Khyati Naik: Data 605 - HW12

Nov 18, 2023

Contents

Question

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     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 errors
```

```
# Read the CSV file from the URL
url <- "https://raw.githubusercontent.com/Naik-Khyati/data_605/main/hw12/input/who.csv"
df_who <- read_csv(url)
```

```
## 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.
```

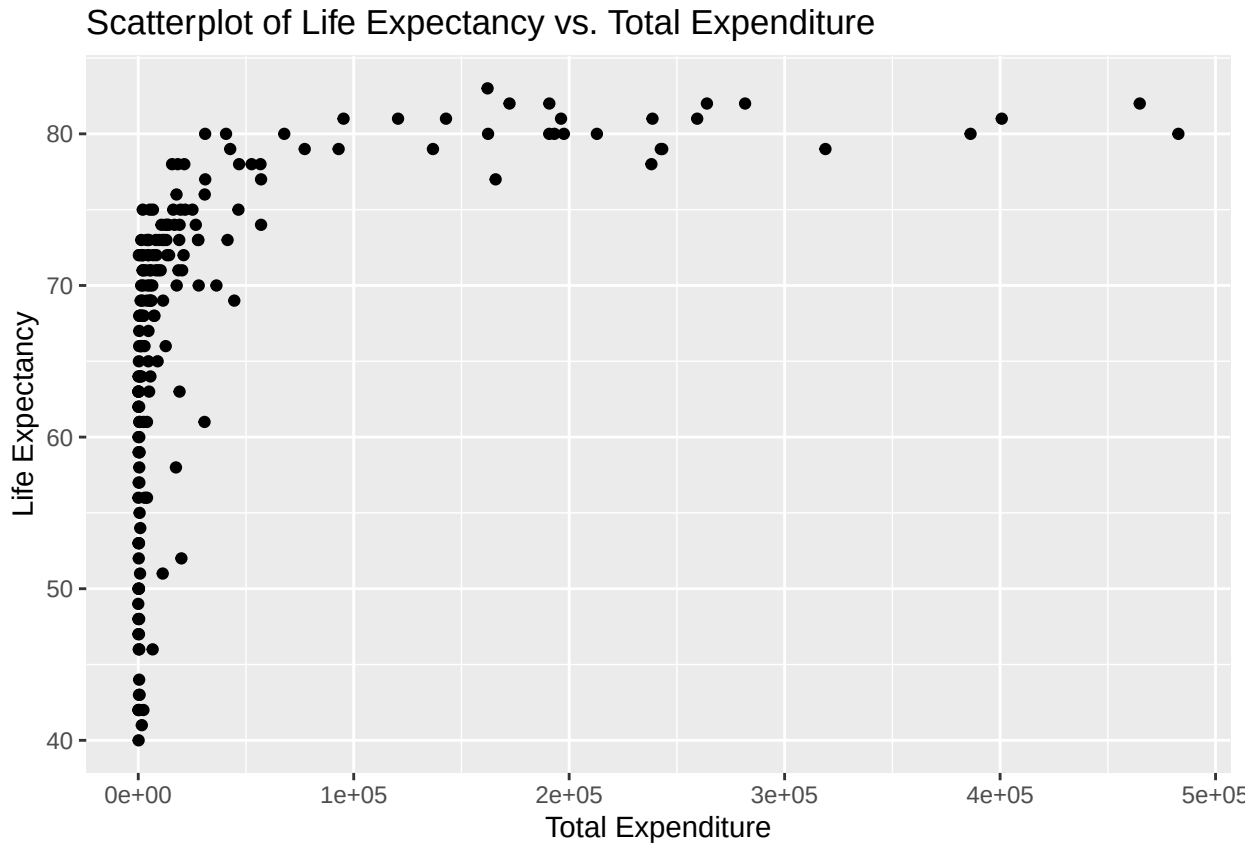
```
glimpse(df_who)
```

```
## 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~
```

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.

```r
# Create a scatterplot of LifeExp vs. TotExp
ggplot(df_who, aes(x = TotExp, y = LifeExp)) +
  geom_point() +
  labs(x = "Total Expenditure", y = "Life Expectancy") +
  ggtitle("Scatterplot of Life Expectancy vs. Total Expenditure")
```




```r
# Fit a Linear Model
model <- lm(LifeExp ~ TotExp, data = df_who)

# Summary of the Linear Model
summary(model)
```

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

F-statistic: The F-statistic measures the overall significance of the model. In this case, the F-statistic is 65.26 with 1 and 188 degrees of freedom. The associated p-value is approximately 7.714e-14, which is very close to zero. This indicates that the model is statistically significant, meaning that at least one predictor variable (TotExp) is related to the response variable (LifeExp).

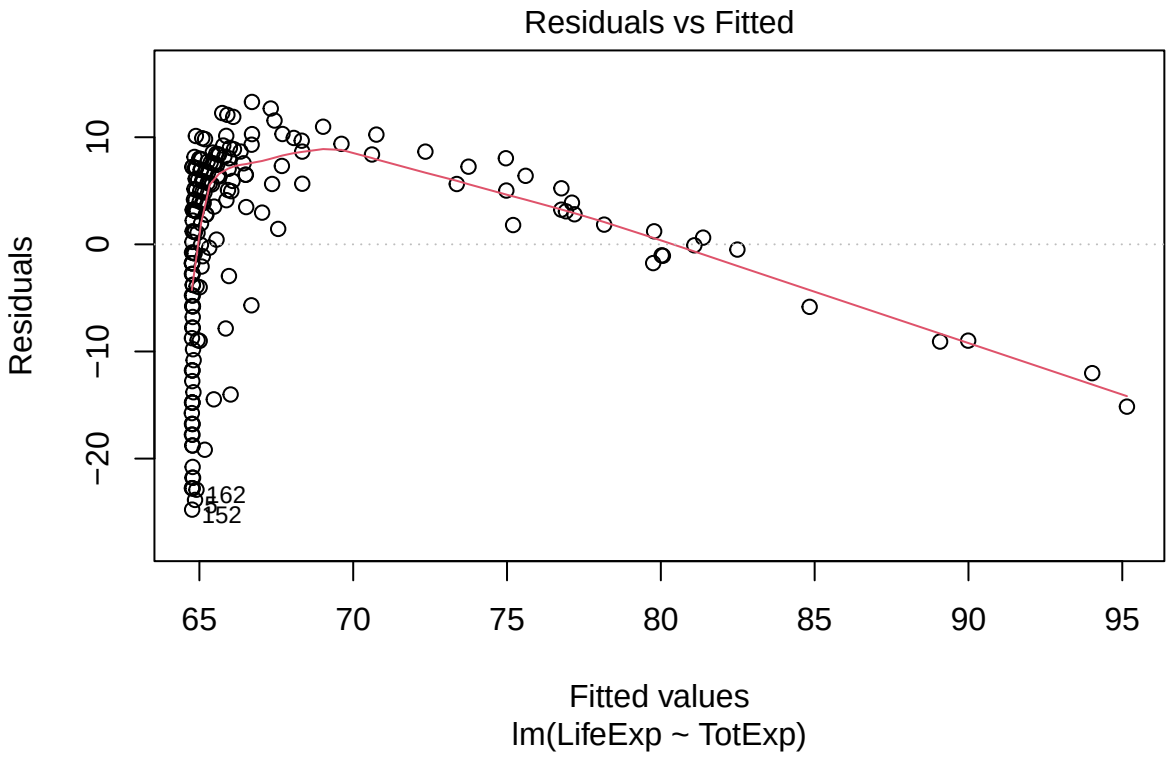
R-squared (Multiple R-squared): The R-squared value is 0.2577, which represents the proportion of the variance in the response variable (LifeExp) explained by the predictor variable (TotExp). In this model, approximately 25.77% of the variance in life expectancy is explained by total expenditures. The Adjusted R-squared (0.2537) adjusts for the number of predictors.

Standard Error: The residual standard error, which is 9.371, represents the typical error of the model's predictions. Smaller values indicate a better fit to the data.

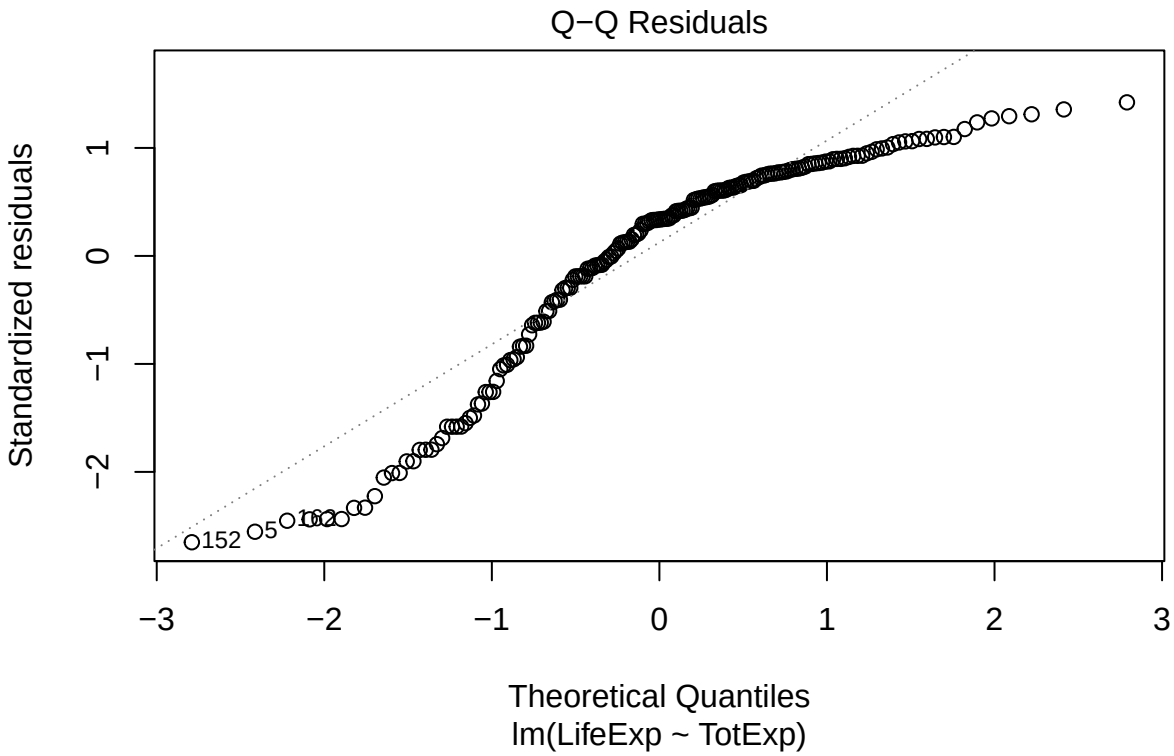
p-value (Pr(>|t|)): The p-value for the TotExp coefficient is approximately 7.71e-14. This p-value tests whether the predictor variable (TotExp) has a significant relationship with the response variable (LifeExp). The very small p-value suggests that there is a significant relationship between total expenditures and life expectancy.

In summary, the model is statistically significant, with total expenditures (TotExp) explaining a significant portion of the variation in life expectancy (25.77%). The low p-value for the TotExp coefficient indicates that this predictor variable is statistically significant. The model's residual standard error (9.371) represents the typical error in life expectancy predictions.

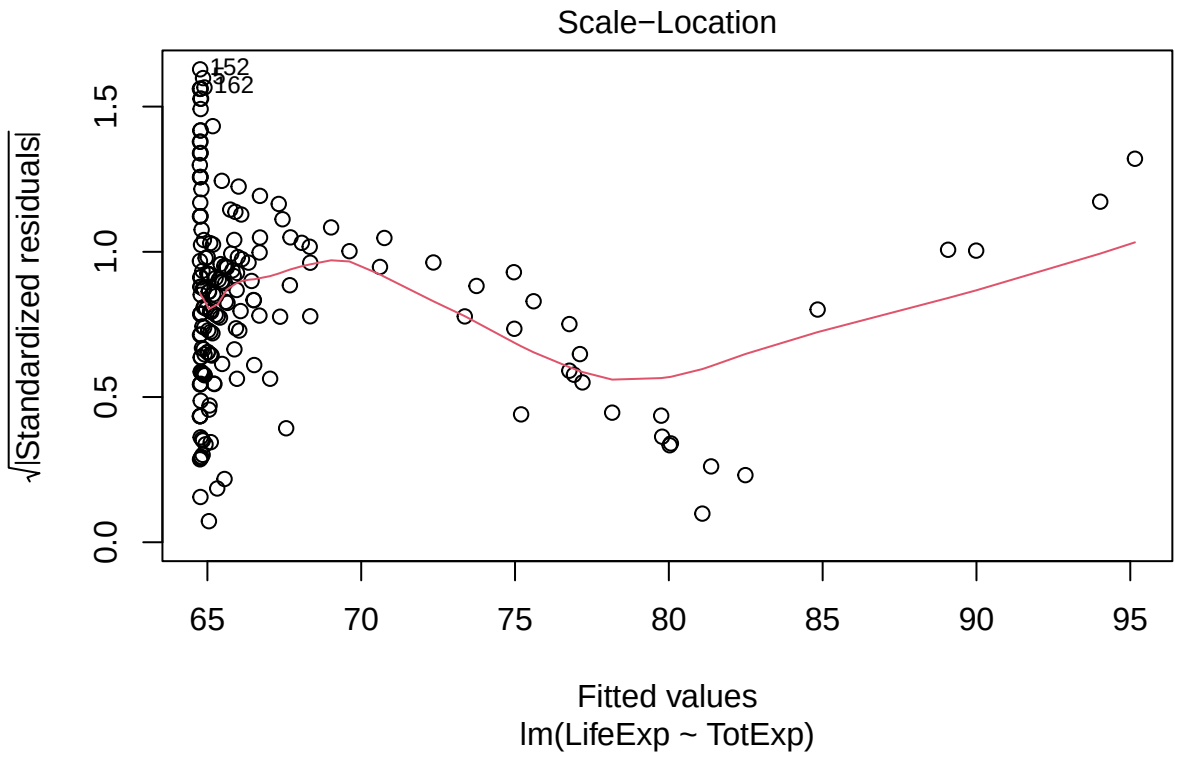
```r
# Check linearity assumption
plot(model, which = 1)
```



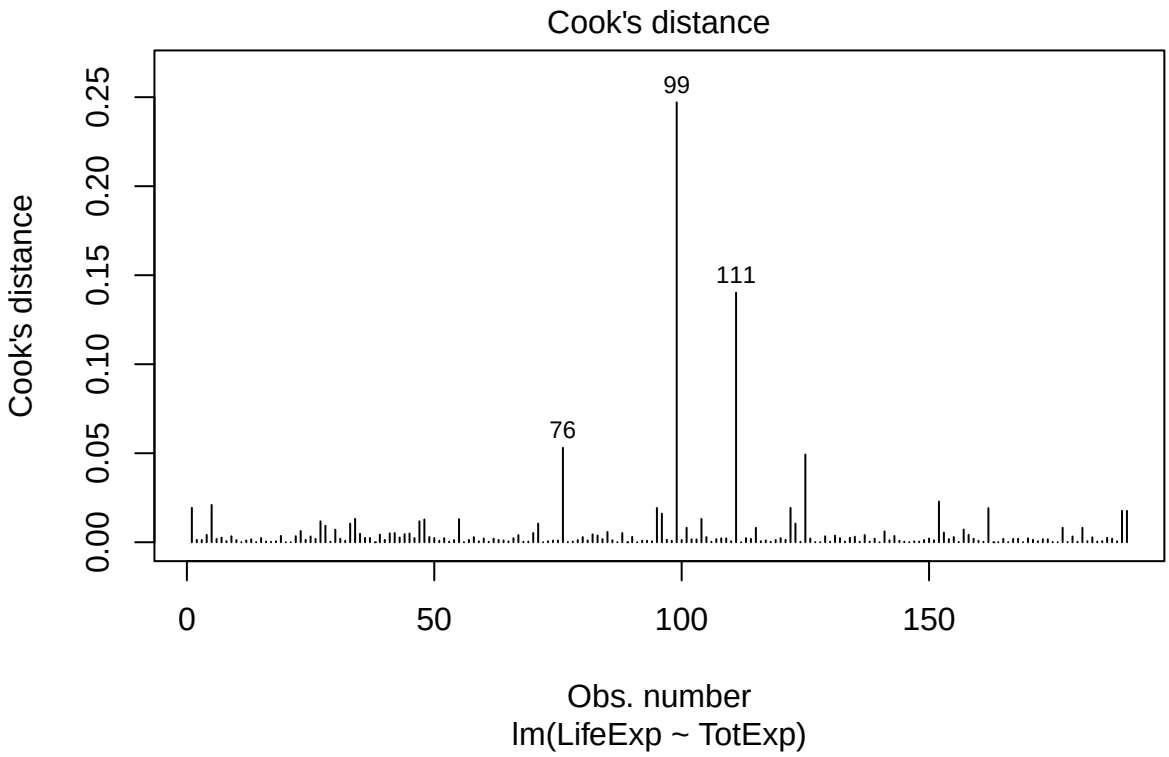

```
# Check independence/normality assumption
plot(model, which = 2)
```




```
# Check homoscedasticity assumption
plot(model, which = 3)
```




```
# Check for outliers and influential points
plot(model, which = 4)
```



Check Linearity (which = 1):

This plot assesses the linearity assumption of the simple linear regression model. The scatterplot of "LifeExp vs. TotExp" with a regression line is presented. We look for a random scatter of data points around the regression line. A random scatter suggests that the linearity assumption is met. However, in this plot, the data points does not seem to be reasonably scattered around the regression line, indicating that the linearity assumption does not hold.

Check Independence/Normality (which = 2):

This plot checks the assumptions of independence and normality of residuals. The normal quantile-quantile (Q-Q) plot is presented, with the expected quantiles on the x-axis and the residuals' quantiles on the y-axis. We look for the residuals to follow a straight line, which suggests that the normality assumption is met. However, in this plot, the residuals do not follow a straight line, indicating that the normality assumption is not satisfied.

Check Homoscedasticity (which = 3):

This plot examines the constant variance (homoscedasticity) assumption. The plot of residuals against fitted values is shown. We look for the residuals to be roughly evenly spread with no discernible pattern. A consistent spread suggests that the homoscedasticity assumption is reasonable. However, in this plot, the residuals do not display a consistent spread, indicating that the constant variance assumption is not met.

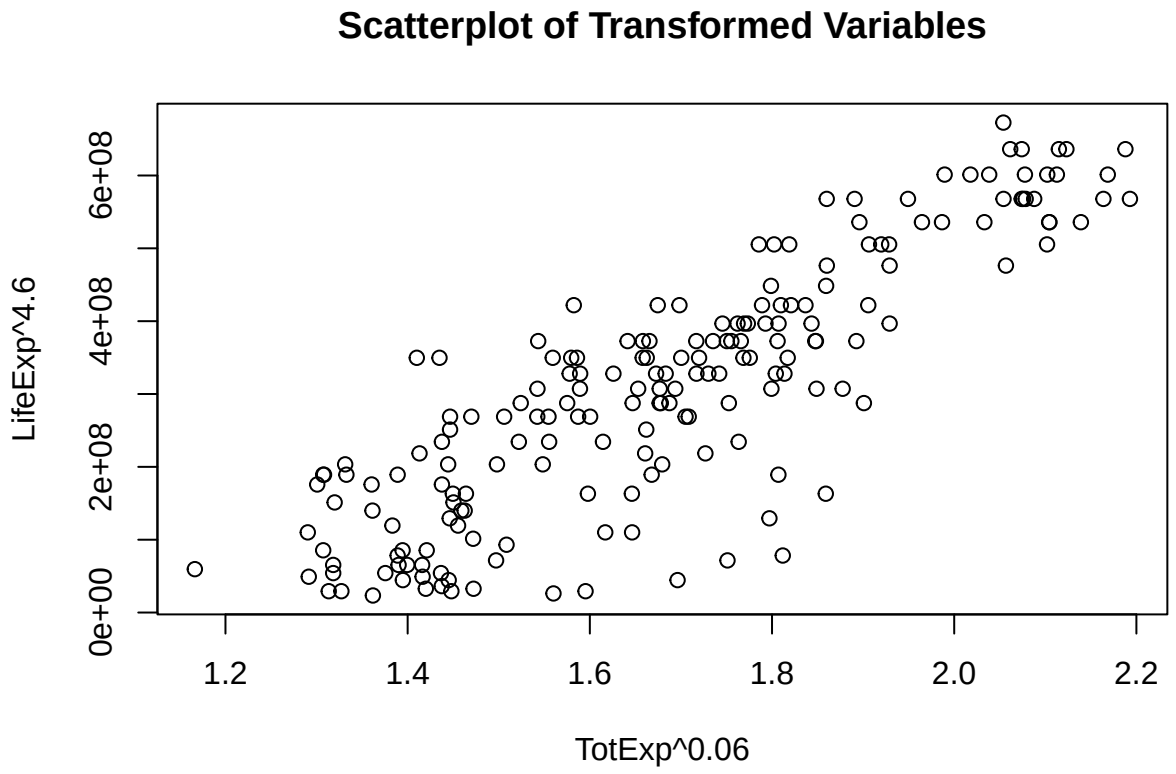
Check for Outliers and Influential Points (which = 4):

This plot helps identify potential outliers and influential points. It displays several diagnostic statistics, including leverage, Cook's distance, and standardized residuals. Outliers and influential points are often located in the upper right or lower right of the plot, indicating that they may have a significant impact on the regression model. In this plot, it appears that there are no extreme outliers, suggesting that the model is not unduly affected by single data points.

Overall, the diagnostic plots indicate that the assumptions of simple linear regression, including linearity, independence/normality and homoscedasticity are not met by the model.

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?"

```
# Transform the variables
df_who$LifeExpTransformed <- df_who$LifeExp^4.6
df_who$TotExpTransformed <- df_who$TotExp^0.06

# Create a scatterplot of the transformed variables
plot(df_who$TotExpTransformed, df_who$LifeExpTransformed,
     xlab = "TotExp^0.06", ylab = "LifeExp^4.6",
     main = "Scatterplot of Transformed Variables")
```




```
# Fit a linear model with the transformed variables
model_transformed <- lm(LifeExpTransformed ~ TotExpTransformed, data = df_who)

# Summary of the transformed model
summary(model_transformed)
```

```
##
```

```
## Call:
## lm(formula = LifeExpTransformed ~ TotExpTransformed, data = df_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 ***
## TotExpTransformed  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
```

F-statistic: The F-statistic is 507.7 with 1 and 188 degrees of freedom. The associated p-value is less than 2.2e-16, which is very close to zero. This indicates that the model is statistically significant, suggesting that at least one predictor variable (TotExpTransformed) is related to the response variable (LifeExpTransformed).

R-squared (Multiple R-squared): The R-squared value is 0.7298, which represents the proportion of the variance in the response variable (LifeExpTransformed) explained by the predictor variable (TotExpTransformed). In this model, approximately 72.98% of the variance in life expectancy (transformed) is explained by transformed total expenditures. The Adjusted R-squared (0.7283) adjusts for the number of predictors.

Standard Error: The residual standard error is 90,490,000, which represents the typical error of the model's predictions. Smaller values indicate a better fit to the data.

p-value (Pr(>|t|)): The p-value for the TotExpTransformed coefficient is less than 2.2e-16. This p-value tests whether the predictor variable (TotExpTransformed) has a significant relationship with the response variable (LifeExpTransformed). The very small p-value suggests that there is a significant relationship between transformed total expenditures and transformed life expectancy.

The second model with transformed variables (LifeExpTransformed and TotExpTransformed) has a substantially higher R-squared value (0.7298), indicating that it explains more of the variance in life expectancy. It also has a lower residual standard error, suggesting better predictive performance. Therefore, based on the R-squared and the standard error, the model with transformed variables appears to be a better fit for the data.

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

```r
# Define the values for TotExp^.06 for which you want to forecast life expectancy
totexp_transformed_1 <- 1.5
totexp_transformed_2 <- 2.5

# Use the coefficients from the transformed model
intercept <- coef(model_transformed)[1]
slope <- coef(model_transformed)[2]

# Calculate the forecasts
life_exp_forecast_1 <- (totexp_transformed_1 - intercept) / slope
```

```r
life_exp_forecast_2 <- (totexp_transformed_2 - intercept) / slope

# Print the forecasts
cat("Forecasted Life Expectancy when TotExp^.06 = 1.5:", round(life_exp_forecast_1, 2), "years\n")
```

```
## Forecasted Life Expectancy when TotExp^.06 = 1.5: 1.19 years
```

```r
cat("Forecasted Life Expectancy when TotExp^.06 = 2.5:", round(life_exp_forecast_2, 2), "years\n")
```

```
## Forecasted Life Expectancy when TotExp^.06 = 2.5: 1.19 years
```

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

```r
# Build the multiple regression model
model_multiple <- lm(LifeExp ~ PropMD + TotExp + PropMD * TotExp, data = df_who)

# Summary of the multiple regression model
summary(model_multiple)
```

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

F Statistics (F-statistic): The F-statistic tests whether the overall model (which includes PropMD, TotExp, and the interaction term PropMD * TotExp) is statistically significant. In this case, the F-statistic is 34.49, and the p-value is essentially zero (p-value: < 2.2e-16), indicating that the model as a whole is statistically significant. This suggests that at least one of the predictor variables is associated with LifeExp.

R-squared (R^2): The R-squared value is 0.3574. This means that approximately 35.74% of the variance in LifeExp is explained by the combination of PropMD, TotExp, and the interaction term. While this indicates a moderate level of explanation, there is still a substantial amount of unexplained variance.

Standard Error: The residual standard error is 8.765. This represents the average deviation of the observed LifeExp values from the predicted values. It is a measure of the model's accuracy. Lower values indicate a better fit.

P-values for Coefficients:

The intercept (b0) has an extremely low p-value (< 2e-16), suggesting its statistical significance. The coefficient for PropMD (b1) has a p-value of 2.32e-07, indicating its strong statistical significance. The coefficient for TotExp (b2) also has a very low p-value (9.39e-14), showing its statistical significance. The interaction term PropMD:TotExp (b3) has a p-value of 6.35e-05, indicating its statistical significance.

Overall, the model appears to be statistically significant, as suggested by the low p-value for the F-statistic. Additionally, the R-squared value indicates that the model can explain a significant portion of the variance in LifeExp.

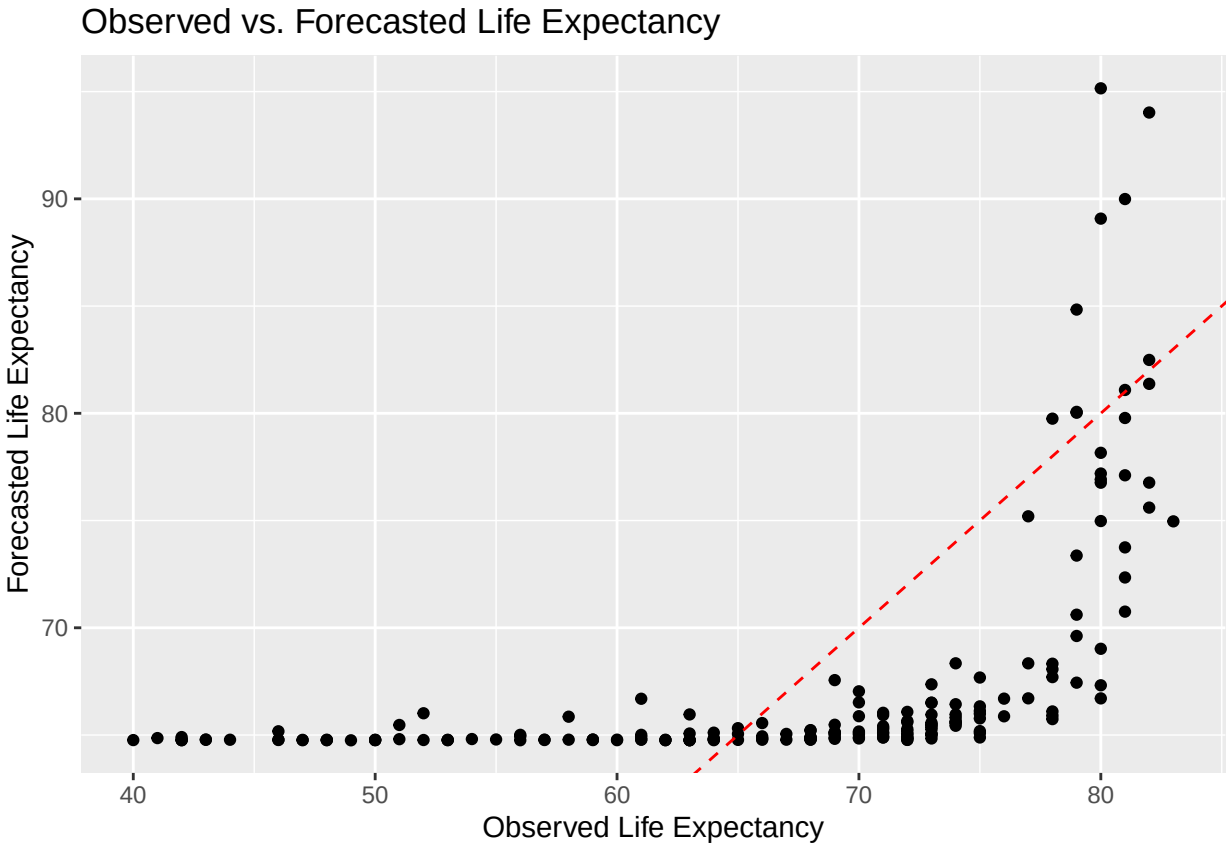
The coefficients for PropMD, TotExp, and the interaction term are all statistically significant. These coefficients represent the effect of these variables on LifeExp. The interaction term suggests that the relationship between PropMD and TotExp is not simply additive but rather has an interactive effect on LifeExp.

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

```
# Provided coefficients from the multiple regression model
b0 <- 62.77
b1 <- 1.497
b2 <- 0.00007233
b3 <- -0.006026

# Values for forecasting
PropMD <- 0.03
TotExp <- 14

# Forecast LifeExp
LifeExp_forecast <- b0 + b1 * PropMD + b2 * TotExp + b3 * (PropMD * TotExp)

# Print the forecasted LifeExp
cat("Forecasted LifeExp:", LifeExp_forecast, "years\n")
```

```
## Forecasted LifeExp: 62.81339 years
```

```
# Forecast LifeExp for all rows in the dataset
df_who$LifeExp_forecast <- predict(model, newdata = df_who)

# Compare forecasts to observed values
ggplot(df_who, aes(x = LifeExp, y = LifeExp_forecast)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, color = "red", linetype = "dashed") +
  labs(title = "Observed vs. Forecasted Life Expectancy",
       x = "Observed Life Expectancy",
       y = "Forecasted Life Expectancy")
```


Observed vs. Forecasted Life Expectancy
Forecasted Life Expectancy
90
80
70
40
50
60
70
80
Observed Life Expectancy