

Life Expectancy Trends (2000-2015)

Marissa Alfieri

2025-11-11

Introduction

Life expectancy says a lot about how people live and the conditions they face around the world. In this project, I wanted to look at how life expectancy differs across continents and how it changes over time. In order to establish a baseline, I started by checking whether continents already had different average life expectancies in the year 2000. Then I tested if the average life expectancies changed from the years 2000 to 2015. Lastly, I wanted to see if life expectancy correlates to CO2 emissions.

The main goal is to understand whether there are significant differences in life expectancy across regions and time, which could reflect global inequality.

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.

```
data = read.csv("Life_Expectancy.csv")
```

Section 1 – Comparing Continents in 2000

I wanted to start by establishing a baseline in life expectancy across continents at the start of the 21st century.

The hypotheses were:

H_0 : The average life expectancies are the same across all continents.

H_a : At least one continent differs.

Checking The Assumptions

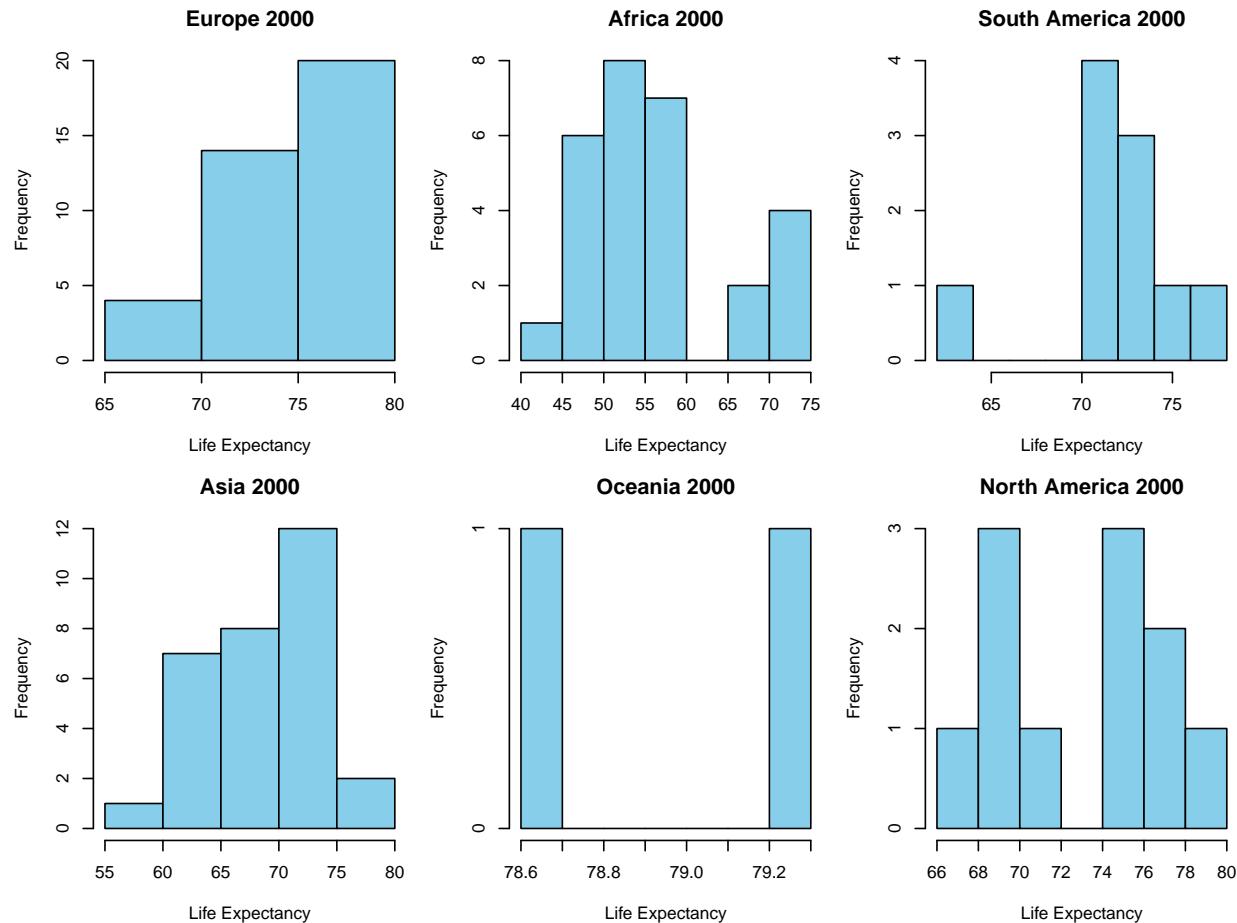
I planned to test this using a one-way ANOVA, so I first checked whether the assumptions of normality and homoscedasticity were met.

1. Normality

H_0 : The life expectancy values within each continent are normally distributed.

H_a : The life expectancy values within at least one continent deviate from normality.

For normality, I visually inspected histograms of life expectancy for each continent.



Visual inspection of the histograms suggested that the distributions were varied, with some having approximately bell shaped curves, some with unique distributions, and some skewed. To confirm normality statistically, I used a Monte Carlo Anderson–Darling test.

This test measures how far the observed data deviates from what would be expected under a normal distribution with the same mean and standard deviation. Comparing the observed statistic to this simulated one let me estimate the p-value and decide whether the normality assumption held.

```
set.seed(123123)
nmc = 10000
n = length(x)
mc_ad = c()

x_sorted = sort(x)
F_emp = (1:n) / (n+1)
F_null = pnorm(x_sorted, mean(x_sorted), sd(x_sorted))
AD_obs = sum(((abs(F_emp - F_null))2) / (F_emp * (1 - F_emp)))

for(k in 1:nmc){
  smc = sort(rnorm(n, 0, 1))
  F_emp_mc = (1:n)/(n+1)
  F_null_mc = pnorm(smc, 0, 1)
  mc_ad = c(mc_ad, sum((F_emp_mc - F_null_mc)2 / (F_emp_mc * (1 - F_emp_mc))))
```

```

}

alpha = 0.10
ad_crit = quantile(mc_ad, 1 - alpha)
emp_pval = mean(mc_ad >= AD_obs)

## Anderson-Darling Critical Value: 1.929625
## Empirical P-Value: 4e-04

```

The Anderson–Darling test gave an empirical p-value of 0.0004, which is below the alpha level of 0.10. This means I reject the null hypothesis of normality. So, the life expectancy data for 2000 aren't normally distributed. This makes sense because some continents have much higher averages while others are lower, so the data are naturally skewed instead of bell-shaped. Before deciding how to move forward, I also wanted to check whether the variability between groups was similar, because large differences in spread can also affect ANOVA results.

2. Testing Homoscedasticity Equal variances (homoscedasticity) matters because ANOVA assumes that all groups have about the same level of variability. If one continent's life expectancy values are way more spread out than another's, the results can be misleading.

$$H_0 : \sigma_{Africa}^2 = \sigma_{Europe}^2 = \sigma_{Asia}^2 = \sigma_{North America}^2 = \sigma_{South America}^2 = \sigma_{Oceania}^2$$

$$H_a : \text{At least one group's variance differs.}$$

I used the Bartlett statistic, which tests whether the variances of life expectancy across groups (j) are equal. It compares each group's sample variance (s_j^2) to the pooled variance (s_p^2) that would be expected if all groups had the same variability. The Bartlett statistic (B_{stat}) is computed as:

$$B_{\text{stat}} = \frac{v * \ln(s_p^2) - (\sum_{j=1}^g v_j * \ln(s_j^2))}{1 + \frac{1}{3(g-1)} * [(\sum_{j=1}^g \frac{1}{v_j}) - \frac{1}{v}]}$$

First, I split the dataset by continent so that I could calculate each group's mean, variance, and sample size:

```

xs = split(x, data_2000$Continent)
xbarj = sapply(xs, mean)
s2j = sapply(xs, var)
nj = sapply(xs, length)

```

Each continent's variance (s_j^2) represents how spread out its life expectancy values are. Then I calculated each group's degrees of freedom ($v_j = n_j - 1$) and counted the number of groups (g):

```

vj = nj - 1
g = length(unique(data_2000$Continent))

```

Next, I computed the pooled variance, which combines the group variances into a single weighted value based on their degrees of freedom:

$$s_p^2 = \frac{\sum_{j=1}^g v_j s_j^2}{\sum_{j=1}^g v_j}$$

```
sp2 = sum(vj * s2j) / sum(vj)
```

After defining the components for each continent (sample variance s_j^2 , degrees of freedom v_j , and the number of groups g), I computed the Bartlett statistic:

```
bastat = ((sum(vj) * log(sp2)) - sum(vj * log(s2j))) /
  (1 + 1/(3*(g-1)) * (sum(1/vj) - 1/sum(vj)))

pval = pchisq(bastat, g-1, lower.tail= F)

## Bartlett's Test Statistic: 26.5106
## Bartlett's Test P-Value: 7.102463e-05
```

The Bartlett's test returned a test statistic of 26.51 and a p-value of 0.0001, which is below the alpha level of 0.10. This means I reject the null hypothesis and conclude that the variances of life expectancy are not equal across continents.

Welch's One-Factor ANOVA

Because Bartlett's test indicated that the variances across continents were not equal, I used Welch's ANOVA, which adjusts for unequal variances and sample sizes.

$$H_0 : \mu_{\text{Africa}} = \mu_{\text{Asia}} = \mu_{\text{Europe}} = \mu_{\text{North America}} = \mu_{\text{South America}} = \mu_{\text{Oceania}}$$

$$H_a : \text{At least one mean differs.}$$

Welch's F statistic is calculated as:

$$F^* = \frac{\sum_{j=1}^g w_j (\bar{X}_j - \bar{X}_w)^2 / (g - 1)}{1 + \frac{2(g-2)}{g^2-1} \sum_{j=1}^g \frac{(1-w_j / \sum_{k=1}^g w_k)^2}{n_j-1}}$$

where $w_j = \frac{n_j}{s_j^2}$ are the weights for each group.

```
nj = sapply(xs, length)
s2j = sapply(xs, var)
wj = nj/s2j
xbarbar_h = sum(wj*xbarj)/sum(wj)
g = length(unique(data_2000$Continent))
n = length(x)

ssm_h = sum(wj*(xbarj-xbarbar_h)^2)
msm_h = ssm_h/g
v_h = ((sum(nj/s2j))^2) / (sum((nj^2) / (s2j^2 * (nj - 1)))) 

Fstat_h = msm_h/(1+2 * (g-2)/v_h)
p = pf(Fstat_h, g-1, v_h, lower.tail=F)

## F Statistic: 11.8083
## P Value: 0.06788656
```

At an alpha level of 0.10, the p-value (0.0679) is slightly below the cutoff. Therefore, I fail to reject the null hypothesis, meaning that when adjusting for unequal variances, there is no statistically significant difference in mean life expectancy across continents.

However, since the data were not normally distributed, I could not fully rely on the results of Welch's ANOVA.

Kruskal–Wallis Test

To confirm whether the previous conclusion holds, I used the Kruskal–Wallis test, a non-parametric alternative that compares the median ranks instead of the means.

The hypotheses are:

H_0 : The distributions of life expectancy are the same across all continents.

H_a : At least one continent differs.

I used the Kruskal–Wallis statistic, which tests whether the distributions of life expectancy across groups (j) are the same. It ranks all observations together, then compares the average rank within each group (\bar{R}_j) to the overall mean rank (\bar{R}). If the group ranks differ a lot, it suggests that at least one continent's distribution is different. The Kruskal–Wallis statistic (H) is computed as:

$$H = \frac{12}{N(N+1)} \sum_{j=1}^g n_j (\bar{R}_j - \bar{R})^2$$

```
R_all = rank(unlist(xs))
Rj = split(R_all, rep(names(xs), times = sapply(xs, length)))
nj = sapply(Rj, length)
N = sum(nj)
g = length(Rj)
Rbar_j = sapply(Rj, mean)
Rbar = (N + 1) / 2

H_stat = (12 / (N * (N + 1))) * sum(nj * (Rbar_j - Rbar)^2)
pval_kw = pchisq(H_stat, g - 1, lower.tail = FALSE)

## Kruskal-Wallis H Statistic: 65.0293
## P-Value: 1.105162e-12
```

The Kruskal–Wallis test returned an H statistic of 65.03 with a p-value of 1.10×10^{-12} , which is below the alpha level of 0.10. This means I reject the null hypothesis and conclude that life expectancy distributions differ significantly across continents in the year 2000.

This confirmed that regional inequality in life expectancy was already present at the start of the century.

Section 2 – Comparing Continents from 2000-2015

After finding that life expectancy differed across continents in 2000, I wanted to see whether those differences changed over time from 2000 to 2015.

The hypotheses were:

$H_0 : \mu_{2000} = \mu_{2001} = \mu_{2002} = \dots = \mu_{2015}$

H_a : At least one mean differs.

Checking The Assumptions

I planned to test this using a repeated-measures ANOVA, so I first checked whether the assumptions of normality and sphericity were met.

1. Normality

H_0 : The life expectancy values across years are normally distributed.

H_a : The life expectancy values deviate from normality.

To test this, I used the Kolmogorov–Smirnov test, which compares the sample's cumulative distribution to a normal distribution with the same mean and standard deviation.

The KS statistic (D) is computed as:

$$D = \max |F_{\text{emp}}(x) - F_{\text{null}}(x)|$$

In R, it was calculated as:

```
x = data$Life.Expectancy
x_sorted = sort(x)
set.seed(123123)
nmc = 10000
n = length(x_sorted)

F_emp = (1:n) / (n+1)
F_null = pnorm(x_sorted, mean(x_sorted), sd(x_sorted))
KS_obs = max(abs(F_emp - F_null))

KS_mc = c()
for(k in 1:nmc){
  smc = sort(rnorm(n, 0, 1))
  F_emp_mc = (1:n) / (n+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)

## Kolmogorov-Smirnov Critical Value: 0.02790051
## Empirical P-Value: 0
```

The Kolmogorov–Smirnov test gave a critical value of 0.0279 and an empirical p-value of 0, which is below the alpha level of 0.10. This means I reject the null hypothesis of normality. Before deciding how to move forward, I also wanted to check the assumption of sphericity, which deals with the relationships between repeated measurements over time.

2. Sphericity Sphericity assumes that the variances of the differences between every pair of years are roughly equal.

H_0 : The variances of the differences between all pairs of years are equal.

H_a : At least one pair of years differs in variance of differences.

I computed Mauchly's test to evaluate the sphericity assumption. This test measures whether the covariance matrix of repeated measures is close to spherical.

The Mauchly statistic (W) is calculated as:

$$W = \frac{|\Sigma|}{\left(\frac{\text{tr}(\Sigma)}{k}\right)^k}$$

where $|\Sigma|$ is the determinant of the covariance matrix and $\text{tr}(\Sigma)$ is its trace.

The test statistic (M_{stat}) is then computed as:

$$M_{\text{stat}} = -(n - 1) \ln(W)$$

```
k = ncol(x)
n = nrow(x)
c = cov(x)
tr = sum(diag(c))
det = det(c)

W = det / ((tr / k)^k)
Mstat = -(n-1) * log(W)
df = (k*(k-1)/2)
Mcrit = qchisq(1-0.10, df)
Mpval = pchisq(Mstat, df, lower.tail = F)
```

```
## Mauchly's Test Statistic: 2119.985
## Mauchly's Test P-Value: 0
```

The Mauchly's test of sphericity returned a test statistic of 2119.99 and a p-value of 0, which is below the alpha level of 0.10. This means I reject the null hypothesis of sphericity. So, the variances of the differences between years are not equal, meaning that the correlations between time points vary.

Non-Spherical Repeated-Measures ANOVA

Since Mauchly's test showed that the data violated sphericity, I used the Greenhouse–Geisser correction to adjust the repeated-measures ANOVA.

$H_0 : \mu_{2000} = \mu_{2001} = \mu_{2002} = \dots = \mu_{2015}$

H_a : At least one mean differs.

The corrected F statistic is calculated as:

$$F = \frac{MS_{\text{Year}}}{MS_{\text{Error(Year)}}}$$

but both degrees of freedom are multiplied by the Greenhouse–Geisser epsilon (ε_{GG}):

$$df_1^* = \varepsilon_{GG}(k - 1), \quad df_2^* = \varepsilon_{GG}(k - 1)(n - 1)$$

```
xbarj = colMeans(x)
xbari = rowMeans(x)
xbarbar = mean(x)
s2all = var(x)

n = nrow(x)
g = ncol(x)
Y = x - rowMeans(x)
c = cov(Y)

SSM = n*sum((xbarj - xbarbar)^2)
E = x - xbari - matrix(xbarj, n, g, byrow=T) + xbarbar
SSE = sum(E^2)
Fstat = (SSM/(g-1))/(SSE/((n-1)*(g-1)))

# Greenhouse-Geisser Correction
eps_GG = ((sum(diag(c)))^2) / ((g - 1) * sum(c^2))
df1_GG = eps_GG * (g - 1)
df2_GG = eps_GG * (g - 1) * (n - 1)

p_GG = pf(Fstat, df1_GG, df2_GG, lower.tail = FALSE)

## F Statistic: 22.95869
## P-Value: 0.004653024
```

After applying the correction, the results showed a significant main effect of year, $p = 0.0047$. This means I reject the null hypothesis and conclude that mean life expectancy differed across at least one year between 2000 and 2015.

Bootstrap Confidence Interval

Since the Kolmogorov–Smirnov test showed that the data were not normally distributed, I can't fully rely on the repeated-measures ANOVA results. So, I used a bootstrap confidence interval to check whether mean life expectancy changed between 2000 and 2015.

By resampling the data 10,000 times and recalculating the mean difference each time, the bootstrap method builds an empirical distribution of possible mean changes. This lets me estimate a confidence interval without assuming that the data are normal.

```
set.seed(123123)
data_2000 = subset(data, Year == 2000)$Life.Expectancy
data_2015 = subset(data, Year == 2015)$Life.Expectancy
obs_diff = mean(data_2015) - mean(data_2000)

nboot = 10000
boot_diffs = c()
for (i in 1:nboot) {
  s1 = sample(data_2000, replace = TRUE)
```

```

    s2 = sample(data_2015, replace = TRUE)
    boot_diffs[i] = mean(s2) - mean(s1)
}

alpha = 0.10
ci_lower = quantile(boot_diffs, alpha / 2)
ci_upper = quantile(boot_diffs, 1 - alpha / 2)

## Bootstrap CI: 3.157008 to 6.605049

```

The bootstrap interval ranged from 3.16 to 6.61 years, which excludes 0, confirming that mean life expectancy differed between 2000 and 2015.

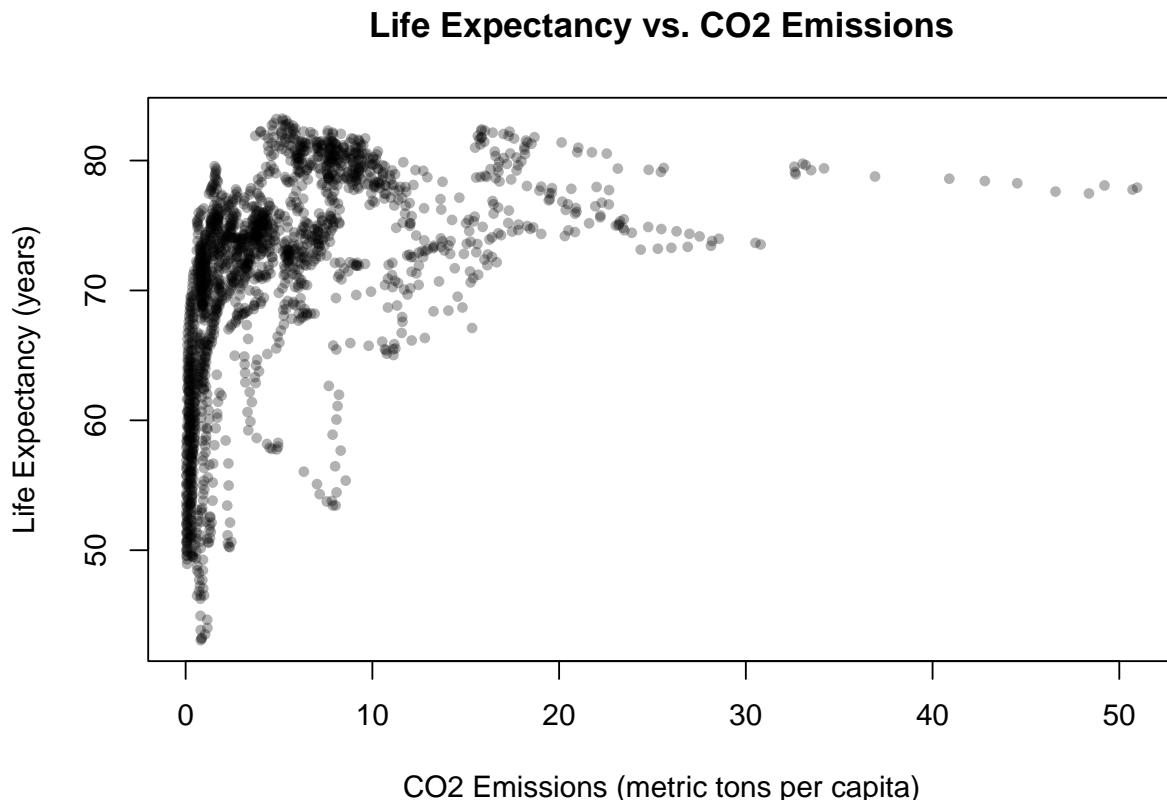
Section 3 – Life Expectancy and CO2 Emissions

After examining differences across continents and changes over time, I wanted to understand what factors might influence life expectancy. CO₂ emissions can serve as an indicator of industrial development, which typically brings better infrastructure and healthcare. I used bootstrap regression to estimate the relationship between CO₂ emissions and life expectancy.

The hypotheses were:

$$H_0 : \beta_1 = 0 \text{ (no relationship between CO}_2 \text{ and life expectancy)}$$

$$H_a : \beta_1 \neq 0 \text{ (a relationship exists)}$$



Bootstrap Confidence Intervals

I used bootstrap resampling to construct confidence intervals for the regression coefficients. This non-parametric method does not assume normality of residuals.

For a linear regression model:

$$\hat{y} = \beta_0 + \beta_1 x$$
$$\beta_1 = r_{xy} \cdot \frac{s_y}{s_x} \quad \text{and} \quad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

```
set.seed(123123)
nboot = 10000

X = data$CO2.emissions
Y = data$Life.Expectancy
n = length(X)

mybhat1 = c()
mybhat0 = c()

myindex = seq(1, n, 1)

for(k in 1:nboot){
  sboot = sample(myindex, n, replace = T)
  Xboot = X[sboot]
  Yboot = Y[sboot]

  bhat1 = cor(Xboot, Yboot) * sd(Yboot) / sd(Xboot)
  bhat0 = mean(Yboot) - bhat1 * mean(Xboot)

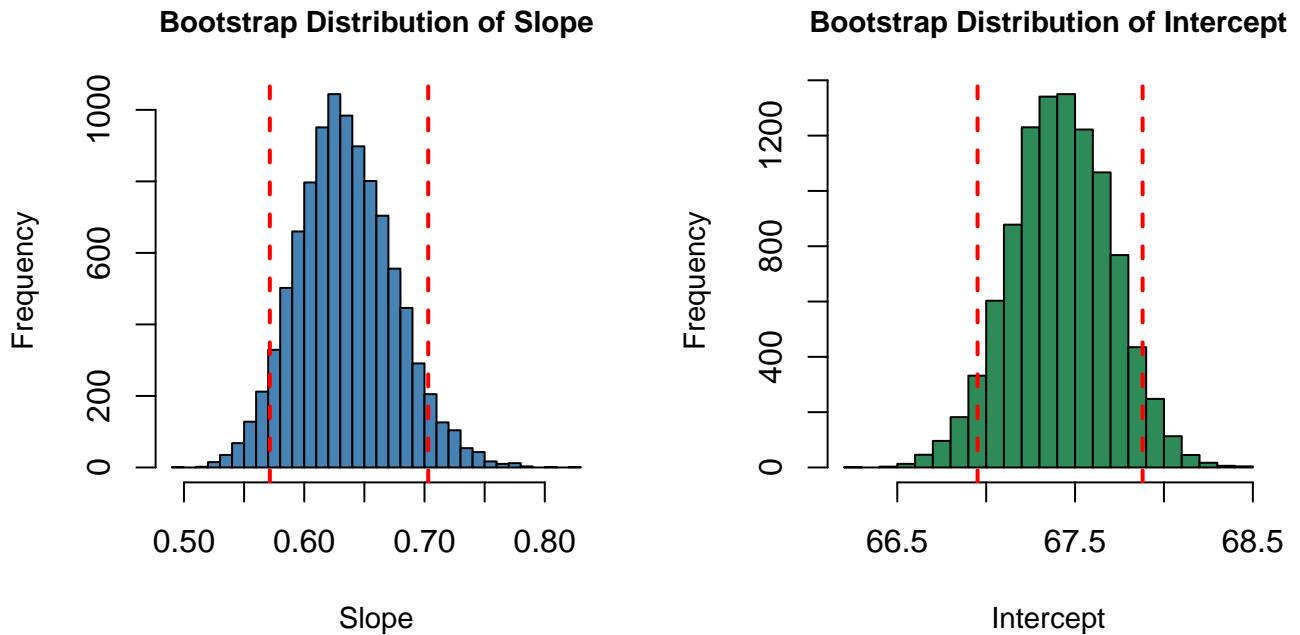
  mybhat1 = c(mybhat1, bhat1)
  mybhat0 = c(mybhat0, bhat0)
}

alpha = 0.10
cibhat1 = quantile(mybhat1, c(alpha/2, 1-alpha/2))
cibhat0 = quantile(mybhat0, c(alpha/2, 1-alpha/2))

## Confidence Interval for Slope:
## [ 0.5713824 , 0.7030796 ]

## Confidence Interval for Intercept:
## [ 66.95175 , 67.8799 ]
```

Bootstrap Distributions



The histograms show the distributions of 10,000 bootstrap estimates. The red dashed lines indicate the 90% confidence interval bounds.

The 90% confidence interval for the slope does not contain zero, indicating a statistically significant positive relationship between CO₂ emissions and life expectancy. Countries with higher CO₂ emissions tend to have longer life expectancies, likely because emissions correlate with industrial development, which brings better healthcare infrastructure and living standards.

Conclusion

Overall, the analysis provides strong evidence that life expectancy varied substantially across continents. The outcomes highlight the ongoing global challenge of health inequality and underscores the importance of continued investment in public health to improving life expectancy. The Life Expectancy (2000–2015) dataset not only illustrates differences across continents but also reflects the broader social and economic realities shaping human life expectancy worldwide. Together, these analyses show that global life expectancy had changed over time from 2000–2015, differs across continents, and is correlated to broader factors.