Linear Regression

Linear regression is the most fundamental tool in econometrics that models the relationship between variables.

So how does a linear regression model capture causality ?

Simple Linear Regression : The Model

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

Y: regressand or dependent variable

x : regressor or independent variable

β_0: intercept parameter

β_1: slope parameter

u: error term (unobserved determinants of Y

$\beta_0 + \beta_1 x$ is the *population regression line*

An Empirical Example

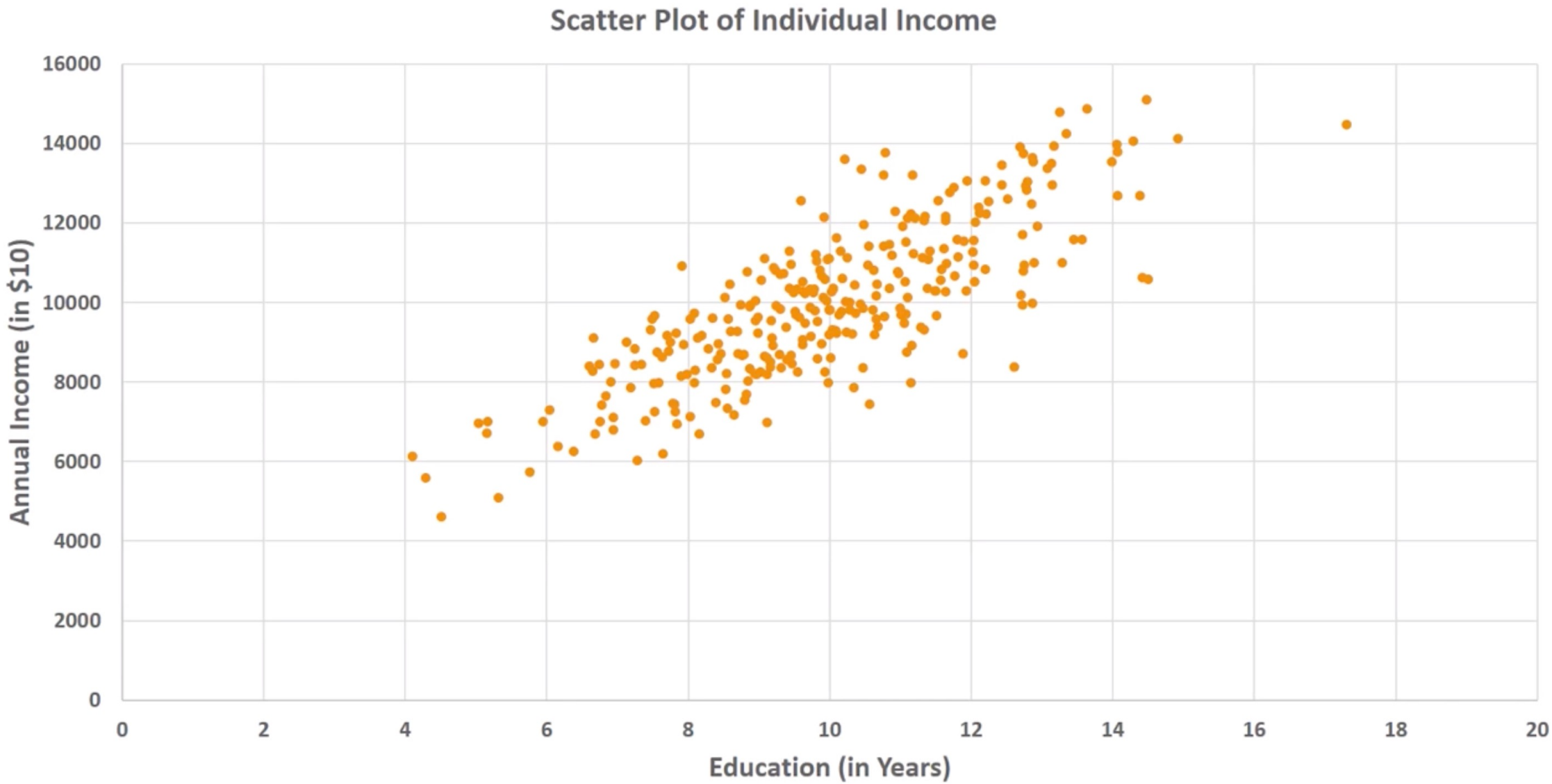
Suppose we want to answer the question: Does education affect earnings ?

The model becomes

$$Income_i = \beta_0 + \beta_{1i} Education\ Level_i + unobserved\ ability_{\ i}$$

If we had all the data for the population, our model would predict the income for any educational level

We do econometrics because we want to find out what β_0 and β_1 are in reality in the population!


Scatter Plot of Individual Income
Annual Income (in $10)
16000
14000
12000
10000
8000
6000
4000
2000
0
0
2
4
6
8
10
12
14
16
18
20
Education (in Years)


Scatter Plot of Individual Income
Annual Income (in $10)
16000
14000
12000
10000
8000
6000
4000
2000
0
0
2
4
6
8
10
12
14
16
18
20
Education (in Years)

Estimation

However, we never have all the data for everyone in the population, so we need to *estimate* β_0 and β_1 .

We find $\widehat{\beta_0}$ and $\widehat{\beta_1}$ from a sample of data which are estimates of β_0 and β_1 in the population.

Now, how do we find $\widehat{\beta_0}$ and $\widehat{\beta_1}$, and are they good estimates of β_0 and β_1 ?

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

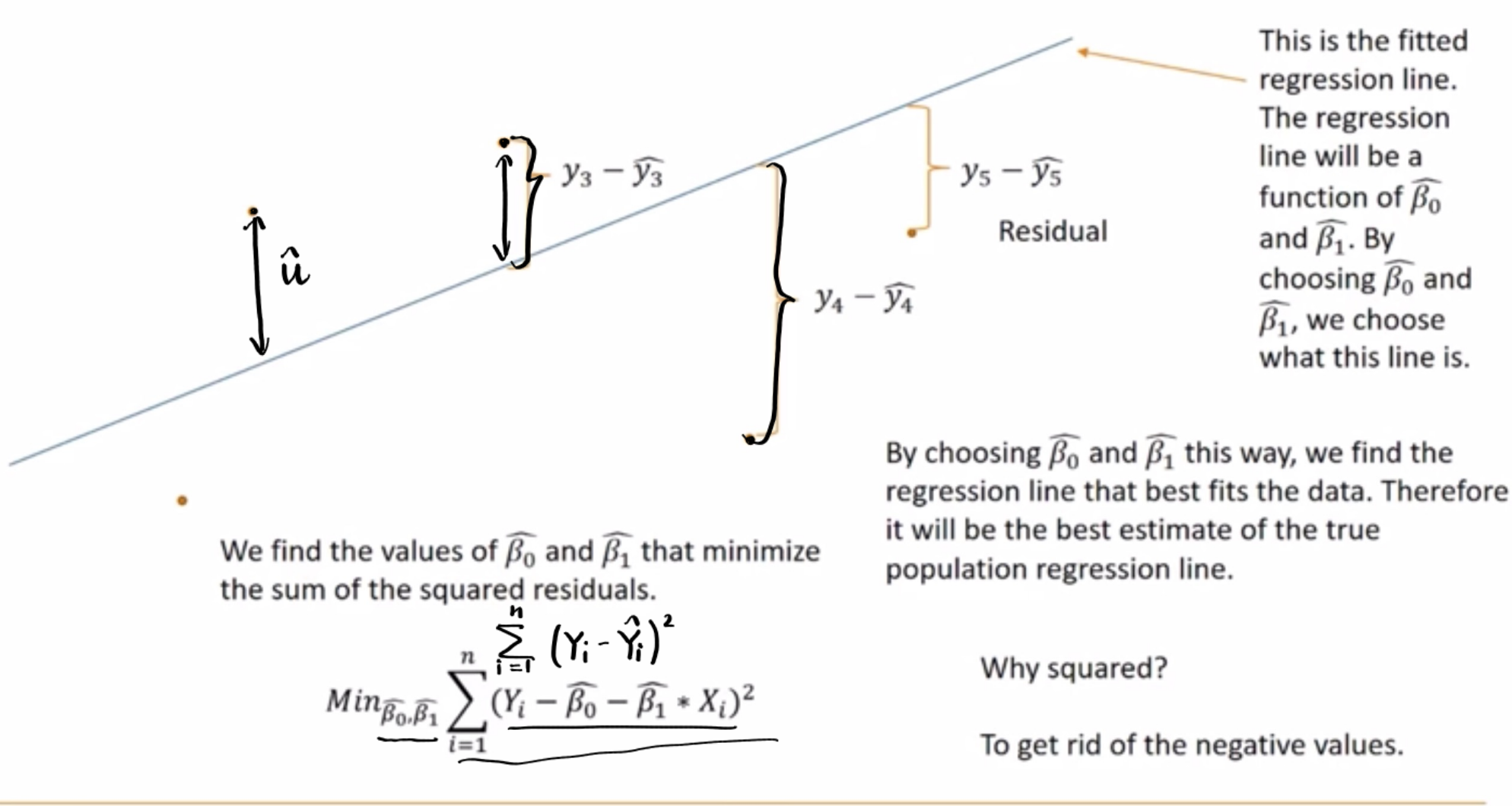
$$y = \hat{\beta}_0 + \hat{\beta}_1 x + u$$

OLS Estimator

Multiple estimated lines could be drawn through the scatterplot. How, then, should you choose among the many possible lines?

By far the most common way is to choose the line that produces the "least squares" fit to these data—that is, to use the **ordinary least squares (OLS) estimator.**

The OLS estimator chooses the regression coefficients so that the estimated regression line is as close as possible to the observed data, where closeness is measured by the sum of the squared residuals made in predicting *Y* given *X*.


$\hat{u}$
$y_3 - \widehat{y_3}$
$y_4 - \widehat{y_4}$
$y_5 - \widehat{y_5}$
Residual
This is the fitted regression line. The regression line will be a function of $\widehat{\beta_0}$ and $\widehat{\beta_1}$. By choosing $\widehat{\beta_0}$ and $\widehat{\beta_1}$, we choose what this line is.
We find the values of $\widehat{\beta_0}$ and $\widehat{\beta_1}$ that minimize the sum of the squared residuals.
$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$
$Min_{\widehat{\beta_0},\widehat{\beta_1}} \sum_{i=1}^{n} (Y_i - \widehat{\beta_0} - \widehat{\beta_1} * X_i)^2$
By choosing $\widehat{\beta_0}$ and $\widehat{\beta_1}$ this way, we find the regression line that best fits the data. Therefore it will be the best estimate of the true population regression line.
Why squared?
To get rid of the negative values.

How do we calculate $\widehat{\boldsymbol{\beta_0}}$ and $\widehat{\boldsymbol{\beta_1}}$?

First, define residual – difference between estimated value of Y and observed value of Y.

$\hat{u}_i = Y_i - \hat{Y}_i$

We know that $\hat{Y}_i = \widehat{\beta_0} + \widehat{\beta_1}\, x_i$

Plug into $\hat{u}_i$ to get $\hat{u}_i = Y_i - \widehat{\beta_0} - \widehat{\beta_1}\, x_i$ (residual we want to minimize across all i)

How do we calculate $\widehat{\boldsymbol{\beta_0}}$ and $\widehat{\boldsymbol{\beta_1}}$?

Now sum the squared residuals for all individuals:

$$\sum_{i=1}^{n} (Y_i - \widehat{\beta_0} - \widehat{\beta_1}\, x_i)^2$$

We want to find the values of $\widehat{\boldsymbol{\beta_0}}$ **and** $\widehat{\boldsymbol{\beta_1}}$ that minimizes this equation

Assumptions

There are certain assumptions for the OLS method.

If these assumptions hold, $\widehat{\beta_0}$ and $\widehat{\beta_1}$ will be the best possible linear estimates that are unbiased.

E[U] = 0

E[XU] = 0

Var[X] finite

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

β_0, β_1: we never know

Want to find $\hat{\beta}_0$, $\hat{\beta}_1$ that estimates β_0, β_1

Assumption 1 : $E[u] = 0$

If $E[u] \neq 0$, we can still normalize β_0 so that $E[\text{new error term}] = 0$

$$y = \underbrace{\beta_0 + E[u]}_{\beta_0'} + \beta_1 x + \underbrace{u - E[u]}_{u'}$$

$$E[u'] = E[u - E[u]] = E[u] - \underbrace{E[E[u]]}_{=E[u]} = 0$$

Assumption 2 : $E[u|x] = 0$

Example :

$E[\text{ability} \mid x = 8] = E[\text{ability} \mid x = 7] = 0$

$E[u|x] = E[u] = 0$

$$E[y] = E[\beta_0 + \beta_1 x + u]$$
$$= E[\beta_0] + E[\beta_1 x] + \cancel{E[u]}^{\,0}$$
$$= \beta_0 + \beta_1 E[x]$$

$$E[Y|X] = \beta_0 + \beta_1 \underbrace{E[X|X]}_{X}$$

$$E[Y|X] = \beta_0 + \beta_1 X \quad \text{population}$$

Dataset

$\{(x_i, y_i) : i = 1, 2, \ldots n\}$ random sample of size n

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

$y_i = \beta_0 + \beta_1 x_i + u_i$

Using assumption 1, $E[u] = 0$

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

$E[y - \beta_0 - \beta_1 x] = 0$ (1)

Using assumption 2,

$E[u|x] = 0$

* Note: $E[E[Y|X]] = E[Y]$ (inner $\downarrow$ xu)

$$\text{Cov}(xu) = \underline{E[xu]} - \cancel{E[x]}E[u] = \underline{E[E[xu|x]]} = E[x\,E[u|x]]$$

(the underlined $E[xu]$ and $E[E[xu|x]]$ are equal by (*))

$$= E[x]\,\cancel{E[u|x]}^{0} = 0$$

$$\Rightarrow E[xu] = 0$$

$$E[x(y - \beta_0 - \beta_1 x)] = 0 \quad (2)$$

$$\left.\begin{aligned} &(1)\ \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ &(2)\ \frac{1}{n}\sum_{i=1}^{n}\left(x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)\right) = 0 \end{aligned}\right\}$$ System of lin. eq.

$$(1)\ \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} y_i}_{\bar{y}} - \underbrace{\frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_0}_{n\hat{\beta}_0} - \underbrace{\frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_1 x_i}_{\hat{\beta}_1 \cdot \frac{1}{n}\sum_{i=1}^{n} x_i = \hat{\beta}_1 \bar{x}}$$

$$= \bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x} = 0$$

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

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

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

$$f(\hat{\beta}_0, \hat{\beta}_1) = \min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$\hat{\beta}_0 : \frac{\partial f(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_0} = 0$$

$$\sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \cdot (-1) = 0 \quad (1)$$

$$\hat{\beta}_1 : \sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \cdot (-x_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (2)$$

$$\Rightarrow \sum_{i=1}^{n} y_i x_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2 = 0$$

$$\sum_{i=1}^{n} y_i x_i - x_i(\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i^2 = 0$$

$$\sum_{i=1}^{n} y_i x_i - x_i \bar{y} + \hat{\beta}_1 x_i \bar{x} - \hat{\beta}_1 x_i^2 = 0$$

$$\sum_{i=1}^{n} y_i x_i - \bar{y} \sum_{i=1}^{n} x_i + \hat{\beta}_1 \bar{x} \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\sum_{i=1}^{n} x_i y_i - \bar{y} n \bar{x} + \hat{\beta}_1 \bar{x}^2 n - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$$

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

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

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

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

$$\sum_{i=1}^{n} x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x}\bar{y}$$
$$= \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \bar{y} - \sum_{i=1}^{n} \bar{x} y_i + \sum_{i=1}^{n} \bar{x}\bar{y}$$
$$= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y}$$
$$= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

$$\sum_{i=1}^{n} x_i^2 - 2x_i\bar{x} + \bar{x}^2$$
$$= \sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i \bar{x} + \sum_{i=1}^{n} \bar{x}^2$$
$$= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2$$
$$= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

R-squared

R-squared measures how well the points are fitted to the regression line and takes a value between 0 and 1.

R^2 is defined as

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS},$$

where

$$TSS = \sum_{1 \leqslant i \leqslant n} (Y_i - \bar{Y}_n)^2$$

$$ESS = \sum_{1 \leqslant i \leqslant n} (\hat{Y}_i - \bar{Y}_n)^2$$

$$SSR = \sum_{1 \leqslant i \leqslant n} \hat{U}_i^2 .$$

Extended Model

Previously, we used the model

$Income_i = \beta_0 + \beta_{1i} Education\ Level_i + \varepsilon_i$ to estimate the effect of education on income

But is this sufficient?

What other factors may affect income ?

Multivariable Linear Regression

$$Income_i = \beta_0 + \beta_{1i} Education\ Level_i + \beta_{2i} Experience_i + \beta_{3i} nonwhite + \beta_{4i} female + \varepsilon_i$$

There may be factors such as experience, race, and gender that affects income that are important enough to not be included in the error term

Interaction Terms

Independent variables may interact with each other. For example, does education matter less if you have more experience?

$$Income_i = \beta_0 + \beta_{1i} Education\ Level_i + \beta_{2i} Experience_i + \beta_{3i} nonwhite + \beta_{4i} female + \boldsymbol{\beta_{5i} Education\ Level_i \times Experience_i} + \varepsilon_i$$

Reading a regression table

$H_0 : \hat{\beta}_1 = 0$

$H_1 : \hat{\beta}_1 \neq 0$

	Dependent variable:		
	log(wage)		
	(1)	(2)	(3)
education	0.077***	0.078***	0.077***
	(0.008)	(0.008)	(0.008)
age	0.063***	0.063***	0.080***
	(0.011)	(0.011)	(0.015)
age squared	−0.001***	−0.001***	−0.001***
	(0.0001)	(0.0001)	(0.0002)
age × female			−0.039*
			(0.022)
age squared × female			0.0004
			(0.0003)
female		−0.254***	0.582
		(0.039)	(0.410)
Constant	−0.303	−0.219	−0.563**
	(0.222)	(0.214)	(0.278)
Observations	534	534	534
R^2	0.241	0.298	0.307
Note:			*p<0.1; **p<0.05; ***p<0.01