Econometric

Topic 2

Universidad Carlos III

Getafe, Madrid

UC3M

The Simple Linear Regression Model// Definition and Elements

Goal: Basic Econometric Model. Some basic problems that we must address first:

- How do we allow or address that the relationship between x and y is not perfect (deterministic), How do we allow that other factors affect y? There is never an exact relationship between two variables so we need to allow for the remaining factors that affect the endogenous variable
- Which is functional form that related x and y? The functional relationship between the endogenous (dependent) and exogenous (independent/explanatory) variables needs to be specified.
- How do we ensure that we are estimating a ceteris paribus relationship?

Example: saving rate and family income

- We want to study the relationship between family's savings (Y) and family income (X).
- We have data for 1027 families in the US for the period 1960–62 (Goldberger, Chapter 1).
- We have discrete value for Y and X and whose joint distribution can be represented for the following table:

Example: saving rate and family income

Joint Distribution X and Y

$P(X,Y)$	X (Income Thousand of Dollars)					
Y (saving rate)	**1.4**	**3.0**	**4.9**	**7.8**	**14.2**	$P(Y)$ (Row sum)
0.45	0.015	0.026	0.027	0.034	0.033	*0.135*
0.18	0.019	0.032	0.057	0.135	0.063	*0.306*
0.05	0.059	0.066	0.071	0.086	0.049	*0.331*
-0.11	0.023	0.035	0.045	0.047	0.015	*0.165*
-0.25	0.018	0.016	0.016	0.008	0.005	*0.063*
$P(X)$ (Column sum)	*0.134*	*0.175*	*0.216*	*0.310*	*0.165*	1.000

Example: saving rate and family income

- Can we observe a determinist relationship between savings and income?
- We have a distribution of values for each income level. That is, for each level of income there are some families that save and others are getting into debts.
- Nevertheless, we observe a higher proportion of families saving as we move to higher income levels.
- In order to see that we will use the conditional distributions.

Example: saving rate and family income

Conditional Distribution of Y
for each value of X

$P(Y\|X)$	X (family income in thousand of dollars)				
Y (saving rate)	**1.4**	**3.0**	**4.9**	**7.8**	**14.2**
0.45	0.112	0.149	0.125	0.110	0.200
0.18	0.142	0.183	0.264	0.435	0.382
0.05	0.440	0.377	0.329	0.277	0.297
-0.11	0.172	0.200	0.208	0.152	0.091
-0.25	0.134	0.091	0.074	0.026	0.030
Cond. Mean $\widehat{\mu}_{Y\|X}$	**0.045**	**0.074**	**0.079**	**0.119**	**0.156**

Example: saving rate and family income

- ... then how do reconciliate theory (deterministic relationships) and empirical evidence (stochastic relationships).
- when theory will state a relationship between Y and X, we should understand that mean value of Y is a function of X.
- what does this mean in terms of our previous example.

Before ... let's talk a little about prediction.

- Given the joint distribution of (Y, X) we are asked about the value of Y for an individuals selected randomly from the population.
- Let assume that we have as target to measure the *quality* of our prediction $c(X)$ minimize $E(U^2)$ and $U = Y - c(X)$ as the error in the prediction.

Best constant prediction.

- We cannot make use of information about X. That is, we can use only information about Y in order to select a prediction c
- Which information do we have?

Y (saving rate)	$P(Y)$
0.45	0.135
0.18	0.306
0.05	0.331
-0.11	0.165
-0.25	0.063

- Then the prediction error is $U = Y - c$ and selecting best constant prediction, c, means looking for a value of c that minimize $E(U^2) = \sum_k (Y_k - c)^2 p_k$

Best constant prediction.

- Therefore the best constant prediction is $c = E(Y) = \mu_Y$
- Assuming that in our previous example the distribution correspond to the population one, the best constant predictions is:

$$
\begin{aligned}
E(Y) &= 0.45 \times 0.135 + 0.18 \times 0.306 + 0.05 \times 0.331 \\
&\quad -0.11 \times 0.165 - 0.25 \times 0.063 \\
&= 0.09848 = 9.85\%
\end{aligned}
$$

Best linear prediction.

- Now we can use information of X but in specific

$$c(X) = c_0 + c_1 X,$$

  with c_0 and c_1 constants.

- Then the prediction error will be $U = Y - c_0 - c_1 X$. As we did before, we will choose c_0 and c_1 in order to minimize $E(U^2) = \sum_k (Y_k - c_0 - c_1 X)^2 p_k$.
- We can show that,

$$\begin{aligned} c_0 &= \alpha_0 = E(Y) - \beta E(X) = \mu_Y - \alpha_1 \mu_X, \\ c_1 &= \alpha_1 = \frac{C(X,Y)}{V(X)} = \frac{\sigma_{XY}}{\sigma_X^2}. \end{aligned}$$

Best linear prediction.

- **The linear function** $\alpha_0 + \alpha_1 X$ is the **linear projection (**or **linear prediction)** of Y given X

$$L(Y \mid X) = \alpha_0 + \alpha_1 X$$

- In our example (saving-income) we will need to calculate $V(X) = E(X^2) - E(X)^2, E(X), E(Y)$ and

$$C(X, Y) = E(XY) - E(X)E(Y)$$

Best linear prediction.

- In specific in order to calculate $E(X * Y) = \sum_{i=1}^{5} \sum_{j=1}^{5} X_i Y_j \Pr(XY = X_i Y_i) = 0.782607$ which comes from the following table,

Marginal distribution of XY

XY	$P(XY)$	XY	$P(XY)$	XY	$P(XY)$	XY	$P(XY)$
-3.55	0.005	**-0.75**	0.016	**0.07**	0.059	**0.54**	0.032
-1.95	0.008	**-0.54**	0.045	**0.15**	0.066	**0.63**	0.015
-1.56	0.015	**-0.35**	0.018	**0.25**	0.071	**0.71**	0.049
-1.23	0.016	**-0.33**	0.035	**0.25**	0.019	**0.88**	0.057
-0.86	0.047	**-0.15**	0.023	**0.39**	0.086	**1.35**	0.026

Best linear prediction.

- and

$$E(X) = 1.4 \times 0.134 + 3.0 \times 0.175 + 4.9 \times 0.216$$
$$+7.8 \times 0.310 + 14.2 \times 0.165 = 6.532$$

therefore,

$$C(X, Y) = 0.782607 - 6.532 \times 0.09848 = 0.13934.$$

- Then,

$$E\left(X^2\right) = 1.4^2 \times 0.134 + 3.0^2 \times 0.175 + 4.9^2 \times 0.216$$
$$+7.8^2 \times 0.310 + 14.2^2 \times 0.165 = 59.155$$

therefore

$$V(X) = E\left(X^2\right) - [E(X)]^2 = 59.155 - 6.532^2 = 16.488$$

Best linear prediction.

- finally based on these series of moments

$$
\begin{aligned}
c_1 &= \alpha_1 = \frac{C(X,Y)}{V(X)} = \frac{0.13934}{16.488} = 0.008451 \\
c_0 &= \alpha_0 = E(Y) - \beta E(X) = 0.09848 - 0.008451 \times 6.532 = 0.04327
\end{aligned}
$$

Then the linear projection in our example is

$$L(Y \mid X) = 0.043278 + 0.008451X$$

- but for the discrete values of X, we can write the linear projection as

$$
L(Y|X) = \begin{cases}
0.043278 + 0.008451 \times 1.4 = 0.055 & \text{if } X = 1.4 \\
0.043278 + 0.008451 \times 3.0 = 0.069 & \text{if } X = 3.0 \\
0.043278 + 0.008451 \times 4.9 = 0.085 & \text{if } X = 4.9 \\
0.043278 + 0.008451 \times 7.8 = 0.1092 & \text{if } X = 7.8 \\
0.043278 + 0.008451 \times 14.2 = 0.1633 & \text{if } X = 14.2
\end{cases}
$$

Best prediction.

- Now let assume that we can use any function of X.
- As we did before, we will choose $c(X)$ such as we minimize $E(U^2)$. The results is $c(X) = E(Y \mid X)$.
- The **best predictor** of Y given X is its **conditional mean**, $E(Y|X)$.
  - Only when the conditional mean is linear, the linear projection $L(Y|X)$ and the conditional mean $E(Y|X)$ are the same.
  - Nevertheless, when the conditional mean is not linear the linear projection is not the best predictor, **the best linear approximation** to the conditional mean.

Best prediction.

	Prediction for the saving rate		
X (income in thousand of dollars)	C	$L(Y \mid X)$	$E(Y \mid X)$
1.4	0.0985	0.055	0.045
3.0	0.0985	0.069	0.074
4.9	0.0985	0.085	0.079
7.8	0.0985	0.1092	0.119
14.2	0.0985	0.1633	0.156

Best prediction.

- As we observe in the previous table, $L(Y|X) \neq E(Y|X)$
- Nevertheless $L(Y|X)$ provide a good approximation to $E(Y|X)$. Therefore $L(Y|X)$ can be used in some circumstances as a good predictor although does not match the $E(Y|X)$.
- BUT while $E(Y|X)$ characterize population moments (conditional means) of Y given X, the $L(Y|X)$ NOT.
- This last imply that while $E(Y|X)$ can have causal interpretation, the $L(Y|X)$ NOT.

The Simple Linear Regression Model

- Consider the following equation which is assumed to hold in the population of interest:
- $Y = \beta_0 + \beta_1 X + u$
- Does this relation address the problems we defined in the previous slide?.
- Element of the Model:
  - Variables (y, x) and Error Term (u).
  - Functional Relationship.
  - Parameters

The Simple Linear Regression Model. Functional Form

- If we hold the rest of the factors in u fixed so that the change in u is zero, $\Delta u = 0$, then x has linear impact on y
- $\Delta y = \beta_1 \Delta x$ if $\Delta u = 0$
- Is Linearity a realistic assumption in Economics?

The Simple Linear Regression Model. Parameters.

- β_1: denotes the slope parameter in the relationship between y and x holding everything else constant (factors in u fixed). When we multiply it for the change in x denotes a change y associate to the change in x. It is the key parameter when we study the relationship between x and y.
- β_0: is known as the intercept parameter (i.e. it is the value of y when x and u are equal to zero)
- Example Mincer Equation

Analysis Ceteris paribus?

- β_1: Impact of x on y keeping everything else (captured in u) Constant. But, How can we keep these other factors constant in order to get this conclusion?
- To get reliable estimates of the model parameters(β_0 y β_1) we have to make an assumption restricting the manner in which x is related to the error term u. Otherwise we will not be able to estimate the ceteris paribus effect.
- Since x and u are random variables we need a concept based on their distribution.

Constant Term: β_0?

- Initial Assumption: As long as the intercept, β_0, is included in the model, nothing is lost by assuming that the average value of u in the population is zero:

$$E(u) = 0.$$

  - The assumption in does not say anything about the relationship between u and x. It simply makes a statement about the distribution of the unobservables in the population
  - It is only a normalization: we rescale the impact of other factors to be zero.

The Simple Linear Regression Model. Relationship between x and u

- We define the independence of x and u from the perspective of the conditional distribution of u given x :

  $E(u|x) = E(u) = 0.$
- For all possible values of x, the average of u is always the same, 0.
- What does it force us to buy in the context of the Mincer Equation?
- Example: Fertilizer

Implication

- The assumption(s) $E(u|x) = E(u) = 0$. implies another useful interpretation of the model. Taking the expected value of y conditional on x,

  $E(y|x) = \beta_0 + \beta_1 X$
- This expression states that the population regression function, $E(y|x)$, (PRF) is a linear function of x Its estimated counterpart is known as the sample regression function (SRF).
- We can also write it as

  $y = E(y|x) + u = \beta_0 + \beta_1 X + u$,
  where $E(y|x) = \beta_0 + \beta_1$ is the explained part by x and u is the unexplained part x.

Assumptions in the Simple Linear Regression (SLR).

- Assumption SLR1: Linear in Parameters
- A SLR2: We have a random sample of size $N, \{(x_i, y_i) : i : 1, 2 \ldots, N\}$, following the population model.
- A SLR3: The sample variable x is not constant. We will see the next class the need of this assumption.
- A SLR4: The error u has an expected value of zero given any value of the explanatory variable.
  $E(u|x) = E(u) = 0.$

- $V(u|X) = \sigma^2 \Rightarrow$
  - $V(u) = \sigma^2$
  - $V(Y|X) = \sigma^2$

Estimation.

- Our goal is estimating population parameters from the available data.
- Giving the previous assumptions and the model

$$Y = \beta_0 + \beta 1X + u$$

$$E(u|x) = 0.$$
$$V(u|x) = \sigma^2$$

and with our data $(y_i^*, x_i^*) i = 1, \ldots, n$ as a random sample of size n from the population of interest

- ...we can write our problem as

$$Y_i = \beta_0 + \beta 1X_i + u_i$$

$$E(u_i|x) = 0.$$
$$V(u_i|x) = \sigma^2$$

Estimation. The analogy principle

- The parameters of interest are population characteristics and they are function of population moments. The principle of analogy consists in using as estimator the sample moment.
- Giving the previous assumptions made for our SLR model

$$E(Y|X) = L(Y|X) = \beta_0 + \beta_1 X$$

with β_0, β_1 as the solution of solving the minimization problem

$E(u^2) = E[(Y - \beta_0 - \beta_1 X)^2]$ and
$\beta_0 = E(Y) - \beta_1 E(X)$
$\beta_1 = \frac{C(X,Y)}{V(X)}$

- Applying the analogy principle we replace in the previous expressions the population moments by the sample ones
$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
$\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{S_{XY}}{S_X^2}$

The OLS principle.

- OLS can be also motivated as the sample analog of the population problem:
  $minE(u^2)$ or

  $min[E(Y - \beta_0 - \beta_1 X)^2]$
- What is the sample analog for the error term (difference between the observed and expected value)? The residual.
  $e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$
- therefore, the sample analog for $minE(u^2)$
  $min_{\hat{\beta}_0, \hat{\beta}_1} \sum e_i^2$
- What are the FOC of this problem?

- $\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X} = \sum_i Y_i - \widehat{\beta}_1 \overline{X}$
- $\widehat{\beta}_1 = \frac{\sum_i x_i Y_i}{\sum_i x_i^2} = \sum_i c_i Y_i, \quad$ where $x_i = X_i - \overline{X}$, $c_i = \frac{x_i}{\sum_i x_i^2}$

Properties of OLS estimators. Unbiasness

- This property holds when the assumptions of linearity and $E\left(\varepsilon|\, X\right) = 0$
  $E\left(\widehat{\beta}_0\right) = \beta_0$ (see problem set 2)
  $E\left(\widehat{\beta}_1\right) = \beta_1$

$$
\begin{aligned}
\widehat{\beta}_1 &= \sum_i c_i Y_i = \sum_i c_i \left(\beta_0 + \beta_1 X_i + \varepsilon_i\right) \\
&= \beta_0 \sum_i c_i + \beta_1 \sum_i c_i X_i + \sum_i c_i \varepsilon_i
\end{aligned}
$$

but

$$
\begin{aligned}
\sum_i c_i &= \frac{1}{\sum_i x_i^2}\left(\sum_i x_i\right) = 0 \text{ because } \sum_i x_i = \sum_i X_i - n\overline{X} = 0 \\
\sum_i c_i X_i &= \sum_i \frac{x_i X_i}{\sum_i x_i^2} = \frac{1}{\sum_i x_i^2} \sum_i x_i^2 = 1
\end{aligned}
$$

Properties of OLS estimators. Unbiasness

- Therefore

$$\widehat{\beta}_1 = \beta_1 + \sum_i c_i \varepsilon_i$$

and

$$\begin{aligned} E\left(\widehat{\beta}_1 \middle| X\right) &= \beta_1 + E\left(\textstyle\sum_i c_i \varepsilon_i \middle| X\right) = \quad \beta_1 + \textstyle\sum_i c_i \underbrace{E\left(\varepsilon_i \middle| X\right)}_{=0} \\ &= \beta_1 \end{aligned}$$

OLS estimators. Variance

- Additional to the assumptions of linearity and $E(\varepsilon|X) = 0$ we will assume ($V(\varepsilon|X) = \sigma^2$ for every X).
- $V\left(\widehat{\beta}_0\right) = \left(\sum_i X_i^2 / n\right) V\left(\widehat{\beta}_1\right)$ (see problem set)
- $V\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{n} E\left(\frac{1}{S_x^2}\right)$.

OLS estimators. Variance

- $$\begin{aligned}
V\left(\widehat{\beta}_1\right) &= E\left[\left(\widehat{\beta}_1-\beta_1\right)\right]^2 = E\left(\sum_i c_i\varepsilon_i\right)^2 = \sum_i E\left(c_i^2\varepsilon_i^2\right) \\
&= E\left[\sum_i E\left(c_i^2\varepsilon_i^2\middle| X\right)\right] = E\left[\sum_i c_i^2 E\left(\varepsilon_i^2\middle| X\right)\right] \\
&= \sigma^2 E\left[\sum_i c_i^2\right] \text{ (by Homoscedasticity)} \\
&= \sigma^2 E\left[\sum_i\left(\frac{x_i}{\sum_i x_i^2}\right)^2\right] = \sigma^2 E\left[\frac{1}{\left(\sum_i x_i^2\right)^2}\sum_i x_i^2\right] \\
&= \sigma^2 E\left[\frac{\frac{1}{n}}{\frac{1}{n}\left(\sum_i x_i^2\right)}\right] = \frac{\sigma^2}{n} E\left(\frac{1}{S_X^2}\right)
\end{aligned}$$

The Gauss-Markov Theorem

- In the context of the linear regression model, and under the assumptions of linearity, $E(\varepsilon|X) = 0$, and Homoscedasticity, $\hat{\beta_0}$ and $\hat{\beta_1}$ are the one with the lowest variance among the linear and unbiased estimators.
- Therefore when the assumptions os the classical model are holding, the OLS estimator is the efficient among the family of linear and unbiased estimator.

Consistency

- The OLS estimators $\hat{\beta_0}$ and $\hat{\beta_1}$ are consistent estimators of β_0 and β_1.

$$\text{p}\lim_{n\to\infty} \widehat{\beta}_j = \beta_j, \qquad j = 0, 1.$$

that is:

$$\lim_{n\to\infty} \Pr\left(\left|\widehat{\beta}_j - \beta_j\right| < \delta\right) = 1, \qquad \forall\delta > 0$$

Estimation of the variance

- The variance of of the OLS estimators, $\widehat{\beta}_0$ and $\widehat{\beta}_1$, depends on $\sigma^2 = V(\varepsilon) = E(\varepsilon^2)$.
- The problem we face is the fact that ε_i $(i = 1, \ldots, n)$ is unobserved.
- After estimating the model we observe the residuals $\widehat{\varepsilon}_i$:

$$\begin{aligned} \widehat{\varepsilon}_i &= Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i = (\beta_0 + \beta_1 X_i + \varepsilon_i) - \widehat{\beta}_0 - \widehat{\beta}_1 X_i \\ &= \varepsilon_i - \left(\widehat{\beta}_0 - \beta_0\right) - \left(\widehat{\beta}_1 - \beta_1\right) X_i \qquad (i = 1, \ldots, n) \end{aligned}$$

- Although $E\left(\widehat{\beta}_0\right) = \beta_0$, $E\left(\widehat{\beta}_1\right) = \beta_1$, $\widehat{\varepsilon}_i \neq \varepsilon_i$. Additionally, $E(\widehat{\varepsilon}_i - \varepsilon_i) \neq 0$.
- For a sample of size n, the errors ε_i $(i = 1, \ldots, n)$, a natural estimator for σ^2 would be the sample analogous $E\left(\varepsilon^2\right)$, that is, $\frac{1}{n}\sum_i \varepsilon_i^2$. But this estimator is not feasible.

Estimation of the variance

- If we replace in the previous expression the sample analogous, the residuals, we could use as estimator of σ^2:

$$\widetilde{\sigma}^2 = \frac{\sum_i \widehat{\varepsilon}_i^2}{n}.$$

- This estimator is feasible but bias. The reason is the fact that residuals were obtained by imposing 2 linear restrictions, $\frac{1}{n}\sum_i \widehat{\varepsilon}_i = 0$ and $\frac{1}{n}\sum_i \widehat{\varepsilon}_i x_i = 0$, therefore there are only $(n-2)$ independent residuals (which is known as **degree of freedom**).
- Alternatively, we could use an unbias estimator:

$$\widehat{\sigma}^2 = \frac{\sum_i \widehat{\varepsilon}_i^2}{n-2}.$$

- Both $\widetilde{\sigma}^2$ and $\widehat{\sigma}^2$ are consistent estimators of σ^2.

Estimation of the variance

- $V\left(\widehat{\beta}_0\right) = \left(\sum_i X_i^2 / n\right) V\left(\widehat{\beta}_1\right)$
- $V\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{n} E\left(\frac{1}{S_x^2}\right)$.
- As estimator of $V\left(\widehat{\beta}_1\right)$, we estimate $E\left(\frac{1}{S_x^2}\right)$ by $\frac{1}{S_x^2}$ as well a consistent estimator of σ^2:

$$\widehat{V}\left(\widehat{\beta}_1\right) = \frac{\widehat{\sigma}^2}{n S_x^2}$$

- Therefor in order to estimate $V\left(\widehat{\beta}_0\right)$,

$$\widehat{V}\left(\widehat{\beta}_0\right) = \frac{\widehat{\sigma}^2 \sum_i X_i^2}{n^2 S_x^2}$$

Goodness of fit

- The first intuitive measure of goodness of fit is using the sample analogous of $E\left(\varepsilon^2\right)$, that is $\widehat{\sigma}^2$ as a measure of failure or success of the model.
- Also is common the use of the square root of $\widehat{\sigma}^2$, $\widehat{\sigma}$, as measure of goodness o fit. This measure is addressed as the standard error of the regression.
- Probably a more popular measure of fit is the R^2 or **determination coefficient**, defined as

$$R^2 = \frac{\sum_i \widehat{y}_i^2}{\sum_i y_i^2} = 1 - \frac{\sum_i \widehat{\varepsilon}_i^2}{\sum_i y_i^2},$$

where $y_i = Y_i - \overline{Y}_i$, $\widehat{y}_i = \widehat{Y}_i - \overline{Y}_i$, $\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i$.

- The R^2 is interpreted as the proportion of the sample variance of Y explained by the model.
- The R^2 satisfy that $0 \leq R^2 \leq 1$.

- it can be shown that

$$R^2 = \widehat{\rho}_{Y\widehat{Y}} = \left(\frac{S_{Y\widehat{Y}}}{S_Y S_{\widehat{Y}}}\right)^2 = \frac{\frac{1}{n}\sum_i \left(Y_i - \overline{Y}\right)\left(\widehat{Y}_i - \overline{Y}\right)}{\sqrt{\frac{1}{n}\sum_i \left(Y_i - \overline{Y}\right)^2}\sqrt{\frac{1}{n}\sum_i \left(\widehat{Y}_i - \overline{Y}\right)^2}}$$

that is: the R^2 is the square of the sample correlation coefficient between Y_i and $\widehat{Y}_i$.