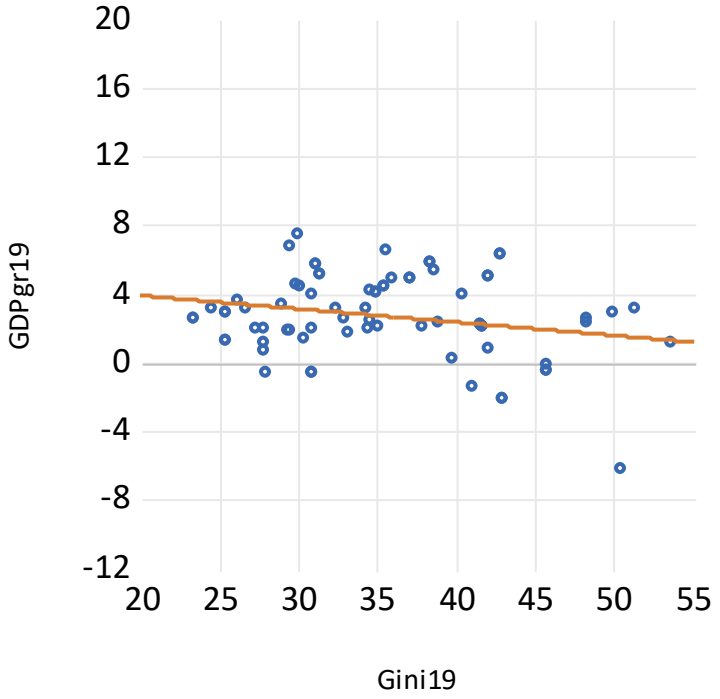
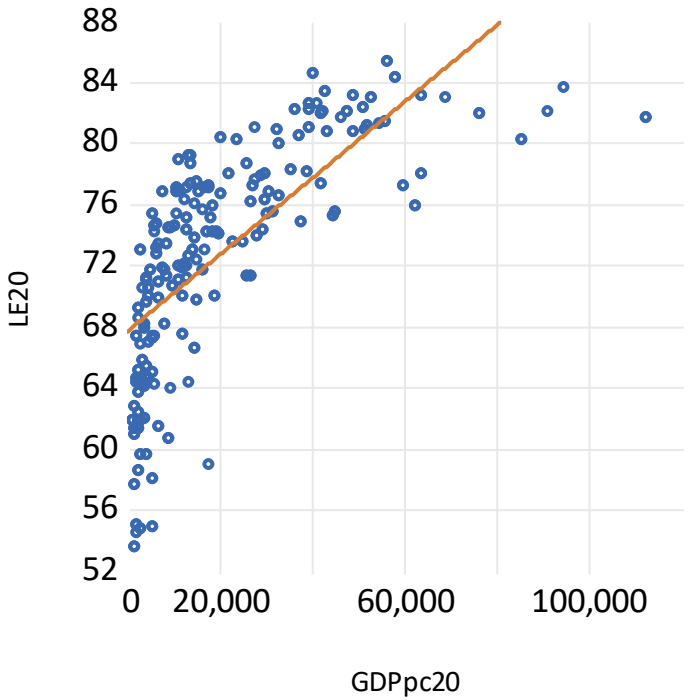
# Linear Relationships Examples: World Development Indicators 2022

**https://databank.worldbank.org/source/world-development-indicators**

The relationships of Gini (2019) and real GDP growth rates (2019), and of Real GDP (PPP) per capita and Life expectancy at birth (2020)



# Terminology for the Linear Regression Model with a Single Regressor

**The linear regression model is :**

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

**where the subscript** $i$ **runs over observations** $i = 1, \ldots, n$;

$Y_i$ **is the *dependent variable,* the *regressand*, of simply the *left-hand variable*;**

$X_i$ **is the *independent variable,* the *regressor*, of simply the *right-hand variable*;**

$\beta_1 + \beta_2 X_i$ **is the *population regression line* or *population regression function* ;**

$\beta_1$ **is the *intercept* of the population regression line;**

$\beta_2$ **is the *slope o*f the population regression line;**

$u_i$ **is the *error term.***

# ASSUMPTIONS FOR MODEL A

**Model A: Cross-sectional data with nonstochastic regressors.**

**A.1 The model is linear in parameters and correctly specified.**

$$Y = \beta_1 + \beta_2 X + u$$

**A.2 There is some variation in the regressor in the sample.**

**A.3 The disturbance term has zero expected value in each observation:**

**for all i (Gauss-Markov 1 condition)** $\mathrm{E}(u_i) = 0$

# ASSUMPTIONS FOR MODEL A
(continued)

**A.4 The disturbance term is homoscedastic**

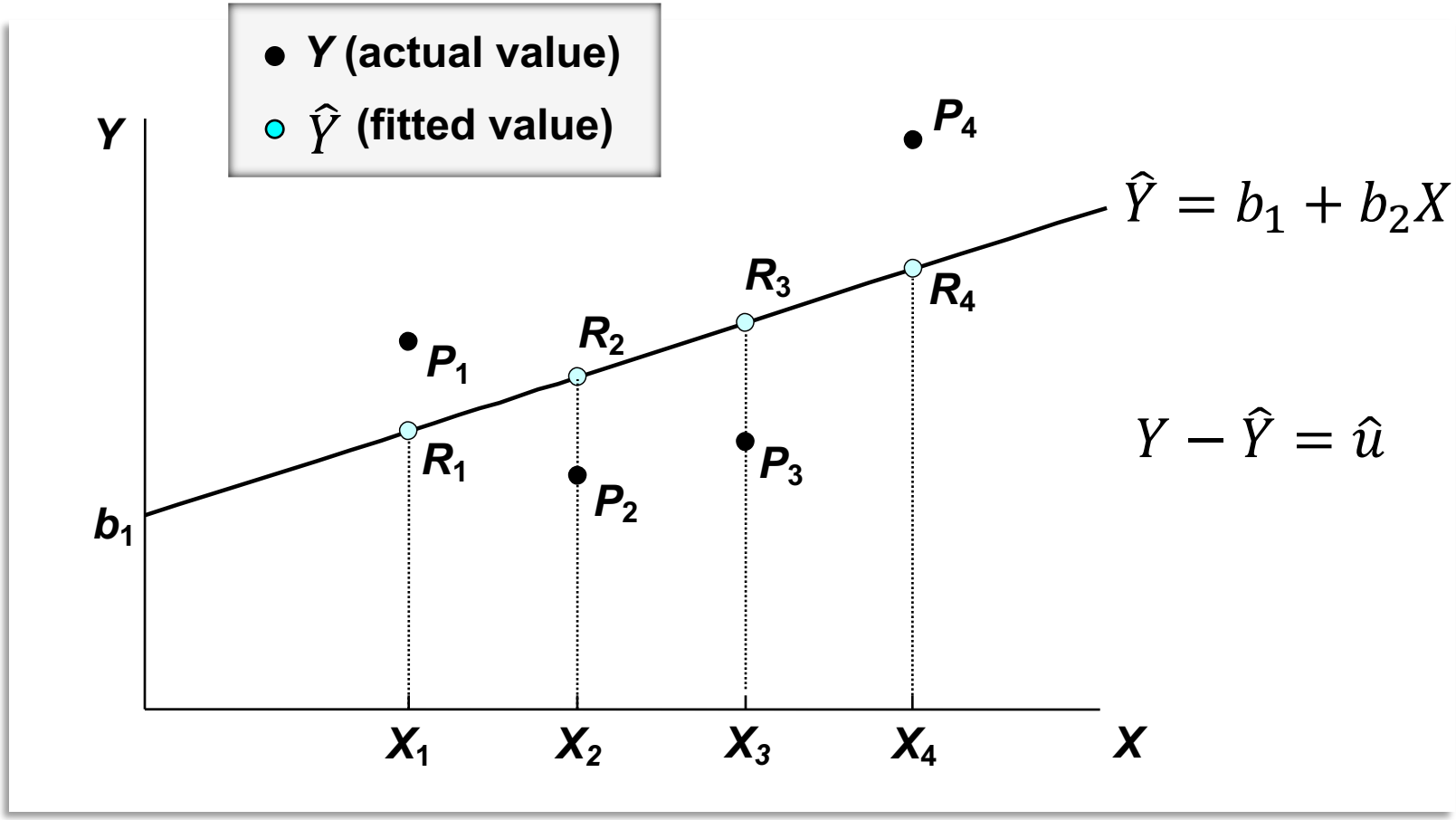
**for all *i*** $\sigma_{u_i}^2 = \sigma_u^2$ **(Gauss-Markov 2 condition)**

**A.5 The values of the disturbance term have independent distributions ($u_i$ and $u_j$ are independent for all $j \neq i$) (Gauss-Markov 3 condition)**

$$\sigma_{u_i u_j} = \mathbf{E}[(u_i - \mu_u)(u_j - \mu_u)] = \mathbf{E}(u_i u_j)$$
$$= \mathbf{E}(u_i)\mathbf{E}(u_j) = 0$$

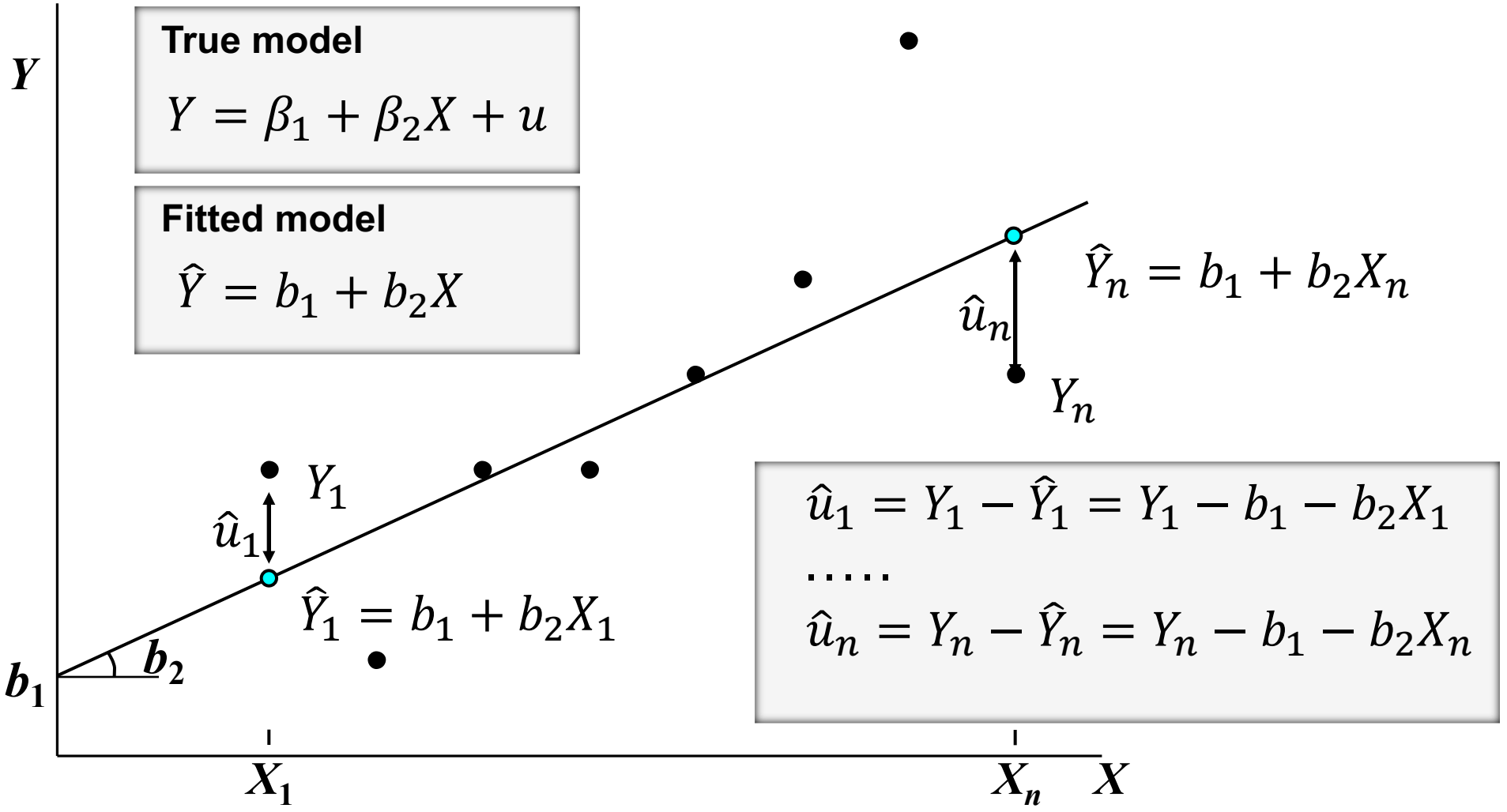
**A.6 The disturbance term has a normal distribution**

# SIMPLE LINEAR REGRESSION MODEL




The line is called the fitted model and the values of *Y* predicted by it are called the fitted values of *Y*. The differences $\hat{u}_i$ are called residuals

# DERIVING LINEAR REGRESSION COEFFICIENTS



# DERIVING LINEAR REGRESSION COEFFICIENTS

$$SSR = \hat{u}_1^2 + \ldots + \hat{u}_n^2 = (Y_1 - b_1 - b_2 X_1)^2 + \ldots + (Y_n - b_1 - b_2 X_n)^2$$

$$= \sum Y_i^2 + n b_1^2 + b_2^2 \sum X_i^2 - 2 b_1 \sum Y_i - 2 b_2 \sum X_i Y_i$$

$$\frac{\partial RSS}{\partial b_1} = 0 \quad \Rightarrow \quad 2 n b_1 - 2 \sum Y_i + 2 b_2 \sum X_i = 0$$

$$b_1 = \bar{Y} - b_2 \bar{X}$$

# DERIVING LINEAR REGRESSION COEFFICIENTS

$$\frac{\partial SSR}{\partial b_2} = 0 \quad \Rightarrow \quad 2b_2 \sum X_i^2 - 2 \sum X_i Y_i + 2b_1 \sum X_i = 0$$

$$b_2 \sum X_i^2 - \sum X_i Y_i + b_1 \sum X_i = 0$$

$$b_2 \sum X_i^2 - \sum X_i Y_i + (\bar{Y} - b_2 \bar{X}) \sum X_i = 0$$

$$b_2 \sum X_i^2 - \sum X_i Y_i * + (\bar{Y} - b_2 \bar{X}) n\bar{X} = 0$$

$$b_2 \sum X_i^2 - \sum X_i Y_i + n\bar{X}\bar{Y} - nb_2 \bar{X}^2 = 0$$

$$b_2 \left( \sum X_i^2 - n\bar{X}^2 \right) = \sum X_i Y_i - n\bar{X}\bar{Y}$$

# DERIVING LINEAR REGRESSION COEFFICIENTS

$$\hat{\beta}_2 \left( \sum X_i^2 - n\bar{X}^2 \right) = \sum X_i Y_i - n\bar{X}\bar{Y}$$

$$\hat{\beta}_2 = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2}$$

$$\hat{\beta}_2 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{\widehat{\mathrm{Cov}}(X, Y)}{\widehat{\mathrm{Var}}(X)}$$

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = \sum X_i Y_i - n\bar{X}\bar{Y}$$

$$\sum (X_i - \bar{X})^2 = \sum X_i^2 - n\bar{X}^2$$

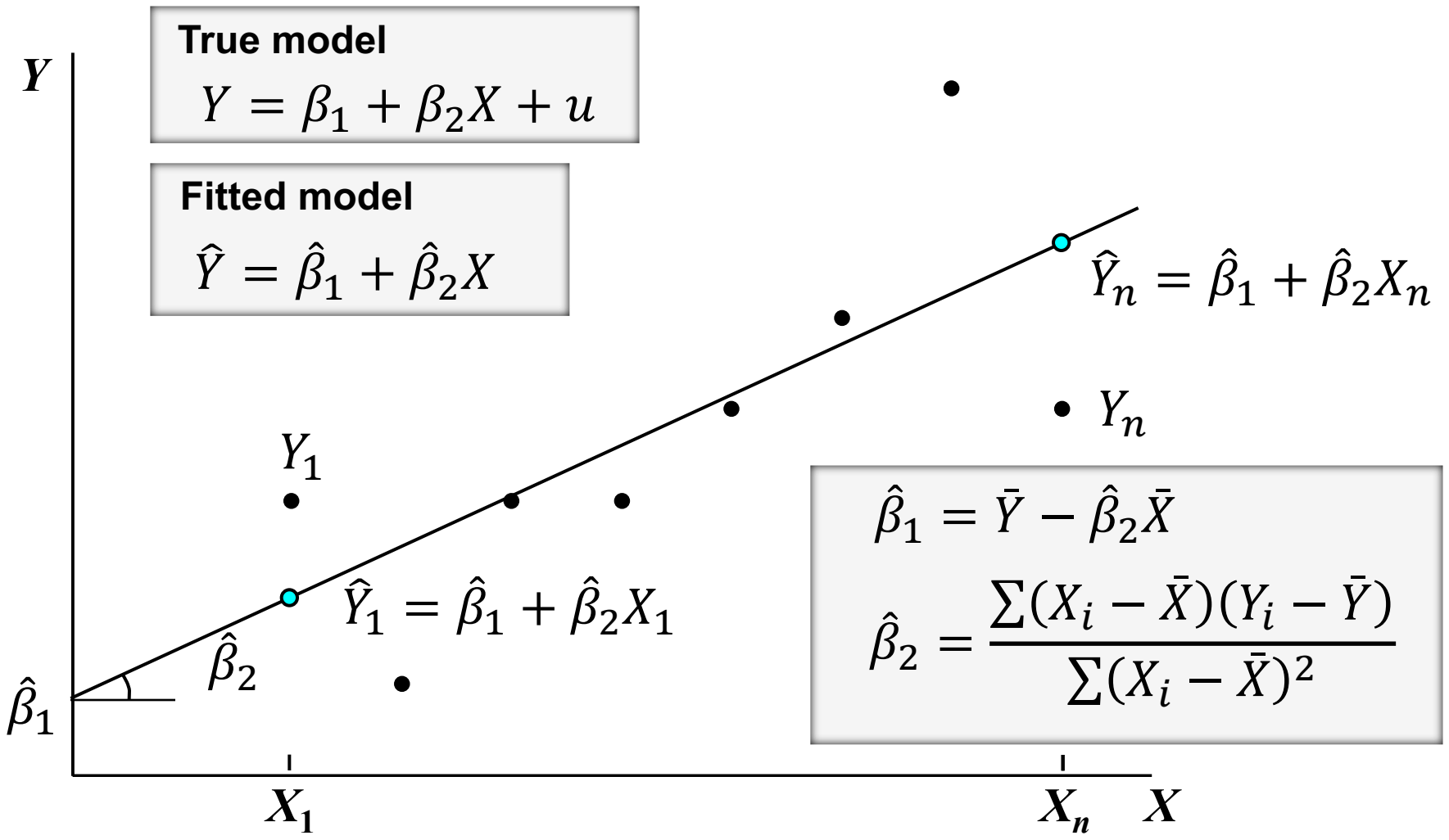
# DERIVED LINEAR REGRESSION COEFFICIENTS

**True model**

$$Y = \beta_1 + \beta_2 X + u$$

**Fitted model**

$$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X$$




$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$$

$$\hat{\beta}_2 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

## LINEAR REGRESSION MODEL WITHOUT INTERCEPT: DERIVING COEFFICIENTS

| True model | Fitted model |
| --- | --- |
| $Y = \beta_2 X + u$ | $\hat{Y} = \hat{\beta}_2 X$ |

$$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - b_2 X_i$$

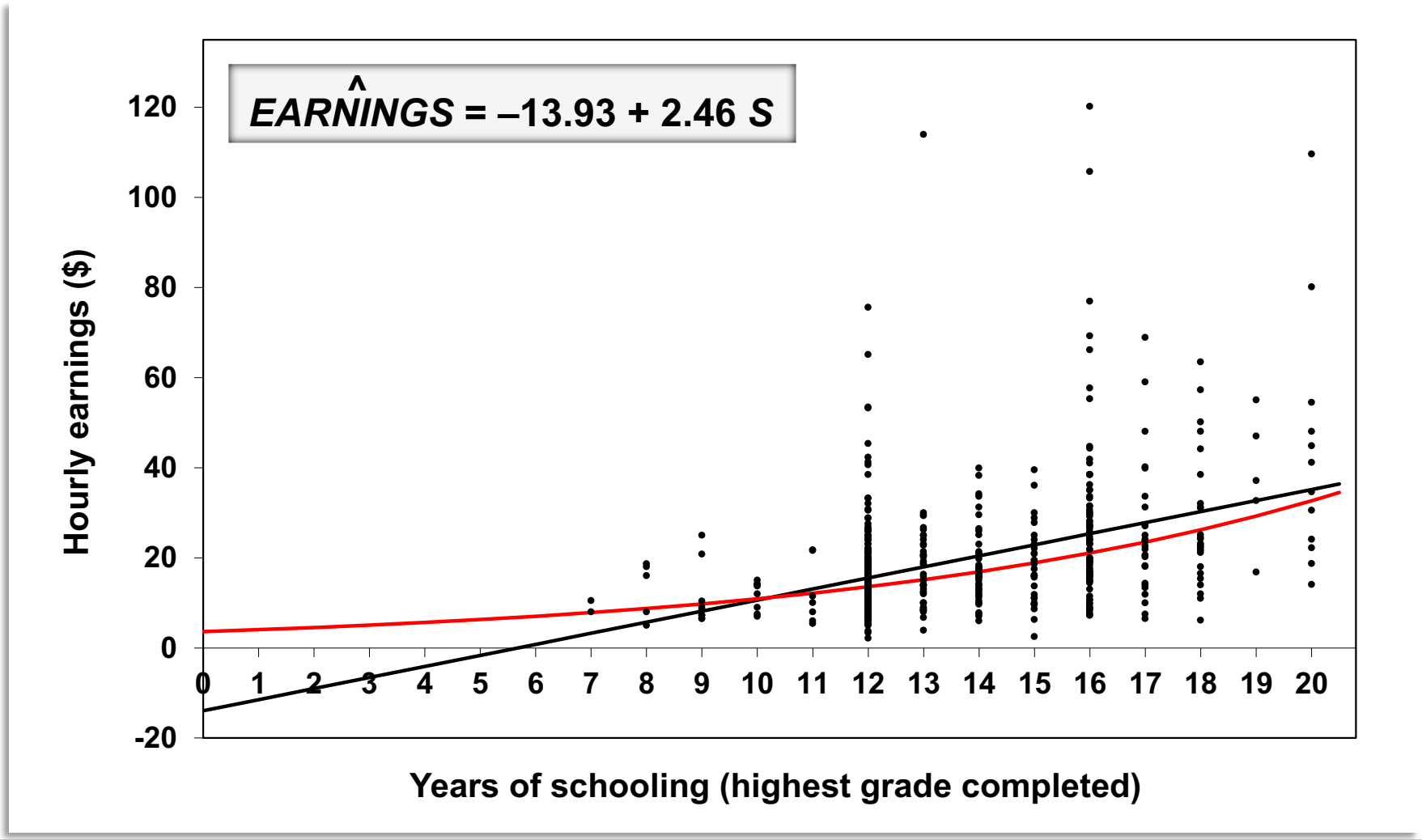
$$SSR = \sum (Y_i - b_2 X_i)^2 = \sum Y_i^2 - 2b_2 \sum X_i Y_i + b_2^2 \sum X_i^2$$

$$\frac{\mathrm{d}SSR}{\mathrm{d}b_2} = 2b_2 \sum X_i^2 - 2 \sum X_i Y_i$$

$$2\hat{\beta}_2 \sum X_i^2 - 2 \sum X_i Y_i = 0 \qquad \hat{\beta}_2 = \frac{\sum X_i Y_i}{\sum X_i^2}$$

$$\frac{\mathrm{d}^2 SSR}{\mathrm{d}b_2^2} = 2 \sum X_i^2 > 0$$

# INTERPRETATION OF A REGRESSION EQUATION




**The slope coefficient implies that hourly earnings increase (on average) by $2.46 for each extra year of schooling.**

**The negative intercept does not make any sense. We limit the interpretation to the range of the sample data. The true relationship may be nonlinear.**

## CHANGES IN THE UNITS OF MEASUREMENT: LINEAR TRANSFORMATION OF Y

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad \hat{\beta}_2 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$
$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

$$Y_i^* = \lambda_1 + \lambda_2 Y_i \qquad \hat{Y}_i^* = \hat{\beta}_1^* + \hat{\beta}_2^* X_i$$

$$\hat{\beta}_2^* = \frac{\sum(X_i - \bar{X})(Y_i^* - \bar{Y}^*)}{\sum(X_i - \bar{X})^2} = \frac{\sum(X_i - \bar{X})([\lambda_1 + \lambda_2 Y_i] - [\lambda_1 + \lambda_2 \bar{Y}])}{\sum(X_i - \bar{X})^2}$$
$$= \frac{\sum(X_i - \bar{X})(\lambda_2 Y_i - \lambda_2 \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{\lambda_2 \sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \lambda_2 \hat{\beta}_2$$

**The effect on the intercept: an exercise.**

**The effect of a change in the units of measurement of *X:* an exercise.**

## CHANGES IN THE UNITS OF MEASUREMENT: X* AS DEVIATION FROM THE MEAN

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

$$\hat{\beta}_2 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

$$X_i^* = X_i - \bar{X}$$

$$\hat{Y}_i = \hat{\beta}_1^* + \hat{\beta}_2^* X_i^*$$

$$\hat{\beta}_2^* = \frac{\sum(X_i^* - \bar{X}^*)(Y_i - \bar{Y})}{\sum\left(X_i^* - \bar{X}^*\right)^2} = \frac{\sum X_i^*(Y_i - \bar{Y})}{\sum {X_i^*}^2}$$

$$= \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \hat{\beta}_2$$

$$\bar{X}^* = 0$$

$$\hat{\beta}_1^* = \bar{Y} - \hat{\beta}_2^* \bar{X}^* = \bar{Y}$$

**The intercept is now the fitted value of *Y* at the sample mean of *X*, this is the sample mean of *Y*.**

# GOODNESS OF FIT

**Four useful results:** 1) $\bar{\hat{u}} = 0$

$$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$$

$$\sum \hat{u}_i = \sum Y_i - n\hat{\beta}_1 - \hat{\beta}_2 \sum X_i$$

$$\frac{1}{n}\sum \hat{u}_i = \frac{1}{n}\sum Y_i - \hat{\beta}_1 - \hat{\beta}_2 \frac{1}{n}\sum X_i$$

$$\begin{aligned}\bar{\hat{u}} &= \bar{Y} - \hat{\beta}_1 - \hat{\beta}_2 \bar{X} \\ &= \bar{Y} - \left(\bar{Y} - \hat{\beta}_2 \bar{X}\right) - \hat{\beta}_2 \bar{X} \\ &= 0\end{aligned}$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$$

**Useful results:** 2) $\bar{\hat{Y}} = \bar{Y}$ 3) $\sum X_i \hat{u}_i = 0$

$$\hat{u}_i = Y_i - \hat{Y}_i \qquad \sum \hat{u}_i = \sum Y_i - \sum \hat{Y}_i$$

$$0 = \frac{1}{n}\sum Y_i - \frac{1}{n}\sum \hat{Y}_i = \bar{Y} - \bar{\hat{Y}} \qquad \bar{\hat{Y}} = \bar{Y}$$

$$\sum X_i \hat{u}_i = \sum X_i \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right)$$

$$= \sum X_i Y_i - \hat{\beta}_1 \sum X_i - \hat{\beta}_2 \sum X_i^2$$

$$\frac{\partial SSR}{\partial b_2} = 0 \quad \Rightarrow \quad 2\hat{\beta}_2 \sum X_i^2 - 2\sum X_i Y_i + 2\hat{\beta}_1 \sum X_i = 0$$

**Useful results:** 4) $\sum \hat{Y}_i \hat{u}_i = 0$

$$\sum \hat{Y}_i \hat{u}_i = \sum (\hat{\beta}_1 + \hat{\beta}_2 X_i)\hat{u}_i$$

$$= \sum \hat{\beta}_1 \hat{u}_i + \sum \hat{\beta}_2 X_i \, \hat{u}_i$$

$$= \hat{\beta}_1 n\bar{\hat{u}} + \hat{\beta}_2 \sum X_i \, \hat{u}_i = 0$$

$\bar{\hat{u}} = 0$ $\sum X_i \hat{u}_i = 0$

## Useful results: *SST=SSE+SSR*

$$\sum (Y_i - \bar{Y})^2 = \sum \left( [\hat{Y}_i + \hat{u}_i] - \bar{Y} \right)^2$$

$$= \sum \left( [\hat{Y}_i - \bar{Y}] + \hat{u}_i \right)^2$$

$$= \sum \left( \hat{Y}_i - \bar{Y} \right)^2 + \sum \hat{u}_i^2 + 2 \sum \left( [\hat{Y}_i - \bar{Y}] \hat{u}_i \right)$$

$$= \sum \left( \hat{Y}_i - \bar{Y} \right)^2 + \sum \hat{u}_i^2 + 2 \sum \hat{Y}_i \hat{u}_i - 2\bar{Y} \sum \hat{u}_i$$

$$\boldsymbol{SST = SSE + SSR}$$

$\sum (Y_i - \bar{Y})^2 = \boldsymbol{SST}$, **sum of squares total**

$\sum \left( \hat{Y}_i - \bar{Y} \right)^2 = \boldsymbol{SSE}$, **sum of squares explained**

$\sum \hat{u}_i^2 = \boldsymbol{SSR}$, **sum of squared residuals**

## Useful results: $R^2$

$$\sum (Y_i - \bar{Y})^2 = \sum \left(\hat{Y}_i - \bar{Y}\right)^2 + \sum \hat{u}_i^2 \qquad SST = SSE + SSR$$

$$R^2 = \frac{SSE}{SST} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}$$

$$R^2 = \frac{SST - SST}{SST} = 1 - \frac{\sum \hat{u}_i{}^2}{\sum (Y_i - \bar{Y})^2}$$

$$r_{Y,\hat{Y}} = \frac{\sum(Y_i - \bar{Y})\left(\hat{Y}_i - \bar{Y}\right)}{\sqrt{\sum(Y_i - \bar{Y})^2 \sum\left(\hat{Y}_i - \bar{Y}\right)^2}} = \frac{\sum\left(\hat{Y}_i - \bar{Y}\right)^2}{\sqrt{\sum(Y_i - \bar{Y})^2 \sum\left(\hat{Y}_i - \bar{Y}\right)^2}}$$

$$= \frac{\sqrt{\sum\left(\hat{Y}_i - \bar{Y}\right)^2}}{\sqrt{\sum(Y_i - \bar{Y})^2}} = \sqrt{\frac{\sum\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum(Y_i - \bar{Y})^2}} = \sqrt{R^2}$$

$$\sum(Y_i - \bar{Y})\left(\hat{Y}_i - \bar{Y}\right) = \sum\left(\left[\hat{Y}_i + \hat{u}_i\right] - \bar{Y}\right)\left(\hat{Y}_i - \bar{Y}\right)$$

$$= \sum\left(\left[\hat{Y}_i - \bar{Y}\right] + \hat{u}_i\right)\left(\hat{Y}_i - \bar{Y}\right)$$

$$= \sum\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum \hat{u}_i \hat{Y}_i$$

$$= \sum\left(\hat{Y}_i - \bar{Y}\right)^2$$

## $R^2$ and Adjusted $R^2$

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

Determination coefficient $R^2$ always grows if an explanatory variable has been added, either significant or not. The adjusted coefficient was introduced which may increase or decrease:

$$R^2{}_{adj} = 1 - \frac{SSR/(n-k)}{SST/(n-1)} =$$

$$= 1 - \frac{SSR/(n-2)}{SST/(n-1)}$$

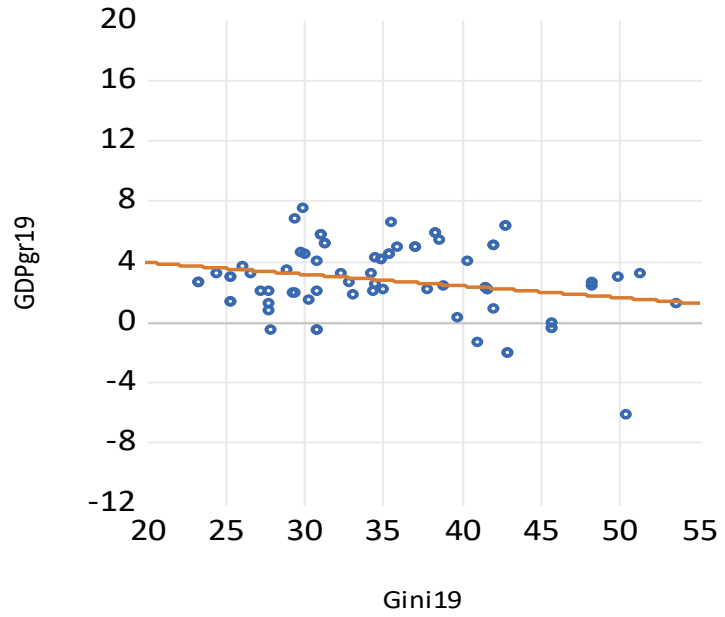
The $R^2_{adj}$ coefficient is not widely used for econometric analysis though available in the regression printouts.

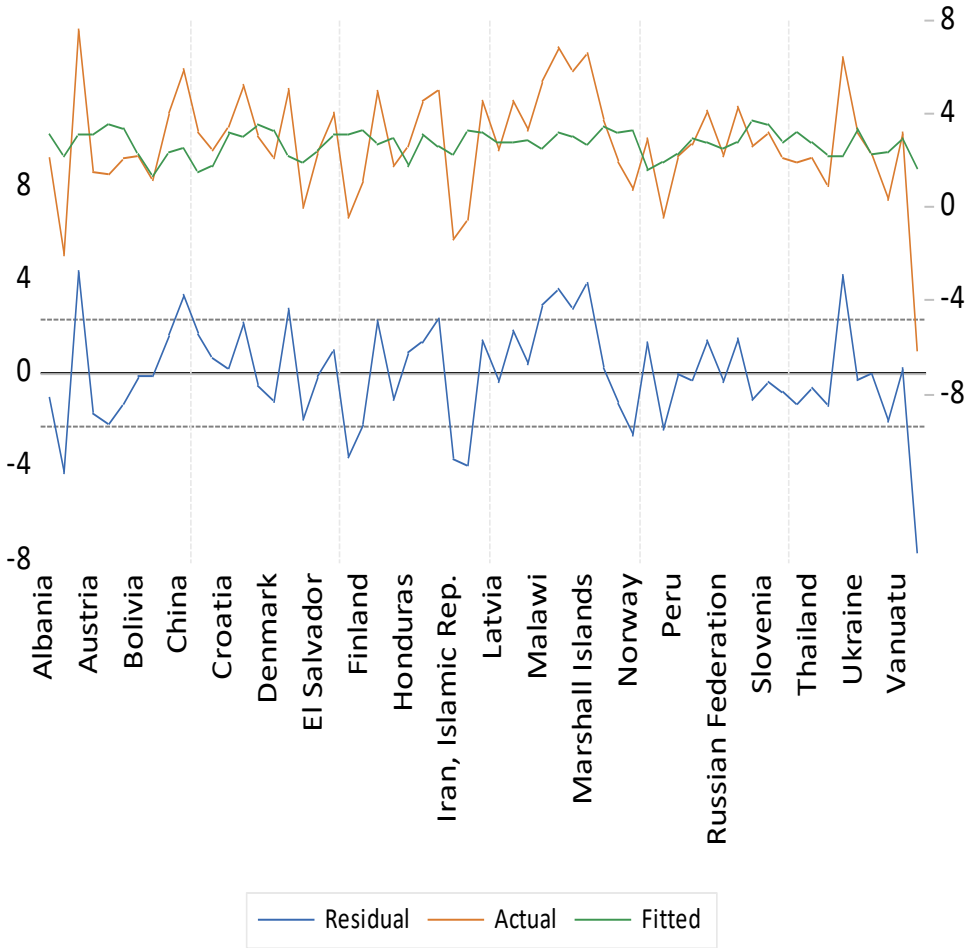
# SLR Model: Gini (2019) and real GDP growth rates (2019), 59 countries

Dependent Variable: GDPGR19
Method: Least Squares
Date: 09/03/22 Time: 16:38
Sample (adjusted): 2 217
Included observations: 59 after adjustments

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 5.475153 | 1.451590 | 3.771832 | 0.0004 |
| GINI19 | -0.077158 | 0.040093 | -1.924476 | 0.0593 |

| | | | |
|---|---|---|---|
| R-squared | 0.061011 | Mean dependent var | 2.742074 |
| Adjusted R-squared | 0.044538 | S.D. dependent var | 2.360560 |
| S.E. of regression | 2.307394 | Akaike info criterion | 4.543425 |
| Sum squared resid | 303.4718 | Schwarz criterion | 4.613850 |
| Log likelihood | -132.0310 | Hannan-Quinn criter. | 4.570916 |
| F-statistic | 3.703609 | Durbin-Watson stat | 2.235485 |
| Prob(F-statistic) | 0.059292 | | |



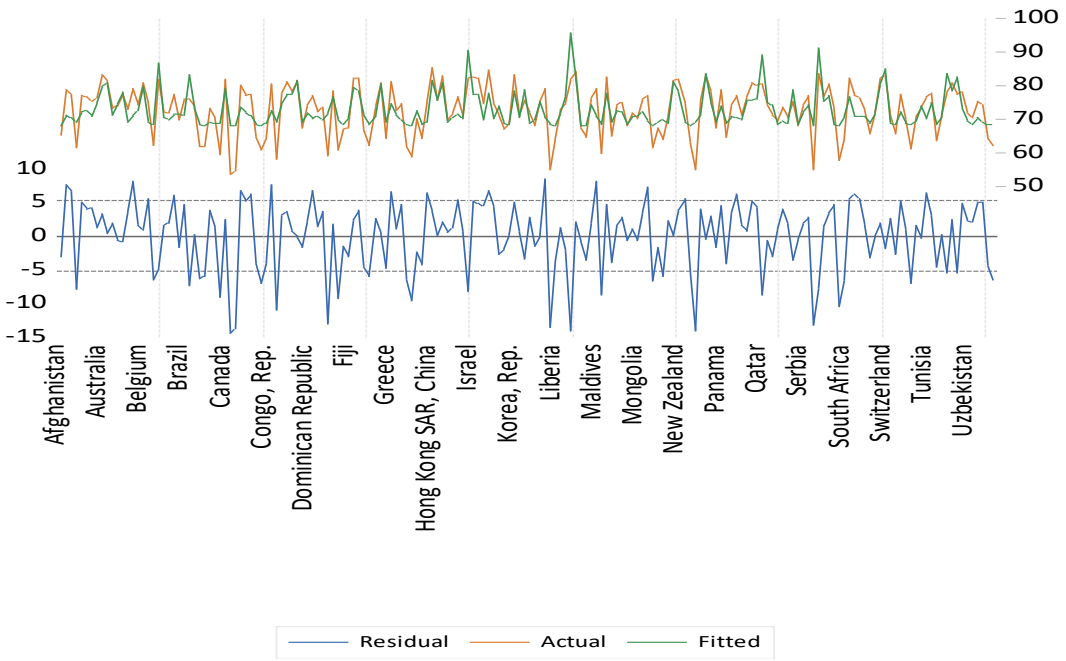



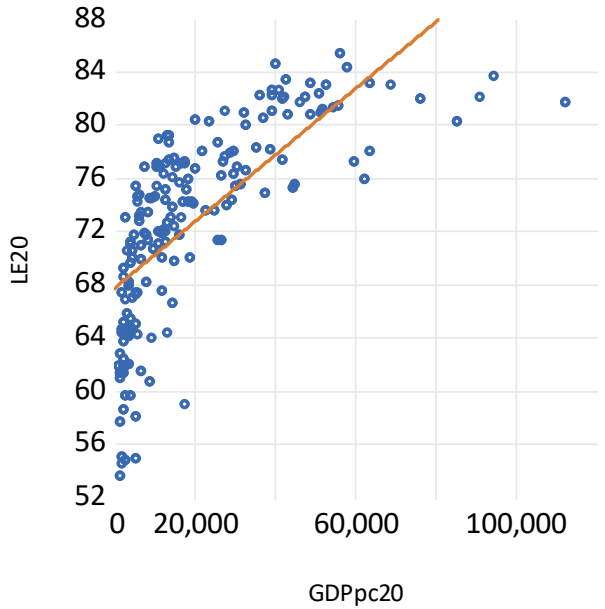
# SLR Model: Real GDP (PPP) per capita and Life expectancy at birth (2020), 182 countries

Dependent Variable: LE20
Method: Least Squares
Date: 09/03/22 Time: 16:48
Sample: 1 217
Included observations: 182

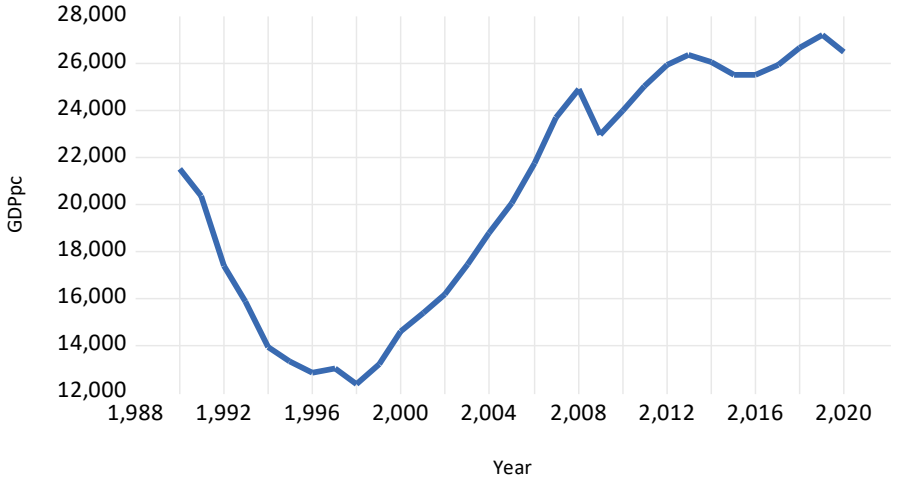
| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 67.76663 | 0.540498 | 125.3780 | 0.0000 |
| GDPPC20 | 0.000249 | 1.87E-05 | 13.33038 | 0.0000 |

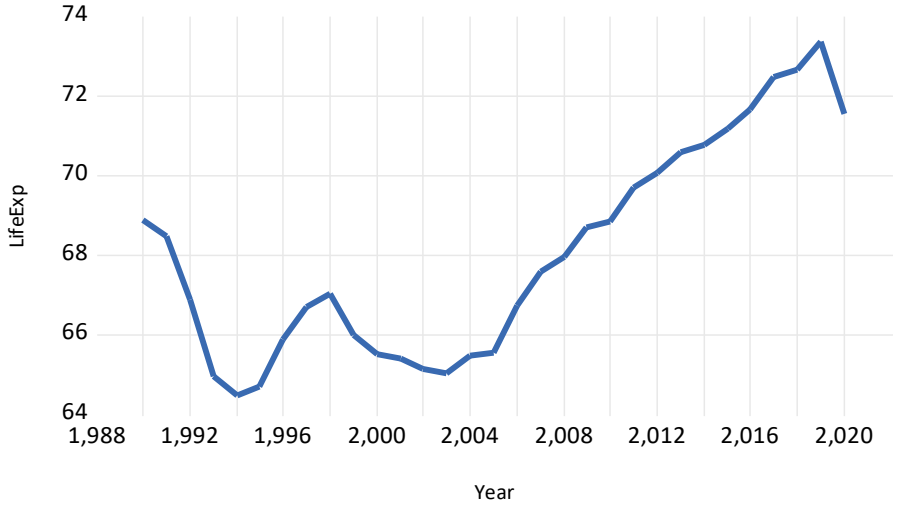
| | | | |
|---|---|---|---|
| R-squared | 0.496784 | Mean dependent var | 72.81049 |
| Adjusted R-squared | 0.493988 | S.D. dependent var | 7.319947 |
| S.E. of regression | 5.207009 | Akaike info criterion | 6.148817 |
| Sum squared resid | 4880.331 | Schwarz criterion | 6.184026 |
| Log likelihood | -557.5423 | Hannan-Quinn criter. | 6.163090 |
| F-statistic | 177.6990 | Durbin-Watson stat | 1.843746 |
| Prob(F-statistic) | 0.000000 | | |



# Real GDP per capita (GDPpc, USD, PPP, 2017) and Life Expectancy at birth (LifeExp), Russia, 1990-2020 (time series!)




Dependent Variable: LIFEEXP
Method: Least Squares
Date: 09/05/21 Time: 16:12
Sample: 1990 2020
Included observations: 31

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 58.91761 | 1.031057 | 57.14290 | 0.0000 |
| GDPPC | 0.000447 | 4.89E-05 | 9.133864 | 0.0000 |

| | | | |
|---|---|---|---|
| R-squared | 0.742056 | Mean dependent var | 68.05202 |
| Adjusted R-squared | 0.733161 | S.D. dependent var | 2.704471 |
| S.E. of regression | 1.397033 | Akaike info criterion | 3.568920 |
| Sum squared resid | 56.59935 | Schwarz criterion | 3.661435 |
| Log likelihood | -53.31825 | Hannan-Quinn criter. | 3.599077 |
| F-statistic | 83.42747 | Durbin-Watson stat | 0.285841 |
| Prob(F-statistic) | 0.000000 | | |