ECON-C4100 - Capstone: Econometrics I

Lecture 6: Multiple regression #1: estimation

Otto Toivanen

Learning outcomes

- At the end of this lectures, you

1 understand what how multivariate regression differs from univariate regression.

2 understand how and why to carry out a multivariate regression analysis.

3 appreciate the assumptions made in multivariate regression analysis.

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

Starting point:

$$Y = f(X_1, X_2, ..., X_k, u)$$

Outcome variable of interest a function of several variables.Observables and unobservables.One or more hypotheses

(needed)?

Income, age and gender

1. Is income affected by age?
2. Do women and men of same age earn differently?

- Let's study these using the open access FLEED data of Statistics Finland.
- These data can be downloaded from the Statistics Finland web page.
- We will use the year 15 (= 2010) cross section data.

Univariate regression

$$Y = f(X, u) = \beta_0 + \beta_1 X + U$$

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

Multivariate/multiple regression

$$Y = f(X_1, X_2, u) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

$$E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

More structure - linear

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

- This is the so called **population regression line (populaatio regressio)**.
- Y = **dependent variable (vastemuuttuja)** or **endogenous variable**.
- X_k = **independent variable** k (selittävä muuttuja) or **exogenous variable** k or *regressor* k.
- $\beta_0, \beta_1, \beta_2$: **parameters of the model**.

Are all Xs born equal?

- Depends...
- Treatment variable = the one of primary interest.
- Control variable(s) = affect(s) Y but we are not (so much) interested in this/these.
- Why include variables that we are not interested in?

What type of control variables matter, how & why?

1. $cov(X_1, X_2) = 0$
2. $cov(X_1, X_2) \neq 0$

- Key is whether the treatment variable and control variable are correlated or not.

What type of control variables matter, how & why?

- Why is this key? Recall

$$\hat{\beta}_1 = \beta_1 + \rho_{xu}\frac{\sigma_u}{\sigma_x}. \tag{1}$$

Rewrite

$$u = \beta_2 X_2 + v \tag{2}$$

What type of control variables matter, how & why?

Assume

$$cov(\boldsymbol{X}, v) = 0$$

Then

$$\hat{\beta}_1 = \beta_1 + \beta_2 \rho_{X_1 X_2} \times \frac{\sigma_{X_2}}{\sigma_{X_1}} \tag{3}$$

Make sure you know how to derive equation (3).

What type of control variables matter, how & why?

If the Xs are correlated, then the bias in β_1 depends on

1. the impact of X_2 on Y.
2. the correlation between the Xs.
3. how much variance X_2 has relative to X_1.

What type of control variables matter, how & why?

- So are we home if $cov(X_1, X_2) = 0$?
- Yes and no.
- If $cov(X_1, X_2) = 0$ then $\hat{\beta}_1 = \beta_1$
- However, adding X_2 decreases the standard error / increases the precision of $\hat{\beta}_1$.

What type of control variables matter, how & why?

- A two-variable model (App 6.2. in S&W).
- 2 explanatory variables and homosc. errors, $\rho_{X_1,X_2} = 0$. Then

$$\sigma^2_{\beta_1} = \frac{1}{n}\frac{\sigma^2_u}{\sigma^2_{X_1}}$$

  which is the variance of $\hat{\beta}_1$.
- Adding X_2 necessarily decreases σ^2_u.

Income modeled as function of age and gender?

- Let's look at the following model:

$$Income_i = \beta_0 + \beta_{AgeMV} Age_i + \beta_{GMV} G_i + u_{MVi}$$

Where

$Age_i =$ age in years.

$G_i =$ dummy for gender.

MV stands for **M**ulti**v**ariate.

Should we suspect that age affects income?

- Experience increases with age.
- In a cross-section such as ours, younger people typically better educated than older (conditional on not being too young).
- Physical condition and mental agility start to decrease relatively early.

Should we suspect that gender affects income?

- Segregation of job market a well known phenomenon.
- Women bear a larger share of household work and stay longer at home after getting a child.
- Educational levels and fields differ by gender.

Some conditional descriptive statistics

```
. tabstat income age  if year == 15, stat(mean sd p50) by(gender)

Summary statistics: mean, sd, p50
  by categories of: gender
```

gender	income	age
0	25478.2	41.5928
	18894.06	16.10072
	24000	42
1	21053.65	42.15153
	14852.83	16.48018
	20000	43
Total	23296.67	41.86563
	17163.61	16.28821
	21000	43

Are age and gender correlated in our data?

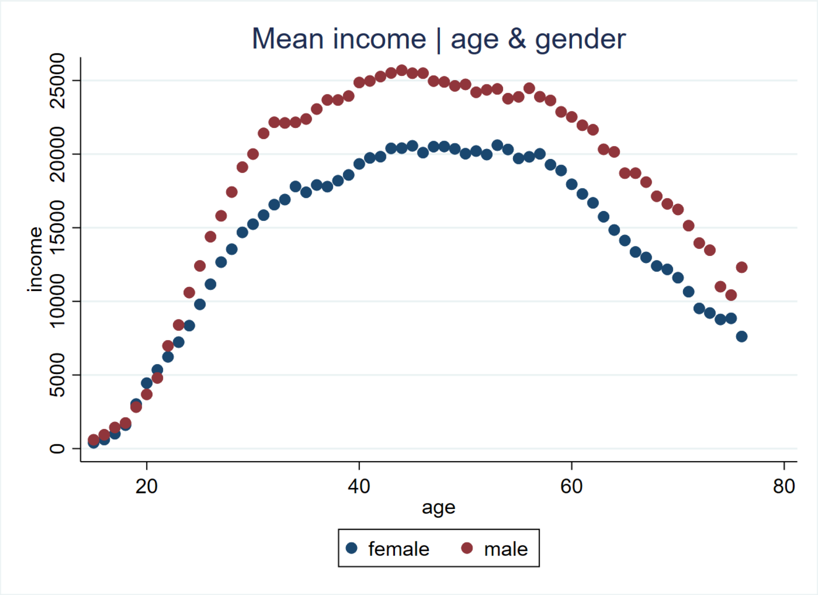
```
. pwcorr age gender if year == 15, sig

             |      age   gender
-------------+------------------
         age |   1.0000 
             |
             |
      gender |   0.0171   1.0000 
             |   0.1755
```

Mean income conditional on age and gender

Stata code

```
bysort age gender: egen income_m_age_g = mean(income)
bysort age gender: gen win_age_g_ind  = _n
twoway  scatter income_m_age_g age if gender == 1 & win_age_g_ind == 1 || ///
    scatter income_m_age_g age if gender == 0 & win_age_g_ind == 1, ///
    legend(lab (1 "female") lab (2 "male")) ///
    graphregion(fcolor(white)) ///
    ytitle("income") ///
    title("Mean income | age & gender")
graph export "mean_income_age_gender.png", replace
```

Mean income conditional on age and gender

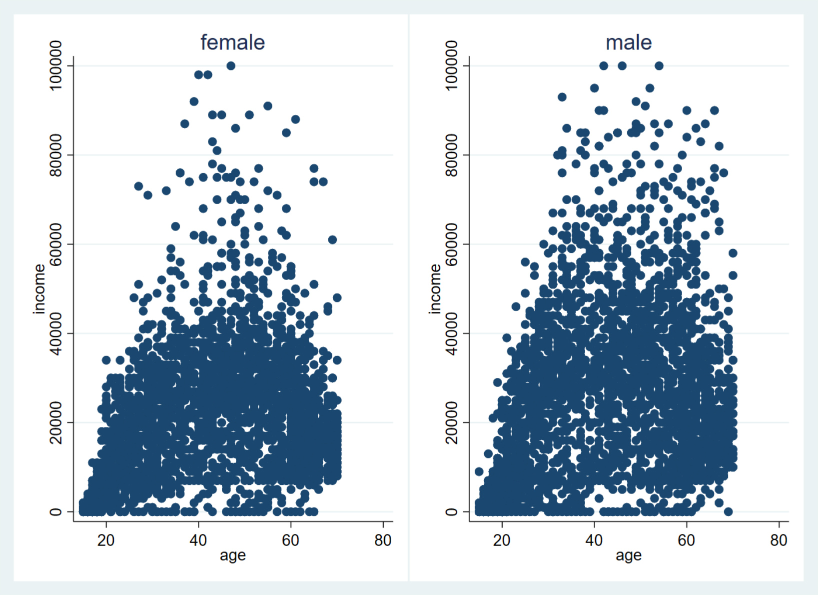


Income-age scatter by gender

Stata code

```
scatter income age if gender == 0 & year == 15, ///
  graphregion(fcolor(white)) ///
  title("male") ///
  saving(income_age_male, replace)
scatter income age if gender == 1 & year == 15, ///
  graphregion(fcolor(white)) ///
  title("female") ///
  saving(income_age_female, replace)
gr combine income_age_female.gph income_age_male.gph
graph export "income_age_gender.png", replace
```

Income-age scatter by gender



How to get $\beta_0, \ \beta_1 \& \beta_2$: OLS

$$min_{\beta_0,\beta_1,\beta_2} \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i})]^2 \tag{4}$$

$$min_{\boldsymbol{\beta}} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^{'} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{5}$$

How to get β_0, β_1&β_2: OLS

$$\hat{\beta}_1 = \frac{\sum x_2^2 \sum x_2 y - \sum x_1 x_2 \sum x_1 y}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2}$$

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

- Note: now not using matrix algebra leads to very cumbersome mathematics; with matrix algebra, the solution stays the same as with univariate regression.
- The expression for $\hat{\beta}_2$ is symmetric with that of $\hat{\beta}_1$.
- Finally, $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$.

Let's compare univariate regressions to multivariate regression

$$Income = \beta_{0AgeUV} + \beta_{AgeUV} Age + u_{AgeUV} \quad (6)$$

$$Income = \beta_{0GUV} + \beta_{GUV} G + u_{GUV} \quad (7)$$

$$Income = \beta_0 + \beta_{AgeMV} Age + \beta_{GMV} G + u_{MV} \quad (8)$$

Univariate regressions

Table: Univariate income regressions

	(1) income	(2) income
Age	296.8***	
	(13.35)	
Gender		-4424.6***
		(440.5)
Constant	10654.7***	25478.2***
	(607.6)	(309.3)
Observations	5973	5973
r2	0.0764	0.0166
F	493.9	100.9

Standard errors in parentheses

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

Multivariate regression

Table: Income regressions

	(1) income	(2) income	(3) income
Age	296.8***		298.5***
	(13.35)		(13.23)
Gender		-4424.6***	-4545.0***
		(440.5)	(422.9)
Constant	10654.7***	25478.2***	12819.2***
	(607.6)	(309.3)	(634.6)
Observations	5973	5973	5973
r2	0.0764	0.0166	0.0939
F	493.9	100.9	309.4

Standard errors in parentheses

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

Issues?

1. How do the individual coefficients compare to univariate results?
2. What explains the difference(s)?
3. What about statistical significance of individual coefficients?
4. What about several / all coefficients?
5. What about R^2?

Issues?

6. What is the interpretation of individual coefficients?
7. (under what assumptions) does OLS work?
8. How to choose which explanatory variables to include / exclude?
9. What if the world is more complicated than linear?
10. What all can go wrong, and how would I know / find out?

Q1 & Q2 multivariate vs. univariate?

1. How do the individual coefficients compare to univariate results?
2. What explains the difference(s)?

Q1 & Q2 multivariate vs. univariate?

- Compare the *Age* coefficient in the univariate to that in the multivariate regression.
- What can you conclude? Recall

$$\hat{\beta}_{AgeUV} = \beta_{AgeMV} + \beta_G MV \rho_{Age,G} \times \frac{\sigma_G}{\sigma_{Age}}$$

Multivariate regression

```
. estout income_age income_gender income_age_gender, cells(b(star fmt(3)) se(par fmt(3)))
```

	income_age b/se	income_gen~r b/se	income_age~r b/se
age	296.754***		298.548***
	(13.353)		(13.228)
gender		-4424.553***	-4545.022***
		(440.537)	(422.935)
_cons	10654.695***	25478.203***	12819.203***
	(607.567)	(309.335)	(634.636)

```
. pwcorr age gender if e(sample)
```

	age	gender
age	1.0000	
gender	0.0126	1.0000

```
. sum age gender if e(sample)
```

Variable	Obs	Mean	Std. Dev.	Min	Max
age	5,973	42.60087	15.9866	15	70
gender	5,973	.4930521	.4999936	0	1

Q1 & Q2 multivariate vs. univariate?

- Plug numbers from the previous slide into the bias formula:

$$Bias_{\beta_{ageUV}} = \beta_{GMV}\rho_{Age,G} \times \frac{\sigma_G}{\sigma_{Age}}$$
$$= -4545.02 \times 0.0126 \times \frac{0.500}{15.987} = -1.791$$

- Compare to

$$\beta_{AgeUV} - \beta_{AgeMV} = 296.754 - 298.548 = -1.794$$

- Do the same for gender.
- What can you conclude?

Q1 & Q2 multivariate vs. univariate?

- Multivariate regression allows the researcher to
  1. control for observable variables and thereby either remove (omitted variable) bias and/or increase efficiency.
  2. test several hypotheses simultaneously.
  3. (as we will see), enrich the main hypotheses to allow for heterogenous effects.

Q3 & Q4 statistical significance, individual coefficients

3 What about statistical significance of individual coefficients?

4 What about the statistical significance of several / all coefficients?

Q3 & Q4 statistical significance, individual coefficients

- Can we reject the null that
  1. $\beta_0 = 0$,
  2. $\beta_{Age} = 0$,
  3. $\beta_G = 0$?

Q3 & Q4 statistical significance, individual coefficients

```
. regr income age gender          if year == 15
```

Source	SS	df	MS
Model	1.6524e+11	2	8.2622e+10
Residual	1.5940e+12	5,970	267009188
Total	1.7593e+12	5,972	294589468

Number of obs	=	5,973
F(2, 5970)	=	309.43
Prob > F	=	0.0000
R-squared	=	0.0939
Adj R-squared	=	0.0936
Root MSE	=	16340

income	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	298.5478	13.22762	22.57	0.000	272.6169	324.4787
gender	-4545.022	422.9347	-10.75	0.000	-5374.127	-3715.917
_cons	12819.2	634.6356	20.20	0.000	11575.09	14063.32

Q3 & Q4 statistical significance, individual coefficients

- Are $\beta_0, \beta_{Age}, \beta_g$ all $= 0$?
- F - test (and others) for the *joint significance.*
- Cannot do this by looking at individual (t-) tests.
- Reason: two or more random variables $\rightarrow$ need their joint distribution.

Q3 & Q4 statistical significance, individual coefficients

- F test (under homosc.). For illustration only.

$$F = \frac{(SSR_{restricted} - SSR_{unrestricted})/q}{SSR_{unrestricted}/(n - k_{restricted} - 1)}$$

$$= \frac{(R^2_{restricted} - R^2_{unrestricted})/q}{(1 - R^2_{unrestricted})/(n - k_{restricted} - 1)}$$

- Modern software calculate the heterosc. robust F-test.

Q3 & Q4 statistical significance, individual coefficients

```
. regr income age gender          if year == 15

      Source |       SS           df       MS      Number of obs   =     5,973
-------------+----------------------------------   F(2, 5970)      =    309.43
       Model |  1.6524e+11         2  8.2622e+10   Prob > F        =    0.0000
    Residual |  1.5940e+12     5,970   267009188   R-squared       =    0.0939
-------------+----------------------------------   Adj R-squared   =    0.0936
       Total |  1.7593e+12     5,972   294589468   Root MSE        =     16340

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   298.5478   13.22762    22.57   0.000     272.6169    324.4787
      gender |  -4545.022   422.9347   -10.75   0.000    -5374.127   -3715.917
       _cons |    12819.2   634.6356    20.20   0.000     11575.09    14063.32
------------------------------------------------------------------------------
```

Q3 & Q4 statistical significance, individual coefficients

- What about $\beta_{Age} = \beta_G = 0$?
- In other words, Null hypothesis is that a subset of parameters are zero.
- Modern software allow this.

Q3 & Q4 statistical significance, individual coefficients

```
. testparm age gender

 ( 1)  age = 0
 ( 2)  gender = 0

       F(  2,  5970) =  309.43
            Prob > F =    0.0000
```

Q3 & Q4 statistical significance, individual coefficients

- What about $\beta_{Age} = \beta_G$?
- Need either a direct test modern software allow this (easily).
- Or a trick (add and substract).

Q3 & Q4 statistical significance, individual coefficients

```
. test age  = gender

 ( 1)  age - gender = 0

       F(  1,  5970) =  130.92
            Prob > F =    0.0000
```

Q3 & Q4 statistical significance, individual coefficients

- With multivariate regression:
  1. Important to check the regression diagnostic statistics (F-test) (more on this to follow).
  2. Rich possibilities to test hypotheses that involve multiple parameters.

Q5 What about R^2?

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

- R^2 increases (almost) surely as you add explanatory variables.
- Adjusted R^2 corrects for this:

$$adjR^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS} = \frac{s_{\hat{u}^2}}{s_Y^2}$$

- n = number of obs; k = number of expl. variables.
- Adjusted R^2 always lower than R^2.

Q5 What about R^2?

```
. regr income age gender         if year == 15

      Source |       SS           df       MS      Number of obs   =     5,973
-------------+----------------------------------   F(2, 5970)      =    309.43
       Model |  1.6524e+11         2  8.2622e+10   Prob > F        =    0.0000
    Residual |  1.5940e+12     5,970   267009188   R-squared       =    0.0939
-------------+----------------------------------   Adj R-squared   =    0.0936
       Total |  1.7593e+12     5,972   294589468   Root MSE        =     16340

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   298.5478   13.22762    22.57   0.000     272.6169    324.4787
      gender |  -4545.022   422.9347   -10.75   0.000    -5374.127   -3715.917
       _cons |    12819.2   634.6356    20.20   0.000     11575.09    14063.32
------------------------------------------------------------------------------
```

Q5 What about R^2?

- High R^2 / an increases in R^2 says nothing about causality.
- High R^2 does not mean your model does not suffer from omitted variable bias.
- High R^2 does not mean you have the right set of explanatory variables.
- High R^2 tells nothing about the economic significance of your results.
- High R^2 means that factors outside your model (= the stuff going into the error term) play a relatively speaking smaller role in the process that determines the value of Y.
- But, as we saw from the F-test formual, (changes in) R^2 are indicative and a certain level of R^2 is needed to reject the Null that all your model parameters are insignificant.