REGRESSION COEFFICIENTS: FIXED AND RANDOM COMPONENTS

True model

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

Fitted model

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

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

$$= \frac{\sum(X_i - \bar{X})([\beta_1 + \beta_2 X_i + u_i] - [\beta_1 + \beta_2 \bar{X} + \bar{u}])}{\sum(X_i - \bar{X})^2}$$

$$= \frac{\sum(X_i - \bar{X})(\beta_2(X_i - \bar{X}) + u_i - \bar{u})}{\sum(X_i - \bar{X})^2}$$

$$= \beta_2 + \frac{\sum(X_i - \bar{X})(u_i - \bar{u})}{\sum(X_i - \bar{X})^2}$$

We decompose $\hat{\beta}_2$ into its nonrandom and random components.

COEFFICIENTS a_i

We denote $x_i = X_i - \overline{X}$, $y_i = Y_i - \overline{Y}$ $\qquad a_i = \dfrac{X_i - \overline{X}}{\sum_{j=1}^{n}\left(X_j - \overline{X}\right)^2}$

Then $\hat{\beta}_2 = \dfrac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \dfrac{\sum_i x_i y_i}{\sum_i x_i^2} = \sum a_i y_i$

a_i cofficients are non−stochastic since they involve only x′s which are non−stochastic by the assumptions of the model A.
These coefficients are convenient for derivation of the formulas of the population variances of $\hat{\beta}_1$ and $\hat{\beta}_2$.
Their properties are used for the proof of Gauss−Markov theorem.

UNBIASEDNESS OF THE REGRESSION COEFFICIENTS

$$\hat{\beta}_2 = \beta_2 + \frac{\sum(X_i - \bar{X})(u_i - \bar{u})}{\sum(X_i - \bar{X})^2} \quad = \beta_2 + \sum a_i u_i$$

$$\sum(X_i - \bar{X})(u_i - \bar{u}) = \sum(X_i - \bar{X})u_i$$

$$\bar{u}\sum(X_i - \bar{X}) = 0 \qquad \text{since} \qquad \sum(X_i - \bar{X}) = 0$$

$$\mathrm{E}(\hat{\beta}_2) = \mathrm{E}(\beta_2) + \mathrm{E}\left(\sum a_i u_i\right)$$

$$= \beta_2 + \sum \mathrm{E}(a_i u_i) = \beta_2 + \sum a_i \mathrm{E}(u_i)$$

The proof of unbiasedness of the intercept estimator will be done below.

PROPERTIES OF COEFFICIENTS a_i

$$1) \sum_{i=1}^{n} a_i = 0$$

$$\text{Proof : } \sum_{i=1}^{n} a_i = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sum_{j=1}^{n} \left(X_j - \bar{X}\right)^2} \right) = \frac{1}{\sum_{j=1}^{n} \left(X_j - \bar{X}\right)^2} \sum_{i=1}^{n} (X_i - \bar{X}) = 0$$

$$\text{Since} \quad \sum_{i=1}^{n} (X_i - \bar{X}) = \sum_{i=1}^{n} X_i - n\bar{X} = n\bar{X} - n\bar{X} = 0$$

PROPERTIES OF COEFFICIENTS a_i

$$2)\ \sum_{i=1}^{n} a_i^2 = \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{1}{\sum_{i=1}^{n} x_i{}^2}$$

Proof :

$$\sum_{i=1}^{n} a_i^2 = \sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sum_{j=1}^{n}\left(X_j - \bar{X}\right)^2}\right)^2 = \frac{1}{\left(\sum_{j=1}^{n}\left(X_j - \bar{X}\right)^2\right)^2}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$= \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{1}{\sum_{i=1}^{n} x_i{}^2}$$

PROPERTIES OF COEFFICIENTS a_i

$$3) \sum_{i=1}^{n} a_i X_i = 1$$

Proof :

$$\sum_{i=1}^{n} a_i X_i = \sum_{i=1}^{n} \frac{(X_i - \bar{X})X_i}{\sum_{j=1}^{n}(X_j - \bar{X})^2} = \frac{1}{\sum_{j=1}^{n}(X_j - \bar{X})^2} \sum_{i=1}^{n} (X_i - \bar{X})X_i = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sum_{j=1}^{n}(X_j - \bar{X})^2} = 1 .$$

Since

$$\sum \bar{X}(X_i - \bar{X}) = \bar{X} \sum (X_i - \bar{X}) = \bar{X}(n\bar{X} - n\bar{X}) = 0$$

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_u}{\sum(X_i - \bar{X})^2}$$

$$\sigma^2_{\hat{\beta}_1} = \sigma^2_u \left\{ \frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2} \right\}$$

For the variance of $\hat{\beta}_2$:

- The larger is the population variance of *u*, the larger is the variance of $\hat{\beta}_2$.

-the larger is the sum of the squared deviations of *X*, the smaller is the variance of $\hat{\beta}_2$

The variance of $\hat{\beta}_1$ is also growing with the (mean X)^2 growth.

PRECISION OF THE REGRESSION COEFFICIENTS: PROOF FOR $\hat{\beta}_2$

$$\sigma^2_{\hat{\beta}_2} = \mathrm{E}\left\{\left(\hat{\beta}_2 - E(\hat{\beta}_2)\right)^2\right\} = \mathrm{E}\left\{\left(\hat{\beta}_2 - \beta_2\right)^2\right\} = \mathrm{E}\left\{\left(\sum_{i=1}^{n} a_i u_i\right)^2\right\} =$$

$$= \mathrm{E}\left\{\sum_{i=1}^{n} a_i^2 u_i^2 + \sum_{i=1}^{n}\sum_{j\neq i} a_i a_j u_i u_j\right\} = \sum_{i=1}^{n} a_i^2 \mathrm{E}\left(u_i^2\right) + \sum_{i=1}^{n}\sum_{j\neq i} a_i a_j \mathrm{E}\left(u_i u_j\right) =$$

$$= \sum_{i=1}^{n} a_i^2 \sigma_u^2 = \sigma_u^2 \sum_{i=1}^{n} a_i^2 = \frac{\sigma_u^2}{\sum_{j=1}^{n}\left(X_j - \bar{X}\right)^2}$$

PRECISION OF THE REGRESSION COEFFICIENTS: PROOF FOR $\widehat{\boldsymbol{\beta}}_1$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2\bar{X} = (\beta_1 + \beta_2\bar{X} + \bar{u}) - \bar{X}(\beta_2 + \sum a_i u_i) =$$
$$= \beta_1 + \frac{1}{n}\sum u_i - \bar{X}\sum a_i u_i = \beta_1 + \sum c_i u_i$$

where $c_i = \frac{1}{n} - a_i\bar{X}$, and hence $\mathrm{E}(b_1) = \beta_1$.

$$\sigma_{b_1}^2 = \mathrm{E}\left[\left(\sum c_i u_i\right)^2\right] = \sigma_u^2 \sum c_i^2 = \sigma_u^2\left(n\frac{1}{n^2} - 2\frac{\bar{X}}{n}\sum a_i + \bar{X}^2\sum a_i^2\right).$$

Since $\sum a_i = 0$ and $\sum a_i^2 = \frac{1}{\sum(X_i - \bar{X})^2}$, we get:

$$\sigma_{b_1}^2 = \sigma_u^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)$$

GAUSS-MARKOV THEOREM

Gauss–Markov theorem states that, provided that the assumptions of Model A are satisfied, the OLS estimators are BLUE: best (most efficient) linear (combinations of the Y_i) unbiased estimators of the regression parameters.

Let $\tilde{\beta}_2 = \sum_{i=1}^{n} g_i Y_i$

Scheme of the Efficiency Proof (see Dougherty's book) :

$$\sigma^2_{\tilde{\beta}_2} = \mathrm{E}\left\{\left(\tilde{\beta}_2 - E\left(\tilde{\beta}_2\right)\right)^2\right\} = \mathrm{E}\left\{\sum_{i=1}^{n}(g_i u_i)^2\right\} = \sigma_u^2 \sum_{i=1}^{n} g_i^2 =$$

$$= \sigma_u^2 \sum_{i=1}^{n}(a_i + h_i)^2 = \sigma_u^2 \left\{\sum_{i=1}^{n} a_i^2 + \sum_{i=1}^{n} h_i^2 + 2\sum_{i=1}^{n} a_i h_i\right\}$$

$$= \sigma_u^2 \left\{\sum_{i=1}^{n} a_i^2 + \sum_{i=1}^{n} h_i^2\right\}$$

STANDARD DEVIATIONS OF THE REGRESSION COEFFICIENTS

$$\sigma_{\hat{\beta}_1}^2 = \sigma_u^2 \left\{ \frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2} \right\} \qquad \sigma_{\hat{\beta}_2}^2 = \frac{\sigma_u^2}{\sum (X_i - \bar{X})^2}$$

$$s_u^2 = \frac{1}{n-2} \sum \hat{u}_i^2$$

S_u - standard error of regression.

Standard errors of regression coefficients:

$$\text{s.e.}(\hat{\beta}_1) = s_u \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2}} \qquad \text{s.e.}(\hat{\beta}_2) = \sqrt{\frac{s_u^2}{\sum (X_i - \bar{X})^2}}$$

We obtain estimates of the standard deviations of the distributions of $\hat{\beta}_1$ and $\hat{\beta}_2$ by substituting s_u^2 for σ_u^2 in the variance expressions and taking the square roots.

t TESTS OF HYPOTHESES RELATING TO REGRESSION COEFFICIENTS

True model	Fitted model
$Y = \beta_1 + \beta_2 X + u$	$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X$

Null hypothesis $H_0: \beta_2 = \beta_2^0,$

Alternative (two-sided) hypothesis $H_1: \beta_2 \neq \beta_2^0$

Test statistic $t = \dfrac{\hat{\beta}_2 - \beta_2^0}{\text{s.e.}(\hat{\beta}_2)}$

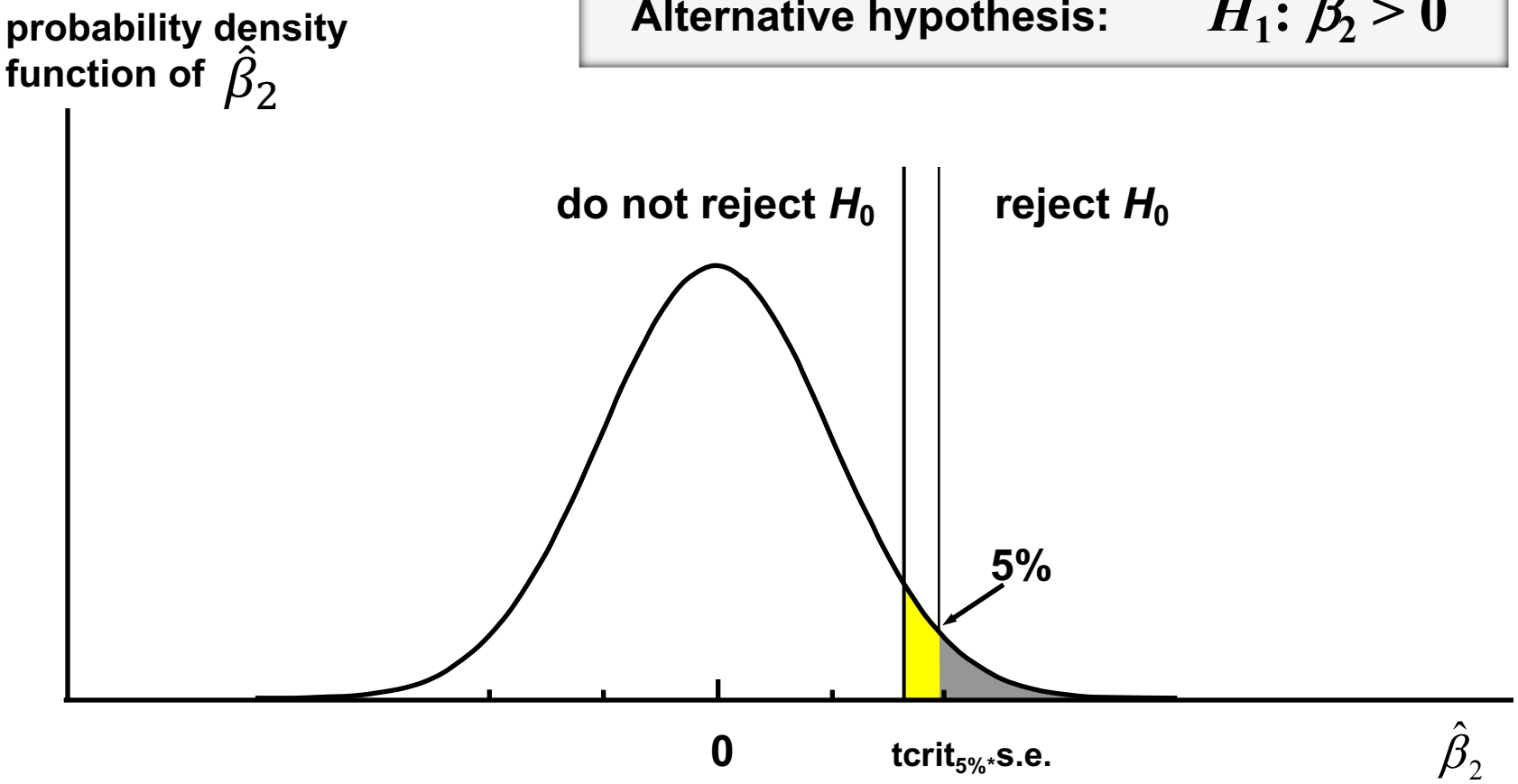
Reject H_0 if $|t| > t_{\text{crit}}$

d.f. = n-2

ONE-SIDED *t* TESTS OF HYPOTHESES RELATING TO REGRESSION COEFFICIENTS

Null hypothesis: $H_0: \beta_2 = 0$

Alternative hypothesis: $H_1: \beta_2 > 0$



CONFIDENCE INTERVALS FOR REGRESSION COEFFICIENTS

Model $\boldsymbol{Y = \beta_1 + \beta_2 X + u}$

Null hypothesis: $\boldsymbol{H_0: \beta_2 = \beta_2^0}$

Alternative hypothesis: $\boldsymbol{H_1: \beta_2 \neq \beta_2^0}$

Reject H_0 if $\dfrac{\hat{\beta}_2 - \beta_2^0}{\text{s.e.}(\hat{\beta}_2)} > t_{\text{crit}}$ or $\dfrac{\hat{\beta}_2 - \beta_2^0}{\text{s.e.}(\hat{\beta}_2)} < -t_{\text{crit}}$

Reject H_0 if $\hat{\beta}_2 - \beta_2^0 > \text{s.e.}(\hat{\beta}_2) \times t_{\text{crit}}$ or $\hat{\beta}_2 - \beta_2^0 < -\text{s.e.}(\hat{\beta}_2) \times t_{\text{crit}}$

Reject H_0 if $\hat{\beta}_2 - \text{s.e.}(\hat{\beta}_2) \times t_{\text{crit}} > \beta_2^0$ or $\hat{\beta}_2 + \text{s.e.}(\hat{\beta}_2) \times t_{\text{crit}} < \beta_2^0$

Do not reject H_0 if $\hat{\beta}_2 - \text{s.e.}(\hat{\beta}_2) \times t_{\text{crit}} \leq \beta_2 \leq \hat{\beta}_2 + \text{s.e.}(\hat{\beta}_2) \times t_{\text{crit}}$

$(\hat{\beta}_2 - \text{s.e.}(\hat{\beta}_2) \times t_{\text{crit}}; \hat{\beta}_2 + \text{s.e.}(\hat{\beta}_2) \times t_{\text{crit}})$ - Confidence interval

F TEST OF GOODNESS OF FIT

Demonstration that $F = t^2$ IN THE SLR MODEL

$$F(k-1, n-k) = \frac{SSE/(k-1)}{SSR/(n-k)} = \frac{\frac{SSE}{SST}/(k-1)}{\frac{SSR}{SST}/(n-k)} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

$$F = \frac{SSE}{SSR/(n-2)} = \frac{\sum\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum \hat{u}_i^2/(n-2)}$$

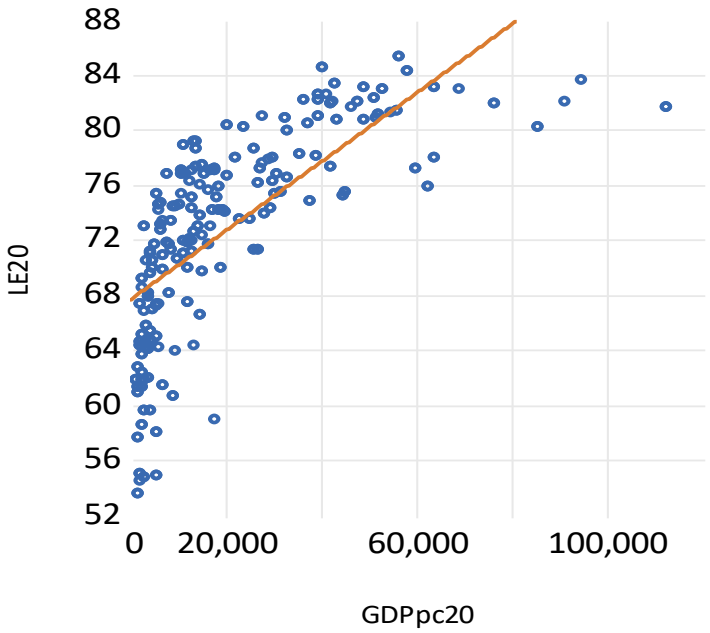
$$= \frac{\sum\left([\hat{\beta}_1 + \hat{\beta}_2 X_i] - [\hat{\beta}_1 + \hat{\beta}_2 \bar{X}]\right)^2}{s_u^2} = \frac{1}{s_u^2}\sum \hat{\beta}_2^2 (X_i - \bar{X})^2$$

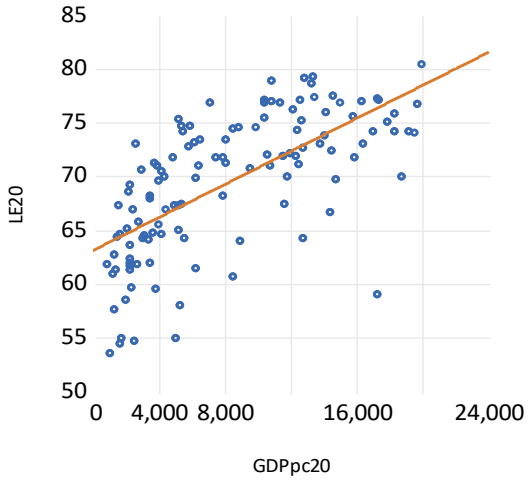
$$= \frac{\hat{\beta}_2^2}{s_u^2}\sum (X_i - \bar{X})^2 = \frac{\hat{\beta}_2^2}{s_u^2/\sum (X_i - \bar{X})^2} = \frac{\hat{\beta}_2^2}{\left(s.e.(\hat{\beta}_2)\right)^2} = t^2$$

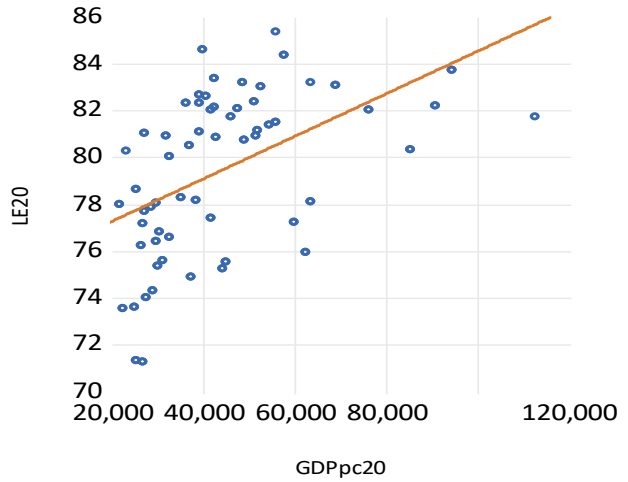
The *F* test does not have its own role in the SLR model; it will do in the multiple regression.

The of Real GDP (PPP) per capita and Life expectancy at birth (2020)

World Development Indicators







Dependent Variable: LE20
Method: Least Squares
Date: 09/03/22 Time: 17:20
Sample: 26 152
Included observations: 121

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	63.22122	0.785302	80.50566	0.0000
GDPPC20	0.000763	7.95E-05	9.597118	0.0000

R-squared	0.436299	Mean dependent var	69.46350
Adjusted R-squared	0.431562	S.D. dependent var	6.420147
S.E. of regression	4.840460	Akaike info criterion	6.008288
Sum squared resid	2788.177	Schwarz criterion	6.054499
Log likelihood	-361.5014	Hannan-Quinn criter.	6.027056
F-statistic	92.10467	Durbin-Watson stat	1.792756
Prob(F-statistic)	0.000000		

Dependent Variable: LE
Method: Least Squares
Date: 09/12/20 Time: 19:21
Sample: 162 217
Included observations: 54

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	76.51574	0.980857	78.00904	0.0000
GDP_PC	6.86E-05	1.84E-05	3.727162	0.0005

R-squared	0.210827	Mean dependent var	79.85820
Adjusted R-squared	0.195650	S.D. dependent var	3.255438
S.E. of regression	2.919657	Akaike info criterion	5.017143
Sum squared resid	443.2687	Schwarz criterion	5.090809
Log likelihood	-133.4629	Hannan-Quinn criter.	5.045553
F-statistic	13.89174	Durbin-Watson stat	1.708254
Prob(F-statistic)	0.000479		

The Simple Regression Model: some nonlinearities

Semi-logarithmic form

Regression of log wages on years of education

$$\log(wage) = \beta_0 + \beta_1 educ + u$$

Natural logarithm of wage

This changes the interpretation of the regression coefficient:

$$\beta_1 = \frac{\Delta \log(wage)}{\Delta educ} = \frac{1}{wage} \cdot \frac{\Delta wage}{\Delta educ} = \frac{\frac{\Delta wage}{wage}}{\Delta educ}$$

Percentage change of wage

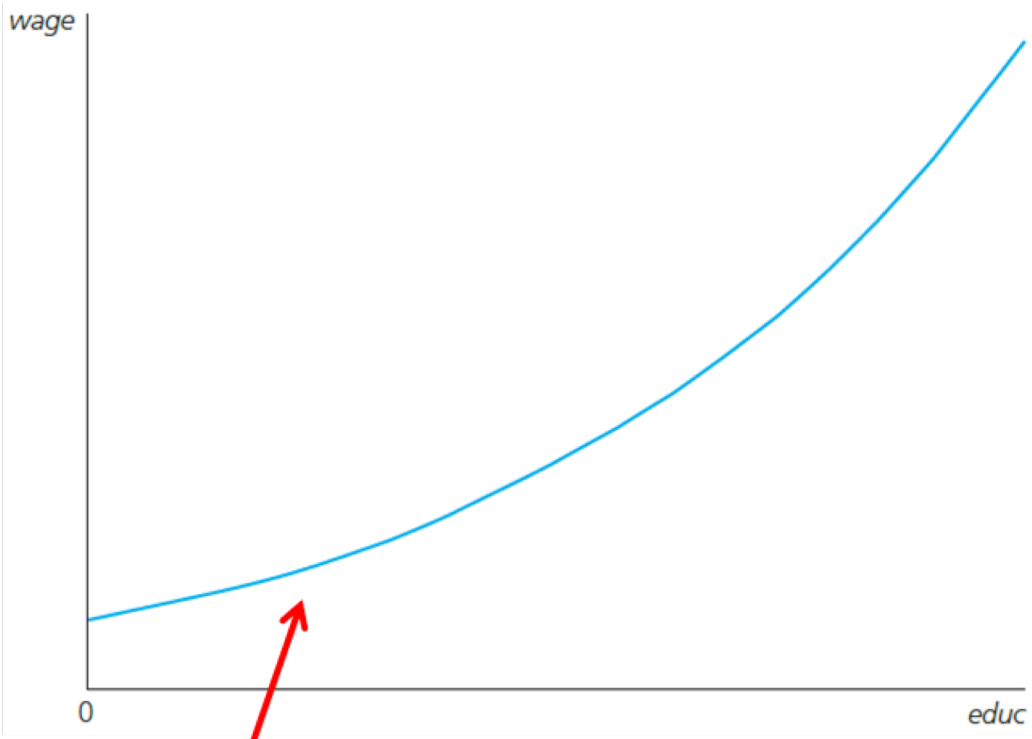
... if years of education are increased by one year

The Simple Regression Model: some nonlinearities

Fitted regression

$$\widehat{\log}(wage) = 0.584 + 0.083\ educ$$

The wage increases by 8.3% for
every additional year of education
(= return to another year of education)



Growth rate of wage is 8.3%
per year of education

The Simple Regression Model: some nonlinearities

Incorporating nonlinearities: Log-logarithmic form

Chief Executive Officer (CEO) salary and firm sales

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + u$$

Natural logarithm of CEO salary

Natural logarithm of his/her firm's sales

This changes the interpretation of the regression coefficient:

$$\beta_1 = \frac{\Delta \log(salary)}{\Delta \log(sales)} = \frac{\frac{\Delta salary}{salary}}{\frac{\Delta sales}{sales}}$$

Percentage change in salary if sales increase by 1%

Logarithmic changes are always percentage changes

The Simple Regression Model: some nonlinearities

CEO salary and firm sales: fitted regression

$$\widehat{\log}(salary) = 4.822 + 0.257 \log(sales)$$

$+1\%\ sales \rightarrow +.257\%\ salary$

The double log form means a constant elasticity model, whereas the semi-log form assumes the relation between the absolute and relative changes