ECON-C4100 - Capstone: Econometrics I

Lecture 3: Univariate regression

Otto Toivanen

Learning outcomes

- At the end of lectures 3 - 5, you

1 understand what one learns from a (univariate) regression analysis.

2 understand how to carry out a regression analysis.

3 appreciate the assumptions made in standard regression analysis.

4 are aware of the most common pitfalls in regression analysis.

The effect of X on Y

- At the end of lectures 3 - 5, you have an idea how to approach answering question such as the following:
- Does having a PhD (in science) help to innovate?
- Is website design A better than design B in terms of sales? By how much?
- Are branded pharmaceuticals more expensive than generic products?
- Are promotions of substitute products of the same firm at the same time effective?

Modeling

- Q1: what is the object you want to model ("explain")?
- Let's call this Y.
- Q2: what is the object whose effect on Y you want to understand?
- Let's call this X.

Modeling

- Where do these (decisions) come from?
- Theory.
- What is theory?
    - Mathematical model.
    - Conseptualization of existing **qualitative** knowledge.
    - Conseptualization of existing **quantitative** knowledge.

Let's look at the relationship between income and age

- Variables
  1. *income* = income in euros
  2. *age* = age in years
- We use the same FLEED data as in lecture 2, i.e., it comes from one year.
- These data are an example of **cross-section** data where each observation unit is observed only once and there is no (meaningful) time (second) dimension to the data besides the individuals.

Descriptive statistics

Descriptive statistics			
variable	mean	sd	median
income	23 297	17 163	21 000
age	41.87	16.29	43

- For brevity, I do not show conditional descriptive statistics as we have already seen them in lecture 2.

Modeling the relationship between *income* and *age*

$$Y = f(X) \tag{1}$$

- What do we know about $f(X)$?
- How can we learn about it?

Quick aside - correlation

$$corr(Y, X) = \frac{cov(Y, X)}{\sqrt{var(X)}\sqrt{var(Y)}} \tag{2}$$

More structure - linear

$$Y = \beta_0 + \beta_1 X \tag{3}$$

- This is the so called population regression line. (**populaatio regressio**).
- Y is called the **dependent variable** or **endogenous variable** (**vastemuuttuja**).
- X is called the **independent** or the **exogenous variable** or **regressor** (**selittävä muuttuja**).
- β_0, β_1 are the **parameters** of the model (**(malli)parametrit**).

More structure - linear

$$Y = \beta_0 + \beta_1 X \tag{4}$$

- $\beta_0, \ \beta_1$ interpretation?
- Intercept, slope.
- What is now assumed about what can influence Y?

How to allow for other factors?

$$Y = f(X, u) = \beta_0 + \beta_1 X + u \tag{5}$$

- u is called the **error term** or **residual** (**virhetermi**). Why such a name?

How to allow for other factors?

$$Y = f(X, u) = \beta_0 + \beta_1 X + u \tag{5}$$

- u is called the **error term** or **residual** (**virhetermi**). Why such a name?

1. It shows how much our model misses in terms of determining Y.

How to allow for other factors?

$$Y = f(X, u) = \beta_0 + \beta_1 X + u \tag{5}$$

- u is called the **error term** or **residual** (**virhetermi**). Why such a name?

1. It shows how much our model misses in terms of determining Y.
2. It measures those things that 1) affect Y and 2) we don't observe.

What is known about u?

- How large should the error be on average?
- Zero. Why?

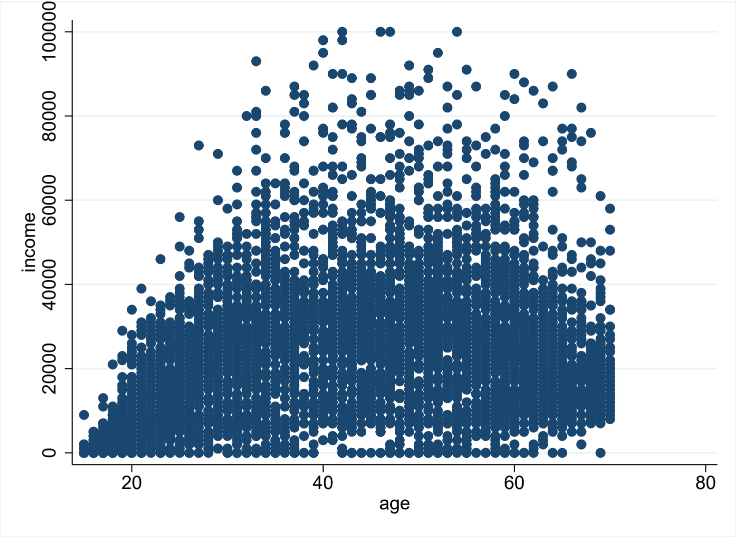
  $\rightarrow E[u|X] = 0$

How to get β_0, β_1?

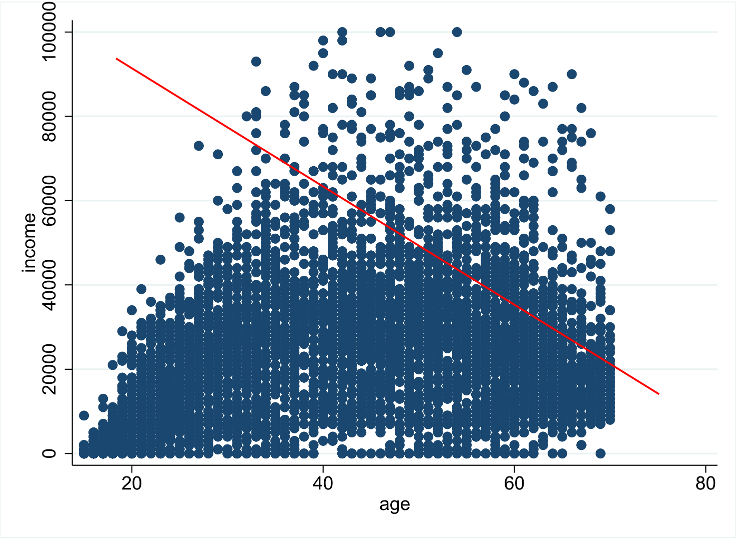
Stata code

```
scatter income age if year == 15 & income != . , ///
  xtitle("age") ///
  ytitle("income") ///
  graphregion(fcolor(white))
```

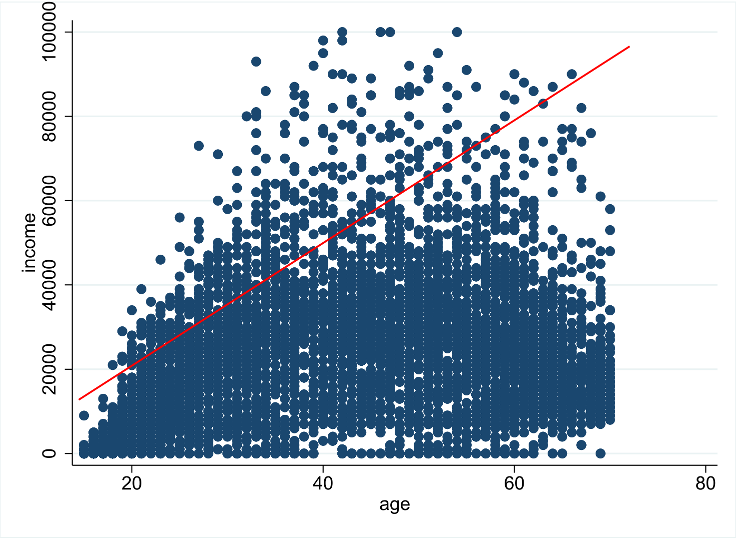
How to get β_0, β_1?



How to get β_0, β_1?



How to get β_0, β_1?



How to get β_0, β_1: OLS

- Ordinary Least Squares.

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{pmatrix}, \quad \boldsymbol{U} = \begin{pmatrix} u_1 \\ u_2 \\ . \\ . \\ . \\ u_n \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ . & \\ . & \\ . & \\ 1 & X_n \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1' \\ \boldsymbol{X}_2' \\ . \\ . \\ . \\ \boldsymbol{X}_n' \end{pmatrix} , \tag{6}$$

and $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$.

How to get β_0, β_1: OLS

- Ordinary Least Squares.

$$Y_i = \beta_0 + \beta_1 X_i + u = \boldsymbol{X_i'}\boldsymbol{\beta} + u_i \tag{7}$$

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U} \tag{8}$$

$$\mathbb{E}[Y - (\beta_0 + \beta_1 X)] = \mathbb{E}[u|X] = \mathbb{E}[Y - \boldsymbol{X_i'}\boldsymbol{\beta}] = 0 \tag{9}$$

$$\mathbb{E}[\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}] = \mathbb{E}[\boldsymbol{U}|\boldsymbol{X}] = 0 \tag{10}$$

How to get β_0, β_1: OLS

$$min_{\beta_0,\beta_1} \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_i)]^2 \tag{11}$$

$$min_{\boldsymbol{\beta}} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})' (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{12}$$

- Suggestion: Do the derivation w/out using matrix algebra. It helps you understand the formula for β_1.

How to get β_0, β_1: OLS

- Notice link to estimation of mean and set $\beta_1 = 0$.

$$\sum_{i=1}^{n} [Y - (\beta_0)]^2 \tag{13}$$

- Now $\beta_0 = m = \mathbb{E}[\mu_Y]$.

How to get β_0, β_1: OLS

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} XY - \bar{X}\bar{Y}}{\frac{1}{n}\sum_{i=1}^{n} XX - \bar{X}\bar{X}} = \frac{cov(Y,X)}{var(X)} = \frac{cov(Y,X)}{\sqrt{var(X)}\sqrt{var(X)}} \tag{14}$$

Note: compare to the formula for correlation.

$$\hat{\beta}_0 = \bar{Y} - \frac{\frac{1}{n}\sum_{i=1}^{n} XY - \bar{X}\bar{Y}}{\frac{1}{n}\sum_{i=1}^{n} XX - \bar{X}\bar{X}}\bar{X} = \bar{Y} - \hat{\beta}_1\bar{X} \tag{15}$$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y}) \tag{16}$$

How to get β_0, β_1: OLS

Predicted value (ennuste): $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ or $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{\beta}} \boldsymbol{X}$.

Prediction error (ennustevirhe): $\hat{u}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ or $\hat{\boldsymbol{U}} = \boldsymbol{Y} - \hat{\boldsymbol{\beta}} \boldsymbol{X}$.

Back to income and age...

- So let's run the regression:

Stata code

```
label var age "Age"
reg income age if year == 15 & income != .
estimates store lin_est
estimates table lin_est , b(%7.3f) se(%7.3f) p(%7.3f) stat(r2)
esttab , scalar(F) r2 label ///
  title(Regression of income on age) ///
  nonumbers mtitles("Model A") ///
  addnote("Data: teaching FLEED, Statistics Finland")
esttab using income_age.tex , scalar(F) r2 label replace booktabs ///
   alignment(D{.}{.}{-1}) width(0.8\hsize)        ///
   title(Income and age\label{tab1})
```

Regular Stata output table

```
. reg income age if year == 15 & income != .

      Source |       SS           df       MS      Number of obs   =     5,973
-------------+----------------------------------   F(1, 5971)      =    493.91
       Model |  1.3441e+11         1  1.3441e+11   Prob > F        =    0.0000
    Residual |  1.6249e+12     5,971   272128687   R-squared       =    0.0764
-------------+----------------------------------   Adj R-squared   =    0.0762
       Total |  1.7593e+12     5,972   294589468   Root MSE        =     16496

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   296.7539   13.35276    22.22   0.000     270.5776    322.9301
       _cons |    10654.7   607.5672    17.54   0.000     9463.644    11845.75
------------------------------------------------------------------------------
```

Coefficients / economic significance

```
. reg income age if year == 15 & income != .

      Source |       SS           df       MS      Number of obs   =     5,973
-------------+----------------------------------   F(1, 5971)      =    493.91
       Model |  1.3441e+11         1  1.3441e+11   Prob > F        =    0.0000
    Residual |  1.6249e+12     5,971   272128687   R-squared       =    0.0764
-------------+----------------------------------   Adj R-squared   =    0.0762
       Total |  1.7593e+12     5,972   294589468   Root MSE        =     16496

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   296.7539   13.35276    22.22   0.000     270.5776    322.9301
       _cons |    10654.7   607.5672    17.54   0.000     9463.644    11845.75
------------------------------------------------------------------------------
```

Standard errors etc., statistical significance of individual parameters

```
. reg income age if year == 15 & income != .

      Source |       SS           df       MS      Number of obs   =     5,973
-------------+----------------------------------   F(1, 5971)      =    493.91
       Model |  1.3441e+11         1  1.3441e+11   Prob > F        =    0.0000
    Residual |  1.6249e+12     5,971   272128687   R-squared       =    0.0764
-------------+----------------------------------   Adj R-squared   =    0.0762
       Total |  1.7593e+12     5,972   294589468   Root MSE        =     16496

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   296.7539   13.35276    22.22   0.000     270.5776    322.9301
       _cons |    10654.7   607.5672    17.54   0.000     9463.644    11845.75
------------------------------------------------------------------------------
```

Regression level statistical measures

```
. reg income age if year == 15 & income != .

      Source |       SS           df       MS      Number of obs   =     5,973
-------------+----------------------------------   F(1, 5971)      =    493.91
       Model |  1.3441e+11         1  1.3441e+11   Prob > F        =    0.0000
    Residual |  1.6249e+12     5,971   272128687   R-squared       =    0.0764
-------------+----------------------------------   Adj R-squared   =    0.0762
       Total |  1.7593e+12     5,972   294589468   Root MSE        =     16496

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   296.7539   13.35276    22.22   0.000     270.5776    322.9301
       _cons |    10654.7   607.5672    17.54   0.000     9463.644    11845.75
------------------------------------------------------------------------------
```

A formatted version with the requested information only

```
. estimates store lin_est

. estimates table lin_est, b(%7.3f) se(%7.3f) p(%7.3f) stat(r2)

--------------------------
    Variable |  lin_est
-------------+------------
         age |  296.754
             |   13.353
             |    0.000
       _cons |  1.1e+04
             |  607.567
             |    0.000
-------------+------------
          r2 |    0.076
--------------------------
                legend: b/se/p
```

A LaTeX version of the same table

Table: Income and age

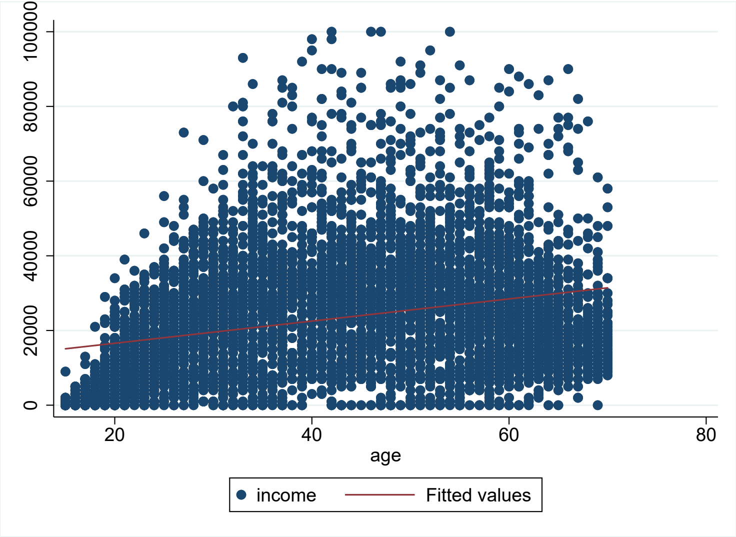
	(1) income
Age	296.8***
	(22.22)
Constant	10654.7***
	(17.54)
Observations	5973
R^2	0.076
F	493.9

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

What are these numbers?

- What do β_0 and β_1 mean?



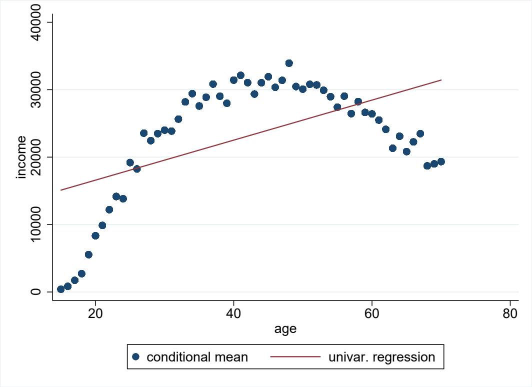
What are these numbers?

- $\beta_0 =$ the intercept.
- $\beta_1 =$ the slope of the regression line.

$$\mathbb{E}[Y|X = x] = \beta_0 + \beta_1 x \tag{17}$$

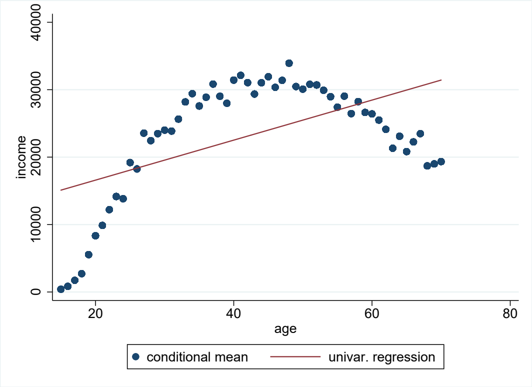
- Regression allows you to study the (changes in) the **conditional mean**.
- Thus, β_1 is the derivative of Y wrt. X.

What are these numbers?



- Why are the two conditional mean presentations in the figure different?

What are these numbers?



- Why are the two conditional mean presentations in the figure different?
- The regression "forces" the relationship to be linear, i.e., we chose the relationship to be linear.

What are these numbers?

- How good is the model's fit? How much does it explain?
- Of what..? Answer: of the variation in Y.

Explained sum of squares (ESS): $\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$.

Total sum of squares (TSS): $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$.

Residual sum of squares (RSS): $\sum_{i=1}^{n}(u_i)^2$.

What are these numbers?

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \in [0, 1] \tag{18}$$

- R^2 "close to one" = "almost all" variation in Y captured by the model (= variation in X).
- R^2 "close to zero" = "almost no" variation in Y captured by the model (= variation in X).
- Note #1: R^2 has not effect on the interpretation of β.
- Note #2: R^2 will have an effect on whether we reject the model or not, on statistical grounds.