ECON-C4100 - Capstone: Econometrics I

Lecture 5: Univariate regression

Otto Toivanen

OLS assumptions

- Recap from previous lecture:

1. Strict exogeneity: $\mathbb{E}(u_i|X_i) = 0$.
2. $(X_i, Y_i),\ i = 1, ..., n$ are independent and identically distributed across observations.
3. X_i and Y_i have finite *fourth* moments.
4. Auxiliary: u_i is homoskedastic.

Learning objectives of this lecture

- At the end of lectures 3 - 5, you understand what one learns from a (univariate) regression analysis. In this lecture you will learn about the following:

1. how OLS results are affected if the data/model violate Assumption #2, i.i.d sampling.
2. how OLS results are affected if the data/model violate Assumption #1, strict exogeneity.
3. point estimates
4. standard errors
5. t-statistics
6. p-values
7. critical values
8. statistical significance
9. confidence intervals

OLS Assumption #1

$$\mathbb{E}(u_i|X_i) = 0$$

- Implies that u and X are uncorrelated.
- $\mathbb{E}(u_i|X_i) = 0 \implies cov(u, X) = 0.$
- Not the other way round because correlation is about a linear relationship only.

OLS Assumption #2

- $(X_i, Y_i),\ i = 1, ..., n$ are i.i.d.
- The same concept as before, but now over a joint distribution of two variables.
- Experiments where X chosen.
- Time series.

Assumption 2: What if the data are not i.i.d?

- Let's discuss the assumption that the observations are i.i.d. in the context of our FLEED income - age data.
- In other words, let's pretend that the 5 973 observations are the **population** rather than a **random sample**.
- In those data, i.i.d. means that
  1. All observations are equally likely to end up in our random sample.
  2. Observing one observation is not informative about the other observations.
- Random sampling means that in expectation, the sample has the same distribution as the population.

What if the data are not i.i.d?

- How could the assumption be violated?
- Well, by doing non-random sampling.
- Example: let's systematically choose young / old with a higher probability than their population share.
- Let's start with the whole data, then drop individuals in the lowest z age deciles, $z = 1, ..., 5$.

What if the data are not i.i.d?

Table: Regression table with non-iid sample

	(1) all data	(2) drop 1d	(3) 2d	(4) 3d	(5) 4d	(6) 5d
age	296.8***	30.16	-157.9***	-325.2***	-458.1***	-569.3***
	(22.22)	(1.78)	(-7.38)	(-11.76)	(-13.32)	(-12.87)
Constant	10654.7***	24836.6***	35227.4***	44790.8***	52623.2***	59377.6***
	(17.54)	(30.15)	(32.25)	(30.16)	(27.43)	(23.16)
Observations	5973	5113	4453	3807	3299	2758

t statistics in parentheses

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

- Notice how the parameters change.
- Note however that for each subsample, the parameters are unbiased. It is just that they are not so for the whole population.

What to do about non-i.i.d samples?

- The example is a rather innocent one, and a case of **selection on observables**.
- In other words, the selection of individuals into the sample took place based on a variable we econometricians observe.
- Such sampling is sometimes done on purpose, e.g., excluding too old individuals as they are close to retirement / some of them have already retired.
- A more often encountered and a more difficult problem is **selection on unobservables**. This is a topic for a more advanced course.
- As an example, think of the current discussion in Finland on the *home care allowance (kotihoidontuki)*: are those parents that use / do not use it a randomly selected group of parents of young children?

Assumption 1: Strict exogeneity: $\mathbb{E}(u_i|X_i) = 0$

- Recall that this rules out that X and U are correlated.
- If this is the case, we say that X is **endogenous**.
- What would go wrong if this was the case?
- Let's think of the case of $corr(X, U) > 0$.
- Let's denote $\tilde{y}_{i0} = \beta_0 + \beta_1 x_i$.

Strict exogeneity

- Given positive correlation between X and U, what do we know of

$$\mathbb{E}(u_i|X_i)$$

as a function of X?

Strict exogeneity

- Given positive correlation between X and U, what do we know of

  $$\mathbb{E}(u_i|X_i)$$

  as a function of X?
- Yes, it is increasing in X.

Strict exogeneity

- This implies that as X increases, we go from a situation where

$$\mathbb{E}(y_i|x_i) < \tilde{y}_{i0} = \beta_0 + \beta_1 x_i$$

for low values of X to a situation where

$$\mathbb{E}(y_i|x_i) > \tilde{y}_{i0} = \beta_0 + \beta_1 x_i$$

for large values of X.

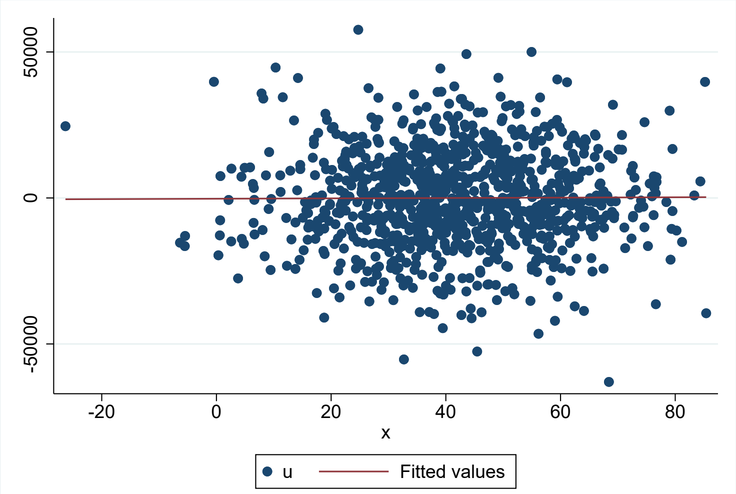
Strict exogeneity

- What does this imply about our regression? It is systematically biased.
- When you do your Least Squares calculation, you end up choosing the parameters so that you get zero errors on average (that is what OLS does).
- You will end up with a regression line that has a slope β_1 that is too large.

Strict exogeneity

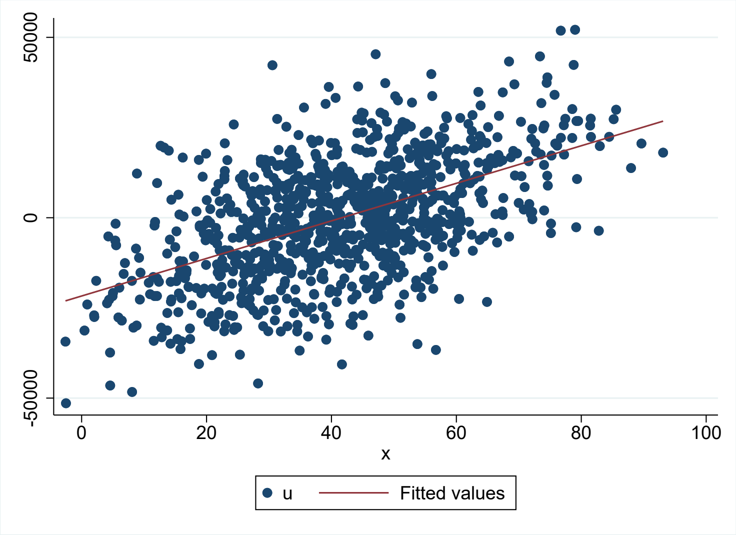
- Let's illustrate how the "data", i.e., the Us look different when
  1. X is exogenous
  2. X is endogenous.

X and U uncorrelated, $\rho_{XU} = 0$



the prediction line is not exactly horizontal because the sample is one random draw

X and U correlated, $\rho_{XU} = 0.5$



Strict exogeneity

- Let's have a look at what happens to β_0 and β_1 when we increase the correlation between X and U.

$\rho_{XU} = 0.1, 0.2, ..., 0.9$

Table: Effect of endogeneity

ρ_{XU}	β_1	se_{β_1}
0	298.32	32.11
0.1	397.14	31.89
0.2	506.35	31.30
0.3	601.94	30.56
0.4	702.25	29.44
0.5	797.81	27.72
0.6	906.04	25.50
0.7	1003.15	22.94
0.8	1107.05	19.27
0.9	1209.41	14.03

Can we say something about this analytically?

- Recall the formula for $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{cov(Y, X)}{var(X)}$$

- Recall also that we have an equation for Y:

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

- Let's plug the latter into the former and see what we get, maintaining Assumptions 2 and 3 (i.i.d and finite variances).

Omitted variable bias

- By splitting the covariance between Y and X into its constituent parts and simplifying, we arrive at:

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

Note: σ_i = standard deviation of variable i, ρ_{ij} = correlation coefficient of variables i and j.

- This equation is the formula for **omitted variable bias**.
- This problem is arguably the most severe of the ones we've discussed thus far.

Omitted variable bias

$$\hat{\beta}_1 = \beta_1 + \rho_{xu} \frac{\sigma_U}{\sigma_X}$$

- Our estimate $\hat{\beta}_1$ is equal (in expectation) to β_1 if the second term on the right is zero.

Omitted variable bias

$$\hat{\beta}_1 = \beta_1 + \rho_{xu} \frac{\sigma_U}{\sigma_X}$$

- Our estimate $\hat{\beta}_1$ is equal (in expectation) to β_1 if the second term on the right is zero.
- It consists of three terms, two of which cannot be zero for our regression to work, i.e., σ_u and σ_X.

Omitted variable bias

$$\hat{\beta}_1 = \beta_1 + \rho_{xu}\frac{\sigma_U}{\sigma_X}$$

- Our estimate $\hat{\beta}_1$ is equal (in expectation) to β_1 if the second term on the right is zero.
- It consists of three terms, two of which cannot be zero for our regression to work, i.e., σ_u and σ_X.
- Therefore, it being zero boils down to the value of ρ_{XU}.

Omitted variable bias

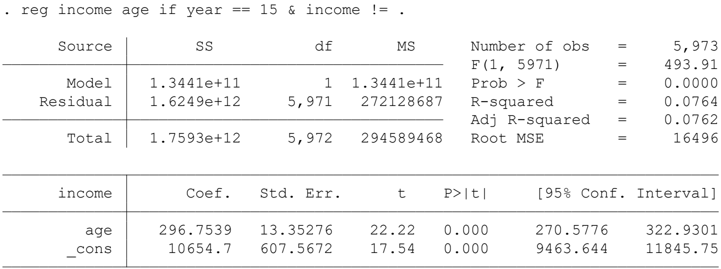
- $\rho_{XU} = 0$ is an **untestable** assumption.
- This is so because u_i are unobserved and different from $\hat{u}_i$.
- To evaluate whether the assumption that $\mathbb{E}[U|X] = 0$ holds, you need to think of whether you are interested in
  1. getting unbiased estimates of the parameters and
  2. being able to predict Y well.
- In the first case, strict exogeneity needs to hold. In the second case, it may or may not matter.

Hypothesis testing and statistical significance using regression

- Today's question: How likely is it that age has an effect on income / that income varies with age?

  i.e. test the null hypothesis that age has no effect.

Our estimation results



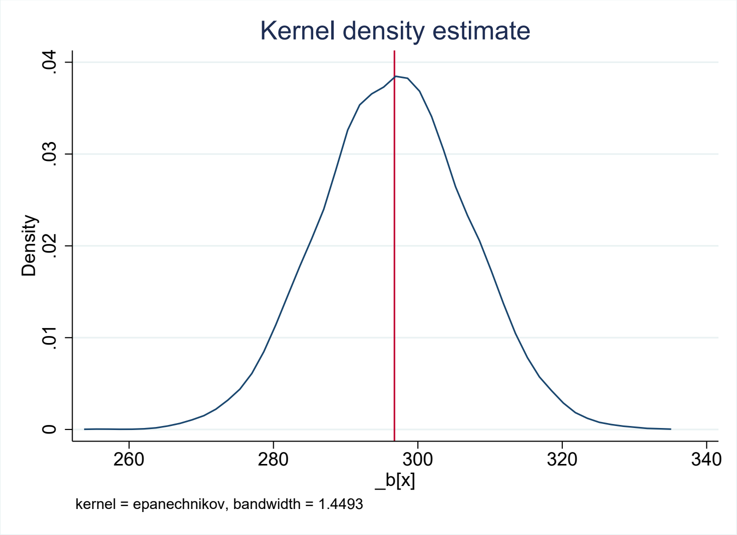
```
. reg income age if year == 15 & income != .

      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
------------------------------------------------------------------------------
```

Point estimates

- Our **point estimate** of $\beta_1 = 296.8$.
- β_1 is a random variable, i.e., it has a distribution.
- How likely is it that we by chance get a value this high or that we would get a number as small or smaller than zero?
- We can approach this question by going back to our Monte Carlo exercise.
- Alternatively, we can approach the question analytically and derive the (parameters of) its distribution. This is what is done when **asymptotic** test-statistics are used.

Distribution of β_1

- Let's first have a look at the distribution of β_1.



Standard error

- Our estimate of the **standard error** of $\beta_1 = 13.4$.
- Standard error is the **standard deviation of a statistic.**
- How can we calculate the standard error?
- By the CLT, we know that the distribution of β_1 becomes (approximately) normal as the sample size increases.
- As a normal distribution is completely characterized by its mean and variance (standard deviation), it is sufficient that we calculate the standard error of β_1.
- In practice, we estimate the standard error from the data. It is a function X and the estimated $\hat{U}$.

Standard error

- The standard error reflects the fact that β_1 is a random variable.
- How does this affect the value our estimate $\hat{\beta}_1$ takes?
- Starting from the formula for $\hat{\beta}_1$ that we derived one can derive the following:

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum(x_i - \bar{X})u_i}{\frac{1}{n}\sum(x_i - \bar{X})^2} \qquad (2)$$

- See Appendix 4.3 in SW for the derivation.
- It can be shown that the second term on the right is zero in expectation (hence $\hat{\beta}_1$ is unbiased).
- The second term on the right is a random variable and gives $\hat{\beta}_1$ its distribution and hence the basis for the standard error of $\hat{\beta}_1$.

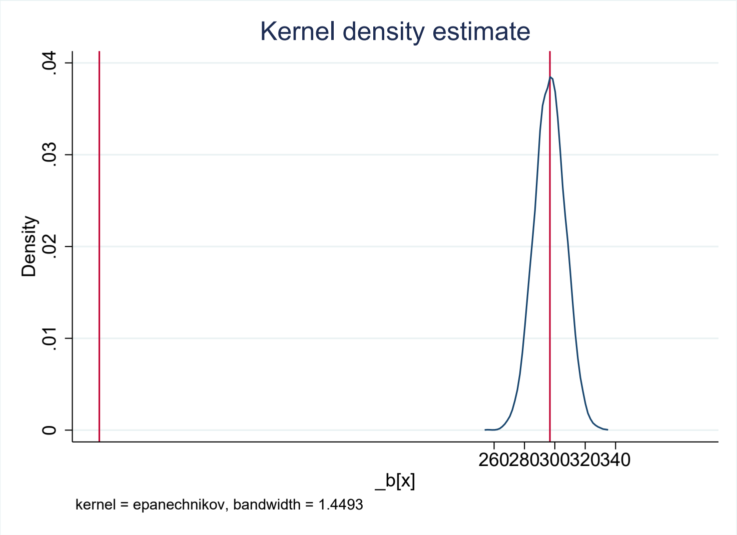
t-statistics, p-value, critical value, and statistical significance

$$t = \frac{\hat{\beta}_1}{\hat{\sigma}_{\beta_1}}$$

- The **t-statistic** is the ratio of the point estimate and its standard error, in the case of β_1 it is 22.22.
- We can then calculate the **p-value** for the t-statistic.
- It tells us how likely it is that a **standard normal variable** would be larger than the t-statistic.
- If this probability is low, below our chosen **critical value** such as 5% or 0.05, we reject the Null hypothesis and say that β_1 is **statistically significant**.
- Critical values are *norms* or rules of thumb.

Visualizing the p-value

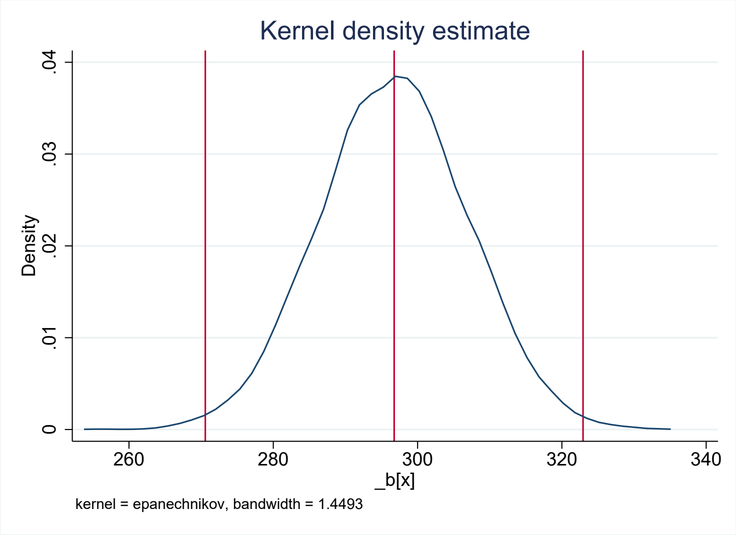
- Let's then ask how likely it is that we could get a value as low as or lower than zero.



Confidence interval

- Instead of asking how likely it is that our estimate is larger or smaller than some comparison value (our Null hypothesis value), we could ask how likely it is that the true parameter value is **within some interval**.
- This is the idea behind **confidence intervals**.
- Based on our estimates (point estimate and its standard error), we can calculate such intervals.
- Let's use as an example the most common, i.e., 95% confidence interval.
- For $\hat{\beta}_1$, this is $[270.6, 322.9]$.

Visualizing the confidence interval



Some final observations

- Here, we concentrated on how to measure and test the statistical significance of a parameter (coefficient).
- We will discuss how to measure and test the statistical significance of the estimation model when we deal with multivariate regression.
- Researchers often use phrases like β_1 *is statistically significant* somewhat loosely. P-values and confidence intervals are "hard facts" that help interpretation of such statements.
- It is important to understand these concepts. We will return to the meaning of statistical significance when we talk about causality. There, important concepts are
  1. false positives and negatives.
  2. statistical power.
  3. minimum detectable effect size.