Session 2

Conditional Expectations / Simple Linear Regression

Anthony Tay

This Version: 13 Aug 2025

Session 2

- Conditional Expectations
- **Ordinary Least Squares (OLS)** estimation of the **Simple Linear Regression (SLR)** model
- Specific example:
  - Estimate relationship between hourly earnings and years of schooling
  - Population of interest: US non-institutional employed civilians aged 16 & above in 2018

Session 2

- Model the population using a joint probability distribution function $f_{X,Y}(x, y)$ such that
  - X, Y discrete: $\Pr(X = x, Y = y) = f_{X,Y}(x, y)$
  - X, Y continuous: $\Pr(a \leq X \leq b\,,\, c \leq Y \leq d) = \int_a^b \int_c^d f_{X,Y}(x, y)\, dy\, dx$
- Obtain a random sample of $\{X_i, Y_i\}_{i=1}^n$ from this population
- Estimate population relationship between $earn$ and $educ$ using linear regression model

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

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

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

$\rightarrow$

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

$\rightarrow$ $E(Y) = 4.5$, $Var(Y) = 0.625$

$X \downarrow$

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

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

Covariance and Correlation

One way to describe relationship between X and Y is to use correlation coefficient

$$\sigma_{X,Y} = Cov(X,Y) = E((X - E(X))(Y - E(Y)))$$

In our example, we have

$$\begin{aligned} Cov(X,Y) = &(5-3)(6.0-4.5)\tfrac{1}{20} + \\ &(4-3)(5.5-4.5)\tfrac{1}{20} + (5-3)(5.5-4.5)\tfrac{2}{20} + \\ &(3-3)(5.0-4.5)\tfrac{1}{20} + (4-3)(5.0-4.5)\tfrac{2}{20} + (5-3)(5.0-4.5)\tfrac{1}{20} + \\ &(2-3)(4.5-4.5)\tfrac{1}{20} + (3-3)(4.5-4.5)\tfrac{2}{20} + (4-3)(4.5-4.5)\tfrac{1}{20} + \\ &(1-3)(4.0-4.5)\tfrac{1}{20} + (2-3)(4.0-4.5)\tfrac{2}{20} + (3-3)(4.0-4.5)\tfrac{1}{20} + \\ &(1-3)(3.5-4.5)\tfrac{2}{20} + (2-3)(3.5-4.5)\tfrac{1}{20} + \\ &(1-3)(3.0-4.5)\tfrac{1}{20} \\ = &1 \end{aligned}$$

Covariance and Correlation

Covariance is not invariant to scale

If X is currently measured in thousands of dollars, and re-scaled to dollars (multiply X by 1000, then the covariance becomes

$$\begin{aligned} Cov(1000X, Y) &= E((Y - E(Y))(1000X - E(1000X))) \\ &= 1000E((Y - E(Y))(X - E(X))) \\ &= 1000\,Cov(X, Y). \end{aligned}$$

For this reason, the correlation coefficient

$$\rho_{X,Y} = Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}},$$

which is invariant to scale and always lies between -1 and 1, is more informative

Covariance and Correlation

Given a sample $\{X_i, Y_i\}_{i=1}^n$ from a joint pdf $f_{X,Y}(x, y)$, we can estimate the covariance using the **sample covariance**

$$\hat{\sigma}_{X,Y} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$$

To estimate the correlation coefficient, we can divide the sample covariance by the sample standard deviations, i.e.,

$$\hat{\rho}_{X,Y} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}}$$

Covariance and Correlation

The following properties of means, variances and covariances are easy to show: if a and b are constants, we have

- $E(aX + bY) = aE(X) + bE(Y)$
- $Var(aX + bY) = a^2\,Var(X) + b^2\,Var(Y) + 2ab\,Cov(X, Y)$
- $Cov(X, Y) = E(XY) - E(X)E(Y)$
- $Cov(X, X) = Var(X)$

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

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

$\rightarrow$

X	1	2	3	4	5
$E(Y \mid X)$	3.5	4	4.5	5	5.5
$Var(Y \mid X)$	0.125	0.125	0.125	0.125	0.125

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

$$E(Y) = c \quad \text{and} \quad Cov(X, Y) = 0$$

Proof:

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

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

Law of Iterated Expectations

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

$$\beta_0 = E(Y) - \beta_1 E(X) \quad \text{and} \quad \beta_1 = \frac{Cov(X,Y)}{Var(X)}$$

Proof:

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

$$E(YX) = E(E(YX \mid X)) = E(XE(Y \mid X)) = E(X(\beta_0 + \beta_1 X)) = \beta_0 E(X) + \beta_1 E(X^2)$$

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

	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:

	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?

Simple Linear Regression

Suppose we want to estimate $E(Y \mid X)$ for a certain population

If

- $E(Y \mid X) = \beta_0 + \beta_1 X$
- and we define $\epsilon = Y - \beta_0 - \beta_1 X$

we can write

$$Y = \beta_0 + \beta_1 X + \epsilon\,,\ E(\epsilon \mid X) = 0$$

since

$$E(\epsilon \mid X) = E(Y - \beta_0 - \beta_1 X \mid X) = E(Y \mid X) - \beta_0 - \beta_1 X = 0$$

Simple Linear Regression

Note that

$$E(\epsilon \mid X) = 0 \implies E(\epsilon) = 0 \quad \text{and} \quad E(\epsilon X) = 0$$

Furthermore

$$E(Y \mid X) = \beta_0 + \beta_1 X \implies \beta_1 = \frac{Cov(X, Y)}{Var(X)} \quad \text{and} \quad \beta_0 = E(Y) - \beta_1 E(X)$$

We want to estimate β_0 and β_1 using a sample $\{Y_i, X_i\}_{i=1}^n$ drawn from the population

Simple Linear Regression

If sample is representative of the population, we can write

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i\,,\ E(\epsilon_i \mid X_i) = 0 \ \text{ for } \ i = 1, \ldots, n\,.$$

If sample is iid, we can extend this to

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i \\ E(\epsilon_i \mid X_1, \ldots, X_n) &= 0 \\ E(\epsilon_i \epsilon_j \mid X_1, \ldots, X_n) &= 0 \ \text{ for } \ i, j = 1, \ldots, n\,,\ i \neq j\,. \end{aligned}$$

This is our **simple linear regression** model

Simple Linear Regression

- $Y_i \sim$ "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 intercept term or "constant" term

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

Simple Linear Regression

Let $\hat{\beta}_0$ and $\hat{\beta}_1$ denote estimators for β_0 and β_1

Define

- Fitted values: $\widehat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
- Residuals: $\hat{\epsilon}_i = Y_i - \widehat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$

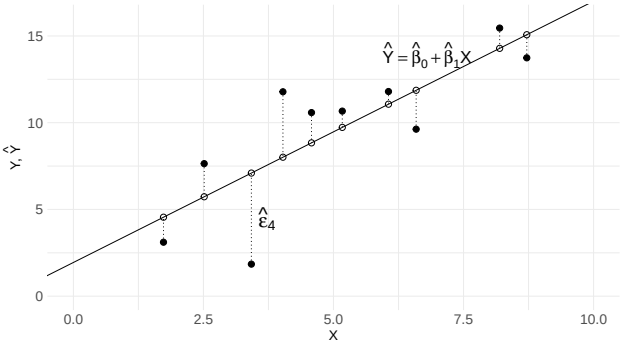
for all $i = 1, \ldots, n$

Of course, we have

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\epsilon}_i = \widehat{Y}_i + \hat{\epsilon}_i$$

.

The Simple Linear Regression Model



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
```

Black dots (X_i, Y_i), Hollow dots $(X_i, \widehat{Y_i})$, Black line: estimated regression line

A Method of Moments Approach

Since (given our assumptions) X, Y and ϵ satisfy

$$E(\epsilon) = 0 \quad \text{and} \quad E(\epsilon X) = 0$$

the method of moments approach suggests to choose β_0^{mm} and β_1^{mm} to satisfy

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^{mm} &= 0 \\ \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^{mm} X_i &= 0 \end{aligned} \quad \Longrightarrow \quad \begin{aligned} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0^{mm} - \hat{\beta}_1^{mm} X_i) &= 0 \\ \sum_{i=1}^{n} (Y_i - \hat{\beta}_0^{mm} - \hat{\beta}_1^{mm} X_i) X_i &= 0 \end{aligned}$$

where $\hat{\epsilon}_i^{mm} = (Y_i - \hat{\beta}_0^{mm} - \hat{\beta}_1^{mm} X_i)$

Least Squares Approach

Choose $\hat{\beta}_0$ and $\hat{\beta}_1$ so minimize residual sum of squares

$$RSS = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

i.e., choose

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

This is the **ordinary least squares (OLS)** approach

Least Squares Approach

Elementary optimization theory says that $\hat{\beta}_0^{ols}, \hat{\beta}_1^{ols}$ must satisfy the necessary first-order conditions

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

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

Notice that the OLS conditions are the same as the method of moments conditions, so both approaches lead to the same estimator (for now we refer to this as the OLS estimator)

Least Absolute Deviation Approach

Yet another approach is the **least absolute deviation (LAD)** approach:

$$\widehat{\beta}_0^{lad}, \widehat{\beta}_1^{lad} = \underset{\widehat{\beta}_0\,,\widehat{\beta}_1}{\arg\min} \sum_{i=1}^{n} \mid Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i \mid .$$

This approach leads to a different set of estimators

Left to more advanced courses

Ordinary Least Squares (Details)

Solving OLS conditions gives

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

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

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

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

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

Can show that the second order condition also holds (omitted)

Ordinary Least Squares

Other ways of writing $\hat{\beta}_1^{ols}$

$$\begin{aligned}
\hat{\beta}_1^{ols} &= \frac{\sum_{i=1}^n (Y_i - \overline{Y})X_i}{\sum_{i=1}^n (X_i - \overline{X})X_i} \\
&= \frac{\sum_{i=1}^n (X_i - \overline{X})Y_i}{\sum_{i=1}^n (X_i - \overline{X})X_i} \\
&= \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \frac{\frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2} = \frac{\textit{Sample Cov}(X_i, Y_i)}{\textit{Sample Var}(X_i)}
\end{aligned}$$

There are other ways which we will come to shortly

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 $\hat{\epsilon}_i^{ols} = Y_i - \hat{Y}_i^{ols} = Y_i - \hat{\beta}_0^{ols} - \hat{\beta}_1^{ols} X_i,\ i = 1, \ldots, n$

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

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

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

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

- $\hat{\beta}_0^{ols} = \overline{Y} - \hat{\beta}_1^{ols}\overline{X}$
- $\hat{\beta}_1^{ols} = \frac{\textit{Sample Cov}(X_i, Y_i)}{\textit{Sample Var}(X_i)}$

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

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

OLS Standard Errors

Standard errors should be calculated for all estimators

For $\hat{\beta}_1^{ols}$, we have

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

$$Var(\hat{\beta}_1^{ols} \mid X_1, \dots, X_n) = \frac{\sum_{i=1}^n (X_i - \overline{X})^2 \, Var(\epsilon_i \mid X_1, \dots, X_n)}{(\sum_{i=1}^n (X_i - \overline{X})^2)^2}$$

OLS Standard Errors

If we are willing to assume that

$$Var(\epsilon_i \mid X_1, \dots, X_n) = \sigma^2$$

then $Var(\hat{\beta}_1^{ols} \mid X_1, \dots, X_n)$ simplifies to

$$Var(\hat{\beta}_1^{ols} \mid X_1, \dots, X_n) = \frac{\sum_{i=1}^n (X_i - \overline{X})^2 \overbrace{Var(\epsilon_i \mid X_1, \dots, X_n)}^{\sigma^2}}{(\sum_{i=1}^n (X_i - \overline{X})^2)^2} = \frac{\sigma^2 \sum_{i=1}^n (X_i - \overline{X})^2}{(\sum_{i=1}^n (X_i - \overline{X})^2)^2}$$

$$= \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2}$$

OLS Standard Errors

$\hat{\beta}_1$ has a smaller variance if

- σ^2 is small (the data is less noisy)
- n is larger (since the denominator is a sum of n non-negative terms) and
- if there is more variation in your X_i sampl

To get a numerical estimate for the variance, we have to get an estimate for σ^2

OLS Standard Errors

The assumption

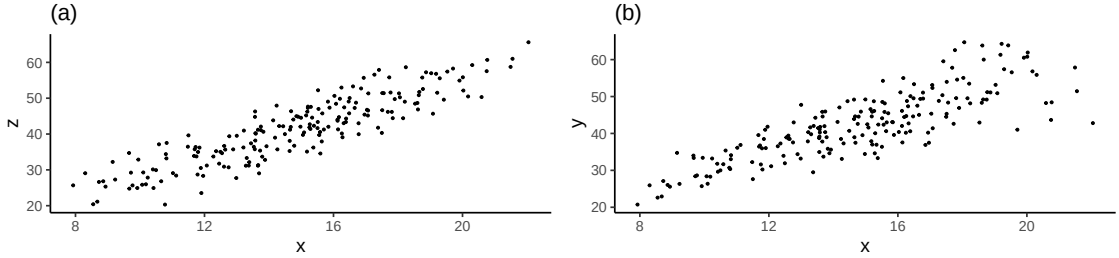
$$Var(\epsilon_i \mid X_1, \ldots, X_n) = \sigma^2$$

is called **homoskedasticity** (otherwise **heteroskedasticity**)

- It will hold if $Var(\epsilon \mid X)$ is constant in population and you have a representative iid sample from the population
- "noisiness" of the data does not depend on any of the X observations

OLS Standard Errors

Figure shows data from the data set `heterosk.csv`. In (a), noise terms appear homoskedastic whereas in (b) the noise terms appear to be heteroskedastic (variance appears to be increasing in X)



OLS Standard Errors

$$Var(\hat{\beta}_1^{ols} \mid X_1, \ldots, X_n) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2}$$

is valid only in the homoskedasticity case

In that case, it can be shown that an unbiased estimator for σ^2 is

$$\widehat{\sigma^2} = \frac{1}{n-2} \sum_{i=1}^n \hat{\epsilon}_{i,ols}^2$$

We define the standard error of $\hat{\beta}_1$ as

$$\text{s.e.}(\hat{\beta}_1) = \sqrt{\frac{\widehat{\sigma^2}}{\sum_{i=1}^n (X_i - \overline{X})^2}}$$

OLS Standard Errors

Furthermore, if $\epsilon_i \sim \text{Normal}(0, \sigma^2)$, we have

$$\text{t-stat} = \frac{\hat{\beta}_1^{ols} - \beta_1}{\text{s.e.}(\hat{\beta}_1)} \quad \sim \quad \text{t}(n-2)$$

If not, we can appeal to the asymptotic result

$$\text{t-stat} = \frac{\hat{\beta}_1^{ols} - \beta_1}{\text{s.e.}(\hat{\beta}_1)} \quad \sim \quad \text{Normal}(0, 1)$$

The t-stat can be used to test hypotheses on β_1

Example: Returns to Schooling

Population of interest: all US non-institutional working civilians aged 16 or over in 2018

You have a representative random (iid) sample from this population, stored in the file `earnings2019.csv`

Your sample size is $n = 4946$

You wish to learn about relationship between earnings and years of schooling

In particular, you want to estimate

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

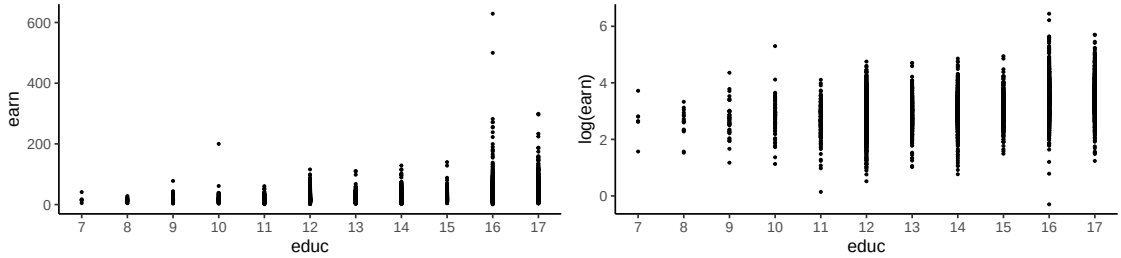
Example: Returns to Schooling

```
options(width=120)
dat1 <- read_csv("data\\earnings2019.csv", show_col_types=FALSE)
dat1 <- dat1 %>%
  mutate(                                  # race variable is
    race_white = if_else(race=="White", 1, 0),   # "White", "Black", "Other"
    race_black = if_else(race=="Black", 1, 0),   # Convert to three dummy var,
    race_other = if_else(race=="Other", 1, 0)    # one for each race
  ) %>%                                    # then
  select(-race)                            # remove race variable
dat1 %>% summary()                         # and produce summary
```

```
      age            height          educ            feduc           meduc           tenure          wexp      
 Min.   :19.00   Min.   :40.00   Min.   : 7.00   Min.   : 0.000   Min.   : 0.000   Min.   : 1.000   Min.   : 1.000  
 1st Qu.:33.00   1st Qu.:64.00   1st Qu.:12.00   1st Qu.: 4.000   1st Qu.: 4.000   1st Qu.: 3.000   1st Qu.: 3.000  
 Median :40.00   Median :67.00   Median :14.00   Median : 4.000   Median : 4.000   Median : 6.000   Median : 7.000  
 Mean   :41.99   Mean   :67.45   Mean   :14.31   Mean   : 5.425   Mean   : 5.523   Mean   : 9.177   Mean   : 9.251  
 3rd Qu.:51.00   3rd Qu.:70.00   3rd Qu.:16.00   3rd Qu.: 7.000   3rd Qu.: 7.000   3rd Qu.:13.000   3rd Qu.:13.000  
 Max.   :82.00   Max.   :83.00   Max.   :17.00   Max.   :26.000   Max.   :26.000   Max.   :54.000   Max.   :51.000  
      male            earn            totalwork      race_white      race_black      race_other    
 Min.   :0.0000   Min.   :  0.7428   Min.   :1000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.: 15.5048   1st Qu.:1936   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000  
 Median :0.0000   Median : 22.9995   Median :2080   Median :1.0000   Median :0.000   Median :0.0000  
 Mean   :0.4646   Mean   : 29.2315   Mean   :2182   Mean   :0.5623   Mean   :0.311   Mean   :0.1268  
 3rd Qu.:1.0000   3rd Qu.: 35.0235   3rd Qu.:2428   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:0.0000  
 Max.   :1.0000   Max.   :628.9308   Max.   :5824   Max.   :1.0000   Max.   :1.000   Max.   :1.0000  
```

Example: Returns to Schooling

```
p1 <- ggplot(dat1, aes(y=earn, x=educ)) + geom_point(size=0.5) +
  scale_x_continuous(breaks=7:17) + theme_classic()
p2 <- ggplot(dat1, aes(y=log(earn), x=educ)) + geom_point(size=0.5) +
  scale_x_continuous(breaks=7:17) + theme_classic()
p1 | p2
```



Example: Returns to Schooling

Why did we assume

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

instead of

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

- log-linear form seems to better describe the data than the fully linear specification
- fully linear form implies that expected earnings go up by β_1 dollars for each additional year of `educ`, regardless of level of schooling, i.e., increasing schooling from 7 to 8 leads to the same dollar increase in earnings as increasing schooling from 15 to 16 years (seems unrealistic)
- log-linear form says that expected earnings go up by $100\beta_1$ percent for each additional year of `educ` (seems more realistic)

Example: Returns to Schooling

```
slr <- function(y, x){
  n <- length(y)
  ybar <- mean(y)
  xbar <- mean(x)
  beta1hat <- sum((x-xbar)*y)/sum((x-xbar)*x)
  beta0hat <- ybar - beta1hat*xbar
  yhat <- beta0hat + beta1hat*x
  ehat <- y - yhat
  rss <- sum(ehat^2)
  beta1hat_se <- sqrt((rss/(n-2))/sum((x-xbar)^2))
  beta1hat_t <- beta1hat / beta1hat_se
  cat("beta0hat:", round(beta0hat,3), "\n")
  cat("beta1hat:", round(beta1hat,3),
      "  s.e.:", round(beta1hat_se,3),
      "  t-stat:", round(beta1hat_t,3),
      "  p-val:", round(2*pt(-abs(beta1hat_t), n-2),3))
  results <- list(beta0hat=beta0hat, beta1hat=beta1hat,
                  beta1hat_se=beta1hat_se, ehat=ehat, yhat=yhat)
}
```

Example: Returns to Schooling

```
cat("Regression of ln earn on educ, with intercept\n\n")
mdl1 <- slr(y=log(dat1$earn), x=dat1$educ)
```

```
Regression of ln earn on educ, with intercept

beta0hat: 1.32
beta1hat: 0.128   s.e.: 0.004   t-stat: 32.17   p-val: 0
```

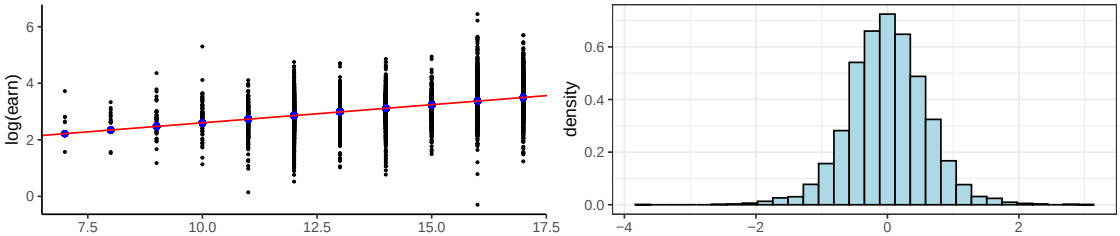
Or use `lm()` function

```
mdl1a <- lm(log(earn)~educ, dat1)
cat("Dependent variable: ", as.character(formula(mdl1a))[2], "\n")
mdl1a %>% summary() %>% coef()
cat("n =", nobs(mdl1a))
```

```
Dependent variable:  log(earn)
             Estimate  Std. Error  t value      Pr(>|t|)
(Intercept) 1.3199894 0.057541299 22.93986 9.401431e-111
educ        0.1279965 0.003978753 32.17000 2.442343e-206
n = 4946
```

Example: Returns to Schooling

```
regdat <- data.frame(educ=dat1$educ, earn=dat1$earn, fitted=mdl1$yhat, resid=mdl1$ehat)
p1 <- ggplot(data=regdat) +
        geom_point(aes(y=log(earn), x=educ), size=0.5) +
        geom_point(aes(y=fitted, x=educ), size=1.5, color='blue') +
        geom_abline(intercept=mdl1$beta0hat, slope=mdl1$beta1hat, color='red') +
        theme_classic()
p2 <- ggplot(regdat, aes(x = resid)) +
  geom_histogram(aes(y = after_stat(density)), fill = "lightblue", color="black", bins=30) +
  xlab("residuals") + theme_bw()
p1 | p2
```



Heteroskedasticity-Robust Standard Errors

What if you are unsure if your errors are homoskedastic?

Recall

$$Var(\hat{\beta}_1^{ols} \mid X_1, \ldots, X_n) = \frac{\sum_{i=1}^n (X_i - \overline{X})^2 \, Var(\epsilon_i \mid X_1, \ldots, X_n)}{(\sum_{i=1}^n (X_i - \overline{X})^2)^2}$$

Cannot simplify under general heteroskedasticity

It can be shown that under quite general conditions that

$$\widehat{Var}_{HC}(\hat{\beta}_1^{ols}) = \frac{\sum_{i=1}^n (X_i - \overline{X})^2 \, \hat{\epsilon}_{i,ols}^2}{\left(\sum_{i=1}^n (X_i - \overline{X})^2\right)^2} \xrightarrow{p} Var(\hat{\beta}_1^{ols})$$

Square root of above is **Heteroskedasticity-robust** / **heteroskedasticy-consistent** / **White standard errors**

Heteroskedasticity-Robust Standard Errors

```
slr_hc0 <- function(y, x){
  n <- length(y)
  ybar <- mean(y)
  xbar <- mean(x)
  beta1hat <- sum((x-xbar)*y)/sum((x-xbar)*x)
  beta0hat <- ybar - beta1hat*xbar
  yhat <- beta0hat + beta1hat*x
  ehat <- y - yhat
  hc0 <- sum((x-xbar)^2*ehat^2)/sum((x-xbar)^2)^2
  beta1hat_se <- sqrt(hc0)
  beta1hat_t <- beta1hat / beta1hat_se
  cat("beta0hat:", round(beta0hat,3), "\n")
  cat("beta1hat:", round(beta1hat,3),
      "  s.e. (het. robust):", round(beta1hat_se,4),
      "  t-stat:", round(beta1hat_t,3),
      "  p-val:", round(2*pt(-abs(beta1hat_t), n-2),4))
  results <- list(beta0hat=beta0hat, beta1hat=beta1hat,
                  beta1hat_se=beta1hat_se, ehat=ehat, yhat=yhat)
}
```

Heteroskedasticity-Robust Standard Errors

```
mdl2 <- slr_hc0(y=log(dat1$earn), x=dat1$educ)
```

```
beta0hat: 1.32
beta1hat: 0.128   s.e. (het. robust): 0.0041   t-stat: 31.584   p-val: 0
```

Alternatively, use `coeftest()` from `lmtest` package with `vcovHC()` from `sandwich` package

```
library(lmtest)     # lmtest::coeftest to calculate t-test
library(sandwich)   # using robust s.e. calculated with sandwich::vcovHC
coeftest(mdl1a, vcov=vcovHC, type="HC0")  # Robust s.e., mdl1a estimated earlier using lm()
```

```
t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) 1.3199894  0.0581594  22.696 < 2.2e-16 ***
educ        0.1279965  0.0040525  31.584 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Prediction

Can use estimated model for predictive purposes

$$\widehat{\ln earn} = \underset{(0.0582)}{1.32} + \underset{(0.0041)}{0.128}\, educ$$

- Imagine going back to the population and sampling one more individual with, say, 12 years of `educ`
- What is the model's prediction for this individual's `ln(earn)`?

```
cat("Pred. ln avg. hr. earn for individual with 12 years educ is ")
cat(round(mdl1$beta0hat + mdl1$beta1hat*12,2), "\"log-dollars\"")
```

```
Pred. ln avg. hr. earn for individual with 12 years educ is 2.86 "log-dollars"
```

Prediction

- Scatterplot of $\ln earn$ vs $educ$ suggests potential pred. error will be quite high
- Many factors that affect earnings left out
- Since parameter estimates are unbiased, prediction is also unbiased, but
- *prediction std.err.* (we consider this later in the course) expected to be quite high
- Model gives predictions for `ln earn`. To convert to `earn` you can take exponentials to get "conditional median prediction" rather than "conditional expectation prediction".

```
cat("Pred. ave. hr. earn for individual with 12 years educ is $")
cat(round(exp(mdl1$beta0hat + mdl1$beta1hat*12),2), "(cond. median pred.).")
```

```
Pred. ave. hr. earn for individual with 12 years educ is $17.39 (cond. median pred.).
```

Measuring direct causal effects?

Estimated model predicts that individual with one more year in `educ` will have 12.8% higher average hourly earnings

- Can we take this as additional year of schooling directly "causing" average hourly earnings to go up by 12.8%?
- The problem is that there are many factors that affect earnings, and your regression is only taking into account only the effect of `educ`
- All of the other factors' influence on `ln earn` may end up being attributed to `educ`

Measuring direct causal effects?

Suppose there are two factors X and Z that affect Y

Suppose conditional expectation of Y given X, and of Y given X and Z are

$$E(Y \mid X) = \alpha_0 + \alpha_1 X$$
$$E(Y \mid X, Z) = \beta_0 + \beta_1 X + \beta_2 Z\,.$$

To capture the direct effect of X on Y, then you would want to estimate β_1, not α_1

β_1 tells you how expected value of Y differs with X for two individuals with the same Z:

$$E(Y|X = x_0 + 1, Z = z_0) - E(Y|X = x_0, Z = z_0) = \beta_1$$

We say that β_1 measures the effect of X on Y **controlling** for Z

Measuring direct causal effects?

Simple linear regression of Y on X gives you an unbiased estimator for α_1, not β_1

If $\beta_2 \neq 0$ and $Cov(X, Z) \neq 0$, then $\alpha_1 \neq \beta_1$

$$\begin{aligned} E_{Y|X}(Y \mid X) &= E_{Z|X}(E_{Y|X,Z}(Y \mid X, Z)) \\ &= E_{Z|X}(\beta_0 + \beta_1 X + \beta_2 Z) \\ &= \beta_0 + \beta_1 X + \beta_2 E_{Z|X}(Z \mid X) \end{aligned}$$

Suppose that $E_{Z|X}(Z \mid X) = \delta_0 + \delta_1 X$. Then

$$\begin{aligned} E(Y \mid X) &= \beta_0 + \beta_1 X + \beta_2(\delta_0 + \delta_1 X) \\ &= (\beta_0 + \beta_2\delta_0) + (\beta_1 + \beta_2\delta_1)X \end{aligned}$$

Measuring direct causal effects?

We know that if $E_{Z|X}(Z \mid X) = \delta_0 + \delta_1 X$, then

$$\delta_0 = E(Z) - \delta_1 E(X) \quad \text{and} \quad \delta_1 = \frac{Cov(X, Z)}{Var(X)}$$

Therefore

$$E(Y \mid X) = (\beta_0 + \beta_2 \delta_0) + \left(\beta_1 + \beta_2 \frac{Cov(X, Z)}{Var(Z)}\right) X$$

Comparing with $E(Y \mid X) = \alpha_0 + \alpha_1 X$, we see that

$$\alpha_1 = \beta_1 + \beta_2 \frac{Cov(X, Z)}{Var(Z)}$$

so $\alpha_1 \neq \beta_1$ unless $\beta_2 = 0$ or $Cov(X, Z) = 0$. A simple linear regression of Y on X gives you an unbiased estimator for α_1, but a biased estimator for β_1

Measuring direct causal effects?

Here is another way of looking at this issue. Suppose we have

$$E(Y \mid X, Z) = \beta_0 + \beta_1 X + \beta_2 Z$$

Define $u = Y - \beta_0 - \beta_1 X - \beta_2 Z$, we can write

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + u\,,\ E(u \mid X, Z) = 0$$

Suppose X and Z are correlated. Then if we write

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where the β_0 and β_1 here are the same as the β_0 and β_1 in equation with Z, then

$$\epsilon = \beta_2 Z + u$$

Measuring direct causal effects?

So we have

$$Y = \beta_0 + \beta_1 X + \epsilon\,,\ \epsilon = \beta_2 Z + u$$

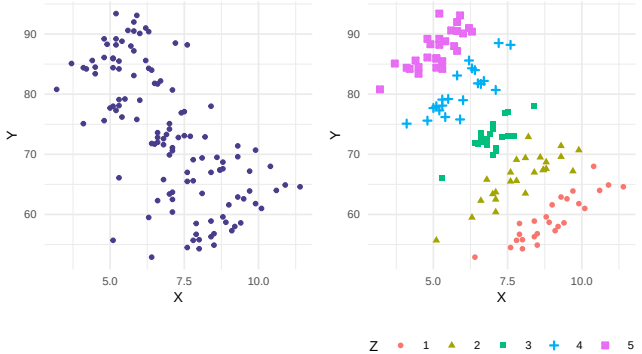
If X and Z are correlated, then

- ϵ and X must be correlated
- we no longer have $E(\epsilon \mid X) = 0$.

This causes the OLS estimator for β_1 in the simple linear regression of Y on X to be biased for β_1

Measuring direct causal effects?

Example: `multireg_eg.csv` contains observations on variables X, Y and Z. Variable Z takes on integer values 1 to 5 only. Figure shows (a) a scatterplot of Y against X, (b) same scatterplot, but with value of Z corresponding to each (X,Y) pair indicated.



Measuring direct causal effects?

Panel (a): strong negative relationship between X and Y

Panel (b): for any fixed level of Z, the effect of X on Y is actually **positive**

Imagine, for instance, that

$Y \sim$ final exam scores in a certain course

$Z \sim$ background preparedness of students ranging from very poor (1) to excellent (5)

$X \sim$ study hours per week

Measuring direct causal effects?

Given any level of preparedness, studying clearly test scores

Background preparedness is also positively correlated with final exam scores

But students who have strong background preparedness spend less time studying for the course

If we look at test scores vs study time only, it will appear that studying reduces test scores.

Measuring direct causal effects?

In this example, clearly α_1 is negative in

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

and a simple linear regression of Y on X produces

```
lm(Y~X, data=df) %>% summary() %>% coef()
```

```
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 102.86406  3.2199095 31.94626 6.373574e-60
X            -4.23213  0.4474755 -9.45779 3.857133e-16
```

Measuring direct causal effects?

The separate effects of X and Z on Y are captured in the conditional expectation

$$E(Y \mid X, Z) = \beta_0 + \beta_1 X + \beta_2 Z \,.$$

Next session, we see that a *multiple linear regression* of Y on X and Z manages to disentangle the effects

```
lm(Y~X+Z, data=df) %>% summary() %>% coef()
```

```
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 21.188337  2.0816539 10.17861 8.212834e-18
X            3.114465  0.2054364 15.16024 2.219609e-29
Z           10.109717  0.2379845 42.48057 5.955647e-73
```