# GDP per Capita and Life Expectancy: Statistical Analysis

Denys Chepeliuk

08.09.25

## Contents



### Abstract

I study the relationship between GDP per capita and life expectancy at birth across countries, and I examine the time trend for the Czech Republic. In cross-sectional 2020 data ($n = 194$), I find a strong positive association: Pearson $r = 0.624$, $p \approx 2.64 \times 10^{-22}$. A linear model life_expectancy $\sim \log_{10}(\text{GDP per capita})$ explains about $67\%$ of cross-country variation. For the Czech Republic (2000–2020), life expectancy increased $\approx 2.24$ years per decade; a paired $t$-test of year-over-year changes indicates a positive mean increase ($t = 2.996$, $p = 0.0074$).

# 1 Data

- **Life expectancy:** `data/lex.csv` (wide format by year).
- **GDP per capita:** `data/gdp_pcap.csv` (wide format by year; strings such as "24.5k" are parsed to numbers).
- I reshape both datasets to long format (country–year) and merge by `country` and `year`.
- For the 2020 cross-section I filter out invalid/extreme entries:
  - GDP per capita $\leq 0$ or $> 10^6$ removed
  - Life expectancy $\leq 10$ removed

# 2 Methods

- **Pearson correlation (2020):** association between *raw* GDP per capita and life expectancy.

- **Linear regression (2020):**
  Life expectancy $= \beta_0 + \beta_1 \cdot \log_{10}(\text{GDP per capita})$.
- **Time trend (Czech Republic, 2000–2020):**
  Life expectancy $= \alpha_0 + \alpha_1 \cdot \text{Year}$.
- **Paired $t$-test (CZ):** one-sample $t$-test of year-over-year differences against $\mu = 0$.
- Two-sided tests with $\alpha = 0.05$ unless otherwise stated.

# 3 Hypotheses, Results, and Decisions

## 3.1 Cross-sectional correlation (2020)

**Null hypothesis $H_0$:** population correlation $\rho = 0$ (no linear correlation).
**Alternative $H_1$:** $\rho \neq 0$.

- $n = 194$ countries (after filtering)
- Pearson $r = 0.624$
- Test statistic: $t = r\sqrt{\frac{n-2}{1-r^2}} \Rightarrow t \approx 11.06$, df $= 192$
- $p \approx 2.64 \times 10^{-22}$

**Decision:** Reject $H_0$ at $\alpha = 0.05$.
**Interpretation:** Countries with higher GDP per capita tend to have higher life expectancy.

## 3.2 Linear regression on $\log_{10}(\text{GDP})$ (2020)

**Model:** life_expectancy $= \beta_0 + \beta_1 \cdot \log_{10}(\text{GDP per capita})$
**Null hypothesis $H_0$:** $\beta_1 = 0$ (no linear effect of log-GDP).
**Alternative $H_1$:** $\beta_1 \neq 0$.

OLS estimates ($n = 194$): - $\beta_1 = 10.876$, SE $= 0.551 \Rightarrow t \approx 19.77$, $p \ll 0.001$
95% CI for $\beta_1$: $[9.790,\ 11.962]$
- $\beta_0 = 28.226$ (95% CI $[23.809,\ 32.644]$)
- $R^2 = 0.670$ (Adj. $R^2 = 0.669$)

**Decision:** Reject $H_0$ at $\alpha = 0.05$.
**Interpretation:** A $10\times$ increase in GDP per capita is associated with about $+10.9$ **years** higher life expectancy (diminishing returns captured by the log transform).

## 3.3 Time trend for the Czech Republic (2000–2020)

**Model:** life_expectancy $= \alpha_0 + \alpha_1 \cdot \text{Year}$
**Null hypothesis $H_0$:** $\alpha_1 = 0$ (no temporal trend).
**Alternative $H_1$:** $\alpha_1 \neq 0$.

OLS estimates ($n = 21$ years): - **Slope** $\alpha_1 = 0.224$ years per year ($\approx +2.24$ years per decade)
- $R^2 = 0.947$
- (From the OLS summary, the slope is highly significant; equivalent $t/F$ tests give $p \ll 0.001$.)

**Decision:** Reject $H_0$ at $\alpha = 0.05$.
**Interpretation:** Czech life expectancy rose steadily over 2000–2020.

## 3.4 Paired $t$-test of Czech year-over-year changes

Let $\Delta_i$ be the change in life expectancy from year $i$ to $i + 1$, for 2000–2020 (20 differences).

**Null hypothesis** $H_0$**:** $\mu_\Delta = 0$ (no average year-over-year change).
**Alternative** $H_1$**:** $\mu_\Delta \neq 0$.

- Mean $\Delta \approx +0.18$ years/year; SD $\approx 0.269$; $n = 20$
- $t = 2.996$, df $= 19$, $p = 0.0074$

**Decision:** Reject $H_0$ at $\alpha = 0.05$.
**Interpretation:** On average, life expectancy increased from year to year.

# 4 Figures

- `results/scatter_2020.png` — 2020 scatter of life expectancy vs GDP per capita
- `results/regression_2020.png` — regression fit using $\log_{10}(\text{GDP})$
- `results/trend_czech.png` — Czech Republic 2000–2020 trend line

# 5 Assumptions & Diagnostics (brief)

- **Correlation:** assumes approximate linearity and bivariate normality; the cross-section shows a clear positive pattern, with curvature handled in the regression via the log of GDP.
- **OLS:** linearity, independent errors, and roughly homoskedastic, normal residuals for inference. Cross-country data may exhibit heteroskedasticity; robust SEs would be a natural extension.
- **Time series (CZ):** short series; autocorrelation could affect standard errors, but the trend is large.

# 6 Reproducibility

- Python 3.9+, `pip install -r requirements.txt`, then `python src/main.py`.
- Code is modular (`src/`), and plots/results are written to `results/`.

# 7 Limitations

- Cross-sectional associations do not imply causation.
- Data quality varies across countries/years; extreme/invalid entries were filtered.