Regression Model Basics

Econometrics

Péter Elek

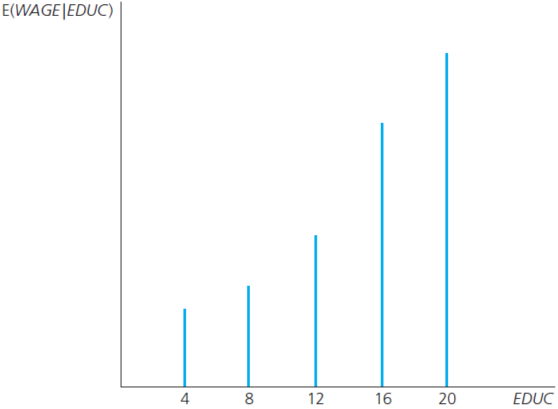
Table of Contents

Conditional Expectation

Conditional Expectation Function (CEF)

- Conditional expectation: $E(y|x_1, x_2, \ldots, x_k) = E(y|\mathbf{x})$
- It shows the expected value of y if the explanatory variables take on given values.
- Example:
  - y : wage
  - x_1 : education
  - x_2 : tenure (seniority)

Example: Education and Wage



Source: Wooldridge Fig. B.5.

Example: Nonlinear CEF

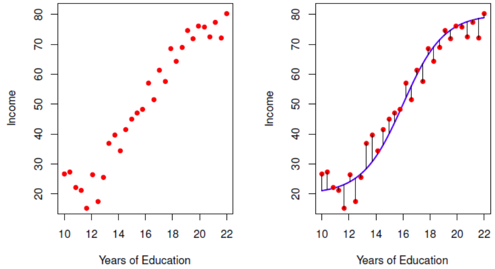


FIGURE 2.2. *The* Income *data set.* Left: *The red dots are the observed values of* income *(in tens of thousands of dollars) and* years of education *for* 30 *individuals.* Right: *The blue curve represents the true underlying relationship between* income *and* years of education*, which is generally unknown (but is known in this case because the data were simulated). The black lines represent the error associated with each observation. Note that some errors are positive (if an observation lies above the blue curve) and some are negative (if an observation lies below the curve). Overall, these errors have approximately mean zero.*

Source: ISLR Fig. 2.2.

Example: CEF with Two Explanatory Variables

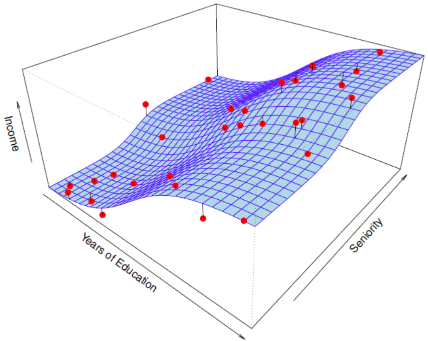


FIGURE 2.3. *The plot displays* `income` *as a function of* `years of education` *and* `seniority` *in the* `Income` *data set. The blue surface represents the true underlying relationship between* `income` *and* `years of education` *and* `seniority`*, which is known since the data are simulated. The red dots indicate the observed values of these quantities for 30 individuals.*

Source: ISLR Fig. 2.3.

A Few Basic Properties of CEF

- $E(g(x) \mid x) = g(x)$ for any $g(x)$ function of the r.v. x.
- $E(g(x)y \mid x) = g(x)E(y \mid x)$ for any r.v. x, y and a $g(x)$ function.
- If $\mathbf{x}$ and y are independent then $E(y \mid \mathbf{x}) = E(y)$.
- If $E(y \mid \mathbf{x}) = E(y)$ then $cov(\mathbf{x}, y) = 0$. Also, $cov(g(\mathbf{x}), y) = 0$ for every $g(\mathbf{x})$ function.
- Law of iterated expectations: $E(E(y \mid \mathbf{x})) = E(y)$.
- CEF decomposition: $y = E(y \mid \mathbf{x}) + \varepsilon$, where ε is a r.v. such that $E(\varepsilon \mid \mathbf{x}) = 0$.

CEF Prediction Property

- CEF is the best (i.e. minimum mean squared error) approximation of y as a function of $\mathbf{x}$.
- Formally: for all $g(\mathbf{x})$ functions of the explanatory variables,

$$E\left((y - E(y \mid \mathbf{x}))^2 \mid \mathbf{x}\right) \leq E\left((y - g(\mathbf{x}))^2 \mid \mathbf{x}\right)$$

$$E\left((y - E(y \mid \mathbf{x}))^2\right) \leq E\left((y - g(\mathbf{x}))^2\right)$$

- In the unconditional case: $E(y)$ minimizes the mean squared error, i.e. for all m, $E((y - E(y))^2) \leq E((y - m)^2)$.
- Proof: during class or e.g. in Wooldridge Appendix B.4.
- Hence CEF is important for prediction purposes.

Linear Regression

Linear Regression Model

- Linear regression model:

$$E(y \mid \mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

- Equivalently:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

- where u is the error term (capturing all unobserved influences)
- and the zero conditional mean assumption holds:

$$E(u \mid \mathbf{x}) = 0.$$

- This is a population model: the parameters $(\beta_0, \beta_1, \ldots, \beta_k)$ are defined in the full population; data allow us to estimate them.

Linearity in Parameters

- The model is linear in parameters. But the explanatory variables may contain nonlinear functions e.g. polynomials $(x^2, x^3, \dots)$, logarithmic function $(\log(x))$ or interactions $(x_1 x_2)$. See later at functional forms.

- Moreover, even if the true $E(y \mid \mathbf{x})$ is nonlinear, the function $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ provides the minimum mean squared error linear approximation to the CEF.

- Hence assuming CEF linearity is not as restrictive as it first looks.

Parameter Interpretation

- β_j is the expected difference of y in case of a one unit difference of x_j if all other x_m-s are unchanged (i.e. all other x_m-s are controlled for):

$$\beta_j = \frac{dE(y \mid \mathbf{x})}{dx_j}.$$

- Predictive interpretation: β_j compares the expected y values of different subjects with different x_j values but all other x_m-s being the same (i.e. after controlling for all other x_m-s).
- Not (necessarily) causal interpretation: no information on the change of y if x_j changes within a subject

Examples

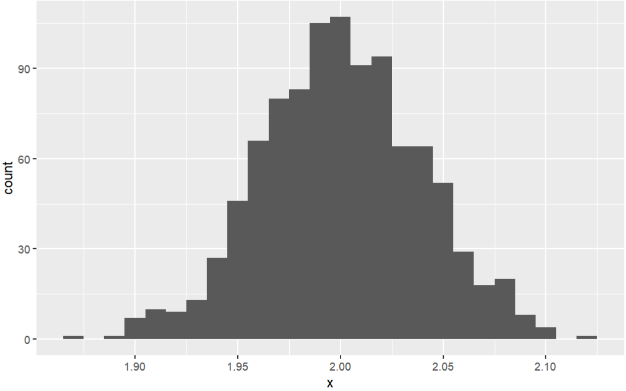
- Example 1: y wage, x_1 years education, x_2 tenure etc.
- Predictive interpretation: β_1 shows the expected difference of wage of individuals with one year difference in education and the same control variables (e.g. tenure).
- Does not necessarily show the expected wage difference that would occur as a result of an increase of an individual's level of education (no causal interpretation).

- Example 2: does smoking during pregnancy decrease birth weight? No straightforward answer based on observational data (without experiments).

OLS Expected Value

Simulation of the Sampling Distribution of the OLS Estimator

Model: $y = 3 + 2x + u$ where $x \sim N(0, 9)$ and $u \mid x \sim N(0, 36)$



Source: Cunningham Fig. 2.3

Unbiasedness of OLS

- Suppose that we have
  1. a linear regression model $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$,
  2. random (i.i.d.) sampling,
  3. $E(u \mid \mathbf{x}) = 0$ and
  4. there is no perfect collinearity between the variables in the sample.
- Then the OLS estimator is unbiased, i.e.

$$E(\hat{\beta}_j) = \beta_j \quad (j = 0, 1, \ldots, k).$$

Proof in the simple regression case ($k = 1$)

- (Throughout we use that $\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n}(y_i - \bar{y}) = 0$.)
- The OLS estimator for β_1 is:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

- Substituting $y_i = \beta_0 + \beta_1 x_i + u_i$:

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Proof (cont.)

- (We denote the sample measurements of x-s by X.)
- Random sampling and the zero conditional mean assumption imply:

$$E\left((x_i - \bar{x})\, u_i \mid X\right) = (x_i - \bar{x})\, E\left(u_i \mid x_i\right) = 0.$$

- Hence

$$E\left(\hat{\beta}_1 \mid X\right) = \beta_1 + \frac{\sum_{i=1}^{n} E\left((x_i - \bar{x})\, u_i \mid X\right)}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \beta_1.$$

- Then the law of iterated expectations concludes the proof.

Proof in the multiple regression case $(k > 1)$

- Proof in the multiple regression case follows from the OLS anatomy formula, using $\sum_{i=1}^{n} x_{ji}\hat{r}_{1i} = 0$ $(j = 2, \ldots, k)$ and then the same tricks as in the $k = 1$ case:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \hat{r}_{1i} y_i}{\sum_{i=1}^{n} \hat{r}_{1i}^2} = \beta_1 + \frac{\sum_{i=1}^{n} \hat{r}_{1i} u_i}{\sum_{i=1}^{n} \hat{r}_{1i}^2}.$$

- Details e.g. in Wooldridge Appendix 3A.3.

Omitted Variable Bias

- True model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, \quad E(u \mid x_1, x_2) = 0.$
- Estimated model: $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$
- OLS omitted variable formula: $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \hat{\delta}_1.$
- (We denote the sample measurements of the explanatory variables as $\mathbf{X}$.)
- OLS is unbiased in the true model: $E(\hat{\beta}_1 \mid \mathbf{X}) = \beta_1$ and $E(\hat{\beta}_2 \mid \mathbf{X}) = \beta_2.$
- Hence the conditional expectation of $\tilde{\beta}_1$:

$$E(\tilde{\beta}_1 \mid \mathbf{X}) = E(\hat{\beta}_1 + \hat{\beta}_2 \hat{\delta}_1 \mid \mathbf{X}) = \beta_1 + \beta_2 \hat{\delta}_1.$$

Omitted Variable Bias (cont.)

- Suppose that β_1 is of interest (association of x_1 and y after controlling for x_2).
- Then:
- $\tilde{\beta}_1$ is unbiased if only if:
  - $\beta_2 = 0$ (the omitted variable has no effect on y) OR
  - $\hat{\delta}_1 = 0$ (x_{1i} and x_{2i} are uncorrelated in the sample).
- $\tilde{\beta}_1$ is upward biased if
  - $\beta_2 > 0$ and $\hat{\delta}_1 > 0$ OR
  - $\beta_2 < 0$ and $\hat{\delta}_1 < 0$.
- $\tilde{\beta}_1$ is downward biased if
  - $\beta_2 > 0$ and $\hat{\delta}_1 < 0$ OR
  - $\beta_2 < 0$ and $\hat{\delta}_1 > 0$.

Exogeneity and Endogeneity

- Regression model: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$
- The explanatory variables are called exogenous if $E(u \mid \mathbf{X}) = 0$ in the regression.
  - Then $cov(x_j, u) = 0$ for $j = 1, \ldots, k$.
- An explanatory variable x_j is called endogenous if $cov(x_j, u) \neq 0$.

- Omitted variable is a frequent source of endogeneity.
  - Suppose that β_1 is of interest, but we estimate $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$.
  - Writing this regression with an error term: $y = \beta_0 + \beta_1 x_1 + v$ where $cov(x_1, v) = cov(x_1, \beta_2 x_2 + u) = \beta_2 cov(x_1, x_2)$ may be different from zero.
  - Not surprising: OLS excluding x_2 estimates $y = \gamma_0 + \gamma_1 x_1 + w$ where $E(w \mid x_1) = 0$. But in general $\gamma_1 \neq \beta_1$ (γ_1 does not control for, β_1 controls for x_2).

Example: Education, Ability and Wage

- y : wage, x_1 : education, x_2 : ability (and there may be further control variables)
- β_1 is of primary interest (association of education and wage after controlling for ability)
- If ability is unobserved, only the smaller model (without ability) can be estimated.
- Possible sign of bias of OLS in the smaller model?
- We can also say that the smaller model estimates a different quantity (where ability is not controlled for). It may or may not be of interest.

Example: Black-White Wage Differences

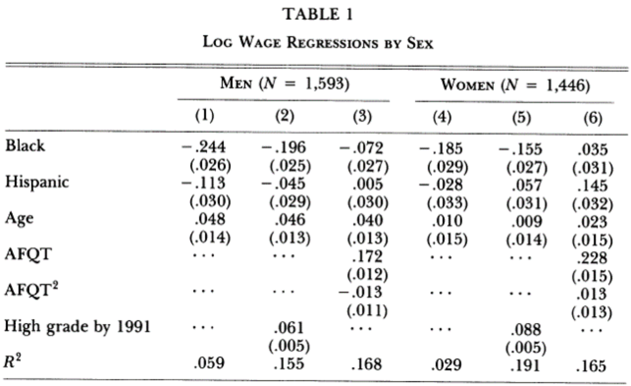


TABLE 1

Log Wage Regressions by Sex

	Men (N = 1,593)			Women (N = 1,446)		
	(1)	(2)	(3)	(4)	(5)	(6)
Black	−.244	−.196	−.072	−.185	−.155	.035
	(.026)	(.025)	(.027)	(.029)	(.027)	(.031)
Hispanic	−.113	−.045	.005	−.028	.057	.145
	(.030)	(.029)	(.030)	(.033)	(.031)	(.032)
Age	.048	.046	.040	.010	.009	.023
	(.014)	(.013)	(.013)	(.015)	(.014)	(.015)
AFQT	. . .	. . .	.172	. . .	. . .	.228
			(.012)			(.015)
AFQT^2	. . .	. . .	−.013	. . .	. . .	.013
			(.011)			(.013)
High grade by 1991	. . .	.061	. . .	. . .	.088	. . .
		(.005)			(.005)	
R^2	.059	.155	.168	.029	.191	.165

Source: Neal and Johnson (1996): The Role of Premarket Factors in Black-White Wage Differences. Journal of Political Economy 104, 869-895.
AFQT: Armed Forces Qualification Test (taken at teenage years)

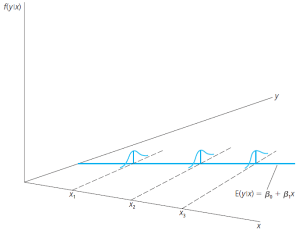
OLS Variance

Homoskedasticity

Homoskedasticity: conditional variance of the error term is constant

$$Var(u|\mathbf{x}) = \sigma_u^2$$

FIGURE 2.8 The simple regression model under homoskedasticity.



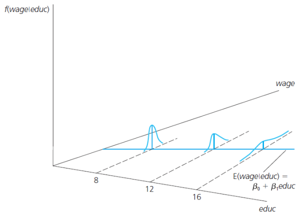
Source: Wooldridge Fig. 2.8.

Heteroskedasticity

Heteroskedasticity: conditional variance of the error term is not constant

$$Var(u|\mathbf{x}) \quad \text{not constant}$$

FIGURE 2.9 Var(*wage*|*educ*) increasing with *educ*.



Source: Wooldridge Fig. 2.9.

Variance of OLS Estimator Under Homoskedasticity

- If conditions (1)-(4) (needed for unbiasedness) and (5) homoskedasticity hold then

$$Var(\beta_j \mid \mathbf{X}) = \frac{\sigma_u^2}{SSR_j} = \frac{\sigma_u^2}{\sum_{i=1}^{n}(x_{ji} - \bar{x}_j)^2 * (1 - R_j^2)} = \frac{\sigma_u^2}{n * \hat{Var}(x_j) * (1 - R_j^2)}$$

- where SSR_j and R_j^2 are, respectively, the residual sum of squares and R-squared of the regression of x_j on the other explanatory variables.

- Specifically, in the simple regression case ($k = 1$), there is no other explanatory variable so $R_1^2 = 0$ and hence

$$Var(\beta_1 \mid X) = \frac{\sigma_u^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma_u^2}{n * \hat{Var}(x_j)}.$$

Proof

- By the OLS anatomy formula and i.i.d. sampling:

$$Var\left(\hat{\beta}_j \mid \mathbf{X}\right) = Var\left(\frac{\sum_{i=1}^n \hat{r}_{ji} u_i}{\sum_{i=1}^n \hat{r}_{ji}^2} \mid \mathbf{X}\right) = \frac{\sum_{i=1}^n \hat{r}_{ji}^2 Var\left(u_i \mid \mathbf{X}\right)}{\left(\sum_{i=1}^n \hat{r}_{ji}^2\right)^2} = \frac{\sigma_u^2}{\sum_{i=1}^n \hat{r}_{ji}^2} = \frac{\sigma_u^2}{SSR_j}.$$

OLS Variance (cont.)

- Determinants of OLS Variance
  - σ_u^2 (variance of the error term)
  - n (sample size)
  - $\hat{Var}(x_j)$ (variance of x_j)
  - R_j^2 (strength of linear relationship between the explanatory variables)

- Multicollinearity: R_j^2 is "close to one"
- Perfect collinearity: $R_j^2 = 1$ (OLS is not defined)

Standard Error of $\hat{\beta}_j$

- An unbiased (and consistent) estimator of σ_u^2 is

$$\hat{\sigma}_u^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n - k - 1}$$

- where division by $n - k - 1$ occurs because, in the OLS estimation, $k + 1$ parameters are estimated ($k + 1$ degrees of freedom are "lost").
  - Doesn't matter much if $n >> k$

- Hence the standard error (s.e.) of $\hat{\beta}_j$:

$$s.e.(\beta_j \mid \mathbf{X}) = \frac{\hat{\sigma}_u}{\sqrt{\sum_{i=1}^{n} (x_{ji} - \bar{x}_j)^2 * (1 - R_j^2)}}.$$

Bias-Variance Tradeoff

- Adding more explanatory variables to a regression
  - may reduce omitted variable bias
  - but may increase the variance of parameter estimates
  - (R_j^2 may go up, although σ_u^2 may go down)

Heteroskedasticity-Robust Standard Error of OLS Estimator

- Under assumptions (1)-(4) (but not assuming homoskedasticity), in large samples:

$$Var\left(\hat{\beta}_j \mid \mathbf{X}\right) = \frac{\sum_{i=1}^{n} \hat{r}_{ji}^2 Var\left(u_i \mid \mathbf{X}\right)}{\left(\sum_{i=1}^{n} \hat{r}_{ji}^2\right)^2} \approx \frac{\sum_{i=1}^{n} \hat{r}_{ji}^2 \hat{u}_i^2}{SSR_j^2}.$$

- This is called heteroscedasticity-robust (or White- or Eicker-Huber-White-) variance estimation.
- It can always be used in large samples (irrespective to heteroskedasticity).
- Under heteroskedasticity, the conventional standard errors usually underestimate the true standard errors. However, the difference is only modest in most cases.
- Homoskedasticity can be tested (see later). If homoskedasticity holds, conventional standard errors are more precise, especially in small samples.

Material

- Cunningham 2.11-2.12, 2.16-2.19, 2.21-2.23, 2.25-2.26
- Wooldridge 2.1, 2.5, 3.1-3.4, 3.6, 8.2.