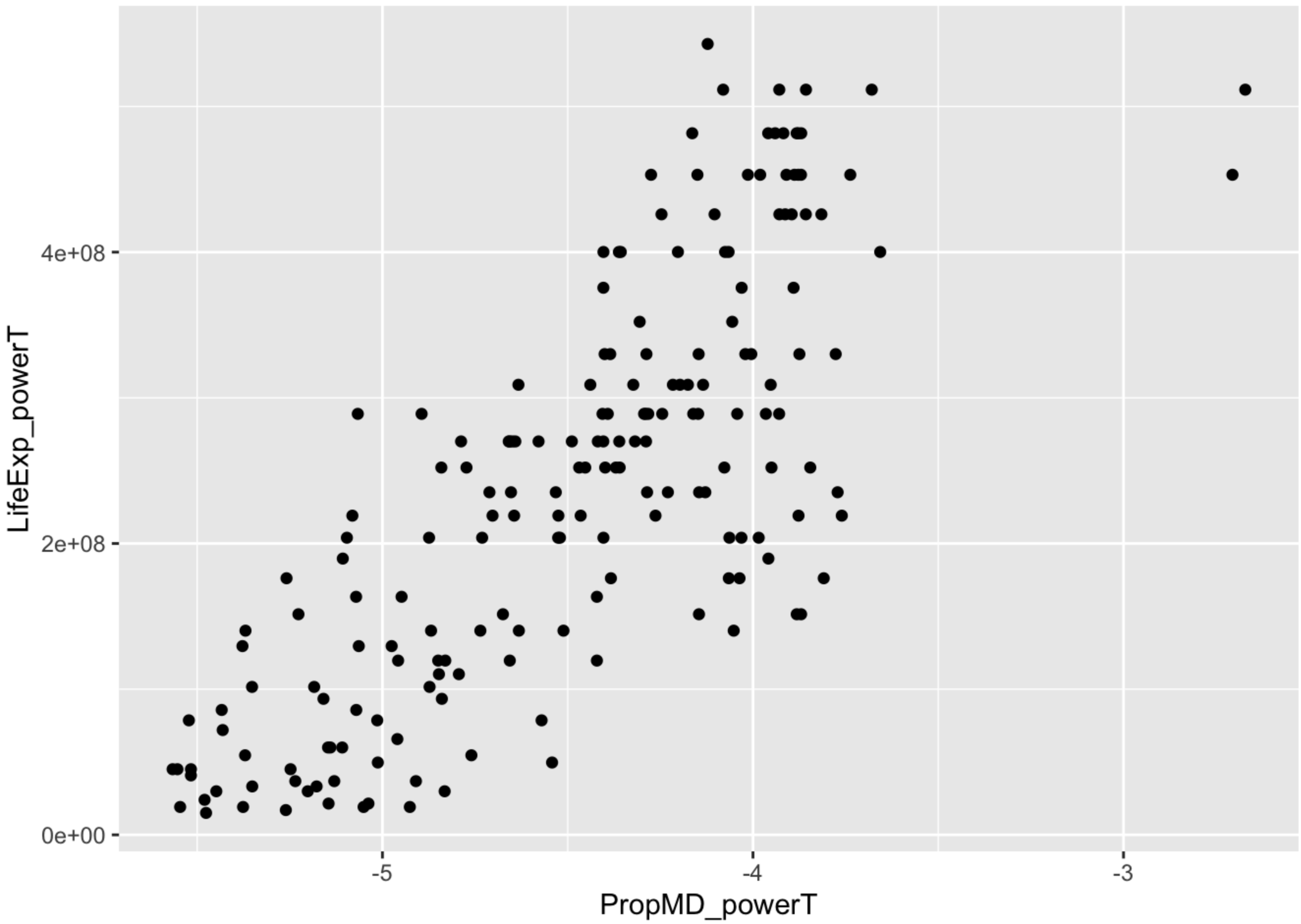
```
who %>%
  filter(PropMD == max(who$PropMD))
```

Country	LifeExp	InfantSurvival	Under5Survival	TBFree	PropMD	PropRN	Pers…
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
San Marino	82	0.997	0.997	0.99995	0.03512903	0.07083871	349

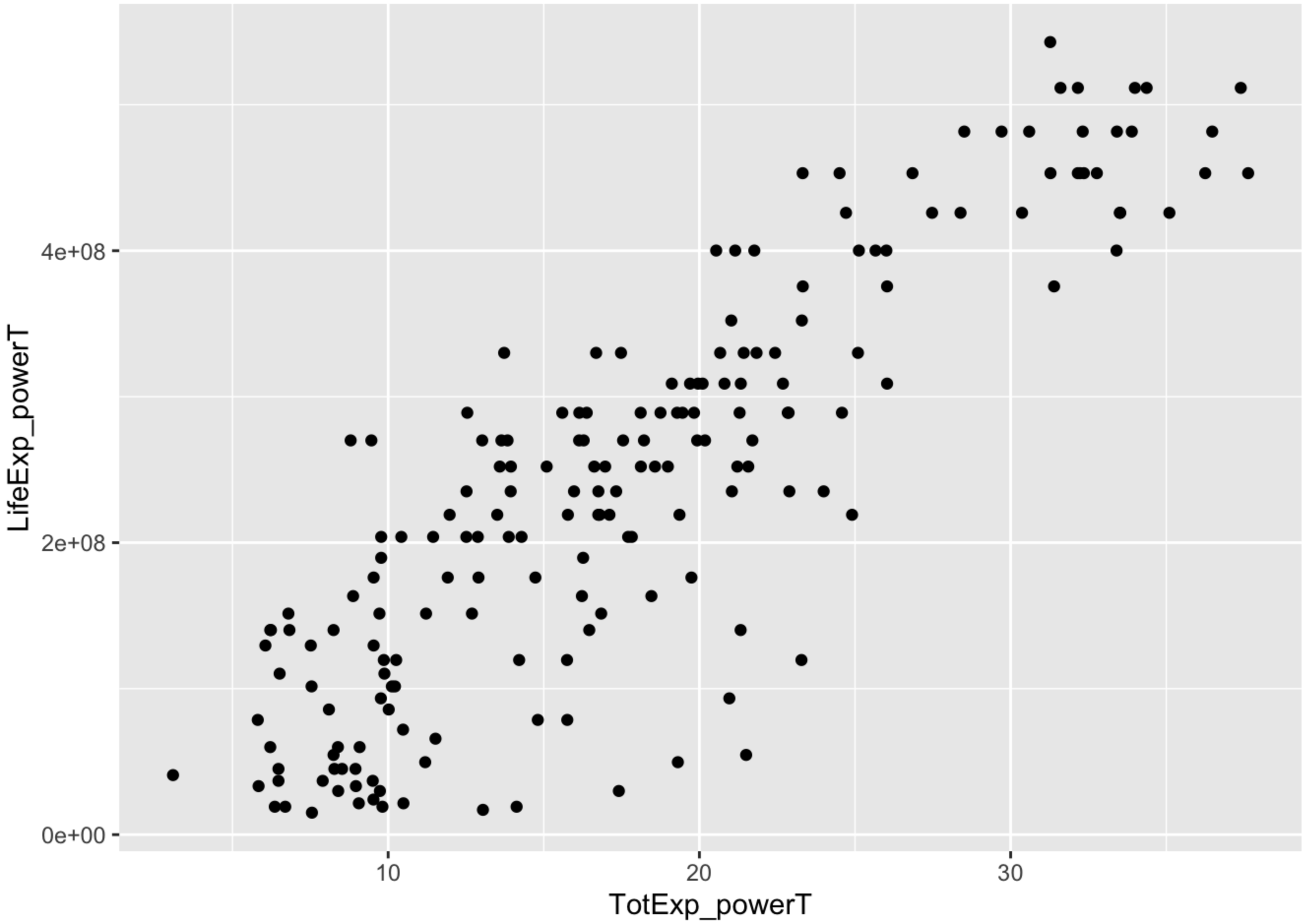
1 row | 1-9 of 12 columns

So, it seems this unrealistic prediction is an indication of problems with the model. This point aligns with the residual problems and sizeable standard errors we identified previously. Let's see if we can apply some transformations to improve the fit and produce a better prediction.

```
y_transformation <- powerTransform(model)
y_transformed <- bcPower(who$LifeExp, y_transformation$lambda)

x1_transformation <- powerTransform(PropMD ~ 1, data = who)
x1_transformed <- bcPower(who$PropMD, x1_transformation$lambda)

x2_transformation <- powerTransform(TotExp ~ 1, data = who)
x2_transformed <- bcPower(who$TotExp, x1_transformation$lambda)

who <- who %>%
  mutate(
    LifeExp_powerT = y_transformed,
    PropMD_powerT = x1_transformed,
    TotExp_powerT = x2_transformed
  )

who %>%
  ggplot(aes(PropMD_powerT, LifeExp_powerT)) +
  geom_point()
```



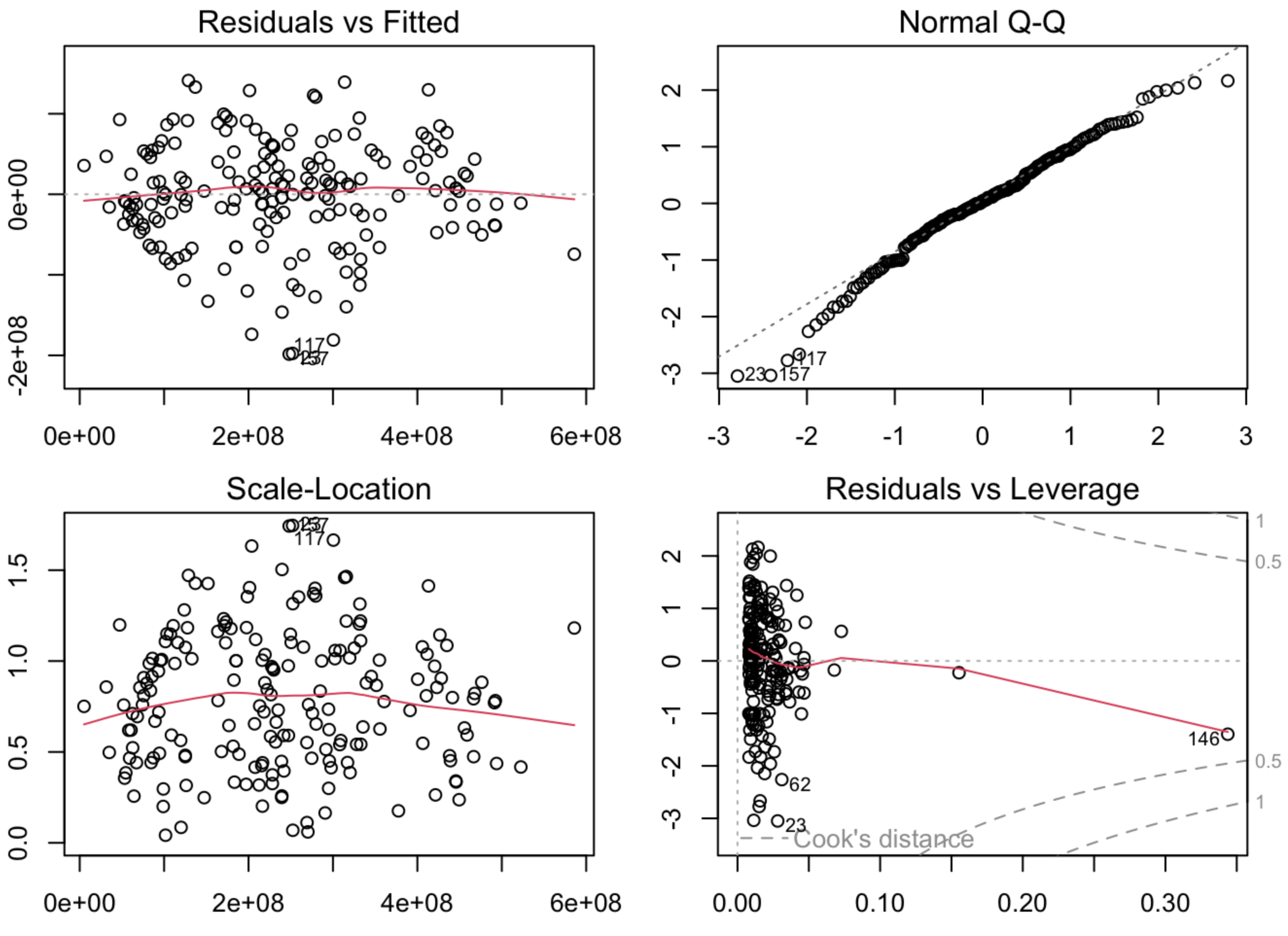
```
who %>%
  ggplot(aes(TotExp_powerT, LifeExp_powerT)) +
  geom_point()
```



```
model <- lm(
  LifeExp_powerT ~ PropMD_powerT + TotExp_powerT + PropMD_powerT:TotExp_powerT,
  data = who
)
print(summary(model))
```

```
##
## Call:
## lm(formula = LifeExp_powerT ~ PropMD_powerT + TotExp_powerT +
##     PropMD_powerT:TotExp_powerT, data = who)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -198499574  -36253931    2636851   44765106  141254397
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  375377245   99388034   3.777 0.000214 ***
## PropMD_powerT                 72617287   20733416   3.502 0.000577 ***
## TotExp_powerT                 13591172    5013972   2.711 0.007343 **
## PropMD_powerT:TotExp_powerT     682009    1180615   0.578 0.564184
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 65720000 on 186 degrees of freedom
## Multiple R-squared:  0.7867, Adjusted R-squared:  0.7832
## F-statistic: 228.6 on 3 and 186 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2), mar = c(2,2,2,2))
plot(model)
```



```
newdata = data.frame(
  PropMD_powerT = 0.03^x1_transformation$lambda,
  TotExp_powerT = 14^x2_transformation$lambda
)
predict(model, newdata = newdata)^(1/y_transformation$lambda)
```

```
##       Y1
## 57.35217
```

After applying transformations, our model fit appears much better. Residuals are roughly normal with reasonable constant variance, and the R^2 is much higher. Our prediction is also much more reasonable. It's very low, but not outside what is seen in the dataset (e.g. Sierra Leone has a Life Expectancy of only 40 years). These transformations appear very effective!