Simple Univariate Regression: Part 1

August 19, 2020

Plan:

- simple univariate regression analysis for the next two weeks
- multivariate regression after that

Population and Sample

Population

A set of ALL individuals, items, phenomenon, that you are interested in learning about

Example

- Suppose you are interested in the impact of eduction on income across the U.S. Then, the population is all the individuals in U.S.
- Suppose you are interested in the impact of water pricing on irrigation water demand for farmers in NE. Then, your population is all the farmers in NE.

Important

Population differs depending on the scope of your interest

Population and Sample

Sample

Sample is a subset of population that you observe

Example

- data on education, income, and many other things for 300 individuals from each State
- data on water price, irrigation water use, and many other things for 500 farmers who farm in the Upper Republican Basin (southwest corner of NE)

Question

Are the samples representative of the population you are interested in learning about?

Econometrics

Econometrics

learn about the population using sample

Simple linear regression model

Consider a phenomenon in the population that is correctly represented by the following model (This is the model you want to learn about using sample),

A simple model in the population

$$y = \beta_0 + \beta_1 x + u$$

- y: to be explained by x (dependent variable)
- x: explain y (independent variable, covariate)
- u: parts of y that cannot be explained by x (error term)
- β_0 and β_1: real numbers that gives the model a quantitative meaning (parameters)

Simple linear model

A simple model

$$y = \beta_0 + \beta_1 x + u$$

What does β_1 measure?

$$\Delta y = \beta_1 \Delta x + \Delta u$$

If you change x by 1 unit ($\Delta x = 1$) while holding u (everything else) constant ($\Delta u = 0$),

$$\Delta y = \beta_1$$

That is, y changes by β_1. We call β_1 the ceteris paribus (with everything else fixed) causal impact of x on y.

Why do we fish for ceteris paribus causal impact?

Example: Quality of College

You

- have been admitted to University A (better) and B (worse)
- are trying to decide which school to attend
- are interested in knowing a boost in your future income to make a decision

Why do we fish for ceteris paribus causal impact?

You have found the following data

	A	B
average income after graduation	131K	90K
sample size	500	500

Why do we fish for ceteris paribus causal impact?

You have found the following data

	A	B
average income after graduation	131K	90K
sample size	500	500

Question

Should you assume the difference $(41K)$ is the expected boost you would get if you are to attend A instead of B?

Why do we fish for ceteris paribus causal impact?

Question

Should you assume the difference $(41K)$ is the expected boost you would get if you are to attend A instead of B?

What would you be interested in?

Let's say your ability score is 6 out of 10 (the higher, the better),

(1) $E[inc|A, ability = 9]$
$- E[inc|B, ability = 6]$?

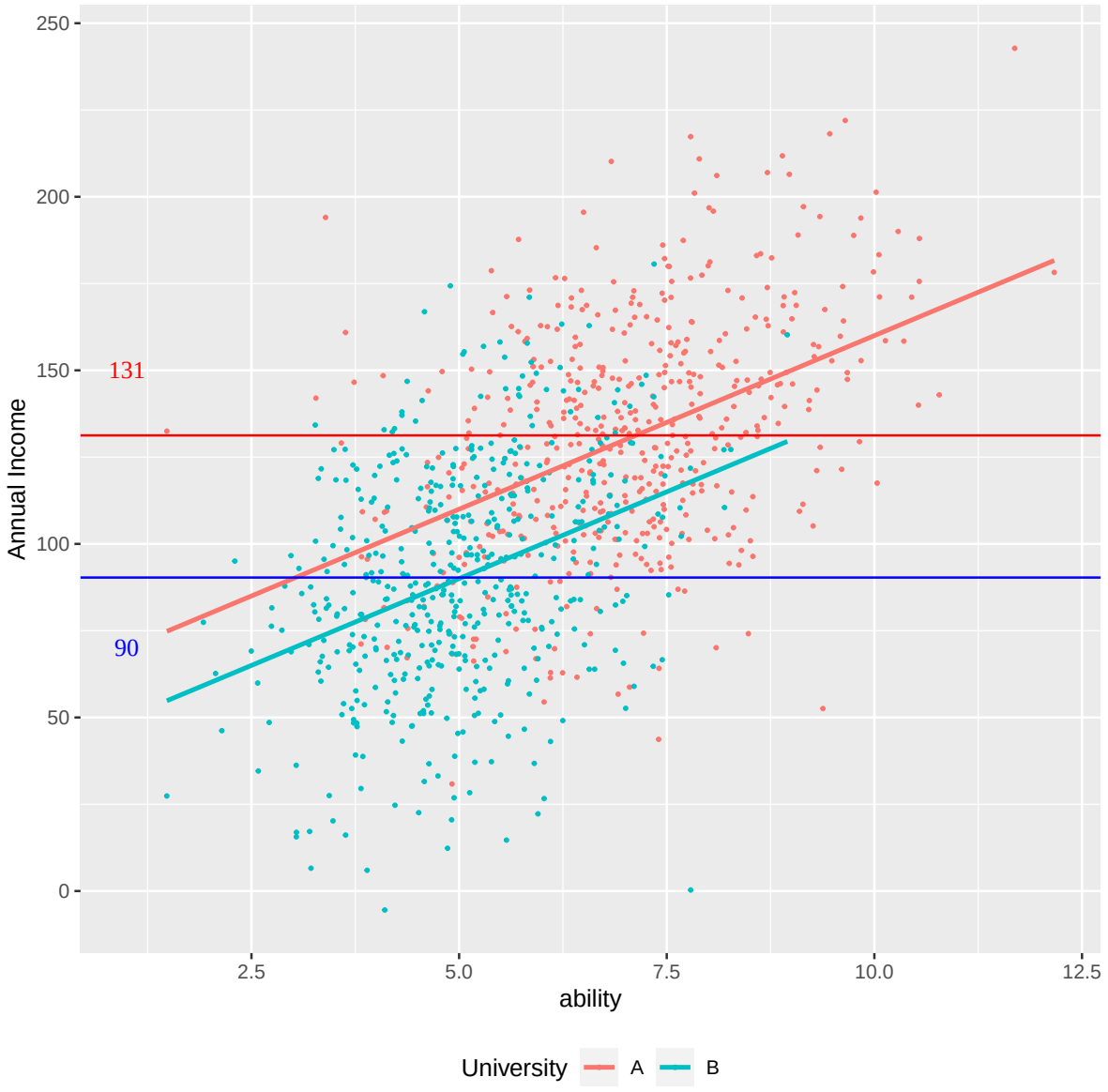
(2) $E[inc|A, ability = 6]$
$- E[inc|B, ability = 6]$?

Ceteris Paribus Impact of School Quality

Why ceteris paribus impact?

- you want ability (an unobservable) to stay fixed when you change the quality of school because your innate ability is not going to miraculously increase by simply attending school A
- you don't want the impact of school quality to be confounded with something else

Why do we want ceteris paribus causal impact?



What does β_0 measure?

A simple model

$$y = \beta_0 + \beta_1 x + u$$

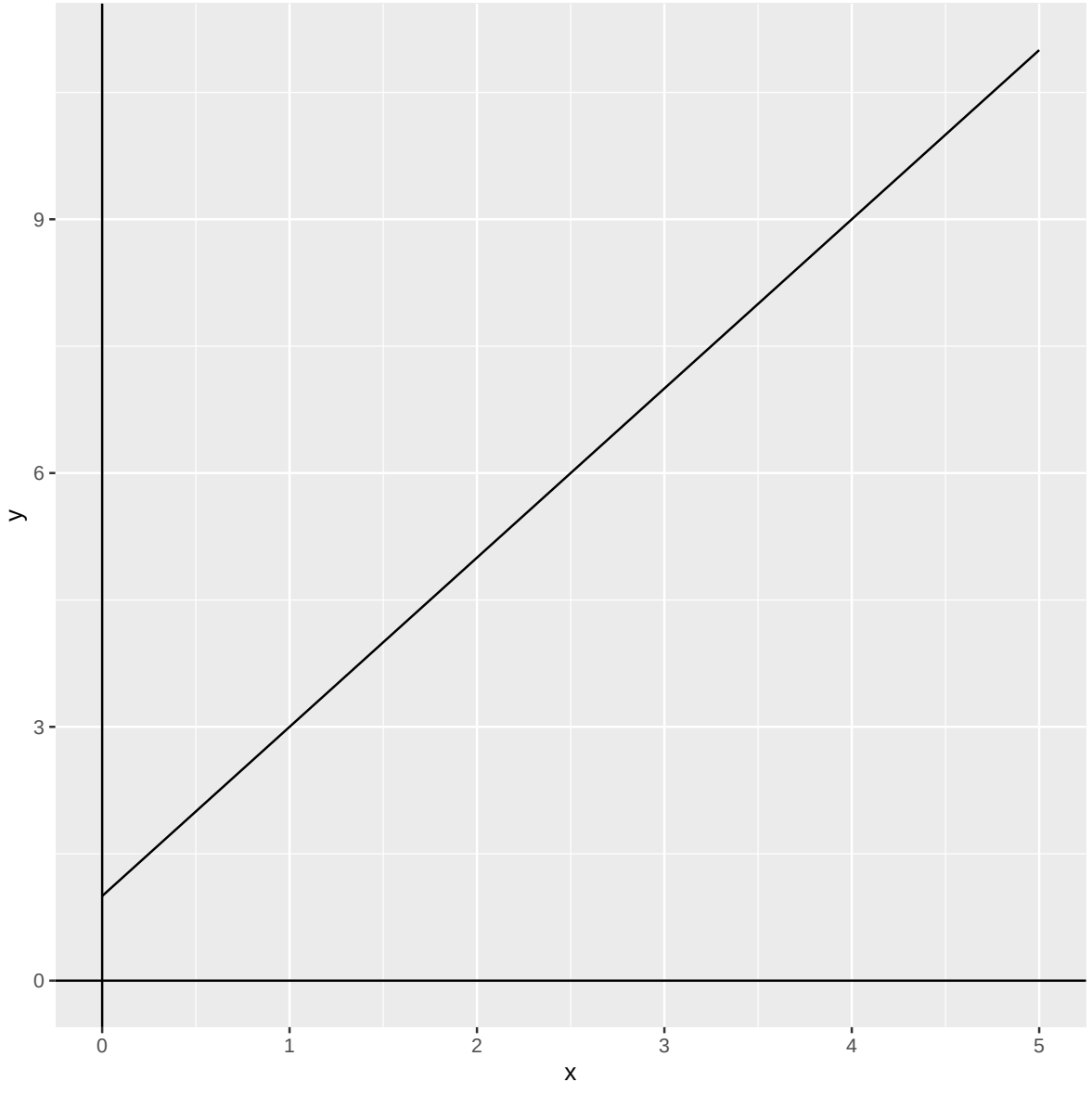
What does β_0 measure?

When $x = 0$ and $u = 0$,

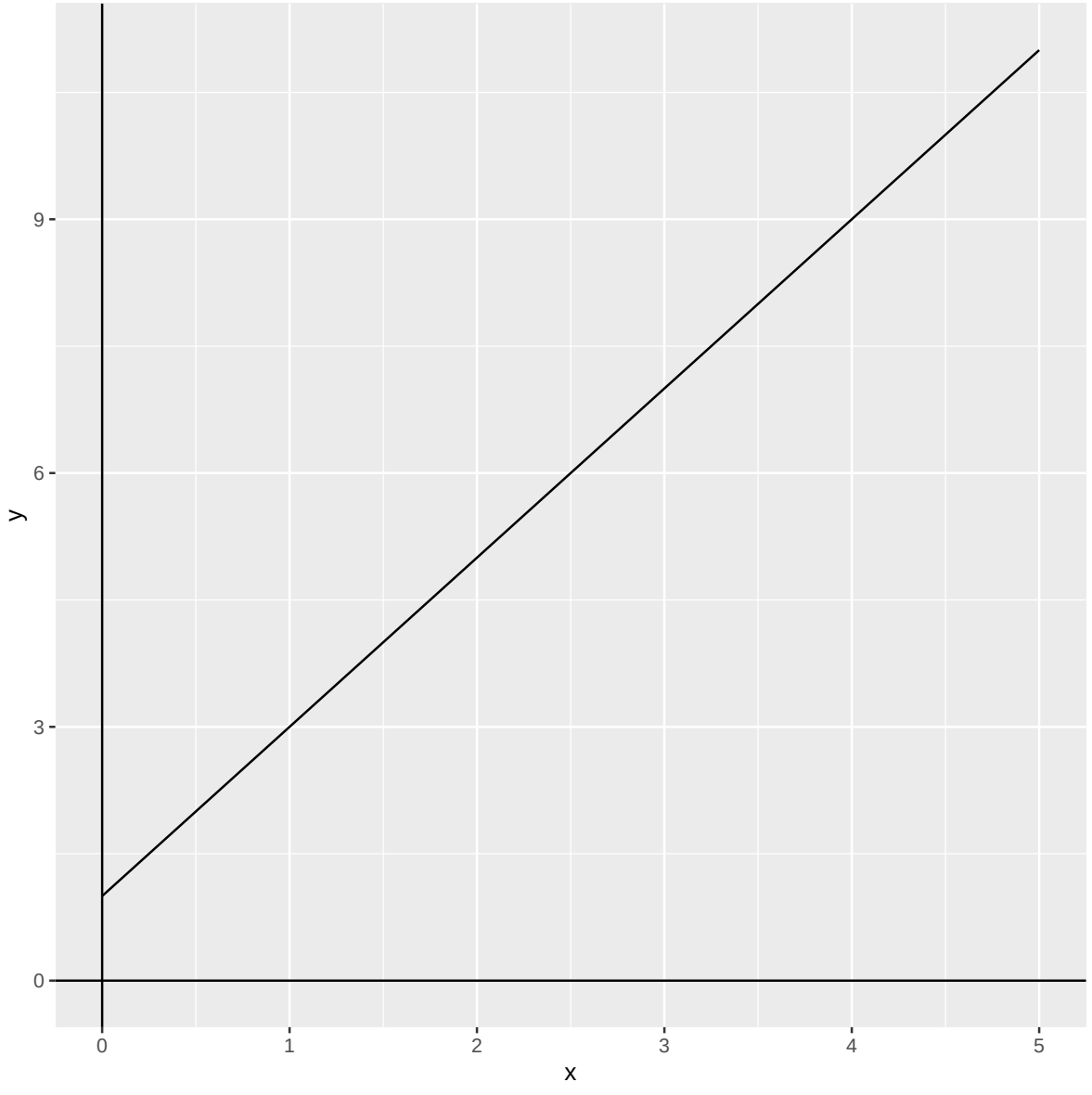
$$y = \beta_0$$

So, β_0 represents the intercept (let's see this graphically).

Graphical representation of the model



Graphical representation of the model



β_0: intercept

β_1: coefficient (slope)

Example of a simple linear model

Corn yield and fertilizer

$$yield = \beta_0 + \beta_1 fertilizer + u$$

Questions

- what is in the error term?
- are you comfortable with this model?

Estimating β_1 using sample

$$yield = \beta_0 + \beta_1 fertilizer + u$$

- you do not know β_0 and β_1, and would like to estimate them
- you observe a series of $\{yield_i, fertilizer_i\}$ combinations $(i = 1, \ldots, n)$
- you would like to estiamte β_1, the impact of fertilizer on yield, **ceteris paribus** (with everything else fixed)

Estimating β_1 using sample

Question

How could we possibly find the **ceteris paribus** impact of fertilizer on yield when we do not observe whole bunch of other factors (error term)?

Crucial conditions to identify the ceteris paribus impact

Before that...

You can always assume $E(u) = 0$ as long as an intercept is included in the model.

$$
\begin{aligned}
y =& \beta_0 + \beta_1 x + u_1, \quad \text{where} \quad E(u_1) = \alpha \\
=& \beta_0 + \alpha + \beta_1 x + u_1 - \alpha \qquad (1) \\
=& \gamma_0 + \beta_1 x + u_2, \qquad (2)
\end{aligned}
$$

where, $\gamma_0 = \beta_0 + \alpha$ and $u_2 = u_1 - \alpha$. Now, $E[u_2] = 0$.

Crucial conditions to identify the ceteris paribus impact

Remember

You are trying to find the **ceteris paribus** impact of x (fertilizer) on y (yield), while not observing whole bunch of other factors, u

It turns out, the following condition between x and u needs to be satisfied,

Mean independence (Important)

- mathematically:

$$E(u|x) = E(u)$$

- verbally: the average value of the unobservables is the same at any value of x, and that the common average is equal to the average of u over the entire population

Crucial conditions to identify the ceteris paribus impact

Combined with $E(u) = 0$

mean independence: $E(u|x) = E(u)$

$\Rightarrow$ zero conditional mean: $E(u|x) = 0$

Correlation and Mean Independence

In practice

We use correlation and mean independence interchangeably (though not entirely correct)

In the context of yield-fertilizer relationship,

$$yield = \beta_0 + \beta_1 fertilizer + u$$

Questions:

(Think about what you would do if you are a famer)

- What's in u?
- Is it correlated with fertilizer?

Going back to the college-income example

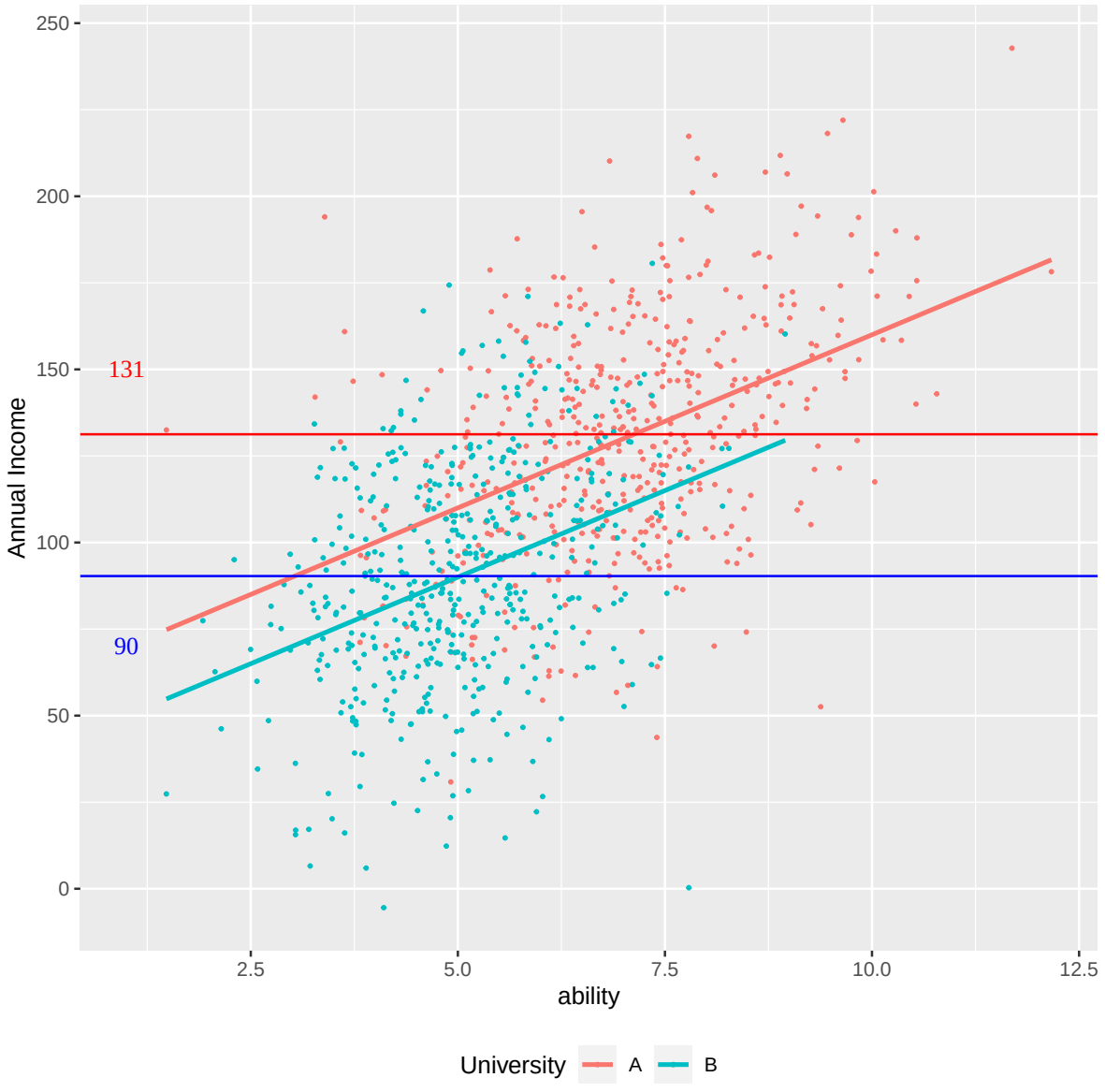
$$Income = \beta_0 + \beta_1 College\ \ A + u$$

where $College\ \ A$ is 1 if attending college A, 0 if attending college B, and u is the error term that includes ability.

Zero conditional mean?

$$E[u(ability)|collegeA] = 0?$$

Going back to the college-income example: $E[u|x] \neq 0$



Going back to the college-income example: $E[u|x] = 0$