ECON-C4100 - Capstone: Econometrics I

Lecture 7: Multiple regression #2: estimation

Otto Toivanen

Learning outcomes

- At the end of this lecture, 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.

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?

Q6 What is the interpretation of individual coefficients?

- Our estimation equation is:

$$Income = \beta_0 + \beta_{AgeMV} Age + \beta_{GMV} G + u_{MV} \tag{1}$$

Interpretation of individual coefficients

- Regression yields the **conditional expectation** of the dependent variable Y:

$$\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... \tag{2}$$

$$\mathbb{E}[Income_i|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_{AgeMV} Age_i + \beta_{GMV} G_i \tag{3}$$

- By plugging into the regression those values of $\boldsymbol{X}$ that we are interested we get the conditional expectation of Y.

Interpretation of individual coefficients

- Example: the expected income of a woman of 35 years of age is:

$$\mathbb{E}[Income_i|\boldsymbol{X} = 35, 1] =$$
$$12819.2 + 298.55 \times 35 - 4545.02 \times 1 = 18723.4$$

- Example: the expected income of a man of 55 years of age is:

$$\mathbb{E}[Income_i|\boldsymbol{X} = 55, 0] =$$
$$12819.2 + 298.55 \times 55 - 4545.02 \times 0 = 29239.45$$

Interpretation of individual coefficients

- Coefficients as partial derivatives:

$$\frac{\partial \mathbb{E}[Income_i | \boldsymbol{X}]}{\partial X_k} = \beta_k \tag{4}$$

- With discrete explanatory variables cannot take derivatives, so a coefficient measures the change in Y from a one unit change in X_k:

$$\beta_k = \mathbb{E}[Income_i | \boldsymbol{X}, X_k = m] - \mathbb{E}[Income_i | \boldsymbol{X}, X_k = m - 1] \tag{5}$$

- Notice that in both, we fix all other variables (their effect on Y).
- So β_G is the effect of gender on income, **keeping the effect of** *Age* **constant**.

Q7 What are the OLS assumptions?

1. Strict exogeneity: $\mathbb{E}(u|\boldsymbol{X}) = 0$.
2. $(\boldsymbol{X_i}, Y),\ i = 1, ..., n$ are independent and identically distributed across observations.
3. $\boldsymbol{X_i}$ and Y_i (u_i) have finite *fourth* moments.
4. No **perfect multicollinearity** ($\boldsymbol{X}$ has **full column rank**).
5. Auxiliary: u_i is homoscedastic.

Perfect multicollinearity

- Analog: To solve a system of equations, you need as many equations as you have unknowns.
- Two variables are perfectly (multi)collinear if one is a perfect linear function of the other.
- Example: Think of a phenomenon with two **mutually exclusive and exhaustive** outcomes, A and B.
- A dummy taking value 1 if A is true and 0 otherwise: $D_A = 1 - D_B$, where D_B is the dummy taking value 1 if B is true and 0 otherwise ("**dummy variable trap**)".
- Perfect collinearity = correlation $+/-1$.

Collinearity of two variables

- Recall from previous lecture the two-variable model (App 6.2. in S&W).
- 2 explanatory variables and homosc. errors, $\rho_{X_1,X_2} \neq 0$. Then

$$\sigma^2_{\beta_1} = \frac{1}{n}\frac{1}{1-\rho_{X_1,X_2}}\frac{\sigma^2_u}{\sigma^2_{X_1}}$$

Collinearity of two variables

- Notice what happens when $\rho_{X_1,X_2} \to 1$.
- Collinearity ($=$ "high" ρ_{X_1,X_2}) increases the standard error(s) of the other coefficient(s).
- Correlation between the $\boldsymbol{X}$s is a two-edged sword:
  1. It removes omitted variable bias.
  2. It reduces the efficiency gains from introducing a further explanatory variable.

Collinearity vs. multicollinearity

- Collinearity refers to the (high) correlation between two variables.
- **Multi**collinearity is a characteristic of a matrix (vector) $\boldsymbol{X}$.
- while the pair-wise correlations between elements of $\boldsymbol{X}$ may be "not so high", the aggregate effect of them may lead to inflated standard errors.

Q8 How to choose the explanatory variables?

- Too few explanatory variables $\rightarrow$ possible omitted variable bias.
- Too many variables may lead to multicollinearity and inflated se's.
- Note: "too many" requires correlation among explanatory variables.
- Can one test one's way out of this?
- No, but tests do help.

How to choose the explanatory variables?

- There are tests of individual and of joint significance. Why cannot I run these on autopilot?
- Case #1: start from a small model, add variables according to some (statistical) criterion.
- Case #2: start from a large model, drop variables according to some (statistical) criterion.
- Case #3: use machine learning methods (for later). Designed especially for the case where number of variables $>$ number of observations.

How to choose the explanatory variables?

- What goes wrong?
  1. Statistical significance $\neq$ economic significance.
  2. Statistical significance $\neq$ economic relevance.
  3. You may end up with variables that are highly correlated with Y, but have no real connection to it.
  4. Multiple testing leads to wrong (too good) test results.

How to choose the explanatory variables?

- The principled approach:
  1. Before touching your data, write down a protocol.
  2. Base explanatory (control) variables on theory and existing knowledge.
  3. Specify a testing protocol.
  4. Execute.

How to choose the explanatory variables?

- The practical approach:
  1. Try to be as close to the principled approach as possible.
  2. Learning allowed and encouraged $\rightarrow$ new/respecification.
  3. Robustness testing.

How to choose the explanatory variables?

- Robustness testing:
  1. It is rarely the case that there is a "right model" that you can (re)cover.
  2. Ask: are your results sensitive to small, well-justified changes to your model?
     1. Adding (meaningful) variables.
     2. Deleting variables.
     3. Changing functional form.
     4. Changing assumptions about the error term.

Q8 What if the world is not linear?

- Well, nothing prevents us from making our model nonlinear.
  1. Keep the $\boldsymbol{X}$ base-variables the same, but make the function $f(\boldsymbol{X})$ more complicated.
  2. Transform the variables.
- Let's start by making $f(\boldsymbol{X})$ more complicated.
- Let's remind ourselves of what the income - age graph looks like.
- But before that let's remind ourselves of what polynomial functions are.

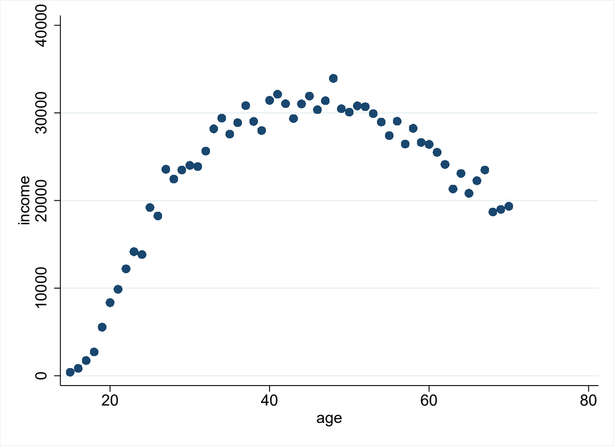
What if the world is not linear?

- Polynomial of order k (of one variable):

$$\begin{aligned} p(X) &= \sum_{i=0}^{k} \alpha_i X^i \\ &= \alpha_0 + \alpha_1 X + \alpha_2 X^2 + ... \alpha_k X^k \end{aligned}$$

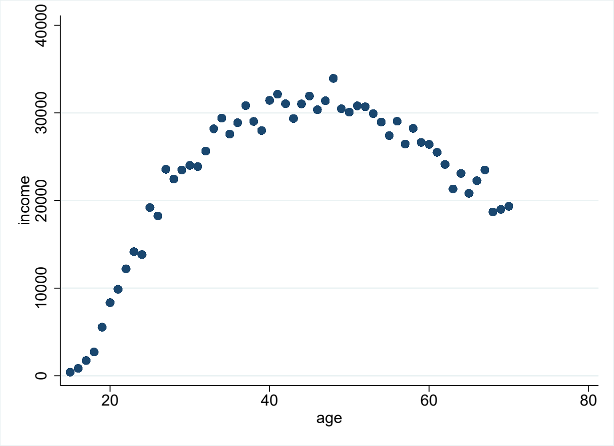
- Polynomials may consist of several variables.

What if the world is not linear?



- The figure suggests that an inverted - U shaped function could be a good fit.

What if the world is not linear?



- The figure suggests that an inverted - U shaped function could be a good fit.
- Let's try a quadratic function of age.

What if the world is not linear?

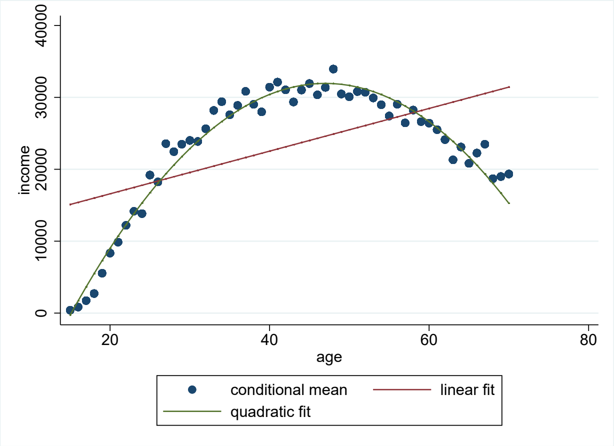
Table: Polynomial income regressions

	(1)	(2)
	income	income
Age	296.8***	2958.2***
	(13.35)	(74.84)
Age2		-31.48***
		(0.874)
Constant	10654.7***	-37549.0***
	(607.6)	(1446.6)
Observations	5973	5973
r2	0.0764	0.241
F	493.9	949.9

Standard errors in parentheses

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

What if the world is not linear?



What if the world is not linear?

- How to test for the order of the polynomial?

  1. Start from an order that is a reasonable (high) one, such as 3 or 4.
  2. Test down, i.e., whether the high(er) order term(s) are (jointly) significant.

- Notice: here you have a prior plan.

- Notice #2: a more modern version of this would involve a **semi-** or **non-parametric** approach (for later courses).

- Let's test whether a second order polynomial is sufficient by adding a third order term.

- Notice that for pedagogical purposes I am doing things in the **wrong** order.

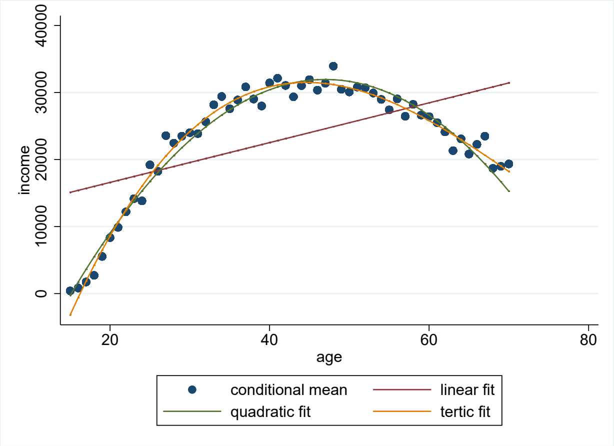
What if the world is not linear?

Table: Polynomial income regressions

	(1) income	(2) income	(3) income
Age	296.8***	2958.2***	4685.0***
	(13.35)	(74.84)	(317.0)
Age2		-31.48***	-75.69***
		(0.874)	(7.934)
Age3			0.346***
			(0.0618)
Constant	10654.7***	-37549.0***	-57597.8***
	(607.6)	(1446.6)	(3856.7)
Observations	5973	5973	5973
r2	0.0764	0.241	0.245
F	493.9	949.9	646.9

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

What if the world is not linear?



What if the world is not linear?

- Can you do more than use polynomials?
- Yes... though polynomials give very good approximations.

  $Y = f(X) + u$

- Give $f(X)$ any shape you like.
- We will skip this for now (semi- and nonparametric estimation).

Q9 Interactions

- What if there is reason to believe that the effect of X_1 depends on the value of X_2?
- Examples:
  1. Returns to experience (=age) and/or education different by gender.
  2. Effect of R&D subsidies different by firm size.

Example: Effect of age on income depends on gender

$$
\begin{aligned}
Income &= f(Age, G, u) = \beta_0 + \beta_{Age} \times Age + \beta_G \times G + u \\
Income &= f(Age, G, u) = \beta_0 + \beta_{Age} \times Age + \beta_g \times G \\
&\quad + \beta_{AgeG} \times Age \times G + u
\end{aligned}
$$

- What is now the expected income | gender?
- What is now the expected income | age?

Q9 Example: Effect of age on income depends on gender

- How to calculate the effect of age on income?
- Now depends on the value of G directly.
- Notice
  1. without the interaction $Age \times G$ the effect of age on income independent of G ($=$ the same no matter what value G takes).
  2. not true any more with the interaction.
- **Note:** if you add an interaction, make sure to have the original variables in the specification as well.

What if the world is not linear?

Table: Polynomial income regressions

	(1) income	(2) income
Age	298.5***	333.5***
	(13.23)	(18.80)
Gender	-4545.0***	-1598.7
	(422.9)	(1203.3)
Age_G		-69.16**
		(26.44)
Constant	12819.2***	11336.3***
	(634.6)	(850.8)
Observations	5973	5973
r2	0.0939	0.0950
F	309.4	208.8

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Q9 Transformations of variables

- What is a transformation of a variable?
- Use some $g(X)$ instead of X.
- Most often use (natural) log of X.
- Sometimes $\frac{1}{X}$.
- Always use a *monotonic* transformation!

Which variable to transform?

- Y, X, or both (all)?
- Using logs *smooths* the data, i.e., decreases the differences across different values that the variable takes.
- Taking logs allows negative values for a non-negative variable (if value < 1)
- On the other hand, cannot take logs if < 0.

Log approximation to % change

$$\ln(Y + \Delta Y) - \ln(Y) \cong \frac{\Delta Y}{Y}$$

Which variable to transform?

1. Only Y

$$\begin{aligned}\ln \mathit{Income} &= \beta_0 + \beta_{Age} \times Age + \beta_G \times G + u \\ \mathit{Income} &= e^{\beta_0 + \beta_{Age} \times Age + \beta_G \times G + u} \\ &= e^{\beta_0} e^{\beta_{Age} \times Age} e^{\beta_G \times G} e^{u}\end{aligned}$$

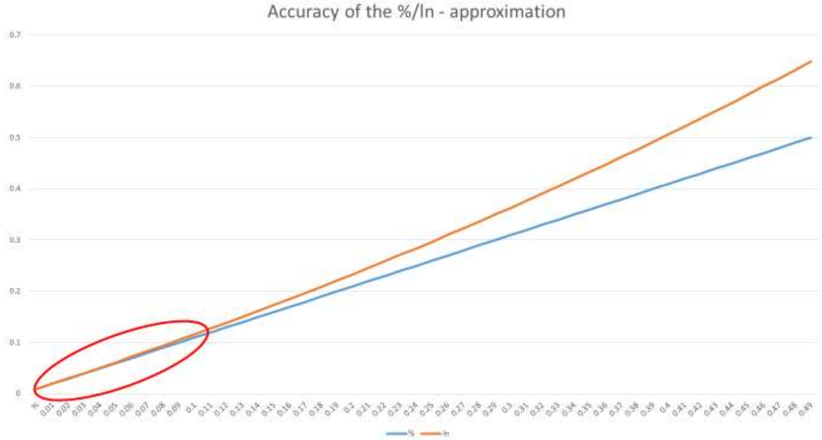
Which variable to transform?

2. Only X

$$Income = \beta_0 + \beta_{\ln Age} \times \ln Age + \beta_G \times G + u$$

- Interpretation of $\beta_{\ln Age}$?
- A 1% increase in Age is associated with at $0.01 \times \beta_{\ln Age}$ change in income.

Log approximation to % change



Which variable to transform?

3. Both Y and X

$$\ln Income = \beta_0 + \beta_{\ln Age} \times \ln Age + \beta_G \times G + u$$

- Interpretation of $\beta_{\ln Age}$?
- $\beta_{\ln Age} =$ %-change in income due to a 1% increase in Age.
- In other words, $\beta_{\ln Age}$ is the **age elasticity of income**.

Which variable to transform?

Stata code

```
gen lnincome = ln(income)
gen lnage = ln(age)
regr income age gender if year == 15 & income != . & income_age_m != ., robust
    eststo linear
regr lnincome age gender if year == 15 & income != . & income_age_m != ., robust
    eststo loglin
regr income lnage gender if year == 15 & income != . & income_age_m != ., robust
    eststo linlog
regr lnincome lnage gender if year == 15 & income != . & income_age_m != ., robust
    eststo loglog
estout linear loglin linlog loglog, cells(b(star fmt(3)) se(par fmt(2))) ///
stats(r2 r2_a F N, fmt(%9.5f %9.5f %9.0g))
```

Log-linear transformations

Table: Polynomial income regressions

	(1) income	(2) lnincome	(3) income	(4) lnincome
Age	298.5***	0.0206***		
	(13.23)	(0.000723)		
Gender	-4545.0***	-0.143***	-4503.1***	-0.140***
	(422.9)	(0.0226)	(412.7)	(0.0217)
lnAge			14059.7***	0.982***
			(486.4)	(0.0269)
Constant	12819.2***	8.970***	-26075.9***	6.236***
	(634.6)	(0.0350)	(1806.8)	(0.100)
Observations	5973	5695	5973	5695
r2	0.0939	0.130	0.137	0.194
F	309.4	425.3	475.3	686.8

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Interpretations of $\beta_{Age}, \beta_{\ln Age}$ (mean age = 43)

- Linear: a 1 unit = 1 year ($\approx 2.3\%$) increase in age increases income by

  300€

- Log-linear: a 1 unit = 1 year increase in age increases income by

  $2.1\% \rightarrow 0.021 \times 23296.67$€$\approx 490$€

- Linear-log: a 1% (note: $\ln(1.01) \approx 0.01$) increase in age (≈ 0.43 years) increases income by

  0.01×14059€≈ 140€$\rightarrow$ effect of 1 year increase $\approx 2.5 \times 140$€≈ 350€

- Log-log: a 1% increase in age increases income by 0.982%. Effect of a 1 year increase

  $\approx 2.5 \times 0.00982 \times 23296.67$€$\approx 570$€

Interpretations of $\beta_{Age}, \beta_{\ln Age}$ (mean age $= 43$)

- Using the log of the dependent variable $\rightarrow$ coefficient interpretation in **percent**.
- Typically, in economic data, using log of explanatory variable leads to higher R^2.
- Many economic variables have a lower limit (income cannot be negative), but OLS assumes that the support is the real line.

  $\rightarrow$ log transformation allows coverage of the real line.

  $\rightarrow$ log transformation necessitates $Y(X) > 0$.

Q10 What can go wrong?

1. Internal validity.
2. External validity.

Internal validity

1. Omitted variable bias.
2. Functional form misspecification (mistake).
3. Measurement error in variable(s).
4. Sample selection (OVB).
5. Simultaneous (reverse) causality (OVB).
6. Non-homoskedastic errors.

Internal validity 1. - Omitted variable bias

- The relevant condition the one we have already discussed.
- "Judicious" choice of controls.
- Add variables.
- There are further solutions. We will get to these.

Internal validity 2. - Functional form

- How can you be sure?

  1. Tests between the functional forms you try.
  2. Note: can easily test only those functional forms that are "nested".

  Example #1: 1st and 2nd order polynomial **nested** (= one is a restricted version of the other).

  Example #2: log-log and linear are non-nested.

- Try out different ones and check robustness of your results (see earlier).

Internal validity 3. - Measurement error in variables

- Case #1: Y measured with error, error random.

$$Y_{observed} = Y + error$$

- Let's have a look at our regressions:

  Regression we'd like to estimate:

$$Y = \beta_0 + \beta_1 X + u$$

  Regression we can estimate:

$$\begin{aligned} Y_{observed} &= Y + error = \beta_0 + \beta_1 X + u + error \\ &= \beta_0 + \beta_1 X + v \end{aligned}$$

Internal validity 3. - Measurement error in variables

- Measurement error in Y not a big problem (as long as random).
- Leads to higher standard errors, but no bias.

Internal validity 3. - Measurement error in variables

- Case #2: X measured with error, error random.

$$X_{observed} = X + error,\ \rho_{X,error} = 0$$

- This is the case of so-called **classical errors-in-variables**. This case is "well-behaved".
- Let's have a look at our regression:

Internal validity 3. - Measurement error in variables

- We would like to estimate

$$Y = \beta_0 + \beta_1 X + u$$

- However we only observe $X_{observed} = X + error$. Hence we need to rewrite

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + u \\ &= \beta_0 + \beta_1 X_{observed} - \beta_1 (X_{observed} - X) + u \\ &= \beta_0 + \beta_1 X_{observed} - \beta_1 error + u \\ &= \beta_0 + \beta_1 X_{observed} + v \end{aligned}$$

Internal validity 3. - Measurement error in variables

- One can show (see SW ch. 9.2):

$$\hat{\beta}_1 = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_{error}^2}\beta_1$$

- $\frac{\sigma_X^2}{\sigma_X^2+\sigma_{error}^2}$ = **signal to noise** - ratio.
- The larger is the role of the error, i.e., the larger is σ_{error}^2 relative to σ_X^2, the more biased is $\hat{\beta}_1$.
- This is so-called **Attenuation bias**.

Internal validity 3. - Measurement error in variables

- Solution #1: Get better measures.
- Solution #2: Get a measure of σ^2_{error}.
- Solution #3: A technical solution (**instrumental variables**) that we will get to later.

Internal validity 4. - Sample selection

- Your observations are not a random sample of the underlying population.
- Example #1: Estimate the returns to entrepreneurship using 5 year old firms.
- The non-profitable entrants exit.
- Example #2: Estimate the returns to graduating quickly.
- Those who graduate quickly have unobservable skills that make them (un)attractive to employers.

Internal validity 4. - Sample selection

- Example #3: estimate effects of R&D subsidies.
- Firms that get subsidies not avg. firms.
- Rule: think through and understand selection into your sample.
- Model selection into the sample.
- We will discuss this later, but in general is an advanced topic.

Internal validity 4. - Sample selection

- Sample selection can threaten internal validity - the parameters you obtain for the population of interest are biased.
- Sample selection can also threated the external validity of your exercise, i.e., even if you get unbiased estimates for the population in question, your results do not generalize.

Internal validity 5. - Simultaneous causality

- Think of the determination of prices and quantities.
- Price affects how much is sold and produced.
- How much is bought and produced affects the price.

  $\rightarrow$ simultaneous causality.
- We will come back to this.

Internal validity 6. - Heteroskedasticity

- Deviations from homoskedasticity can take different forms depending on the data.
- With sequential observations, maybe also correlation over time (**autocorrelation**).
- With e.g. geographical data, correlation across observation units (**clustering**).
- Affects statistical precision (=standard errors) of individual coefficients, nothing else.
- Can be corrected by using **robust** standard errors (with potential loss of efficiency - but robust se's can be smaller than homosc. se's).
- In data with relevant other dimensions (e.g. geographical locations), clustered se's may be more appropriate than regular robust se's.

Internal validity 6. - heteroskedasticity

Stata code

```
regr income age gender if year == 15 & income != . & income_age_m != .
    eststo linear
regr income age gender if year == 15 & income != . & income_age_m != ., robust
    eststo linear_het
```

Internal validity 6. - heteroskedasticity

Table: Polynomial income regressions

	(1) homosc.	(2) heterosk. robust
Age	298.5***	298.5***
	(13.23)	(11.82)
Gender	-4545.0***	-4545.0***
	(422.9)	(421.6)
Constant	12819.2***	12819.2***
	(634.6)	(566.5)
Observations	5973	5973
r2	0.0939	0.0939
F	309.4	369.3

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

External validity - Recall from Lecture 1

- Econometrics:

  A branch of economics in which economic theory and statistical methods are fused in the analysis of numerical and institutional data

  Hood, W. & Koopmans, T. (1953). Studies in econometric method. *Cowles Commission Monograph no. 14, Wiley*

External validity

- External validity = results generalize to other settings than the one studied.
- Any (material) change to any of the components of your study jeopardizes external validity.

1. Differences in (applicable) theory.
2. Differences in statistical method.
3. Differences in data (including in populations).
4. Differences in institutions.

External validity

- Example: Do our income - age results hold for (an)other year in the FLEED data?
- Let's compare results from current year 15 to year 10.

External validity: comparison of results from two very similar data

Table: Polynomial income regressions

	(1) year 15	(2) year 10
Age	298.5***	216.4***
	(13.23)	(11.74)
Gender	-4545.0***	-4765.7***
	(422.9)	(361.7)
Constant	12819.2***	12167.0***
	(634.6)	(568.5)
Observations	5973	5779
r2	0.0939	0.0803
F	309.4	252.2

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

External validity

- Is any study externally valid?
- Yes and no.
- Best to ensure internal validity, and conduct many studies.