OLS Algebra

Econometrics

Péter Elek

Setup

- y : dependent variable
- $x_1, x_2, \ldots, x_k$: explanatory variables
- Given a sample $\{(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i) : i = 1, \ldots, n\}$.
- n: number of observations
- k : number of explanatory variables
- Example:
  - y : birth weight
  - x_1 : number of daily cigarette smoking during pregnancy
  - x_2 : family income

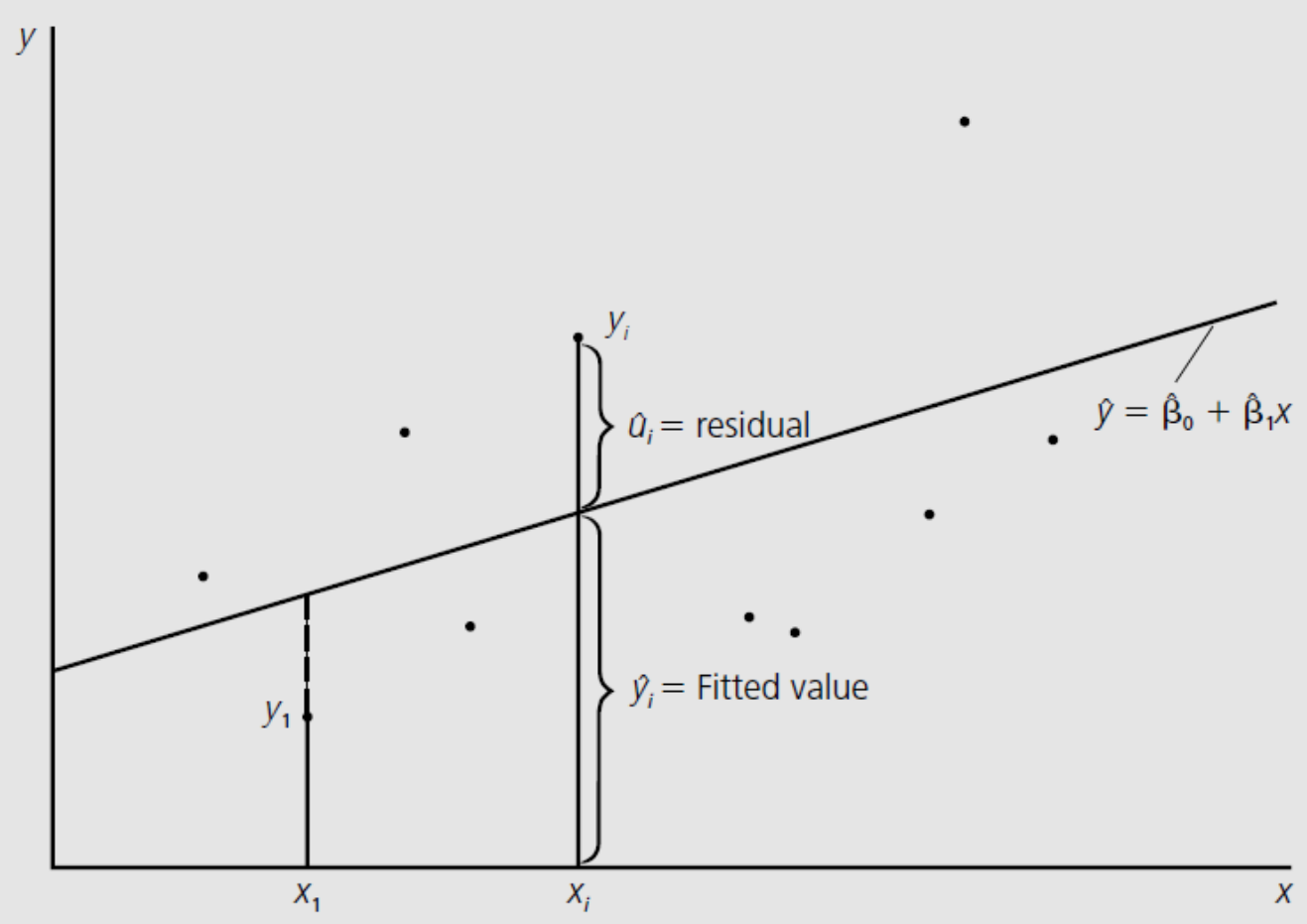
Ordinary Least Squares (OLS)

- Search for the best fitting linear function ($k = 1$: best fitting line), i.e. approximate y with the explanatory variables $x_1, x_2, \ldots, x_k$
- Minimize:

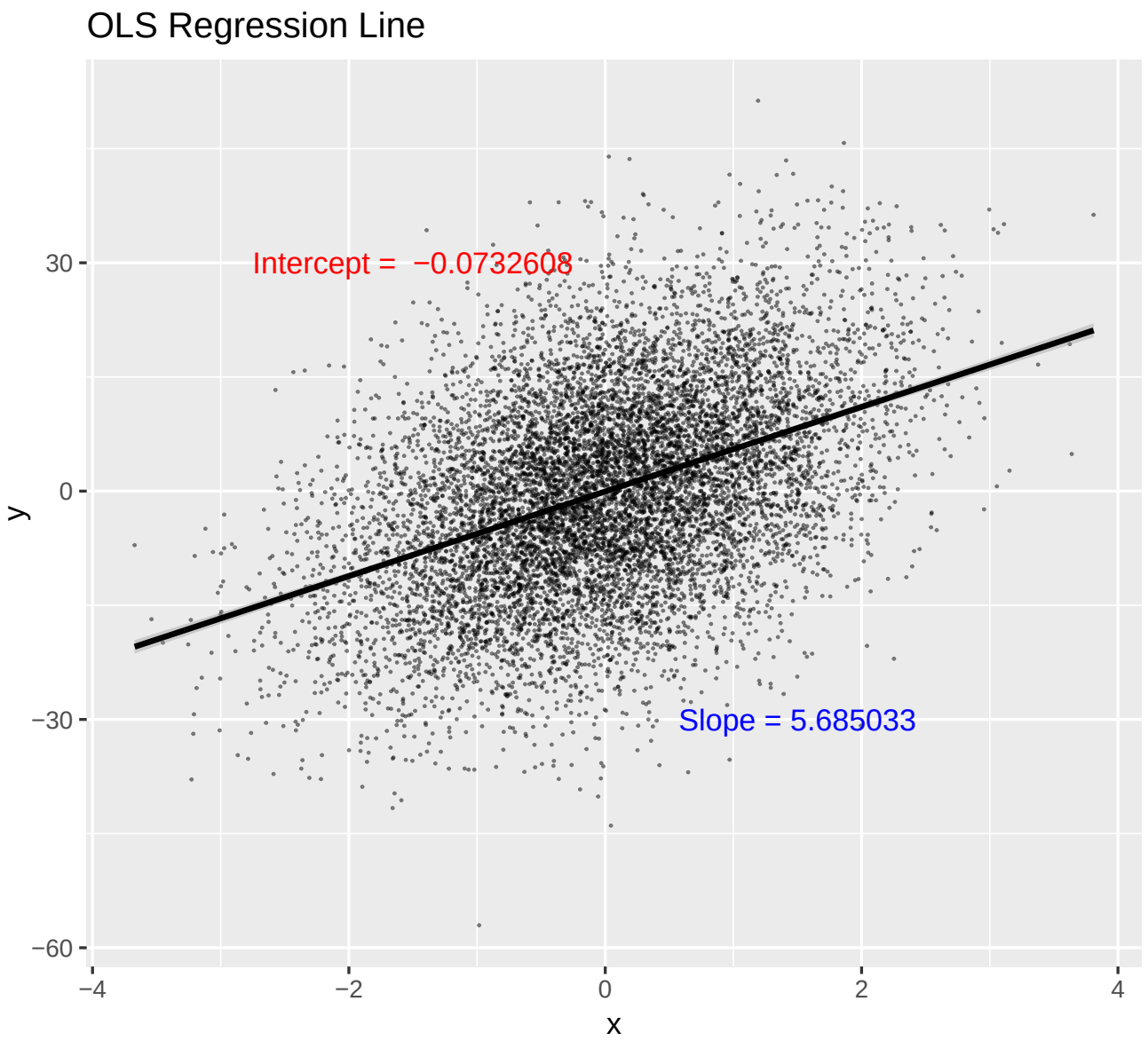
$$Q(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k) = \sum_{i=1}^{n} \left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki} \right) \right)^2 = \sum_{i=1}^{n} \hat{u}_i^2$$

- where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki}$ is the fitted value
- and $\hat{u}_i$ is the residual.

Fitted values and residuals



Fitted regression line



First-order conditions

- Taking derivatives, the OLS estimates $(\hat{\beta}_0, \hat{\beta}_1 \ldots, \hat{\beta}_k)$ satisfy:

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \beta_k x_{ki}) = 0$$

$$\sum_{i=1}^{n} x_{ji}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \beta_k x_{ki}) = 0 \quad (j = 1, \ldots, k)$$

- Proof: during class or e.g. in Wooldridge Appendix 2A, 3A.1
- $k + 1$ equations, $k + 1$ unknown parameters
- Can be solved if there is no perfect collinearity, i.e. there is no perfect linear relationship between the variables in the sample.

Case of a single explanatory variable ($k = 1$)

- For simplicity we use the notation $x_i = x_{1i}$.
- OLS estimate of the slope parameter (if $k = 1$):

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\hat{cov}(x, y)}{\hat{\sigma}_x^2} = \hat{corr}(x, y)\frac{\hat{\sigma}_y}{\hat{\sigma}_x}$$

- where we use $\hat{cov}, \hat{corr}, \hat{\sigma}$ for the sample covariance, correlation, standard deviation.
- OLS estimate of the intercept:

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

Example: "regression to the mean"

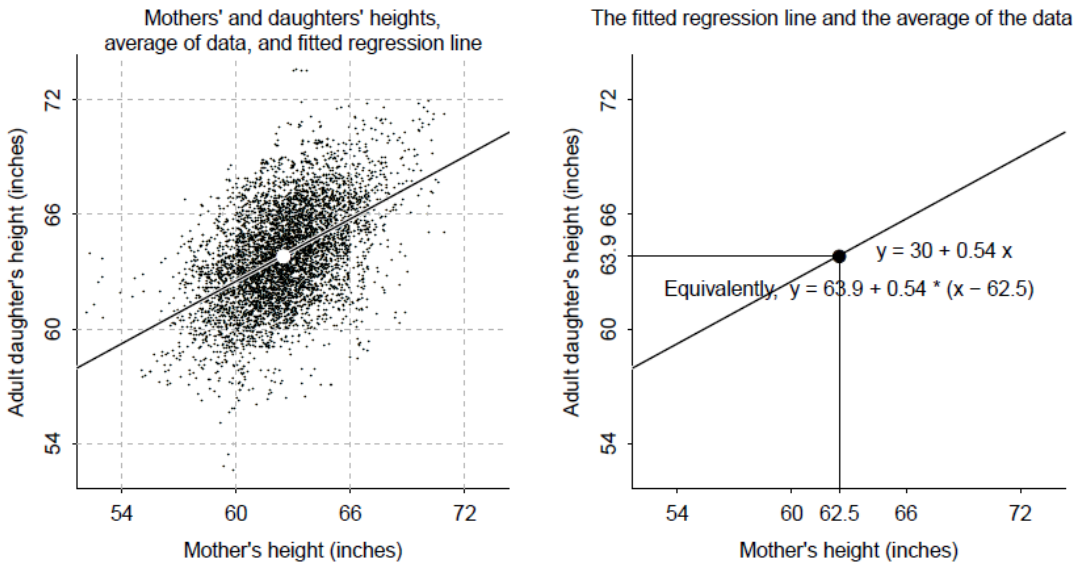


Figure 6.3 *(a) Scatterplot adapted from data from Pearson and Lee (1903) of the heights of mothers and their adult daughters, along with the* regression line *predicting daughters' from mothers' heights. (b) The regression line by itself, just to make the pattern easier to see. The line automatically goes through the mean of the data, and it has a slope of 0.54, implying that, on average, the difference of a daughter's height from the average (mean) of women's heights is only about half the difference of her mother's height from the average.*

- adapted from Gelman et al. (2025) and Pearson and Lee (1903).
- The original example of Galton (1886) was similar but covered sons and parents.
- If $\sigma_x \approx \sigma_y$, then $\hat{\beta}_1 \approx \hat{corr}(x, y)$.

Algebraic Properties of OLS

Key sample-level properties by the derivation of the OLS solution:

- Residuals sum to zero:

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

- Sample covariance between explanatory variables and residuals is zero:

$$\sum_{i=1}^{n} x_{ji} \hat{u}_i = 0 \quad (j = 1, \ldots, k)$$

- The point of averages $(\bar{x}_1, \ldots, \bar{x}_k, \bar{y})$ lies on the OLS regression line:

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

- Fitted values and residuals are uncorrelated in the sample:

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

R-squared

- Total sum of squares: $\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$
- Explained sum of squares: $\text{SSE} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- Residual sum of squares: $\text{SSR} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\hat{u}^2$
- $SST = SSE + SSR$ (Proof: see below)
- R-squared measure of goodness of fit:

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

- $0 \leq R^2 \leq 1$
  - $R^2 = 1$ if perfect linear relationship
  - $R^2 = 0$ if no (linear) relationship at all

Proof: SST = SSE + SSR

- We know that $y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$.
- Squaring both sides and summing over all i:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}).$$

- The cross-product term is zero by the properties of the residuals:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum_{i=1}^{n}\hat{u}_i(\hat{y}_i - \bar{y}) = \sum_{i=1}^{n}\hat{u}_i\hat{y}_i - \bar{y}\sum_{i=1}^{n}\hat{u}_i = 0.$$

- Hence:

$$\text{SST} = \text{SSE} + \text{SSR}.$$

Omitted variable formula

- Fitted regression with two explanatory variables:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{u}_i.$$

- Fitted regression with a single explanatory variable:

$$y_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_{1i} + \tilde{u}_i.$$

- Fitted regression between the two explanatory variables:

$$x_{2i} = \hat{\delta}_0 + \hat{\delta}_1 x_{1i} + \hat{v}_i.$$

- Then

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \hat{\delta}_1.$$

- Hence $\tilde{\beta}_1 = \hat{\beta}_1$ if and only if $\hat{\beta}_2 = 0$ or $\hat{\delta}_1 = 0$.

Proof of the omitted variable formula

- Using the previous equations:

$$\begin{aligned} y_i &= \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2(\hat{\delta}_0 + \hat{\delta}_1 x_{1i} + \hat{v}_i) + \hat{u}_i \\ &= (\hat{\beta}_0 + \hat{\beta}_2\hat{\delta}_0) + (\hat{\beta}_1 + \hat{\beta}_2\hat{\delta}_1)x_{1i} + \hat{\beta}_2\hat{v}_i + \hat{u}_i. \end{aligned}$$

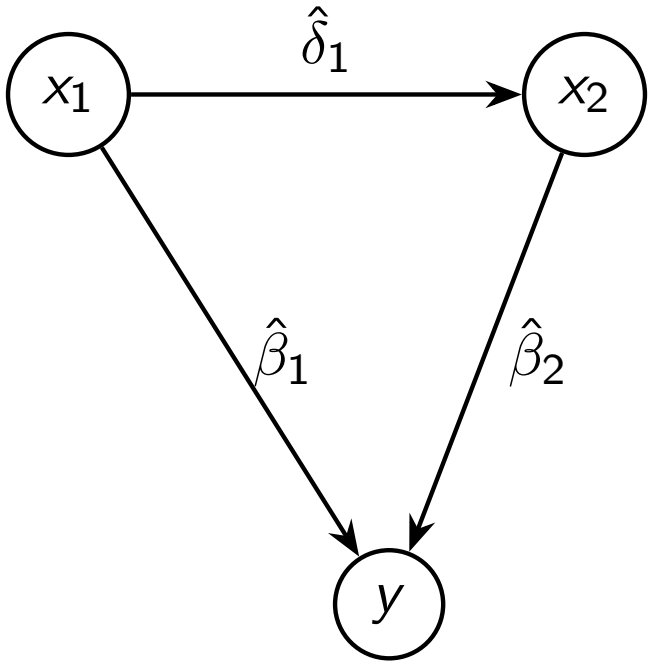
- By the properties of the OLS residuals $\hat{u}$ and $\hat{v}$, the combined residuals are uncorrelated with x_1 :

$$\sum_{i=1}^{n} x_{1i}\left(\hat{\beta}_2\hat{v}_i + \hat{u}_i\right) = \hat{\beta}_2\sum_{i=1}^{n} x_{1i}\hat{v}_i + \sum_{i=1}^{n} x_{1i}\hat{u}_i = 0.$$

- Hence the above equation is the OLS fit of y on x_1, thus:

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2\hat{\delta}_1.$$

Illustration: $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2\hat{\delta}_1$



Example: smoking and birth weight

Dep. var: birth weight (gr)	**(1)** (bwghtgr $\sim$ cigs)	**(2)** (bwghtgr $\sim$ cigs + faminc)
Intercept	3395.53***	3316.22***
	(16.23)	(29.74)
cigarettes per day	-14.57***	-13.14***
	(2.57)	(2.60)
family income ($1000)	—	2.63***
		(0.83)
R^2	0.0227	0.0298
N	1388	1388

- Source: Wooldridge bwght data
- Standard errors in parentheses, *** : $p < 0.01$

Example (cont.)

- $\widehat{\text{bwght}} = 3395.53 - 14.57 \cdot \text{cigs}$
- $\widehat{\text{bwght}} = 3316.22 - 13.14 \cdot \text{cigs} + 2.63 \cdot \text{faminc}$
- $\widehat{\text{faminc}} = 30.16 - 0.54 \cdot \text{cigs}$

- Formula: $-14.57 - (-13.14) = -1.42 = 2.63 \times (-0.54)$

OLS anatomy ("partialling out" formula)

- Fitted regression with two explanatory variables:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$$

- Fitted regression of x_1 on x_2:

$$x_{1i} = \hat{\gamma}_0 + \hat{\gamma}_1 x_{2i} + \hat{r}_{1i}$$

- Then $\hat{\beta}_1$ is also the slope of a regression of y on the $\hat{r}_1$ residuals:

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

- Proof: during class or e.g. in Wooldridge Appendix 3A.2

Material

- Cunningham 2.13-2.15, 2.24
- Wooldridge 2.2-2.3, 3.2