2. Treatment effects

Econometrics II
Winter 2020
Osaka U

Shuhei Kitamura

Summary of the last lecture

Correlation doesn't imply causation

Endogeneity problem

- Reverse causality/simultaneity
- Omitted-variable bias
- Measurement error
- Selection bias

Outline

Treatment effect

Average treatment effect (ATE)

- Independence assumption
- Average Effect of Treatment on the Treated (ATT)

Is the effect statistically significant?

- Statistical significance
- Confidence interval

Is the effect large?

- Effect size

Conducting an experiment

Power calculation, MDE

Regression analysis

- Conditional expectation function
- Conditional independence assumption (CIA)
- Regression and matching
- Omitted-variable bias
- Bad control
- Measurement error

Treatment effect

Let Y_i be an outcome and $Y_i(\hat{D}_i)$ be a potential outcome for person i.

$Y_i(1)$ is the potential outcome when i is treated, while $Y_i(0)$ is the potential outcome when i is not treated.

- Unit i can be anything such as person, country, village, cell, etc.

E.g., let Y_i be the probability of malaria infection for individual i.

	$Y_i(0)$ If person doesn't take a malaria pill	$Y_i(1)$ If person takes a malaria pill	τ_i Treatment effect
Person A	0.8	0.7	0.1
Person B	0.5	0.5	0

Treatment effect (cont.)

A treatment effect or a causal effect for unit i is defined by

$$\tau_i = Y_i(1) - Y_i(0). \tag{1}$$

However...

> The foundamental problem of causal inference is that since it is impossible to observe both outcomes for the same unit (in this example, person) i at any given time, we cannot directly measure the causal effect of D on Y for unit i.

What can we do?

How about...

How about comparing a person who indeed takes a malaria pill and a person doesn't?

Suppose Person A takes a pill and Person B doesn't

$$Y_A - Y_B = 0.2. \tag{2}$$

Can we say that the causal effect is 0.2?

How about...

Rewrite the equation

$$\begin{aligned} Y_A - Y_B &= Y_A(1) - Y_B(0) && (3) \\ &= \{Y_A(1) - Y_A(0)\} + \{Y_A(0) - Y_B(0)\}. && (4) \end{aligned}$$

The first comparison is the causal effect of a malaria pill for Person A. The second comparison captures Person A's susceptibility to malaria relative to Person B's.

- We know that the second term is not zero from the above table.

The second term is called selection bias.

In other words, unless selection bias is zero, we cannot get a causal effect by just comparing the outcomes of two individuals!

What should we do?

Average treatment effect

Idea: If we compare groups rather than individuals, we can perhaps alleviate selection bias.

How?

Define the Average Treatment Effect (ATE) by

$$
\begin{aligned}
Avg_n[Y_i(1) - Y_i(0)] &:= \frac{1}{n}\sum_{i=1}^{n}[Y_i(1) - Y_i(0)] && (5)\\
&= \frac{1}{n}\sum_{i=1}^{n} Y_i(1) - \frac{1}{n}\sum_{i=1}^{n} Y_i(0) && (6)\\
&= Avg_n[Y_i(1)] - Avg_n[Y_i(0)]. && (7)
\end{aligned}
$$

Average treatment effect (cont.)

Next, define *actual* status using a dummy variable

$$D_i = \begin{cases} 1 & \text{if } i \text{ takes a malaria pill} \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

Define $Avg_n[Y_i|D_i = 1]$ be the average of outcome Y_i among those who take a malaria pill, and $Avg_n[Y_i|D_i = 0]$ be the average of outcome Y_i among those who don't.

Difference in group average is written by

$$Avg_n[Y_i|D_i = 1] - Avg_n[Y_i|D_i = 0] \tag{9}$$

$$= \quad Avg_n[Y_i(1)|D_i = 1] - Avg_n[Y_i(0)|D_i = 0]. \tag{10}$$

This second equation is what we observe in the data.

Q. Why can we derive the second equation from the first one?

Average treatment effect (cont.)

Define a constant term γ by

$$\gamma = Y_i(1) - Y_i(0). \tag{11}$$

γ means not only the individual causal effect, but also the average causal effect (i.e., ATE) of taking a malaria pill.

- That is, this γ is what we want to identify using data.

Then

$$\begin{aligned} & Avg_n[Y_i(1)|D_i = 1] - Avg_n[Y_i(0)|D_i = 0] \\ = \quad & Avg_n[\gamma + Y_i(0)|D_i = 1] - Avg_n[Y_i(0)|D_i = 0] \qquad (12) \\ = \quad & \gamma + \{Avg_n[Y_i(0)|D_i = 1] - Avg_n[Y_i(0)|D_i = 0]\}. \qquad (13) \end{aligned}$$

The second term is selection bias. Can you observe it in the data?

If the second term is positive, we call it positive selection bias, while if the second term is negative, we call it negative selection bias.

Moving from sample average to population average

So far, we have focused on sample average $Avg_n[Y_i]$ for some Y_i. Denote this as $\bar{Y}$ for simplicity. We want to move from there to population average, denoted by $E[Y_i]$.

Before doing so, let's begin by learning the meaning of $E[x]$, or a mathematical expectation.

Expectation

Suppose you play dice. Given a fair die, roll it once and save the result. Repeat the same thing and compute the average of these results. (If you get $\{4, 5\}$, you get 4.5.) Repeat this process several times.

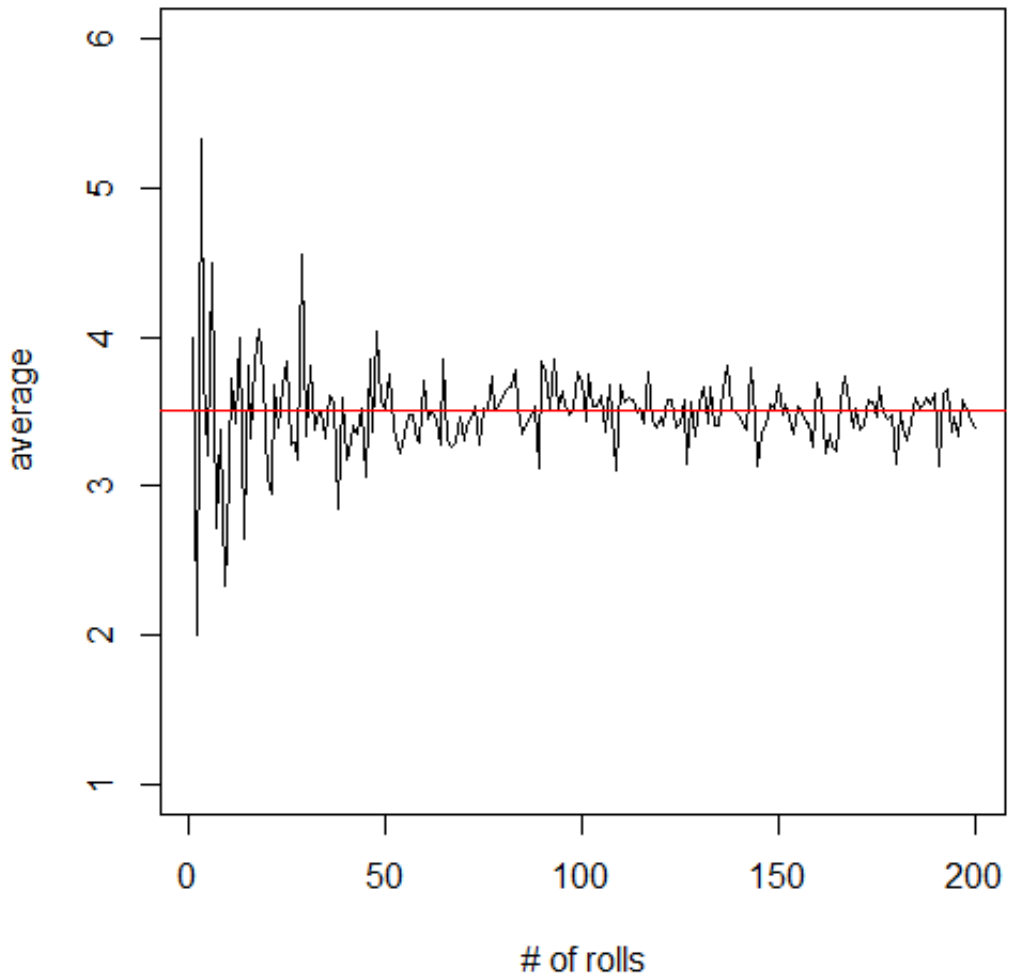
Because the die is fair, you should get each number equally likely. Thus, the sample average gets closer and closer to

$$\frac{1+2+3+4+5+6}{6} = 3.5 \tag{14}$$

This 3.5 is called a mathematical expectation.

The statistical property that a sample average converges to a mathematical expectation (or the average of population from which the sample is drawn) is called the Law of Large Numbers (LLN).

Sample and population average



Expectation (cont.)

Let's move from sample average to population average!

Why?

In randomized experiments, we often randomly divide sample into treatment and control groups (say, by a coin toss).

If the LLN is in action, and the sample size for both groups are large enough, we may have similar individuals in each group.

- Individuals should be very similar in every aspect (educational attainment, age, etc.) between those two groups.

Person A and Person B are quite different. But the average of many people in either group is likely to be similar.

ATE

The population version of the Average Treatment Effect (ATE) is written by

$$E[Y_i(1) - Y_i(0)]. \tag{15}$$

Removing selection bias

Let $E[Y_i|D_i = 1]$ be the population average of Y_i for individuals with $D_i = 1$, and $E[Y_i|D_i = 0]$ for individuals with $D_i = 0$.

- These expressions are called conditional expectations. You can think of these as subgroup averages.

Difference in expectation is written by

$$
\begin{aligned}
& E[Y_i|D_i = 1] - E[Y_i|D_i = 0] && (16)\\
= \; & E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0] && (17)\\
= \; & E[Y_i(1) - Y_i(0)|D_i = 1] && (18)\\
& + \{E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0]\} && (19)\\
= \; & \gamma + \{E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0]\}. && (20)
\end{aligned}
$$

If the sample is so large that the LLN is in action (so that (13) can be replaced by (20)), selection bias can be removed by the random assignment.

- Thus we can get a causal effect in the end!

Removing selection bias (cont.)

Formally, we need the independence assumption to remove selection bias

$$D_i \perp (Y_i(1), Y_i(0)), \tag{21}$$

Which states that D_i is independent of potential outcomes, i.e., it is randomly assigned.

Heterogeneity

So far we have assumed that $\gamma = Y_i(1) - Y_i(0)$ is constant across individuals.

- This assumption implies $E[Y_i(1) - Y_i(0)|D_i = 1] = \gamma$.

In a more general case (e.g., random coefficient models), the ATE and the following effect are not the same

$$E[Y_i(1) - Y_i(0)|D_i = 1]. \tag{22}$$

This effect is called the Average Effect of Treatment on the Treated (ATT) or Treatment on the Treated (TOT).

- If the causal effect is homogeneous (i.e., the effect is the same between the treatment and control group), we can get ATT = ATE.

Heterogeneity (cont.)

Rewrite the difference in means by

$$
\begin{aligned}
& E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0] && (23)\\
= \quad & \text{ATT} + \text{Selection Bias} && (24)\\
= \quad & \text{ATE} + (\text{ATT} - \text{ATE}) + \text{Selection Bias} && (25)
\end{aligned}
$$

The second term in parentheses captures selection on returns to treatment.

- If treated individuals have higher returns to the treatment, this term becomes positive.
- This is different from selection bias!

Statistical significance

Suppose that the group means of the malaria pill experiment is given by

	Treatment	Control	Difference
Prob(malaria)	0.7	0.8	-0.1
Female	0.55	0.56	0.01
Age	0.7	0.3	0.4
Education	0.4	0.5	0.1
Sample size	5,000	1,000	

We see differences in group means, but we do not know whether the differences are also *statistically significant*.

Ideally, we want the difference in the outcome (Prob(malaria)) to be statistically significant, while the difference in the rest of variables to be statistically insignificant.

- Why?

How can we check them?

Statistical significance (cont.)

	Treatment	Control	Difference
Prob(malaria)	0.7	0.8	-0.1
			(???)
Female	0.55	0.56	0.01
			(???)
Age	0.7	0.3	0.4
			(???)
Education	0.4	0.5	0.1
			(???)
Sample size	5,000	1,000	

In order to check whether a mean difference is statistically significant or not, we need to compute standard errors.

How?

Get standard errors

Let $\bar{Y}^1 = Avg_n[Y_i|D_i = 1]$ and $\bar{Y}^0 = Avg_n[Y_i|D_i = 0]$. Assume that they are statistically independent.

Denote σ^2_Y be the population variance of Y, and n_0 and n_1 be the sample size for each group.

The variance of a difference is written by

$$
\begin{aligned}
V(\bar{Y}^1 - \bar{Y}^0) &= V(\bar{Y}^1) + V(\bar{Y}^0) && (26)\\
&= V(\frac{1}{n_1}\sum_i Y_i^1) + V(\frac{1}{n_0}\sum_j Y_j^0) && (27)\\
&= \frac{1}{n_1^2}\sum_i V(Y_i^1) + \frac{1}{n_0^2}\sum_j V(Y_j^0) && (28)\\
&= \frac{\sigma^2_{Y^1}}{n_1} + \frac{\sigma^2_{Y^0}}{n_0}. && (29)
\end{aligned}
$$

Get standard errors (cont.)

Where I use the rule that the variance of the sum of variances is the sum of variances and that the variance of a difference is the sum of variances for statistically independent variables.

- Does this always hold?

Assume $\sigma^2_{Y^1} = \sigma^2_{Y^0} = \sigma^2_Y$.

- What does this mean?

Then, the standard error of the difference is

$$SE(\bar{Y}^1 - \bar{Y}^0) = \sigma_Y \sqrt{\frac{1}{n_1} + \frac{1}{n_0}}. \tag{30}$$

Get standard errors (cont.)

In practice, we can estimate σ_Y and get an estimated standard error

$$\hat{SE}(\bar{Y}^1 - \bar{Y}^0) = S(Y_i)\sqrt{\frac{1}{n_1} + \frac{1}{n_0}}, \tag{31}$$

Where $S(Y_i)$ is the pooled sample standard deviation

$$S(Y_i) = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_0 - 1)s_0^2}{n_1 + n_0 - 2}}, \tag{32}$$

Where s_i is the standard deviation of group i.

- This expression assumes that two groups have the same population variance. (If not, use Welch's t-test.)

Get standard errors (cont.)

Under the null hypothesis that $\mu_1 - \mu_0 = \mu$, where μ_0 and μ_1 are particular values of the population means, the t-statistic for a difference in means is written by

$$t(\mu) = \frac{\bar{Y}^1 - \bar{Y}^0 - \mu}{\hat{SE}(\bar{Y}^1 - \bar{Y}^0)}. \tag{33}$$

For $\mu = 0$ (i.e., the difference is zero in population), this expression simply means the difference in sample means divided by the estimated standard error.

If $t(\mu)$ is large enough, we say that the estimated difference is statistically significant.

- But how large should it be? (You should already know the answer!)

Statistical significance (cont.)

Consider a simpler case.

Let $\bar{Y}$ be the sample mean of, and $E[Y_i]$ be the population mean of, Y_i. Assume that the population mean takes a value μ.

The t-statistic is written by

$$t(\mu) = \frac{\bar{Y} - \mu}{\hat{SE}(\bar{Y})}, \tag{34}$$

Where

$$\hat{SE}(\bar{Y}) = \frac{S(Y_i)}{\sqrt{n}}. \tag{35}$$

$S(Y_i)$ is the standard deviation of Y_i and n is the sample size.

The t-statistic has a t-distribution.

If the sample size is large enough, by the Central Limit Theorem (CLT), the sampling distribution of $t(\mu)$ becomes very close to a standard normal distribution.

Flipping a coin

What does it mean by "the sampling distribution of $t(\mu)$ becomes very close to a standard normal distribution"?

Let's randomly flip a fair coin for n times.

- Let Head $= 1$ and Tail $= 0$.
- You get n observations of 1s and 0s.

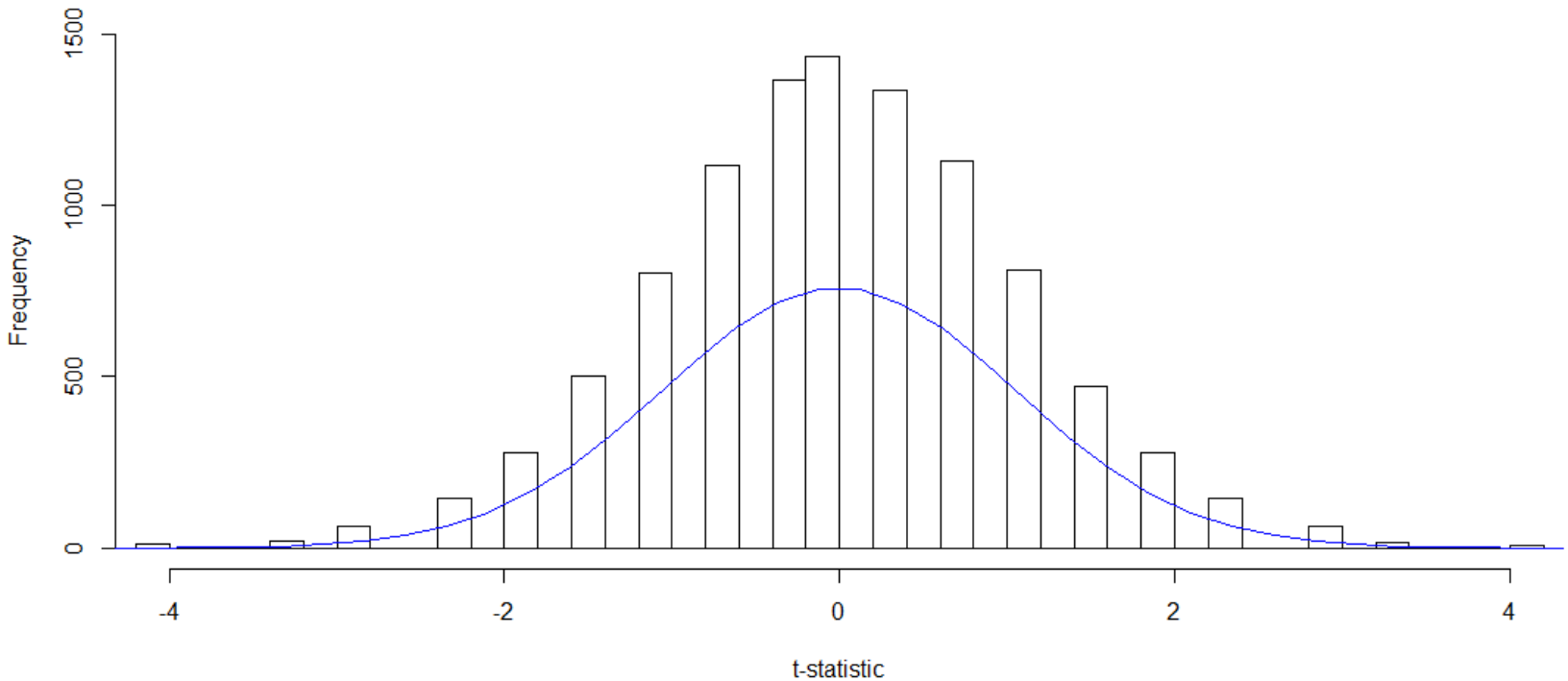
Compute a t-statistic

- What is the population mean μ?

Repeat it for 10,000 times.

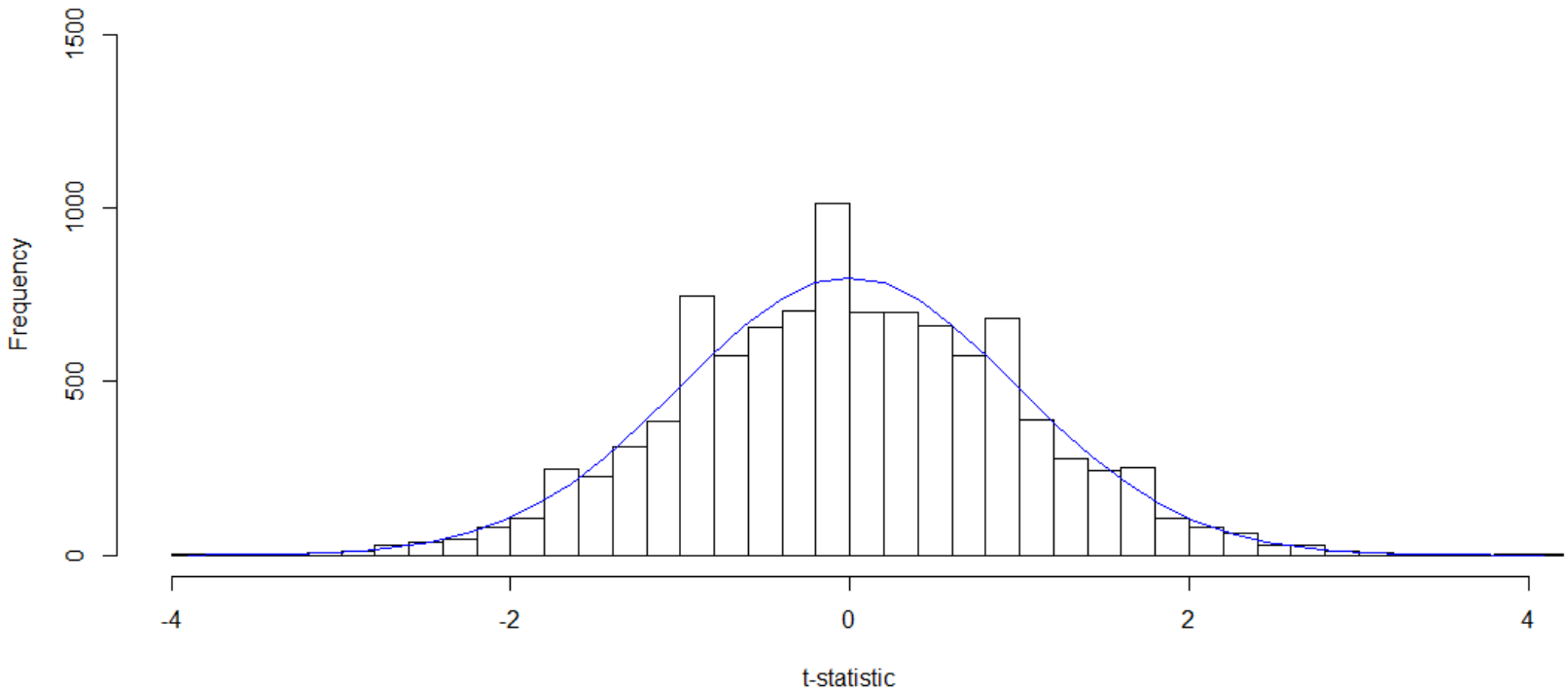
- Thus you get 10,000 t-statistics.

Distribution of t-statistic



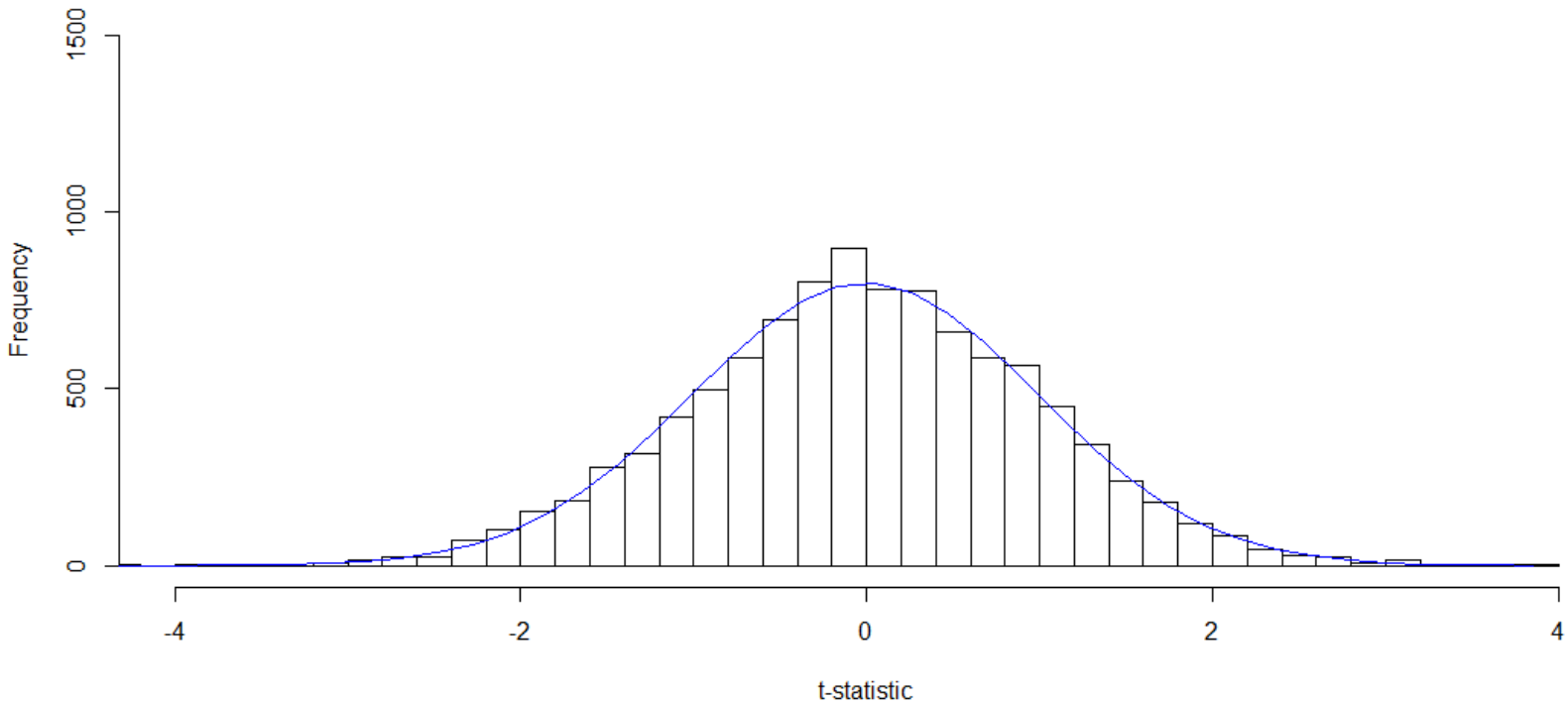
n = 30.

Distribution of t-statistic (cont.)



n = 500.

Distribution of t-statistic (cont.)



n = 5000.

Statistical significance (cont.)

For a significance level α, we can define $t_{\alpha/2}$ such that

$$Prob\left(-t_{\alpha/2} < \frac{\bar{Y} - \mu}{\sigma_{\bar{Y}}} < t_{\alpha/2}\right) = 1 - \alpha. \tag{36}$$

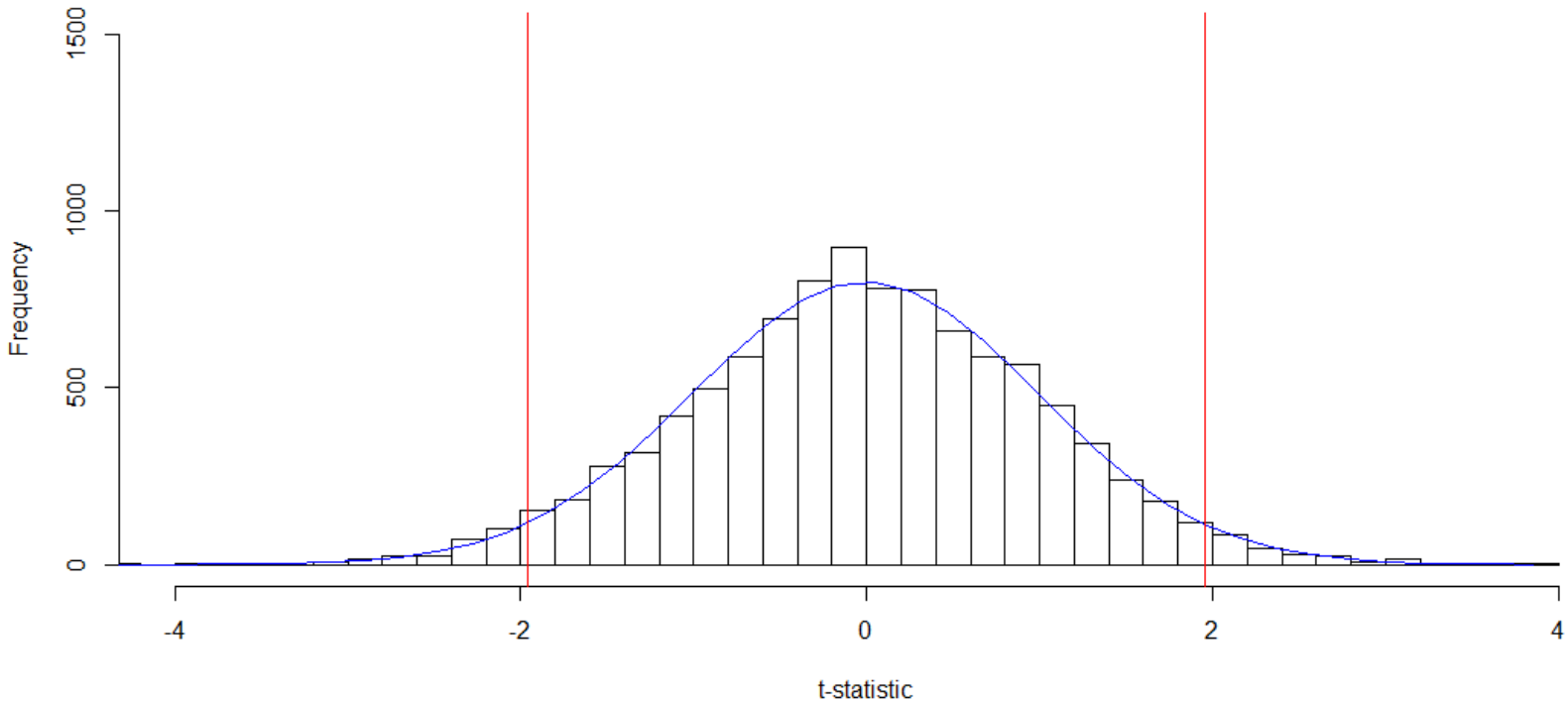
This is the probability that a t-statistic falls within $(-t_{\alpha/2}, t_{\alpha/2})$

For $\alpha = 0.05$, we get $t_{\alpha/2} = 1.96$.

That is, when the t-statistic is larger than about 2 in absolute value, we say that the sample mean $\bar{Y}$ is statistically different from μ.

- If you get an OLS estimate $\hat{\beta}$, you want to know whether the estimate is statistically different from zero.
- If you get a difference in means, you want to know whether the difference is statistically different from zero.

Statistical significance (cont.)



Statistical significance (cont.)

Let's go back to the above example.

Suppose you get the following standard errors.

	Treatment	Control	Difference
Prob(malaria)	0.7	0.8	-0.1
			(**0.04**)
Female	0.55	0.56	0.01
			(**0.01**)
Age	0.7	0.3	0.4
			(**0.35**)
Education	0.4	0.5	0.1
			(**0.09**)
Sample size	5,000	1,000	

What can you conclude from this table? Recall

$$t(\mu) = \frac{\bar{Y}^1 - \bar{Y}^0}{\hat{SE}(\bar{Y}^1 - \bar{Y}^0)}.$$

Job training and labor-market outcomes

A labor training program, the National Supported Work (NSW), was implemented by the Manpower Demonstration Research Corporation (MDRC) in ten sites in the U.S. during the mid-1970s.

- The program provided work experience (e.g., restaurant and construction work) to disadvantaged workers for 9-18 months.
- Disadvantaged workers: ex-drug addicts, ex-criminal offenders, high school dropouts, etc.

Candidates were randomized into the program between March 1975 and July 1977.

- Only treated individuals received all the benefits of the program.

What was the impact of the NSW program on earnings?

Robert J. LaLonde (1986) has explored this question.

Impact on annual earnings

TABLE 3—ANNUAL EARNINGS OF NSW MALE TREATMENTS, CONTROLS, AND SIX CANDIDATE COMPARISON GROUPS FROM THE *PSID* AND *CPS-SSA*

			Comparison Group[a,b]					
Year	Treatments	Controls	*PSID*-1	*PSID*-2	*PSID*-3	*CPS-SSA*-1	*CPS-SSA*-2	*CPS-SSA*-3
1975	$3,066	$3,027	19,056[a]	7,569	2,611	13,650	7,387	2,729
	(283)	(252)	(272)	(568)	(492)	(73)	(206)	(197)
1976	$4,035	$2,121	20,267	6,152	3,191	14,579	6,390	3,863
	(215)	(163)	(296)	(601)	(609)	(75)	(187)	(267)
1977	$6,335	$3,403	20,898	7,985	3,981	15,046	9,305	6,399
	(376)	(228)	(296)	(621)	(594)	(76)	(225)	(398)
1978	$5,976	$5,090	21,542	9,996	5,279	14,846	10,071	7,277
	(402)	(227)	(311)	(703)	(686)	(76)	(241)	(431)
Number of Observations	297	425	2,493	253	128	15,992	1,283	305

[a]The Comparison Groups are defined as follows: *PSID*-1: All male household heads continuously from 1975 through 1978, who were less than 55-years-old and did not classify themselves as retired in 1975; *PSID*-2: Selects from the *PSID*-1 group all men who were not working when surveyed in the spring of 1976; *PSID*-3: Selects from the *PSID*-1 group all men who were not working when surveyed in either spring of 1975 or 1976; *CPS-SSA*-1: All males based on Westat's criteria, except those over 55-years-old; *CPS-SSA*-2: Selects from *CPS-SSA*-1 all males who were not working when surveyed in March 1976; *CPS-SSA*-3: Selects from the *CPS-SSA*-1 unemployed males in 1976 whose income in 1975 was below the poverty level.

[b]All earnings are expressed in 1982 dollars. The numbers in parentheses are the standard errors. The number of observations refer only to 1975 and 1978. In the other years there are fewer observations. The sample of treatments is smaller than the sample of controls because treatments still in Supported Work as of January 1978 are excluded from the sample, and in the young high school target group there were by design more controls than treatments.

Source: Lalonde (1986).

Impact on annual earnings (cont.)

Columns (1) and (2) of the table show the pattern of average earnings for males in the treatment group and the control group, respectively.

- 1975: Pre-treatment
- 1976, 1977: Treatment
- 1978: Post-treatment

Why did he also check pre-treatment earnings?

According to the table, the earnings diverged during the program, and converged to some extent after the program ended.

R exercise

Let's replicate Lalonde (1986).

Launch RStudio.

R exercise (cont.)

Type

```
jt <- wooldridge::jtrain2
```

The data contain 445 males who are treated between December 1975 and January 1978.

- This is a subset of Lalonde's original data for which earnings are available for 1974.

R exercise (cont.)

Let's compare mean earnings in 1978 between the treatment and the control group.

Type

```
re78_t <- with(jt, mean(re78[train==1]))
re78_c <- with(jt, mean(re78[train==0]))
diff_mean <- re78_t - re78_c
diff_mean
```

`diff_mean` is $(\bar{Y}^1 - \bar{Y}^0)$.

What information do we need to check whether the difference is statistically significant?

R exercise (cont.)

Let's compute standard errors. Recall

$$\hat{SE}(\bar{Y}^1 - \bar{Y}^0) = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}. \tag{37}$$

Type

```
n1 <- with(jt, length(re78[train==1]))
n2 <- with(jt, length(re78[train==0]))
sd1 <- with(jt, sd(re78[train==1]))
sd2 <- with(jt, sd(re78[train==0]))
sd_p <- sqrt(((n1-1)*sd1^2 + (n2-1)*sd2^2)/(n1+n2-2))
sd <- sd_p * sqrt(1/n1 + 1/n2)
```

R exercise (cont.)

Finally, you get a t-statistic by typing

```
t <- diff_mean/sd
t
```

You should get about 2.84.

Is the difference statistically significant?

- You should already know the answer!
- Alternatively, compute a p-value by typing `2 * (1 - pt(t, df=n1+n2-1))` (for two-sided test).

R exercise (cont.)

You could easily test the significance by using `t.test`.

```
res <- t.test(re78 ~ train, data = jt, var.equal=TRUE)
res
```

If the population variances are not the same, use Welch's t-test by changing `var.equal` to `FALSE`.

Confidence interval

Instead of checking whether the sample is consistent with a specific value of μ, you can compute a set of all values of μ that are consistent with the sample.

- Such a set is called a confidence interval (CI).

A 95% CI for the population mean $E[Y_i]$ is written by

$$[\bar{Y} - 2 \times \hat{SE}(\bar{Y}), \bar{Y} + 2 \times \hat{SE}(\bar{Y})], \tag{38}$$

Where $\bar{Y}$ is the sample mean, $\hat{SE}(\bar{Y})$ is the estimated standard error.

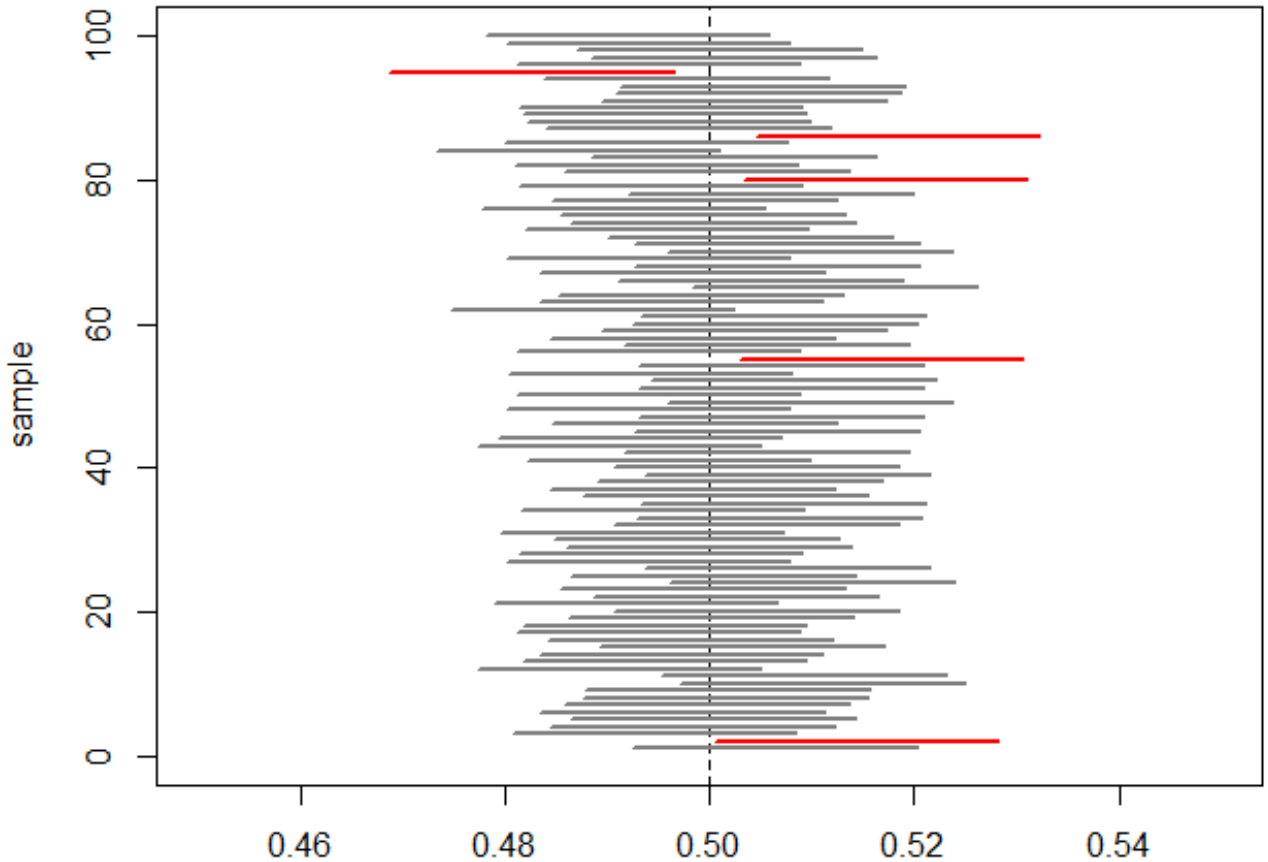
The interval should contain $E[Y_i]$ about 95% of the time.

- In other words, if you draw a sample and compute a CI, and repeat it for 100 times, 5/100 CIs do not contain $E[Y_i]$.

The interval gets narrower as the sample size increases!

- The true value is likely to be found in that narrow interval.

Confidence interval (cont.)



Take first 100 confidence intervals out of 10,000 in the coin example. The red color indicates CIs which do not contain $E[Y_i]$.

Effect size

How do we know whether an effect is large or small?

There are several ways to measure effect size.

Reporting the coefficient size is just fine, but the standardization of the size is often very useful. One example is

$$\theta = \frac{\bar{Y}^1 - \bar{Y}^0}{S(Y_i)}, \tag{39}$$

Which is called *Cohen's d*.

The difference in means is divided by the pooled standard deviation, so that we can interpret effect size *in standard deviation*.

- Notice similarity between Cohen's d and t-statistic.

Suppose $\theta = 0.2$. This means that the treatment moves the outcome by 0.2 standard deviations.

- 0.2 is often used as the benchmark in practice.
- 0.2 small, 0.5 medium, 0.8 large (Cohen 1988).

Effect size (cont.)

Let's go back to the malaria example.

	Treatment	Control	Difference
Prob(malaria)	0.7	0.8	-0.1
			(0.04)
Female	0.55	0.56	0.01
			(0.01)
Age	0.7	0.3	0.4
			(0.35)
Education	0.4	0.5	0.1
			(0.09)
Sample size	5,000	1,000	

How large is Cohen's d? Can you say that the effect is large?

- How can we check it?

R exercise (cont.)

Let's compute Cohen's d using the job training data.

Type

```
cohen_d <- diff_mean/sd_p
cohen_d
```

Is effect size large?

Conducting a field experiment

Before the experiment

- Make hypotheses. **Calculate the sample size**. Write a plan
- Get the approval of the Institutional Review Board (IRB)

During the experiment

- Conduct a baseline, endline, (and possibly) followup survey

After the experiment

- Analyze data and summarize results

Power calculation

How can we derive the sample size?

To do so, power calculation is needed.

What is (statistical) power? How can we compute the sample size?

What is power?

Recall a table that you might have seen in a statistics class.

	H_0 is True	H_0 is False
Fail to reject	Correct	Type II error (β)
Reject	Type I error (α)	Correct

You want to minimize two statistical errors.

- Type I error is the significance level α. That is, it is the probability of rejecting the null hypothesis when it is true. α is often set to 5%.
- Type II error is the probability of failing to reject the null when it is false, and is denoted by β. Power is $1-\beta$. Power is often set to 80%. (In the following slides, I use the notation κ for power in order to avoid any confusion with a regression coefficient β.)

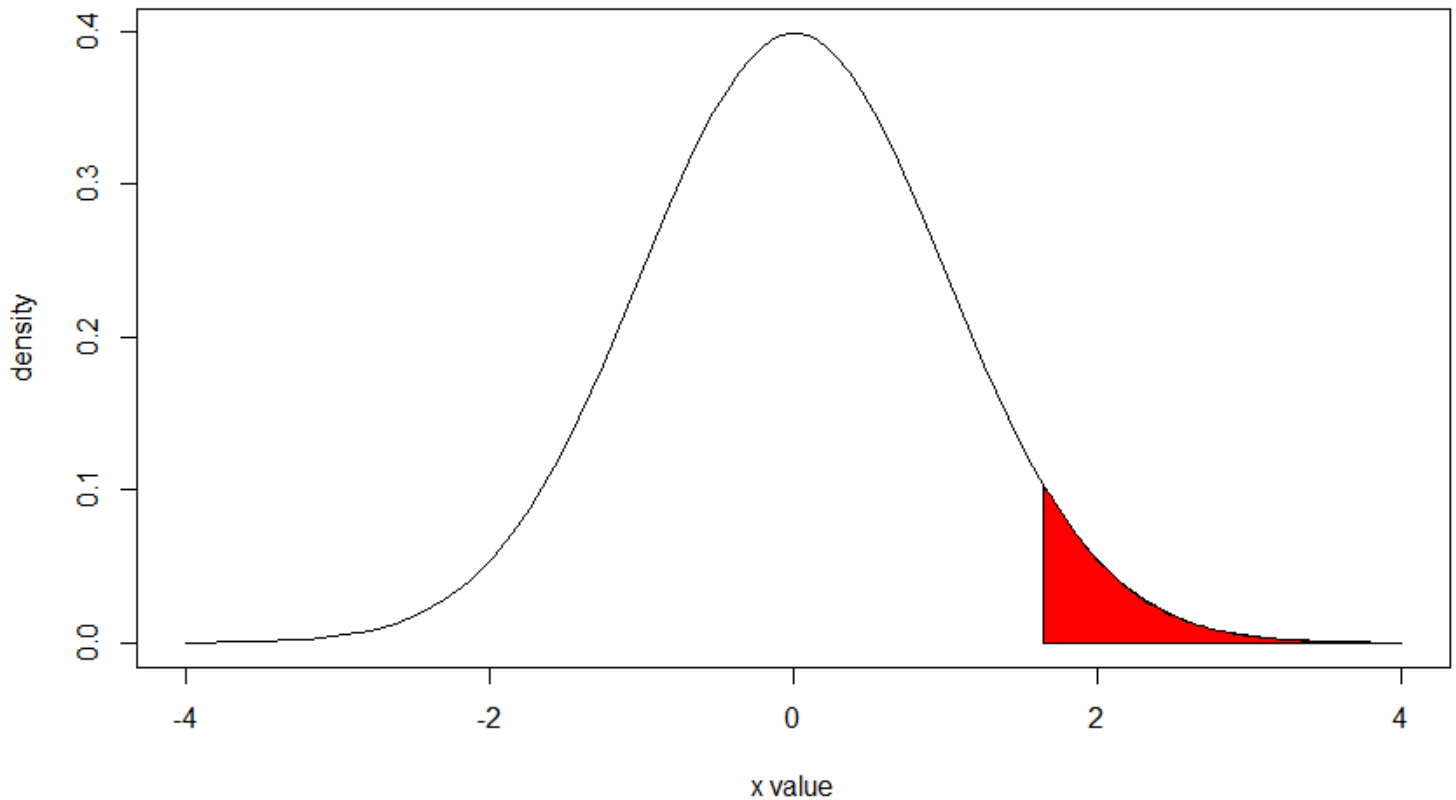
Get $\hat{\beta}$

Suppose you get

$$\hat{\beta} = \bar{Y}^1 - \bar{Y}^0 \quad \text{and} \quad \hat{SE}(\bar{Y}^1 - \bar{Y}^0), \tag{40}$$

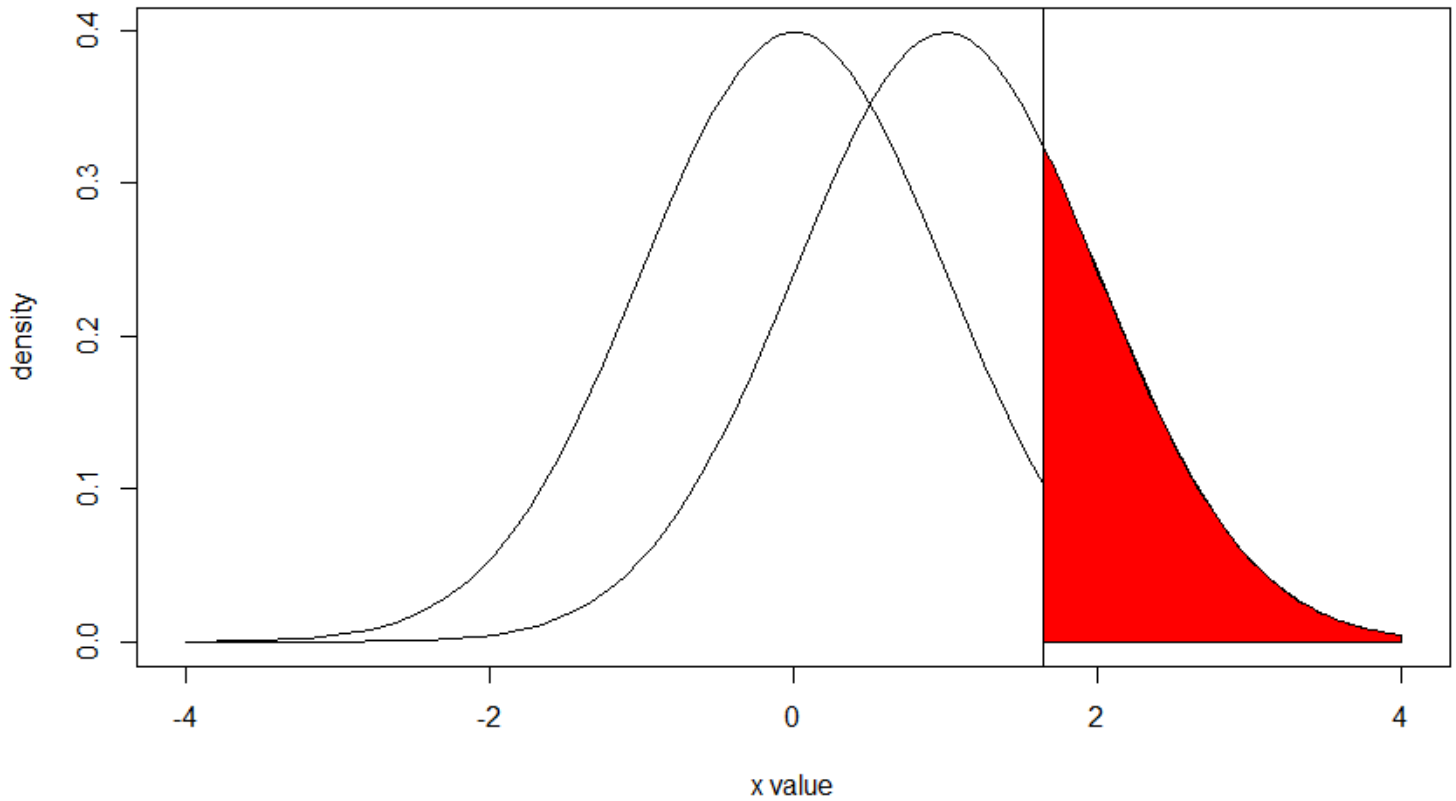
So that you can calculate t-statistic.

Graphical example



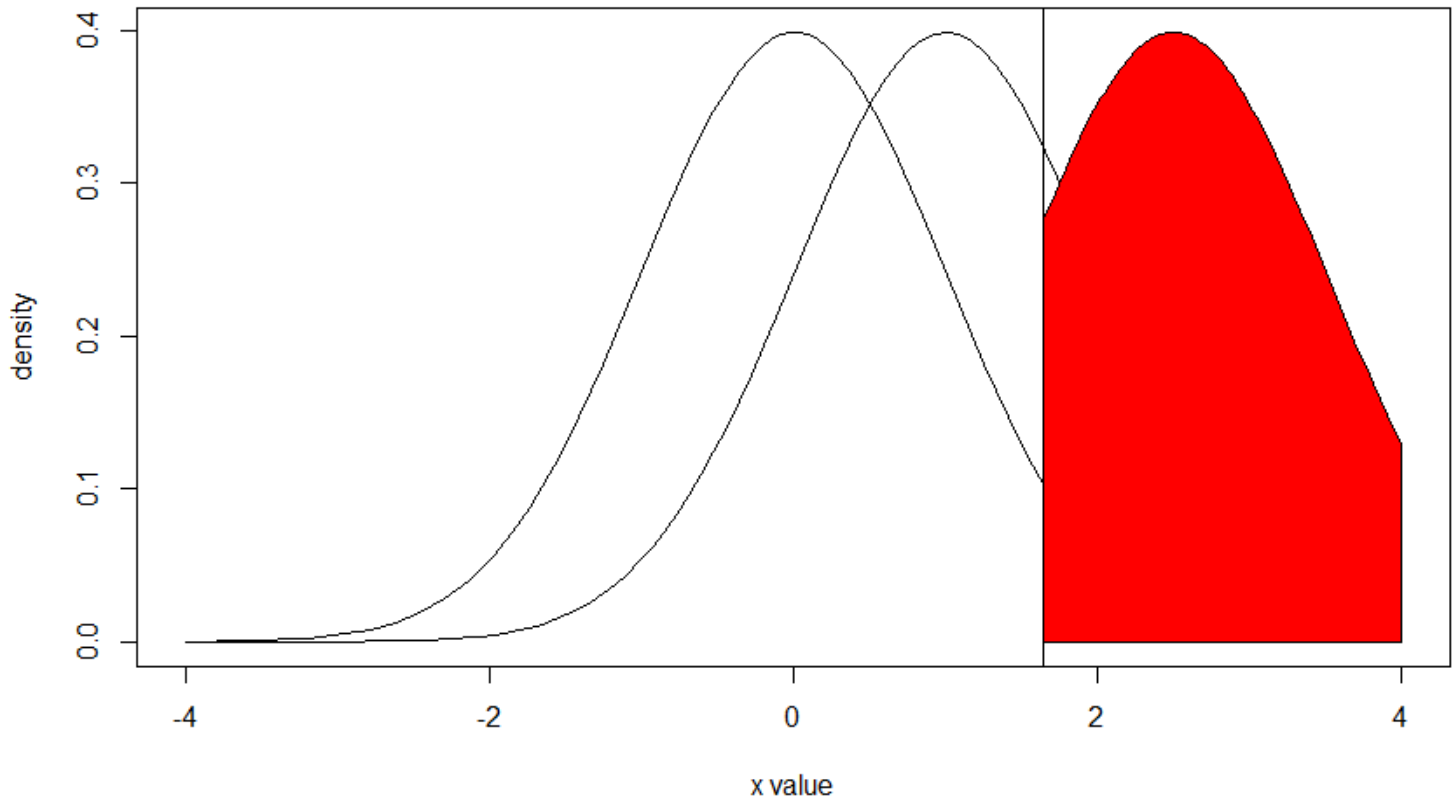
The red shaded area measures size α ($\alpha/2$ for two-tailed test). In this example, $t_\alpha = t_{0.05} = 1.65$. The standard normal distribution is assumed.

Graphical example (cont.)



The true value is assumed to be 1 (so that the tipping point of the right distribution is located where x-axis is 1). The red shaded area measures power κ. In this example, power is 0.26.

Graphical example (cont.)



The true value is assumed to be 2.5. In this example, power is 0.8.

MDE

Power is defined by (for two-sided case)

$$\kappa = \text{Prob}\left(\frac{\hat{\beta}}{\sigma_{\hat{\beta}}} \leq -t_{\alpha/2}|H_1\right) + \text{Prob}\left(\frac{\hat{\beta}}{\sigma_{\hat{\beta}}} \geq t_{\alpha/2}|H_1\right). \tag{41}$$

The first term is almost zero under the alternative hypothesis.

- Why?

Thus

$$\begin{aligned}
\kappa &= \text{Prob}\left(\frac{\hat{\beta}}{\sigma_{\hat{\beta}}} \geq t_{\alpha/2}|H_1\right) && (42)\\
&= 1 - \text{Prob}\left(\frac{\hat{\beta}}{\sigma_{\hat{\beta}}} < t_{\alpha/2}|H_1\right) && (43)\\
&= 1 - \text{Prob}\left(\frac{\hat{\beta}-\beta}{\sigma_{\hat{\beta}}} < t_{\alpha/2} - \frac{\beta}{\sigma_{\hat{\beta}}}\right) && (44)\\
1-\kappa &= \text{Prob}\left(\frac{\hat{\beta}-\beta}{\sigma_{\hat{\beta}}} < t_{\alpha/2} - \frac{\beta}{\sigma_{\hat{\beta}}}\right). && (45)
\end{aligned}$$

MDE (cont.)

Now $\frac{\hat{\beta}-\beta}{\sigma_{\hat{\beta}}}$ of

$$1-\kappa = \text{Prob}\left(\frac{\hat{\beta}-\beta}{\sigma_{\hat{\beta}}} < t_{\alpha/2} - \frac{\beta}{\sigma_{\hat{\beta}}}\right) \tag{46}$$

Has a t-distribution.

- The normal distribution if the sample size is large.

Also, at the threshold, we have

$$t_\kappa = t_{\alpha/2} - \frac{\beta}{\sigma_{\hat{\beta}}}. \tag{47}$$

Since $-t_\kappa = t_{1-\kappa}$, we have

$$\beta_{MDE} = (t_{1-\kappa} + t_{\alpha/2}) \times \sigma_{\hat{\beta}}. \tag{48}$$

How to determine the sample size

This is called the Minimum Detectable Effect Size (MDE)

$$\beta_{MDE} = (t_{1-\kappa} + t_{\alpha/2}) \times SE(\bar{Y}^1 - \bar{Y}^0). \quad (49)$$

MDE depends on

- Significance level
- Power
- Sample size in the treatment and control group
- Variance of the population outcomes

We can use $\hat{SE}(\bar{Y}^1 - \bar{Y}^0)$ for $SE(\bar{Y}^1 - \bar{Y}^0)$.

- Recall (31).

For $\kappa = 0.8$ and $\alpha = 0.05$

$$\beta_{MDE} = (0.84 + 1.96) \times \hat{SE}(\bar{Y}^1 - \bar{Y}^0) \quad (50)$$

$$= 2.8 \times \hat{SE}(\bar{Y}^1 - \bar{Y}^0). \quad (51)$$

You could also say that MDE is such that t-statistic is equal to 2.8, or more precisely 2.83.

How to determine the sample size (cont.)

If we decide MDE, κ, α, and know sample standard deviations, we can compute the sample size.

- Baseline: $MDE = 0.2$, $\kappa = 0.8$, and $\alpha = 0.05$.
- Sample standard deviations are obtained from the pilot study or from the literature.

How to determine the sample size (cont.)

In practice, we compute n in the following way.

Suppose a half of the sample is treated ($n_1 = n_2 = n/2$) and the population variance is the same. Then rewriting (46) yields

$$0.8 = 1 - \text{Prob}\left(z < 1.96 - \frac{0.2}{2 * S(Y_i)/\sqrt{n}}\right) \tag{52}$$

$$0.2 = \text{Prob}\left(z < 1.96 - \frac{0.2}{2 * S(Y_i)/\sqrt{n}}\right), \tag{53}$$

Where $S(Y_i)$ is

$$S(Y_i) = \sqrt{\frac{(n/2-1)s_1^2 + (n/2-1)s_2^2}{n-2}}.$$

Derive these expressions.

Given $\{s_1, s_2\}$, find n that satisfies the equation (53) using a t-distribution (or standard normal distribution if n is large).

Power calculation in practice

Let's compute power in the job training example.

In RStudio, type

```
alpha <- 0.05
qu <- qt(alpha/2, df=n1+n2-2, lower=FALSE)
power <- 1 - pt(qu, df=n1+n2-2, ncp=diff_mean/sd) +
pt(-qu, df=n1+n2-2, ncp=diff_mean/sd)

power
```

Where $\alpha = 0.05$ and two-sided test are assumed.

Regression analysis

Let's move on to regression analysis.

Conditional expectation function

But first, let's refresh the memory of the conditional expectation function...

A conditional expectation

$$E[Y_i|X_i = x] \tag{54}$$

Tells us the population average of Y_i given that X_i takes a specific value x.

- E.g., The population average of earnings when educational attainment is x years.

In contrast,

$$E[Y_i|X_i] \tag{55}$$

Is called the conditional expectation function (CEF).

- The CEF is a function, while the conditional expectation is a value.

You can add more than one conditioning variable

$$E[Y_i|X_{1i}, X_{2i}, ..., X_{ki}]. \tag{56}$$

Conditional expectation function (cont.)

Suppose that the CEF of earnings is a *linear* function of college education (X_i equals to one if i has a college degree)

$$E[Y_i|X_i] = \alpha + \beta X_i. \tag{57}$$

Then the regression

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \tag{58}$$

With $E[\varepsilon_i|X_i] = 0$ recovers this linear function.

To get β, take a first-difference

$$E[Y_i|X_i = 1] - E[Y_i|X_i = 0] = \beta. \tag{59}$$

To get α

$$E[Y_i|X_i = 0] = \alpha. \tag{60}$$

Conditional expectation function (cont.)

What happens if $E[Y_i|X_{1i}, X_{2i}, ..., X_{ki}]$ is nonlinear?

In that case, the regression of Y_i on $X_{1i}, X_{2i}, ..., X_{ki}$ gives the best linear approximation to the CEF.

Potential outcome model

Consider a potential outcome model

$$Y_i = Y_i(0) + (Y_i(1) - Y_i(0))D_i, \tag{61}$$

Where Y_i is earnings and D_i is a college dummy.

One can rewrite this model by

$$Y_i = \alpha + \beta D_i + \varepsilon_i, \tag{62}$$

Where $\alpha = E[Y_i(0)]$, $\beta = Y_i(1) - Y_i(0)$, and $\varepsilon_i = Y_i(0) - E[Y_i(0)]$.

We already know that β is the causal effect.

Can we get a causal effect by running the regression (62)?

Potential outcome model (cont.)

Take the conditional expectations of (62)

$$E[Y_i|D_i=1] = \alpha+\beta+E[\varepsilon_i|D_i=1] \quad (63)$$
$$E[Y_i|D_i=0] = \alpha+E[\varepsilon_i|D_i=0], \quad (64)$$

Which implies

$$E[Y_i|D_i=1]-E[Y_i|D_i=0] = \beta+\{E[\varepsilon_i|D_i=1]-E[\varepsilon_i|D_i=0]\}. \quad (65)$$

The second term is selection bias. To see this

$$E[\varepsilon_i|D_i=1]-E[\varepsilon_i|D_i=0] = E[Y_i(0)|D_i=1]-E[Y_i(0)|D_i=0], \quad (66)$$

Where I used $\varepsilon_i = Y_i(0)-E[Y_i(0)]$.

Potential outcome model (cont.)

Thus, selection bias is correlation between ε_i and D_i!

As long as there is such correlation, we cannot get a causal effect by regressing Y_i on D_i.

Is there any way to mitigate the problem?

Regression-control framework

The conditional independence assumption (CIA) asserts that, conditional on X_i, treatment (e.g., going to college) becomes independent of potential outcomes

$$D_i|X_i \perp (Y_i(1), Y_i(0)). \tag{67}$$

If the CIA is satisfied, we get a causal effect of going to college on earnings by running

$$Y_i = \alpha + \beta D_i + \lambda X_i + \varepsilon_i. \tag{68}$$

Q. Discuss potential candidates for X_i.

Regression-control framework (cont.)

It's good to know what X_i can do.

Let's investigate this using the job training data.

We use the non-experimental version of the data (`jtrain3`). The data contain more individuals in the control group who are not randomly assigned.

Type

```
res1 <- compareGroups(train ~ age + educ + black + hisp
+ married + re74 + re75 + re78 + unem74 + unem75 +
unem78, data=jt)
res2 <- compareGroups(train ~ age + educ + black + hisp
+ married + re74 + re75 + re78 + unem74 + unem75 +
unem78, data=jtrain3)

createTable(res1)
createTable(res2)
```

Are these variables balanced in both data?

Regression-control framework (cont.)

Let's check whether estimates would change by using the non-experimental data and whether adding control variables can solve the problem.

Type

```
reg1 <- lm(re78 ~ train, data = jt)
reg2 <- lm(re78 ~ train + age + educ + black + hisp +
married + re74 + re75, data=jt)
reg3 <- lm(re78 ~ train, data = jtrain3)
reg4 <- lm(re78 ~ train + age + educ + black + hisp +
married + re74 + re75, data=jtrain3)

stargazer(reg1, reg2, reg3, reg4, type="text")
```

Columns (1) and (2) use experimental data, while columns (3) and (4) use non-experimental data (closer to observational data).

What do you find?

Regression-control framework (cont.)

Since the program provided work experience to disadvantaged workers, we may need to find a better control group so that it is more comparable to the treatment group.

Idea: What about restricting the sample to those who had been unemployed before the program started?

Type

```
reg5 <- lm(re78 ~ train, data=jtrain3,
subset=c(unem75==1 & unem74==1))
reg6 <- lm(re78 ~ train + age + educ + black + hisp +
married + re74 + re75, data=jtrain3, subset=c(unem75==1
& unem74==1))

stargazer(reg1,reg2,reg5,reg6,type="text")
```

What do you find?

Regression-control framework (cont.)

This example tells you that

- (1) Having a good comparison group is important. Using experimental data are preferred if available.
- (2) Adding control variables does not always solve the problem. Omitted variables are not always available for econometricians.

It is likely that you have only observational data.

- Ask yourself: Can you say that the effect you find is likely a causal effect?

Regression and matching

What does it mean by "adding X_i to a regression"?

Suppose that the CIA is satisfied. Then a causal effect is

$$\begin{aligned}
\beta &= E[Y_i(1) - Y_i(0)|D_i = 1] && (69)\\
&= E\{E[Y_i(1) - Y_i(0)|X_i, D_i = 1]|D_i = 1\} && (70)\\
&= E\{E[Y_i(1)|X_i, D_i = 1] - E[Y_i(0)|X_i, D_i = 1]|D_i = 1\} && (71)\\
&= E\{E[Y_i(1)|X_i, D_i = 1] - E[Y_i(0)|X_i, D_i = 0]|D_i = 1\} && (72)\\
&= E[\delta_X|D_i = 1]. && (73)
\end{aligned}$$

Where I use the law of iterated expectations ($E[Y_i] = E[E[Y_i|X_i]]$) to get (70) and the CIA assumption to get (72).

Intuitively, we take the (weighted) average of difference in Y_i over groups defined by X_i.

Regression and matching (cont.)

Suppose that X_i is a discrete variable (e.g., going to university).

The matching estimand (the population quantities to be estimated) is written as the weighted average of university-non university difference in mean earnings (δ_X) for each group.

$$E[Y_i(1) - Y_i(0)|D_i = 1] = \sum_x \delta_x \text{Prob}(X_i = x|D_i = 1). \quad (74)$$

This is ATT, but ATE can be written similarly.

Regression and matching (cont.)

In the sample, one can replace $\text{Prob}(X_i = x | D_i = 1)$ by

$$\frac{I_x N_x^1}{\sum_x I_x N_x^1}, \tag{75}$$

Where $I_x = I(N_x^1 > 0, N_x^0 > 0)$ is the indicator variable taking value one if $N_x^1 > 0$ and $N_x^0 > 0$, and zero otherwise.

In other words, the weights are defined for all x's for which δ_x can be defined.

- What happens to the weight for $N_{\tilde{x}}^0 = 0$, i.e., there is no one in the control group for some particular group $\tilde{x}$?

Regression and matching (cont.)

What happens when you have more than one control variable?

Suppose you estimate the effect of union membership on earnings, and add two control variables, sex and age group.

$$Y_i = \alpha + \beta D_i + \lambda_1 Sex_i + \lambda_2 AgeGroup_i + \varepsilon_i. \tag{76}$$

Say age group takes five values (1 = 20s, ..., 5 = 60s). Then you have $2 * 5 = 10$ groups.

The matching estimand can be written as the weighted sum over these 10 groups.

Regression and matching (cont.)

Let's compare a regression estimate and a matching estimate using the job training data.

In RStudio, type

```
reg <- lm(re78 ~ train + black, data=jt)
stargazer(reg, type="text")
```

Note that we include the `black` dummy as a control.

Check the size of the estimate for `train`.

Regression and matching (cont.)

Type

```
black_t <- with(jt, mean(re78[black==1 & train==1]))
black_c <- with(jt, mean(re78[black==1 & train==0]))
diff_black <- black_t - black_c

noblack_t <- with(jt, mean(re78[black==0 & train==1]))
noblack_c <- with(jt, mean(re78[black==0 & train==0]))
diff_noblack <- noblack_t - noblack_c
```

(cont...)

Regression and matching (cont.)

```
n_black_t <- with(jt, length(black[black==1 &
train==1]))
n_noblack_t <- with(jt, length(black[black==0 &
train==1]))

w_black <- n_black_t/(n_black_t + n_noblack_t)
w_noblack <- n_noblack_t/(n_black_t + n_noblack_t)

matching <- w_black*diff_black + w_noblack*diff_noblack

matching
```

Endogeneity issue

There are several endogeneity issues such as

- Omitted-variable bias
- Measurement error
- Selection bias
- Reverse causality/simultaneity

Dealing with the endogeneity issue is the heart of rigorous empirical analysis.

How can we alleviate endogeneity?

OLS estimator

Before going through them, it's worthwhile to recall the derivation of regression coefficients.

Regression coefficients (α, β) of a *bivariate* regression (where there is a single independent variable)

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \tag{77}$$

Are obtained by minimizing the residual sum of squares with respect to b and a

$$E[(Y_i - a - bX_i)^2]. \tag{78}$$

The solution is

$$\beta = \frac{Cov(Y_i, X_i)}{Var(X_i)}, \tag{79}$$

$$\alpha = E[Y_i] - \beta E[X_i]. \tag{80}$$

OLS estimator (cont.)

What happens if there are more than one independent variable?

Suppose that one can estimate the effect of going to college using

$$Y_i = \alpha + \beta D_i + \lambda X_i + \varepsilon_i, \tag{81}$$

Where Y_i is earnings, D_i is a college dummy, and X_i is SAT scores.

The estimand for β is

$$\beta = \frac{Cov(Y_i, e_i)}{Var(e_i)}, \tag{82}$$

Where e_i is the residual from regressing D_i on X_i.

$$D_i = \psi_0 + \psi_1 X_i + e_i. \tag{83}$$

(82) is called the regression anatomy formula.

That is, you get the same coefficient by regressing Y_i on e_i.

- However, the standard errors needn't be the same.

OLS estimator (cont.)

In general, if the model has K variables

$$Y_i = \alpha + \beta_1 X_{1i} + ... + \beta_K X_{Ki} + \varepsilon_i, \tag{84}$$

The estimand for β_k is

$$\beta_k = \frac{Cov(Y_i, e_{ki})}{Var(e_{ki})}, \tag{85}$$

Where e_{ki} is the residual from a regression of X_{ki} on the remaining X_{ji}'s.

OLS estimator (cont.)

Let's check this using the job training data.

Type

```
reg1 <- lm(re78 ~ educ + age, data = jt)
reg2 <- lm(educ ~ age, data = jt)
resid <- resid(reg2)
reg3 <- lm(re78 ~ resid, data = jt)

stargazer(reg1, reg2, reg3, type="text")
```

Omitted-variable bias

Suppose that the variable X_i is omitted from the above equation and you estimate

$$Y_i = \alpha^s + \beta^s D_i + \varepsilon_i^s. \tag{86}$$

The estimate of β^s is most likely biased.

- This is called the omitted variable bias.

But in which direction?

Omitted-variable bias (cont.)

We already know that the coefficient β^s is written by

$$\beta^s = \frac{Cov(Y_i, D_i)}{Var(D_i)}. \tag{87}$$

Substituting (81) into Y_i yields

$$\begin{aligned} \beta^s &= \frac{Cov(\alpha + \beta D_i + \lambda X_i + \varepsilon_i, D_i)}{Var(D_i)} \\ &= \beta + \lambda \frac{Cov(X_i, D_i)}{Var(D_i)} \\ &= \beta + \lambda \pi_1, \end{aligned} \tag{88}$$

Where the last equation uses the coefficient from the regression of X_i on D_i

$$X_i = \pi_0 + \pi_1 D_i + u_i. \tag{89}$$

(88) is called the omitted-variable bias (OVB) formula.

Omitted-variable bias (cont.)

The omitted variable is most likely unobservable to econometricians.

However, you can make an educated guess on the direction of bias using the OVB formula. What you need to know is the signs of λ and π_1 in

$$\beta^s = \beta + \lambda \pi_1.$$

Discuss the potential direction of bias in the college-earning example.

Omitted-variable bias (cont.)

Consider a regression

$$Y_i = \alpha + \beta X_{1i} + \lambda X_{2i} + \varepsilon_i, \tag{90}$$

Where Y_i is earnings, X_{1i} is education, and X_{2i} is age.

If you do not have data on age and the variable is omitted, what could be the potential bias for β?

OLS estimator (cont.)

Let's check your prediction using the job training data.

Type

```
reg1 <- lm(re78 ~ educ, data = jt)
reg2 <- lm(re78 ~ educ + age, data = jt)

stargazer(reg1, reg2, type="text")
```

Is your prediction correct?

Bad control

Notice that X_i can be a good control or a bad control.

- Bad controls are the outcomes of the experiment
- Good controls are measured before the experiment, and do not vary during the experiment (e.g., sex, parents' education)

You *should not* use a bad control unless necessary.

Bad control: An example

Consider a regression

$$Y_i = \alpha + \beta D_i + \lambda X_i + \varepsilon_i, \tag{91}$$

Where Y_i is earnings, D_i is going to college, and X_i is occupation (e.g., blue collar and white collar jobs).

- Thus, with this regression, we estimate the effect of going to college on earnings, conditional on occupation.

However, including X_i in the above regression does not provide a causal effect (even if D_i is randomly assigned). .

Let's see why.

Bad control: An example (cont.)

Consider an example

	$Y_i(0)$	$Y_i(1)$	$X_i(0)$	$X_i(1)$
Person A	600	650	B	B
Person B	700	850	B	W
Person C	825	800	W	W

Where B means blue collar, while W stands for white collar.

ATE is about 58.

- Compute it by yourself.

However, including occupation makes it zero, because within-group comparison gives no difference between the treatment and control groups (i.e., $Y_i(1) - Y_i(0) = 0$) for white and blue collars.

- Check it by yourself. Recall the matching estimator.

Measurement error

Consider the following model

$$Y_i = \alpha + \beta X_i^* + \varepsilon_i, \tag{92}$$

But you cannot directly observe X_i^*.

Assume that the variable X_i^* has measurement error

$$e_i = X_i - X_i^*, \tag{93}$$

Such that $E(e_i) = 0$ and you only observe X_i.

Then, what you estimate is

$$Y_i = \alpha + \beta X_i + u_i, \tag{94}$$

Where $u_i = \varepsilon_i - \beta e_i$.

Measurement error (cont.)

Consider two extreme cases.

Assumption 1: e_i and observed X_i are not correlated.

$$Cov(X_i, e_i) = 0. \tag{95}$$

Assumption 2: e_i and unobserved X_i^* are not correlated

$$Cov(X_i^*, e_i) = 0. \tag{96}$$

With Assumption 1, one can still get an unbiased estimate.

With Assumption 2, we get a biased estimate.

Measurement error (cont.)

Suppose that Assumption 2 holds.

The regression coefficient is

$$\begin{aligned}
\frac{Cov(Y_i, X_i)}{Var(X_i)} &= \frac{Cov(\alpha + \beta X_i + (\varepsilon_i - \beta e_i), X_i)}{Var(X_i)} \\
&= \beta - \beta \frac{Cov(e_i, X_i)}{Var(X_i)} \\
&= \beta \left(1 - \frac{Cov(e_i, X_i)}{Var(X_i)}\right) \\
&= \beta \left(\frac{\sigma^2_{x^*}}{\sigma^2_{x^*} + \sigma^2_e}\right) \qquad (97)
\end{aligned}$$

Where the derivation of the final equation uses
$Cov(e_i, X_i) = Var(e_i) = \sigma^2_e$ and
$Var(X_i) = Var(X^*_i) + Var(e_i) = \sigma^2_{x^*} + \sigma^2_e$.

Measurement error (cont.)

Since the value inside parentheses in (97) is always less than one, the regression coefficient is biased towards zero.

This is called the attenuation bias.

- Can you see that measurement error is a type of omitted-variable bias? (See, e.g., (88).)

How to cope with endogeneity issue

Conduct a randomized experiment

Find an instrument (IV)

- IV is a textbook method for alleviating measurement error, OVB, and simultaneity.

Find a discontinuity (RDD)

Find a change in slope (DID)

Carefully add control variables + α

Always

- Be aware of bad controls!
- Discuss the direction of potential bias!

Summary

Treatment effect

Average treatment effect (ATE)

- Independence assumption
- Average Effect of Treatment on the Treated (ATT)

Is the effect statistically significant?

- Statistical significance
- Confidence interval

Is the effect large?

- Effect size

Conducting an experiment

Power calculation, MDE

Regression analysis

- Conditional expectation function
- Conditional independence assumption (CIA)
- Regression and matching
- Omitted-variable bias
- Bad control
- Measurement error