USC, Fall 2016, Economics 513

Lecture 3: The Classical Linear Regression Model I

Fitting linear relations and mathematical statistics

- OLS is a method that gives the best fitting linear relation between a dependent variable and a set of independent variables.
- Mathematical statistics develops methods for the analysis of data generated by a random experiment in order to learn about that random experiment.
- What is the random experiment that generates the observations on the dependent and independent variables?
- Least squares can be used as a pure fitting tool. It was developed as such in the late 18-s by Carl Friedrich Gauss and even in early econometrics, e.g. Tinbergen's work on Keynesian macro-models, there was no reference to mathematical statistics.

Is there randomness in economic relations?

- Earnings equation: relation between earnings and education, work experience, gender...
- Macro consumption function: relation between (national) consumption and (national) income.
- Any role for randomness?

Starting point: All economic relations are essentially deterministic, i.e. there is a set of independent variables $x_1, \ldots, x_W$ such that

$$y = f(x_1, \ldots, x_W)$$

Hence, if we have data $y_i, x_{i1}, \ldots, x_{iW}, i = 1, \ldots, n$ then

$$y_i = f(x_{i1}, \ldots, x_{iW}), i = 1, \ldots, n$$

Two issues

- The function f is possible nonlinear.
- We may not observe all W independent variables.

Nonlinearity

Let $\bar{x}_1, \ldots, \bar{x}_W$ be the sample averages of the variables and assume that f is sufficiently many times differentiable to have a Taylor series expansion around $\bar{x}_1, \ldots, \bar{x}_W$, i.e. a polynomial approximation:

$$y_i = \beta_0^* + \beta_1(x_{i1} - \bar{x}_1) + \cdots + \beta_W(x_{iW} - \bar{x}_W) + \cdots$$
$$\cdots + \gamma_1(x_{i1} - \bar{x}_1)^2 + \cdots + \gamma_W(x_{iW} - \bar{x}_W)^2 + \cdots$$
$$\cdots + \delta_1(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) + \cdots$$

with $\beta_0^* = f(\bar{x}_1, \ldots, \bar{x}_W)$.

We divide the independent variables $x_1, \ldots, x_W$ into three groups

1. Variables that do not vary in the sample (take this to be the last $W - V$ variables), i.e. for $i = 1, \ldots, n$, $x_{i,V+1} = \bar{x}_{V+1}, \ldots, x_{iW} = \bar{x}_W$. Example: gender if we consider a sample of women.
2. Variables in the relation that are omitted or cannot be included because they are unobservable. Let this be the next $V - K + 1$ variables.
3. Variables included in the relation, i.e. $x_1, \ldots, x_{K-1}$.

If we choose to include only the linear part we have for $i = 1, \ldots, n$

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{K-1} x_{i,K-1} + \varepsilon_i$$

with

$$\beta_0 = \beta_0^* - \sum_{j=1}^{K-1} \beta_j \bar{x}_j$$

The remainder term contains all the omitted terms

$$\varepsilon_i = \beta_K(x_{iK} - \bar{x}_K) + \cdots + \beta_V(x_{iV} - \bar{x}_V) +$$
$$+ \gamma_1(x_{i1} - \bar{x}_1)^2 + \cdots + \gamma_V(x_{iV} - \bar{x}_V)^2 +$$
$$+ \delta_1(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) + \cdots$$

We call ε_i the disturbance or error (of the exact linear relation).

Note that

$$\beta_j = \frac{\partial f}{\partial x_j}(\bar{x}_1, \ldots, \bar{x}_W), j = 1, \ldots, K-1$$

$$\beta_0 = f(\bar{x}_1, \ldots, \bar{x}_W) - \sum_{j=1}^{K-1} \beta_j \bar{x}_j$$

Therefore

1. The slope coefficient β_j is the partial effect of x_j on y.
2. If linear relation is an approximation, the partial effects depend on the average value of all variables that affect y. Only if the relation is truly linear (all quadratic etc. terms have 0 coefficients), the partial effects are independent of these average values.
3. Implication: if the relation is an approximation we expect the partial effects to be different in different populations, e.g. men versus women.
4. The intercept always depends on the average value of all variables and is unlikely to be 0 (even if the relation is truly linear).

Why the dependent variable is a random variable

Consider the following experiment: Prediction of y on the basis of $x_1, \ldots, x_{K-1}$.

If we have observed the values of $x_1, \ldots, x_{K-1}$, this does not tell us much about the disturbance ε that depends on (many) variables beside $x_1, \ldots, x_{K-1}$. Hence, even if we know the coefficients $\beta_0, \ldots, \beta_{K-1}$, we cannot predict with certainty what y is.

A variable with a value that is unknown before the experiment is performed is a random variable. We can think of the observations $y_i, x_{i1}, \ldots, x_{i,K-1}, i = 1, \ldots, n$ as the outcomes of n repetitions of a random experiment, in which (i) $x_{i1}, \ldots, x_{i,K-1}$ is drawn from some joint distribution (which does not play much of a role), (ii) y_i is generated by the linear relation

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{K-1} x_{i,K-1} + \varepsilon_i$$

for some realization of the random variable ε_i.

As in classical random experiments (flipping a coin, rolling a die) randomness is due to lack of knowledge. The outcome of a roll of a die is fully deterministic, if we know surface characteristics, angle, force, etc. However there are too many factors involved for an accurate prediction on the basis of these variables if we apply sufficient force.

Mean independence of disturbance and right hand side variables

In general, $x_1, \ldots, x_{K-1}$ does not tell us much about ε. We make the apparently extreme assumption that ε cannot be predicted using knowledge of $x_1, \ldots, x_{K-1}$.

Assumption 1: $E(\varepsilon | x_1, \ldots, x_{K-1}) = 0$

In words: the disturbance ε is mean-independent of $x_1, \ldots, x_{K-1}$, i.e. the mean of ε is the same whatever the values of the independent variables $x_1, \ldots, x_{K-1}$.

In other words: on average there is no relation between ε and the independent variables $x_1, \ldots, x_{K-1}$.

What happens if Assumption 1 does not hold?

We consider the consequence of such a failure in an example. Write the random error as (this is always possible)

$$\varepsilon_i = \beta_K(x_{iK} - \bar{x}_K) + \eta_i$$

and assume that x_K was not included, but is related to x_1 as

$$x_{iK} - \bar{x}_K = \gamma(x_{i1} - \bar{x}_1) + \zeta_i \qquad (1)$$

so that

$$\varepsilon_i = \beta_K \gamma(x_{i1} - \bar{x}_1) + \beta_K \zeta_i + \eta_i$$

Assumption 1 applies to η_i and ζ_i so that upon substitution

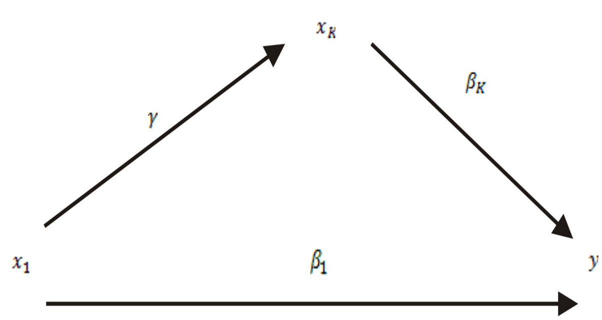
$$y_i = \beta_0 - \beta_K \gamma \bar{x}_1 + (\beta_1 + \gamma \beta_K) x_{i1} + \cdots + \beta_{K-1} x_{i,K-1} + \beta_K \zeta_i + \eta_i$$

Remember that

$$\beta_1 = \frac{\partial f}{\partial x_1}(\bar{x}_1, \ldots, \bar{x}_W)$$

so that if (1) holds and $\beta_K \neq 0$ then the coefficient of x_1 is not equal to the partial effect of x_1 on y.

Failure of Assumption 1 is a failure of the ceteris paribus condition in the sample: a change in x_1 has two effects on y , a direct effect β_1 and an indirect effect $\gamma\beta_K$. The latter effect is because in a sample we cannot hold other relevant variables like x_K fixed/constant. Hence we only measure the partial effect if the omitted variables are not related with x_1.



Measuring partial effects is the goal of most (but not all) empirical research in economics (and other social sciences), both because partial affects can be related to predictions made by economic theory, and because they have a causal interpretation. The biggest challenge in empirical research is to ensure that Assumption 1 holds.

There are cases that we are not necessarily interested in a partial effect. Consider a homeowner who is interested in the relation between house price and square footage of his/her house.

- If he/she wants to predict the sales price of the house only the strength of the relation between house price and square footage matters and there is no reason to be concerned about the interpretation of the regression coefficient as partial effect.
- If he/she wants to evaluate the investment in an addition to the house the estimation of the partial effect is essential.

Two strategies to estimate the partial effect of x_1

- Include all variables that are correlated with x_1 in the relation.
- Assign x_1 randomly, i.e. using a random experiment that is independent of anything, e.g. by flipping a coin if x_1 is dichotomous.

Empirical application

An important issue in labor economics is the effect of unearned income on work effort/labor earnings. Unearned income is e.g. income from assets, income of spouse or welfare benefits. If leisure is a normal good, we expect labor supply to go down if unearned income increases. How big is this income effect?

Notation: y_i is labor earnings of i (in a particular year), x_{i1} is unearned income of i (in a particular year)

Linear relation

$$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$$

with β_1 the income effect.
Usual approach to estimate β_1: use for x_{i1} spousal income or asset income. Are these variables potentially related to omitted variables in the relation?

Problems

- Asset income: Asset accumulation due to low preference for leisure.
- Spousal income: High because spouse prefers work and preferences of partners may be correlated.

Potential solution: Use randomly assigned unearned income.

Idea in 'Estimating the effect of unearned income on labor earnings, savings, and consumption: Evidence from a survey of lottery players', by Imbens, Rubin, and Sacerdote, AER, 2001.

- Data is sample of all people who won a major prize in the Megabucks lottery in Massachusetts in years 1984-1988. Survey was in 1996.
- Major prize is between \$22000 and \$9,696,000 and is paid out over 20 years.
- Assumption: amount of prize randomly assigned among the major prize winners.
- Prize possibly correlated with number of tickets bought. Effect of number of tickets on labor earnings?
- Bigger problem is non-response: of 802 winners there are usable data on 237. Response possibly correlated with amount of prize and other variables as gender, education and past earnings, so that among respondents the prize and these variables are correlated.

Variables

- *pearn* is the average of the yearly social security earnings in 6 years after winning (thousands of dollars).
- *xearn* yearly social security earnings in the 6 years before winning.
- *yearwon* year in which lottery was won.
- *tixbot* number of tickets in typical week at the time of winning.
- *agewon* age at which lottery was won.
- *male* indicator for gender.
- *educ* years of education.
- *workthen* indicator of working at the time of winning.
- *prize* yearly prize (one twentieth of total prize).

Sample statistics

	Mean	Std. dev.	Min	Max	Median
PEARN	10.9364	12.3520	0.0000	43.1016	5.9884743
AGEWON	46.9451	13.7970	23.0000	85.0000	47.000000
EDUC	12.9705	2.1909	8.0000	17.0000	13.000000
MALE	0.5781	0.4949	0	1	
PRIZE	55.1955	61.8035	1.1390	484.7890	31.748000
TIXBOT	4.5696	3.2820	0.0000	10.0000	4.0000000
WORKTHEN	0.8017	0.3996	0	1	
XEARN6	11.9651	11.7900	0.0000	36.7830	10.223730
XEARN5	12.1153	11.9923	0	39.2840	9.4733465
XEARN4	12.0374	12.0813	0	39.8737	8.6078591
XEARN3	12.8196	12.6539	0	40.3360	9.9304562
XEARN2	13.4787	12.9646	0	42.0000	10.786230
XEARN1	14.4676	13.6236	0	42.2570	12.530676
YEARWON	1986.0591	1.2940	1984.0000	1988.0000	1986.0000

Some OLS estimates

- I: No controls
- II: Controls
- III: Differenced earnings, no controls

	I	II	III
CONST	12.38	2347.58	-0.757
PRIZE	-0.0262	-0.0420	-0.0503
AGEWON		-0.184	
EDUC		-0.0275	
MALE		1.212	
TIXBOT		-0.00406	
WORKTHEN		1.900	
XEARN6		-0.143	
XEARN5		-0.118	
XEARN4		0.0396	
XEARN3		0.390	
XEARN2		0.0772	
XEARN1		0.286	
YEARWON		-1.176	

This is the effect of a temporary (20 year) increase in unearned income. Smaller effect than a permanent change. IRS suggest to multiply by 1.1 to obtain effect of permanent change.