AREC422 Jan31 Thursday Notes

Youpei Yan

January 31, 2019

1. Introduction of the course

1.1. The course, econometrics, describes economic relationships using statistical methods. It's similar to Statistics, but we add more economic meaning and interpretations into the relationships.

How does a college degree increase our wage? What level of precipitation is the best for the crop yield? How close to a garbage incinerator will my house price drop?

All of the things around us run with their own nature laws, and it would be interesting to describe their causal relationships and quantify these relationships with some data.

Our goals of the course: In the end, we can have some understandings of describing the causal relationship using econometric models. We can estimate the models and get the corresponding coefficients, and interpret the coefficients using economic jargon.

1.2. Grading method

Either >85 or the top 3 to 4 students will get A or A+
Either <50 or the bottom 1 to 2 students will get C- to C+
Anyone in between will get B- to B+.

1.3. Coding in R

Installation: download R first, then Rstudio. Rstudio is more like a user-friendly interface.

1.4. Textbook: Wooldridge's Introductory Econometrics, Ch1-7 & 9. Some statistic background can be found in the Appendices A to C.

1.5. HW submission

Coding part in R: email me directly
Calculation and interpretation question (regular econometric question): write down with steps (if needed).
The exam will be similar to HW questions, but no coding in R. I will show you the R result and ask you to interpret the meaning or do additional calculation based on it.

2. Basic concepts for today

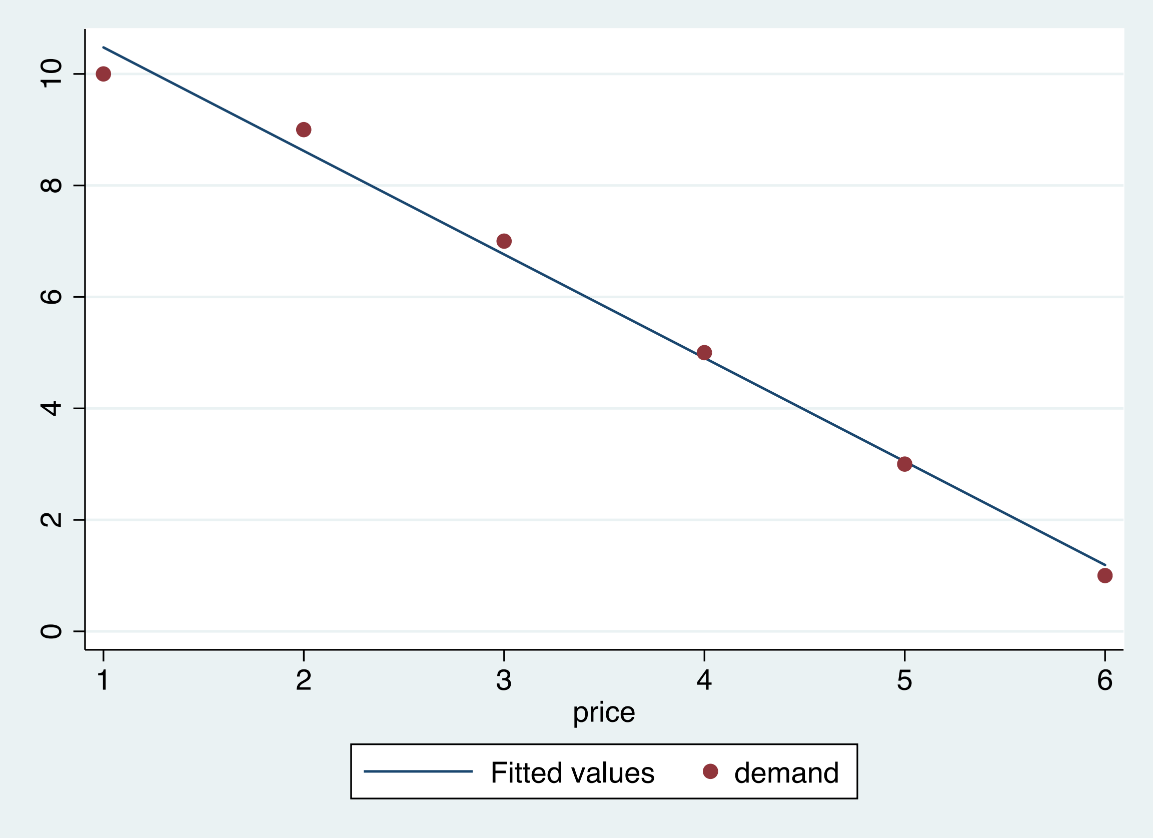
2.1. What's the meaning of using "linear" models to analyze data?

The simplest example in consumer theory: price and demand.

Say, we have the following data:

price	1	2	3	4	5	6
demand	10	9	7	6	3	1

We can plot the data:



The fitted line (in blue) gives us a relationship between demand and price: $\widehat{demand} = 12.40 - 1.83 price$

To summarize what we did here:
(1) We "make up" a linear relationship: $y = a + bx + error$
(2) We force the data points we have to be as close to this line as possible. (We want to find the best fit of the line. In the future, we'll know that it gives the smallest sum of the squared errors.)
(3) We find the slope and the intercept of the line. We call these two estimated numbers "coefficients" in our class.

This is an example of a Simple Linear Regression (with just one regressor on the right hand side (RHS).)

We can also complicate the model: add log functions, quadratic terms, more variables, variable interactions, etc. to better describe the data and the relationships.

Also, all the models have their flaws. We'll learn to compare them and choose between models. We'll perform hypothesis tests and use more advanced models if data permitted (for instance, panel data).

Some econometric terms for today: marginal effect and partial effect.

In the previous example, we have the following relationship: with 1 unit increase in price, the number of units you purchase decreases by 1.83.

Mathematically, Δ demand = -1.83 Δ price. We call this -1.83 "the marginal effect of the price", or the slope of the price.

If we have more regressors on the RHS: say, $demand = a + b_1 price + b_2 distance + error$, where *distance* is the distance to a store from your place, we'll have a Multiple Regression Model. Here, we have Δ demand = $b_1 \Delta$ price + $b_2 \Delta$ distance

If the store's location is *fixed*, we'll have the same "Δ demand = $b_1 \Delta$ price" relationship, but b_1 is called the partial effect here.

2.2. Statistics

Say, we have a variable x with n numbers in it: $\{x_1, x_2, ..., x_n\}$:

$$mean = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = E(x) = \mu$$

$$variance = var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 = E[(x-\mu)^2] = E(x^2) - E(x)^2$$

$$sd = \sqrt{var(x)}$$

$$correlation = corr(x,y) = \frac{E[(x-\mu_x)(y-\mu_y)]}{sd(x) \cdot sd(y)}$$

Sample vs. Population:

We always assume that our dataset is randomly drawn from a population distribution. We can never collect the whole world's population, we just use a *sample* of data to "best" estimate the world.

Say, $\{x_1, x_2, ..., x_n\}$ is a sample of x drawn from a population distribution. What's the best prediction of the population mean (μ, which is something unknown)?
Answer:
Sample mean. Or $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, which is an estimator. It is also an unbiased estimator because $E(\bar{x}) = \mu$ in theory.

What is the sample variance s^2?
Answer:

$$s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Note that it is not the same as the population variance σ^2:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

The sample variance loses one degree of freedom because it uses $\bar{x}$ to estimate μ first.