

Socioeconomic and Environmental Determinants of HIV Prevalence: A Cross-Country Regression Analysis

Phan Anh Le

2024

Introduction and Motivation

This project examines multiple countries variation in HIV prevalence through the lens of socioeconomic and environmental indicators. Motivated by the idea that population disease burden reflects broader structural conditions, the analysis uses regression modeling to explore associations between HIV prevalence and factors such as economic growth, urbanization, political stability, and industrialization. The project emphasizes data preparation, model selection, and diagnostic testing, illustrating the strengths and limitations of applying econometric methods to population health data.

Hypothesis

This study tests two related hypotheses concerning the structural determinants of HIV prevalence. First, countries with higher levels of economic development, urbanization, political stability, and industrialization are expected to be associated with lower levels of HIV prevalence when the outcome is treated as a continuous variable. Second, when HIV prevalence is recoded as a categorical outcome distinguishing relatively low and high prevalence, countries with higher levels of economic development and urbanization are expected to be more likely to fall into the low-prevalence category. These hypotheses are tested using linear and logistic regression models, respectively.

Data & Variables

The analysis uses a dataset covering approximately 40 countries. The unit of analysis is the country. The primary dependent variable is HIV prevalence among individuals aged 15–49. Independent variables include measures of economic development, urbanization, political stability, and CO₂ emissions, the latter used as a proxy for industrialization. Control variables are included to capture additional country-level characteristics. All variables are measured at a single point in time, forming a cross-sectional dataset.

Data Cleaning and Open Package

Open Package

```
library(dplyr)
library(ggplot2)
library(car)
library(sandwich)
library(lmtest)
```

```
library(leaps)
library(readr)
library(caret)
Data <- read.csv("C:/Users/admin/Downloads/C3 AIDS.csv")
```

Data Cleaning

```
str(Data)
```

```
## 'data.frame': 41 obs. of 26 variables:
## $ Country : chr "Angola" "Benin" "Burkina Faso" "Burundi" ...
## $ Literacy.Rate.....2005. : chr "67.40%" "34.70%" "23.60%" "59%" ...
## $ GDP.per.Capita..2006. : chr "1970" "530" "440" "100" ...
## $ Population.Density..2006. : chr "13" "79" "52" "318" ...
## $ Political.Stability..120.pt.scale..2006 : chr "88.3" "72" "89.7" "96.7" ...
## $ CO2.Emissions..Metric.Tons.per.Capita..2004. : chr "0.51" "0.29" "0.08" "0.03" ...
## $ GDP.Growth..2006. : chr "18.60%" "4.10%" "6.40%" "5.10%" ...
## $ Aid.as...of.GNI..2005. : chr "1.50%" "8.20%" "12.80%" "46.80%" ...
## $ Urban.Population.Percentage..2006. : chr "54%" "41%" "19%" "10%" ...
## $ HIV.Prevalence...of.Individuals.ages.15.49 : num 3.7 1.8 2 3.3 10.7 3.5 0.1 3.2 3.1 3.2 ...
## $ X : logi NA NA NA NA NA NA .
..
## $ X.1 : logi NA NA NA NA NA NA .
..
## $ X.2 : logi NA NA NA NA NA NA .
..
## $ X.3 : logi NA NA NA NA NA NA .
..
## $ AIDS : chr "Literacy Rate" "GDP per Capita" "Population Density" "Gov't Stability" ...
## $ X.4 : chr "World Bank Education Statistics Database. Collected using 2005 data for percentage of adults ages 15 and over."|__truncated__ "World Development Indicators (World Bank Data). Collected using 2006 data in current US $ http://web.worldbank"|__truncated__ "Health, Nutrition, and Population Statistics (World Bank Data). Collected using 2006 data in people per km2 ht"|__truncated__ "Fund for Peace (Failed State Index). Collected using 2006 data on a 120 point scale (0=Stable, 120=Unstable) ht"|__truncated__ ...
## $ X.5 : logi NA NA NA NA NA NA .
..
## $ X.6 : logi NA NA NA NA NA NA .
```

```

..
## $ X.7 : logi NA NA NA NA NA NA .
..
## $ X.8 : logi NA NA NA NA NA NA .
..
## $ X.9 : logi NA NA NA NA NA NA .
..
## $ X.10 : logi NA NA NA NA NA NA .
..
## $ X.11 : logi NA NA NA NA NA NA .
..
## $ X.12 : logi NA NA NA NA NA NA .
..
## $ X.13 : logi NA NA NA NA NA NA .
..
## $ X.14 : logi NA NA NA NA NA NA .
..

Data$GDP.Growth..2006. <- sub("%", "", Data$GDP.Growth..2006.)
Data$Literacy.Rate.....2005. <- sub("%", "", Data$Literacy.Rate.....2005.)
Data$Aid.as...of.GNI..2005. <- sub("%", "", Data$Aid.as...of.GNI..2005.)
Data$Urban.Population.Percentage..2006. <- sub("%", "", Data$Urban.Population.P
ercentage..2006.)

Data$GDP.Growth..2006. <- as.numeric(as.character(Data$GDP.Growth..2006.))

Data$Literacy.Rate.....2005. <- as.numeric(as.character(Data$Literacy.Rate..
.....2005.))

Data$GDP.per.Capita..2006. <- as.numeric(as.character(Data$GDP.per.Capita..20
06.))

Data$Population.Density..2006. <- as.numeric(as.character(Data$Population.Den
sity..2006.))

Data$Political.Stability..120.pt.scale..2006 <- as.numeric(as.character(Data$
Political.Stability..120.pt.scale..2006))

Data$CO2.Emissions..Metric.Tons.per.Capita..2004. <- as.numeric(as.character(
Data$CO2.Emissions..Metric.Tons.per.Capita..2004.))

Data$GDP.Growth..2006. <- as.numeric(as.character(Data$GDP.Growth..2006.))
Data$Aid.as...of.GNI..2005. <- as.numeric(as.character(Data$Aid.as...of.GNI..
2005.))

Data$Urban.Population.Percentage..2006. <- as.numeric(as.character(Data$Urban
.Population.Percentage..2006.))

Data <- Data %>%
  select(-Country)%>%
  select(-(X:X.14))
summary(Data)

```

```
## Literacy.Rate.....2005. GDP.per.Capita..2006. Population.Density..2006.
## Min. :23.60 Min. : 100.0 Min. : 3.0
## 1st Qu.:37.60 1st Qu.: 272.5 1st Qu.: 17.5
## Median :54.00 Median : 435.0 Median : 49.0
## Mean :54.57 Mean : 720.0 Mean : 112.2
## 3rd Qu.:68.05 3rd Qu.: 735.0 3rd Qu.: 89.5
## Max. :90.00 Max. :8510.0 Max. :1198.0
## NA's :10 NA's :3 NA's :1
## Political.Stability..120.pt.scale..2006
## Min. : 66.10
## 1st Qu.: 83.25
## Median : 89.70
## Mean : 90.01
## 3rd Qu.: 96.60
## Max. :112.30
## NA's :2
## CO2.Emissions..Metric.Tons.per.Capita..2004. GDP.Growth..2006.
## Min. :0.0100 Min. : -5.600
## 1st Qu.:0.0750 1st Qu.: 4.100
## Median :0.1500 Median : 5.200
## Mean :0.2151 Mean : 5.412
## 3rd Qu.:0.2400 3rd Qu.: 7.400
## Max. :1.0300 Max. :18.600
## NA's :2 NA's :1
## Aid.as...of.GNI..2005. Urban.Population.Percentage..2006.
## Min. : -37.80 Min. :10.00
## 1st Qu.: 7.35 1st Qu.:20.00
## Median :12.65 Median :30.00
## Mean :14.49 Mean :31.46
## 3rd Qu.:20.20 3rd Qu.:39.00
## Max. :54.10 Max. :87.00
## NA's :7
## HIV.Prevalence...of.Individuals.ages.15.49
## Min. : 0.050
## 1st Qu.: 0.900
## Median : 2.200
## Mean : 3.793
## 3rd Qu.: 3.500
## Max. :23.200
##
```

Rename the variables

```
names(Data) <- c("Literacy", "GDP", "Pol_Dens", "Pol_Stab", "CO2", "GDP_Grow",
, "AIDS_GNI", "Urban_Pop", "HIV_Prev")
```

Remove the columns with too many NAs

```
Data$Literacy <- NULL  
Data$AIDS_GNI <- NULL
```

Remove NAs from the remaining column

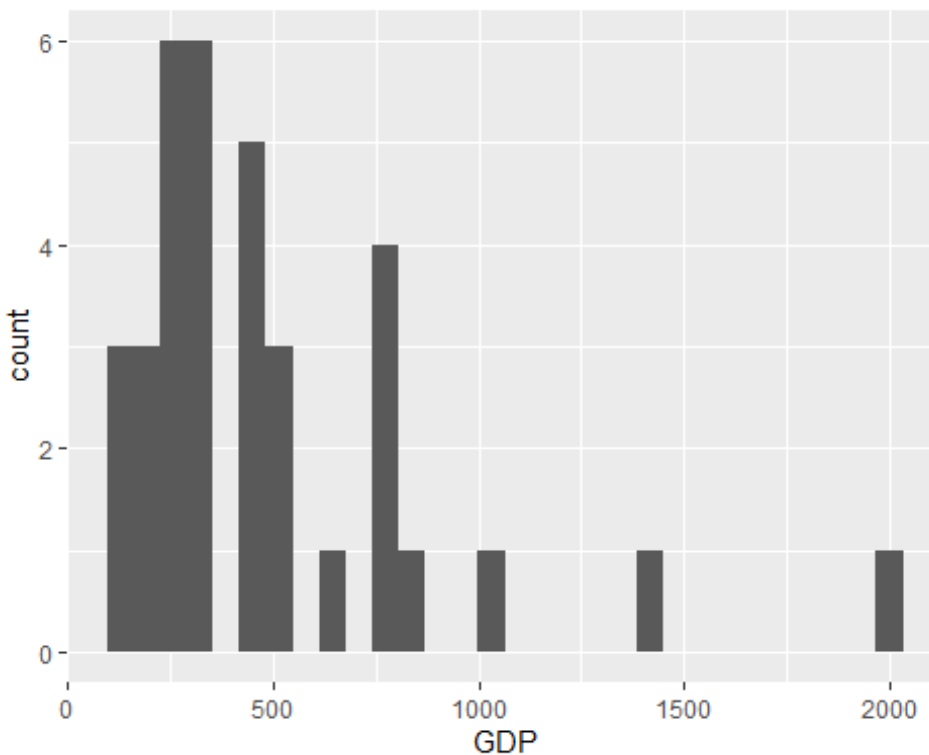
```
Data_new <- na.omit(Data)
```

Variable Examination

First of all, we begin by examining the histograms of each variable to determine if log transformation is necessary for any of them based on their distribution.

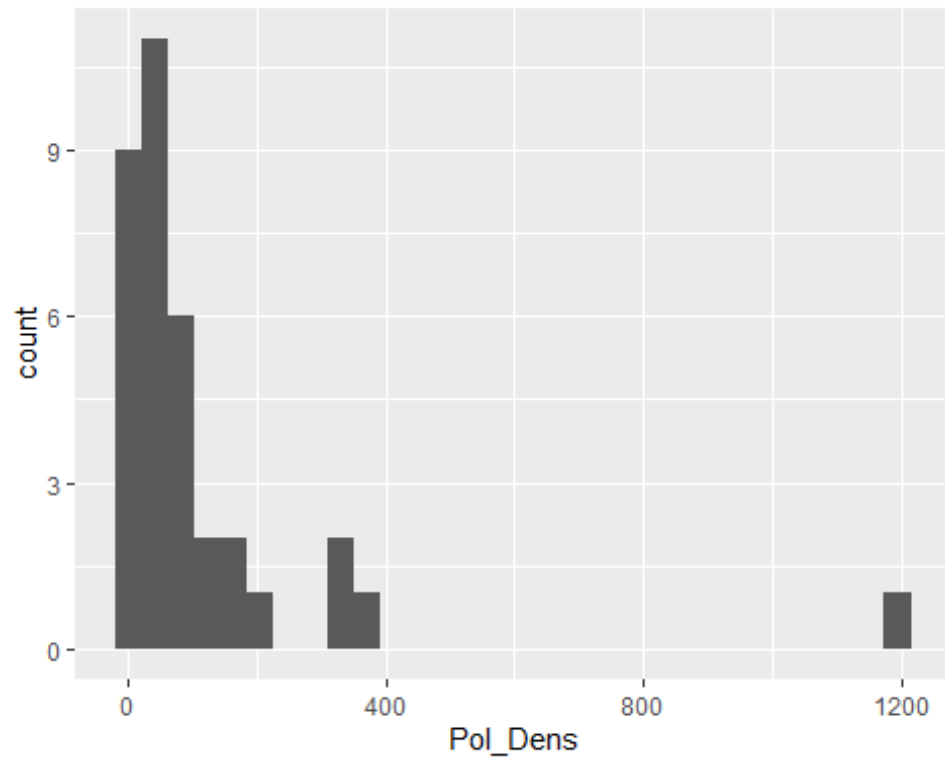
```
Data_new %>% ggplot(aes(GDP))+  
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



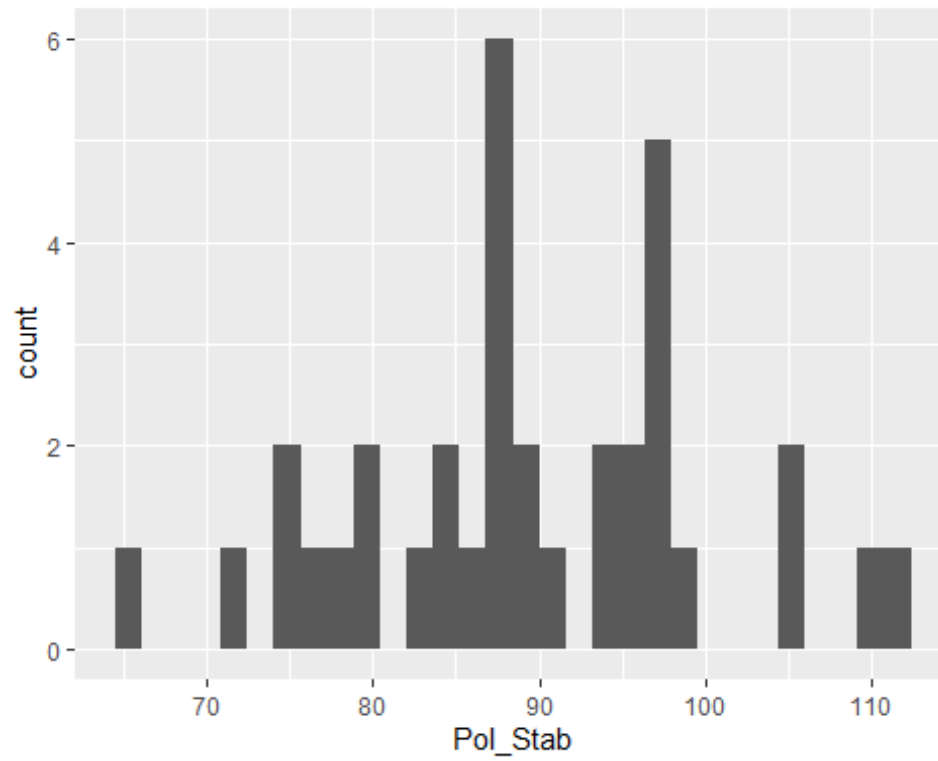
```
Data_new %>% ggplot(aes(Pol_Dens))+  
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



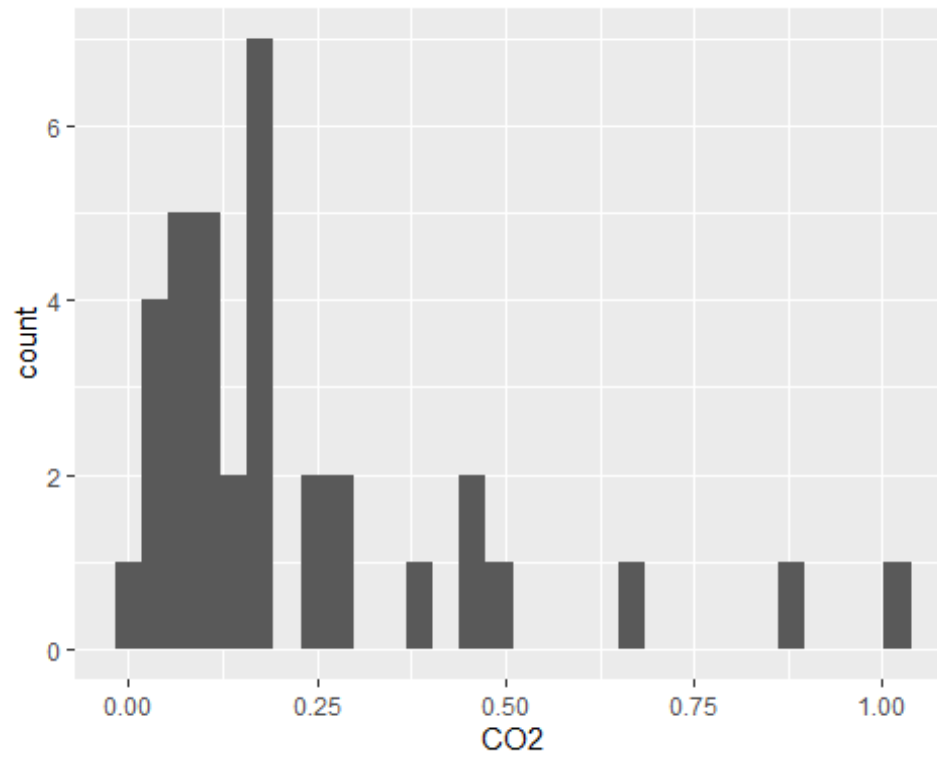
```
Data_new %>% ggplot(aes(Pol_Stab))+  
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



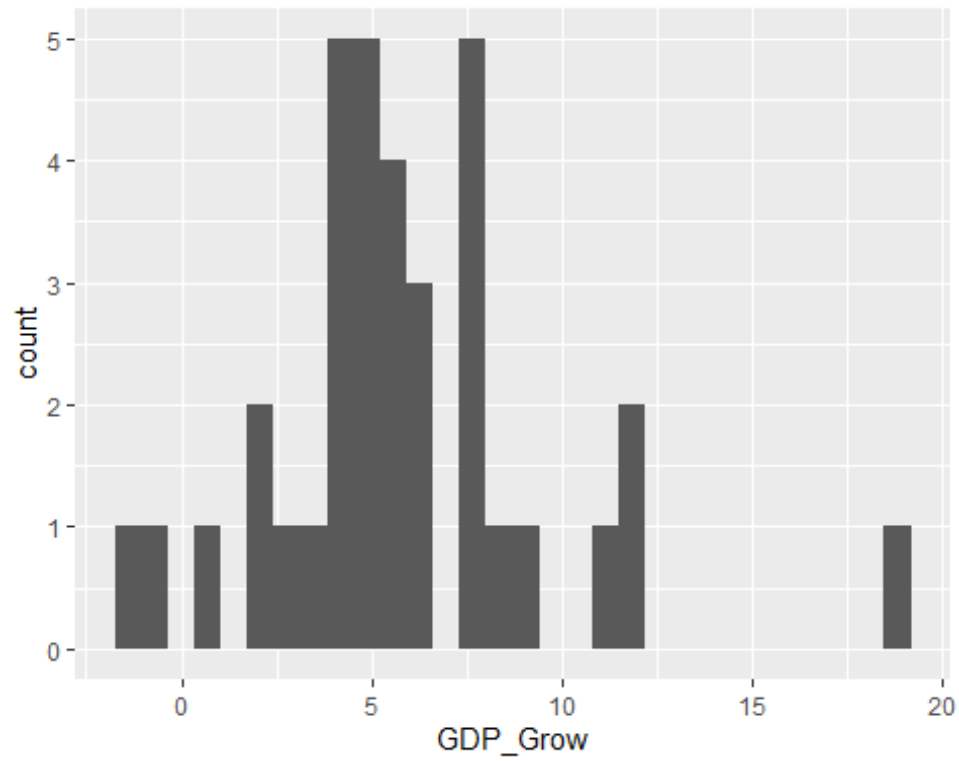
```
Data_new %>% ggplot(aes(C02))+  
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



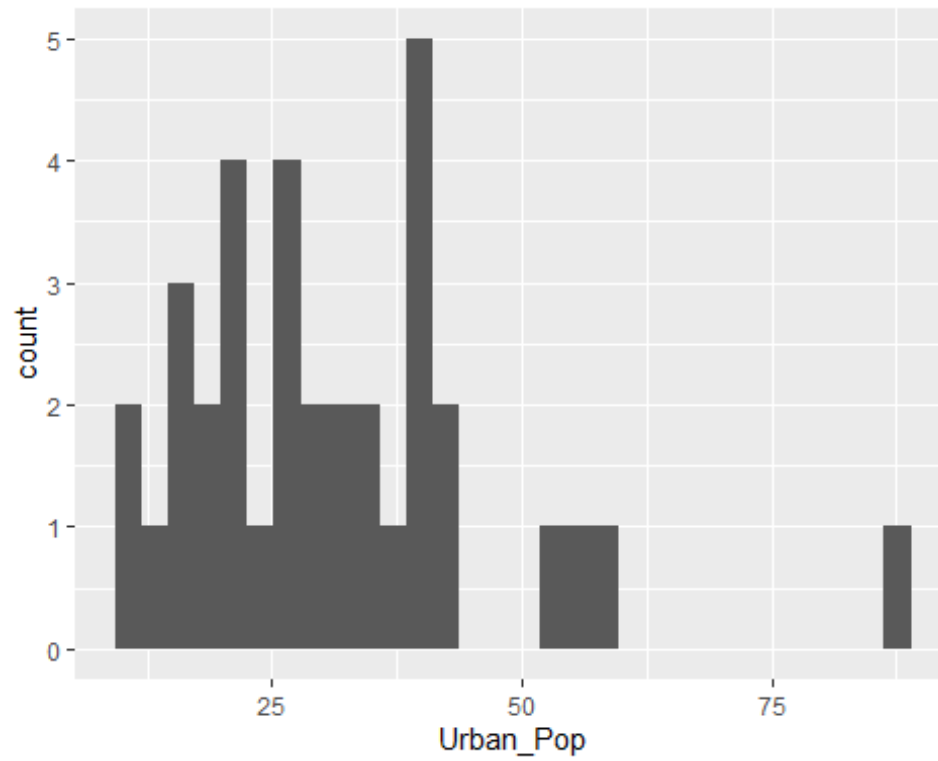
```
Data_new %>% ggplot(aes(GDP_Grow))+  
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



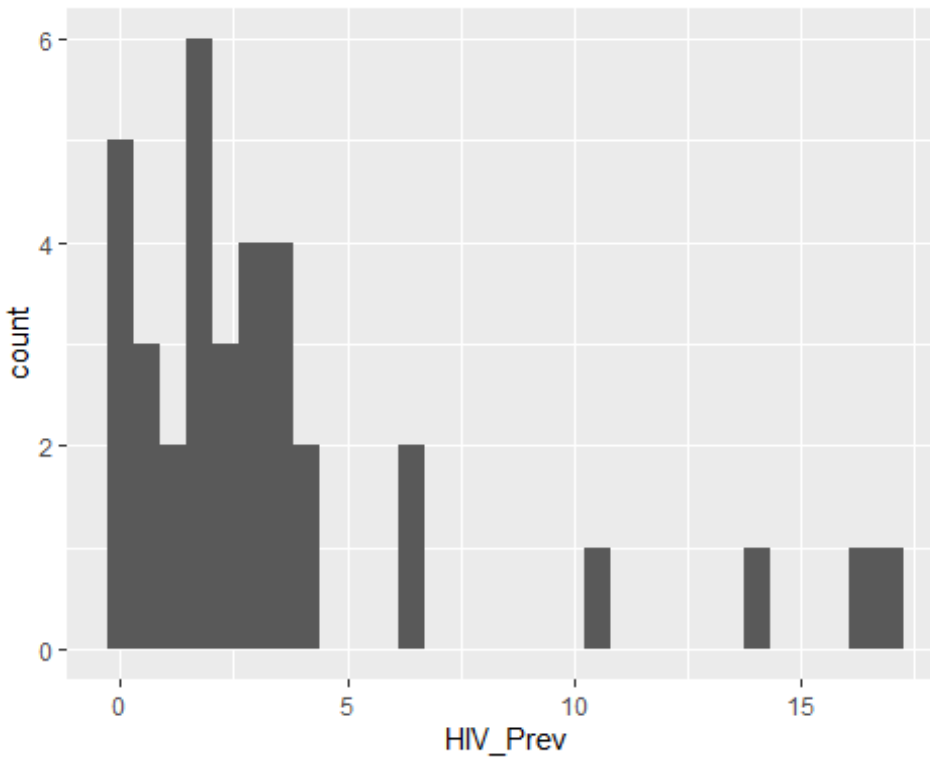
```
Data_new %>% ggplot(aes(Urban_Pop))+  
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
Data_new %>% ggplot(aes(HIV_Prev))+  
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

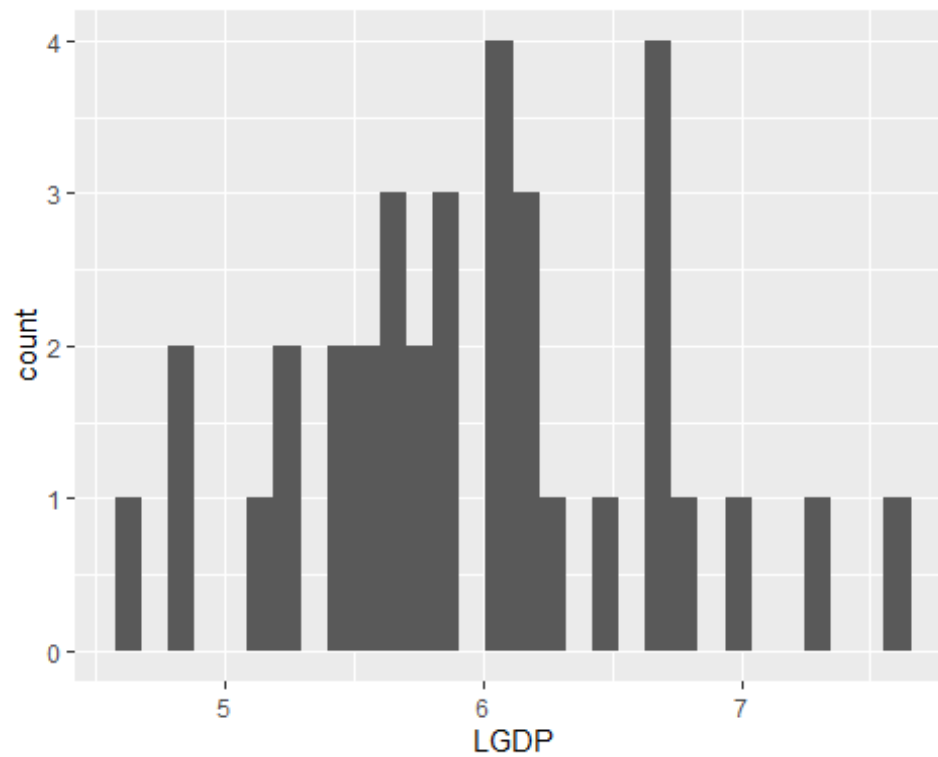


Upon reviewing the histograms, it becomes evident that the distributions of GDP, Pol_Dens, CO2, and HIV_Prev are significantly right-skewed. To normalize these distributions, we will apply log transformation to each of them.

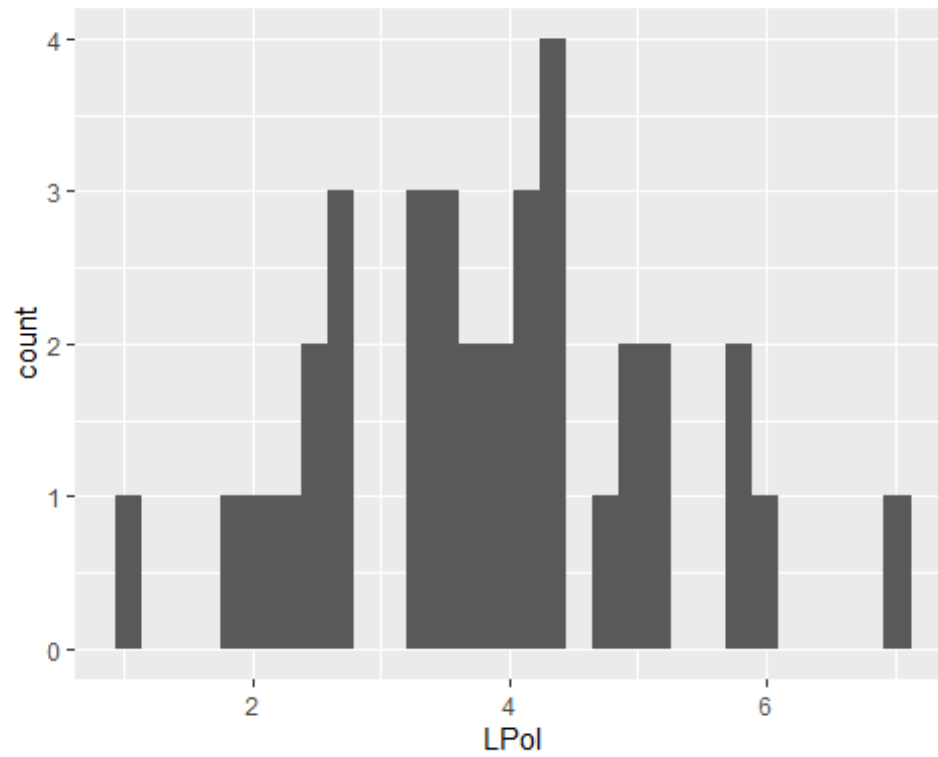
```
Data_new1 <- Data_new %>% mutate(LGDP = log(GDP), LPol = log(Pol_Dens), LCO2 = log(CO2), LHIV = log(HIV_Prev) )
```

```
Data_new1 %>%  
  ggplot(aes(LGDP))+  
  geom_histogram()
```

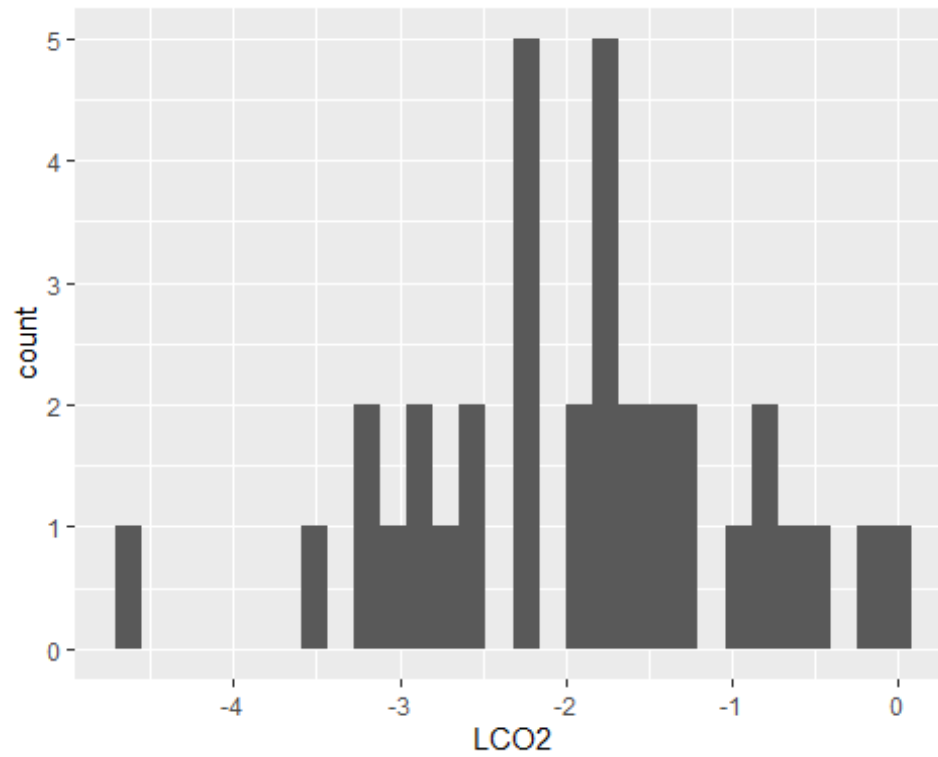
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



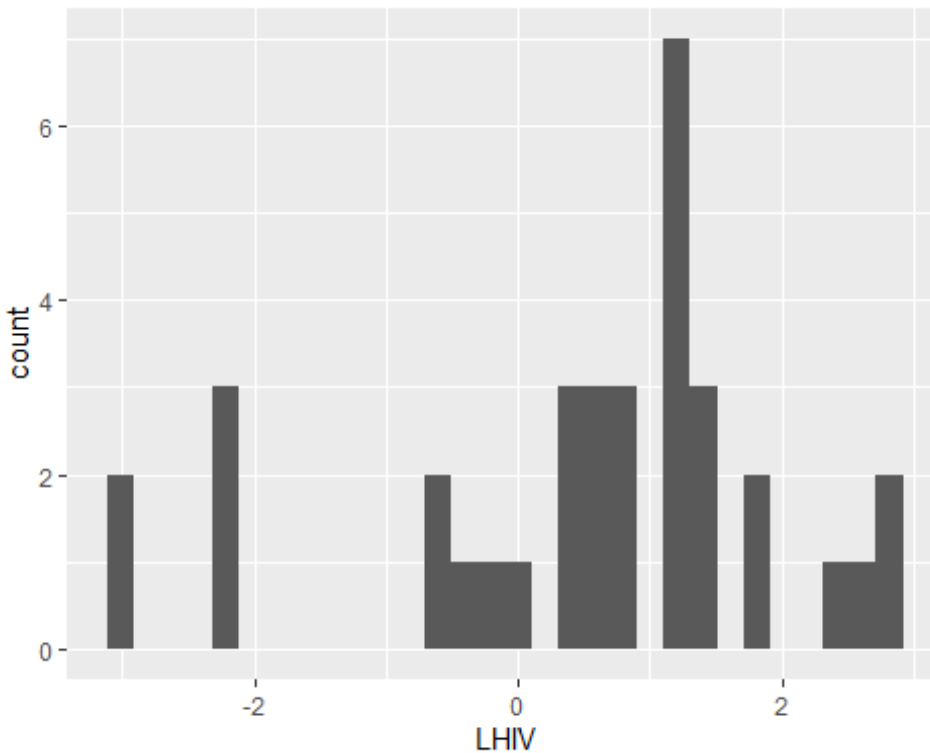
```
Data_new1 %>%  
  ggplot(aes(LPo1))+  
  geom_histogram()  
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
Data_new1 %>%  
  ggplot(aes(LC02))+  
  geom_histogram()  
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
Data_new1 %>%  
  ggplot(aes(LHIV))+  
  geom_histogram()  
  
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Given that the histograms of the log-transformed variables demonstrate clear normality, we will retain and utilize these variables in our model.

```
Data_new2 <- Data_new1 %>%
  select(Pol_Stab, GDP_Grow, Urban_Pop, LGDP:LHIV)
```

Using Best Subset and Stepwise Regression

To select the optimal variables for our model, we will compare the outcomes from the best subset method with those obtained through stepwise regression. Let's start by examining the best subsets identified for this model.

```
models <- regsubsets(LHIV~., data = Data_new2, nvmax = 6)
summary(models)

## Subset selection object
## Call: regsubsets.formula(LHIV ~ ., data = Data_new2, nvmax = 6)
## 6 Variables (and intercept)
##           Forced in Forced out
## Pol_Stab      FALSE      FALSE
## GDP_Grow      FALSE      FALSE
## Urban_Pop     FALSE      FALSE
## LGDP          FALSE      FALSE
## LPol          FALSE      FALSE
## LCO2          FALSE      FALSE
## 1 subsets of each size up to 6
## Selection Algorithm: exhaustive
```

```
##           Pol_Stab GDP_Grow Urban_Pop LGDP LPol LCO2
## 1  ( 1 ) " "      " "      " "      " " " " "*"
## 2  ( 1 ) " "      " "      "*"      " " " " "*"
## 3  ( 1 ) " "      "*"      "*"      " " " " "*"
## 4  ( 1 ) " "      "*"      "*"      "*" " " " "*"
## 5  ( 1 ) "*"      "*"      "*"      "*" " " " "*"
## 6  ( 1 ) "*"      "*"      "*"      "*" "*" " " "*"

```

```
sumData <- summary(models)
```

The code to check for the highest R squared model

```
AdR2 = which.max(sumData$adjr2)
head(AdR2)
```

```
## [1] 5
```

The findings indicate that the most suitable model includes five variables: Pol_Stab, GDP_Grow, Urban_Pop, LGDP, and LCO2. Next, we will examine the outcomes of our two-way stepwise regression analysis to see if it corroborates these results.

```
stepwise <- step(lm(LHIV ~. , Data_new2), direction="both")
```

```
## Start:  AIC=18.57
## LHIV ~ Pol_Stab + GDP_Grow + Urban_Pop + LGDP + LPol + LCO2
##
##           Df Sum of Sq    RSS    AIC
## - LPol      1    0.1479 40.029 16.699
## <none>                        39.881 18.570
## - Pol_Stab  1    3.8248 43.706 19.775
## - LGDP      1    4.9571 44.839 20.670
## - GDP_Grow  1    5.4940 45.375 21.087
## - Urban_Pop 1   12.6720 52.553 26.227
## - LCO2      1   13.4806 53.362 26.761
##
## Step:  AIC=16.7
## LHIV ~ Pol_Stab + GDP_Grow + Urban_Pop + LGDP + LCO2
##
##           Df Sum of Sq    RSS    AIC
## <none>                        40.029 16.699
## - Pol_Stab  1    3.9913 44.021 18.026
## + LPol      1    0.1479 39.881 18.570
## - LGDP      1    4.9404 44.970 18.772
## - GDP_Grow  1    6.0560 46.085 19.630
## - Urban_Pop 1   13.3059 53.335 24.744
## - LCO2      1   14.7212 54.750 25.660

```

```
summary(stepwise)
```

```
##
## Call:
## lm(formula = LHIV ~ Pol_Stab + GDP_Grow + Urban_Pop + LGDP +

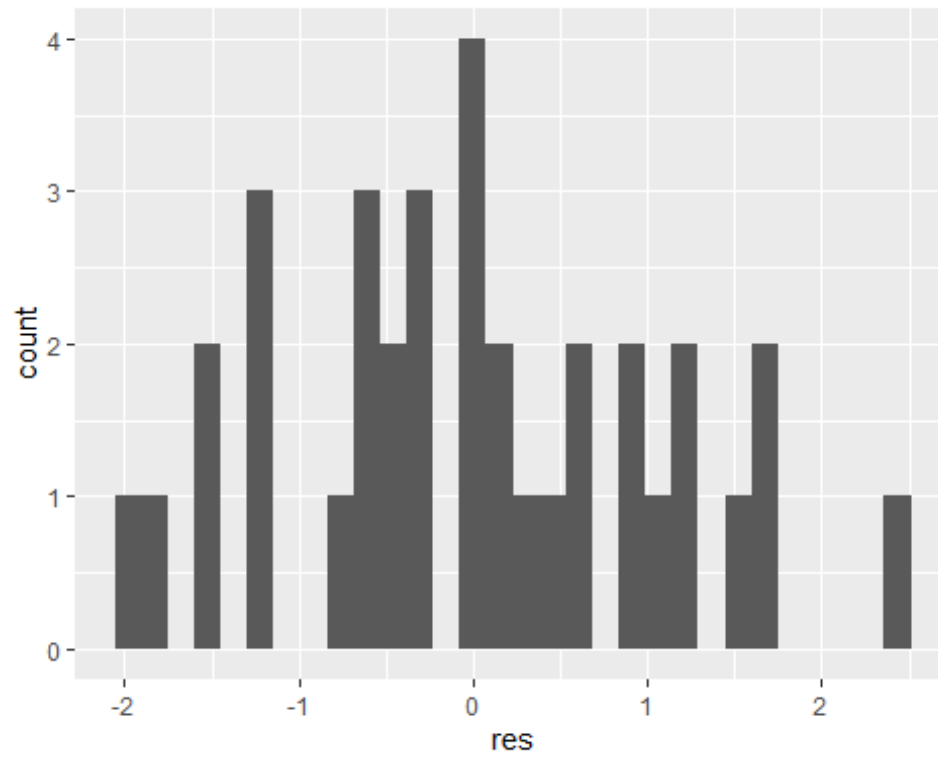
```

```
##      LC02, data = Data_new2)
##
## Residuals:
##      Min        1Q      Median        3Q        Max
## -2.05372 -0.60866 -0.05828  0.79922  2.35836
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.05339    3.12763   1.296  0.20520
## Pol_Stab     -0.03377    0.01986  -1.700  0.09975 .
## GDP_Grow      0.11864    0.05664   2.095  0.04506 *
## Urban_Pop     0.04346    0.01400   3.105  0.00423 **
## LGDP         -0.71162    0.37615  -1.892  0.06853 .
## LC02         -0.86676    0.26541  -3.266  0.00280 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.175 on 29 degrees of freedom
## Multiple R-squared:  0.5025, Adjusted R-squared:  0.4167
## F-statistic: 5.857 on 5 and 29 DF,  p-value: 0.000739
```

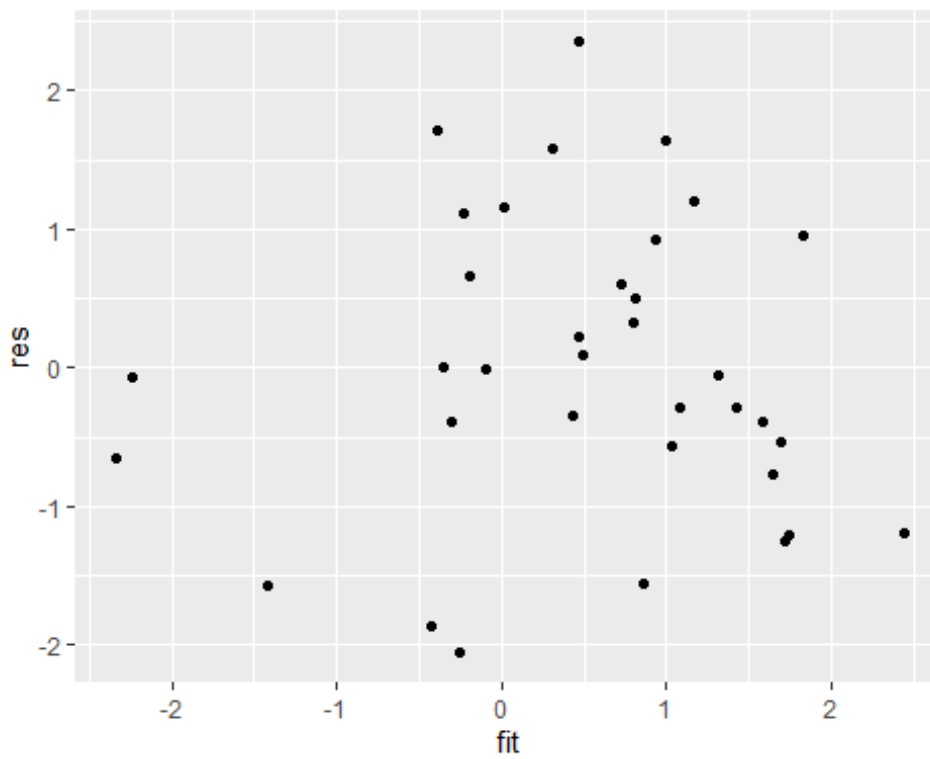
Observing the final model from the stepwise regression analysis, it suggests removing the log-transformed LPol from the variables, aligning with the selection made by the best subsets method. Therefore, we'll proceed to update our model accordingly and then generate a histogram of the residuals and a scatterplot to visualize the model's performance.

```
A <- lm(LHIV ~ Pol_Stab+ GDP_Grow+ Urban_Pop+ LGDP+ LC02, Data_new2)
Data_new3 <- Data_new2 %>%
  mutate(res=residuals(A), fit=fitted.values(A))
Data_new3 %>%
  ggplot(aes(res))+
  geom_histogram()

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
Data_new3 %>%  
  ggplot(aes(fit,res))+  
  geom_point()
```



Check for the mathematical assumptions

```
shapiro.test(Data_new3$res)
```

```
##  
##  Shapiro-Wilk normality test  
##  
## data:  Data_new3$res  
## W = 0.98481, p-value = 0.8994
```

```
ncvTest(A)
```

```
## Non-constant Variance Score Test  
## Variance formula: ~ fitted.values  
## Chisquare = 0.2378819, Df = 1, p = 0.62574
```

Normality: The residual histogram is expected to show a normal distribution. The observed plot depicts a distribution that resembles a bell shape but lacks perfect symmetry, displaying several peaks. This indicates the residuals may not align perfectly with a normal distribution. Nonetheless, the deviation from normality isn't pronounced, fulfilling this assumption.

Linearity: A linear relationship would be indicated by a scatterplot of points that appears random and without a specific pattern. The plot we have shows such a random distribution of residuals, which is encouraging. However, there's noticeable variation in the spread of residuals across different values, indicating some inconsistency.

Independence: The assumption of independence among residuals implies no apparent pattern should exist in their plot. The observed randomness in the scatter of residuals affirms this assumption, indicating that residuals do not influence each other.

Homoscedasticity: The Non-constant Variance Score Test results, showcasing a Chi-square value of 0.2378819 with one degree of freedom and a significantly high p-value, lead us to retain the null hypothesis that asserts constant variance of residuals. This lack of statistical evidence against constant variance across different fitted values corroborates the homoscedasticity assumption, a cornerstone for the validity of linear regression analysis.

Do Robust Standard Errors

```
coeftest(A, vcov = vcovHC(A, type = "HC4"))
```

```
##  
## t test of coefficients:  
##  
##           Estimate Std. Error t value Pr(>|t|)  
## (Intercept)  4.053386   2.716807  1.4920 0.1465073  
## Pol_Stab     -0.033765   0.019299 -1.7496 0.0907707 .  
## GDP_Grow      0.118637   0.061603  1.9258 0.0639800 .  
## Urban_Pop     0.043459   0.011710  3.7113 0.0008708 ***  
## LGDP          -0.711622   0.364219 -1.9538 0.0604275 .  
## LC02          -0.866765   0.213723 -4.0556 0.0003440 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The term for the Intercept stands out as markedly negative, with a value of -3.457424. This indicates a low log of HIV prevalence when other variables are at zero, establishing a statistically significant baseline at the 0.01 significance level.

For the LogCO2 variable, there's a notable negative impact, evidenced by a coefficient of -0.969377. This suggests a decrease in the log of HIV prevalence as CO2 emissions rise, a relationship confirmed to be highly significant at the 0.001 level.

The effect of the Pop_per variable is positively significant, with a coefficient of 0.048547. This implies that areas with a denser population percentage tend to exhibit higher rates of HIV prevalence, a result significant at the 0.05 level.

Regarding GDP_Growth, its coefficient of 0.063214 indicates a positive correlation with the log of HIV prevalence. However, with a p-value of 0.1785839, the statistical significance of this relationship is uncertain, pointing to a less reliable impact of GDP growth on HIV prevalence rates.

However, these observations are made under the condition of established homoscedasticity, which implies that the variability of the residual terms is consistent across the model.

Make the predictions

```
HIV_new <- data.frame(
  Pol_Stab = c(60, 65, 70, 75, 80),
  GDP_Grow = c(1, 2, 4, 5, 6),
  Urban_Pop = c(15, 20, 33, 54, 22),
  LGDP = c(1, 5, 4, 7, 3),
  LCO2 = c(-1, -4, 0, -2, 0)
)

predictData <- predict(A, newdata = HIV_new, interval = "prediction", level = 0.95)
print(predictData)
```

	fit	lwr	upr
## 1	2.9531471	-1.944705609	7.851000
## 2	2.8740619	0.005644797	5.742479
## 3	0.7520393	-2.600656205	4.104735
## 4	1.2131532	-1.480488976	3.906795
## 5	0.8852371	-2.961814372	4.732289

Given that the HIV prevalence data has undergone log transformation, an additional step to reverse this transformation is necessary to obtain accurate results. This process will allow us to interpret the outcomes in their original scale, providing a clearer understanding of the HIV prevalence rates.

```

predictUNLOG <- as.data.frame(predictData)
predictUNLOG$fit <- exp(predictUNLOG$fit)
predictUNLOG$lwr <- exp(predictUNLOG$lwr)
predictUNLOG$upr <- exp(predictUNLOG$upr)
HIV_new_predictions <- cbind(predictUNLOG)
print(HIV_new_predictions)

```

```

##           fit           lwr           upr
## 1 19.166177 0.14302932 2568.30103
## 2 17.708804 1.00566076 311.83649
## 3  2.121322 0.07422486  60.62666
## 4  3.364076 0.22752641  49.73930
## 5  2.423559 0.05172498 113.55514

```

The broad span of these intervals indicates a significant level of uncertainty in these forecasts. Despite this, the predictions themselves seem to align closely with the actual data, even in light of the extensive intervals.

Do the confidence interval

```
confint(A)
```

```

##           2.5 %           97.5 %
## (Intercept) -2.343345292 10.450117593
## Pol_Stab    -0.074376008  0.006845905
## GDP_Grow     0.002796627  0.234478032
## Urban_Pop    0.014831064  0.072086737
## LGDP         -1.480932084  0.057688945
## LCO2         -1.409592609 -0.323937034

```

Intercept: The substantial interval from -2.34 to 10.45 reflects a high degree of uncertainty in predicting LHIV when all independent variables are held constant. This uncertainty could stem from data variability or limited observations.

Pol_Stab: The interval ranging slightly from -0.074 to 0.0068 envelops zero, suggesting inconclusive evidence regarding its impact on LHIV. This indicates that any effect political stability may have on the log of HIV prevalence is minor and not conclusively proven by statistics.

GDP_Grow: The interval indicating a positive relationship between 0.0028 and 0.2345 points to a significant correlation with LHIV, hinting that an increase in GDP growth could potentially elevate HIV prevalence. However, this interpretation needs to be contextualized within wider economic and health scenarios.

Urban_Pop: The interval lying between 0.0148 and 0.0721, not crossing zero, signifies a statistically significant positive impact, suggesting that a greater urban population percentage may lead to higher HIV prevalence rates.

LGDP: The broad interval from -1.481 to 0.0577, which includes zero, presents an ambiguous effect of LGDP on LHIV, indicating a lack of a definitive statistical relationship or varying impacts across different settings.

LCO2: The interval ranging from -1.41 to -0.324, excluding zero, reveals a statistically significant negative correlation with LHIV. This might suggest that increased CO2 emissions, perhaps indicative of greater industrial activity, are associated with reduced HIV prevalence.

Final Model

summary(A)

```
##
## Call:
## lm(formula = LHIV ~ Pol_Stab + GDP_Grow + Urban_Pop + LGDP +
##      LCO2, data = Data_new2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.05372 -0.60866 -0.05828  0.79922  2.35836
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.05339     3.12763   1.296  0.20520
## Pol_Stab    -0.03377     0.01986  -1.700  0.09975 .
## GDP_Grow     0.11864     0.05664   2.095  0.04506 *
## Urban_Pop     0.04346     0.01400   3.105  0.00423 **
## LGDP        -0.71162     0.37615  -1.892  0.06853 .
## LCO2        -0.86676     0.26541  -3.266  0.00280 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.175 on 29 degrees of freedom
## Multiple R-squared:  0.5025, Adjusted R-squared:  0.4167
## F-statistic: 5.857 on 5 and 29 DF, p-value: 0.000739
```

Residuals: The residuals span a broad range from -2.05 to 2.36, indicating that while the model captures some of the variance in LHIV, substantial prediction errors exist for certain data points.

Coefficients:

Intercept (4.05339): The model estimates a baseline log of HIV prevalence at 4.05339 in the absence of all predictors. However, with a p-value of 0.20520, this estimate lacks statistical significance, casting doubt on the predicted LHIV without the influence of predictors.

Pol_Stab (-0.03377): An increase in Political Stability is linked to a decrease in LHIV. This association nearly reaches significance (p-value = 0.09975), hinting at a potential decrease in HIV prevalence with improved political stability.

GDP_Grow (0.11864): A significant positive correlation with LHIV is observed (p-value = 0.04506), indicating that higher GDP growth rates are associated with an increase in HIV prevalence.

Urban_Pop (0.04346): The significant positive coefficient (p-value = 0.00423) implies that a higher urban population percentage leads to an increase in LHIV.

LGDP (-0.71162): This negative coefficient suggests a possible inverse relationship between GDP and LHIV, nearing statistical significance (p-value = 0.06853).

LCO2 (-0.86676): A significant negative relationship (p-value = 0.00280) indicates that higher CO2 emissions, potentially indicating higher industrial activity, are linked to lower LHIV.

Model Fit:

The model's Multiple R-squared of 0.5025 indicates it explains approximately 50.25% of LHIV's variance, demonstrating a moderate fit.

The Adjusted R-squared of 0.4167, which accounts for the number of predictors and sample size, suggests a slight adjustment to the model's explanatory power.

The F-statistic of 5.857, with 5 and 29 degrees of freedom, along with a p-value of 0.000739, confirms the model's statistical significance. This means the predictors collectively have a meaningful linear relationship with LHIV.

Conclusion

This study examined multiple countries variation in HIV prevalence using socioeconomic, political, and environmental indicators within a regression framework. Using log-transformed HIV prevalence as the primary outcome, the analysis assessed whether structural country-level conditions are associated with differences in HIV prevalence. Overall, the results provide partial support for the proposed hypothesis.

The final regression model indicates that several variables are significantly associated with HIV prevalence. Urban population share shows a positive and statistically significant relationship with HIV prevalence, suggesting that more urbanized countries tend to experience higher prevalence levels. In contrast, CO₂ emissions, used as a proxy for industrialization, display a statistically significant negative association with HIV prevalence, indicating lower prevalence in more industrialized economies. GDP per capita and political stability are negatively associated with HIV prevalence but do not consistently reach conventional levels of statistical significance, while GDP growth is positively associated with HIV prevalence.

Diagnostic tests suggest that the assumptions of linear regression are reasonably satisfied, and results are robust to the use of heteroskedasticity-consistent standard errors. Nevertheless, the analysis is subject to important limitations, including a small sample size, reliance on aggregate country-level data, etc. Despite these constraints, the findings demonstrate how regression-based methods can be used to explore population health outcomes and highlight the complexity of structural influences on HIV prevalence.