DATA_605_Assignment_3_Fox

Amanda Fox

2025-04-13

Preparation

Load libraries:

```
# load libraries
library(tidyverse)
library(ggplot2)
library(scales)
```

Problem 1: Transportation Safety

1. Data Visualization:

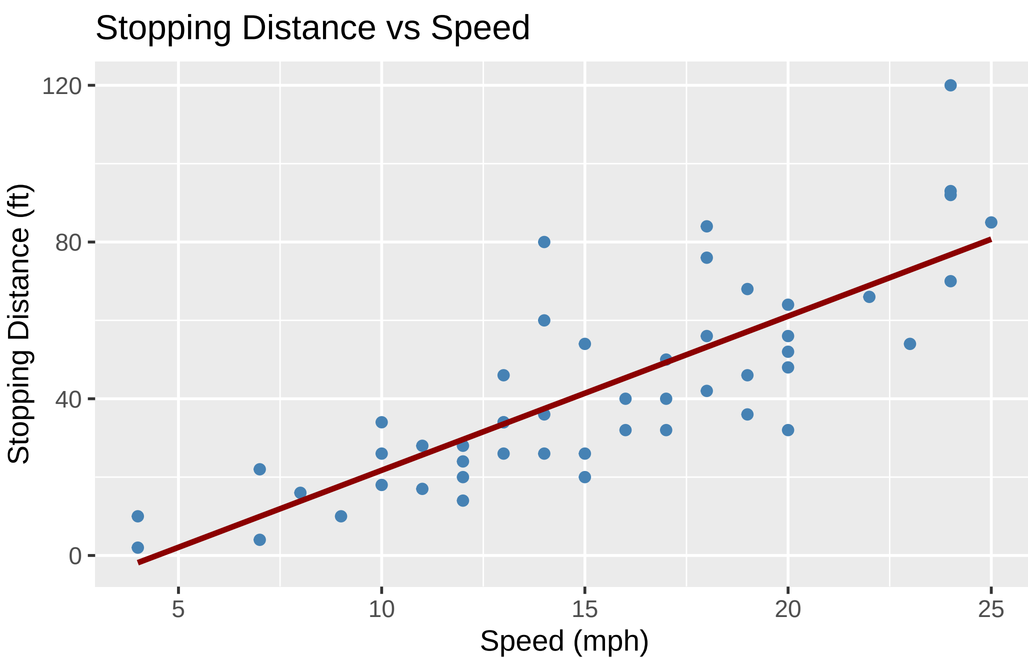
Create a scatter plot of stopping distance (dist) as a function of speed (speed). Add a regression line to the plot to visually assess the relationship.

Stopping distance increases with speed in a roughly linear pattern, with a few possible outliers and more variability at higher speeds.

```
# Load data
data("cars")
glimpse(cars)
```

```
Rows: 50
Columns: 2
$ speed <dbl> 4, 4, 7, 7, 8, 9, 10, 10, 10, 11, 11, 12, 12, 12, 12, 13, 13, 13~
$ dist  <dbl> 2, 10, 4, 22, 16, 10, 18, 26, 34, 17, 28, 14, 20, 24, 28, 26, 34~
```

```r
plt1 <- cars %>%
ggplot(aes(x = speed, y = dist)) +
  geom_point(color = "steelblue") +
  geom_smooth(method = "lm",
              se = FALSE,
              color = "darkred") +
  labs(title = "Stopping Distance vs Speed",
       x = "Speed (mph)", y = "Stopping Distance (ft)")

plt1
```



2. Build a Linear Model:

Construct a simple linear regression model where stopping distance (dist) is the dependent variable and speed (speed) is the independent variable. Summarize the model to evaluate its coefficients, R-squared value, and p-value

Stopping Distance = 3.932409 * speed - 17.58

R-squared = 0.6511 which indicatest that speed explains 65.11% of variation in stopping distance.

p-value = 1.49e-12 which indicates that this relatioship is very unlikely to be due to chance.

```r
# Model
mod_dist <- lm(dist ~ speed, data = cars)

# Summarize
summary(mod_dist)
```

```

Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max 
-29.069  -9.525  -2.272   9.215  43.201 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -17.5791     6.7584  -2.601   0.0123 *  
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,	Adjusted R-squared:  0.6438 
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

```r
# Add predicted values (verified new plot matches above)
df_cars_model <- cars %>%
  mutate(predicted_dist = predict(mod_dist))

glimpse(df_cars_model)
```

```
Rows: 50
Columns: 3
$ speed          <dbl> 4, 4, 7, 7, 8, 9, 10, 10, 10, 11, 11, 12, 12, 12, 12, 1~
$ dist           <dbl> 2, 10, 4, 22, 16, 10, 18, 26, 34, 17, 28, 14, 20, 24, 2~
$ predicted_dist <dbl> -1.849460, -1.849460, 9.947766, 9.947766, 13.880175, 17~
```

3 & 4: Model Quality Evaluation & Residual Analysis

Calculate and interpret the R-squared value to assess the proportion of variance in stopping distance explained by speed.

R-squared is 0.6511 so 65.11% of the variation in stopping distance is explained by speed.

Perform a residual analysis to check the assumptions of the linear regression model, including linearity, homoscedasticity, independence, and normality of residuals. Plot the residuals versus fitted values to check for any patterns. Create a Q-Q plot of the residuals to assess normality. Perform a Shapiro-Wilk test for normality of residuals. Plot a histogram of residuals to further check for normality.

Assumptions are generally well met. The relationship is roughly linear and residual tests below show a reasonable fit with roughly normal distribution of residuals and some heteroscedasticity.

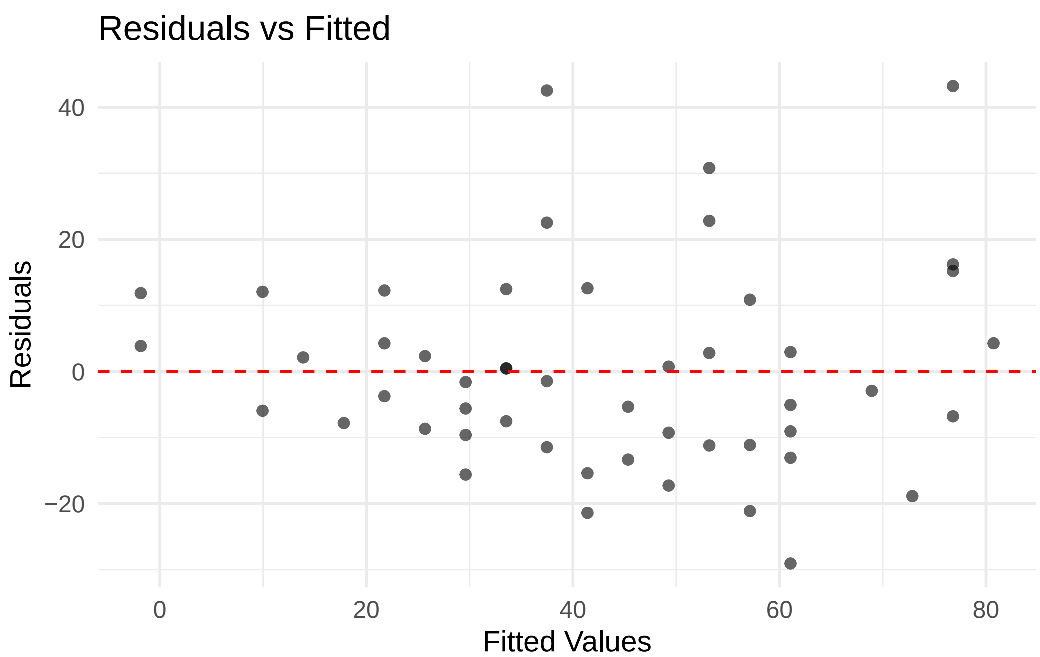
In the context of the small dataset size and visual examination of the scatterplots above, these results are acceptable. Note that independence is assumed since the behavior of one car should not impact another.

1. Residuals vs Fitted: Fan shape suggests some heteroscedasticity with increasing variance at higher speeds
2. Histogram of Residuals: Broadly normally distributed with a slight right skew
3. Q-Q: Follows the line generally but deviation at both ends, particularly the upper end, suggests some non-normality/outliers
4. Shapiro-Wilkes test: $p<0.05$ suggests non-normality (null hypothesis = data is normal, which is rejected).

```r
# Add to dataframe
df_cars_model <- df_cars_model %>%
  mutate(
    .fitted = fitted(mod_dist),
    .resid = resid(mod_dist),
    .std_resid = rstandard(mod_dist)
  )

glimpse(df_cars_model)
```

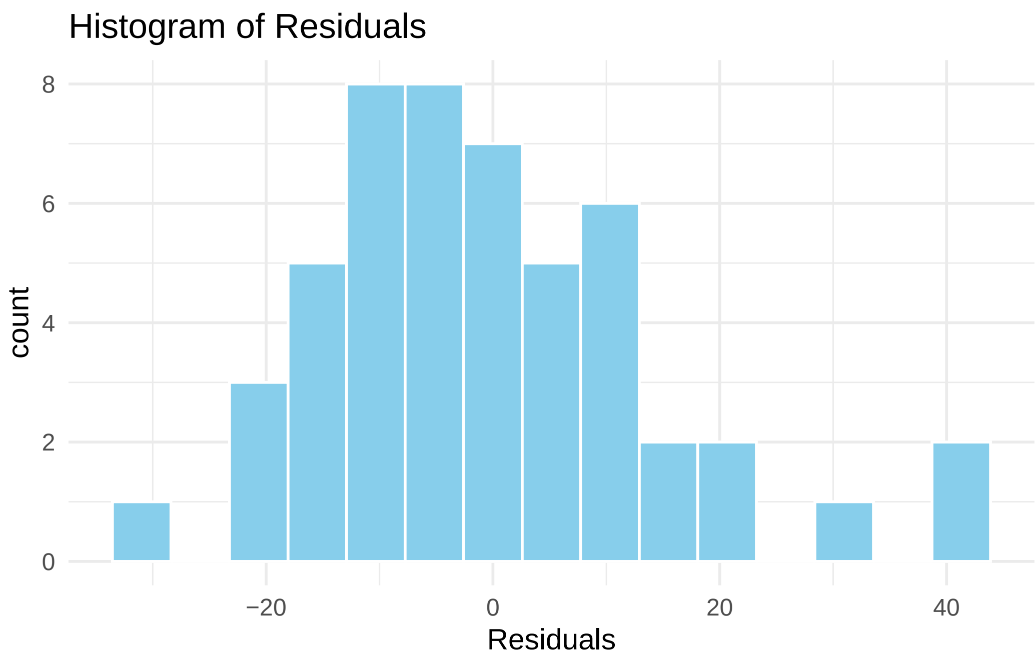
```
Rows: 50
Columns: 6
$ speed          <dbl> 4, 4, 7, 7, 8, 9, 10, 10, 10, 11, 11, 12, 12, 12, 12, 1~
$ dist           <dbl> 2, 10, 4, 22, 16, 10, 18, 26, 34, 17, 28, 14, 20, 24, 2~
$ predicted_dist <dbl> -1.849460, -1.849460, 9.947766, 9.947766, 13.880175, 17~
$ .fitted        <dbl> -1.849460, -1.849460, 9.947766, 9.947766, 13.880175, 17~
$ .resid         <dbl> 3.849460, 11.849460, -5.947766, 12.052234, 2.119825, -7~
$ .std_resid     <dbl> 0.26604155, 0.81893273, -0.40134618, 0.81326629, 0.1421~
```

```r
#---------------
# Diagnostic Plots**
#---------------

# Residuals vs Fitted
plot_resid <- df_cars_model %>%
  ggplot(aes(x = .fitted, y = .resid)) +
  geom_point(alpha = 0.6) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(
    title = "Residuals vs Fitted",
    x = "Fitted Values",
    y = "Residuals"
  ) +
  theme_minimal()

plot_resid
```

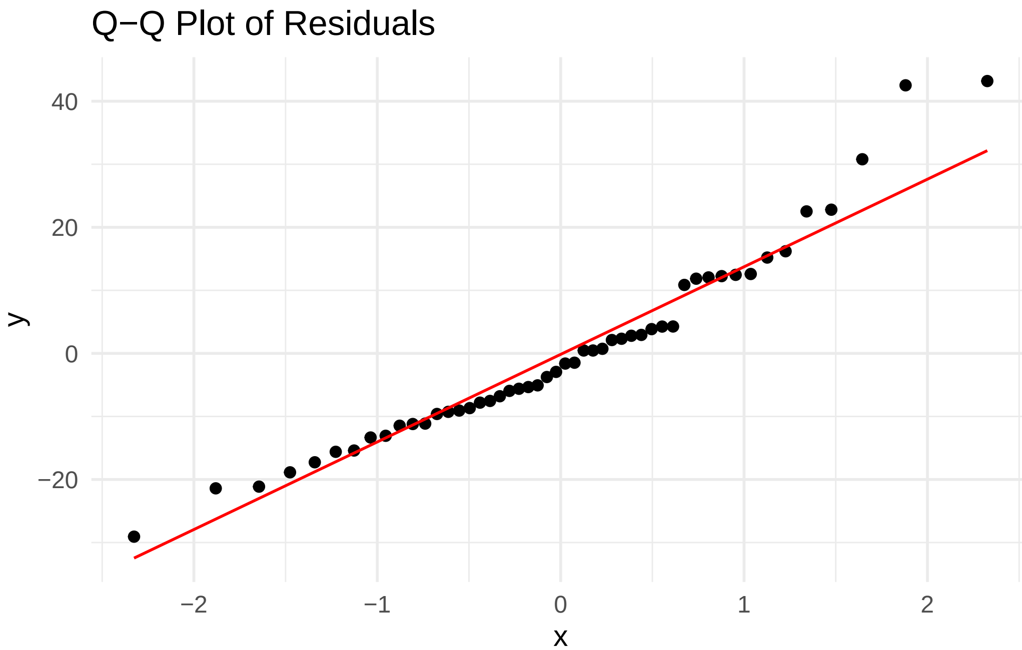



```r
# Histogram of Residuals
plot_resid_hist <- df_cars_model %>%
  ggplot(aes(x = .resid)) +
  geom_histogram(bins = 15, fill = "skyblue", color = "white") +
```

```r
  labs(title = "Histogram of Residuals", x = "Residuals") +
  theme_minimal()

plot_resid_hist
```




```r
# Q-Q
plot_qq <- df_cars_model %>%
  ggplot(aes(sample = .resid)) +
  stat_qq() +
  stat_qq_line(color = "red") +
  labs(title = "Q-Q Plot of Residuals") +
  theme_minimal()

plot_qq
```




```r
# shapiro-wilkes test
df_cars_model$.resid %>%
  shapiro.test()
```

```
	Shapiro-Wilk normality test

data:  .
W = 0.94509, p-value = 0.02152
```

5. Conclusion

Based on the model summary and residual analysis, determine whether the linear model is appropriate for this data. Discuss any potential violations of model assumptions and suggest improvements if necessary.

Based on the model summary and residual analysis, the linear model is acceptable. The R-squared value is meaningful but not strong; model fit might improve by adding more predictors such as weight, tire specs, etc.

Linearity, normality, and constant variance are generally met, with acceptable slight heteroscedasticity and non-normality in residuals. Independence was not formally tested but reasonable to assume since each car's stopping distance is unrelated to others.

Problem 2: Health Policy Analyst

1. Initial Assessment of Healthcare Expenditures and Life Expectancy

Task: Create a scatterplot of LifeExp vs. TotExp to visualize the relationship between healthcare expenditures and life expectancy across countries.

```
#---------------
# Load data
#---------------

df_who_raw <- read_csv("https://raw.githubusercontent.com/AmandaSFox/DATA605_Math/main/Assig

glimpse(df_who_raw)
```

```
Rows: 190
Columns: 12
$ Country        <chr> "Afghanistan", "Albania", "Algeria", "Andorra", "Angola~
$ LifeExp...2    <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~
$ ...10          <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ TotExp         <dbl> 112, 3297, 5292, 172314, 1656, 13046, 19654, 1944, 1907~
$ LifeExp...12   <dbl> 42, 71, 71, 82, 41, 73, 75, 69, 82, 80, 64, 74, 75, 63,~
```

```
#---------------
# Clean data
#---------------

# check for NAs
df_who_raw %>%
  summarise(across(everything(), ~sum(is.na(.))))
```

```
# A tibble: 1 x 12
  Country LifeExp...2 InfantSurvival Under5Survival TBFree PropMD PropRN PersExp
    <int>       <int>          <int>          <int>  <int>  <int>  <int>   <int>
```

```
1       0            0               0               0      0      0      0       0
# i 4 more variables: GovtExp <int>, ...10 <int>, TotExp <int>,
#   LifeExp...12 <int>
```

```
# check for duplicate rows
nrow(df_who_raw) - nrow(distinct(df_who_raw))
```

```
[1] 0
```

```
# count unique values in each column
df_who_raw %>%
  summarise(across(everything(), n_distinct))
```

```
# A tibble: 1 x 12
  Country LifeExp...2 InfantSurvival Under5Survival TBFree PropMD PropRN PersExp
    <int>       <int>          <int>          <int>  <int>  <int>  <int>   <int>
1     190          43             83             99    140    188    189     161
# i 4 more variables: GovtExp <int>, ...10 <int>, TotExp <int>,
#   LifeExp...12 <int>
```

```
# verify col 10 includes only NA values
unique(df_who_raw$...10)
```

```
[1] NA
```

```
# verify both LifeExp columns are identical
all(df_who_raw$`LifeExp...2` == df_who_raw$`LifeExp...12`)
```

```
[1] TRUE
```

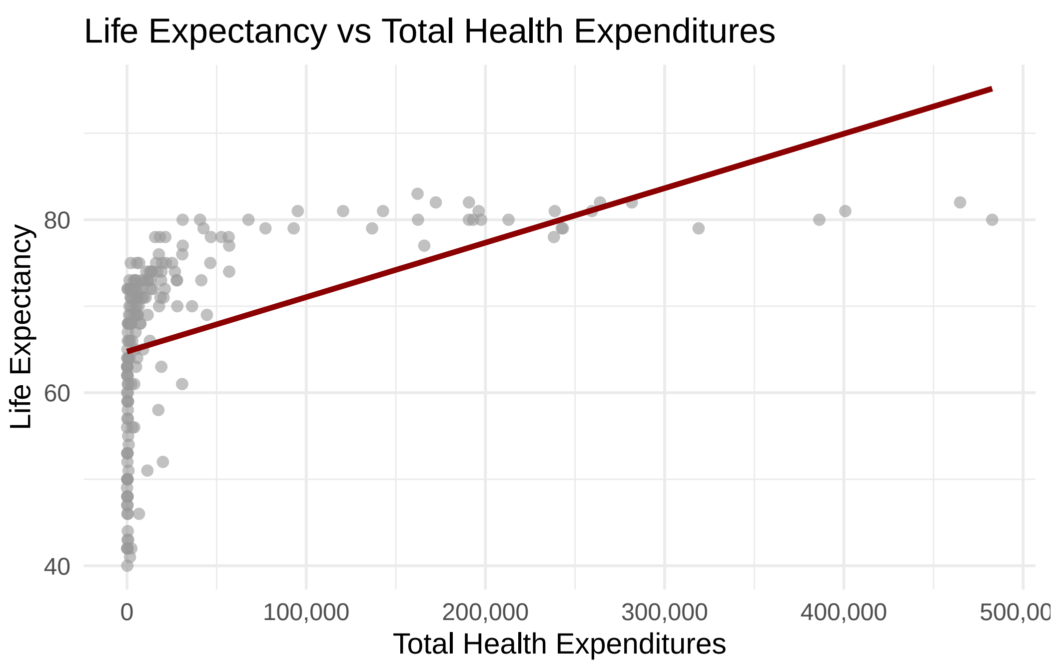
```
# drop second LifeExp and NA columns, rename LifeExp
df_who_clean <- df_who_raw %>%
  select(- LifeExp...12,- ...10) %>%
  rename(LifeExp = `LifeExp...2`)

glimpse(df_who_clean)
```

```
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~
```

```r
#---------------
# Summary and Plot
#---------------

# summarize
summary(df_who_clean)
```

```
   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
```

```r
# Scatterplot with linear regression line
plt3 <- df_who_clean %>%
  ggplot(aes(x = TotExp, y = LifeExp)) +
  geom_point(alpha = 0.6, color = "gray60") +
  geom_smooth(method = "lm", se = FALSE, color = "darkred") +
  scale_x_continuous(labels = scales::comma)+
    labs(
    title = "Life Expectancy vs Total Health Expenditures",
    x = "Total Health Expenditures",
    y = "Life Expectancy"
  ) +
  theme_minimal()

plt3
```



Run a simple linear regression with LifeExp as the dependent variable and TotExp as the independent variable (without transforming the variables).

$\text{LifeExp} = 64.75 + 0.00006297 \times \text{TotExp}$ For every additional $1 in total healthcare spending, life expectancy increases by about 0.000063 years or about 0.023 days above a "baseline" of 64.75 years at $0 spending.

Provide and interpret the F-statistic, R-squared value, standard error, and p-values.

- F-statistic and p value: 65.26 on 1 and 188 DF with p of 7.7e-14 indicates the relationship between total expenditure and life expectancy is extremely unlikely to be due to chance
- R-squared: 0.2577 indicates that total healthcare expenditure explains only 25.77% of the variance in life expectancy, which is weak
- Standard error: 9.37 indicates that average error is 9.37 years which is significant in this case as the interquartile range is 13.75 years.

Discuss whether the assumptions of simple linear regression (linearity, independence, homoscedasticity, and normality of residuals) are met in this analysis.

Overall the model violates all assumptions except independence (countries are independent observations so independence is assumed). Linearity: The relationship is clearly non-linear in the scatterplot Homoscedasticity: The residual plot shows an uneven distribution of residuals around the line Normality: The histogram has a long left tail, and the qq plot varies widely around the line. The Shapiro-Wilkes small p value also indicates non normality.

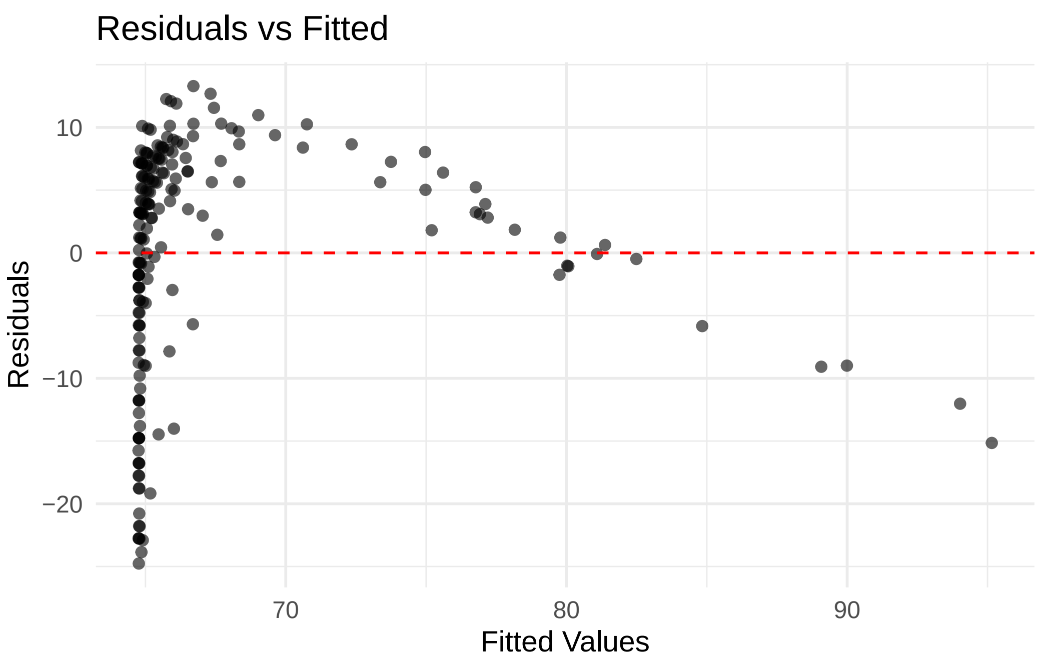
```r
#---------------
# Model
#---------------
mod_life <- lm(LifeExp ~ TotExp, data = df_who_clean)

summary(mod_life)
```

```

Call:
lm(formula = LifeExp ~ TotExp, data = df_who_clean)

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
```

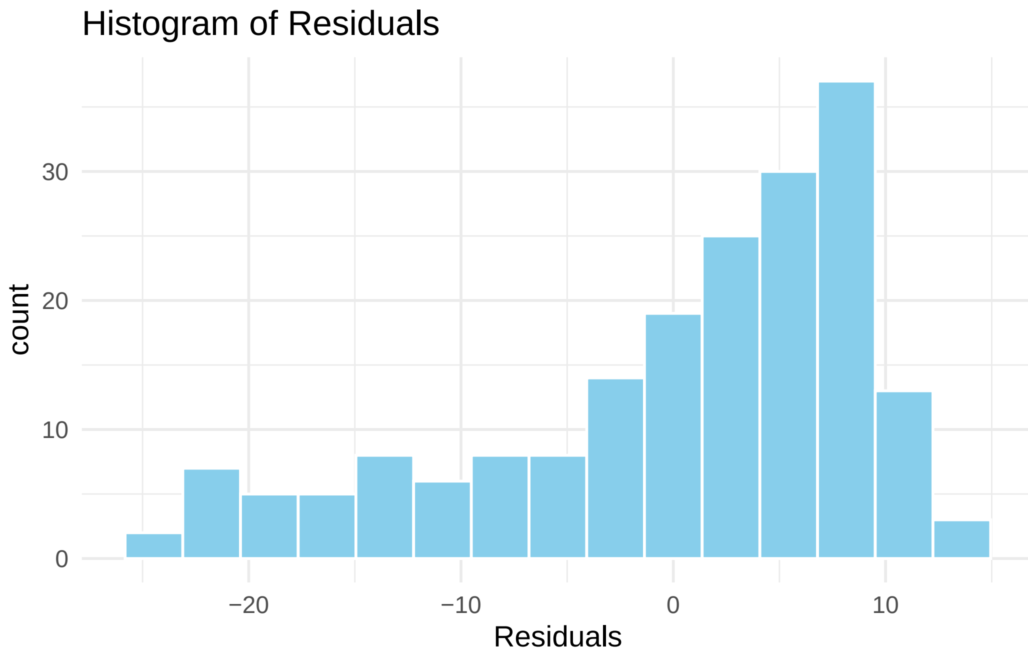
```r
#---------------
# Diagnostics
#---------------

# add predicted and residuals to df
df_who_model <- df_who_clean %>%
  mutate(predicted_life = predict(mod_life),
         .fitted = fitted(mod_life),
         .resid = resid(mod_life),
         .std_resid = rstandard(mod_life))

glimpse(df_who_model)
```

```
Rows: 190
Columns: 14
$ 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~
$ predicted_life <dbl> 64.76043, 64.96099, 65.08661, 75.60402, 64.85765, 65.57~
$ .fitted        <dbl> 64.76043, 64.96099, 65.08661, 75.60402, 64.85765, 65.57~
$ .resid         <dbl> -22.7604272, 6.0390128, 5.9133872, 6.3959804, -23.85765~
$ .std_resid     <dbl> -2.43668924, 0.64646820, 0.63298727, 0.68842673, -2.554~
```

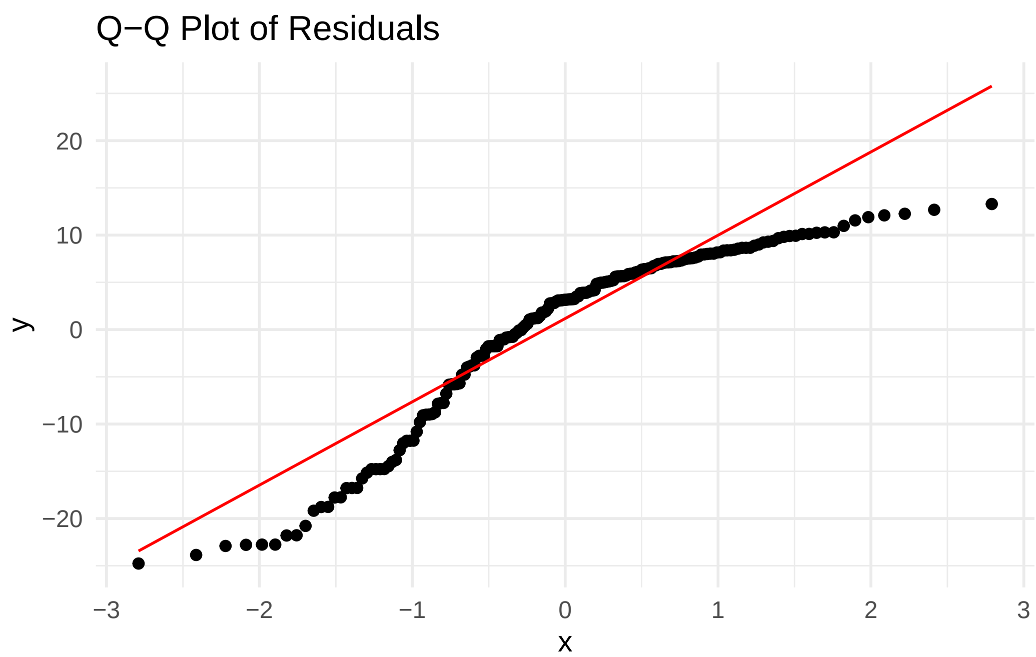
```r
# Residuals vs Fitted
plot_resid_who <- df_who_model %>%
  ggplot(aes(x = .fitted, y = .resid)) +
  geom_point(alpha = 0.6) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Residuals vs Fitted", x = "Fitted Values", y = "Residuals") +
  theme_minimal()

plot_resid_who
```




```r
# Histogram of Residuals
plot_hist_who <- df_who_model %>%
  ggplot(aes(x = .resid)) +
  geom_histogram(bins = 15, fill = "skyblue", color = "white") +
  labs(title = "Histogram of Residuals", x = "Residuals") +
  theme_minimal()

plot_hist_who
```




```
# QQ plot
plot_qq_who <- df_who_model %>%
  ggplot(aes(sample = .resid)) +
  stat_qq() +
  stat_qq_line(color = "red") +
  labs(title = "Q-Q Plot of Residuals") +
  theme_minimal()

plot_qq_who
```




```
# Shapiro Wilkes
shapiro.test(df_who_model$.resid)
```

```
    Shapiro-Wilk normality test

data:  df_who_model$.resid
W = 0.89146, p-value = 1.609e-10
```

Discussion: Consider the implications of your findings for health policy. Are higher healthcare expenditures generally associated with longer life expectancy? What do the assumptions of the regression model suggest about the reliability of this relationship?

Higher healthcare expenditures are related to longer life expectancy but the relationship is not linear (see scatterplot above). The linear model is significant (low p value) but the fit is poor with a low R-squared and high standard error.

To make a final conclusion on which to base policy, additional work is needed such as transformations or a non-linear model.

2. Transforming Variables for a Better Fit

Task: Recognizing potential non-linear relationships, transform the variables as follows: Raise life expectancy to the 4.6 power (LifeExp^4.6). Raise total expenditures to the 0.06 power (TotExp^0.06), which is nearly a logarithmic transformation. Create a new scatterplot with the transformed variables and re-run the simple linear regression model.

The transformed data now has a visually linear relationship on the scatterplot with a more even distribution around the linear regression line.

Provide and interpret the F-statistic, R-squared value, standard error, and p-values for the transformed model.

- F-statistic and p value: 65.26 on 1 and 188 DF with p of 7.7e-14 indicates the relationship between total expenditure and life expectancy is extremely unlikely to be due to chance
- R-squared: 0.2577 indicates that total healthcare expenditure explains only 25.77% of the variance in life expectancy, which is weak
- Standard error: 9.37 indicates that average error is 9.37 years which is significant in this case as the interquartile range is 13.75 years.

Compare this model to the original model (from Question 1). Which model provides a better fit, and why?

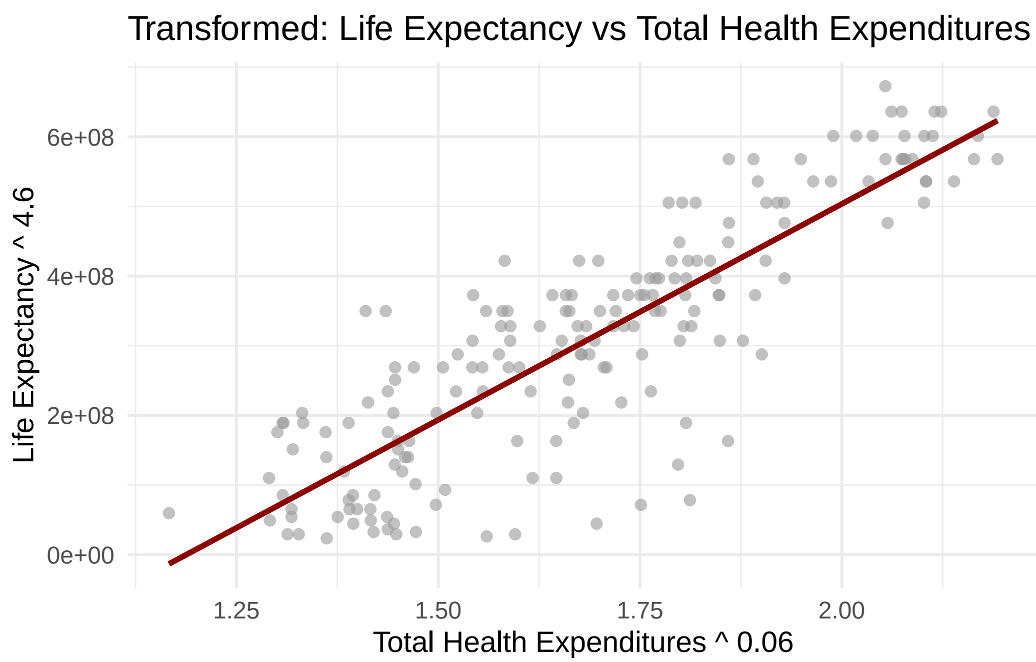
- F-statistic and p value: 65.26 on 1 and 188 DF with p of 7.7e-14 indicates the relationship between total expenditure and life expectancy is extremely unlikely to be due to chance
- R-squared: 0.2577 indicates that total healthcare expenditure explains only 25.77% of the variance in life expectancy, which is weak
- Standard error: 9.37 indicates that average error is 9.37 years which is significant in this case as the interquartile range is 13.75 years.


```r
#---------------
# Transform and Plot
#---------------

df_who_transform <- df_who_clean %>%
  mutate(LifeExp_46 = LifeExp^4.6,
         TotExp_006 = TotExp^0.06)

glimpse(df_who_transform)
```

```
Rows: 190
Columns: 12
$ 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~
$ LifeExp_46     <dbl> 29305338, 327935478, 327935478, 636126841, 26230450, 37~
$ TotExp_006     <dbl> 1.327251, 1.625875, 1.672697, 2.061481, 1.560068, 1.765~
```

```r
# Scatterplot with linear regression line
plt4 <- df_who_transform %>%
  ggplot(aes(x = TotExp_006, y = LifeExp_46)) +
  geom_point(alpha = 0.6, color = "gray60") +
  geom_smooth(method = "lm", se = FALSE, color = "darkred") +
  scale_x_continuous(labels = scales::comma)+
    labs(
    title = "Transformed: Life Expectancy vs Total Health Expenditures",
    x = "Total Health Expenditures ^ 0.06",
    y = "Life Expectancy ^ 4.6"
  ) +
  theme_minimal()

plt4
```




```r
#---------------
# Model
#---------------

mod_life_transform <- lm(LifeExp_46 ~ TotExp_006, data = df_who_transform)

summary(mod_life_transform)
```

```
Call:
lm(formula = LifeExp_46 ~ TotExp_006, data = df_who_transform)

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_006   620060216   27518940   22.53   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

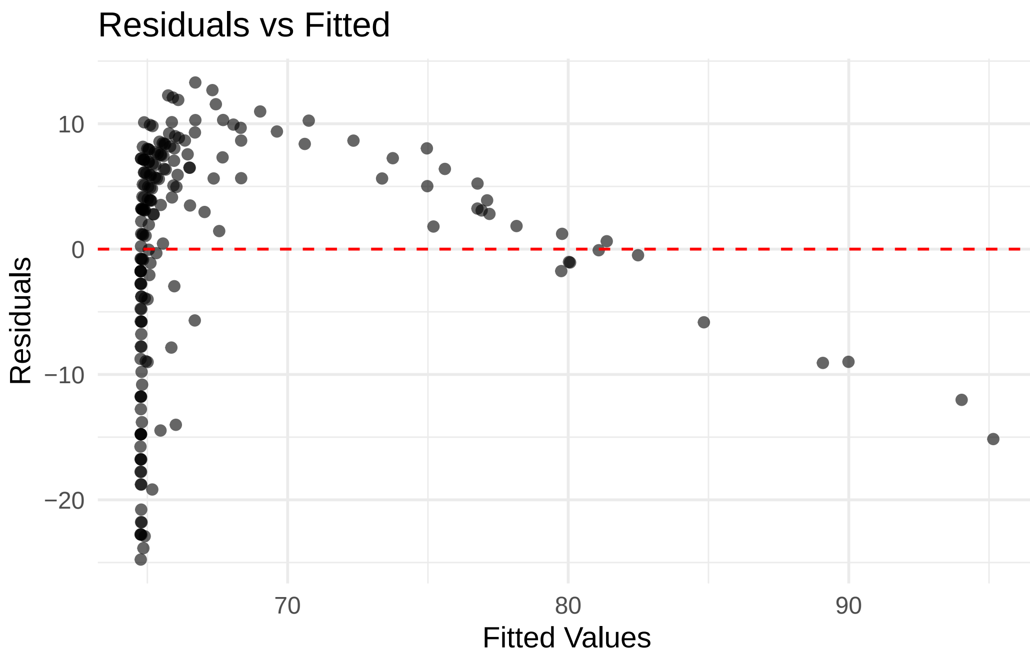
```
Residual standard error: 90490000 on 188 degrees of freedom
Multiple R-squared:  0.7298,	Adjusted R-squared:  0.7283 
F-statistic: 507.7 on 1 and 188 DF,  p-value: < 2.2e-16
```

```r
#---------------
# Diagnostics
#---------------

# add predicted and residuals to df
df_who_model <- df_who_clean %>%
  mutate(predicted_life = predict(mod_life),
         .fitted = fitted(mod_life),
         .resid = resid(mod_life),
         .std_resid = rstandard(mod_life))

glimpse(df_who_model)
```

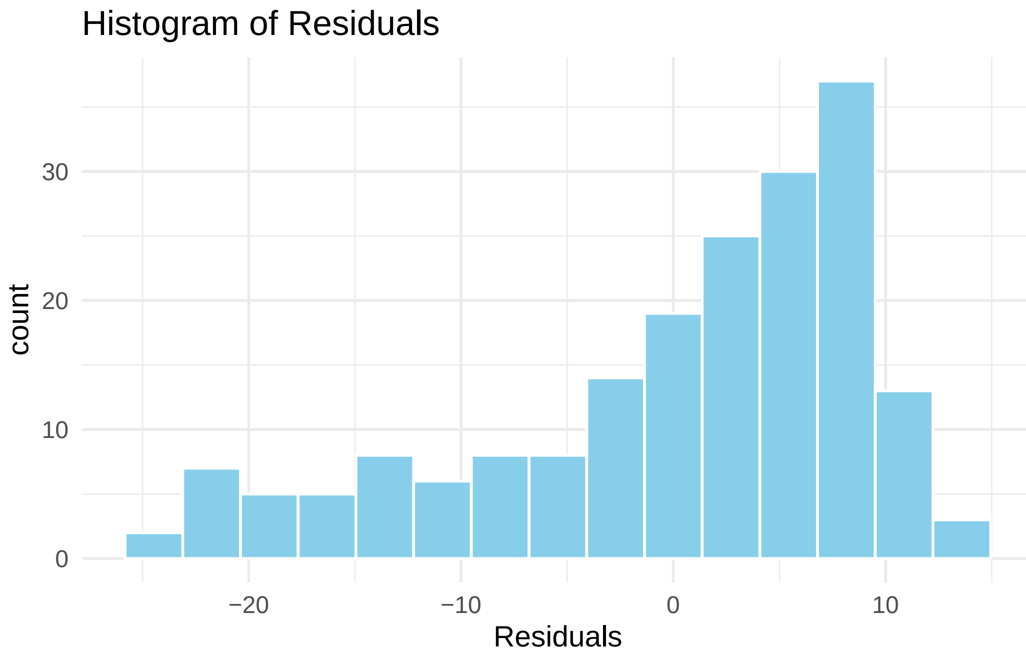
```
Rows: 190
Columns: 14
$ Country        <chr> "Afghanistan", "Albania", "Algeria", "Andorra", "Angola~
$ LifeExp        <dbl> 42, 71, 71, 82, 41, 73, 75, 69, 82, 80, 64, 74, 75, 63,~
$ InfantSurvival <dbl> 0.835, 0.985, 0.967, 0.997, 0.846, 0.990, 0.986, 0.979,~
$ Under5Survival <dbl> 0.743, 0.983, 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~
$ predicted_life <dbl> 64.76043, 64.96099, 65.08661, 75.60402, 64.85765, 65.57~
$ .fitted        <dbl> 64.76043, 64.96099, 65.08661, 75.60402, 64.85765, 65.57~
$ .resid         <dbl> -22.7604272, 6.0390128, 5.9133872, 6.3959804, -23.85765~
$ .std_resid     <dbl> -2.43668924, 0.64646820, 0.63298727, 0.68842673, -2.554~
```

```r
# Residuals vs Fitted
plot_resid_who <- df_who_model %>%
  ggplot(aes(x = .fitted, y = .resid)) +
  geom_point(alpha = 0.6) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Residuals vs Fitted", x = "Fitted Values", y = "Residuals") +
  theme_minimal()
```

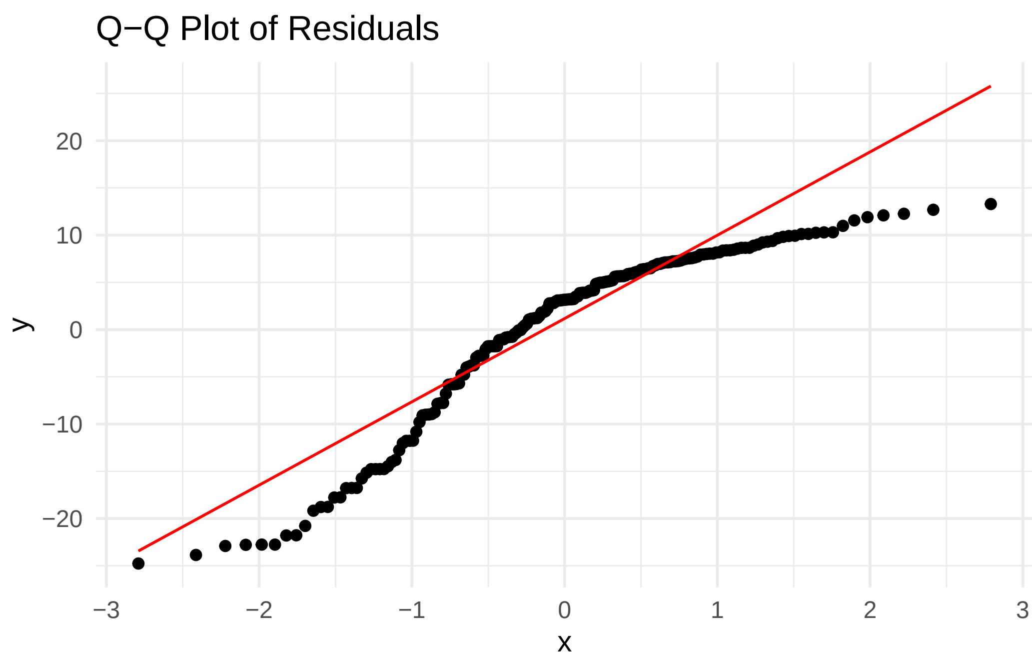
```
plot_resid_who
```




```
# Histogram of Residuals
plot_hist_who <- df_who_model %>%
  ggplot(aes(x = .resid)) +
  geom_histogram(bins = 15, fill = "skyblue", color = "white") +
  labs(title = "Histogram of Residuals", x = "Residuals") +
  theme_minimal()

plot_hist_who
```




```
# QQ plot
plot_qq_who <- df_who_model %>%
  ggplot(aes(sample = .resid)) +
  stat_qq() +
  stat_qq_line(color = "red") +
  labs(title = "Q-Q Plot of Residuals") +
  theme_minimal()

plot_qq_who
```




```r
# Shapiro Wilkes
shapiro.test(df_who_model$.resid)
```

```
	Shapiro-Wilk normality test

data:  df_who_model$.resid
W = 0.89146, p-value = 1.609e-10
```

Discussion: How do the transformations impact the interpretation of the relationship between healthcare spending and life expectancy? Why might the transformed model be more appropriate for policy recommendations?