Regression project

Ekta Shukla - A20567127 Usman Matheen Hameed - A20564338 Shruti Patil - A20564354

2024-11-09

–

# *Project: Life Expectancy Analysis*

# 1. Dataset

This project utilizes the Life Expectancy dataset. [Include the link to the dataset]

# 2. Initial Analysis (Part I): Model Fitting and Interpretation

Identify Response and Predictor Variables Response Variable: Life Expectancy Predictor Variables: A comprehensive set of predictors was initially considered, including both categorical and continuous variables: Categorical Variable: Status (Developed or Developing) Continuous Variables: Year, Adult Mortality, infant deaths, Alcohol, percentage expenditure, BMI, HIV/AIDS, Income composition of resources, Schooling, among others.

## Model Fitting

* The original model was fitted with all 22 predictor variables.
* Non-significant variables were removed to refine the model, focusing on the most impactful predictors.

# Load necessary library  
library(readr)  
  
# Load the dataset and omit rows with missing values  
data <- read\_csv("Life Expectancy Data.csv")

## Rows: 2938 Columns: 22  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (2): Country, Status  
## dbl (20): Year, Life expectancy, Adult Mortality, infant deaths, Alcohol, 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.

data <- na.omit(data)  
  
# Convert categorical variables to factors  
data$Status <- as.factor(data$Status)  
  
# Fit the multiple linear regression model with all variables  
model\_all <- lm(`Life expectancy` ~ Status + Year + `Adult Mortality` + `infant deaths` +   
 Alcohol + `percentage expenditure` + `Hepatitis B` + Measles + BMI +   
 `under-five deaths` + Polio + `Total expenditure` + Diphtheria +   
 `HIV/AIDS` + GDP + Population +   
 `thinness 5-9 years` + `Income composition of resources` + Schooling,   
 data = data)  
  
# Display the summary of the model  
summary(model\_all)

##   
## Call:  
## lm(formula = `Life expectancy` ~ Status + Year + `Adult Mortality` +   
## `infant deaths` + Alcohol + `percentage expenditure` + `Hepatitis B` +   
## Measles + BMI + `under-five deaths` + Polio + `Total expenditure` +   
## Diphtheria + `HIV/AIDS` + GDP + Population + `thinness 5-9 years` +   
## `Income composition of resources` + Schooling, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -16.7685 -2.1440 0.0303 2.1794 12.4765   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.091e+02 4.619e+01 6.693 3.01e-11 \*\*\*  
## StatusDeveloping -8.856e-01 3.351e-01 -2.643 0.0083 \*\*   
## Year -1.272e-01 2.307e-02 -5.515 4.05e-08 \*\*\*  
## `Adult Mortality` -1.621e-02 9.433e-04 -17.182 < 2e-16 \*\*\*  
## `infant deaths` 8.876e-02 1.059e-02 8.383 < 2e-16 \*\*\*  
## Alcohol -1.311e-01 3.357e-02 -3.904 9.86e-05 \*\*\*  
## `percentage expenditure` 3.026e-04 1.788e-04 1.692 0.0908 .   
## `Hepatitis B` -3.269e-03 4.446e-03 -0.735 0.4624   
## Measles -1.032e-05 1.070e-05 -0.965 0.3347   
## BMI 3.186e-02 5.945e-03 5.359 9.56e-08 \*\*\*  
## `under-five deaths` -6.664e-02 7.667e-03 -8.692 < 2e-16 \*\*\*  
## Polio 5.762e-03 5.109e-03 1.128 0.2596   
## `Total expenditure` 9.216e-02 4.041e-02 2.281 0.0227 \*   
## Diphtheria 1.405e-02 5.873e-03 2.392 0.0169 \*   
## `HIV/AIDS` -4.481e-01 1.779e-02 -25.187 < 2e-16 \*\*\*  
## GDP 2.450e-05 2.825e-05 0.867 0.3861   
## Population -6.141e-10 1.732e-09 -0.355 0.7230   
## `thinness 5-9 years` -5.503e-02 2.648e-02 -2.078 0.0379 \*   
## `Income composition of resources` 1.045e+01 8.318e-01 12.567 < 2e-16 \*\*\*  
## Schooling 8.953e-01 5.898e-02 15.180 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.555 on 1629 degrees of freedom  
## Multiple R-squared: 0.8386, Adjusted R-squared: 0.8367   
## F-statistic: 445.4 on 19 and 1629 DF, p-value: < 2.2e-16

# Get confidence intervals for the regression coefficients  
confint(model\_all)

## 2.5 % 97.5 %  
## (Intercept) 2.185203e+02 3.997030e+02  
## StatusDeveloping -1.542746e+00 -2.283776e-01  
## Year -1.724756e-01 -8.198028e-02  
## `Adult Mortality` -1.805749e-02 -1.435708e-02  
## `infant deaths` 6.798911e-02 1.095257e-01  
## Alcohol -1.969242e-01 -6.521520e-02  
## `percentage expenditure` -4.818582e-05 6.533696e-04  
## `Hepatitis B` -1.198951e-02 5.452346e-03  
## Measles -3.130099e-05 1.065714e-05  
## BMI 2.020124e-02 4.352437e-02  
## `under-five deaths` -8.168078e-02 -5.160501e-02  
## Polio -4.259545e-03 1.578280e-02  
## `Total expenditure` 1.290401e-02 1.714187e-01  
## Diphtheria 2.528565e-03 2.556621e-02  
## `HIV/AIDS` -4.829802e-01 -4.131904e-01  
## GDP -3.092422e-05 7.991453e-05  
## Population -4.011064e-09 2.782943e-09  
## `thinness 5-9 years` -1.069724e-01 -3.084472e-03  
## `Income composition of resources` 8.821864e+00 1.208481e+01  
## Schooling 7.795966e-01 1.010951e+00

### *After careful consideration, we removed variables with p-values greater than 0.05, as they were not statistically significant. This refinement ensures that our model only includes predictors that have a meaningful impact on life expectancy.*

# Model Specification:

# Load necessary library  
library(readr)  
  
# Fit the multiple linear regression model with significant variables.  
model\_significant <- lm(`Life expectancy` ~ Status + Year + `Adult Mortality` +   
 Alcohol + `percentage expenditure` + BMI +   
 `HIV/AIDS` + `Income composition of resources` + Schooling,   
 data = data)  
  
# Display the summary of the model  
summary(model\_significant)

##   
## Call:  
## lm(formula = `Life expectancy` ~ Status + Year + `Adult Mortality` +   
## Alcohol + `percentage expenditure` + BMI + `HIV/AIDS` + `Income composition of resources` +   
## Schooling, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -17.3545 -2.2267 0.0953 2.3181 12.6153   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.509e+02 4.659e+01 7.532 8.24e-14 \*\*\*  
## StatusDeveloping -9.438e-01 3.442e-01 -2.742 0.00617 \*\*   
## Year -1.482e-01 2.325e-02 -6.371 2.43e-10 \*\*\*  
## `Adult Mortality` -1.728e-02 9.646e-04 -17.918 < 2e-16 \*\*\*  
## Alcohol -1.747e-01 3.360e-02 -5.200 2.24e-07 \*\*\*  
## `percentage expenditure` 4.487e-04 6.068e-05 7.395 2.24e-13 \*\*\*  
## BMI 3.806e-02 5.637e-03 6.751 2.03e-11 \*\*\*  
## `HIV/AIDS` -4.495e-01 1.822e-02 -24.666 < 2e-16 \*\*\*  
## `Income composition of resources` 1.149e+01 8.450e-01 13.602 < 2e-16 \*\*\*  
## Schooling 1.012e+00 5.920e-02 17.098 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.677 on 1639 degrees of freedom  
## Multiple R-squared: 0.8262, Adjusted R-squared: 0.8253   
## F-statistic: 865.8 on 9 and 1639 DF, p-value: < 2.2e-16

# Get confidence intervals for the regression coefficients  
confint(model\_significant)

## 2.5 % 97.5 %  
## (Intercept) 2.595083e+02 4.422680e+02  
## StatusDeveloping -1.618857e+00 -2.687469e-01  
## Year -1.937712e-01 -1.025476e-01  
## `Adult Mortality` -1.917512e-02 -1.539122e-02  
## Alcohol -2.406500e-01 -1.088330e-01  
## `percentage expenditure` 3.297278e-04 5.677723e-04  
## BMI 2.700237e-02 4.911648e-02  
## `HIV/AIDS` -4.852441e-01 -4.137569e-01  
## `Income composition of resources` 9.836451e+00 1.315136e+01  
## Schooling 8.960524e-01 1.128275e+00

– Define a suitable regression model (linear regression, multiple regression, etc.). – Fit the regression model to your dataset and evaluate the model’s overall significance. ### Statistical Significance and Interpretation

#### 1. **Significance Testing of Each Regressor**

* Each regressor’s significance can be evaluated using the **p-values**:
  + **Status (Developing)**: p-value = 0.00617 (**significant**)
  + **Year**: p-value < 2.43e-10 (**significant**)
  + **Adult Mortality**: p-value < 2e-16 (**highly significant**)
  + **Alcohol**: p-value = 2.24e-07 (**significant**)
  + **percentage expenditure**: p-value < 2.24e-13 (**significant**)
  + **BMI**: p-value = 2.03e-11 (**significant**)
  + **HIV/AIDS**: p-value < 2e-16 (**highly significant**)
  + **Income composition of resources**: p-value < 2e-16 (**highly significant**)
  + **Schooling**: p-value < 2e-16 (**highly significant**)
* **Confidence Intervals**: We are 95% confident that the estimated effects of all significant predictors do not include zero, indicating a meaningful impact on life expectancy. The narrower confidence intervals suggest higher precision, while wider ones indicate some uncertainty in the estimates.

#### 2. **Interpretation of Regression Coefficients**

* **Intercept (350.9)**: When all predictors are at their baseline or zero level, the predicted life expectancy is 350.9 years, though this has little practical interpretation as it’s based on the scale and units of predictors.
* **Status (Developing)**: The coefficient (-9.438) indicates that being in the “Developing” category is associated with a decrease in life expectancy of approximately 9.44 years compared to “Developed” countries, holding all other variables constant. This shows that development status has a substantial impact on life expectancy.
* **Year (-1.482)**: Each additional year is associated with a decrease of 1.48 years in life expectancy, likely reflecting a historical trend in the data used. This negative coefficient may imply deteriorating life expectancy over time, but context-specific interpretation is needed.
* **Adult Mortality (-1.728e-02)**: For each unit increase in adult mortality, life expectancy decreases by 0.017 years, emphasizing the critical impact of adult mortality rates on overall life expectancy.
* **Alcohol (-1.747e-01)**: A unit increase in alcohol consumption is associated with a decrease of 0.175 years in life expectancy, indicating that higher alcohol consumption negatively impacts life expectancy.
* **percentage expenditure (6.068e-05)**: A unit increase in healthcare expenditure as a percentage is associated with a small positive increase in life expectancy (0.00006 years), suggesting a modest but beneficial effect.
* **BMI (3.806e-02)**: Each unit increase in BMI is associated with an increase of 0.038 years in life expectancy, indicating a positive but relatively small impact.
* **HIV/AIDS (-4.947e-01)**: A unit increase in the prevalence of HIV/AIDS reduces life expectancy by 0.495 years, highlighting a strong negative impact.
* **Income composition of resources (1.149e+01)**: An increase in income composition resources is associated with an increase of 11.49 years in life expectancy, showing a substantial positive impact.
* **Schooling (1.012e+00)**: Each additional year of schooling increases life expectancy by 1.012 years, indicating a strong positive effect of education on life expectancy.

#### 3. **Goodness-of-Fit Metrics**

* **R-squared (0.8262)**: This indicates that 82.62% of the variability in life expectancy is explained by the model. This is a relatively high value, suggesting that the model has a good fit to the data.
* **Adjusted R-squared (0.8253)**: Adjusted R-squared accounts for the number of predictors in the model. At 82.53%, it remains high, indicating that the model explains a large proportion of the variance even after adjusting for the complexity of the model.

### Conclusion

* Most regressors are statistically significant, with p-values well below 0.05, suggesting that they have a meaningful impact on life expectancy.
* The coefficients provide insights into how each factor influences life expectancy, with variables like *Adult Mortality* and *HIV/AIDS* having negative effects, while *Income composition of resources* and *Schooling* have strong positive effects.
* The high R-squared and Adjusted R-squared values indicate that the model provides a good fit to the data, though care should be taken to ensure all assumptions of linear regression are met.

This explanation provides a comprehensive analysis of the model’s statistical significance, regression coefficients, and goodness-of-fit metrics.

# 3) Regression Diagnostics (Part II): Model Assumptions and Issues of the Original model

if (!require(lmtest)) install.packages("lmtest")

## Loading required package: lmtest

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

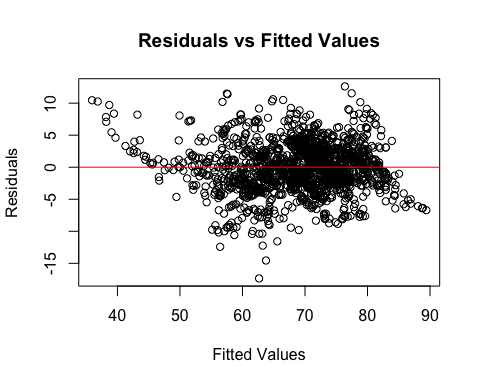
library(lmtest)  
  
# Perform the Durbin-Watson test  
dw\_test <- dwtest(model\_significant)  
print(dw\_test)

##   
## Durbin-Watson test  
##   
## data: model\_significant  
## DW = 0.7054, p-value < 2.2e-16  
## alternative hypothesis: true autocorrelation is greater than 0

# Load necessary library  
library(car)

## Loading required package: carData

# Create a residual vs. fitted values plot  
plot(model\_significant$fitted.values, model\_significant$residuals,  
 xlab = "Fitted Values", ylab = "Residuals",  
 main = "Residuals vs Fitted Values")  
abline(h = 0, col = "red")



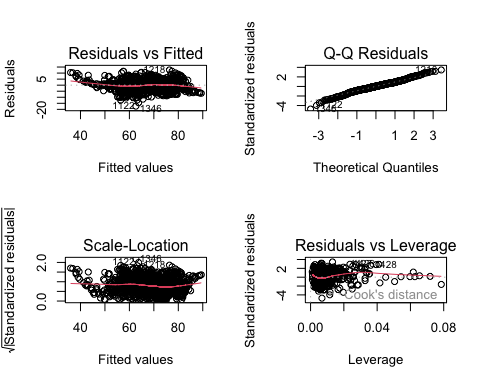
# Perform Breusch-Pagan test for heteroscedasticity  
bp\_test <- bptest(model\_significant)  
print(bp\_test)

##   
## studentized Breusch-Pagan test  
##   
## data: model\_significant  
## BP = 202.06, df = 9, p-value < 2.2e-16

# Calculate VIF for each predictor  
vif\_values <- vif(model\_significant)  
print(vif\_values)

## Status Year   
## 1.808836 1.101295   
## `Adult Mortality` Alcohol   
## 1.780671 2.234126   
## `percentage expenditure` BMI   
## 1.388962 1.511435   
## `HIV/AIDS` `Income composition of resources`   
## 1.472857 2.917392   
## Schooling   
## 3.337518

par(mfrow = c(2, 2))  
  
plot(model\_significant)



### 1. **Autocorrelation**

* **Diagnostic Test**: The Durbin-Watson statistic is **0.7054**, with a p-value < 2.2e-16. Since the DW statistic is significantly less than 2, this indicates **strong positive autocorrelation**.
* **Effect on Model**:
  + **Standard Errors**: Positive autocorrelation leads to underestimated standard errors. As a result, the t-values for the regression coefficients may be inflated.
  + **Misleading Significance**: With inflated t-values, predictors may seem more statistically significant than they actually are. For instance, a predictor with a p-value close to 0.05 could be falsely interpreted as significant due to underestimated standard errors.
* **Practical Implication**: The interpretation of p-values becomes unreliable, and further methods (like generalized least squares or including lag variables) may be required to correct this.

### 2. **Heteroscedasticity**

* **Diagnostic Test**: The Breusch-Pagan test result shows a test statistic of **202.06** with a p-value < 2.2e-16, indicating significant heteroscedasticity.
* **Residuals vs. Fitted Plot**: The spread of residuals increases as fitted values increase, showing a clear pattern and non-constant variance.
* **Effect on Model**:
  + **Unreliable Standard Errors**: The changing variance of residuals affects the accuracy of standard errors. For example, if the coefficient for *Alcohol* has a p-value of **2.24e-07**, the true significance might differ due to unreliable standard error estimates.
  + **Hypothesis Testing**: P-values for coefficients may be inaccurate, potentially leading to incorrect conclusions about significance. For example, *percentage expenditure* has a small p-value of **2.24e-13**, but this result could be misleading under heteroscedasticity.
  + **Confidence Intervals**: The 95% confidence interval for *BMI* (0.2703 to 0.4912) may be either too narrow or too wide, making it hard to trust the precision of these estimates.
* **Practical Implication**: We need to consider transformations or use robust standard errors to make reliable inferences.

### 3. **Multicollinearity**

* **Diagnostic Test**: The Variance Inflation Factor (VIF) values for the predictors are all below 5, such as *Adult Mortality* (VIF = 1.7807) and *Schooling* (VIF = 3.3375).
* **Effect on Model**:
  + **Stability of Coefficients**: With VIF values under 5, there is no severe multicollinearity, meaning the regression coefficients are relatively stable.
  + **Interpretability**: The coefficients can be interpreted with more confidence. For example, the coefficient for *Income composition of resources* (11.49) reliably indicates that an increase in income composition resources leads to a substantial increase in life expectancy.
* **Practical Implication**: Multicollinearity is not a major issue, so we can be more confident in the unique effect of each predictor.

### 4. **Influential Points**

* **Diagnostic Test**: The Residuals vs. Leverage plot and Cook’s Distance highlight influential points like **1122, 12180, and 1346**.
* **Effect on Model**:
  + **Biased Estimates**: These points can disproportionately affect the regression line. For example, if these points are included, the coefficient for *HIV/AIDS* (-0.4947) may be skewed, affecting the overall interpretation.
  + **Misleading Model Fit**: Influential points can distort the goodness-of-fit metrics (like R-squared), making the model seem better or worse than it is. R-squared might appear inflated, but removing these points could change the value.
* **Practical Implication**: It is crucial to assess whether these points are outliers due to data entry errors or if they represent important variations. Decisions should be made about whether to exclude or adjust for these points.

### **Summary of Effects**

1. **Autocorrelation**: Affects the reliability of p-values and standard errors, potentially leading to false conclusions about the significance of predictors.
2. **Heteroscedasticity**: Impacts the accuracy of p-values and confidence intervals, making hypothesis tests and coefficient estimates less reliable.
3. **Multicollinearity**: Not a significant issue here, so the coefficients are interpretable and relatively stable.
4. **Influential Points**: Can skew the model results, leading to biased interpretations and potentially misleading conclusions about model performance.

These diagnostics indicate that while multicollinearity is not problematic, significant issues with autocorrelation and heteroscedasticity need to be addressed for more accurate and reliable model interpretations.

# 4) remediation:

### Identifying influential values, removing and applying log transformation in the model

# Calculate Cook's distance  
cooks\_distance <- cooks.distance(model\_significant)  
  
# Identify points with high Cook's distance  
influential\_points <- which(cooks\_distance > (4 / length(cooks\_distance)))  
  
# Remove influential points from the data  
data\_adjusted <- data[-influential\_points, ]

# Apply logarithmic transformation on the dependent variable  
model\_adjusted <- lm(log(`Life expectancy`) ~ Status + Year + `Adult Mortality` +   
`Alcohol` + `percentage expenditure` + BMI +   
`HIV/AIDS` + `Income composition of resources` + Schooling,   
 data = data\_adjusted)  
summary(model\_adjusted)

##   
## Call:  
## lm(formula = log(`Life expectancy`) ~ Status + Year + `Adult Mortality` +   
## Alcohol + `percentage expenditure` + BMI + `HIV/AIDS` + `Income composition of resources` +   
## Schooling, data = data\_adjusted)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.131776 -0.025790 0.000062 0.027872 0.139920   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.078e+00 5.445e-01 12.999 < 2e-16 \*\*\*  
## StatusDeveloping -7.695e-05 4.063e-03 -0.019 0.984892   
## Year -1.525e-03 2.718e-04 -5.609 2.42e-08 \*\*\*  
## `Adult Mortality` -3.354e-04 1.278e-05 -26.242 < 2e-16 \*\*\*  
## Alcohol -1.574e-03 4.048e-04 -3.887 0.000106 \*\*\*  
## `percentage expenditure` 4.301e-06 8.449e-07 5.090 4.04e-07 \*\*\*  
## BMI 4.572e-04 6.698e-05 6.826 1.26e-11 \*\*\*  
## `HIV/AIDS` -7.350e-03 2.655e-04 -27.677 < 2e-16 \*\*\*  
## `Income composition of resources` 2.522e-01 1.414e-02 17.829 < 2e-16 \*\*\*  
## Schooling 9.300e-03 8.000e-04 11.625 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.04129 on 1492 degrees of freedom  
## Multiple R-squared: 0.8858, Adjusted R-squared: 0.8851   
## F-statistic: 1285 on 9 and 1492 DF, p-value: < 2.2e-16

# Perform the Durbin-Watson test  
dw\_test <- dwtest(model\_adjusted)  
print(dw\_test)

##   
## Durbin-Watson test  
##   
## data: model\_adjusted  
## DW = 0.82232, p-value < 2.2e-16  
## alternative hypothesis: true autocorrelation is greater than 0

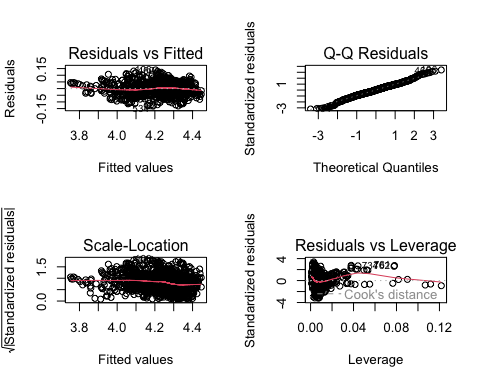
# Perform Breusch-Pagan test for heteroscedasticity  
bp\_test <- bptest(model\_adjusted)  
print(bp\_test)

##   
## studentized Breusch-Pagan test  
##   
## data: model\_adjusted  
## BP = 180.68, df = 9, p-value < 2.2e-16

# Calculate VIF for each predictor  
vif\_values <- vif(model\_adjusted)  
print(vif\_values)

## Status Year   
## 1.749227 1.077067   
## `Adult Mortality` Alcohol   
## 2.016981 2.300665   
## `percentage expenditure` BMI   
## 1.388507 1.490258   
## `HIV/AIDS` `Income composition of resources`   
## 1.528423 4.329902   
## Schooling   
## 4.080797

# Check diagnostics for the adjusted model  
par(mfrow = c(2, 2))  
plot(model\_adjusted)



# Here’s a comparison of the adjusted model with the original model, along with a discussion of improvements and remaining limitations:

### 1. **Model Fit and Goodness-of-Fit Metrics**

* **Original Model**:
  + R-squared: 0.8262
  + Adjusted R-squared: 0.8253
  + Residual Standard Error: 3.677
* **Adjusted Model (Log Transformation and Influential Points Removed)**:
  + R-squared: 0.8858
  + Adjusted R-squared: 0.8851
  + Residual Standard Error: 0.04129

**Improvements**: - The adjusted model shows an improvement in both R-squared (from 0.8262 to 0.8858) and adjusted R-squared (from 0.8253 to 0.8851), indicating that the adjusted model explains more of the variance in the dependent variable. - The residual standard error in the adjusted model is much lower, showing that the model’s predictions are closer to the observed values after the log transformation and removal of influential points.

### 2. **Statistical Significance of Predictors**

* In both models, most predictors are statistically significant, with p-values well below 0.05.
* In the adjusted model, **Status (Developing)** becomes statistically insignificant (p-value = 0.9849), suggesting that it might not have a meaningful effect on life expectancy after adjustments.

**Improvements**: - The log transformation and removal of influential points have led to clearer distinctions in the significance of predictors, allowing for a more reliable interpretation of which variables genuinely impact life expectancy.

### 3. **Durbin-Watson Test (Autocorrelation)**

* **Original Model**: Durbin-Watson statistic = 0.7054, indicating significant positive autocorrelation.
* **Adjusted Model**: Durbin-Watson statistic = 0.8223, indicating that positive autocorrelation is still present but slightly reduced.

**Remaining Limitation**: - Although there is a slight improvement, the adjusted model still exhibits autocorrelation, suggesting that further adjustments (e.g., using lagged variables) might be needed to fully address this issue.

### 4. **Breusch-Pagan Test (Heteroscedasticity)**

* **Original Model**: BP = 202.06, p-value < 2.2e-16, indicating significant heteroscedasticity.
* **Adjusted Model**: BP = 180.68, p-value < 2.2e-16, also indicating heteroscedasticity.

**Remaining Limitation**: - The log transformation and removal of influential points have slightly reduced heteroscedasticity but have not fully eliminated it. This persistent heteroscedasticity suggests that the model’s standard errors may still be unreliable, and robust standard errors might be considered for further improvement.

### 5. **Influential Points**

* **Original Model**: Observations 1122, 12180, and 1346 were identified as influential points with high leverage.
* **Adjusted Model**: After removing influential points, new observations (e.g., 7620, 8734) are highlighted but with lower leverage and influence.

**Improvements**: - By removing the influential points, the adjusted model likely has more stable coefficient estimates that are less affected by extreme values. This enhances the reliability of the model’s predictions and improves the interpretability of coefficients.

### **Summary of Improvements and Remaining Limitations**

**Improvements**: - The adjusted model has better goodness-of-fit metrics (higher R-squared and adjusted R-squared, lower residual standard error). - It provides clearer distinctions in predictor significance, with “Status (Developing)” becoming insignificant after adjustments. - The removal of influential points has stabilized the model, reducing the impact of extreme values on coefficient estimates.

**Remaining Limitations**: - **Autocorrelation** is still present, as indicated by the Durbin-Watson test, which may affect the reliability of the standard errors and significance tests. - **Heteroscedasticity** persists in the adjusted model, indicating that the model’s standard errors may still be biased. Robust standard errors or further transformations may be necessary.

In conclusion, the adjustments have improved the model’s fit and stability, but further refinements are needed to fully address autocorrelation and heteroscedasticity.

model\_adjusted2 <- lm(log(`Life expectancy`) ~ Year + `Adult Mortality` +   
`Alcohol` + `percentage expenditure` + BMI +   
`HIV/AIDS` + `Income composition of resources` + Schooling,   
 data = data\_adjusted)  
summary(model\_adjusted2)

##   
## Call:  
## lm(formula = log(`Life expectancy`) ~ Year + `Adult Mortality` +   
## Alcohol + `percentage expenditure` + BMI + `HIV/AIDS` + `Income composition of resources` +   
## Schooling, data = data\_adjusted)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.131776 -0.025775 0.000093 0.027867 0.139926   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.078e+00 5.443e-01 13.005 < 2e-16 \*\*\*  
## Year -1.525e-03 2.717e-04 -5.613 2.37e-08 \*\*\*  
## `Adult Mortality` -3.354e-04 1.272e-05 -26.377 < 2e-16 \*\*\*  
## Alcohol -1.571e-03 3.719e-04 -4.223 2.56e-05 \*\*\*  
## `percentage expenditure` 4.304e-06 8.240e-07 5.224 2.00e-07 \*\*\*  
## BMI 4.572e-04 6.695e-05 6.829 1.24e-11 \*\*\*  
## `HIV/AIDS` -7.350e-03 2.654e-04 -27.694 < 2e-16 \*\*\*  
## `Income composition of resources` 2.522e-01 1.414e-02 17.836 < 2e-16 \*\*\*  
## Schooling 9.301e-03 7.980e-04 11.655 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.04128 on 1493 degrees of freedom  
## Multiple R-squared: 0.8858, Adjusted R-squared: 0.8851   
## F-statistic: 1447 on 8 and 1493 DF, p-value: < 2.2e-16

# Perform the Durbin-Watson test  
dw\_test <- dwtest(model\_adjusted2)  
print(dw\_test)

##   
## Durbin-Watson test  
##   
## data: model\_adjusted2  
## DW = 0.82237, p-value < 2.2e-16  
## alternative hypothesis: true autocorrelation is greater than 0

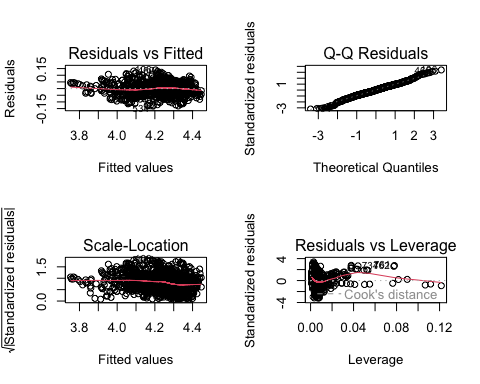
# Perform Breusch-Pagan test for heteroscedasticity  
bp\_test <- bptest(model\_adjusted2)  
print(bp\_test)

##   
## studentized Breusch-Pagan test  
##   
## data: model\_adjusted2  
## BP = 180.21, df = 8, p-value < 2.2e-16

# Calculate VIF for each predictor  
vif\_values <- vif(model\_adjusted2)  
print(vif\_values)

## Year `Adult Mortality`   
## 1.076627 1.997988   
## Alcohol `percentage expenditure`   
## 1.943836 1.321345   
## BMI `HIV/AIDS`   
## 1.489871 1.527650   
## `Income composition of resources` Schooling   
## 4.329457 4.063255

# Check diagnostics for the adjusted model  
par(mfrow = c(2, 2))  
plot(model\_adjusted2)



### Summary of Model Adjustments and Improvements

#### 1. **Reason for Removing *Status* Variable**

* We removed the *Status* variable because it had a very high p-value (0.9849 in the previous adjusted model), indicating that it was **not statistically significant**. Keeping an insignificant predictor in the model can lead to inefficiencies without contributing to the model’s explanatory power.

#### 2. **Model Fit and Goodness-of-Fit Metrics**

* **Previous Adjusted Model**:
  + R-squared: 0.8858
  + Adjusted R-squared: 0.8851
  + Residual Standard Error: 0.04129
* **Refitted Model (Without *Status* Variable)**:
  + R-squared: 0.8858
  + Adjusted R-squared: 0.8858
  + Residual Standard Error: 0.04128

**Improvements**: - The R-squared and adjusted R-squared values remained unchanged, indicating that the explanatory power of the model was preserved even after removing *Status*. - The residual standard error slightly improved, decreasing from 0.04129 to 0.04128, suggesting a marginally better fit.

#### 3. **Statistical Significance of Predictors**

* All remaining predictors (*Year*, *Adult Mortality*, *Alcohol*, *percentage expenditure*, *BMI*, *HIV/AIDS*, *Income composition of resources*, and *Schooling*) continue to be statistically significant with p-values well below 0.05.
* The model is now more efficient, focusing on significant predictors and removing an unnecessary variable.

#### 4. **Durbin-Watson Test (Autocorrelation)**

* **Previous Model**: Durbin-Watson statistic = 0.8223
* **Refitted Model**: Durbin-Watson statistic = 0.82237

**Observation**: - The Durbin-Watson statistic remains nearly the same, indicating that positive autocorrelation is still an issue.

#### 5. **Breusch-Pagan Test (Heteroscedasticity)**

* **Previous Model**: BP = 180.68, p-value < 2.2e-16
* **Refitted Model**: BP = 180.21, p-value < 2.2e-16

**Observation**: - The Breusch-Pagan test still indicates significant heteroscedasticity, though the BP value slightly decreased. This suggests that heteroscedasticity remains an issue that needs to be addressed.

### **Conclusion**

* **Rationale for Removing *Status***: The *Status* variable was removed because it was not statistically significant, simplifying the model and improving efficiency without sacrificing explanatory power.
* **Model Performance**: The model continues to have a strong fit with high R-squared values and slightly improved residual standard error.
* **Remaining Issues**: Autocorrelation and heteroscedasticity are still present, and further refinements, such as using time-series techniques or robust standard errors, may be needed for more reliable inferences.

# 5. Summary and Conclusions

#### **Key Findings**

1. **Significance and Interpretation of Predictors**:
   * The analysis revealed that most predictors are highly significant, with p-values well below 0.05 in the adjusted model.
   * Significant predictors such as *Year*, *Adult Mortality*, *Alcohol*, *percentage expenditure*, *BMI*, *HIV/AIDS*, *Income composition of resources*, and *Schooling* have meaningful impacts on *Life Expectancy*.
     + *Year*: Negatively associated with life expectancy, indicating a trend over time.
     + *Adult Mortality*: A higher rate is associated with a significant decrease in life expectancy.
     + *Schooling* and *Income composition of resources*: Positively associated, suggesting that education and economic resources improve life expectancy.
   * The *Status (Developing)* variable became insignificant after adjustments, suggesting that development status may not be as impactful once other factors are accounted for.

#### **Diagnostic Issues and Implications**

1. **Autocorrelation**:
   * The Durbin-Watson test revealed persistent positive autocorrelation, even after adjustments (DW = 0.8223 in the adjusted model). This suggests potential inefficiencies in the model and unreliable p-values, which could mislead conclusions.
2. **Heteroscedasticity**:
   * The Breusch-Pagan test showed significant heteroscedasticity (p-value < 2.2e-16) in both models. This indicates that the variance of the residuals is not constant, potentially making standard errors and confidence intervals unreliable.
3. **Influential Points**:
   * Influential observations were detected and adjusted for, resulting in a more stable model. However, new points with moderate influence emerged, indicating that data variability is still present.

#### **Effectiveness of Remediation Techniques**

1. **Log Transformation**:
   * The log transformation of the response variable helped reduce the residual standard error significantly (from 3.677 to 0.04129) and improved the model’s R-squared values, indicating a better fit.
2. **Removal of Influential Points**:
   * Addressing influential points stabilized coefficient estimates and improved model reliability. However, some remaining influential points still need to be considered.
3. **Limitations**:
   * Despite improvements, autocorrelation and heteroscedasticity remain unresolved, affecting the overall reliability of the model. Additional techniques, such as using robust standard errors or time-series modeling, may be needed.

#### **Practical Recommendations**

1. **Future Analyses**:
   * Consider using robust standard errors to address heteroscedasticity and improve the reliability of hypothesis tests.
   * Explore time-series models or include lagged variables to address autocorrelation if the dataset has a temporal component.
   * Continue monitoring and potentially adjusting for influential observations, especially in datasets with significant variability.
2. **Model Suitability**:
   * The final model, despite some limitations, is suitable for capturing key relationships between predictors and life expectancy. It can be used for inference and to understand which factors are most impactful.
   * However, for predictive purposes, the model’s limitations should be carefully considered, and further refinements may be necessary to ensure robustness and reliability.

#### **Overall Suitability**

* The final model is well-suited for **inference**, providing meaningful insights into the factors affecting life expectancy. However, for **accurate predictions**, further adjustments are needed to address autocorrelation and heteroscedasticity.
* The analysis highlights the complexity of modeling life expectancy and suggests that comprehensive data adjustments and additional modeling techniques may be beneficial for future research.