ECON-C4100 - Capstone: Econometrics I

Lecture 4: Univariate regression

Otto Toivanen

Learning objectives of this lecture

- In this lecture you will learn about the following:

1. Any regression analysis rests on a set of assumptions
2. Knowing the assumptions you impose and understanding the consequences of violating them are key to an informed analysis.
3. The OLS assumptions.
4. The consequences of violating two of them.

What are the numbers produced by univariate regression?

- We postpone further discussion of regression level diagnostics to the lectures on multivariate regression.
- The reason for this is that they (adjusted R^2, F-test), are more meaningful in the multivariate context.

What are these numbers?

- Economic interpretation & significance is of key importance.
- What about statistical significance?
- Recall the discussion on the properties of the sample mean:
  1. It is a random variable (every random sample produces its own mean to be used as an estimate of the population mean).
  2. It is unbiased and consistent
  3. It has a distribution that we can characterize (with large n, becomes / approaches a normal distribution).

What are these numbers?

- Similarly, the parameters we "find" or estimate with OLS depend on the random sample available to us.
- Were you to draw a different random sample, you would get different parameter estimates.
- In other words, β_0 and β_1 are also random variables.
- We'd like to know their properties, i.e., how they are distributed, and how different things affect that distribution.
- Under assumptions that we'll discuss in a moment, β_0 and β_1 are (bivariate) normally distributed with a known mean and variance.

What are these numbers?

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are
  1. unbiased
  2. consistent and
  3. efficient (with an extra assumption).

  under a set of assumptions.

OLS assumptions

- One needs to understand the assumptions that allow a particular interpretation of the results.
- Crucial to understand the assumptions & their implications.
- Crucial to form an opinion or test the validity of assumptions and/or the robustness of results to those assumptions.

OLS assumptions

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 homoscedastic.

We next discuss each of these in turn.

OLS Assumption #1

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

- Implies that u and X are uncorrelated.
- $E(u_i|X_i) = 0 \implies cov(u, X) = 0.$
- 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.

OLS Assumption #3

- X_i and Y_i have finite *fourth* moments: $\mathbb{E}(X)^4, \mathbb{E}(Y)^4 < \infty$.

  $=$ they have finite kurtosis.

- Needed to ensure that the standard errors are from a normal distribution (4th moment $\approx$ variance of variance).

- Means that large outliers are (extremely) unlikely.

OLS Assumption #4

- $var(u_i|X_i = x) = \sigma^2$ for $i = 1, ..., n$.
- u_i is homoscedastic (as opposed to heteroskedastic).
- Alternative: $var(u_i|X_i = x) = \sigma_i^2$.

The Gauss-Markov Theorem

- The Gauss-Markov Theorem states that:

  If A.1 - A.4 hold, then OLS is BLUE (Best Linear Conditionally Unbiased Estimator).

- You can find the proof in your textbook.

Let's have a look at the effects of the OLS assumptions

- To understand the effects of the OLS assumptions, let's study the following estimation equation:

$$Y_i = \beta_0 + \beta_1 X_i + u = \boldsymbol{X_i'}\beta + u_i \tag{1}$$

- Let's vary different aspects of the **Data Generating Process (GDP)**.
- How do we do this?

(Monte Carlo) simulation

- Let's use artificial data that has "appealing" features.
- Artificial data: ask the computer to generate it.

  $\rightarrow$ the researcher chooses what the data looks like.
- Monte Carlo simulation: repeat a statistical model S times on artificial data, look at means and distributions of parameters.
- We are going to generate artificial data that has the key properties of our FLEED data and use it to illustrate the effects of the OLS assumptions.

How to generate data?

1. Decide the properties you want the data to have.
2. Choose the parameters of the model.
3. Use a random number generator to generate the exogenous variables, including the error term.
4. Generate the dependent variable using the parameters and the exogenous variables.

How to generate data that looks like FLEED?

Stata code

```
regr income age if year == 15 & income != .
predict u_hat, res
sum income age u_hat
matrix beta    = e(b)
matrix list beta
scalar beta0  = beta[1,2]
scalar beta1  = beta[1,1]
qui sum u_hat   if e(sample)
scalar u_sd   = r(sd)
qui sum age
scalar age_m  = r(mean)
scalar age_sd   = r(sd)
```

How to generate data that looks like FLEED?

Stata code

```
drop all
global age_m          = age_m
global b0           = beta0
global b1           = beta1
global age_m          = age_m
set seed 987345
capture program drop myprog_sim

program define myprog_sim
  drop _all
  set obs  10000
  gen x               = $age_m + rnormal(0,age_sd)
  scalar beta0          = $b0
  scalar beta1          = $b1
  gen u           = rnormal(0,u_sd)
  qui sum u
  scalar u_mean         = r(mean)
  replace u           = u - u_mean
  gen y                 = beta0 + beta1 * x + u
  regr y x
end

simulate _b _se, saving(myprog_sim, replace) reps(1000): myprog_sim
display "OLS nobs 10000"
sum
```

Assumption 4: Homoskedastic versus heteroskedastic u

- To study the role of the variance of the error term, let's create data sets with different types of variances.
- The data generating process:
- Case #1: $u = rnormal(0, \sigma_u^2)$
- Case #2: $u_{het} = rnormal(0, \sigma_u^2) \times (1 + z \times age)$
  - $z =$ a multiplier chosen by the modeller.
- Notice both cases satisfy $\mathbb{E}(u|age) = 0$.

Homoskedastic versus heteroskedastic u

$$Income_i = \hat{\beta}_0 + \hat{\beta}_1 age_i + u_i$$

$$Income_{het,i} = \hat{\beta}_0 + \hat{\beta}_1 age_i + u_{het,i}$$

- Let's vary $z = 0.1, ..., 1$
- Sample size 1000.

Benchmark: the actual regression results

```
. estimates store lin_est

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

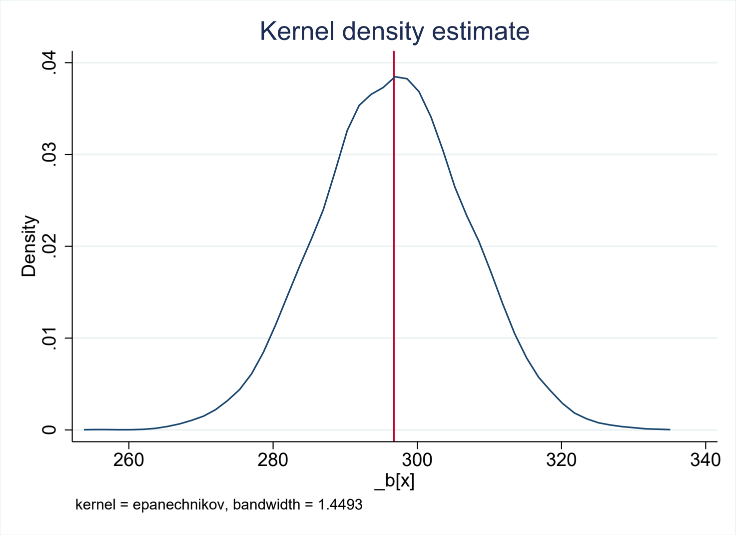
Variable	lin_est
age	296.754
	13.353
	0.000
_cons	1.1e+04
	607.567
	0.000
r2	0.076

legend: b/se/p

Let's first check how this works with $z = 0$: Estimates

Variable	Obs	Mean	Std. Dev.	Min	Max
_b_x	1,000	296.5223	10.36742	267.9936	325.9853
_b_cons	1,000	10664.63	441.6165	9412.67	11876.38
_se_x	1,000	10.31445	.1038239	9.895374	10.73409
_se_cons	1,000	469.3358	4.696664	449.9725	488.3702

Let's first check how this works with $z = 0$: Distribution of β_1



Then increase z

Table: Effect of heteroskedasticity

z	β_1	se_{β_1}	β_0	se_{β_0}
1	293.58	174.12	10741.91	7815.36
2	337.49	318.82	9111.05	14304.02
3	297.95	462.01	10402.12	20770.64
4	261.93	606.49	11762.07	27183.05
5	326.50	747.35	9278.87	33530.71
6	374.37	895.85	7290.83	40201.69
7	176.23	1042.51	15146.60	46913.89
8	397.26	1186.82	6880.53	53272.23
9	320.64	1328.07	9650.81	59678.00
10	358.15	1470.39	8538.93	66067.65
"Truth"	296.754		10664	

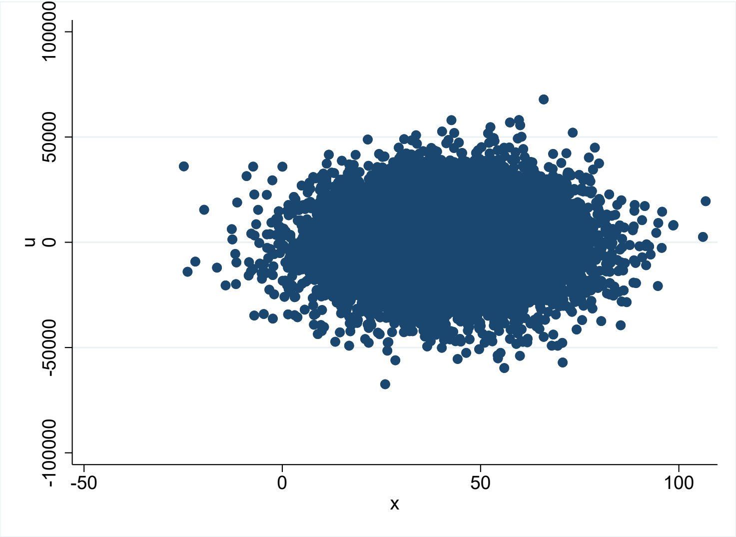
What happens to estimated parameters and their standard errors?

- The parameter estimates vary from row to row in the table of the previous slide, but are on average correct.
- The standard errors however are monotonically increasing as we go down the rows, i.e., as we increase the degree of heteroskedasticity.

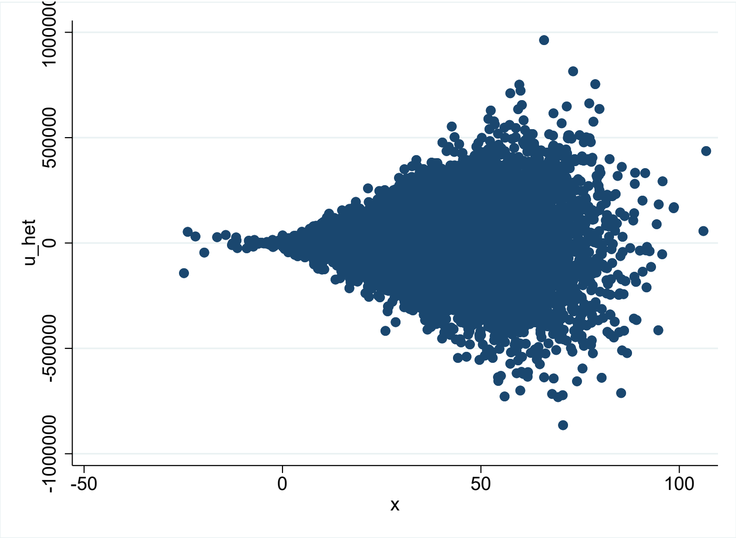
What happens to the distribution of u?

- $\mathbb{E}[u|X] = 0$ holds for all the samples.
- But the variance of u becomes an increasing function of *age*.
- This leads to a very different looking distribution.

Distribution of homoskedastic *u*



Distribution of heteroskedastic u



What to do about heteroskedasticity?

- In practice, data have/lead to heteroskedastic errors almost always.

  $\rightarrow$ easy and efficient ways to correct for heteroskedasticity.

- Modern default is to use (heteroskedasticity) robust standard errors.

- Wrong assumption on variance of the error term biases standard errors, *not coefficients*.

Assumption #3: No large outliers

- (large) outliers may lead to biased estimates.
- Difficulty is of course to determine what is large.
- For illustration, let's replace a few values of *age* with much larger values.
- First, *age* of one individual multiplied by 10 ("typo") in a sample of 1000 observations.
- Second, same done for 10 individuals.

Introduce outliers

Table: Effect of outliers

% obs. changed	β_1	se_{β_1}	β_0	se_{β_0}
0.1	207.16	26.63	14348.99	1247.97
1	47.29	12.71	20958.43	795.41
True estimates	296.754	13.353	10664	607.567

What to do about outliers?

- Always check your data for outliers.
- If you find any, check whether they are typos or real.
- Check that your results are robust to excluding the outlier - observations from your estimation sample.
- Using richer functional forms for $Y = f(X, u)$ (=i.e., multiple regression) may also help.
- ADVANCED: a technique called winzorising allows a systematic study of the effect of outliers. In winzorizing, extreme values are replaced by "less extreme" values, e.g. the 1^{st} and 99^{th} percentile.