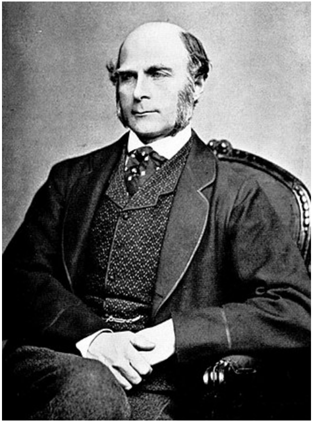
The Simple Regression Model: Definition, Estimation, and Statistical Properties

1. Definition
2. Deriving the Ordinary Least Squares (OLS) Estimates
3. Properties of OLS

Definition of the Simple Regression Model

Sir Francis Galton



https://en.wikipedia.org/wiki/Francis_Galton

Regression Towards Mediocrity in Hereditary Stature (1886)

Definition of the Simple Regression Model (cont.)

- Cross-sectional analysis
- Objective : estimate the effect of

X on Y

e.g. educ. wage

trade pollution

indep. var. dep. var.

explanatory var. explained var.

regressor regressand

Definition of the Simple Regression Model (cont.)

β_1 : how y changes when x changes
keeping u fixed

slope

- The simple linear regression model

$$y = \beta_0 + \beta_1 x + u$$

unobserved or error term

intercept

$$\Delta y = \beta_1 \Delta x + \Delta u$$

wage

educ.

ability

Definition of the Simple Regression Model (cont.)

$$y = (\beta_0 + 150) + \beta_1 x + (u - 150)$$

e.g. if $E(u) = 150$

- Objective: estimate β_0 and β_1
- Two assumptions

$$E(u) = 0$$

$$E(u|x) = E(u)$$

No!

implying $E(u|x) = 0$

$\text{corr.}(x, u) = 0$

$E(x.u) = 0$

Is this likely to be satisfied?

what if yes!

educ. or trade were randomly allocated?

Deriving the Ordinary Least Squares Estimates

$E(u) = 0$

$E(xu) = 0$

- Two equations

$$E(y - \beta_0 - \beta_1 x) = 0$$

$$E\, x(y - \beta_0 - \beta_1 x) = 0$$

- Cross-sectional data $\{(x_i, y_i) : i = 1, 2, ..., n\}$
- Sample analogs

$$\frac{1}{n} \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

- $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\frac{1}{n} \sum_i x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

estimates of β_0 and β_1

Deriving the Ordinary Least Squares Estimates (cont.)

Example

y (wage)	x (educ)
3.1	11
3.2	12
3	11
6	8
5.3	12
8.8	16
11	18
5	12
3.6	12
18	17

Deriving the Ordinary Least Squares Estimates (cont.)

$$\begin{aligned} \bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x} &= 0 \\ \overline{xy} - \hat{\beta}_0 \bar{x} - \hat{\beta}_1 \overline{x^2} &= 0 \end{aligned}$$

$$\begin{aligned} 6.742 - \hat{\beta}_0 - 12.9\hat{\beta}_1 &= 0 \\ 97.234 - 12.9\hat{\beta}_0 - 175.1\hat{\beta}_1 &= 0 \end{aligned}$$

$$\begin{aligned} \hat{\beta}_0 &= -8.492 \\ \hat{\beta}_1 &= 1.181 \end{aligned}$$

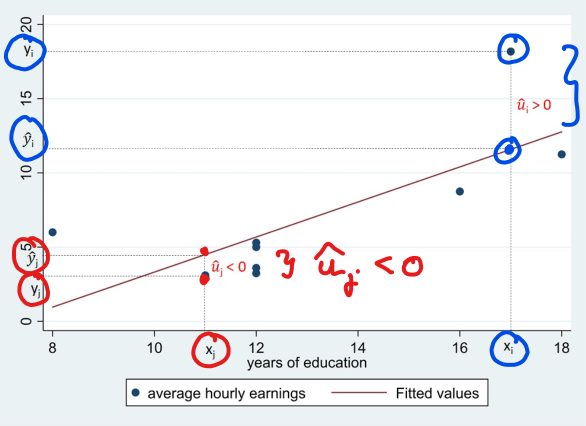
$\bar{x}$, $\bar{y}$, $\overline{xy}$, and $\overline{x^2}$: sample average of x, y, xy, and x^2

Deriving the Ordinary Least Squares Estimates (cont.)

https://www.stata.com/giftshop/beta-hat/

Deriving the Ordinary Least Squares Estimates (cont.)

- For the ith observation
  - Value of the dependent variable: y_i
  - Fitted value: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
  - Residual: $\hat{u}_i = y_i - \hat{y}_i$



$\hat{u}_i > 0$

Deriving the Ordinary Least Squares Estimates (cont.)

y (wage)	x (educ)	$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$	$\hat{u} = y - \hat{y}$
3.1	11	4.498	−1.398
3.2	12	5.679	−2.439
3	11	4.498	−1.498
6	8	0.955	5.045
5.3	12	5.679	−0.379
8.8	16	10.403	−1.653
11	18	12.765	−1.515
5	12	5.679	−0.679
3.6	12	5.679	−2.079
18	17	11.584	6.596

Deriving the Ordinary Least Squares Estimates (cont.)

$$= \sum_i (y_i - \hat{y}_i)^2$$

- Sum of squared residuals

$$\sum_{i=1}^{n} \hat{u}_i^2 = \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- $\hat{\beta}_0$ and $\hat{\beta}_1$ *minimize* the sum of squared residuals

Deriving the Ordinary Least Squares Estimates (cont.)

- The **ordinary least squares** (OLS) slope estimate

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

- The OLS intercept estimate

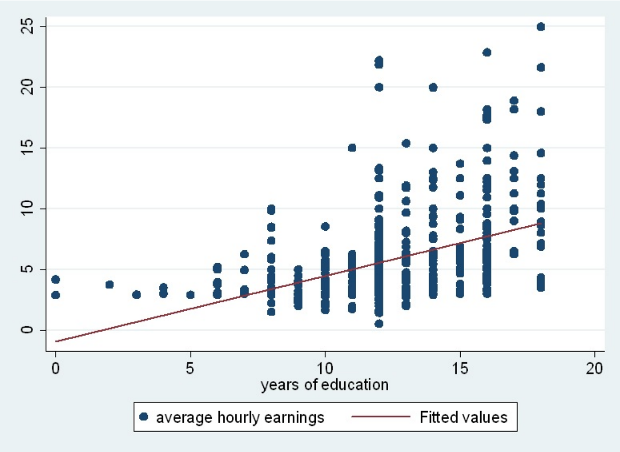
$$\hat{\beta}_0 =$$

$$\bar{y} - \hat{\beta}_1 \bar{x}$$

Deriving the Ordinary Least Squares Estimates (cont.)

Data: wage1

- *wage*: dollars per hour; *educ*: years of education; $n = 526$
- Estimated equation: $\widehat{wage} = -0.90 + 0.54\ educ$



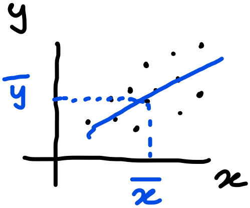
Properties of OLS

1. Sum (and thus average) of OLS residuals

$$\sum_{i=1}^{n} \hat{u}_i = 0$$

2. Sample correlation between x and $\hat{u}$

$$\sum_{i=1}^{n} x_i \hat{u}_i = 0$$



Correlation between the fitted values and residuals

$$\sum_{i=1}^{n} \hat{y}_i \hat{u}_i = 0$$

3. $(\bar{x}, \bar{y})$ is on the OLS regression line

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

Properties of OLS (cont.)

y (wage)	x (educ)	$\hat{y}$	$\hat{u}$	$x\hat{u}$	$\hat{y}\hat{u}$
3.1	11	4.498	-1.398	-15.381	-6.290
3.2	12	5.679	-2.439	-29.270	-13.852
3	11	4.498	-1.498	-16.481	-6.740
6	8	0.955	5.045	40.356	4.820
5.3	12	5.679	-0.379	-4.550	-2.153
8.8	16	10.403	-1.653	-26.446	-17.194
11	18	12.765	-1.515	-27.265	-19.335
5	12	5.679	-0.679	-8.150	-3.857
3.6	12	5.679	-2.079	-24.950	-11.808
18	17	11.584	6.596	112.136	76.409

Properties of OLS (cont.)

$$\hat{u}_i = y_i - \hat{y}_i$$

Goodness-of-Fit

- For each observation

$$y_i = \hat{y}_i + \hat{u}_i$$

- Total, explained, and residual - sum of squares

SST SSE SSR

$$SST = \sum_i (y_i - \bar{y})^2$$
$$SSE = \sum_i (\hat{y}_i - \bar{y})^2$$
$$SSR = \sum_i \hat{u}_i^2$$

Properties of OLS (cont.)

- Can show

$$SST = SSE + SSR$$

- Fraction of the variation in y explained by x

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

- Equals the *square* of the correlation between y and $\hat{y}$

$$0 \leq R^2 \leq 1$$

Properties of OLS (cont.)

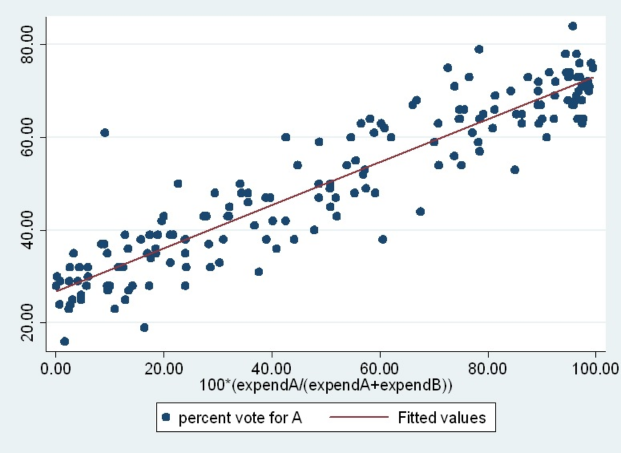
poor fit of OLS line
very little of variation in Y
captured by " " $\hat{Y}$

- $R^2 = 0$
- $R^2 = 1$
- High R^2 : not the ultimate objective in causal analysis

perfect fit of OLS line
all data pts. on same line

Properties of OLS (cont.)

Data: vote1



Properties of OLS (cont.)

y (wage)	x (educ)	$\hat{y}$	$\hat{u}$	$(y - \overline{y})^2$	$(\hat{y} - \overline{y})^2$	$\hat{u}^2$
3.1	11	4.498	−1.398	13.264	5.034	1.955
3.2	12	5.679	−2.439	12.264	1.130	5.950
3	11	4.498	−1.498	14.003	5.034	2.245
6	8	0.955	5.045	0.551	33.484	25.447
5.3	12	5.679	−0.379	2.079	1.130	0.144
8.8	16	10.403	−1.653	4.032	13.402	2.732
11	18	12.765	−1.515	20.322	36.273	2.294
5	12	5.679	−0.679	3.035	1.130	0.461
3.6	12	5.679	−2.079	9.872	1.130	4.323
18	17	11.584	6.596	130.828	23.443	43.510

$SST = 210.249 \qquad SSE = 121.188$

$SSR = 89.061 \qquad R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} = 0.576$