Econometrics 1

Chapter 1: The Fundamentals of Linear Regressions

Xavier D'Haultfœuille and Elia Lapenta

CREST-ENSAE

- We are interested in predicting a variable $Y \in \mathbb{R}$ by other variables $X = (X^1, \ldots, X^k)' \in \mathbb{R}^k$.
- Important: X is a column vector. We denote with X^j (and not X_j) the jth component of X.
- X=covariates, explanatory variables, independent variables.
- Y=outcome, explained variable, dependent variable, response variable.
- We study here "the (linear) regression of Y on X", in particular its definition and basic properties.
- We assume to have cross-sectional data of n units. In particular, we assume the sample $(X_i, Y_i)_{i=1\ldots n}$ to be i.i.d., with $(X_i, Y_i) \sim (X, Y)$.

Outline

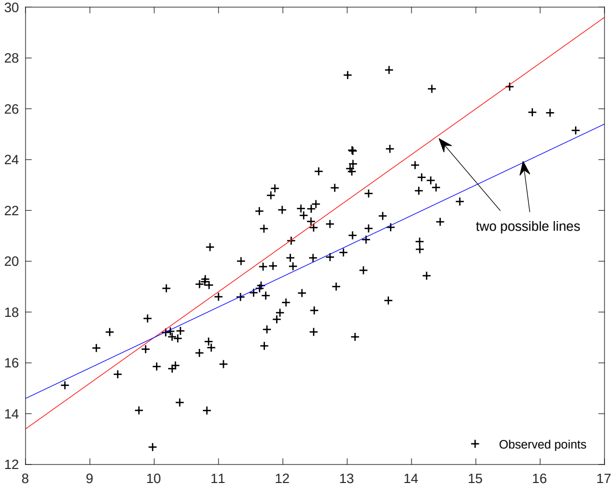
The OLS estimator

- We begin by the simple case where $k = 2$: $X = (1, D)'$, where $D \in \mathbb{R}$.
- Assume hereafter that $(D_1, ..., D_n)$ are not all equal.
- Then the OLS estimator $(\widehat{\alpha}, \widehat{\beta}_D)$ in the "regression of Y on D" is defined as:

$$\left(\widehat{\alpha}, \widehat{\beta}_D\right) = \arg \min_{(a,b)\in\mathbb{R}^2} \sum_{i=1}^{n} (Y_i - a - D_i b)^2. \tag{1}$$

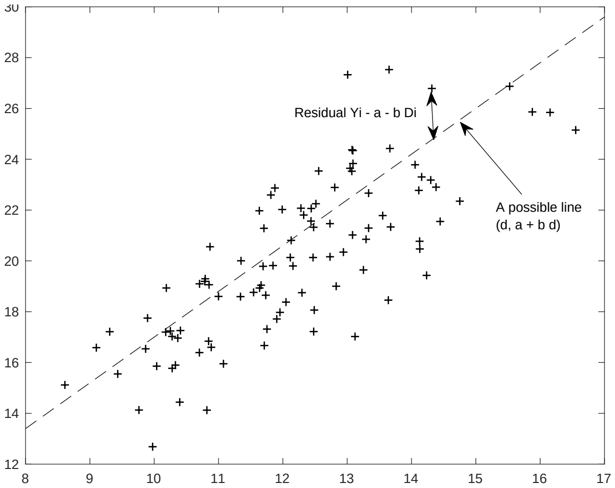
  As we shall see, the minimum does exist and is unique.
- Let $\widehat{Y}_i := \widehat{\alpha} + D_i \widehat{\beta}_D$. $\widehat{Y}_i$ is called the predicted value of Y_i (importantly: non-causal prediction!).
- Then $(\widehat{Y}_1, ..., \widehat{Y}_n)$ is the best linear approximation (with the Euclidean norm) of $\boldsymbol{Y} = (Y_1, ..., Y_n)'$ based on the vector $\boldsymbol{D} = (D_1, ..., D_n)'$.
- $\widehat{\varepsilon}_i := Y_i - \widehat{Y}_i$ is called the residual of obs. i.
- $d \mapsto \widehat{\alpha} + \widehat{\beta}_D d$ is called the regression line.

Geometric interpretation



Geometric interpretation

Among all lines $y = a + bd$, that with $(a, b) = (\widehat{\alpha}, \widehat{\beta}_D)$ is minimizing the sum of the squares of the residuals $Y_i - a - bD_i$.



Properties of the OLS estimator

- For any random variables (r.v.) A, B, (and $(A_i, B_i)_{i=1,\dots,n}$ an iid sample with $(A_i, B_i) \sim (A, B)$) we let hereafter:

$$\overline{A} = \frac{1}{n}\sum_{i=1}^{n} A_i,$$

$$\widehat{V}(A) = \frac{1}{n-1}\sum_{i=1}^{n}(A_i - \overline{A})^2,$$

$$\widehat{\text{Cov}}(A, B) = \frac{1}{n-1}\sum_{i=1}^{n}(A_i - \overline{A})(B_i - \overline{B})$$

Proposition 1

Assume that $(D_1, ..., D_n)$ are not all equal. Then:

1. $\left(\widehat{\alpha}, \widehat{\beta}_D\right)$ *are well-defined and satisfy*

$$\widehat{\beta}_D = \frac{\widehat{Cov}(D, Y)}{\widehat{V}(D)}, \ \widehat{\alpha} = \overline{Y} - \overline{D}\widehat{\beta}_D.$$

2. $Y_i = \widehat{\alpha} + \widehat{\beta}_D D_i + \widehat{\varepsilon}_i$, *with* $\overline{\widehat{\varepsilon}} = \overline{D\widehat{\varepsilon}} = 0$.

Proof of Proposition 1

1. Let $f(a,b) = \sum_{i=1}^n (Y_i - a - D_i b)^2$. Its hessian H satisfies

$$H = 2\begin{pmatrix} n & \sum_{i=1}^n D_i \\ \sum_{i=1}^n D_i & \sum_{i=1}^n D_i^2 \end{pmatrix} >> 0 \text{ (viz., positive definite)},$$

since $\sum_{i=1}^n (D_i - \overline{D})^2 > 0$ by assumption. Hence, f is strictly convex. Thus, (1) has at most one solution given by the first-order conditions (FOC)

$$\sum_{i=1}^n (Y_i - \widehat{\alpha} - D_i \widehat{\beta}_D) = 0,$$

$$\sum_{i=1}^n D_i (Y_i - \widehat{\alpha} - D_i \widehat{\beta}_D) = 0. \qquad (2)$$

Equivalently, $\widehat{\alpha} = \overline{Y} - \overline{D}\widehat{\beta}_D$ and $\widehat{\beta}_D = \widehat{\text{Cov}}(D, Y)/\widehat{V}(D)$.

2. Notice that by definition of $\widehat{\varepsilon}_i$ and the first equality in the FOC's we have

$$\overline{\widehat{\varepsilon}} = \overline{Y - \widehat{\alpha} - D\widehat{\beta}_D} = \overline{Y} - \widehat{\alpha} - \overline{D}\widehat{\beta}_D = 0.$$

$\overline{D\widehat{\varepsilon}} = 0$ follows directly from (2) $\square$

- Often, D_i is binary, $D_i \in \{0,1\}$.
- Then let n_d =card$\{i : D_i = d\}$ and let $\overline{Y}_d = \frac{1}{n_d}\sum_{i:D_i=d} Y_i$ (average of Y for those s.t. $D_i = d$).
- Then $\overline{Y} = \overline{D} \times \overline{Y}_1 + (1-\overline{D})\overline{Y}_0$. Thus:

$$
\begin{aligned}
\widehat{\beta}_D &= \frac{\widehat{\mathrm{Cov}}(D,Y)}{\widehat{V}(D)} \\
&= \frac{\overline{DY} - \overline{D} \times \overline{Y}}{\overline{D^2} - \overline{D}^2} \\
&= \frac{\overline{D}\,\overline{Y}_1 - \overline{D}\left(\overline{D}\,\overline{Y}_1 + (1-\overline{D})\overline{Y}_0\right)}{\overline{D}(1-\overline{D})} \\
&= \overline{Y}_1 - \overline{Y}_0.
\end{aligned}
$$

- With a similar reasoning, we obtain $\widehat{\alpha} = \overline{Y}_0$.
- Intuitive: we predict Y_i by $\widehat{\alpha} = \overline{Y}_0$ if $D_i = 0$, and by $\widehat{\alpha} + \widehat{\beta}_D = \overline{Y}_1$ if $D_i = 1$.

Other properties of the OLS estimator

- Point 1 of Proposition 1 implies that $(\overline{D}, \overline{Y})$ is on the estimated regression line.
- Point 2 of Proposition 1 implies that in the sample, residuals are uncorrelated with predicted values:

$$\begin{aligned}\widehat{\text{Cov}}(\widehat{Y}, \widehat{\varepsilon}) =& \frac{1}{n-1}\sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{Y}\right)\widehat{\varepsilon}_i \\ =& \frac{\widehat{\beta}_D}{n-1}\sum_{i=1}^{n}\left(D_i - \overline{D}\right)\widehat{\varepsilon}_i \\ =& 0.\end{aligned}$$

- Because $Y = \widehat{Y} + \widehat{\varepsilon}$, we have the following variance decomposition:

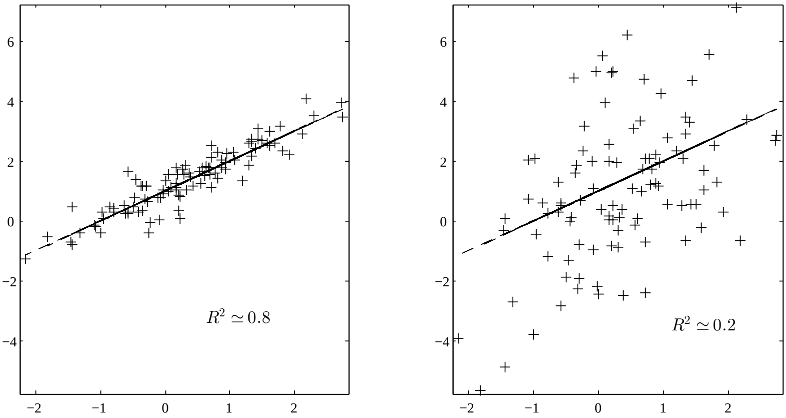
$$\widehat{V}(Y) = \widehat{V}(\widehat{Y}) + \widehat{V}(\widehat{\varepsilon}). \tag{3}$$

- Effect of a location or scale change in D or Y on the OLS estimator?
- If $Y' = Y + c$, then $\widehat{\beta}'_D = \widehat{\beta}_D$ and $\widehat{\alpha}' = \widehat{\alpha} + c \Rightarrow \widehat{Y}' = \widehat{Y} + c$.
- If $Y' = cY$, then $\widehat{\beta}'_D = c\widehat{\beta}_D$ and $\widehat{\alpha}' = c\widehat{\alpha} \Rightarrow \widehat{Y}' = c\widehat{Y}$.
- If $D' = D + c$, then $\widehat{\beta}'_D = \widehat{\beta}_D$ et $\widehat{\alpha}' = \widehat{\alpha} - c\widehat{\beta}_D \Rightarrow \widehat{Y}' = \widehat{Y}$.
- If $D' = cD$, then $\widehat{\beta}'_D = \widehat{\beta}_D / c$ et $\widehat{\alpha}' = \widehat{\alpha} \Rightarrow \widehat{Y}' = \widehat{Y}$.
- Similar rules if we apply affine transforms to Y or D (e.g., $Y' = c_0 + c_1 Y$).

Quality of the prediction

- Let us denote $\widehat{Corr}(A, B) := \widehat{Cov}(A, B)/\sqrt{\widehat{V}(A)\ \widehat{V}(B)}$
- To know whether D predicts accurately Y, we often compute the R^2:
$$R^2 := \frac{\widehat{V}(\widehat{Y})}{\widehat{V}(Y)} = \widehat{\text{Corr}}(Y, \widehat{Y})^2 \in [0, 1] \text{ (by (3))}.$$
- Part of the variance of Y that is explained by (linear functions of) D.
- If $R^2 = 1$, the prediction is perfect $(\widehat{\varepsilon}_1 = ... = \widehat{\varepsilon}_n = 0)$.
- If $R^2 = 0$ $(\Leftrightarrow \widehat{\beta}_D = 0)$, D is useless to predict Y: $\widehat{Y}_i = \overline{Y}$.
- Note: the R^2 is unaffected by any affine change on Y or D.
- In social sciences, it is common to have very low R^2, e.g. around 1%.
- It does not mean that the corresponding regressions would be "wrong"!
- A small R^2 just tells us that D is not very useful to predict Y.

Quality of the prediction



Outline

Motivation

- We now consider the case where $k > 2$: X includes more than a single random variable.
- Oftentimes, we can use several, not just one, variables to predict Y.
- Intuitively, we can improve our prediction of Y by adding explanatory variables.
- Also, adding nonlinear functions of D can be useful if the relationship between D and Y is nonlinear.
- As above, we always assume that X includes the intercept ("variable 1").

The OLS estimator

- We assume hereafter:
$$\frac{1}{n}\sum_{i=1}^{n} X_i X_i' \text{ is invertible.} \tag{Inv}$$
- Then we define the OLS estimator as:
$$\widehat{\beta} = \arg\min_{b\in\mathbb{R}^k} \sum_{i=1}^{n} \left(Y_i - X_i' b\right)^2.$$
- As we shall see, this estimator is well-defined under (Inv).
- The vector $\widehat{\beta}$ generalizes the OLS estimator $(\widehat{\alpha}, \widehat{\beta}_D)'$ defined previously.
- We are still looking for the best prediction of Y_i based on a linear combination of the vector X_i.
- As above, we define $\widehat{Y}_i = X_i'\widehat{\beta}$ and $\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i$.

- If the components of $(X^1, ..., X^k)'$ are not functionally dependent, for every $i = 1, ..., n$ we have $\widehat{\beta}_j = \partial \widehat{Y}_i / \partial X_i^j$.

$\Rightarrow$ Marginal effect of X^j on the prediction of $\widehat{Y}_i$.

- We often refer to the "marginal effect" of X^j.
- This "effect" is not causal in general(!) but it is the effect of X^j on the prediction of Y.
- If $(X^1, ..., X^k)'$ are not functionally dependent, we also have

$$\widehat{\beta}_j = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \widehat{Y}_i}{\partial X_i^j}$$

$\Rightarrow$ <u>Average</u> marginal effect of X^j on the prediction of $\widehat{Y}_i$.

- The components of $(X^1, ..., X^k)$ can be functionally dependent.
- For instance, we can have $X = (1, D, D^2)'$. Then, : $\partial \widehat{Y}_i / \partial D_i = \widehat{\beta}_1 + 2\widehat{\beta}_2 D_i$.

$\Rightarrow$ The marginal effect changes (and it can also change sign) with i.

- In this case, the average marginal effect $\widehat{\Delta}_j$ is :

$$\begin{aligned}\widehat{\Delta}_j &= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \widehat{Y}_i}{\partial X_i^j} \\ &= \widehat{\beta}_1 + 2\widehat{\beta}_2 \times \overline{D}.\end{aligned}$$

Proposition 2

Assume that (Inv) *holds. Then:*

1. $\widehat{\beta}$ *is well-defined and satisfies* $\widehat{\beta} = \left(\frac{1}{n}\sum_{i=1}^n X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n X_i Y_i\right)$;
2. *We have* $Y_i = X_i'\widehat{\beta} + \widehat{\varepsilon}$, *with* $\overline{X\widehat{\varepsilon}} = 0$.

Proof: let $f(b) = \sum_{i=1}^n \left(Y_i - X_i' b\right)^2$. Its Hessian is then $2\sum_{i=1}^n X_i X_i'$, which is symmetric positive by (Inv).

Then f is strictly convex and has at most one minimum, which solves the FOC:

$$\sum_{i=1}^n X_i(Y_i - X_i' b) = 0. \tag{4}$$

Since $\sum_{i=1}^n X_i X_i'$ is invertible, Point 1 follows. Then $Y_i = X_i'\widehat{\beta} + \widehat{\varepsilon}_i$ holds by definition of $\widehat{\varepsilon}$. The last point follows by (4), replacing b by $\widehat{\beta}$ therein $\square$

The full-rank condition (Inv)

- One can show that $\text{rank}(\frac{1}{n}\sum_{i=1}^{n} X_i X_i') \leq \min(n, k)$. Thus (Inv) implies that $n \geq k$: more observations than regressors.
- (Inv) is equivalent to having, for all $\lambda \in \mathbb{R}^k$,

$$X_i'\lambda = 0 \ \forall i \in \{1, ..., n\} \Rightarrow \lambda = 0.$$

- When $X = (1, D)'$ with $D \in \mathbb{R}$, (Inv) $\Leftrightarrow$ $(D_1, ..., D_n)$ not all identical.
- Counterexample of (Inv): D binary and $X = (1, D, 1 - D)'$.
- When $X = (1, D, G)$ with $D \in \mathbb{R}$, $G \in \mathbb{R}$ and $\min(\widehat{V}(D), \widehat{V}(G)) > 0$,

$$\text{(Inv)} \Longleftrightarrow |\widehat{\text{Corr}}(D, G)| < 1.$$

- Similar results more generally: (Inv) holds if we cannot recreate any regressor by a linear combination of the other regressors.

$\Rightarrow$ (Inv) allows for any level of correlation between covariates, except perfect collinearity.

Properties of the OLS estimator

- Point 1 of Proposition 2 generalizes Proposition 1 above: when $X = (1, D)'$, $D \in \mathbb{R}$, we get
$$\widehat{\beta} = \left(\overline{Y} - \frac{\widehat{\text{Cov}}(D, Y)}{\widehat{V}(D)} \overline{D}, \ \frac{\widehat{\text{Cov}}(D, Y)}{\widehat{V}(D)} \right)'.$$
- Point 2 of Proposition 2 implies that $\overline{\widehat{\varepsilon}} = 0$ and thus $(\overline{X}, \overline{Y})$ belongs to the regression "line" (=hyperplane) $\{(x, x'\widehat{\beta}) : x \in \mathbb{R}^k\}$.
- Same invariance properties as with simple, linear regressions.
- We also have $\widehat{\text{Cov}}(\widehat{Y}, \widehat{\varepsilon}) = 0$ and then $\widehat{V}(Y) = \widehat{V}(\widehat{Y}) + \widehat{V}(\widehat{\varepsilon})$. We still define the R^2 by:
$$R^2 = \frac{\widehat{V}(\widehat{Y})}{\widehat{V}(Y)} = \widehat{\text{Corr}}(Y, \widehat{Y})^2 \in [0, 1].$$
- Important: if we add a new explanatory variable, the R^2 necessarily increases.

Outline

- Let $X = (1, D, G')'$, $D \in \mathbb{R}$ and $\widehat{\beta} = (\widehat{\alpha}, \widehat{\beta}_D, \widehat{\beta}'_G)'$. Let also $\widehat{\eta}$ denote the residual of the regression of D on G.

Proposition 3

(Frisch-Waugh Theorem) If (Inv) *holds, then* $\widehat{\beta}_D$ *is the slope coefficient of* $\widehat{\eta}$ *in the linear regression of* Y *on* $\widehat{\eta}$.

Proof: let $\widehat{D}$ denote the predicted D from the regression of D on $(1, G)$. Then $D = \widehat{D} + \widehat{\eta}$, with $\overline{\widehat{\eta}} = \overline{\widehat{D}\widehat{\eta}} = \overline{G\widehat{\eta}} = 0$. The FOC of the reg. of Y on X are:

$$\sum_{i=1}^{n} X_i \left(Y_i - \widehat{\alpha} - (\widehat{D}_i + \widehat{\eta}_i)\widehat{\beta}_D - G'_i \widehat{\beta}_G \right) = 0$$

The same holds replacing X_i by any linear combination of X_i. In particular:

$$\sum_{i=1}^{n} \widehat{\eta}_i (Y_i - \widehat{\eta}_i \widehat{\beta}_D - \widehat{\alpha} - \widehat{D}_i \widehat{\beta}_D - G'_i \widehat{\beta}_G) = 0.$$

The above equality implies $\sum_{i=1}^{n} \widehat{\eta}_i (Y_i - \widehat{\eta}_i \widehat{\beta}_D) = 0$ and thus:

$$\widehat{\beta}_D = \frac{\sum_{i=1}^{n} \widehat{\eta}_i Y_i}{\sum_{i=1}^{n} \widehat{\eta}_i^2} = \frac{\sum_{i=1}^{n} (\widehat{\eta}_i - \overline{\widehat{\eta}})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (\widehat{\eta}_i - \overline{\widehat{\eta}})^2} \quad \square$$

- The second property below is useful to understand the so-called "omitted variable bias" considered in Chapter 4.
- Let again $X = (1, D, G')'$ with $D \in \mathbb{R}$, $G = (G^1, ..., G^p)'$ and let:
  - $\widehat{\beta}_D^S$ = coeff. of D in the simple linear reg. of Y on D;
  - $\widehat{\lambda} = (\widehat{\lambda}_1, ..., \widehat{\lambda}_p)'$, with $\widehat{\lambda}_j$ = coefficient of D in the simple linear reg. of G^j on D.

Proposition 4

If (Inv) *holds, we have* $\widehat{\beta}_D^S = \widehat{\beta}_D + \widehat{\lambda}'\widehat{\beta}_G$.

Proof: we have $\widehat{\beta}_D^S = \widehat{\text{Cov}}(Y, D)/\widehat{V}(D)$ and $Y = \widehat{\alpha} + D\widehat{\beta}_D + G'\widehat{\beta}_G + \widehat{\varepsilon}$, with $\widehat{\text{Cov}}(X, \widehat{\varepsilon}) = 0$. Thus,

$$\widehat{\beta}_D^S = \widehat{\beta}_D + \frac{\widehat{\text{Cov}}(D, G)'}{\widehat{V}(D)}\widehat{\beta}_G.$$

Now, $\widehat{\text{Cov}}(D, G)'/\widehat{V}(D)$ is a vector with jth term equal to $\widehat{\text{Cov}}(D, G_j)/\widehat{V}(D)$ which is the coefficient of D the reg. of G_j on D $\square$

Outline

Predicting wages by education and experience

- How well can we predict wages by education? By education and experience?
- In Wooldridge's dataset wage1.dta on the 1976 U.S. labour force, we observe the following variables:
  - `wage`: hourly wage (in 1976 dollars);
  - `educ`: years of education (starting at 6 years of age);
  - `exper`: years of potential experience: age - (age when education completed) .
- We consider several regressions corresponding to the following Stata code:

```
reg wage educ
gen educ10=max(0,educ-10)
reg wage educ educ10
reg wage educ exper
```

Regression on education only

- Stata output of `reg wage educ`:

Source	SS	df	MS
Model	1179.73204	1	1179.73204
esidual	5980.68225	524	11.4135158
Total	7160.41429	525	13.6388844

F(1, 524)	=	103.36
Prob > F	=	0.0000
R-squared	=	0.1648
Adj R-squared	=	0.1632
Root MSE	=	3.3784

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.5413593	.053248	10.17	0.000	.4367534	.6459651
_cons	-.9048516	.6849678	-1.32	0.187	-2.250472	.4407687

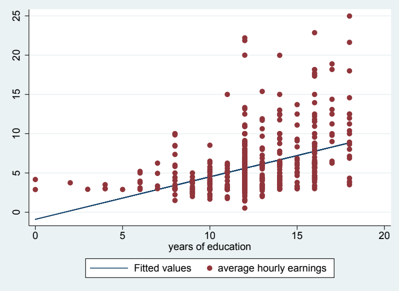
- Which salary can we predict when educ$=10$? When educ$=0$?
- Results of the 2nd regression (`reg wage educ educ10`):

$$\widehat{\texttt{wage}} = 4.37 - 0.068\texttt{educ} + 0.83\texttt{educ10}, \quad R^2 \simeq 0.198$$

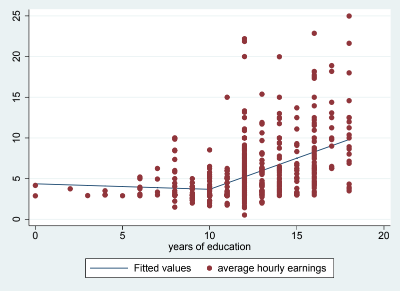
- What is now the predicted value at educ$=0$? At educ$=10$?

Comparison of predictions

Figure 1: Scatter plot with predicted values



`reg wage educ`



`reg wage educ educ10`

Including experience

- Result of the third regression (`reg wage educ exper`):

$$\widehat{\texttt{wage}} = -3.39 + 0.64\texttt{educ} + 0.07\texttt{exper}, \quad R^2 \simeq 0.23.$$

- Reminder on the initial regression without experience:

$$\widehat{\texttt{wage}} = -0.90 + 0.54\texttt{educ}$$

- Why did the coefficient of `educ` increase when including experience?

Outline

Asymptotic behavior of the OLS estimator

Proposition 5

If $E(|Y|^2) < \infty$, $E(||X||^2) < \infty$ and $E[XX']$ is invertible, then:

1. $\widehat{\beta}$ *is well-defined with probability tending to one (wpto) and*

$$\widehat{\beta} \xrightarrow{P} \beta_0 := E[XX']^{-1}E[XY].$$

2. $\beta_0 = \arg\min_b E[(Y - X'b)^2] = \arg\min_b E[(E(Y|X) - X'b)^2]$.

3. *There exists* ε *such that* $Y = X'\beta_0 + \varepsilon$, *with* $E[X\varepsilon] = 0$. *Moreover,* $\widehat{\varepsilon}_i \xrightarrow{P} \varepsilon_i$ *for all* i.

- The OLS estimator converges under weak conditions to some $\beta_0 \in \mathbb{R}^k$.
- ε is called the residual of the theoretical regression of Y on X.

1. By the strong law of large numbers (LLN),

$$\frac{1}{n}\sum_{i=1}^{n} X_i X_i' \xrightarrow{P} E[XX'].$$

Thus, wpto, $\sum_{i=1}^{n} X_i X_i'/n$ is invertible and then $\widehat{\beta}$ is well-defined by Proposition 2.

Moreover, $E(||XY||) \leq [E(||X||^2)E(Y^2)]^{1/2} < \infty$, so that

$$\frac{1}{n}\sum_{i=1}^{n} X_i Y_i \xrightarrow{P} E(XY).$$

Then, by the continuous mapping theorem, since $E[XX']$ is invertible,

$$\widehat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right) \xrightarrow{P} E(XX')^{-1}E(XY).$$

2. By definition of β_0, we have $E[X(Y - X'\beta_0)] = 0$. These are the FOC of the strictly convex program $\min_b E[(Y - X'b)^2]$.

For the 2nd eq., remark that for all f, $E[(Y - E(Y|X))f(X)] = 0$. Then:

$$\begin{aligned} E[(Y - X'b)^2] =& E[(Y - E(Y|X) + E(Y|X) - X'b)^2] \\ =& E[(Y - E(Y|X))^2 + 2E[(Y - E(Y|X))(E(Y|X) - X'b)] \\ &+ E[(E(Y|X) - X'b)^2] \\ =& E[(Y - E(Y|X))^2 + E[(E(Y|X) - X'b)^2]. \end{aligned}$$

Thus, $\beta_0 = \arg\min_b E[(Y - X'b)^2] = \arg\min_b E[(E(Y|X) - X'b)^2]$.

3. Let $\varepsilon = Y - X'\beta_0$. Then $Y = X'\beta_0 + \varepsilon$ and

$$E[X\varepsilon] = E[X(Y - X'\beta_0)] = 0.$$

Finally, we have $\widehat{\varepsilon}_i - \varepsilon_i = -X_i'(\widehat{\beta} - \beta_0) \xrightarrow{P} 0$ for all i $\square$

- β_0 = coefficient of the theoretical regression ($\min_b E[(Y - X'b)^2]$) instead of the data sample regression ($\min_b \sum_{i=1}^n (Y_i - X_i' b)^2$).
- When $X = (1, D)'$, we obtain

$$\widehat{\alpha} \xrightarrow{P} \alpha_0 = E(Y) - \frac{\text{Cov}(D, Y)}{V(D)} E(D),$$
$$\widehat{\beta}_D \xrightarrow{P} \beta_D = \frac{\text{Cov}(D, Y)}{V(D)}.$$

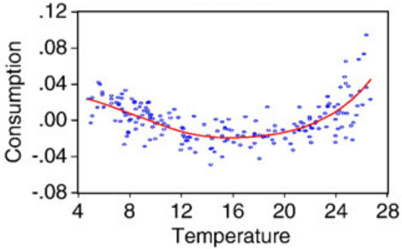
  In particular, when $D \in \{0, 1\}$, $\beta_D = E[Y|D = 1] - E[Y|D = 0]$.
- $\beta_0 = \arg\min_b E[(Y - X'b)^2]$ indicates that $X'\beta_0$ is the best prediction, in the L^2 sense, of Y by linear functions of X.
- $\beta_0 = \arg\min_b E[(E(Y|X) - X'b)^2]$ means that the linear regression $X'\beta_0$ is the best linear approximation of conditional expectation...
- ... But sometimes this approximation is bad!

Linear predictions can be useless

- Assume $X = (1, D)' \in \mathbb{R}^2$ and
$$E(Y|D) = E(Y) + b(D - E(D))^2,$$
where the distribution of D is symmetric around its mean.

- Then $\text{Cov}(D, Y) = b\text{Cov}(D, (D - E(D))^2) = 0$. Thus the linear prediction is just $X'\beta_0 = E(Y)$.

- This may happen in practice: below relationship between temperature (D) and electricity consumption (Y) for Greece:



Source: Bessec and Fouquau, Energy Economics (2008)

Theoretical marginal effects

- If the X^j $(j = 1, ..., k)$ are not functionally dependent, β_j equals :

1) the "marginal effect" of X^j on the theoretical prediction of Y ;
2) also the "average marginal effect" of X^j on the theoretical prediction of Y;

- If the X^j $(j = 1, ..., k)$ are functionally dependent,

  $\beta_j \neq$ the marginal effect and the average marginal effect of X^j in general.

Link between simple and multiple theoretical regressions

- We have the same results on β_0 as Propositions 3-4 on $\widehat{\beta}$.
- Hereafter, $X = (1, D, G')'$, $D \in \mathbb{R}$, $\beta_0 = (\alpha_0, \beta_D, \beta_G')'$ and:
  - η=residual of the theoretical regression of D on G
  - β_D^S= coeff. of D in the theoretical reg. of Y on D;
  - $\lambda = (\lambda_1, ..., \lambda_p)'$, with λ_j= coeff. of D in the theoretical reg. of G^j on D.

Proposition 6

(Frisch-Waugh, v2) If $E(|Y|^2) < \infty$, $E(||X||^2) < \infty$ and $E(XX')$ is invertible, β_D is the coeff. of η the theoretical reg. of Y on η.

Proposition 7

If $E(|Y|^2) < \infty$, $E(||X||^2) < \infty$ and $E(XX')$ is invertible, $\beta_D^S = \beta_D + \lambda'\beta_G$.

Outline

Summary of Chapter 1

- Definition of the OLS estimator in simple and multiple linear regressions.
- Algebraic properties of the OLS.
- Quality of the prediction: R^2.
- Link between “short” and “long” regressions.
- Theoretical regressions, interpretation of the probability limit of the OLS estimator.