ECON207 Session 2

Linear Regression Overview

Anthony Tay

This Version: 31 Jul 2024

Session 2

Deep dive into the **Simple Linear Regression (SLR)** model and **Ordinary Least Squares (OLS)** estimation methodology

Specific example: to estimate the relationship between hourly earnings and years of schooling

- Interested in explaining how hourly earnings vary with years of schooling
  - focus is on variation in hourly earnings and
  - the extent to which years of schooling explains variation in hourly earnings
- Population of interest: US non-institutional employed civilians aged 16 & above in 2018

Session 2

You have a random sample of $n = 4946$ individuals from that population

```
library(tidyverse)  # packages for data wrangling and plotting
library(patchwork)  # ... for composing plots
library(latex2exp)  # ... for LaTeX (math) annotations in plots
dat <- read_csv("data\\earnings2019.csv", show_col_types=FALSE)
head(dat,3)
```

```
# A tibble: 3 x 11
    age height  educ feduc meduc tenure  wexp race   male  earn totalwork
  <dbl>  <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl> <chr> <dbl> <dbl>     <dbl>
1    59     67    12     3     3      5    30 White     0 36.3       1652
2    43     63    10     4     3      7    13 White     1  6.46      1548
3    28     74    12     2     3      6     9 White     1 13.1       2460
```

Session 2

We model our data $\{Y_i, X_i\}_{i=1}^n$ as a random sample from the population

- Y_i and X_i are hourly earnings and years of schooling of individual i respectively

Population represented by some **joint probability distribution function** $f_{X,Y}(x, y)$

$$\text{i.e.,} \quad (X_i, Y_i) \overset{iid}{\sim} f_{X,Y}(x, y)$$

- (X_i, Y_i) independent of (X_j, Y_j), $i \neq j$ (independent across observations)
- X_i and Y_i can be related to each other

Objective will be to learn something about $f_{X,Y}(x, y)$

- E.g., **conditional expectation (conditional mean) of** Y **given** X, $E(Y \mid X)$

Session 2

Agenda for this session

- A bit of math
    - Joint distributions, conditional expectations
    - A bit of optimization theory
- Introduction to the simple linear regression model
    - Estimation via OLS
    - Predictive vs causal interpretations of parameters
    - Application to returns to schooling application
    - *Non-technical intro first, then details*

Session 2.1

Session 2.1 Math Review

- Conditional distributions and expectations
- A bit of optimization theory

Joint and Conditional Distributions

Joint Probability Distribution Function tells us the probability of obtaining various events

Consider bivariate (two variable) case

- Discrete Variables: $f_{X,Y}(x, y) = \Pr(X = x, Y = y)$
- Continuous Variables: $f_{X,Y}(x, y)$ such that

$$\Pr(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f_{X,Y}(x, y)\, dy\, dx$$

In our application later: $earn$ Y is "continuous" whereas $educ$ X is discrete (not a problem)

- $f_{X,Y}(x, y)$ where $Y \in (0, +\infty)$ and $X = 7, 8, \ldots, 17$

Can extend to more than two variables

Joint and Conditional Distributions

We will take a simple, artificial, discrete example to illustrate the concepts

Random Variables X, Y

with possible values

$x =$ 1, 2, 3, 4, 5
$y =$ 3, 3.5, 4, 4.5, 5, 5.5, 6

with Joint PDF
$f_{X,Y}(x, y) = \Pr(X = x, Y = y)$

Y		$X=1$	$X=2$	$X=3$	$X=4$	$X=5$
	6	0	0	0	0	$\frac{1}{20}$
	5.5	0	0	0	$\frac{1}{20}$	$\frac{2}{20}$
	5	0	0	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{1}{20}$
Y	4.5	0	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{1}{20}$	0
	4	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{1}{20}$	0	0
	3.5	$\frac{2}{20}$	$\frac{1}{20}$	0	0	0
	3	$\frac{1}{20}$	0	0	0	0
		1	2	3	4	5
				X		

Marginal (Unconditional) Distributions

Y						
	6	0	0	0	0	$\frac{1}{20}$
	5.5	0	0	0	$\frac{1}{20}$	$\frac{2}{20}$
	5	0	0	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{1}{20}$
Y	4.5	0	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{1}{20}$	0
	4	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{1}{20}$	0	0
	3.5	$\frac{2}{20}$	$\frac{1}{20}$	0	0	0
	3	$\frac{1}{20}$	0	0	0	0
		1	2	3	4	5
				X		

$\rightarrow$

	y	$\Pr(Y=y)$
	6	$\frac{1}{20}$
	5.5	$\frac{3}{20}$
	5	$\frac{4}{20}$
Y	4.5	$\frac{4}{20}$
	4	$\frac{4}{20}$
	3.5	$\frac{3}{20}$
	3	$\frac{1}{20}$

$\rightarrow$

$$\begin{aligned} E(Y) &= 4.5 \\ Var(Y) &= 0.625 \end{aligned}$$

$\downarrow$

x	1	2	3	4	5
$\Pr(X=x)$	$\frac{4}{20}$	$\frac{4}{20}$	$\frac{4}{20}$	$\frac{4}{20}$	$\frac{4}{20}$

$\rightarrow$ $E(X) = 3,\ Var(X) = 2$

Condition Distribution (when $X = 1$)

$$\Pr(Y = 3.0 \mid X = 1) = \frac{1/20}{4/20} = \frac{1}{4}$$

$$\Pr(Y = 3.5 \mid X = 1) = \frac{2/20}{4/20} = \frac{1}{2}$$

$$\Pr(Y = 4.0 \mid X = 1) = \frac{1/20}{4/20} = \frac{1}{4}$$

$$\Pr(Y = 4.5 \mid X = 1) = \frac{0}{4/20} = 0$$

$$\Pr(Y = 5.0 \mid X = 1) = \frac{0}{4/20} = 0$$

$$\Pr(Y = 5.5 \mid X = 1) = \frac{0}{4/20} = 0$$

$$\Pr(Y = 6.0 \mid X = 1) = \frac{0}{4/20} = 0$$

In general, we have

$$\Pr(Y = y \mid X = x) = \frac{\Pr(Y = y, X = x)}{\Pr(X = x)}$$

- or -

$$\Pr(Y = y, X = x) = \Pr(Y = y \mid X = x)\Pr(X = x)$$

We can write

$$f_{X,Y}(x, y) = f_{Y|X}(y \mid x) f_X(x) = f_{X|Y}(x \mid y) f_Y(y)$$

Conditional Distribution / Expectation / Variance

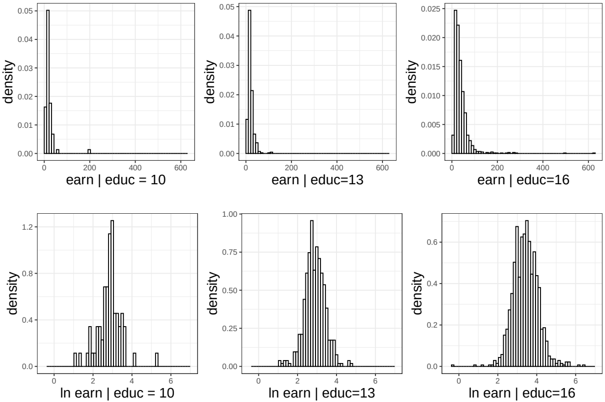
Calculate for each possible value of X

$$
\begin{array}{cc|ccccc}
 & 6 & 0 & 0 & 0 & 0 & \frac{1}{4} \\
 & 5.5 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{2} \\
 & 5 & 0 & 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\
Y & 4.5 & 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 \\
 & 4 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 & 0 \\
 & 3.5 & \frac{1}{2} & \frac{1}{4} & 0 & 0 & 0 \\
 & 3 & \frac{1}{4} & 0 & 0 & 0 & 0 \\
\hline
 & & 1 & 2 & 3 & 4 & 5 \\
 & & & & X & &
\end{array}
\rightarrow
\begin{array}{c|ccccc}
X & 1 & 2 & 3 & 4 & 5 \\
\hline
E(Y \mid X) & 3.5 & 4 & 4.5 & 5 & 5.5 \\
Var(Y \mid X) & 0.125 & 0.125 & 0.125 & 0.125 & 0.125
\end{array}
$$

In this example, $E(Y \mid X)$ varies with X, $Var(Y \mid X)$ is constant for all X

Conditional Distributions / Expectation / Variance

For $earnings$ and $educ$ sample



Suggests

- $Y|X \sim$ Log-normal(μ, σ^2)?
- $\ln Y|X \sim$ Normal(μ, σ^2)?
- Does μ and/or σ^2 depend on X?

Conditional Expectations

For continuous random variables X, Y with joint pdf $f_{X,Y}(x,y)$ we have

- $f_X(x) = \int_Y f_{X,Y}(x,y)\,dy$ and $f_Y(y) = \int_X f_{X,Y}(x,y)\,dx$
- $f_{X,Y}(x,y) = f_{Y|X}(y \mid x) f_X(x) = f_{X|Y}(x \mid y) f_X(y)$
- $E(Y \mid X = x) = \int_Y y f_{Y|X}(y \mid x)\,dy$
- $Var(Y \mid X = x) = \int_Y (y - E_{Y|X}(Y))^2 f_{Y|X}(y \mid x)\,dy$

Conditional Expectations

Manipulating Conditional Expectations (and Conditional Variances)

- Treat conditioning information as fixed

Examples:

- $E(aXY \mid X) = aXE(Y \mid X)$
- $Var(aXY \mid X) = a^2X^2 Var(Y \mid X)$
- $Var(aX \mid X) = 0$ (cf. $Var(aX) = a^2 Var(X)$)

Law of Iterated Expectations

Returning to our example, and "reinstating" the randomness in X

$X = x$	1	2	3	4	5
$E(Y \mid X = x)$	3.5	4	4.5	5	5.5
$\Pr(X = x)$	0.2	0.2	0.2	0.2	0.2

- $E(Y \mid X)$ is a function of X
- Because X is a random variable, so is $E(Y \mid X)$

Here $E(Y \mid X)$ is uniformly distributed over 3.5, 4.0, 4.5, 5.0, 5.5

Law of Iterated Expectations

Can compute mean and variance of $E(Y \mid X)$:

- $E_X(E_{Y|X}(Y \mid X)) = 3.5(0.2) + 4.0(0.2) + \cdots + 5.5(0.2) = 4.5$
- $Var_X(E_{Y|X}(Y \mid X)) = ?$ (Exercise)

Recall $E_Y(Y) = 4.5$

Not a coincidence that $E(Y)$ is the same as $E_X(E_{Y|X}(Y \mid X))$

Law of Iterated Expectations

$$E_Y(Y) = E_X(E_{Y|X}(Y \mid X))$$

Law of Iterated Expectations

Special case of $E_{X,Y}(g(X,Y)) = E_X(E_{Y|X}(g(X,Y)))$

$$\begin{aligned}
E_{X,Y}(g(X,Y)) &= \int_X \int_Y g(x,y) f_{X,Y}(x,y)\,dy\,dx \\
&= \int_X \int_Y g(x,y) f_{Y|X}(y \mid x) f_X(x)\,dy\,dx \\
&= \int_X \left(\int_Y g(x,y) f_{Y|X}(y \mid x)\,dy \right) f_X(x)\,dx \\
&= E_X\left(E_{Y|X}(g(X,Y) \mid X) \right)
\end{aligned}$$

If $g(X,Y) = Y$, we get the law of iterated expectations

Law of Iterated Expectations

Implications of Law of Iterated Expectations

If $E(Y \mid X) = c$, then (a) $E(Y) = c$, (b) $Cov(X, Y) = 0$

If $E(Y \mid X) = \beta_0 + \beta_1 X$, then (c) $\beta_0 = E(Y) - \beta_1 E(X)$, (d) $\beta_1 = \dfrac{Cov(X,Y)}{Var(X)}$

Proof: If $E(Y \mid X) = c$, then

(a) $E(Y) = E(E(Y|X)) = E(c) = c$

(b) $Cov(X, Y) = E(YX) - E(Y)E(X) = E(XE(Y \mid X)) - cE(X) = cE(X) - cE(X) = 0$

If $E(Y \mid X) = \beta_0 + \beta_1 X$, then

(c) $E(Y) = E(E(Y|X)) = E(\beta_0 + \beta_1 X) = \beta_0 + \beta_1 E(X)$

Law of Iterated Expectations

(d) We have

$$\begin{aligned} E(YX) &= E(E(YX \mid X)) = E(XE(Y \mid X)) = E(X(\beta_0 + \beta_1 X)) \\ &= \beta_0 E(X) + \beta_1 E(X^2) \end{aligned}$$

Substituting in $\beta_0 = E(Y) - \beta_1 E(X)$ gives

$$E(YX) = E(Y)E(X) - \beta_1 E(X)^2 + \beta_1 E(X^2) = E(Y)E(X) + \beta_1 Var(X)$$

$$\beta_1 = \frac{E(YX) - E(Y)E(X)}{Var(X)} = \frac{Cov(X,Y)}{Var(X)}$$

Law of Iterated Expectation

Is there a law of iterated variance?

If yes, what does it look like?

We have

$$Var(Y) = E(Var(Y \mid X)) + Var(E(Y|X))$$

Proof: exercise

Law of Iterated Expectations

Y						
	10	0	0	0	0	$\frac{1}{10}$
	9	0	0	0	$\frac{1}{10}$	0
	8	0	0	$\frac{1}{10}$	0	0
	7	0	$\frac{1}{10}$	0	0	0
	6	$\frac{1}{10}$	0	0	0	0
Y	5	$\frac{1}{10}$	0	0	0	0
	4	0	$\frac{1}{10}$	0	0	0
	3	0	0	$\frac{1}{10}$	0	0
	2	0	0	0	$\frac{1}{10}$	0
	1	0	0	0	0	$\frac{1}{10}$
		1	2	3	4	5
				X		

Exercise:

- Find marginal distribution of X and Y
- Find conditional distribution of Y given X
- Find $Cov(X, Y)$
- How is cond. distribution of Y related to X?

Law of Iterated Expectations

X and Y in exercise are uncorrelated but not independent

Two random variables are independent if

$$\Pr(Y = y \mid X = x) = \Pr(Y = y) \text{ for all } x \text{ and } y$$

or

$$\Pr(Y = y, X = x) = \Pr(Y = y)\Pr(X = x) \text{ for all } x \text{ and } y$$

For continuous rv: $f_{Y|X}(y \mid x) = f_Y(y)$ or $f_{Y,X}(y,x) = f_Y(y)f_X(x)$

Independent Random Variables

Suppose Y and X have the following joint pdf:

Y						
	5	0.01	0.04	0.03	0.01	0.01
	4	0.02	0.08	0.06	0.02	0.02
Y	3	0.04	0.16	0.12	0.04	0.04
	2	0.02	0.08	0.06	0.02	0.02
	1	0.01	0.04	0.03	0.01	0.01
		1	2	3	4	5
				X		

Independent? Identically Distributed?

A Little Optimization Theory

- Given $f(x)$, find x^* such that $f(x^*)$ is max or min
- Given $f(x, y)$, find (x^*, y^*) such that $f(x^*, y^*)$ is max or min

Presume you know differentiation and partial differentiation, e.g.,

- $f(x) = x^2 \;\Rightarrow\; f'(x) = 2x$ (if $y = f(x)$, we can write $f'(x)$ as $\dfrac{dy}{dx}$)
- $z = f(x, y) = x^2y^3 \;\Rightarrow\; \dfrac{\partial z}{\partial x} = 2xy^3$ and $\dfrac{\partial z}{\partial y} = 3x^2y^2$

$\dfrac{\partial z}{\partial x}$: how z changes with x, holding y fixed $\qquad$ $\dfrac{\partial z}{\partial y}$: how z changes with y, holding x fixed

A Little Optimization Theory

$z = f(x, y) = x^2 y^3$

First partial derivatives: $\dfrac{\partial z}{\partial x} = 2xy^3$, $\dfrac{\partial z}{\partial y} = 3x^2 y^2$

Second partial derivatives:

$$\frac{\partial^2 z}{\partial x^2} = \frac{\partial}{\partial x}\left(\frac{\partial z}{\partial x}\right) = 2y^3 \qquad \frac{\partial^2 z}{\partial y \partial x} = \frac{\partial}{\partial y}\left(\frac{\partial z}{\partial x}\right) = 6xy^2$$

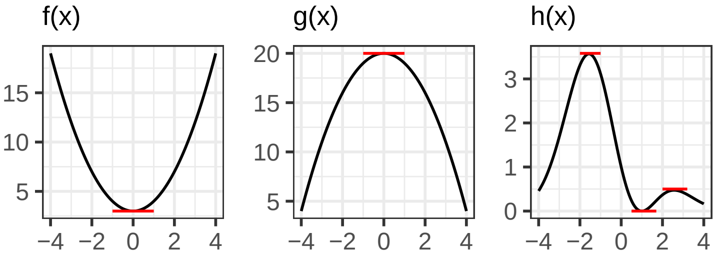
$$\frac{\partial^2 z}{\partial x \partial y} = \frac{\partial}{\partial x}\left(\frac{\partial z}{\partial y}\right) = 6xy^2 \qquad \frac{\partial^2 z}{\partial y^2} = \frac{\partial}{\partial y}\left(\frac{\partial z}{\partial y}\right) = 6x^2 y$$

Notice that $\dfrac{\partial^2 z}{\partial y \partial x} = \dfrac{\partial^2 z}{\partial x \partial y}$. This is true in general ("Young's Theorem")

A Little Optimization Theory

We'll look only at the cases relevant to us

- If $f''(x) > 0$ for all x, then $f'(x^*) = 0 \Rightarrow x^*$ minimizes $f(x)$
- If $f''(x) < 0$ for all x, then $f'(x^*) = 0 \Rightarrow x^*$ maximizes $f(x)$
- $f''(x) > 0$, $f''(x) < 0$ are not necessary for $f(x)$ to have a min/max



Two variables $f(x, y)$

Same idea for $f(x, y)$: f concave implies stationary pt is max, f convex implies stationary pt is min

Stationary point: (x^*, y^*) such that $f'_x(x^*, y^*) = 0$ and $f'_y(x^*, y^*) = 0$

- If $f(x, y)$ is convex; stationary point is minimum point
- If $f(x, y)$ is concave; stationary point is maximum point
- If $v_1^2 \dfrac{\partial^2 f(x,y)}{\partial x^2} + 2 v_1 v_2 \dfrac{\partial^2 f(x,y)}{\partial x \partial y} + v_2^2 \dfrac{\partial^2 f(x,y)}{\partial y^2} < 0$ for all v_1, v_2 not both zero

then $f(x, y)$ is concave

- If $v_1^2 \dfrac{\partial^2 f(x,y)}{\partial x^2} + 2 v_1 v_2 \dfrac{\partial^2 f(x,y)}{\partial x \partial y} + v_2^2 \dfrac{\partial^2 f(x,y)}{\partial y^2} > 0$ for all v_1, v_2 not both zero

then $f(x, y)$ is convex

Two variables $f(x, y)$

Rough explanation: let $x = x_0 + v_1 s$, $y = y_0 + v_2 s$, $v_1^2 + v_2^2 = 1$, and

$$z(s) = f(x(s), y(s))$$

Note that $z(0) = f(x_0, y_0)$, $dx/ds = v_1$, $dy/dx = v_2$

Directional derivative at (x_0, y_0) in direction $v = (v_1, v_2)$ is

$$\frac{dz}{ds} = f'_x(x, y)\frac{dx}{ds} + f'_y(x, y)\frac{dy}{ds} = v_1 f'_x(x, y) + v_2 f'_y(x, y)$$

(x_0, y_0) stationary pt means slope $= 0$ in all directions at the point, i.e.,

$$f'_x(x_0, y_0) = 0, f'_x(x_0, y_0) = 0\,.$$

Two variables $f(x, y)$

Second directional derivative is

$$
\begin{aligned}
\frac{d^2 z}{ds^2} &= f'_x(x,y)\frac{dx}{ds} + f'_y(x,y)\frac{dy}{ds} \\
&= \left[f''_{xx}(x,y)\frac{dx}{ds} + f''_{xy}(x,y)\frac{dy}{ds}\right]\frac{dx}{ds} + f'_x(x,y)\frac{d^2x}{ds^2} \; + \\
&\qquad \left[f''_{yx}(x,y)\frac{dx}{ds} + f''_{yy}(x,y)\frac{dy}{ds}\right]\frac{dy}{ds} + f'_y(x,y)\frac{d^2y}{ds^2} \\
&= v_1^2 f''_{xx}(x,y) + 2v_1 v_2 f''_{xy}(x,y) + v_2^2 f''_{yy}(x,y)
\end{aligned}
$$

since $dx/ds = v_1$, $dy/ds = v_2$, and $d^2x/ds^2 = d^2y/ds^2 = 0$

f convex: slope decreasing in all directions, i.e., $z''(0) < 0$ for all v and for all (x_0, y_0)

f concave: slope increasing in all directions, i.e., $z''(0) > 0$ for all v and for all (x_0, y_0)

Two variables $f(x, y)$

Find minimum point of $f(x, y) = x^2 + xy + y^2$

We have $f_x'(x, y) = 2x + y$ and $f_y'(x, y) = y + 2y$

Therefore

$$\text{FOC:} \quad \begin{aligned} f_x'(x^*, y^*) &= 2x^* + y^* = 0 \\ f_y'(x^*, y^*) &= 2y^* + x^* = 0 \end{aligned} \quad \Rightarrow \quad (x^*, y^*) = (0, 0) \;\; \text{stationary point}$$

Two variables $f(x, y)$

We have $f''_{xx}(x,y) = 2$, $f''_{xy}(x,y) = f''_{yx}(x,y) = 1$ and $f''_{yy}(x,y) = 2$, therefore

$$
\begin{aligned}
& v_1^2 f''_{xx}(x,y) + 2v_1 v_2 f''_{xy}(x,y) + v_2^2 f''_{yy}(x,y) \\
= \;& 2(v_1^2 + v_1 v_2 + v_2^2) \\
= \;& 2[(v_1 + 0.5v_2)^2 + 0.75v_2^2] > 0
\end{aligned}
$$

for all v_1, v_2 not both equal to zero

The function $f(x, y)$ is convex

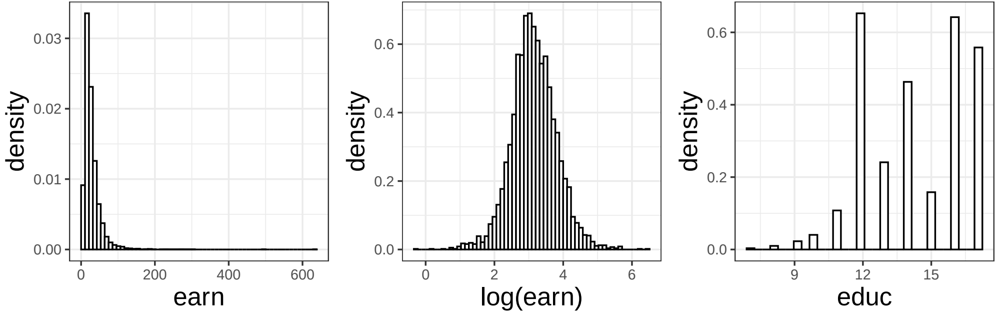
Therefore $(x^*, y^*) = (0, 0)$ is a minimum point of $f(x,y) = x^2 + xy + y^2$

Session 2.2

Session 2.2 Linear Regression Overview

- Non-technical Introduction to the Linear Regression Model

Linear Regression Overview



What is the relationship between `earn` and `educ`?

Linear Regression Overview

Perhaps estimate $Cov(earn, educ)$?

```
dat1 <- dat %>% select(c(earn, educ))
N <- nrow(dat1)
cov_yx <- cov(dat1)[1,2]
cor_yx <- cor(dat1)[1,2]
set.seed(1701)  # Set random number seed for bootstrap standard errors
B <- 200;       # Number of bootstrap replications
bcov <- bcor <- rep(NA, B) # To store bootstrapped covs and cors
for (b in 1:B){
  bcov[b] <- cov(sample_n(dat1, N, replace=TRUE))[1,2]
  bcor[b] <- cor(sample_n(dat1, N, replace=TRUE))[1,2]
}
cat("sample cov(earn, educ):", round(cov_yx,3), " s.e.:", round(sqrt(var(bcov)),3), "\n")
cat("sample cor(earn, educ):", round(cor_yx,3), "  s.e.:", round(sqrt(var(bcor)),3), "\n")
```

```
sample cov(earn, educ): 17.622  s.e.: 0.736
sample cor(earn, educ): 0.321   s.e.: 0.017
```

Linear Regression Overview

Positive correlation $\longrightarrow$ higher `earn` associated with more years of `educ`

Another perspective: estimate $E(earn \mid educ)$

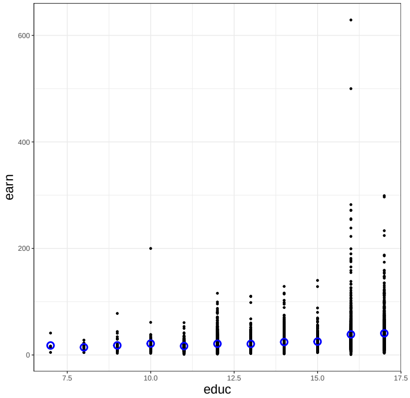
Tells you how `earn` changes with one year change in `educ`,

i.e., $E(earn \mid educ + 1) - E(earn \mid educ)$ at various levels of `educ`

How to estimate $E(earn \mid educ)$?

Linear Regression Overview

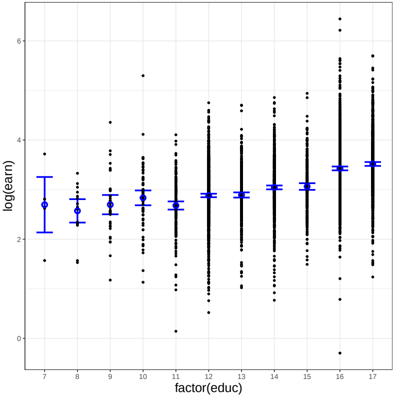
Suppose we estimate mean earn at each level of educ




```
# A tibble: 11 x 4
    educ     n mean_earn se_earn
   <dbl> <int>     <dbl>   <dbl>
 1     7     6      17.8   5.00
 2     8    17      14.4   1.48
 3     9    39      18.0   2.16
 4    10    69      21.4   2.87
 5    11   184      16.8   0.668
 6    12  1113      21.0   0.387
 7    13   411      20.8   0.623
 8    14   790      24.3   0.509
 9    15   270      25.1   1.01
10    16  1095      38.6   1.13
11    17   952      40.7   0.963
```

Linear Regression Overview

Alternative: estimate mean `ln(earn)` at each level of educ $E(\ln earn \mid educ = i)$ for $i = 7, \ldots, 17$




```
# A tibble: 11 x 4
    educ     n mean_logearn se_logearn
   <dbl> <int>        <dbl>      <dbl>
 1     7     6         2.70     0.280
 2     8    17         2.57     0.118
 3     9    39         2.70     0.0974
 4    10    69         2.83     0.0749
 5    11   184         2.68     0.0411
 6    12  1113         2.88     0.0176
 7    13   411         2.89     0.0261
 8    14   790         3.04     0.0195
 9    15   270         3.06     0.0337
10    16  1095         3.43     0.0193
11    17   952         3.52     0.0196
```

Linear Regression Overview

How to interpret differences in $\ln x_2 - \ln x_1$?

Linear approximation of a function $f(x)$ around $x = a$

$$f(x) \approx p(x) = f(a) + f'(a)(x-a) \quad \text{for} \quad x \quad \text{near} \quad a$$

When $f(x) = \ln x$, $f'(x) = 1/x$

Therefore, for all x near x_1, we have

$$\ln x \approx p(x) = \ln x_1 + \frac{1}{x_1}(x - x_1)$$

For x_2 near x_1, we have $\ln x_2 - \ln x_1 \approx \dfrac{x_2 - x_1}{x_1}$

i.e., $\ln x_2 - \ln x_1$ is approx. difference between x_2 and x_1 as percentage of x_1

Linear Regression Overview

Another approach: assume

$$E(\ln earn \mid educ) = \beta_0 + \beta_1 educ$$

Interpretation: β_1 is percentage difference in mean hourly earnings comparing two people with one year difference years of schooling

Use data to estimate β_0 and β_1:

$$\widehat{E}(\ln earn \mid educ) = \hat{\beta}_0 + \hat{\beta}_1 educ$$

Linear Regression Overview

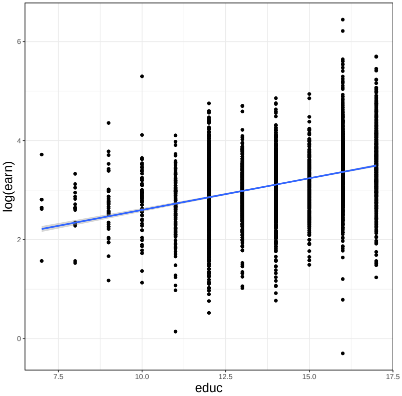
How to estimate β_0 and β_1? Maybe choose $\widehat{\beta}_0$ and $\widehat{\beta}_1$ to minimize

$$\sum_{i=1}^{n}(\ln earn_i - \widehat{\beta}_0 - \widehat{\beta}_1 educ_i)^2$$

"Ordinary Least Squares"

Qn: Is this a good way to estimate β_0 and β_1? Good under what conditions? Do those conditions hold? Are there better ways?

Linear Regression Overview




```
mdl1 <- lm(log(earn) ~ educ, data = dat)
summary(mdl1)$coefficients %>% round(4)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.320     0.0575 22.9399        0
educ           0.128     0.0040 32.1700        0
```

We will study the math later (in detail!!)

Additional one year of schooling is associated with additional mean hourly earnings of around 12.8%

Statistically significant

Linear Regression Overview

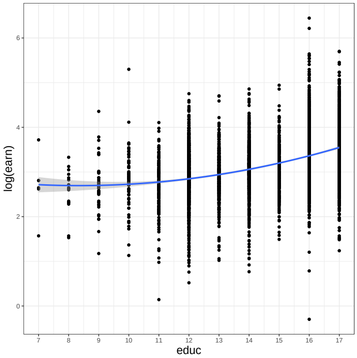
- Linear regression uses a “global approach”
- Uses assumption about form of conditional expectation to “tie” data together
- Uses all observations to estimate two parameters
- Able to give “big perspective” view of relationship between `ln earn` and `educ`
- Previous approach was “local”, and estimated 11 means
- Is assumption about form of conditional expectation correct? (Probably not)

Try

$$E(\ln earn \mid educ) = \beta_0 + \beta_1 educ + \beta_2 educ^2$$

Linear Regression Overview

Choose $\hat\beta_0, \hat\beta_1, \hat\beta_2$ to minimize $\displaystyle\sum_{i=1}^{n}(\ln earn_i - \hat\beta_0 - \hat\beta_1 educ - \hat\beta_2 educ^2)^2$




```
mdl1a <- lm(log(earn) ~ educ + I(educ^2), data = dat)
summary(mdl1a)$coefficients %>% round(4)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.4870     0.3573  9.7607    0e+00
educ         -0.1899     0.0519 -3.6600    3e-04
I(educ^2)     0.0114     0.0019  6.1455    0e+00
```

$$
\begin{aligned}
&\widehat{E}(\ln earn \mid educ+1) - \widehat{E}(\ln earn \mid educ) \\
&= \hat\beta_0 + \hat\beta_1(educ+1) + \hat\beta_2(educ+1)^2 \\
&\qquad - (\hat\beta_0 + \hat\beta_1 educ + \hat\beta_2 educ^2) \\
&= \hat\beta_1 + \hat\beta_2 + 2\hat\beta_2\, educ
\end{aligned}
$$

Linear Regression Overview

Qns:

Is $E(\ln earn \mid educ)$ *really* quadratic?

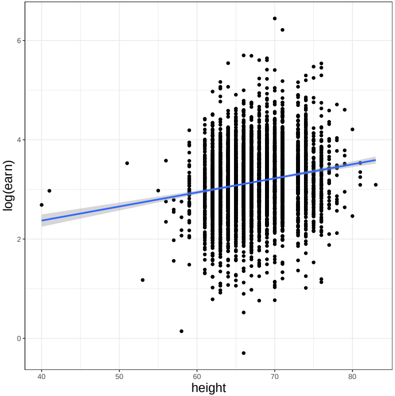
- Probably not, but probably close enough. Nice "big picture" view of relationship
- good Bias-Variance trade-off compared with "local approach"

Why are we estimating $E(Y \mid X)$?

- To use educ to predict hourly earnings?
  - Some predictive ability but not a lot
  - What other factors are relevant?
- To determine if variation in hourly earnings is determined by years of schooling?
  - To what extent can we view estimated relationship as "causal"?

Linear Regression Overview

Consider $E(\ln earn \mid height)$, assuming $E(\ln earn \mid height) = \beta_0 + \beta\, height$




```r
mdl2 <- lm(log(earn) ~ height, data = dat)
summary(mdl2)$coefficients %>% round(4)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.2382     0.1536  8.0617        0
height        0.0284     0.0023 12.4766        0
```

Every additional inch in height is associated with 2.8% more hourly earnings (statistically and economically significant)

Can we say `height` causes higher hourly earnings?

If "no", then how to interpret the estimates?

Linear Regression Overview

What if we include `male` in specification:

$$E(\ln earn \mid height) = \beta_0 + \beta_1\, height + \beta_2\, male$$

```
mdl2a <- lm(log(earn) ~ height + male, data = dat)
summary(mdl2a)$coefficients %>% round(4)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.8140     0.2003  9.0568        0
height        0.0191     0.0031  6.1896        0
male          0.1109     0.0248  4.4668        0
```

Coefficient in `height` now smaller

- Why does including `male` change the coefficient (and standard error) on `height`?
- Does coefficient of `height` now reflect extent of causality of `height` on `earn`?
- What *exactly* happens when we include a new variable?

Linear Regression Overview

A fuller specification

```
mdl3 <- lm(log(earn) ~ height + male + educ + I(educ^2) + wexp + age + I(age^2) + tenure, data = dat)
summary(mdl3)$coefficients %>% round(4)
```

```
              Estimate Std. Error  t value Pr(>|t|)
(Intercept)     0.7950     0.3757   2.1164   0.0344
height          0.0132     0.0026   4.9756   0.0000
male            0.1834     0.0215   8.5439   0.0000
educ           -0.1553     0.0483  -3.2180   0.0013
I(educ^2)       0.0102     0.0017   5.9257   0.0000
wexp           -0.0022     0.0011  -2.0137   0.0441
age             0.0621     0.0046  13.3546   0.0000
I(age^2)       -0.0007     0.0001 -12.9063   0.0000
tenure          0.0147     0.0011  14.0261   0.0000
```

(Height still significant...)

Session 2.3

Session 2.3 The Simple Linear Regression Model

- Derive relevant formulas related to estimation of simple linear regression model
- When OLS gives good estimators and when not
- How to interpret fitted parameters and fitted model
  - Statistical Interpretation
  - Economic Interpretation
  - Predictive vs Causal Interpretation

(Next session: what happens when we include new variables)

The Simple Linear Regression Model

Assumptions about population:

- $E(Y \mid X) = \beta_0 + \beta_1 X$

Assumptions about sample

- $\{X_i, Y_i\}_{i=1}^n$ is a random sample from the population of interest
- $\sum_{i=1}^n (X_i - \overline{X})^2 > 0$, i.e., there is variation in the X_i observations

Objective is to estimate $E(Y \mid X)$, i.e., to obtain an empirical model

$$E(\widehat{Y \mid X}) = \hat{\beta}_0 + \hat{\beta}_1 X$$

Will often write the empirical model as $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

The Simple Linear Regression Model

If we define $\epsilon = Y - E(Y \mid X) = Y - \beta_0 - \beta_1 X$, we can write the model as

$$Y = \beta_0 + \beta_1 X + \epsilon\,,\; E(\epsilon \mid X) = 0 \;\text{ and }\; Var(\epsilon \mid X) = \sigma^2$$

Note that $E(\epsilon \mid X) = 0$ implies

- $E(\epsilon) = 0$ and
- $Cov(\epsilon, X) = 0$

The second of these is often written $E(\epsilon X) = 0$ since

$$Cov(\epsilon, X) = 0 \Leftrightarrow E(\epsilon X) - E(\epsilon)E(X) = 0 \Leftrightarrow E(\epsilon X) = 0$$

The Simple Linear Regression Model

If sample is representative of the population, then the sample satisfies

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \,,\, i = 1, \ldots, n \quad \text{("Linear Regression Model")}$$

- $Y_i \sim$ "Regressand", "Dependent Variable", "Outcome Variable"
- $X_i \sim$ "Regressor", "Independent Variable", "Predictor", "Feature"
- $\epsilon_i \sim$ "Noise" or "Error" term
- β_1 is the slope coefficient or simply "coefficient" on X_i
- β_0 is the (y-) intercept term or "constant" term

In Machine Learning, β_0 is called the "bias". We will **not** use that terminology here.

The Simple Linear Regression Model

The assumptions imply

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \,,\; i = 1, \ldots, n$$

- $E(\epsilon_i \mid X_1, X_2, \ldots, X_n) = 0$ for all $i = 1, \ldots, n$
- $E(\epsilon_i \epsilon_j \mid X_1, X_2, \ldots, X_n) = 0$ for all $i, j = 1, \ldots, n$, $i \neq j$

The first comes from $E(\epsilon \mid X) = 0$ and the iid assumption.

The second comes from the iid assumption.

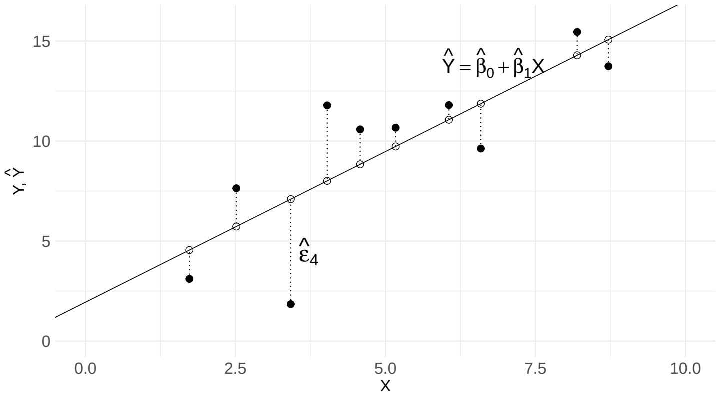
Estimation by Ordinary Least Squares

Given any potential estimator $\hat{\beta}_0$ and $\hat{\beta}_1$, define

- Fitted values: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_0 X_i$
- Residuals: $\hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_0 X_i$
- Sum of Squared Residuals (SSR): $\sum_{i=1}^n \hat{\epsilon}_i^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_0 X_i)^2$

$$\text{OLS}: \quad \hat{\beta}_0^{ols}, \hat{\beta}_1^{ols} = \underset{\hat{\beta}_0, \hat{\beta}_1}{\arg\min} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

Ordinary Least Squares



data in ols01.csv

```
# A tibble: 10 x 2
       X     Y
   <dbl> <dbl>
 1  2.51  7.64
 2  5.17 10.7
 3  1.73  3.11
 4  3.42  1.85
 5  4.03 11.8
 6  4.58 10.6
 7  8.19 15.5
 8  6.59  9.63
 9  8.72 13.7
10  6.06 11.8
```

Ordinary Least Squares

First-Order Condition:

$$(1) \qquad \left.\frac{\partial SSR}{\partial \hat{\beta}_0}\right|_{\hat{\beta}_0^{ols},\, \hat{\beta}_1^{ols}} = -2\sum_{i=1}^{n}(Y_i - \hat{\beta}_0^{ols} - \hat{\beta}_1^{ols} X_i) = 0$$

$$(2) \qquad \left.\frac{\partial SSR}{\partial \hat{\beta}_1}\right|_{\hat{\beta}_0^{ols},\, \hat{\beta}_1^{ols}} = -2\sum_{i=1}^{n}(Y_i - \hat{\beta}_0^{ols} - \hat{\beta}_1^{ols} X_i)X_i = 0$$

Solving gives

$$\hat{\beta}_0^{ols} = \overline{Y} - \hat{\beta}_1^{ols}\overline{X}$$

$$\hat{\beta}_1^{ols} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})X_i}{\sum_{i=1}^{n}(X_i - \overline{X})X_i} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Ordinary Least Squares (Details)

$$(1) \;\Rightarrow\; \sum_{i=1}^{n} Y_i - n\hat{\beta}_0^{ols} - \hat{\beta}_1^{ols}\sum_{i=1}^{n} X_i = 0 \;\Rightarrow\; \overline{Y} - \hat{\beta}_0^{ols} - \hat{\beta}_1^{ols}\overline{X} = 0 \;\Rightarrow\; \hat{\beta}_0^{ols} = \overline{Y} - \hat{\beta}_1^{ols}\overline{X}$$

Substitute $\hat{\beta}_0^{ols}$ into (2), we have

$$\sum_{i=1}^{n}(Y_i - (\overline{Y} - \hat{\beta}_1^{ols}\overline{X}) - \hat{\beta}_1^{ols}X_i)X_i = 0$$

$$\sum_{i=1}^{n}[(Y_i - \overline{Y}) - \hat{\beta}_1^{ols}(X_i - \overline{X})]X_i = 0$$

$$\sum_{i=1}^{n}(Y_i - \overline{Y})X_i - \hat{\beta}_1^{ols}\sum_{i=1}^{n}(X_i - \overline{X})X_i = 0 \;\Rightarrow\; \hat{\beta}_1^{ols} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})X_i}{\sum_{i=1}^{n}(X_i - \overline{X})X_i}$$

Ordinary Least Squares (Details)

Does the second-order condition hold for minimization problem? (i.e., is SSR convex in $\hat{\beta}_0$ and $\hat{\beta}_1$?)

We have $\dfrac{\partial^2 SSR}{\partial \hat{\beta}_0^2} = 2n$, $\dfrac{\partial^2 SSR}{\partial \hat{\beta}_1^2} = 2\sum_{i=1}^n X_i^2$, $\dfrac{\partial^2 SSR}{\partial \hat{\beta}_0\,\partial \hat{\beta}_1} = 2\sum_{i=1}^n X_i$

$$
\begin{aligned}
& v_1^2\left(\frac{\partial^2 SSR}{\partial \hat{\beta}_0^2}\right) + 2v_1v_2\left(\frac{\partial^2 SSR}{\partial \hat{\beta}_0\,\partial \hat{\beta}_1}\right) + v_2^2\left(\frac{\partial^2 SSR}{\partial \hat{\beta}_1^2}\right) \\
&= 2\left(v_1^2 n + 2v_1v_2\sum_{i=1}^n X_i + v_2^2\sum_{i=1}^n X_i^2\right) \\
&= 2n\left(v_1^2 + 2v_1v_2\overline{X} + v_2^2\frac{1}{n}\sum_{i=1}^n X_i^2\right) \\
&= 2n\left[(v_1 + v_2\overline{X})^2 - v_2^2\overline{X}^2 + v_2^2\frac{1}{n}\sum_{i=1}^n X_i^2\right] \\
&= 2n\left[(v_1 + v_2\overline{X})^2 + v_2^2\left(\frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}^2\right)\right] \ > \ 0 \ \text{ for all } \ v_1, v_2 \ \text{ not both zero}
\end{aligned}
$$

Ordinary Least Squares

The estimated model (the "Sample Regression Line") is

$$\hat{Y} = \hat{\beta}_0^{ols} + \hat{\beta}_1^{ols} X$$

- The OLS fitted values are: $\hat{Y}_i^{ols} = \hat{\beta}_0^{ols} + \hat{\beta}_1^{ols} X_i$, $i = 1, \ldots, n$
- The OLS residuals are $\hat{\epsilon}_i^{ols} = Y_i - \hat{Y}_i^{ols} = Y_i - \hat{\beta}_0^{ols} - \hat{\beta}_1^{ols} X_i$, $i = 1, \ldots, n$

(We'll discuss estimator standard errors and other associated statistics in the next class)

For now: Are $\hat{\beta}_0^{ols}$ and $\hat{\beta}_1^{ols}$ good estimators for β_0 and β_1?

Unbiasedness of OLS Estimator

(Focus on β_1) First rewrite $\hat{\beta}_1^{ols}$ as

$$
\begin{aligned}
\hat{\beta}_1^{ols} &= \frac{\sum_{i=1}^n (Y_i - \overline{Y})X_i}{\sum_{i=1}^n (X_i - \overline{X})X_i} = \frac{\sum_{i=1}^n (X_i - \overline{X})Y_i}{\sum_{i=1}^n (X_i - \overline{X})X_i} = \frac{\sum_{i=1}^n (X_i - \overline{X})(\beta_0 + \beta_1 X_i + \epsilon_i)}{\sum_{i=1}^n (X_i - \overline{X})X_i} \\
&= \frac{\beta_0 \sum_{i=1}^n (X_i - \overline{X}) + \beta_1 \sum_{i=1}^n (X_i - \overline{X})X_i + \sum_{i=1}^n (X_i - \overline{X})\epsilon_i)}{\sum_{i=1}^n (X_i - \overline{X})X_i} \\
&= \beta_1 + \frac{\sum_{i=1}^n (X_i - \overline{X})\epsilon_i}{\sum_{i=1}^n (X_i - \overline{X})X_i}
\end{aligned}
$$

$$
\text{Then} \quad E(\hat{\beta}_1^{ols} \mid X_1, \dots, X_n) = \beta_1 + \frac{\sum_{i=1}^n (X_i - \overline{X})E(\epsilon_i \mid X_1, \dots, X_n)}{\sum_{i=1}^n (X_i - \overline{X})X_i} = \beta_1
$$

Unbiasedness

It follows that $E(\hat{\beta}_1^{ols}) = \beta_1$

Intuition: population and sample parallels

$E(\epsilon \mid X) = 0$ implies

- $E(\epsilon) = 0$
- $E(\epsilon X) = 0$

$E(Y \mid X) = \beta_0 + \beta_1 X$ implies

- $\beta_0 = E(Y) - \beta_1 E(X)$
- $\beta_1 = \dfrac{Cov(X,Y)}{Var(X)}$

FOC can be written as

- $\sum_{i=1}^{n} \hat{\epsilon}_i^{ols} = 0$ or $(1/n)\sum_{i=1}^{n} \hat{\epsilon}_i^{ols} = \overline{\hat{\epsilon}^{ols}} = 0$
- $\sum_{i=1}^{n} \hat{\epsilon}_i^{ols} X_i = 0$

OLS estimators are

- $\hat{\beta}_0^{ols} = \overline{Y} - \hat{\beta}_1^{ols}\overline{X}$
- $\hat{\beta}_1^{ols} = \dfrac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \dfrac{smpl.Cov(X_i, Y_i)}{smpl.Var(X_i)}$

Consistency

$\hat{\beta}_1^{ols}$ is also consistent for β_1

Rough argument 1:

$$\hat{\beta}_1^{ols} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Appealing to LLN:

- Numerator in second term converges in probability to population $Cov(X, Y)$
- Denominator in second term converges in probability to population $Var(X)$

$$\hat{\beta}_1^{ols} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2} \xrightarrow{p} \frac{Cov(X, Y)}{Var(X)} = \beta_1$$

Consistency

Rough argument 2:

$$\hat{\beta}_1^{ols} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \beta_1 + \frac{\sum_{i=1}^{n}(X_i - \overline{X})\epsilon_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})\epsilon_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

- Numerator in second term converges in probability to population $Cov(X, \epsilon)$
- Denominator in second term converges in probability to population $Var(X)$

If population $Cov(X, \epsilon) = 0$ and population $Var(X) \neq 0$, then

$$\hat{\beta}_1^{ols} = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})\epsilon_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2} \xrightarrow{p} \beta_1 + \frac{Cov(X, \epsilon)}{Var(X)} = \beta_1$$

Session 2.4

Session 2.4 Causal Interpretations?

- When can you give causal interpretation to β_1?
- Do you only have a predictive relationship?

Causal Interpretations?

If

- $Y = \beta_0 + \beta_1 X + \epsilon$, $E(\epsilon \mid X) = 0$ (or equivalently, $E(Y \mid X) = \beta_0 + \beta_1 X$)
- and you have a representative iid sample from the population
- and you estimate $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ using OLS

Then

- $\hat{\beta}_0^{ols}$ and $\hat{\beta}_1^{ols}$ are unbiased and consistent for β_0 and β_1
- The empirical model $\hat{Y} = \hat{\beta}_0^{ols} + \hat{\beta}_1^{ols} X$ is unbiased and consistent for $E(Y \mid X)$

Question: Can we interpret β_1 as representing the "causal effect" of X on Y? And can we regard $\hat{\beta}_1^{ols}$ as an unbiased and consistent estimate of this causal effect?

Causal Interpretations?

$$E(Y \mid X) = \beta_0 + \beta_1 X \;\Rightarrow\; \beta_1 = \frac{Cov(X,Y)}{Var(X)}$$

- $\hat{\beta}_1^{ols}$ ultimately estimates a correlation
- Predictive relationship
- "Correlation is not Causation"

What do we mean by causal effect?

- Effect of X on Y holding everything else "fixed" (a bit strong!)
- Effect on Y of variation in X that are uncorrelated with other factors affecting Y

Example 1: Omitted Variables

Suppose causal relationship among Y, X and Z are

$$Y = \alpha_0 + \alpha_1 X + \alpha_2 Z + u\,,\ u \sim \text{Normal}(0, \sigma_u^2)$$

$$Z = X + v\,,\ v \sim \text{Normal}(0, \sigma_v^2)$$

where X and the noise terms u and v are independent "exogenous" variables. Then

$$E(Y \mid X) = \alpha_0 + \alpha_1 X + \alpha_2 E(Z \mid X) + E(u \mid X) = \alpha_0 + (\alpha_1 + \alpha_2)X$$

In the regression model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, $\hat{\beta}_1^{ols}$ will be unbiased and consistent for $\alpha_1 + \alpha_2$, not α_1.

- Estimates total effect of X on Y
- α_1 is effect of X on Y "controlling for Z"

Example 1: Omitted Variables

Another way to look at it:

If we write $Y = \alpha_0 + \alpha_1 X + \epsilon$, (i.e., $\epsilon = \alpha_2 Z + u$), then $E(\epsilon \mid X) = 0$ does not hold:

$$E(\epsilon \mid X) = E(\alpha_2 Z + u \mid X) = E(\alpha_2 X + \alpha_2 v + u \mid X) = \alpha_2 X$$

- ϵ and X are correlated because Z is subsumed in ϵ, and Z is correlated with X
- $\hat{\beta}_1^{ols}$ will be biased for α_1
- bias will be in the direction that makes $E(\epsilon \mid X) = 0$
- In the reg. eq. $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, Z_i is an "omitted variable"

Example 2: Omitted Variables

Suppose "true causal relationship" between Y, X and Z is

$$Y = \alpha_0 + \alpha_1 X + \alpha_2 Z + u\,,\; u \sim \text{Normal}(0, \sigma_u^2)$$

$$X = Z + v\,,\; v \sim \text{Normal}(0, \sigma_v^2)$$

Assume Z, u, v all independent exogenous variables, with $Z \sim \text{Normal}(0, \sigma_z^2)$

To take an extreme case, suppose $\alpha_1 = 0$. Then

- X does not "cause" Y
- Z drives both Y and X

Example 2: Omitted Variables

What is $E(Y \mid X)$?

- We have $E(Y \mid X) = \alpha_0 + \alpha_1 X + \alpha_2 E(Z \mid X)$
- It can be shown that $E(Z \mid X) = \dfrac{\sigma_z^2}{\sigma_z^2 + \sigma_u^2} X$
  - $E(Z \mid X)$ is linear, i.e.,
    $E(Z \mid X) = \delta_0 + \delta_1 X$ where $\delta_0 = E(Y) - \delta_1 E(X)$ and $\delta_1 = \frac{Cov(X,Z)}{Var(X)}$
  - $Cov(X, Z) = E(XZ) = E(Z^2) = Var(Z) = \sigma_z^2$
  - $Var(X) = Var(Z) + Var(u) = \sigma_z^2 + \sigma_u^2$

Therefore $E(Y \mid X) = \alpha_0 + \alpha_1 X + \alpha_2 E(Z \mid X) = \alpha_0 + \left(\alpha_1 + \dfrac{\alpha_2 \sigma_z^2}{\sigma_z^2 + \sigma_u^2}\right) X$

Example 2: Omitted Variables

In the regression model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, $\hat{\beta}_1^{ols}$ will be unbiased and consistent for $\left(\alpha_1 + \dfrac{\alpha_2 \sigma_z^2}{\sigma_z^2 + \sigma_u^2}\right)$, not α_1

If we define ϵ so that $Y = \alpha_0 + \alpha_1 X + \epsilon$, then $E(\epsilon \mid X) = 0$ does not hold

- ϵ subsumes Z, which is correlated with X
- ϵ and X are correlated
- $\hat{\beta}_1$ will be biased for α_1 in a direction that makes $E(\epsilon \mid X) = 0$
- Again, in reg. eq. $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, Z_i is an "omitted variable"

Example 3: Simultaneity Bias

Suppose the following describes the market for a good

$$
\begin{aligned}
Q_t^d &= \delta_0 + \delta_1 P_t + \epsilon_t^d && \text{(Demand Eq } \delta_1 < 0\text{)} \\
Q_t^s &= \alpha_0 + \alpha_1 P_t + \epsilon_t^s && \text{(Supply Eq } \alpha_1 > 0\text{)} \\
Q_t^s &= Q_t^d && \text{(Market Clearing)}
\end{aligned}
$$

- Q and P represent log quantities and log prices respectively, so δ_1 and α_1 represent price elasticities of demand and supply respectively
- Suppose the demand shock ϵ_t^d and supply shock ϵ_y^s are iid noise terms with zero means and variances σ_d^2 and σ_s^2 respectively, and are mutually uncorrelated

Example 3: Simultaneity Bias

Market clearing implies

$$\delta_0 + \delta_1 P_t + \epsilon_t^d = \alpha_0 + \alpha_1 P_t + \epsilon_t^s$$

Solving gives

$$P_t = \frac{\alpha_0 - \delta_0}{\delta_1 - \alpha_1} + \frac{\epsilon_t^s - \epsilon_t^d}{\delta_1 - \alpha_1}.$$

Substituting P_t into either the demand or supply equation gives

$$Q_t = \left(\delta_0 + \delta_1 \frac{\alpha_0 - \delta_0}{\delta_1 - \alpha_1}\right) + \frac{\delta_1 \epsilon_t^s - \alpha_1 \epsilon_t^d}{\delta_1 - \alpha_1}.$$

Example 3: Simultaneity Bias

This implies

$$Var(P_t) = \frac{\sigma_s^2 + \sigma_d^2}{(\delta_1 - \alpha_1)^2} \quad \text{and} \quad Cov(P_t, Q_t) = \frac{\delta_1 \sigma_s^2 + \alpha_1 \sigma_d^2}{(\delta_1 - \alpha_1)^2}.$$

Regression of $Q_t = \beta_0 + \beta_1 P_t + \epsilon_t$ gives

$$\hat{\beta}_1 \xrightarrow{p} \frac{Cov(Q_t, P_t)}{Var(P_t)} = \frac{\delta_1 \sigma_s^2 + \alpha_1 \sigma_d^2}{\sigma_s^2 + \sigma_d^2}$$

which is neither the price elasticity of demand nor the price elasticity of supply.

Reason is that the regressor P_t is "endogenous" (so is Q_t)

Session 2.5

Session 2.5 Sampling and Other Issues

- Examples of sampling problems that make sample *not* representative of population
  - Measurement error
  - Truncated samples

Problem for both estimation of predictive and causal relationships

Sampling Problems: Measurement Error

Suppose $Y = \beta_0 + \beta_1 X + \epsilon$ but X is only observed with error,

i.e., you only observe $X^* = X + u$.

Assume measurement error u is independent of X. Then

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X + \epsilon \\
&= \beta_0 + \beta_1 (X^* - u) + \epsilon \\
&= \beta_0 + \beta_1 X^* + (\epsilon - \beta_1 u) \\
&= \beta_0 + \beta_1 X^* + v
\end{aligned}
$$

where $v = \epsilon - \beta_1 u$

Sampling Problems: Measurement Error

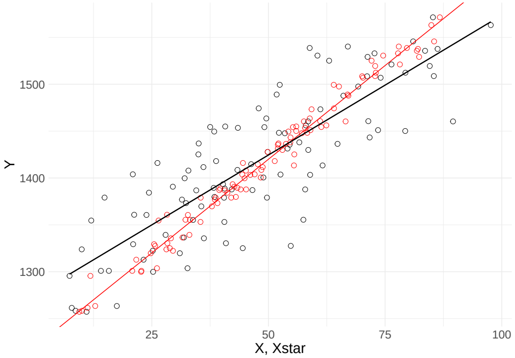
You estimate

$$Y = \beta_0 + \beta_1 X^* + v \quad (1)$$

since only X^* is available to you

However, since u is correlated with X^*

- $v = \epsilon - \beta_1 u$ is correlated with X^*
- $E(v \mid X^*) = 0$ does not hold
- $E(Y \mid X^*) \neq \beta_0 + \beta_1 X^*$
- Coefficient on X^* in $E(Y \mid X^*)$ is not β_1
- $\hat{\beta}_1^{ols}$ from (1) is biased/inconsistent

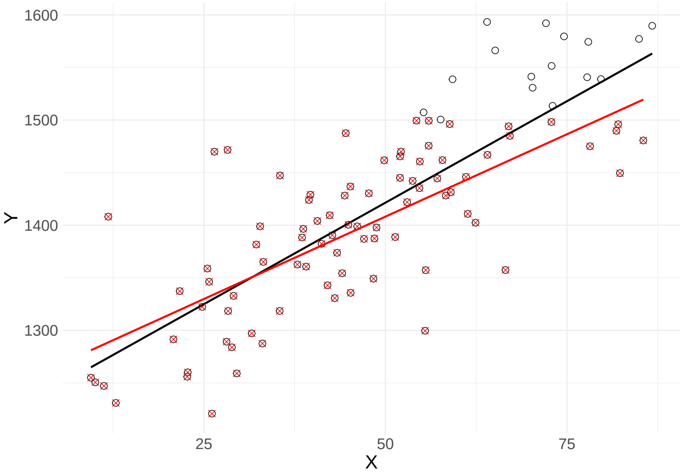


Sampling Problems: Truncated Sampling

Suppose $E(Y \mid X) = \beta_0 + \beta_1 X$ with $\beta_1 > 0$

- But suppose you do **not** have a random sample $\{X_i, Y_i\}_{i=1}^n$
- In particular, suppose you have a "truncated sample" where you cannot observe any observation where $Y_i > c$.
- This induces correlation between ϵ_i and X_i
  - large X_i together with large positive ϵ_i makes $Y_i > c$ more likely
  - large X_i that are observed are those with lower or negative values of ϵ_i
  - implies a negative correlation between X_i and ϵ_i

Sampling Problems: Truncated Sampling



Specification Issues

What if $E(Y \mid X) \neq \beta_0 + \beta_1 X$?

e.g. what if $E(Y \mid X) = \beta_0 + \beta_1 X + \beta_2 X^2$

- $\hat{Y} = \hat{\beta}_0^{ols} + \hat{\beta}_1^{ols} X$ will be a biased estimate of $E(Y \mid X)$
- If $E(Y \mid X) \approx \beta_0 + \beta_1 X$, then $\hat{Y} = \hat{\beta}_0^{ols} + \hat{\beta}_1^{ols} X$ will (hopefully) only be a slightly biased estimate of $E(Y \mid X)$
- "Local approaches" may give unbiased estimates, but may have large standard errors
- The simple linear regression model, though slightly biased, may have smaller standard errors

Roadmap

- (Previous) Session 1: Statistics Review
- **This Session 2: Simple Linear Regression**
- *Next Session 3: Estimator Standard Errors; Multiple Linear Regression*
- Session 4: Matrix Algebra
- Session 5: OLS using Matrix Algebra
- Session 6: Hypothesis Testing
- Session 7: Prediction
- Session 8: Instrumental Variable Regression
- Session 9: Logistic and Other Regressions
- Session 10: Panel Data Regressions
- Session 11: Introduction to Time Series
- Session 12: Time Series Regressions