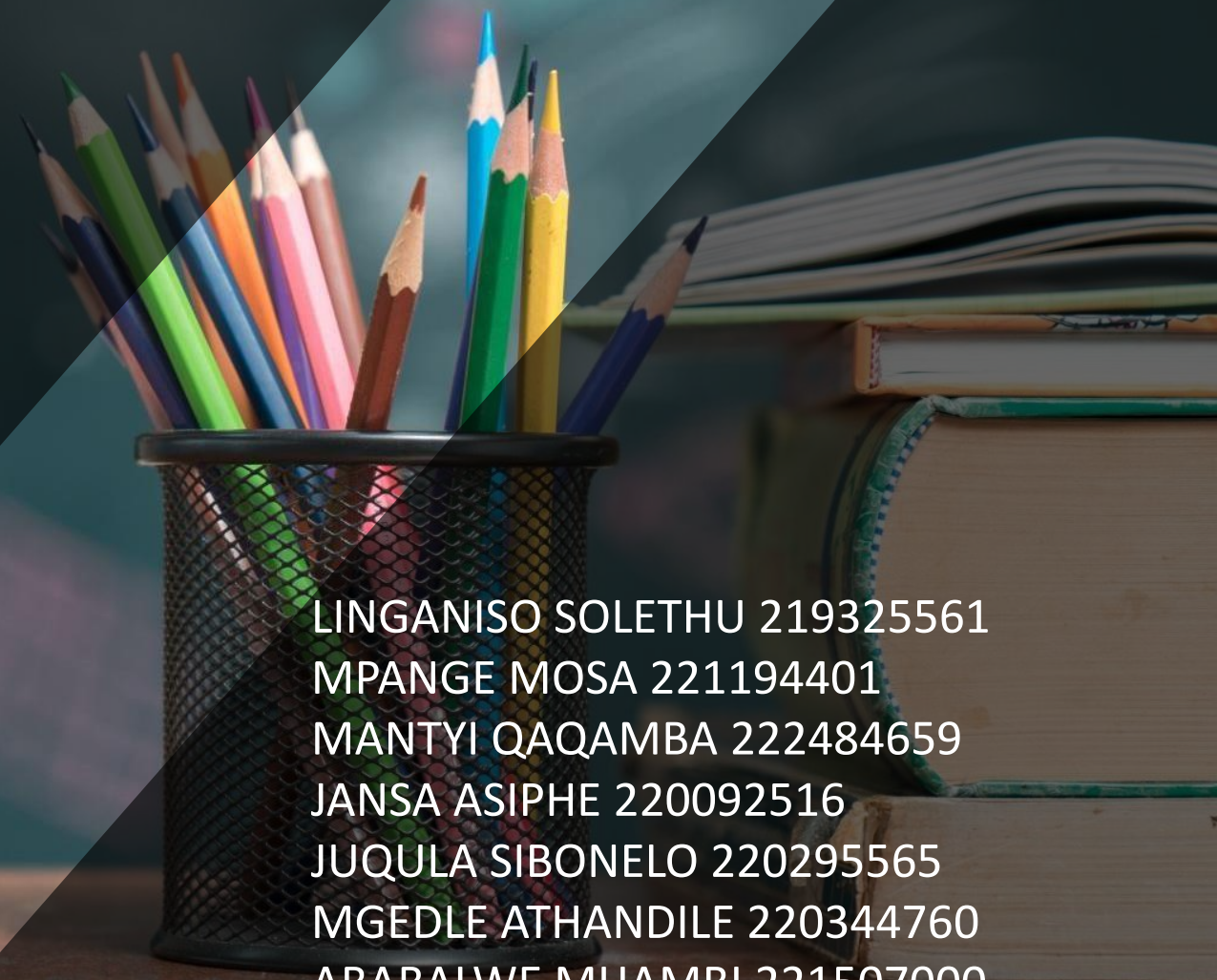
MULTILINEAR REGRESSION RESEARCH ON FACTORS INFLUENCING LIFE EXPECTANCY ON DIFFERENT COUNTRIES



LINGANISO SOLETHU 219325561
MPANGE MOSA 221194401
MANTYI QAQAMBA 222484659
JANSA ASIPHE 220092516
JUQULA SIBONELO 220295565
MGEDLE ATHANDILE 220344760
ABABALWE MHAMBI 221507000

Contents

ABSTRACT

Delving into the labyrinth of life expectancy disparities, we aim to unravel the intricate web of factors shaping longevity across diverse global landscapes. By conducting a comprehensive analysis across diverse countries, our research seeks to offer a nuanced perspective on global life expectancy dynamics, shedding light on the multifaceted determinants shaping both individual and population health outcomes.

To achieve this objective, we will undertake a rigorous examination of life expectancy data sourced from a multitude of authoritative outlets, including governmental records, international repositories, and scholarly literature. Employing robust statistical methodologies such as regression analysis and advanced data visualization techniques, we aim to pinpoint salient variables, unveil intricate relationships, and discern notable trends or anomalies.

Furthermore, our inquiry will extend beyond quantitative analysis to encompass a qualitative exploration of factors influencing life expectancy. This holistic approach will encompass an examination of healthcare infrastructure, socio-economic variables, educational attainment, and lifestyle choices. Through this comprehensive lens, our study aspires to furnish a profound understanding of the multifaceted determinants contributing to disparities in life expectancy across diverse global contexts.

INTRODUCTION

Life expectancy is a very important measure for assessing a country's overall. population health. Predicting the life expectancy helps the policy makers of a country. when making policies for a country hence life expectancy measure is essential. This measure then assists the policy makers of the country to decides the necessary social and health welfare changes in the country.

According to the United Nations, life expectancy at birth numbers corresponds to midyear estimations. They align with the relevant United Nations quinquennial population forecasts and fertility averages. Life tables are created using available mortality and data from civil registration. The World Health Organization supplies all the life tables, using data derived from the year 1,800 life tables based on the logit system.

A multitude of factors shape life expectancy, spanning various demographic, socioeconomic, and health-related aspects. These determinants exhibit notable diversity across nations owing to disparities in economic advancement, healthcare systems, social regulations, cultural practices, and epidemiological characteristics.

Hence, undertaking thorough analyses to pinpoint the primary influencers of life expectancy in diverse contexts is imperative for customizing interventions that efficiently cater to specific population health requirements.

The results of this study carry notable ramifications for public health policy and implementation.

Through understanding the factors influencing life expectancy, policymakers can craft tailored interventions to tackle primary causes of mortality and advocate for healthier behaviours, consequently, fostering enhancements in overall population health and welfare. Moreover, comprehending the determinants of life expectancy can facilitate comparisons between countries, pinpoint intervention areas, and aid in allocating resources effectively to optimize the outcomes of public health endeavours.

By doing this research, we hope to increase our knowledge of the factors that affect. life expectancy and make a valuable contribution to evidence-based strategies for improving population health and well-being in a range of sociodemographic settings.

We seek to promote more resilience against health difficulties in an increasingly interconnected society by facilitating informed decision-making and illuminating the intricate interactions between diseases and broader socioeconomic and environmental factors.

DATA CLEANING

Data Dictionary

Variable name	Variable description	Variable type
Country	Country, we observed	category
Year	Years we observed	category
Status	Developed or developing	category
Life expectancy	Life expectancy in age	numeric

Adult mortality	Adult mortality rate	numeric
Infant mortality	Infant mortality rate	numeric
Alcohol	Per capita alcohol consumption	numeric
Percentage expenditure	As percentage	numeric
Hepatitis B	Immunization coverage	categoric
Measles	Number of cases reported	numeric
BMI	Average body mass index	numeric
Under five deaths	Number of Under five deaths	numeric
Polio	Polio immunization coverage	categoric
Total expenditure	Government health expenditure	numeric
Diphtheria	Diphtheria coverage	categoric
HIV/AIDS	HIV/AIDS deaths	numeric
GDP	GDP/capita	numeric
Population	Population of the country	numeric

TABLE 1

Dataset

The Dataset that we used had 21 variables ,18 which are of numeric and 3 of which are a discrete data (category).

Variable selection criteria

The selection criteria for variables in the research encompass several key considerations to ensure a comprehensive understanding of the factors influencing life expectancy across different countries. Among the variables included are BMI (Body Mass Index), which reflects the population's overall health and nutrition status. Adult mortality rates provide insights into the prevalence of diseases and the effectiveness of healthcare systems in addressing them.
Similarly, infant death rates offer crucial information on maternal and child health, including access to prenatal and neonatal care.

Alcohol consumption rates serve as an indicator of lifestyle and behavioural factors that impact health outcomes. Percentage expenditure on healthcare highlights the level of investment in healthcare infrastructure and services, which directly affects access and quality of care. Disease specific factors such as hepatitis, measles, polio, and diphtheria vaccination rates shed light on the effectiveness of immunization programs and public health initiatives in preventing communicable diseases.

Moreover, the prevalence of HIV/AIDS serves as a critical indicator of the burden of infectious diseases and the effectiveness of prevention and treatment efforts. Economic indicators such as GDP (Gross Domestic Product) provide context on the overall wealth and development status of countries, which influences access to resources and healthcare services. Population size is included to account for demographic variations and scale effects in the analysis.

By incorporating these variables, researchers can explore the multifaceted determinants of life expectancy, ranging from individual health behaviours to broader socio-economic and healthcare system factors. This comprehensive approach enables the identification

of significant predictors and informs evidence-based public health policies aimed at improving population health outcomes worldwide. Then we proceeded to remove all variables that do not provide any additional information on predictors about life expectancy

The Country has a high cardinality problem, there are more countries repeated, every country has economic, social, and health-related data that has been included separately therefore, country variable does not provide any viable information that we can use to the analysis hence we removed it. The variable Year is a time series, and our analysis focuses on time independent economic, social and health related predictors of life expectancy, removing Country and Year helps on making dataset more streamlined for analysis. The Status variable is of no numeric data which would not affect our calculations as well as analysis in life expectancy. We then proceeded to remove Income composition of resource, Thinness of 5-9 years, Thinness 1-18 years, Total expenditure and under five deaths. These Factors were the least contributing factors towards our model.

The Polio, Diphtheria and Hepatitis B are to be regarded as discrete(categorically) since it's the coverage of immunization which is not continuous.

Exploratory Data Analysis

Model fitting.

```
> mydata=read.table("clipboard", header = TRUE)
> attach(mydata)
The following objects are masked from mydata (pos = 3):

    x1, x10, x11, x12, x2, x3, x4, x5, x6, x7, x8, x9, y

> fl = lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x11 + x12)
> fl

Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 +
    x10 + x11 + x12)

Coefficients:
(Intercept)           x1           x2           x3           x4           x5           x6
   x7           x8           x9          x10          x11
  6.379e+01   -2.484e-02   -6.620e-03    2.975e-01   -1.215e-04   -7.287e-03    1.714e-05    9.73
2e-02    2.293e-02    4.225e-02   -4.531e-01    1.463e-04
        x12
  5.963e-09

>
```

Summary of the initial model

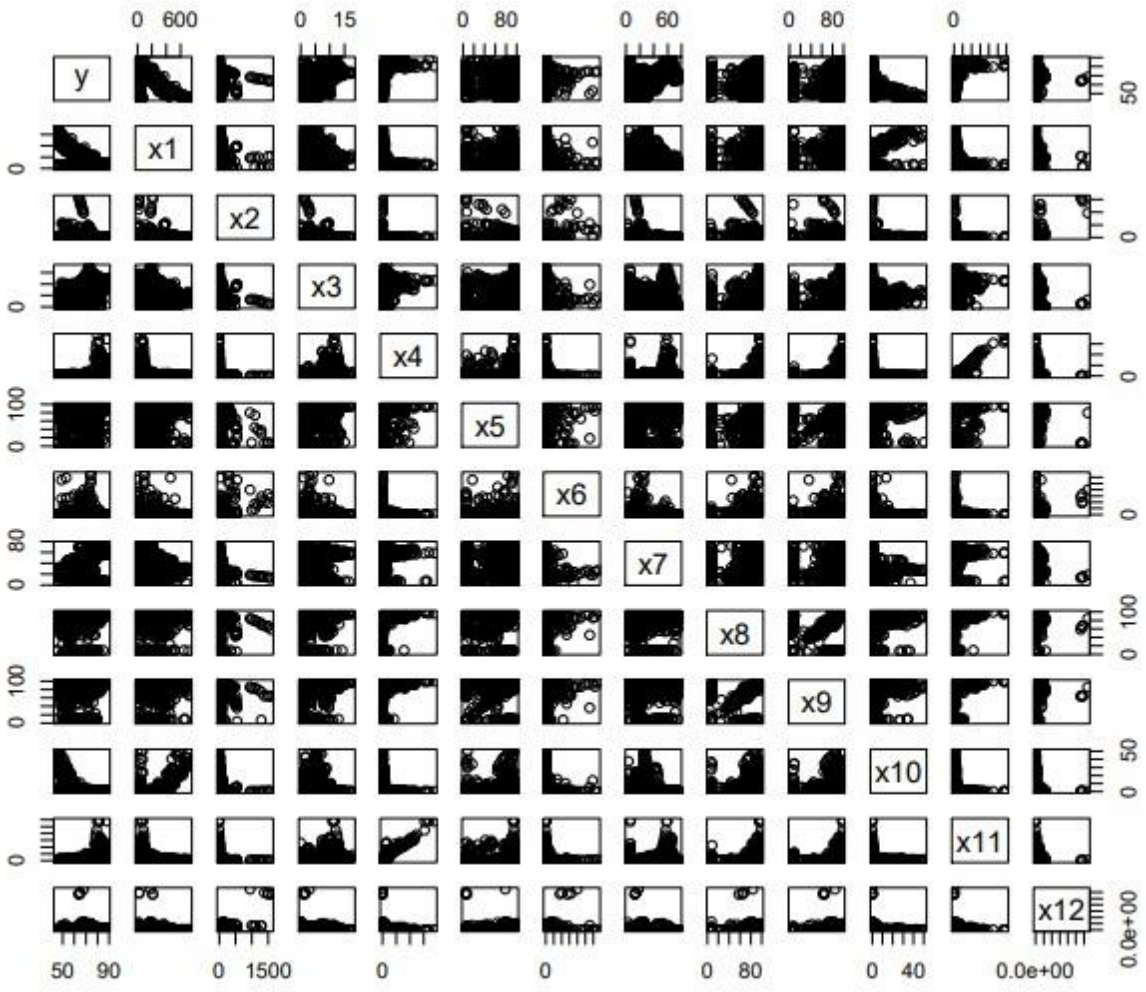
```
> summary(f1)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 +
    x10 + x11 + x12)

Residuals:
     Min       1Q   Median       3Q      Max
-21.2824  -2.5064   0.2694   2.7378  17.6119

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.379e+01  6.194e-01 102.982  < 2e-16 ***
x1          -2.484e-02  1.128e-03 -22.012  < 2e-16 ***
x2          -6.620e-03  1.434e-03  -4.617 4.19e-06 ***
x3           2.975e-01  3.244e-02   9.171  < 2e-16 ***
x4          -1.215e-04  2.230e-04  -0.545 0.585938
x5          -7.287e-03  5.515e-03  -1.321 0.186555
x6           1.714e-05  1.297e-05   1.321 0.186553
x7           9.732e-02  6.488e-03  14.999  < 2e-16 ***
x8           2.293e-02  6.369e-03   3.601 0.000326 ***
x9           4.225e-02  7.268e-03   5.813 7.37e-09 ***
x10         -4.531e-01  2.203e-02 -20.565  < 2e-16 ***
x11          1.463e-04  3.478e-05   4.206 2.74e-05 ***
x12          5.963e-09  2.135e-09   2.792 0.005296 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.472 on 1636 degrees of freedom
Multiple R-squared:  0.7434,    Adjusted R-squared:  0.7415
F-statistic:   395 on 12 and 1636 DF,  p-value: < 2.2e-16

>
```

FIGURE 1

The Summary of the initial fitted model using 12 variables against Life expectancy

$\dot{Y}$ = 6.379e+01 -2.484e-02Adult Mortality -6.620e-01Infant Deaths +2.975e-01Alcohol -1.215e04Percentage Expenditure -7.287e-03Hepatitis B +1.714e-05Measles +9.732e-02BMI +2.293e-02Polio+4.225e-02Diphtheria -4.531e-01HIV/AIDS +1.463e-04GDP +5.963e09Population

```
> mydata=read.table("clipboard", header = TRUE)
> attach(mydata)
The following objects are masked from mydata (pos = 3):

    x1, x10, x11, x12, x2, x3, x4, x5, x6, x7, x8, x9, y

The following objects are masked from mydata (pos = 6):

    x1, x10, x11, x12, x2, x3, x4, x5, x6, x7, x8, x9, y

> head(mydata)
     y  x1 x2   x3        x4 x5   x6   x7 x8 x9 x10       x11      x12
1 65.0 263 62 0.01 71.279624 65 1154 19.1  6 65 0.1 584.25921 33736494
2 59.9 271 64 0.01 73.523582 62  492 18.6 58 62 0.1 612.69651   327582
3 59.9 268 66 0.01 73.219243 64  430 18.1 62 64 0.1 631.74498 31731688
4 59.5 272 69 0.01 78.184215 67 2787 17.6 67 67 0.1 669.95900  3696958
5 59.2 275 71 0.01  7.097109 68 3013 17.2 68 68 0.1  63.53723  2978599
6 58.8 279 74 0.01 79.679367 66 1989 16.7 66 66 0.1 553.32894  2883167
> plot(mydata)
>
```



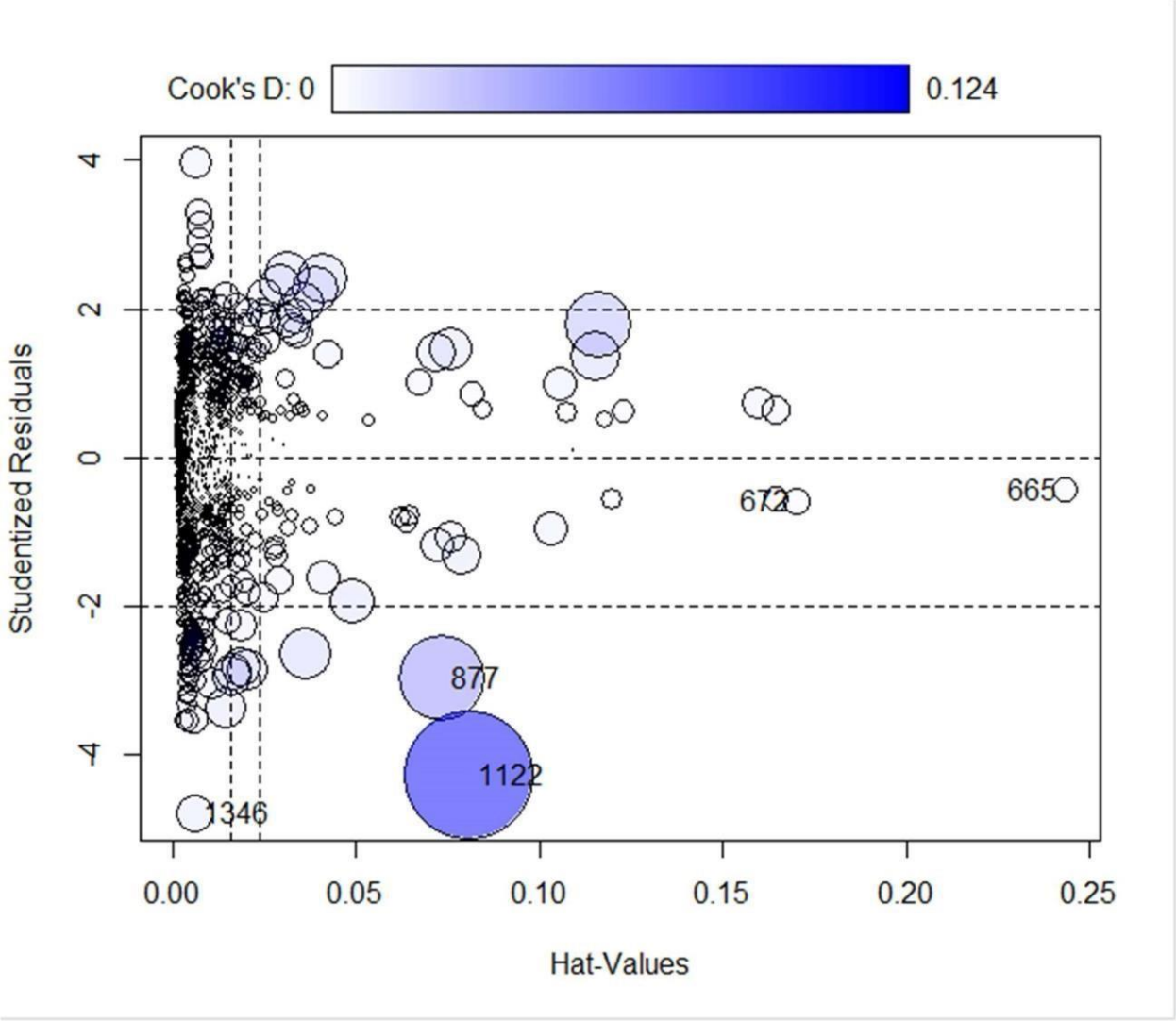
Life Expectancy is negatively related to the variables Adult Mortality, Alcohol, Infant deaths, Percentage Expenditure, Hepatitis B, and Measles, HIV/AIDS and Population and others are positively related to the life expectancy. Infant mortality is linked to life expectancy. The Newborn Deaths variable is expected to have a negative relationship with Life Expectancy since a higher number of baby deaths suggests that many infants die before reaching the age of one, reducing the country's life expectancy.

The Multiple R-squared is at 0.7434 close to 1 suggesting that the model is significant.

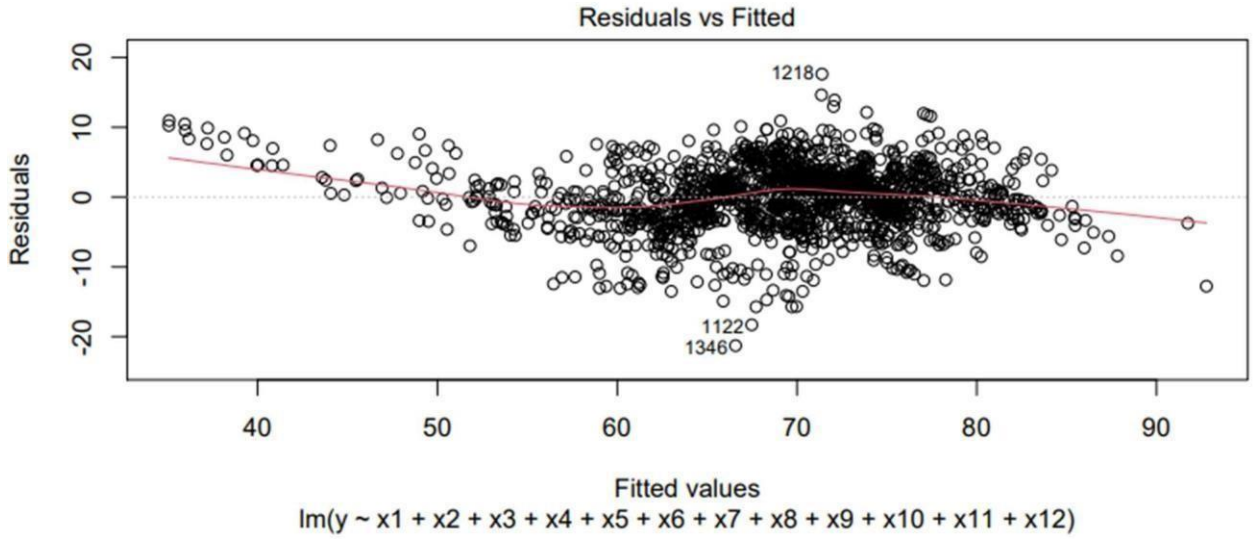
ASSUMPTION CHECKING

We started by checking for outliers using DFFITS and DFBETAS

Influence Plot



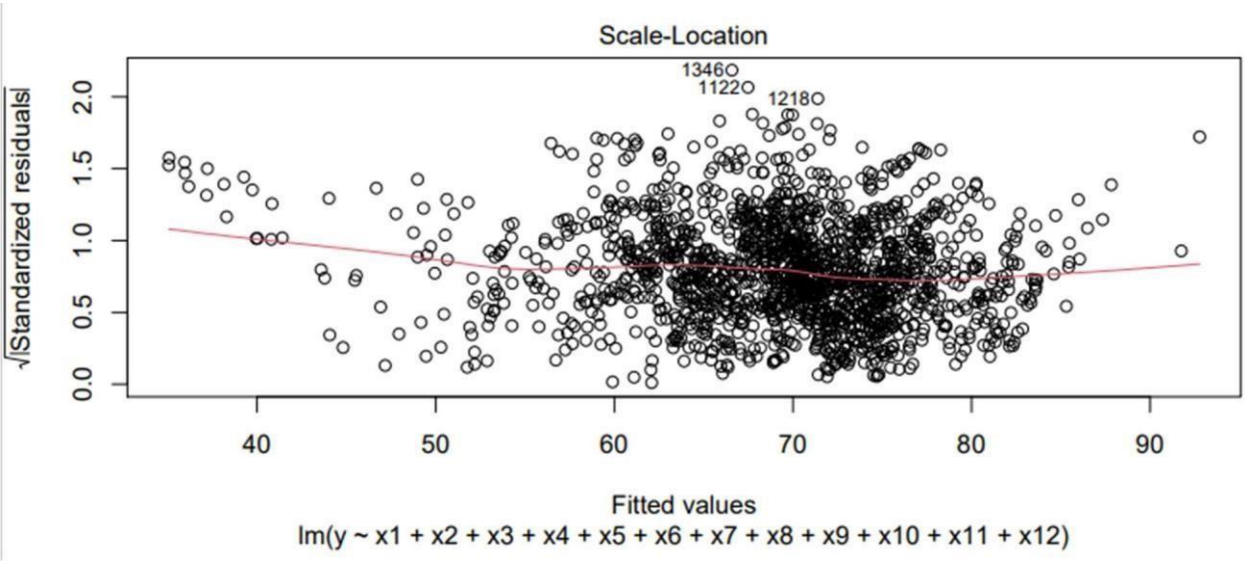
GRAPH 1



GRAPH 2

The scatterplot of the residuals indicates that most data points are closely clustered around the horizontal line at y = 0, suggesting a good fit between the model and the data without any systematic bias. However, a few outliers are observed, indicating instances where the model did not fit the data accurately. Despite these outliers, there's no discernible pattern to their distribution, making it challenging to pinpoint their cause. Overall, the model generally fits the data well.

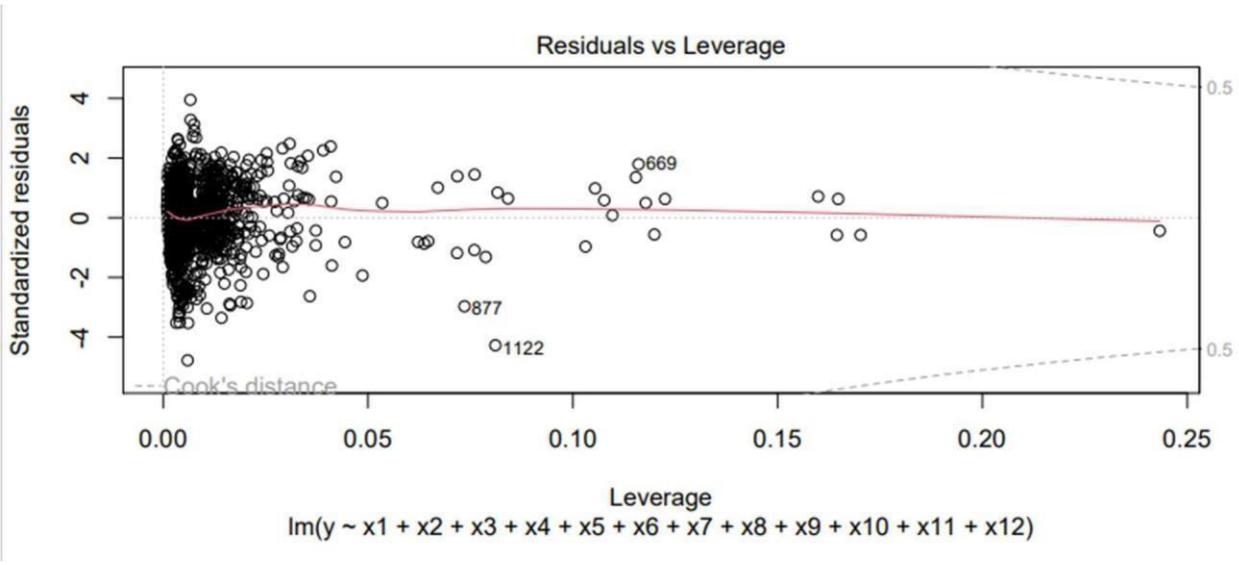
Scale-Location



GRAPH 3

This plot shows a "fanning out" pattern of the residuals, which is typical of a model where the residual (error) variances are not all equal (heteroskedasticity).

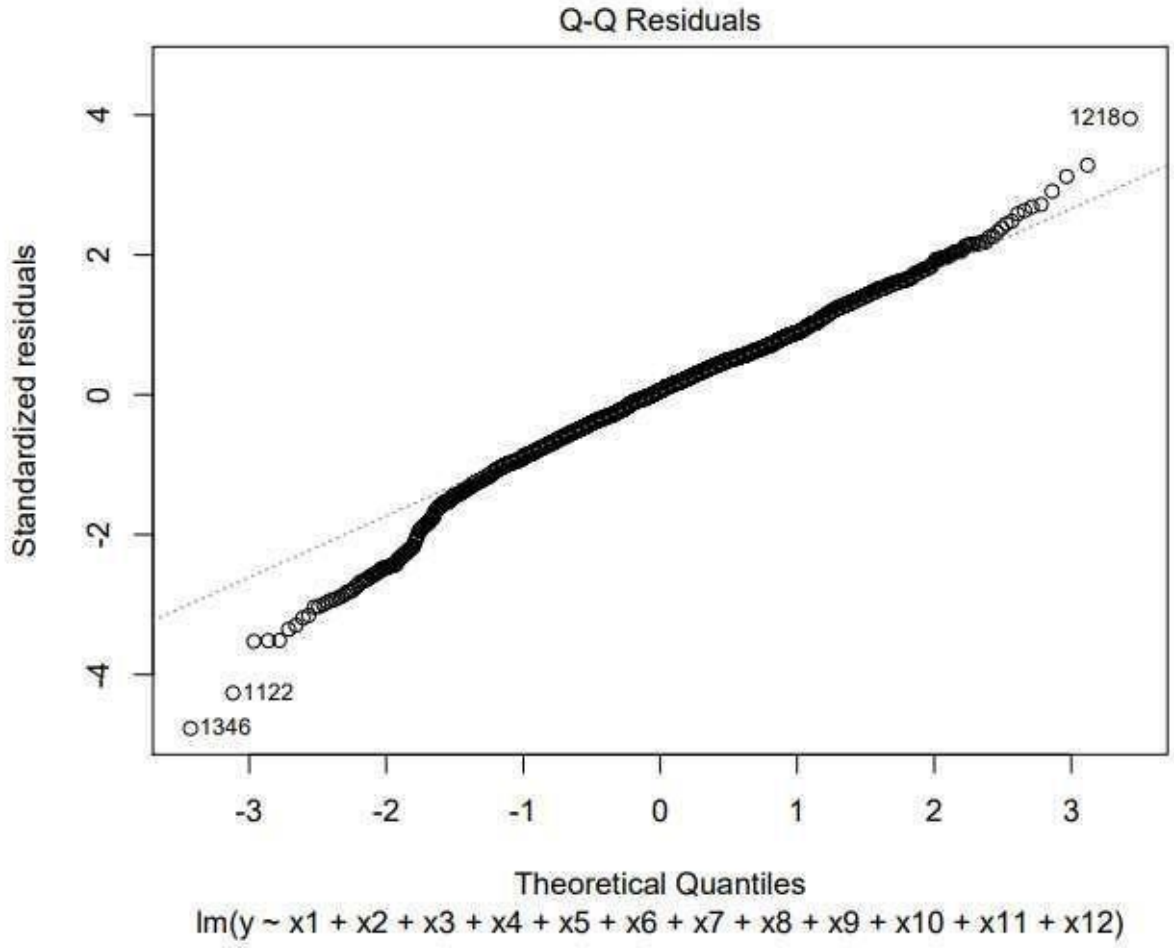
Residuals vs Leverage



GRAPH 4

The plot residual versus leverage indicates that most data points have little impact on the regression line. While there appears to be one point with potentially greater influence, without surpassing the Cook's distance threshold.

The Q-Q residual of the initial model



GRAPH 5

The scatterplot of the residuals shows a generally linear distribution with a few outliers, indicating a departure from perfect normality but not to an extreme extent. The observation that more data points lie above the line than below suggests a subtle positive skew in the residual distribution. Additionally, the Q-Q plot suggests that while the standardized residuals approximate a normal distribution, there are minor deviations from

perfect normality. Overall, while there are indications of non-normality in the residuals, they are not pronounced, implying that the model's assumptions are reasonably met.

```
> f1 = lm(y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12)
> p = 12
> n = length(y)
> f1

Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 +
    x10 + x11 + x12)

Coefficients:
(Intercept)           x1           x2           x3           x4           x5           x6
   x7           x8           x9          x10          x11
  6.379e+01   -2.484e-02   -6.620e-03    2.975e-01   -1.215e-04   -7.287e-03    1.714e-05    9.73
2e-02    2.293e-02    4.225e-02   -4.531e-01    1.463e-04
         x12
  5.963e-09

> p
[1] 12
> n
[1] 1649
> leverage <- hatvalues(f1)
> threshold <- (2*p)/n
> threshold
[1] 0.01455428
> print(which(leverage > threshold))
  12   49   74   80   81   82   83   84   86  121  133  141  145  148  228  229  230  231  232  2
68  310  314  316  317  319  321  322  324  325  341  349  354  355
  12   49   74   80   81   82   83   84   86  121  133  141  145  148  228  229  230  231  232  2
68  310  314  316  317  319  321  322  324  325  341  349  354  355
 356  357  358  359  360  361  362  363  377  393  397  401  417  421  422  437  492  503  511  5
26  558  573  574  576  578  581  583  592  594  605  608  618  645
 356  357  358  359  360  361  362  363  377  393  397  401  417  421  422  437  492  503  511  5
26  558  573  574  576  578  581  583  592  594  605  608  618  645
 665  666  667  668  669  670  671  672  673  674  675  676  704  705  706  712  713  719  803  8
29  837  838  844  845  846  847  848  866  871  872  874  877  879
 665  666  667  668  669  670  671  672  673  674  675  676  704  705  706  712  713  719  803  8
29  837  838  844  845  846  847  848  866  871  872  874  877  879
 881  882  897  903  908  911 1022 1023 1054 1062 1063 1064 1088 1089 1090 1113 1114 1115 1116 11
17 1118 1119 1120 1121 1122 1128 1129 1132 1136 1188 1197 1231 1351
 881  882  897  903  908  911 1022 1023 1054 1062 1063 1064 1088 1089 1090 1113 1114 1115 1116 11
17 1118 1119 1120 1121 1122 1128 1129 1132 1136 1188 1197 1231 1351
1354 1356 1357 1371 1372 1373 1374 1419 1420 1421 1422 1423 1424 1425 1426 1427 1428 1431 1432 14
33 1434 1435 1482 1483 1492 1493 1494 1501 1502 1504 1534 1535 1558
1354 1356 1357 1371 1372 1373 1374 1419 1420 1421 1422 1423 1424 1425 1426 1427 1428 1431 1432 14
33 1434 1435 1482 1483 1492 1493 1494 1501 1502 1504 1534 1535 1558
1561 1566 1573 1575 1580 1581 1596 1629 1631 1632 1639 1642 1643 1644 1645 1646 1647 1648 1649
1561 1566 1573 1575 1580 1581 1596 1629 1631 1632 1639 1642 1643 1644 1645 1646 1647 1648 1649
>
```

FIGURE 2

The observation number 12, 49, 74, ... ,1649 were all identified as outliers, because their hat values are greater than 2p/n.

Non-constant variance

1. Linear relationship between the dependent and independent variables

Ho: Variance is constant

H1: Variance is not constant.

Breusch-Pagan test

```
> library(lmtest)
> bp_test <- bptest(f1)
> print(bp_test)

        studentized Breusch-Pagan test

data:  f1
BP = 266.32, df = 12, p-value < 2.2e-16
```

FIGURE 3

We conducted a test using a Breusch-Pagan test, the Null hypothesis is rejected, the pvalue is less than 0.05 suggesting that the Variance is not constant, and we went on to do another test, White test.

White test

```
> white_test <- bptest(f1, studentize =FALSE)
> print(white_test)

        Breusch-Pagan test

data:  f1
BP = 425.96, df = 12, p-value < 2.2e-16

>
```

FIGURE 4

The Null hypothesis for the White test is rejected since the p-value is less than the 0.05 suggesting that the variance is not constant.

Normality

Normally distributed error component.

Ho: Residuals are Normally distributed

H1: Residuals are not Normally distributed.

Anderson-Darling test

```
> library(nortest)
> resids = fl$residuals
> ad.test(resids)

        Anderson-Darling normality test

data:  resids
A = 5.8909, p-value = 1.698e-14
```

FIGURE 5

Shapiro-Francia test

```
> sf.test(resids)

        Shapiro-Francia normality test

data:  resids
W = 0.98254, p-value = 2.836e-12

>
```

FIGURE 6

Both tests reject the null hypothesis of normally distributed residuals. Since in both test the p-value is less than 0.05, it is then suggested that the residuals are not normally distributed.

Correlation

```
> observed_values <- model1$fitted.values
> correlation <- cor(observed_values, model1$residuals)
> correlation
[1] 0.00000000000000001644619
>
```

FIGURE 7

The correlation of the initial model is close to zero suggesting that the model predictions are not systematically biased in one direction or the other.

Correlated error (auto correction)

Ho: Residuals are uncorrelated

H1: Residuals are not uncorrelated.

Durbin-Watson test

```
> library(car)
> durbinWatsonTest(f1)
 lag Autocorrelation D-W Statistic p-value
   1       0.5612034     0.8735383       0
 Alternative hypothesis: rho != 0
>
```

FIGURE 8

The p-value is extremely low ($p < 0.05$), indicating strong evidence against the null hypothesis. Therefore, we reject the null hypothesis in favour of the alternative, suggesting that there is autocorrelation present in the residuals of your regression model for life expectancy.

The Normal Q-Q plot

Our assumptions about the model suggested that there is, non-linear relationship, heteroscedasticity, the residuals are not normally distributed which led us to go on and try to Transform the variables as well as the model. We started by using the Boxcox transformation and found out that there are little to no changes on our model.

The following are the results after we performed the BOXCOX.

```
> lambda <- bc$x[which.max(bc$y)]
> lambda
[1] 1.59596
> y2=(y^lambda-1)/lambda
> model2=lm (y2~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12)
```

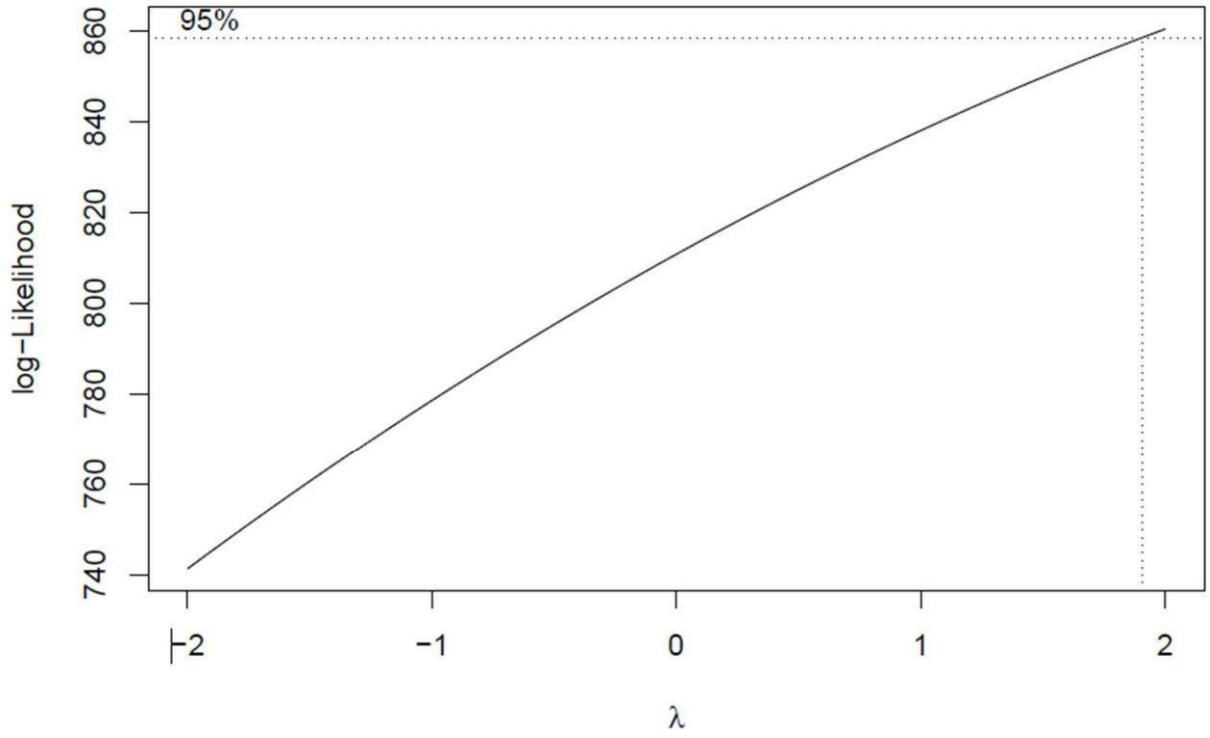
FIGURE 9

This is the summary after we performed the Box-cox transformation to our model, and we now have a new model.

```
> summary(model2)

Call:
lm(formula = y2 ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 +
    x10 + x11 + x12)

Residuals:
     Min       1Q   Median       3Q      Max
-237.981  -32.542    2.841   33.625  237.181

Coefficients:
                    Estimate      Std. Error t value             Pr(>|t|)
(Intercept) 477.52458269022   7.66869970382  62.269 < 0.0000000000000002
x1           -0.30499611294   0.01397067602 -21.831 < 0.0000000000000002
x2           -0.08058145578   0.01775147795  -4.539          0.0000060529
x3            3.85902468412   0.40160531943   9.609 < 0.0000000000000002
x4           -0.00097931225   0.00276036199  -0.355              0.722803
x5           -0.09580425086   0.06828062657  -1.403              0.160778
x6            0.00021522177   0.00016056077   1.340              0.180289
x7            1.19341667038   0.08033250077  14.856 < 0.0000000000000002
x8            0.27746305512   0.07884942338   3.519              0.000445
x9            0.51156068639   0.08998803359   5.685          0.0000000155
x10          -5.01447648302   0.27278199588 -18.383 < 0.0000000000000002
x11           0.00186891452   0.00043062017   4.340          0.0000151169
x12           0.00000006950   0.00000002644   2.629              0.008648

(Intercept) ***
x1          ***
x2          ***
x3          ***
x4
x5
x6
x7          ***
x8          ***
x9          ***
x10         ***
x11         ***
x12         **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 55.37 on 1636 degrees of freedom
Multiple R-squared:  0.7355,    Adjusted R-squared:  0.7336
F-statistic: 379.2 on 12 and 1636 DF,  p-value: < 0.00000000000000022
```

FIGURE 10

The Log-likelihood of the 2nd model after we did Box-cox transformation.



GRAPH 6

LOG TRANSFORMATION

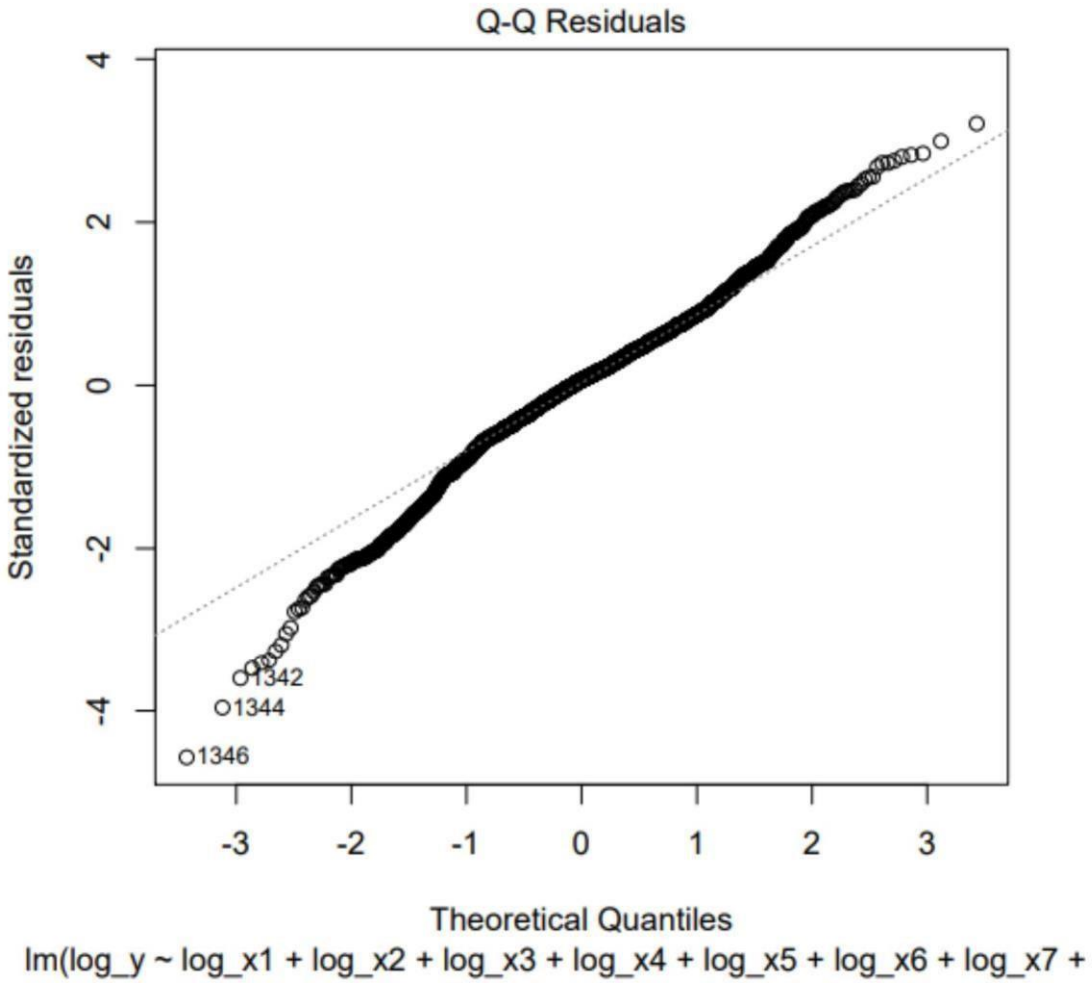
```
> log_y <- log(mydata$y)
> log_x1 <- log(mydata$x1)
> log_x2 <- log(mydata$x2)
> log_x3 <- log(mydata$x3)
> log_x4 <- log(mydata$x4)
> log_x5 <- log(mydata$x5)
> log_x6 <- log(mydata$x6)
> log_x7 <- log(mydata$x7)
> log_x8 <- log(mydata$x8)
> log_x9 <- log(mydata$x9)
> log_x10 <- log(mydata$x10)
> log_x11 <- log(mydata$x11)
> log_x12 <- log(mydata$x12)
>
> f2 = lm(log_y ~ log_x1 + log_x2 + log_x3 + log_x4 + log_x5 + log_x6 + log_x7 + log_x8 + log_
x9 + log_x10 + log_x11 + log_x12)
> f2

Call:
lm(formula = log_y ~ log_x1 + log_x2 + log_x3 + log_x4 + log_x5 +
    log_x6 + log_x7 + log_x8 + log_x9 + log_x10 + log_x11 + log_x12)

Coefficients:
(Intercept)       log_x1       log_x2       log_x3       log_x4       log_x5       log_x6
  log_x7       log_x8       log_x9      log_x10      log_x11
  4.049e+00   -1.195e-02   -1.143e-03    7.477e-03    1.309e-02   -4.034e-03   -1.346e-03    1
.115e-02    3.332e-03    1.451e-02   -5.407e-02   -9.912e-05
    log_x12
  6.659e-04

>
```

```
> summary(f2)

Call:
lm(formula = log_y ~ log_x1 + log_x2 + log_x3 + log_x4 + log_x5 +
    log_x6 + log_x7 + log_x8 + log_x9 + log_x10 + log_x11 + log_x12)

Residuals:
      Min        1Q    Median        3Q       Max
-0.287705 -0.033737  0.004123  0.037379  0.202008

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.049e+00  2.109e-02 191.970  < 2e-16 ***
log_x1      -1.195e-02  1.574e-03  -7.592 5.25e-14 ***
log_x2      -1.143e-03  5.281e-04  -2.164   0.0306 *
log_x3       7.477e-03  7.821e-04   9.560  < 2e-16 ***
log_x4       1.309e-02  1.763e-03   7.423 1.83e-13 ***
log_x5      -4.034e-03  2.632e-03  -1.532   0.1256
log_x6      -1.346e-03  3.003e-04  -4.482 7.93e-06 ***
log_x7       1.115e-02  2.133e-03   5.229 1.93e-07 ***
log_x8       3.332e-03  2.981e-03   1.118   0.2638
log_x9       1.451e-02  3.345e-03   4.337 1.53e-05 ***
log_x10     -5.407e-02  1.149e-03 -47.042  < 2e-16 ***
log_x11     -9.912e-05  2.107e-03  -0.047   0.9625
log_x12      6.659e-04  6.544e-04   1.018   0.3090
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06334 on 1636 degrees of freedom
Multiple R-squared:  0.7813,    Adjusted R-squared:  0.7797
F-statistic:   487 on 12 and 1636 DF,  p-value: < 2.2e-16
```

FIGURE 11

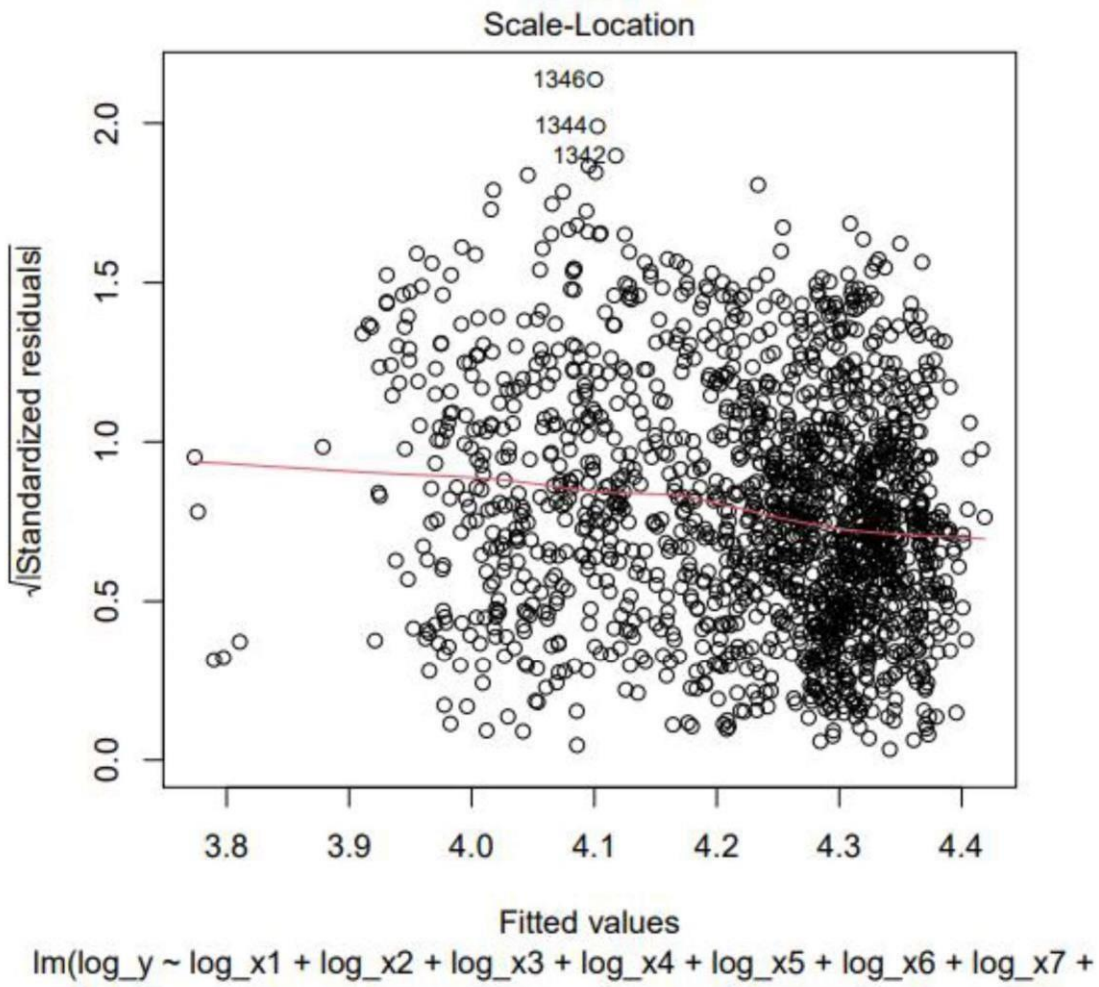
The residuals exhibit a scatter around zero, indicating a generally good fit between the model and the data. Although there are a few outliers suggesting instances where the model's fit is less accurate, their limited number suggests they have a minor impact on overall results. Notably, most residuals are positive, indicating a tendency for the model to underestimate actual values. This slight positive skew in the residuals suggests a consistent pattern of underestimation by the model. Despite these observations, most data points conform well to the model, implying its effectiveness in capturing the underlying relationships within the data.

Q-Q Residual of Log-Transformation



GRAPH 7

Scale-Location of Log-Transformation



GRAPH 8

After we fitted the 3rd model using log Transformation, we then proceeded to test our 3rd model to see if is adequate to be used or not.

We started by using Breusch-Pagan test and White test to test for heteroscedasticity in our fitted model.

The Breusch-Pagan test and White test are both used to detect heteroscedasticity in regression models.

• Null Hypothesis (Ho): There is homoscedasticity in the life expectancy regression model.

• Alternative Hypothesis (H1): There is heteroscedasticity in the life expectancy regression model.

```
> library(lmtest)
> bp_test <- bptest(f2)
> print(bp_test)

        studentized Breusch-Pagan test

data:  f2
BP = 76.294, df = 12, p-value = 2.091e-11

> white_test <- bptest(f2, studentize = FALSE)
> print(white_test)

        Breusch-Pagan test

data:  f2
BP = 110.21, df = 12, p-value < 2.2e-16

>
```

FIGURE 12

The Breusch-Pagan test for 3rd model indicates significant evidence against the null hypothesis of homoscedasticity, suggesting that the residuals in 3rd model exhibit heteroscedasticity. This is supported by a low p-value of 2.091e-11. Similarly, the White test for 3rd model also provides significant evidence against the null hypothesis of homoscedasticity, with an even lower p-value of < 2.2e-16. In summary, both tests confirm that the 3rd model suffers from heteroscedasticity, indicating that the variability of life expectancy is not constant across different values of the independent variables in the model.

The following graphs were plotted after we did a transformation in our variables against fitted.

The Anderson-Darling and Shapiro-Francia tests were conducted on the residuals of the life expectancy regression model to assess their normality.

• Null Hypothesis (Ho): The residuals follow a normal distribution.

• Alternative Hypothesis (H1): The residuals do not follow a normal distribution. Anderson-Darling and Shapiro-Francia test

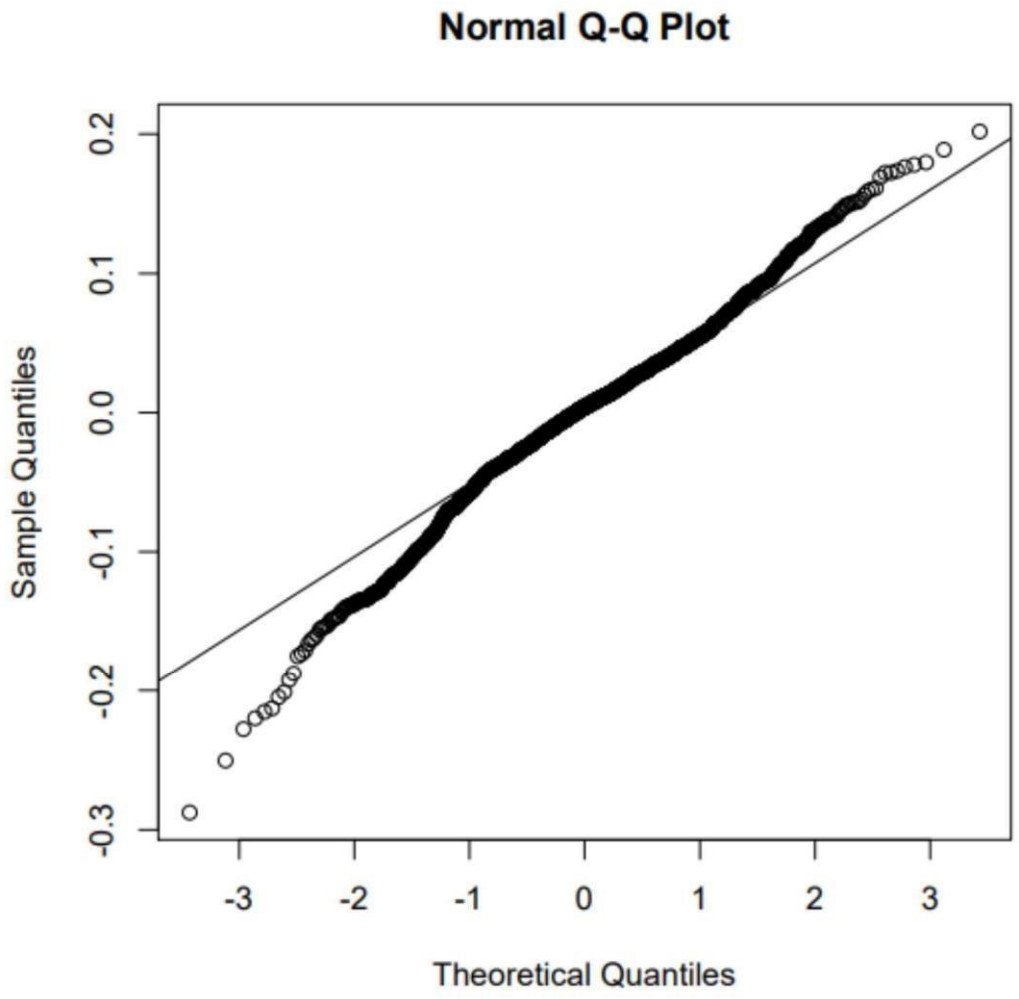
```
> library(nortest)
> ad.test(residuals(f2))

        Anderson-Darling normality test

data:  residuals(f2)
A = 7.1767, p-value < 2.2e-16

> sf.test(residuals(f2))

        Shapiro-Francia normality test

data:  residuals(f2)
W = 0.98669, p-value = 2.107e-10

>
```

FIGURE 13

The normality of the residuals for 3rd model was assessed using the Anderson-Darling and Shapiro-Francia tests. Both tests indicated significant evidence against the null hypothesis of normality. The Anderson-Darling test yielded a p-value that is less than 2.2e-16, while the Shapiro Francia test resulted in a p-value of 2.107e-10. These results suggest that the residuals of the 3rd model deviate significantly from a normal distribution.

The following is the graph after we did log Transformation and obtained 3rd model.

Normal Q-Q Plot of Log-Transformation



GRAPH 9

Since the plotted points are not in a straight line, the residuals appear to be not normally distributed.

We then again did an Autocorrelation test.

The Durbin-Watson test assesses the presence of autocorrelation in the residuals of a regression model.

• Null Hypothesis (Ho): There is no autocorrelation present in the residuals (rho = 0).

• Alternative Hypothesis (H1): There is autocorrelation present in the residuals (rho ≠ 0).

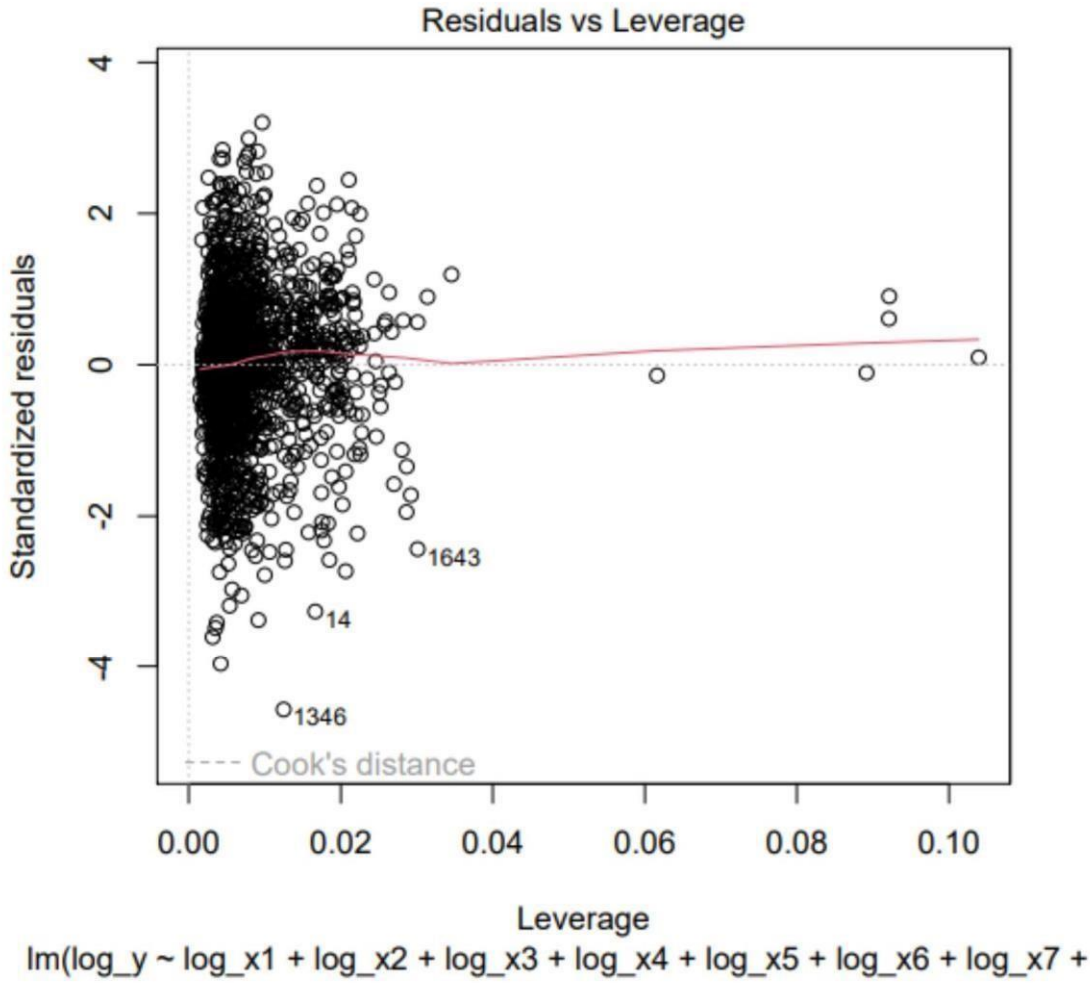
Durbin-Watson test

```
> library(car)
> durbinWatsonTest(f2)
 lag Autocorrelation D-W Statistic p-value
   1       0.6915865     0.6162016       0
 Alternative hypothesis: rho != 0
>
```

FIGURE 14

The Durbin-Watson test was conducted on the 3rd model to assess autocorrelation in the residuals. The test yielded a Durbin-Watson statistic of 0.6162016 with a corresponding p-value of 0, indicating significant evidence against the null hypothesis of no autocorrelation. This result suggests that there is autocorrelation present in the residuals of 3rd model.

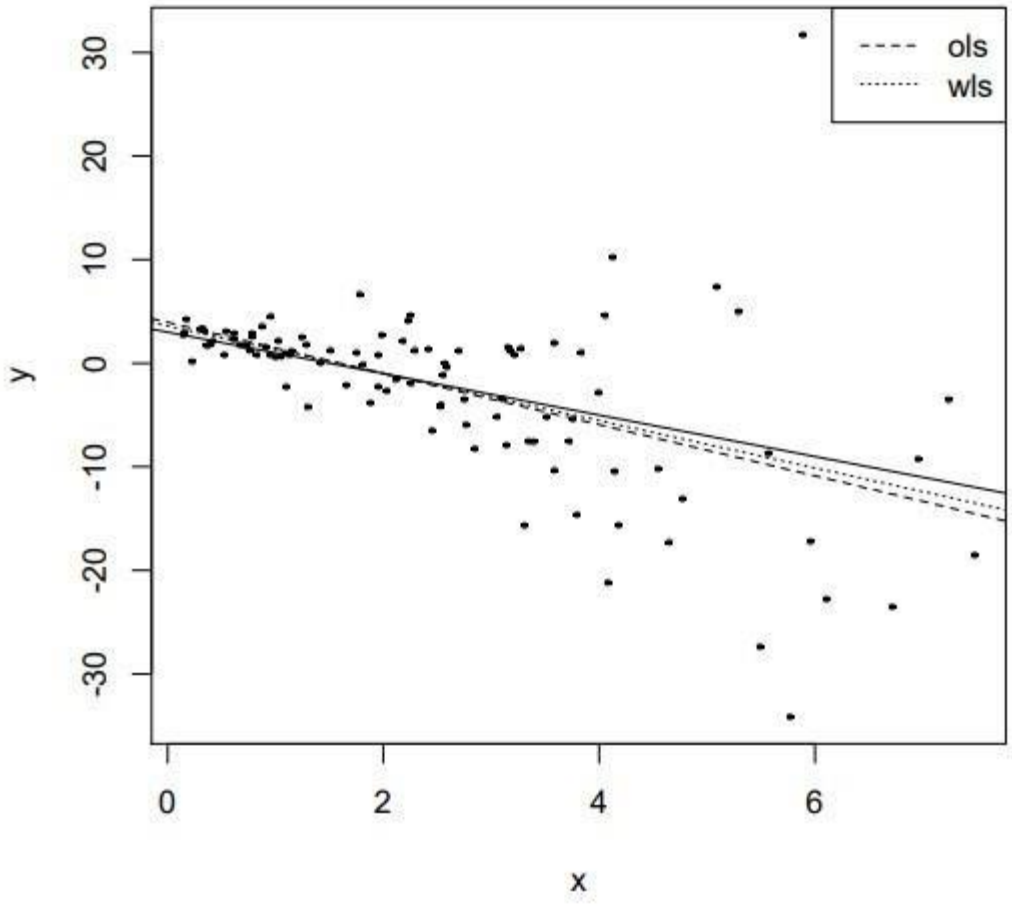
Residuals vs Leverage of Log-Transformation



GRAPH 10

Weighted least squares (WLS)

```
> f1

Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 +
    x10 + x11 + x12)

Coefficients:
(Intercept)           x1           x2           x3           x4           x5           x6
   x7           x8           x9          x10          x11
  6.379e+01   -2.484e-02   -6.620e-03    2.975e-01   -1.215e-04   -7.287e-03    1.714e-05    9.73
2e-02    2.293e-02    4.225e-02   -4.531e-01    1.463e-04
        x12
  5.963e-09

> x=abs(rnorm(100,0,3))
> y=3-2*x+rnorm(100,0,sapply(x,function(x){1+0.5*x^2}))
> plot(x,y,pch=16,cex=0.5)
> abline(a=3,b=-2,col="black")
> fit.ols=lm(y~x)
> abline(fit.ols$coefficients,lty=2)
> fit.wls=lm(y~x,weights=1/(1+0.5*x^2))
> abline(fit.wls$coefficients,lty=3)
> legend("topright",c("ols","wls"),lty=c(2,3))
>
```



The coefficients of the two lines are shown below.

```
> fit.ols$coefficients
(Intercept)           x
   3.933835   -2.467485
>
```

```
> fit.wls$coefficients
(Intercept)           x
   3.596154   -2.285203
>
```

The comparison of coefficients between the OLS and WLS models reveals notable differences, particularly with larger intercept and slope values in the WLS model. This suggests a distinct approach in fitting the data, potentially addressing heteroscedasticity or unequal variances. While OLS assumes equal variance across all observations, WLS adjusts by assigning weights based on variance, giving more weight to observations with lower variance. Therefore, the WLS model places greater emphasis on certain data points, aiming to provide a more accurate fit by mitigating the effects of heteroscedasticity.

Variable selection

Backward Selection

```
> fl = lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x11 + x12)
> summary(fl)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 +
    x10 + x11 + x12)

Residuals:
     Min       1Q   Median       3Q      Max
-21.2824  -2.5064   0.2694   2.7378  17.6119

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.379e+01  6.194e-01 102.982  < 2e-16 ***
x1          -2.484e-02  1.128e-03 -22.012  < 2e-16 ***
x2          -6.620e-03  1.434e-03  -4.617 4.19e-06 ***
x3           2.975e-01  3.244e-02   9.171  < 2e-16 ***
x4          -1.215e-04  2.230e-04  -0.545 0.585938
x5          -7.287e-03  5.515e-03  -1.321 0.186555
x6           1.714e-05  1.297e-05   1.321 0.186553
x7           9.732e-02  6.488e-03  14.999  < 2e-16 ***
x8           2.293e-02  6.369e-03   3.601 0.000326 ***
x9           4.225e-02  7.268e-03   5.813 7.37e-09 ***
x10         -4.531e-01  2.203e-02 -20.565  < 2e-16 ***
x11          1.463e-04  3.478e-05   4.206 2.74e-05 ***
x12          5.963e-09  2.135e-09   2.792 0.005296 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.472 on 1636 degrees of freedom
Multiple R-squared:  0.7434,    Adjusted R-squared:  0.7415
F-statistic:   395 on 12 and 1636 DF,  p-value: < 2.2e-16
```

We fit model with all the parameters, we show estimation details then we remove x4, x5 and x6 because the p-values of these predictor is greater than 0.10.

```
> f2 = lm(y ~ x1 + x2 + x3 + x7 + x8 + x9 + x10 + x11 + x12)
> summary(f2)

Call:
lm(formula = y ~ x1 + x2 + x3 + x7 + x8 + x9 + x10 + x11 + x12)

Residuals:
     Min       1Q   Median       3Q      Max
-21.3473  -2.5123   0.3134   2.7690  17.4916

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.368e+01  6.133e-01 103.829  < 2e-16 ***
x1          -2.495e-02  1.127e-03 -22.139  < 2e-16 ***
x2          -5.616e-03  1.288e-03  -4.362 1.37e-05 ***
x3           3.003e-01  3.241e-02   9.267  < 2e-16 ***
x7           9.672e-02  6.477e-03  14.933  < 2e-16 ***
x8           2.185e-02  6.279e-03   3.479 0.000516 ***
x9           3.829e-02  6.539e-03   5.855 5.76e-09 ***
x10         -4.518e-01  2.201e-02 -20.525  < 2e-16 ***
x11          1.289e-04  1.098e-05  11.741  < 2e-16 ***
x12          5.859e-09  2.131e-09   2.749 0.006046 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.473 on 1639 degrees of freedom
Multiple R-squared:  0.7428,    Adjusted R-squared:  0.7414
F-statistic:   526 on 9 and 1639 DF,  p-value: < 2.2e-16

>
```

After we remove x4, x5 and x6, we fit the model again. We stop the elimination because the pvalues of a remaining predictor is less than 0.1.

Mallows C_p statistic

```
> library(leaps)
> x = model.matrix(f1)[,-1]
> y = y
> f1 = leaps(x,y)
> f1
$which
       1     2     3     4     5     6     7     8     9     A     B     C
1   TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
1  FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
1  FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
1  FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
1  FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
1  FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
1  FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
1  FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
1  FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
1  FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
2   TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
2   TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
2   TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
2   TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
2   TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
2   TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
2  FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE
2   TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
2   TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
2   TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
3   TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE
3   TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
3   TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE
3   TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE
3   TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
3   TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
3   TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE
3   TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
3   TRUE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE
3  FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE
4   TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE
4   TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE
4   TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE
4   TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE  TRUE FALSE FALSE
4   TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE
4   TRUE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE FALSE FALSE
4   TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
4   TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE
4   TRUE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE
4   TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE FALSE
5   TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE
5   TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE
5   TRUE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE  TRUE FALSE
5   TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE  TRUE  TRUE FALSE FALSE
5   TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE
5   TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE  TRUE  TRUE FALSE FALSE
5   TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE  TRUE FALSE FALSE
5   TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE FALSE FALSE
5   TRUE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE
5   TRUE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE
6   TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE
6   TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE FALSE  TRUE  TRUE FALSE FALSE
6   TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE  TRUE FALSE
6   TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE  TRUE FALSE FALSE
6   TRUE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE
6   TRUE FALSE  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE
6   TRUE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE
6   TRUE  TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE
6   TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE
6   TRUE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE
7   TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE
7   TRUE  TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE
7   TRUE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE
```

7	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE
7	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE
7	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE
7	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE
7	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE
7	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE
7	TRUE	FALSE	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE
8	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE
8	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE
8	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE
8	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE
8	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE
8	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE
8	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
8	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE
8	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE
8	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE
9	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
9	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE
9	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE
9	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE
9	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE
9	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE
9	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE
9	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE
9	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE
9	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE
10	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
10	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
10	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
10	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE
10	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE
10	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE
10	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE
10	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE
10	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE
10	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE
11	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
11	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
11	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
11	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE
11	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE
11	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE
11	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
11	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE
11	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
11	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE
12	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

```
$label
 [1] "(Intercept)" "1"           "2"           "3"           "4"           "5"           "6"
     "7"           "8"           "9"           "A"
[12] "B"           "C"

$size
  [1]  2  2  2  2  2  2  2  2  2  2  3  3  3  3  3  3  3  3  3  3  4  4  4  4  4  4  4  4  4
5  5  5  5  5  5  5  5  5  5  6  6  6  6  6  6  6  6  6  6  7  7  7  7
 [55]  7  7  7  7  7  7  8  8  8  8  8  8  8  8  8  8  9  9  9  9  9  9  9  9  9  9 10 10 10 10 1
0 10 10 10 10 10 11 11 11 11 11 11 11 11 11 11 12 12 12 12 12 12 12 12
[109] 12 12 13

$Cp
  [1] 1584.34944 2494.85208 2857.88508 3489.42023 3661.37328 3697.17853 3988.41626 4048.26029 447
6.41296 4549.02730  952.88533 1072.75952 1117.65474 1188.27618
 [15] 1200.38790 1303.15449 1335.68879 1354.36584 1462.48346 1483.96643  590.33501  606.39542  66
3.48121  698.15137  705.58730  714.84278  760.52806  768.13910
 [29]  807.37484  831.84733  269.99442  288.87887  312.11035  404.45966  426.13327  442.25730  45
7.61055  476.69149  502.03337  503.48730  126.86590  141.26357
 [43]  157.72588  158.31521  164.43708  179.48448  188.77387  206.16465  228.68015  246.68753   3
8.02057   56.68904   63.19154   81.39104  108.61458  110.00394


 [57]  116.68241  119.02095  128.30722  128.54673   25.94430   26.95482   39.39877   39.64424   3
9.89629   39.98191   43.23057   45.36696   50.35396   58.04173
 [71]   16.39888   20.95122   27.23117   27.39132   27.43192   27.75488   27.87420   28.29474   2
8.56067   33.53107   10.83889   16.42411   16.94840   18.18874
 [85]   20.95027   22.38773   22.51654   28.02287   28.57663   28.77987   10.97629   11.05817   1
2.59678   17.06276   18.14439   18.74879   22.44006   22.53339
 [99]   23.89373   28.15810   11.29685   12.74608   12.74610   18.79646   23.96764   28.69228   3
2.31816   44.78934   95.09930  235.96706   13.00000

>
```

```
> f1 = lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x11 + x12)
> library(olsrr)
> #  Backward variable selection using AIC
> ols_step_backward_aic(f1)
```

Stepwise Summary

Step	Variable	AIC	SBC	SBIC	R2	Adj. R2
0	Full Model	9634.613	9710.324	4955.160	0.74342	0.74154
1	x4	9632.912	9703.215	4953.429	0.74338	0.74165
2	x5	9632.604	9697.499	4953.092	0.74311	0.74155
3	x6	9632.478	9691.965	4952.941	0.74282	0.74141

x4, x5, x6 eliminated.

```
> #  Forward variable selection using AIC
> ols_step_forward_aic(fl)
```

Stepwise Summary

Step	Variable	AIC	SBC	SBIC	R2	Adj. R2
0	Base Model	11853.803	11864.619	7171.562	0.00000	0.00000
1	x1	10733.977	10750.201	6051.881	0.49354	0.49323
2	x7	10375.913	10397.544	5693.799	0.59289	0.59239
3	x10	10128.373	10155.413	5446.490	0.65006	0.64942
4	x11	9873.091	9905.538	5192.032	0.70061	0.69988
5	x9	9744.706	9782.562	5064.257	0.72337	0.72253
6	x3	9659.512	9702.775	4979.651	0.73762	0.73666
7	x8	9647.579	9696.250	4967.824	0.73983	0.73872
8	x2	9638.063	9692.142	4958.421	0.74164	0.74038
9	x12	9632.478	9691.965	4952.930	0.74282	0.74141

x1, x7, x10, x11, x9, x3, x8, x2, x12 selected.

```
> #  Forward variable selection using AIC
> ols_step_forward_aic(fl)
```

Stepwise Summary

Step	Variable	AIC	SBC	SBIC	R2	Adj. R2
0	Base Model	11853.803	11864.619	7171.562	0.00000	0.00000
1	x1	10733.977	10750.201	6051.881	0.49354	0.49323
2	x7	10375.913	10397.544	5693.799	0.59289	0.59239
3	x10	10128.373	10155.413	5446.490	0.65006	0.64942
4	x11	9873.091	9905.538	5192.032	0.70061	0.69988
5	x9	9744.706	9782.562	5064.257	0.72337	0.72253
6	x3	9659.512	9702.775	4979.651	0.73762	0.73666
7	x8	9647.579	9696.250	4967.824	0.73983	0.73872
8	x2	9638.063	9692.142	4958.421	0.74164	0.74038
9	x12	9632.478	9691.965	4952.930	0.74282	0.74141

DISCUSSIONS

Adult mortality has an inverse relationship with life expectancy, indicating that higher adult mortality rates are associated with shorter average lifespans. Additionally, when infant mortality rates are high, life expectancy is reduced, reflecting a negative relationship.

There is a positive relationship between polio immunization coverage and life expectancy, as people who are immunized tend to live longer compared to those who contract polio without immunization.

People with HIV/AIDS have shorter lifespans, as indicated by our data, which focuses on the prevalence of the virus rather than immunization coverage. This results in a negative relationship with life expectancy, as those infected by the virus tend to die at a younger age. For instance, in South Africa from 2003 to 2008, the high mortality rates among both children and adults due to HIV/AIDS significantly reduced life expectancy.

A country's higher Gross Domestic Product (GDP) indicates that its citizens are likely to enjoy longer life spans, as increased economic production enables the country to invest in healthier initiatives and choices.

Multilinear model assumptions Findings

Our model does not meet the following.

Despite employing Box-cox and Log transformations, the residuals still do not follow a normal distribution. Furthermore, they are not evenly distributed around the line of best fit, as depicted in the figure. This deviation from normality can lead to underestimated standard errors for regression coefficients, inaccurate t-statistics, p-values, and ultimately, incorrect projected values. Given these violations of multi-linear regression assumptions, it is highly advised that we interpret the conclusions of our study with caution. The Heteroscedasticity remains present in our data.

CONCLUSIONS

The multilinear regression model was used to predict the Life expectancy of the country using social, environmental, and health-related factors influencing it. The factors/predictors that were used are as follows, adult mortality, Infant deaths, Alcohol, percentage expenditure, Polio, Hepatitis B, BMI, Measles, Population, Diphtheria, HIV/AIDS, and GDP to predict how they influence the life expectancy. Our predictors are under the significant level of 0.05.

The final model is.

Life expectancy = 6.368e+01 -2.495e-02x15.616e-03x2 +3.003e-01x3 +9.672e-02x7 +2.185e-02x8 +3.829e-02x9 -4.518e-01x10 +1.289e-04x11+ 5.859e-09x12

Multiple R-squared is 74.3%, Adjusted R-squared is 74.2% after we refined the model,

This means that our model is significant since our Multiple R-square is 74.3%, we cannot really trust the model to do accurate predictions, for instance,

let's predict the life expectancy in Afghanistan in 2005, the country is developing.

Adult mortality = 280

Infant deaths = 80

Alcohol = 4.7

Percentage expenditure = 79.67936736

Hepatitis = 66

Measles = 1990

BMI =55.8

Polio = 0

Diphtheria = 0

HIV/AIDS = 0.1

GDP = 189.681557

Population = 2.966463e6

Life expectancy =6.368e+01 -2.495e-02(280)-5.616e-03(80) +3.003e-01(4.7) +9.672e-02(55.8)+2.185e-02(0)+3.829e-02(0)-4.518e-01(0.1)+1.289e-04(189.681557)+5.859e-09(2.966463e6)

Life expectancy = 63 years

HIV/AIDS: HIV/AIDS is a life-threatening chronic illness. Countries with higher HIV/AIDS prevalence have shorter life expectancy. A clear result is the inverse link between adult mortality and life expectancy. A greater adult mortality rate corresponds to a shorter life expectancy. This is because a greater adult mortality rate suggests that the country has more health issues.

The more a country have immunization coverage against polio and diphtheria, the longer people live.

Despite refinement and transformation, the final model does not meet our assumptions. Hence, the residuals are not normally distributed, and the model exhibits heteroscedasticity, and there is no multicollinearity between independent variables. To ensure the accuracy of the p-value and t-statistic, the residuals in the final model are to follow a normal distribution around the fitted line. However, in our analysis, the residuals not only deviate from the fitted line but also show significant dispersion, indicating a lack of normality. This discrepancy may result in incorrect conclusions and needs careful consideration. These are only a few of the significant social, economic, and health-related factors that may be utilized to create a multiple linear regression model to forecast a country's life expectancy. The precise predictors that are most essential may differ based on the nation and the available data. This can result in underestimated standard errors for regression coefficients, incorrect t-statistics and p-values, and, ultimately, incorrect projected values.

GLOSSARY

Life expectancy: refers to the number of years a person is expected to live based on statistical average.

Adult Mortality: refers to the death rate among individuals within a specific age range typically considered to be adults, often defined as individuals aged 15 to 64 or 15 to 69 years old, depending on the context

Infant Deaths: Infant mortality refers to the death of a babies that occurs between the time it is born and 1 year of age.

Quinquennial: refers to a five-year period or event.

Percentage expenditure: refers to the portion or share of total spending allocated to a particular category, activity, or item, expressed as a percentage of the overall budget or expenditure.

Hepatitis B: this is a viral infection that primarily affects the liver. It is caused by the hepatitis B virus (HBV), which is transmitted through contact with the blood or other body fluids of an infected person

Measles: Measles is a very contagious disease that causes fever, a red rash, cough and watery eyes and mostly common in children.

BMI: BMI is used to estimate your body fat and establish whether you are underweight, at your healthy weight, overweight, or have obesity.

Polio immunization coverage: refers to the proportion of individuals within a population who have received the polio vaccine, typically measured as a percentage

Diphtheria immunization coverage: refers to the percentage of individuals within a population who have received vaccination against diphtheria

HIV/AIDS: this is a viral infection that weakens the immune system, making individuals more susceptible to various infections and diseases.

Variable changed name description.

Variable name	Renamed variable name
Life expectancy	y
Adult mortality	X1
Infant death	X2
Alcohol	X3
Percentage expenditure	X4
Hepatitis B	X5
Measles	X6
BMI	X7
Polio	X8
Diphtheria	X9
HIV/AIDS	X10
GDP	X11
Population	X12

REFERENCES

Life expectancy dataset retrieved from the following GitHub website.
https://github.com/selva86/datasets/blob/master/Life_Expectancy_Data.csv

programming language for analysis
R and RStudio

Research was done with the help of the following websites and books.
https://en.wikipedia.org/wiki/Cook%27s_distance https://onlinestatbook.com/2/transformations/box-cox.html https://www.statology.org/how-to-calculate-mallows-cp-in-r/
https://stattrek.com/regression/influential-points#google_vignette
https://book.stat420.org/transformations.html#response-transformations Books

STA37W1 Linear Model notes
Applied Linear Model statistical analysis 5th edition by Richad A Johnson, Dean W. Wichern