Project 2

2025-12-04

Introduction

Life expectancy says a lot about how people live and the conditions they experience around the world. In our first project, we focused on how life expectancy differed across continents, how it changed from 2000 to 2015, and how it related to CO_2 emissions. This project builds on that work by shifting from describing differences in life expectancy to predicting life expectancy using multiple linear regression. Instead of only comparing groups, the goal here is to understand how several environmental, health, and structural factors work together to shape global life expectancy.

The data come from Kaggle's Life Expectancy 2000–2015 dataset (https://www.kaggle.com/datasets/vrec99/life-expectancy-2000-2015), which combines information from the World Health Organization, World Bank, and the United Nations.

Section 1 – Multiple Linear Regression Model for Life Expectancy

In this section, we model life expectancy using multiple predictors in order to better understand how different factors relate to global life expectancy. Rather than analyzing each variable separately, multiple linear regression allows us to examine how these predictors work together to explain variation in life expectancy.

The Hypotheses

This test evaluates whether the regression model provides any predictive value at all:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \cdots = 0$$
$$H_a : \text{At least one } \beta_j \neq 0$$

The response variable is life expectancy, and the predictors include CO emissions, health expenditure, adult obesity, continent, and least-developed status.

The full regression model can be written in matrix form as:

$$Y = X\beta + \varepsilon$$

where Y is the vector of life expectancy values, X is the design matrix containing the predictors, β is the vector of regression coefficients, and ε represents the random error.

Model Construction

```
data= read.csv("Life_Expectancy_00_15.csv", sep = ";")
Y = data$Life.Expectancy
X1 = data$CO2.emissions
X2 = data$Health.expenditure
X3 = data$Obesity.among.adults
C = model.matrix(~ Continent, data)[ , -1]
D  = model.matrix(~ Least.Developed, data)[ , -1]
X = cbind(1, X1, X2, X3, C, D)
```

Here, an intercept column of 1's is included in the design matrix, and dummy variables are created for the categorical predictors (continent and least-developed status) using reference groups.

Checking The Assumptions

We planned to fit this model using multiple linear regression, so we first checked whether the assumptions of linearity, normality, homoscedasticity, and independence were met.

1. Linearity

$H_0 : E(Y|X) = X\beta$ (the relationship between the predictors and the response is linear)

H_a : At least one predictor has a nonlinear relationship with the response

To assess linearity, we examined a residuals versus fitted values plot. If the model is appropriate, the residuals should be randomly scattered around 0 with no clear curved pattern.

Using the least squares estimator,

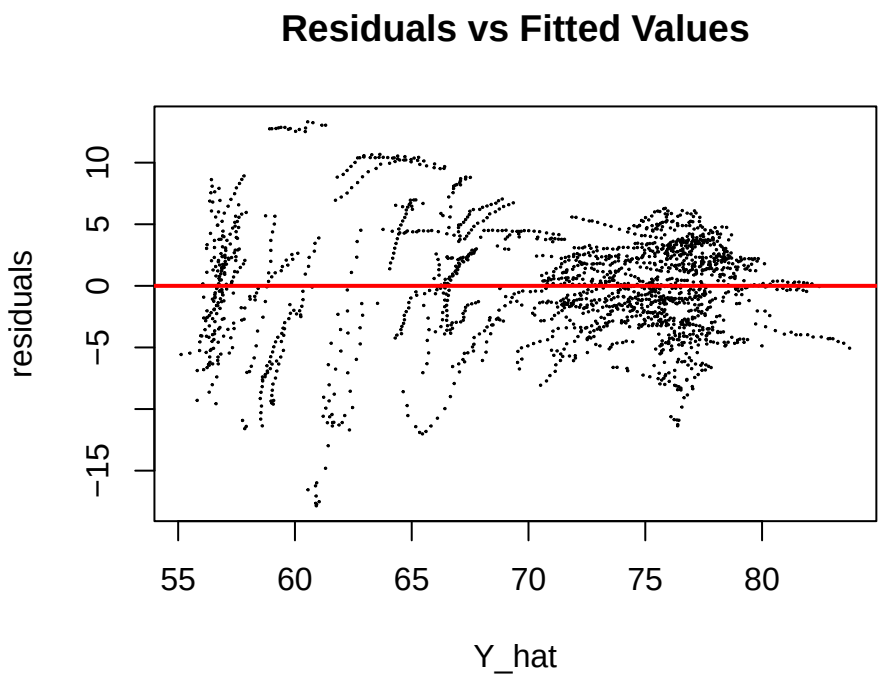
$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

we computed the fitted values and residuals as:

$$\hat{Y} = X\hat{\beta}, \quad \hat{\varepsilon} = Y - \hat{Y}$$

```
beta_hat = solve(t(X) %*% X) %*% (t(X) %*% Y)
Y_hat = X %*% beta_hat
residuals = Y - Y_hat

plot(Y_hat, residuals, cex=0.1, main="Residuals vs Fitted Values")
abline(0, 0, lwd=2, col='red')
```



From the residuals versus fitted values plot, the residuals appear generally centered around zero without a strong global curved pattern. This suggests that the linearity assumption is mostly reasonable for this model.

2. Normality

H_0 : The regression errors are normally distributed.

H_a : The regression errors deviate from normality.

If all four assumptions of multiple linear regression are met, then the standardized residuals should follow a standard normal (Z) distribution. To check this assumption, we computed the standardized residuals using the hat matrix and the estimated error variance.

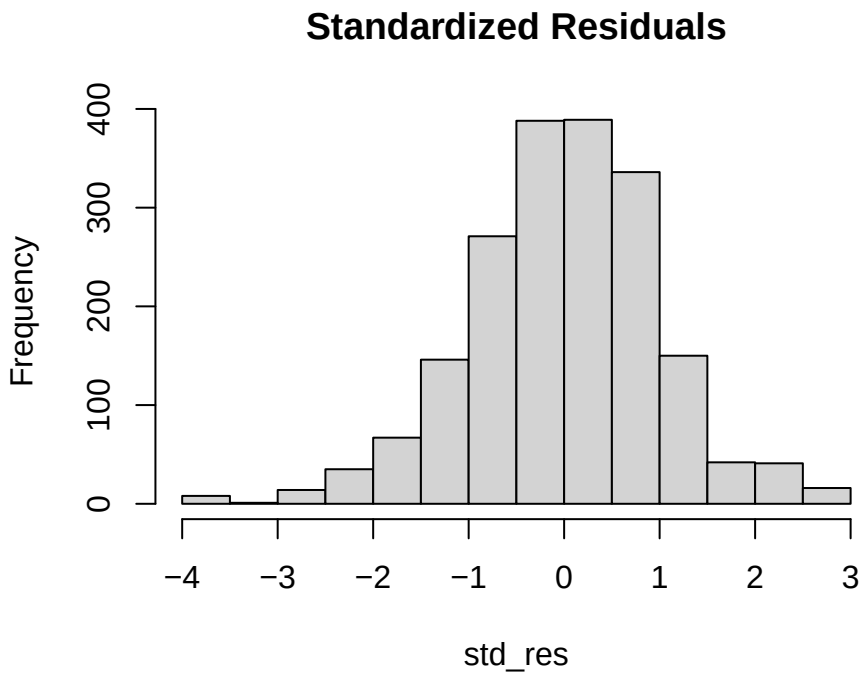
The standardized residuals are defined as:

$$\hat{\varepsilon}_k^S = \frac{\hat{\varepsilon}_k}{\sqrt{\hat{\sigma}^2(1-h_k)}}$$

where $\hat{\sigma}^2$ is the estimated error variance and h_k is the leverage value from the hat matrix.

```
n = length(Y)
p = ncol(X) - 1
SSE = sum(residuals^2)
sigma2_hat = SSE / (n - p - 1)
H = X %*% solve(t(X) %*% X) %*% t(X)
h = diag(H)
std_res = residuals / sqrt(sigma2_hat * (1 - h))

hist(std_res, main = "Standardized Residuals")
```



The histogram of the standardized residuals appears roughly bell-shaped and centered around zero, suggesting that the normality assumption is reasonable at a visual level.

To formally test normality, we applied a Monte Carlo Kolmogorov–Smirnov (KS) test to compare the empirical distribution of the standardized residuals to a standard normal distribution.

```
set.seed(123123)
nmc = 10000
KS_mc = c()

x_sorted = sort(std_res)
nj = length(x_sorted)
F_emp = (1:nj)/(nj+1)
F_null = pnorm(x_sorted, mean(x_sorted), sd(x_sorted))
KS_obs = max(abs(F_emp - F_null))

for(k in 1:nmc){
  smc = sort(rnorm(nj, 0, 1))
  F_emp_mc = (1:nj) / (nj+1)
  F_null_mc = pnorm(smc, 0, 1)
  KS_mc = c(KS_mc, max(abs(F_emp_mc - F_null_mc)))
  }

alpha = 0.10
KS_crit = quantile(KS_mc, 1 - alpha)
emp_pval = mean(KS_mc >= KS_obs)

KS_crit
```

```
##        90%
## 0.02790051
```

```
emp_pval
```

```
## [1] 0.036
```

The Monte Carlo Kolmogorov–Smirnov test produced an empirical p-value of approximately 0.036, which is below the significance level of $\alpha = 0.10$. Therefore, we reject the null hypothesis and conclude that the standardized residuals deviate from perfect normality.

3. Homoscedasticity

$H_0 : \mathrm{Var}(\varepsilon \mid X) = \sigma^2$ (the variance of the errors is constant)

$H_a : \mathrm{Var}(\varepsilon \mid X)$ is not constant

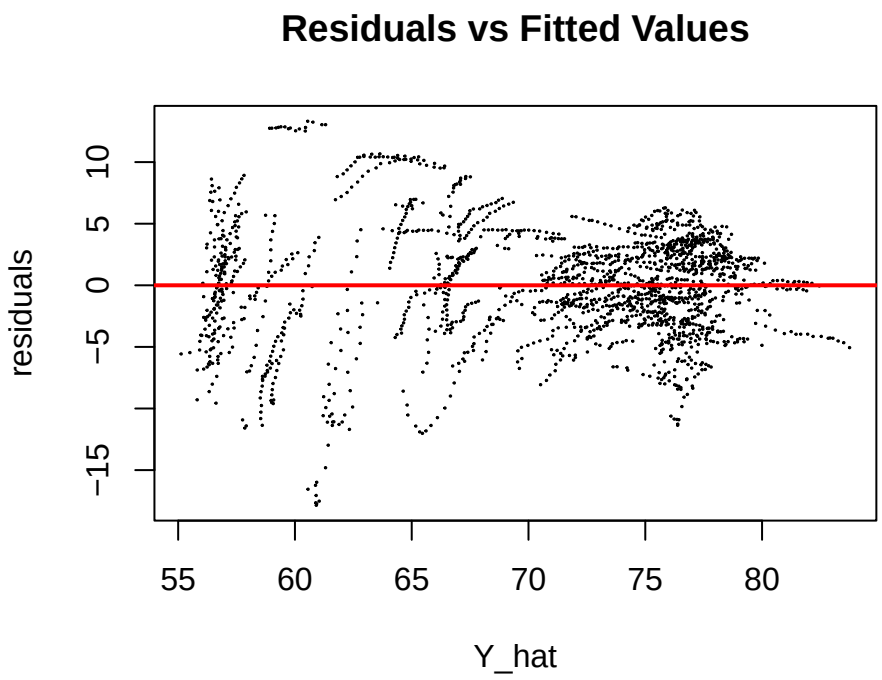
The homoscedasticity assumption states that the variance of the regression errors is constant across all values of the predictors. From the model

$$Y = X\beta + \varepsilon,$$

this assumption means that the spread of the residuals should be approximately the same for all fitted values.

To assess this assumption, we again examined the residuals versus fitted values plot. If the error variance is constant, the residuals should form a horizontal band with roughly the same vertical spread across the range of fitted values.

```r
plot(Y_hat, residuals, cex=0.1, main="Residuals vs Fitted Values")
abline(0, 0, lwd=2, col='red')
```



From this plot, the residuals remain centered around zero, but the vertical spread is not perfectly constant across all fitted values. Specifically, the residuals appear more spread out at lower fitted values and more tightly clustered at higher fitted values. This suggests that the homoscedasticity assumption may be mildly violated, meaning that the error variance is not perfectly constant across the range of the model.

4. Independence

H_0 : The regression errors are independent.

H_a : The regression errors are not independent.

The independence assumption means that one observation's error should not be related to another's. In this data set, each observation represents a country in a given year. Because life expectancy for a given country is naturally related from one year to the next, this creates a time-based structure where observations across years are likely dependent.As a result, the independence assumption is likely violated for this model.

Section 2 – Multicollinearity

Multicollinearity is the strong correlation among the predictors in a regression model. When predictors are highly correlated, the variances of the estimated regression coefficients can become inflated, which makes the estimates unstable.

To assess multicollinearity, we computed the Variance Inflation Factor (VIF) for each predictor. The VIF for predictor X_k is defined as:

$$\text{VIF}_k = \frac{1}{1 - R_k^2}$$

where R_k^2 is the coefficient of determination from regressing X_k on all of the other predictors. The Multicollinearity Index (MCI) is defined as:

$$\text{MCI}_k = \sqrt{\text{VIF}_k}$$

Larger VIF and MCI values indicate more severe multicollinearity.

```
y_CO2 = X[, 2]
X_CO2 = X[, -2]
beta_CO2 = solve(t(X_CO2) %*% X_CO2) %*% (t(X_CO2) %*% y_CO2)
y_hat_CO2 = X_CO2 %*% beta_CO2
SSE_CO2 = sum((y_CO2 - y_hat_CO2)^2)
SST_CO2 = sum((y_CO2 - mean(y_CO2))^2)
R2_CO2 = 1 - SSE_CO2/SST_CO2
VIF_CO2 = 1 / (1 - R2_CO2)
MCI_CO2 = sqrt(VIF_CO2)

y_expend = X[, 3]
X_expend = X[, -3]
beta_expend = solve(t(X_expend) %*% X_expend) %*% (t(X_expend) %*% y_expend)
y_hat_expend = X_expend %*% beta_expend
SSE_expend = sum((y_expend - y_hat_expend)^2)
SST_expend = sum((y_expend - mean(y_expend))^2)
R2_expend = 1 - SSE_expend/SST_expend
VIF_expend = 1 / (1 - R2_expend)
MCI_expend = sqrt(VIF_expend)

y_obesity = X[, 4]
X_obesity = X[, -4]
beta_obesity = solve(t(X_obesity) %*% X_obesity) %*% (t(X_obesity) %*% y_obesity)
y_hat_obesity = X_obesity %*% beta_obesity
SSE_obesity = sum((y_obesity - y_hat_obesity)^2)
SST_obesity = sum((y_obesity - mean(y_obesity))^2)
R2_obesity = 1 - SSE_obesity/SST_obesity
VIF_obesity = 1 / (1 - R2_obesity)
MCI_obesity = sqrt(VIF_obesity)

y_continent = X[, 5]
X_continent = X[, -5]
beta_continent = solve(t(X_continent) %*% X_continent) %*% (t(X_continent) %*% y_continent)
y_hat_continent = X_continent %*% beta_continent
SSE_continent = sum((y_continent - y_hat_continent)^2)
SST_continent = sum((y_continent - mean(y_continent))^2)
R2_continent = 1 - SSE_continent/SST_continent
VIF_continent = 1 / (1 - R2_continent)
MCI_continent = sqrt(VIF_continent)

y_develop = X[, 6]
X_develop = X[, -6]
beta_develop = solve(t(X_develop) %*% X_develop) %*% (t(X_develop) %*% y_develop)
y_hat_develop = X_develop %*% beta_develop
SSE_develop = sum((y_develop - y_hat_develop)^2)
SST_develop = sum((y_develop - mean(y_develop))^2)
R2_develop = 1 - SSE_develop/SST_develop
VIF_develop = 1 / (1 - R2_develop)
```

```
MCI_develop = sqrt(VIF_develop)
```

```
##                Predictor    VIF    MCI
## 1          CO2 Emissions 1.7341 1.3169
## 2     Health Expenditure 1.8450 1.3583
## 3          Adult Obesity 2.6655 1.6326
## 4              Continent 1.8697 1.3674
## 5 Least Developed Status 2.6901 1.6402
```

Overall, all VIF values are well below commonly used cutoffs. This suggests that multicollinearity is not a major concern in this model, and the regression coefficient estimates should be reasonably stable.

Predictor Significance

To determine whether each predictor is significantly related to life expectancy after controlling for the other predictors, we computed partial standardized slopes. These were found using the correlation between:

- the residuals of Y regressed on all other predictors
- and the residuals of X_k regressed on all other predictors

This isolates the unique contribution of each predictor to life expectancy after adjusting for the others.

For each predictor X_k, the standardized slope is:

$$\hat{\beta}_k^* = \text{Cor}(e_{X_k}, e_Y)$$

```
X_others_CO2 = cbind(X[,1], X[,3], X[,4], X[,5], X[,6])
beta_Y_CO2 =
  solve(t(X_others_CO2) %*% X_others_CO2) %*% t(X_others_CO2) %*% Y
beta_X1_CO2 =
  solve(t(X_others_CO2) %*% X_others_CO2) %*% t(X_others_CO2) %*% X[,2]
e_Y_CO2 = Y - X_others_CO2 %*% beta_Y_CO2
e_X1_CO2 = X[,2] - X_others_CO2 %*% beta_X1_CO2
slope_CO2 = cor(e_X1_CO2, e_Y_CO2)

X_others_expend = cbind(X[,1], X[,2], X[,4], X[,5], X[,6])
beta_Y_expend =
  solve(t(X_others_expend) %*% X_others_expend) %*% t(X_others_expend) %*% Y
beta_X2_expend =
  solve(t(X_others_expend) %*% X_others_expend) %*% t(X_others_expend) %*% X[,3]
e_Y_expend = Y - X_others_expend %*% beta_Y_expend
e_X2_expend = X[,3] - X_others_expend %*% beta_X2_expend
slope_expend = cor(e_X2_expend, e_Y_expend)

X_others_obesity = cbind(X[,1], X[,2], X[,3], X[,5], X[,6])
beta_Y_obesity =
  solve(t(X_others_obesity) %*% X_others_obesity) %*% t(X_others_obesity) %*% Y
beta_X3_obesity =
  solve(t(X_others_obesity) %*% X_others_obesity) %*% t(X_others_obesity) %*% X[,4]
e_Y_obesity = Y - X_others_obesity %*% beta_Y_obesity
e_X3_obesity = X[,4] - X_others_obesity %*% beta_X3_obesity
slope_obesity = cor(e_X3_obesity, e_Y_obesity)
```

```r
X_others_continent = cbind(X[,1], X[,2], X[,3], X[,4], X[,6])
beta_Y_continent =
  solve(t(X_others_continent) %*% X_others_continent) %*% t(X_others_continent) %*% Y
beta_C_continent =
  solve(t(X_others_continent) %*% X_others_continent) %*% t(X_others_continent) %*% X[,5]
e_Y_continent = Y - X_others_continent %*% beta_Y_continent
e_C_continent = X[,5] - X_others_continent %*% beta_C_continent
slope_continent = cor(e_C_continent, e_Y_continent)

X_others_develop = cbind(X[,1], X[,2], X[,3], X[,4], X[,5])
beta_Y_develop =
  solve(t(X_others_develop) %*% X_others_develop) %*% t(X_others_develop) %*% Y
beta_D_develop =
  solve(t(X_others_develop) %*% X_others_develop) %*% t(X_others_develop) %*% X[,6]
e_Y_develop = Y - X_others_develop %*% beta_Y_develop
e_D_develop = X[,6] - X_others_develop %*% beta_D_develop
slope_develop = cor(e_D_develop, e_Y_develop)
```

```
##                Predictor Standardized_Slope
## 1          CO2 Emissions             0.0923
## 2     Health Expenditure             0.2403
## 3          Adult Obesity             0.4770
## 4              Continent             0.3477
## 5 Least Developed Status             0.3423
```

Among all predictors, adult obesity has the strongest unique relationship with life expectancy. The positive standardized slope of 0.4770 suggests that, when controlling for the other variables, countries with lower obesity rates tend to have noticeably higher life expectancy. Continent and least developed status also show moderately strong unique effects, with standardized slopes of 0.3477 and 0.3423. This indicates that even after accounting for environmental and health-related variables, regional and developmental differences still play an important role in explaining life expectancy. Health expenditure has a weaker but still meaningful positive relationship with life expectancy (0.2403), suggesting that increased healthcare spending is associated with longer life expectancy, though its effect is not as strong as obesity or structural factors. Finally, CO_2 emissions have the smallest standardized slope (0.0923), indicating that once the other predictors are controlled for, CO_2 emissions have only a weak unique relationship with life expectancy in this model.

Overall, these results suggest that health-related risk (obesity) and structural factors (continent and development status) are the strongest unique predictors of life expectancy, while environmental impact through CO_2 and healthcare spending play a smaller role when considered alongside the other variables.