2 Regression analysis

(1) Bivariate correlation analysis

(2) Simple regression

(3) Estimation and goodness of fit

(4) Hypothesis testing with the t-test

(5) Multiple regression

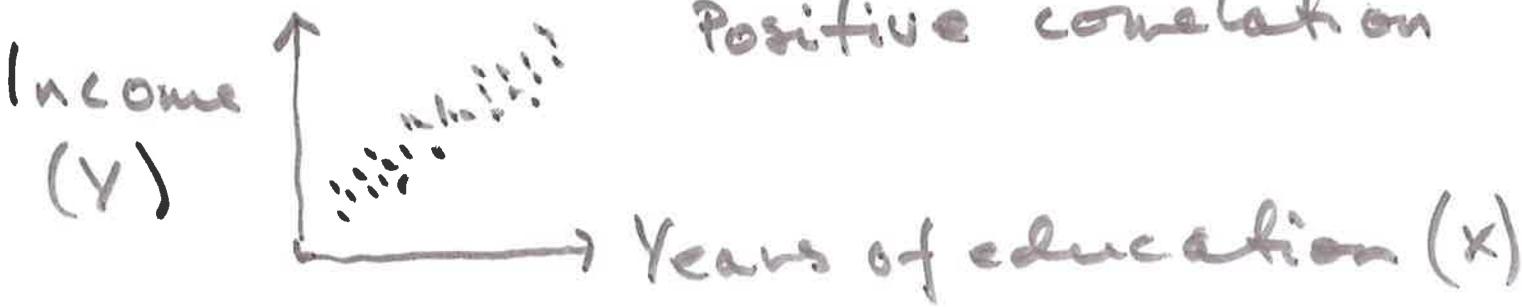
(6) Hypothesis testing with the F-test

(7) Suggested exercises

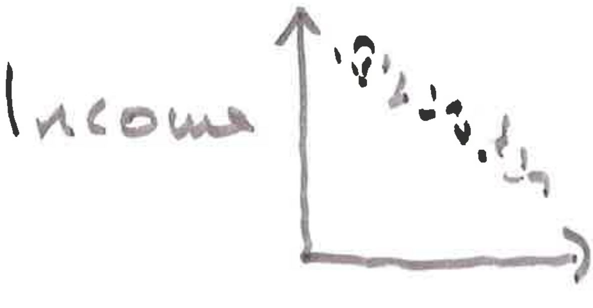
① Bivariate correlation analysis

Def. Bivariate correlation: Statistical association between two variables

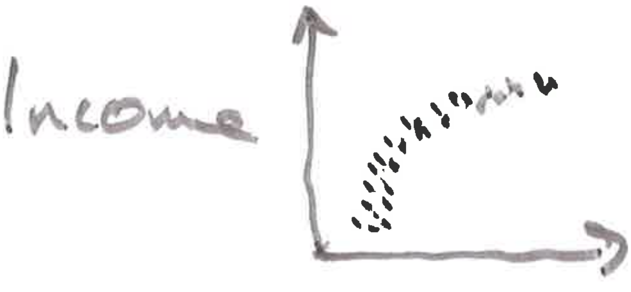
Examples:



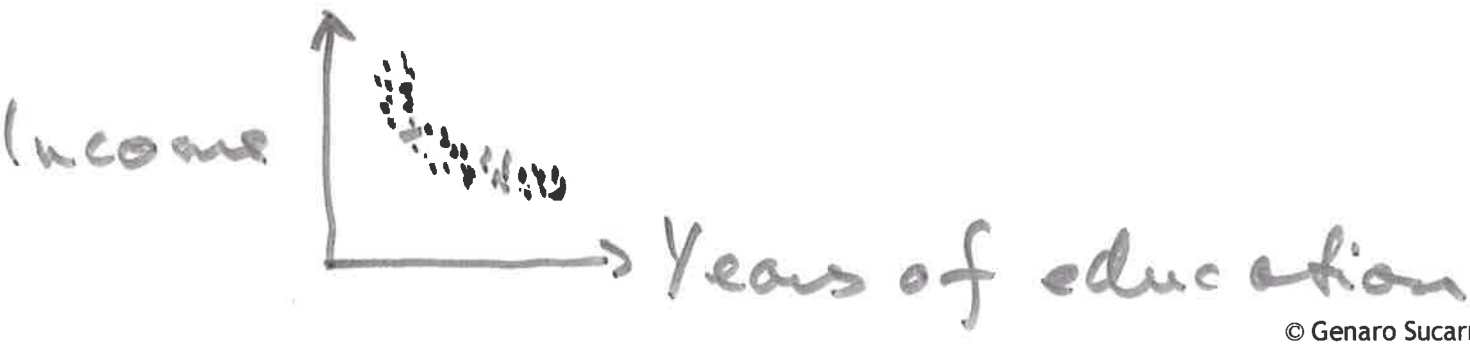
Positive correlation

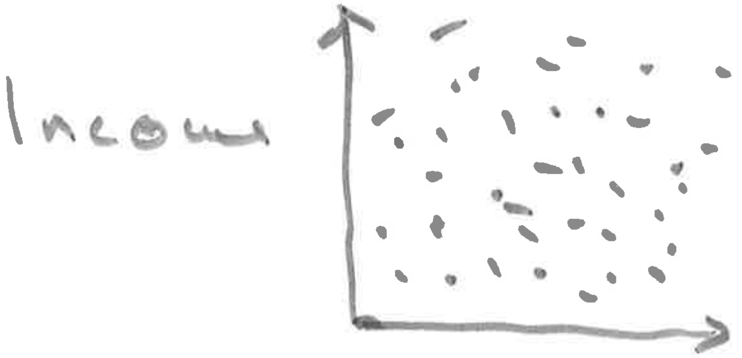


Negative correlation



Nonlinear positive correlation





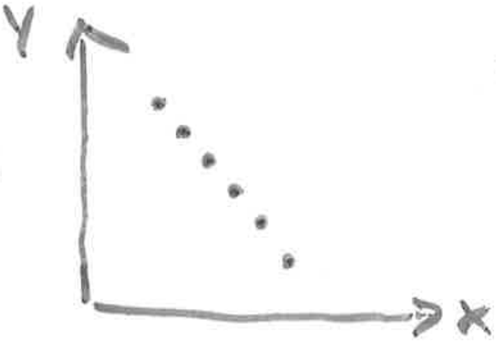
No correlation

Pearson's r ("the" correlation)

→ A linear measure of correlation

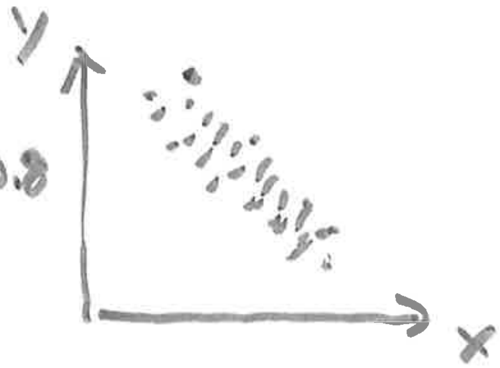
→ Varies between -1 and 1

$r = -1$



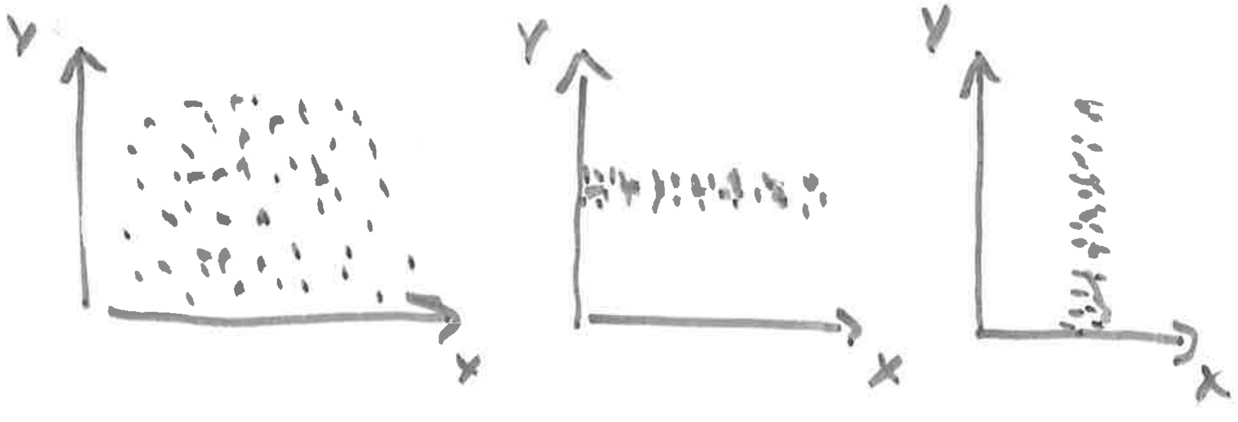
Perfect negative correlation

$r \approx -0.8$

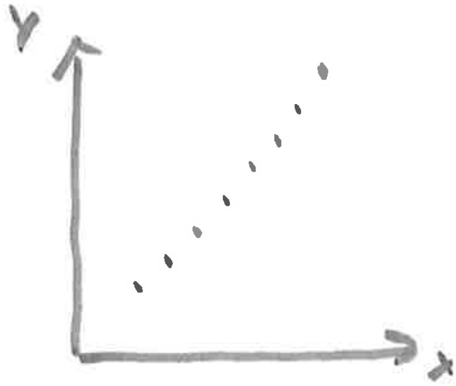


Non-perfect negative correlation

$r \approx 0$

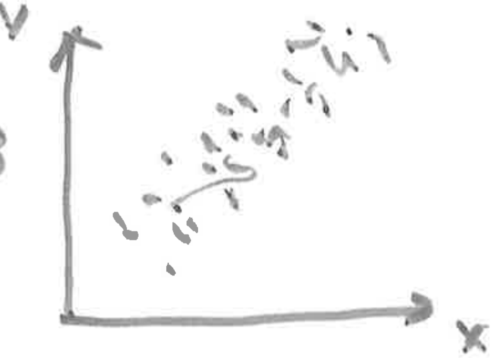


$r = 1$



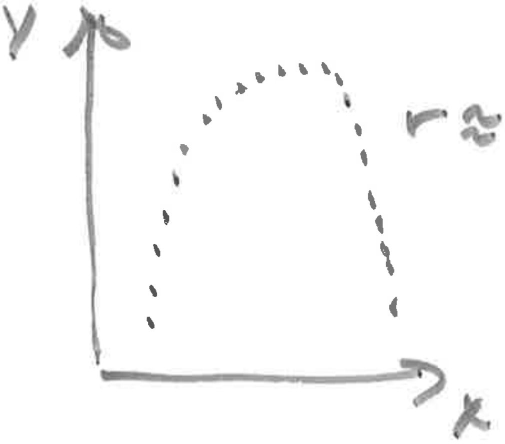
Perfect positive correlation

$r \approx 0.8$

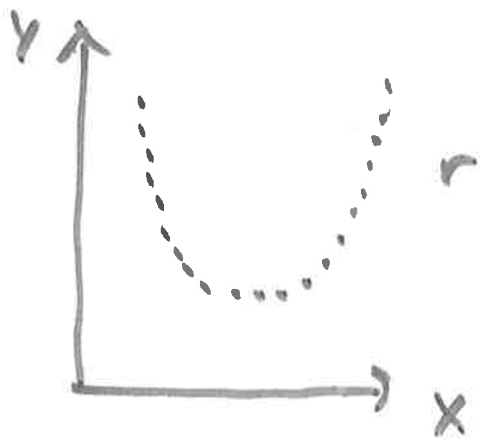


Non-perfect positive correlation

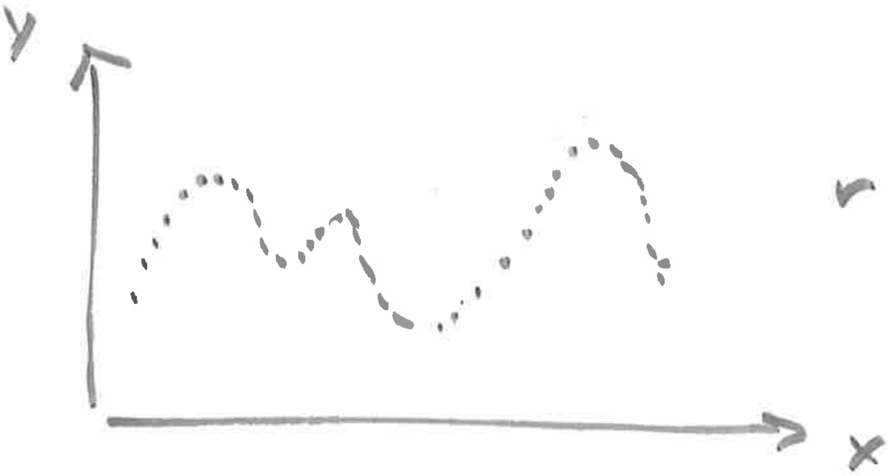
No correlation according to Pearson's r (when there actually is):



$r \approx 0$



$r \approx 0$



$r \approx 0$

Def. Pearson's r ("the" sample correlation):

$$r_{XY} = \frac{S_{XY}}{S_X \cdot S_Y}$$

where

$$S_{XY} = \frac{\sum (x - \bar{x}) \cdot (y - \bar{y})}{n - 1}$$

Example: Q1: Wake-up hour (Y)
Q3: Travel-time (x)

$$Y \xleftarrow{\div} X$$

$\bar{Y} = 7.4 \approx 7h\,24m$ $\quad \bar{X} = 24.8$

What is the correlation between X and Y?

Q1	$(Y-\bar{Y})$	$(Y-\bar{Y})^2$	Q3	$(X-\bar{X})$	$(X-\bar{X})^2$	$(X-\bar{X})\cdot(Y-\bar{Y})$
7.17	-0.23	0.05	30	5.2	27.04	$-0.23 \cdot 5.2$
8.33	0.93	0.86	27.5	2.7	7.29	$0.93 \cdot 2.7$
6.25	⋮	⋮	20	⋮	⋮	⋮
7	etc.	etc.	30	etc.	etc.	etc.
7			2			
6.5			50			
8.07			28			
8.5			20			
9			20			
6			20			
SUMS:		9.5492			1318.65	-12.9

$\bar{Y} = 7.4$, $\bar{X} = 24.8$, $\Sigma(X-\bar{X})^2$

$$r_{XY} = \frac{S_{XY}}{s_X \cdot s_Y} = -0.115$$

S_{XY}: -1.43; s_X: 12.10; s_Y: 1.03

$$s_X^2 = \frac{\Sigma(X-\bar{X})^2}{n-1} = \frac{1318.65}{9} = 146.2$$

$$s_X = \sqrt{s_X^2} = \sqrt{146.2} = \underline{12.10}$$

$$s_y^2 = \frac{\sum(y-\bar{y})^2}{n-1} = \frac{9.5492}{9} = 1.061$$

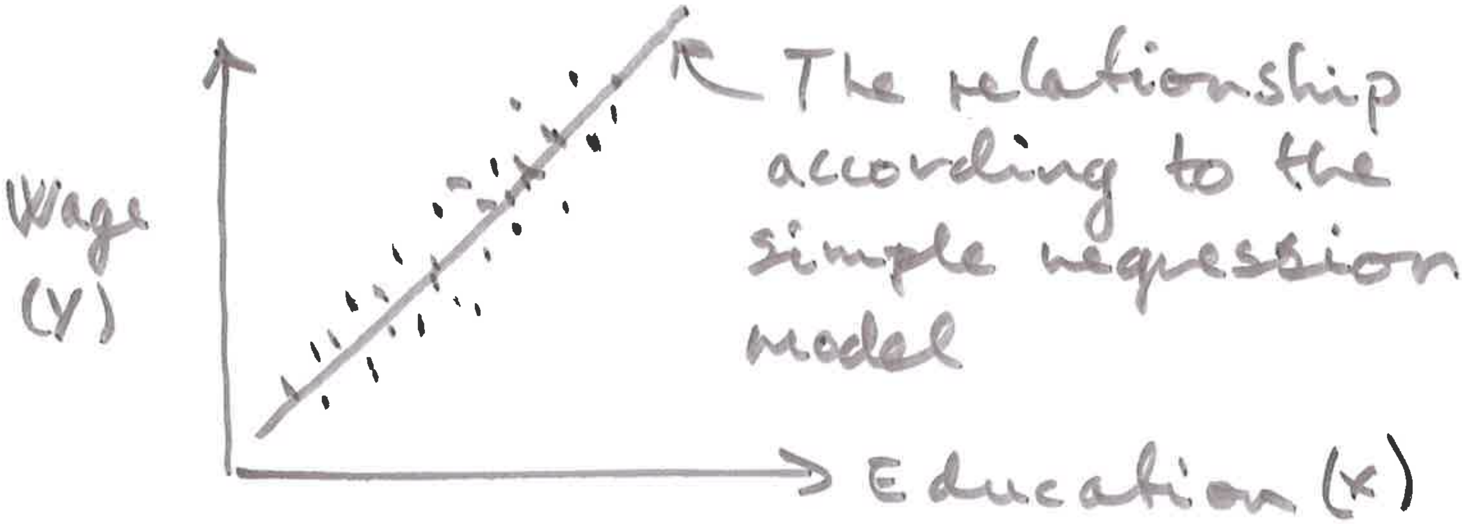
$$s_y = \sqrt{s_y^2} = \sqrt{1.061} = \underline{1.03}$$

$$s_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1} = \frac{-12.9}{9} = \underline{-1.43}$$

Interpretation: The more travel-time, the earlier you wake up

② Simple regression

Simple = One X-variable



Uses: Testing, prediction, counterfactual analysis, intervention analysis

The simple regression model:

e.g. wage

parameters

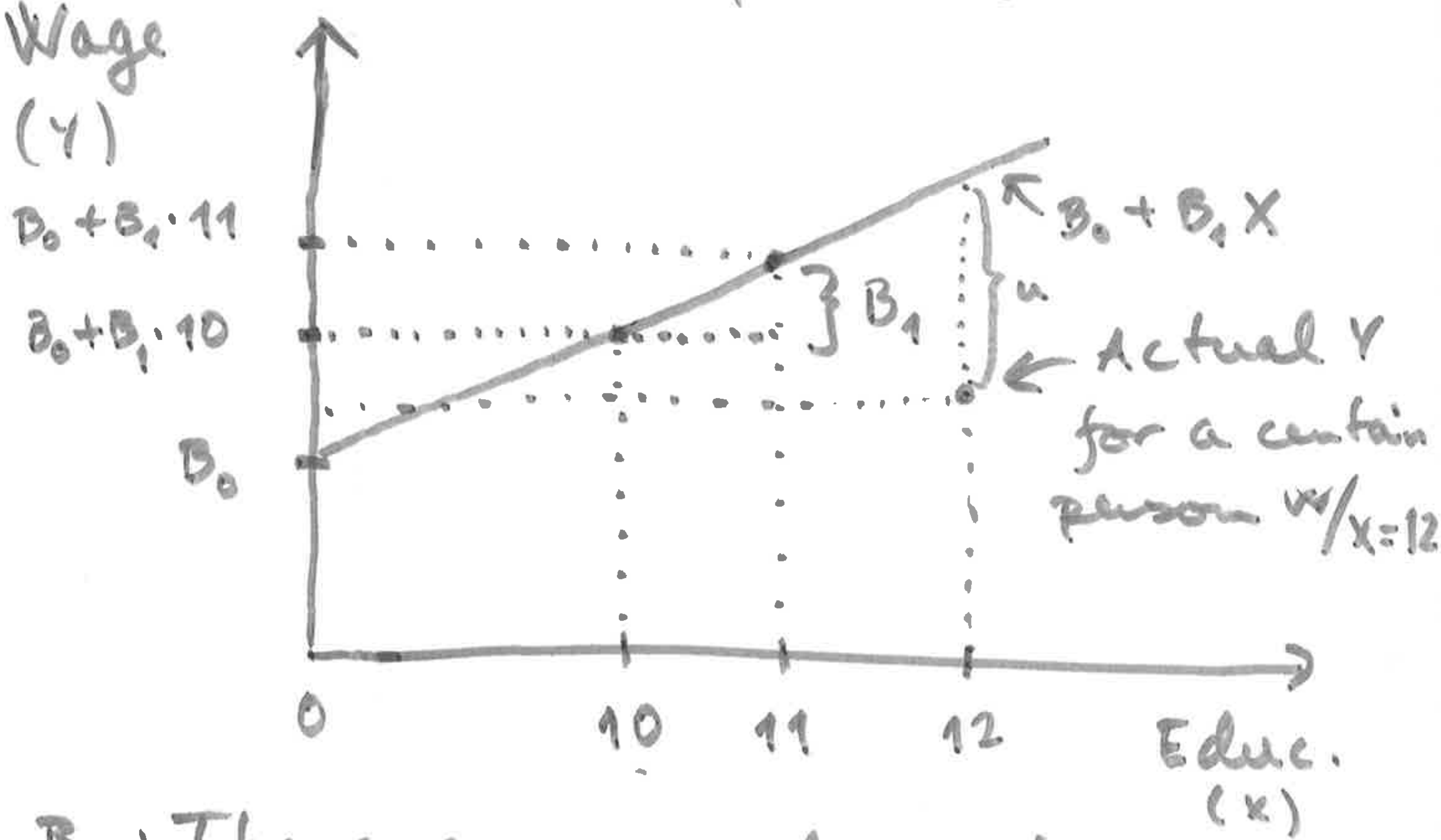
e.g. educ.

error/residual

$$Y = \underbrace{B_0 + B_1 X}_{\text{Explanation/prediction}} + u$$

Intercept

slope coefficient (i.e. the impact of X)



B_0: The average value of Y when $X = 0$

B_1: The slope or effect of X; the average change in Y when X increases with 1 unit

$B_0 + B_1 X$: Prediction / forecast of Y for a value X; the regression line

Example (wage data):

Y = hourly wage in USD

X = years of work experience

$$Y = B_0 + B_1 X + u$$

Estimates: 10.163 0.117

B_0 : Predicted wage for those with no working experience is 10.16 USD

B_1 : One more year of work experience increases on average the hourly wage by 0.12 USD ($\approx$ 12 cents)

Example (wage data):

Y = wage X = Years of education

$$Y = B_0 + B_1 X + u$$

Estimates: -4.474 1.281

B_1: One more year of education increases on average the hourly wage by 1.28 USD

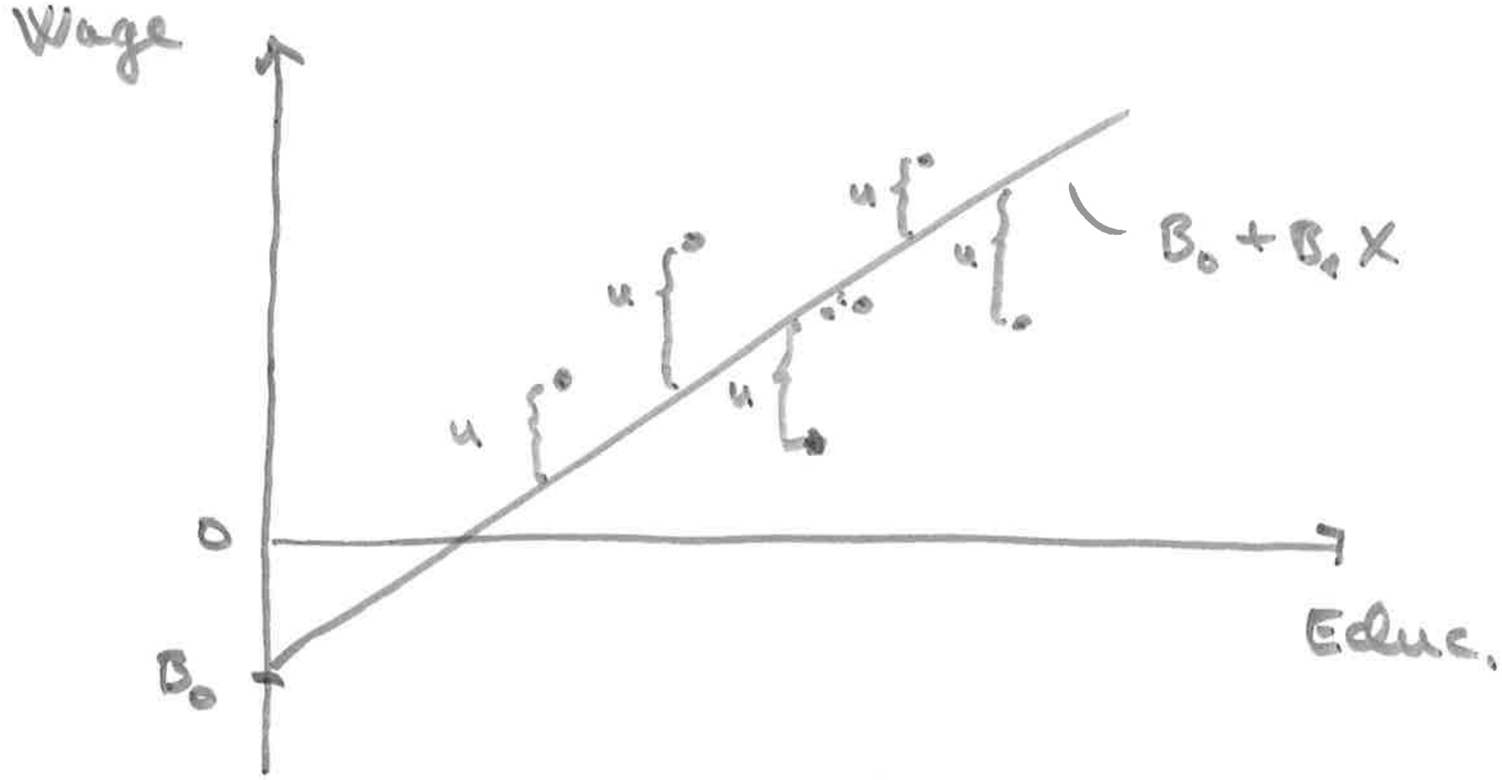
B_0: In this case the economic interpretation of a negative wage does not make sense

3 Estimation and goodness of fit

How do we estimate B_0 and B_1?

OLS!

The Ordinary Least Squares (OLS) method consists of choosing the values B_0 and B_1 such that the sum of the squared prediction errors is minimised.



Estimate of B_1: $\frac{S_{XY}}{S_X^2}$

—— " —— B_0: $\bar{Y} - \hat{B}_1 \bar{X}$

estimate of B_1

Example (wage data):

$Y = \text{wage} \qquad X = \text{exper}$

$\bar{Y} = 12.37 \qquad \bar{X} = 18.79$

$S_{XY} = 15.94 \qquad S_X^2 = 135.92$

Estimate of B_1:

$$\frac{\overset{15.94}{S_{XY}}}{\underset{135.92}{S_X^2}} = \underline{\underline{0.1173}}$$

Estimate of B_0:

$$\underset{12.37}{\bar{Y}} - \underset{0.1173}{\hat{B}_1} \underset{18.79}{\bar{X}} = \underline{\underline{10.1659}}$$

R-squared (R^2):

→ A measure of goodness-of-fit or precision

→ Varies between 0 and 1 (or 0% and 100%):

↳ If 0, then the model explains or predict nothing

↳ If 1, then the model explains 100% of the variation in Y

→ In simple regression:

$$R^2 = (r_{XY})^2 = \left(\frac{S_{XY}}{S_X \cdot S_Y}\right)^2$$

→ R^2 is defined as:

$$R^2 = 1 - \frac{RSS}{TSS}$$

← The unexplained variation

where

TSS = Total Sum of Squares of Y, i.e. $\Sigma(Y-\bar{Y})^2$

RSS = Residual Sum of Squares, i.e. Σu^2

Example (wage data):

Y = wage X = exper Model: $Y = \beta_0 + \beta_1 X + u$

$$\left.\begin{array}{l} TSS = 80\,309.824 \\ RSS = 77\,901.414 \end{array}\right\} \quad \frac{RSS}{TSS} = 0.97$$

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - 0.97 = 0.03$$

Interpretation: Years of work experience explains about 3% of the variation in wage

Note: $|r_{XY}| = \sqrt{R^2} = \sqrt{0.03} = 0.17$

Adjusted R-squared:

→ Problem with R-squared: Never falls when an X-variable is added, even if the X-variable explains nothing

→ Adjusted R-squared: Falls whenever an irrelevant X-variable is added

Def. Adjusted R-squared:

$$1-\left[(1-R^2)\cdot\left(\frac{n-1}{n-k}\right)\right]$$

the number of Bs

④ Hypothesis testing with the t-test

Recall: $Y = B_0 + B_1 X + u$

e.g. wage

e.g. exper

A four-step recipe for testing a single B:

Step 1: Choose α, formulate H_0 and H_A:

- $H_0: B = 0$ $\qquad$ $H_A: B \neq 0$

$H_A: B > 0$

$H_A: B < 0$

2: Identify the rejection area using a t-distribution with $df = n - k$

k = number of Bs

3: Compute the value of the test expression

$$\frac{\overbrace{\text{Estimate of } B}^{\hat{B}} - H_0 \text{ value}}{\underbrace{\text{standard error}(\hat{B})}_{se(\hat{B})}}$$

4: Conclude: Reject H_0 if the test-value lies in the rejection area, otherwise keep H_0

Example (wage data):

Y = wage $\quad X$ = exper

$$Y = B_0 + B_1 X + u \qquad n = 1289$$

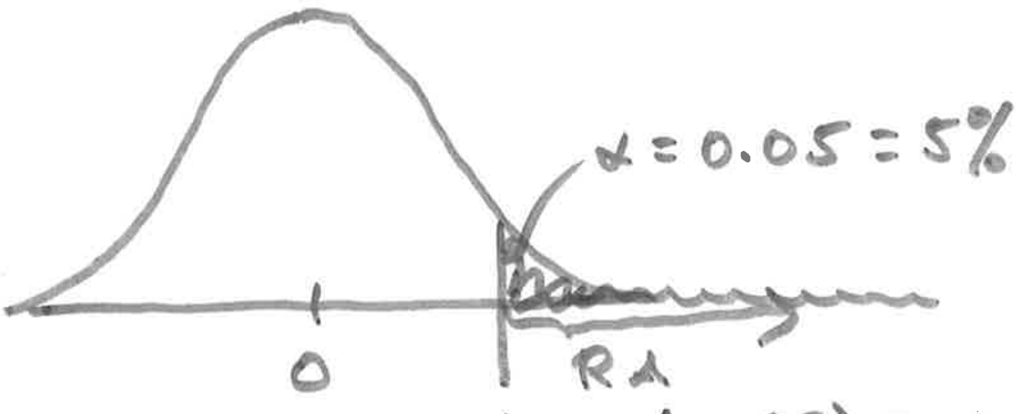
	B_0	B_1
Estimates:	10.63	0.117
$se(\hat{B})$:	0.411	0.019

Does more experience increase the wage-level?

Step 1: $\alpha = 5\%$ $\quad H_0: B_1 = 0 \quad H_1: B_1 > 0$

2: RA:

$$df = n - k = 1289 - 2 = 1287$$



$t_{0.05}(1287)$

$$\approx t_{0.05}(1000) = \underline{1.646}$$

RA: Values greater than 1.646

3. Testvalue:

$$\frac{\overbrace{\hat{B}_1}^{0.117} - \overbrace{H_0 \text{ value}}^{0}}{\underbrace{se(\hat{B}_1)}_{0.019}} = \underline{6.158}$$

4. Conclusion: We reject H_0

⑤ Multiple regression

Simple: wage $\xleftarrow{B_1}$ exper : x

Multiple: wage $\xleftarrow{B_1}$ exper : x_1

$\xleftarrow{B_2}$ educ : x_2

$\xleftarrow{B_3}$ gender : x_3

⋮

etc.

Why not repeated simple regression instead of multiple regression?

Because of "omitted variable bias": If the X-variables are correlated with each other then estimates and tests can be very misleading

The multiple regression model:

Dependent variable | Independent variables | error

$$\underbrace{Y = B_0 + B_1 X_1 + B_2 X_2 + \cdots + B_K X_K}_{\text{Explanation/prediction}} + u$$

Intercept: B_0

slope coefficients / effects of the Xs: $B_1, B_2, \ldots, B_K$

Interpretations:

B_0: The average or predicted value of Y when all Xs are 0

β_1: The average or predicted change in Y when X_1 increases by 1 unit, given that the other Xs do not change

β_2: ——— " ———

—— " —— X_2 —— " ———

——— " ———

⋮

β_K: ——— " ———

—— " —— X_K —— " ———

—— " ———

$\beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K$: Prediction/explanation of Y offered by the model

$$u = Y - \underbrace{\text{prediction}}_{\beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K}$$

Example (wage data):

$$\overset{\text{wage}}{Y} = \beta_0 + \beta_1 \overset{\text{exper}}{X_1} + \beta_2 \overset{\text{educ}}{X_2} + u$$

Estimates: -9.586 $\quad 0.179 \quad 1.415$

$se(\hat{\beta})$: $1.01 \quad 0.02 \quad 0.07$

β_0: Predicted wage is -9.59 USD, which does not make sense economically in this dataset

β_1: The average or predicted increase in wage for 1 more year of work experience is 0.18 USD, given that educ stays the same

β_2: The average or predicted increase in wage for 1 more year of education is 1.42 USD, given that exper stays the same

Does experience have an effect on wage? Does education?

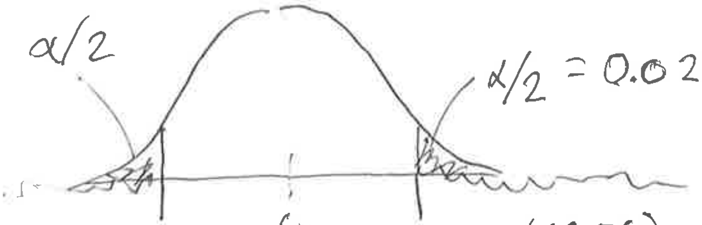
Experience:

1. $\alpha = 5\%$ $\quad H_0: B_1 = 0 \quad H_A: B_1 \neq 0$

2. RA: $df = n - \underbrace{\text{number of } B\text{s}}_{k}$

$= 1289 - 3 = 1286$

$\alpha/2$ $\qquad \alpha/2 = 0.025$



$-1.962 \qquad 0 \qquad t_{0.025}(1286)$

$\approx \underbrace{t_{0.025}(1000)}_{1.962}$

RA = Values higher than 1.962 and values lower than -1.962

3. Testvalue:

$$\frac{\hat{B}_1 - H_0 \text{ value}}{\underbrace{se(\hat{B}_1)}_{0.02}} = \underline{8.95}$$

$\hat{B}_1$: 0.179 $\qquad H_0$ value: 0

4. Conclusion: We reject H_0

Education

1. $\alpha =$

(6) Hypothesis testing with the F-test

Consider: $Y = B_0 + B_1 \, exper + B_2 \, educ + u$

wage

t-tests: Enable us to test the effect of exper and education separately, but not at the same time

F-tests: Enable us to test exper and educ simultaneously ("multiple hypothesis testing")

Why is this of interest?

→ For theoretical reasons it is desirable to use the F-test whenever more than one X-variable is tested

→ This means the F-test can be used to check the result of repeated t-testing

Example of a multiple hypothesis test:

H_0: $B_1 = 0$ and $B_2 = 0$

H_A: One or more of the claims i H_0 are wrong

Ingredients of the F-test:

→ F-distribution

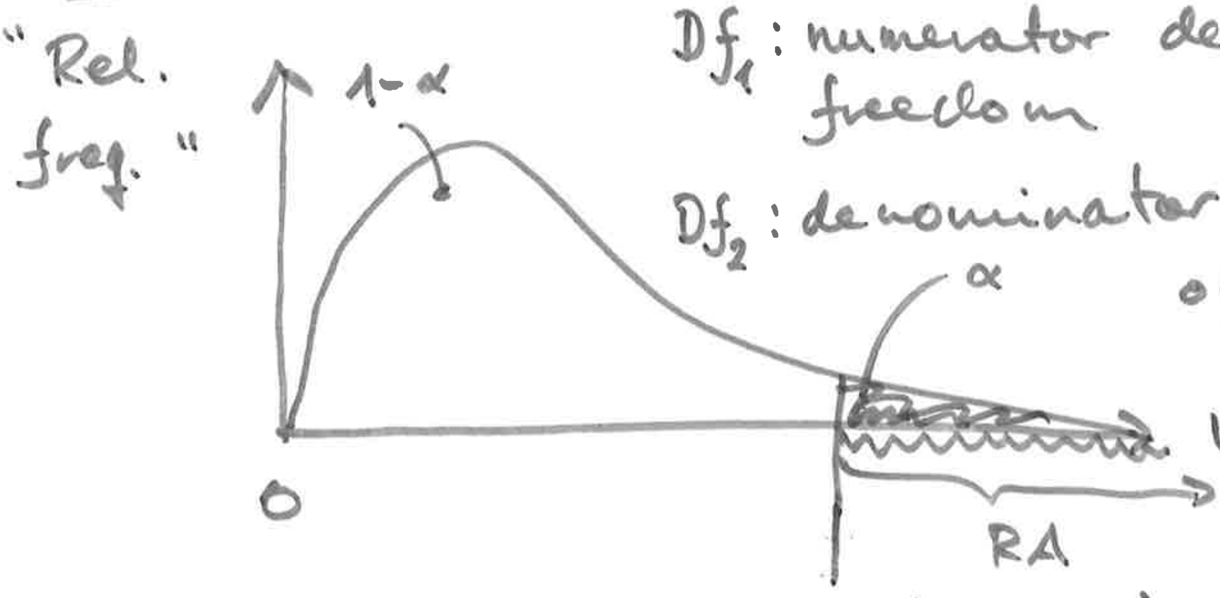
→ The unrestricted model (the H_A model)

→ The restricted model (the H_0 model)

The F-distribution:

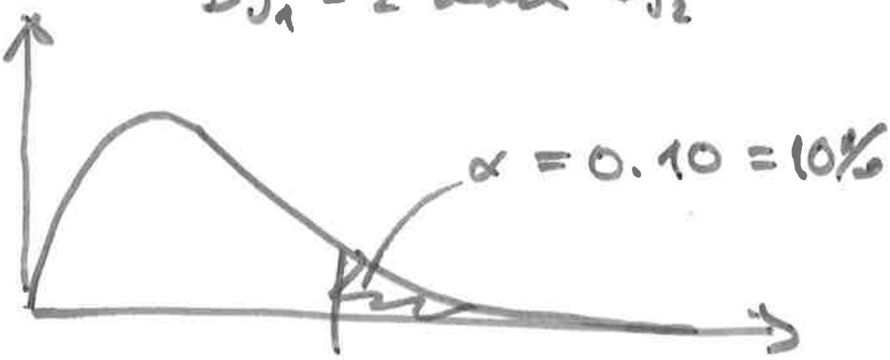
Df_1: numerator degrees of freedom

Df_2: denominator degrees of freedom



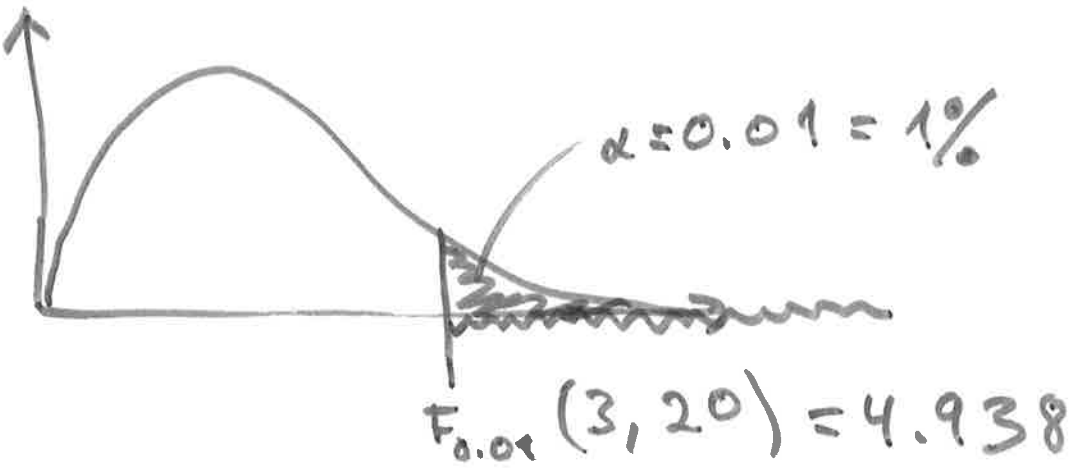
$F_\alpha(Df_1, Df_2)$: Critical value

Example: $\alpha = 10\%$, $Df_1 = 2$ and $Df_2 = 28$

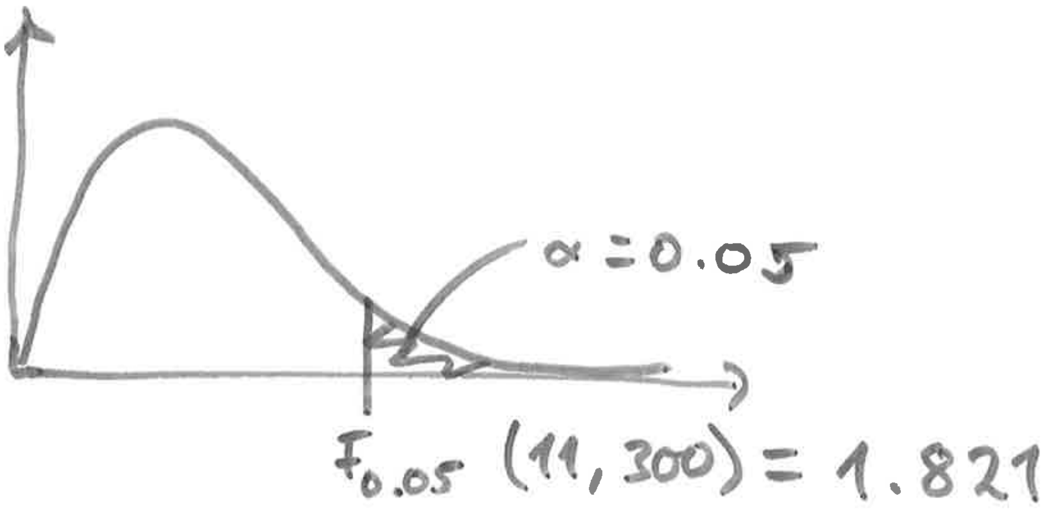


$F_{0.10}(2, 28) = 2.503$

Example: $\alpha = 1\%$ $Df_1 = 3$ and $Df_2 = 20$



Example: $\alpha = 5\%$, $Df_1 = 11$ and $Df_2 = 300$



Models without and with restrictions

→ Unrestricted model (ur): All the Bs in question are feely estimated

→ Restricted model (r): A model in which the values are set or restricted to those in H_0

Example: Consider

$$Y = B_0 + B_1 \text{exper} + B_2 \text{educ} + u \qquad (1)$$

$$\underset{-9.59}{\uparrow} \qquad \underset{0.98}{\uparrow} \qquad \underset{1.42}{\uparrow}$$

If we instead set:

$$B_1 = 0 \text{ and } B_2 = 0$$

and estimate

$$Y = B_0 + u \qquad (2)$$

then (1) can be viewed as ur-model and (2) can be viewed as the r-model

→ The restricted model is associated with H_0

→ The unrestricted model is associated with H_A (in a sense)

Recipe for testing several βs:

Step 1: Choose α, formulate H_0 and H_A:

Example:

H_0: $\beta_0 = 0$ and $\beta_1 = 0$ and $\beta_2 = 1$

H_A: One or more of the claims in H_0 are wrong

2: Identify the rejection area using an F-distribution:

Df_1 = no. of claims ("=") in H_0

Df_2 = n − the number of βs in unr-model

3: Test-value:

$$\frac{(R^2_{ur} - R^2_r)/Df_1}{(1 - R^2_{ur})/Df_2}$$

See 4c) in Ex. set 3 $\left\{\right.$ Note: If the left-hand sides of the ur and r models differ, then it is necessary to use the RSS-version of the test expression

4: Conclusion: Reject H_0 if test value lies in RA

Example (wage data):

ur-model:

$$Y = B_0 + B_1 exper + B_2 educ + u$$

$$R^2_{ur} = 0.276$$

Does experience or education or both have an effect on wage?

Step 1: $\alpha = 0.05$

H_0: $B_1 = 0$ and $B_2 = 0$

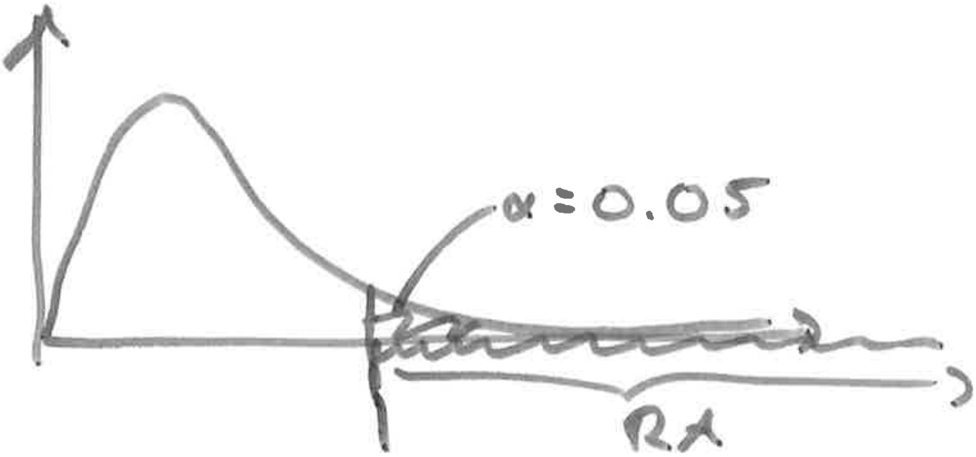
H_A: One or both claims in H_0 are wrong

Restricted model:

$Y = B_0 + u \qquad R_r^2 = 0$

2: Rejection area:

$Df_1 = 2 \qquad Df_2 = n - k = 1286$



$F_{0.05}(2, 1286) \approx \underbrace{F_{0.05}(2, 1000)}_{3.005}$

3: Testvalue:

0 2

0.276 $\dfrac{(R_{ur}^2 - R_r^2)/Df_1}{(1 - R_{ur}^2)/Df_2} = \underline{245.12}$

1286

4: Conclusion: We reject H_0. That is, either exper or educ or both have an effect on wage

7 Suggested exercises

Ex. set 2: Simple regression

- 2a)-k), 3a), c)

Ex. set 3: Multiple regression

- 1a), b), d), 2, 4a), b), d)