PUBL0055: Introduction to Quantitative Methods

Lecture 6: Regression III & Causality II

Michal Ovádek & Indraneel Sircar

Academic year 2024-25

Motivation

Does health insurance improve health outcomes? (Revisited)

Is there a *causal* effect of health insurance on actual levels of health? In week 2, we used this example to show that randomized experiments represent a "gold standard" for causal inference, and that drawing causal conclusions from observational data is complicated by "confounding" relationships.

- Y (Dependent variable): *health*
  - "Would you say your health in general is excellent (5), very good (4), good (3), fair (2), or poor (1)?"
- X (Independent variable): *insured*
  - "Do you have health insurance?" `TRUE` = Insured, `FALSE` = Not insured

Data sources

Observational data

National Health Interview Survey (NHIS, N = 19996): an annual survey of the US population that asks questions about health and health insurance.

Experimental data

RAND Health Insurance Experiment (RAND, N = 2702): an experiment conducted between 1974 and 1982 in the US. In this experiment, researchers *randomly allocated* individuals to receive health insurance.

In both cases we also have information from some of the other questions on the survey (gender, income, race, etc).

Lecture Outline

Regression and randomized experiments

Regression and observational data

Conclusion

Regression and randomized experiments

Randomization and regression

- Revision (1): The difference in means provides an unbiased estimate of the causal effect when our treatment is randomly assigned to units (week 2).
- Revision (2): The coefficient associated with a binary variable in a simple linear regression is equal to the difference in means estimate (week 4).

Randomization and regression

- Revision (1): The difference in means provides an unbiased estimate of the causal effect when our treatment is randomly assigned to units (week 2).
- Revision (2): The coefficient associated with a binary variable in a simple linear regression is equal to the difference in means estimate (week 4).
- Implication: When treatment is randomized, the linear regression coefficient provides an unbiased estimate of the causal effect!

Regression and randomized experiments (example)

Let's calculate the difference in means using the experimental data:

```r
## Mean health level for insured and uninsured individuals
mean_health_insured <- mean(rand$health[rand$insured == TRUE])
mean_health_uninsured <- mean(rand$health[rand$insured == FALSE])
mean_health_insured - mean_health_uninsured
```

```
## [1] -0.01895885
```

Regression and randomized experiments (example)

Let's calculate the difference in means using the experimental data:

```
## Mean health level for insured and uninsured individuals
mean_health_insured <- mean(rand$health[rand$insured == TRUE])
mean_health_uninsured <- mean(rand$health[rand$insured == FALSE])
mean_health_insured - mean_health_uninsured
```

```
## [1] -0.01895885
```

```
## Regression of health on insurance status
lm(health ~ insured, rand)
```

```
...
## (Intercept)  insuredTRUE
##     3.40702     -0.01896
...
```

Implication: The causal effect of insurance on health is very close to zero.

Benefits of using regression to analyse experiments

1. **Heterogeneous treatment effects**
   - Do the effects of the treatment vary by type of unit?
   - You can already do this: interactions!

Benefits of using regression to analyse experiments

1. **Heterogeneous treatment effects**
   - Do the effects of the treatment vary by type of unit?
   - You can already do this: interactions!

2. **Non-binary treatments**
   - Is the treatment you care about continuous? Or categorical?
   - You can already do this: factor/continuous variables in regression!

Benefits of using regression to analyse experiments

1. **Heterogeneous treatment effects**
   - Do the effects of the treatment vary by type of unit?
   - You can already do this: interactions!

2. **Non-binary treatments**
   - Is the treatment you care about continuous? Or categorical?
   - You can already do this: factor/continuous variables in regression!

3. **Increasing the "precision" of our estimates**
   - Control for other factors that determine the outcome can make the estimates of the treatment effects more precise
   - We will cover this in future weeks!

Heterogeneous treatment effects

In week 2, we were concerned with estimating the **average treatment effect**

- What is the average difference in potential outcomes under treatment and control *across all units* in our sample?

Heterogeneous treatment effects

In week 2, we were concerned with estimating the **average treatment effect**

- What is the average difference in potential outcomes under treatment and control *across all units* in our sample?

However, we may care about whether the treatment has **different** effects for different types of individuals.

- Does insurance status matter more for low income than high income individuals?

Heterogeneous treatment effects

In week 2, we were concerned with estimating the **average treatment effect**

- What is the average difference in potential outcomes under treatment and control *across all units* in our sample?

However, we may care about whether the treatment has **different** effects for different types of individuals.

- Does insurance status matter more for low income than high income individuals?

Fortunately, we can use **interactions between explanatory variables** to answer this.

Interactions (revision)

- An **interaction** exists when the relationship between one explanatory variable (X_1) and our dependent variable (Y) depends on the value of another explanatory variable (X_2)
- We can add interactions into the linear regression model by including the **product** of two explanatory variables in our model:

$$Y_i = \alpha + \beta_1 X1 + \beta_2 X_2 + \beta_3 (X_1 * X_2) + \epsilon_i$$

- We can interpret interactions by calculating the **fitted values** for various cases of interest

Heterogeneous treatment effects

```r
heterogeneous_effects_model <- lm(health ~ insured * income, rand)
summary(heterogeneous_effects_model)
```

```
...
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.195337   0.072245  44.229  < 2e-16 ***
## insuredTRUE         0.098969   0.079757   1.241 0.214754
## income              0.006551   0.001963   3.338 0.000855 ***
## insuredTRUE:income -0.003536   0.002181  -1.621 0.105126
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.767 on 2698 degrees of freedom
## Multiple R-squared:  0.007878,   Adjusted R-squared:  0.006775
## F-statistic: 7.141 on 3 and 2698 DF,  p-value: 8.932e-05
...
```

Heterogeneous treatment effects

```r
# Treatment effect for low-income individuals
predict(heterogeneous_effects_model,
        newdata = data.frame(insured = c(not_insured = F, insured = T),
                             income =  10)) # income in thousands
```

```
## not_insured     insured
##     3.26085     3.32446
```

Heterogeneous treatment effects

```r
# Treatment effect for low-income individuals
predict(heterogeneous_effects_model,
        newdata = data.frame(insured = c(not_insured = F, insured = T),
                             income =  10)) # income in thousands
```

```
## not_insured     insured
##     3.26085     3.32446
```

```r
# Treatment effect for high-income individuals
predict(heterogeneous_effects_model,
        newdata = data.frame(insured = c(not_insured = F, insured = T),
                             income = 40)) # income in thousands
```

```
## not_insured     insured
##    3.457387    3.414922
```

Heterogeneous treatment effects

```r
# Treatment effect for low-income individuals
predict(heterogeneous_effects_model,
        newdata = data.frame(insured = c(not_insured = F, insured = T),
                             income =  10)) # income in thousands
```

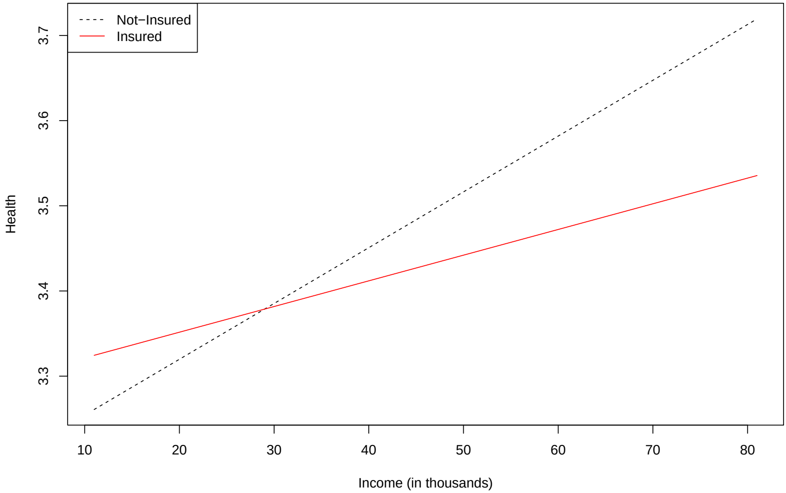
```
## not_insured     insured
##     3.26085     3.32446
```

```r
# Treatment effect for high-income individuals
predict(heterogeneous_effects_model,
        newdata = data.frame(insured = c(not_insured = F, insured = T),
                             income = 40)) # income in thousands
```

```
## not_insured     insured
##    3.457387    3.414922
```

Implication: Some suggestion that the treatment *slightly increases* reported health for low-income individuals, but not for high-income individuals.

Heterogenous treatment effects

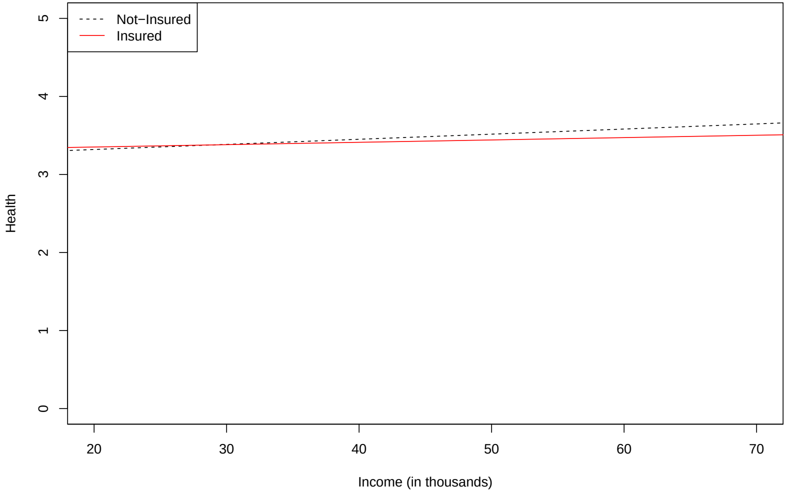


Heterogenous treatment effects

But don’t forget about data range!

Heterogenous treatment effects

But don't forget about data range!



Non-binary treatments

We have generally been focused on treatments that are **binary** – i.e. where units are either assigned to treatment or to control.

However, there are many examples where our (randomly assigned) treatment may not be binary. For instance:

1. Effect of different spending levels on budget approval (continuous)
2. Effect of types of campaign materials on voter turnout (categorical)

In our healthcare example, the treatment in the experiment was actually **categorical**.

Non-binary treatments

```r
table(rand$insured,rand$plantype)
```

```
##
##         Catastrophic Coinsurance Deductible Free
##   FALSE          491           0          0    0
##   TRUE             0         727        593  891
```

- "Catastrophic" – Individuals pay for all health costs
- "Coinsurance" – Individuals pay 25-50% of costs
- "Deductible" – Costs capped at $150
- "Free" – Individuals pay nothing

Question: Are the causal effects the same for all three treatment conditions?

Non-binary treatments (factor)

```r
categorical_treatment_model <- lm(health ~ plantype, rand)
summary(categorical_treatment_model)
```

```
...
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.407016   0.034684  98.231   <2e-16 ***
## plantypeCoinsurance  0.044112   0.044893   0.983   0.3259
## plantypeDeductible  -0.009908   0.046893  -0.211   0.8327
## plantypeFree        -0.076444   0.043196  -1.770   0.0769 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7685 on 2698 degrees of freedom
## Multiple R-squared:  0.003769,   Adjusted R-squared:  0.002661
## F-statistic: 3.403 on 3 and 2698 DF,  p-value: 0.01701
...
```

Non-binary treatments (factor)

```r
# Treatment effects for different plans
predict(
  categorical_treatment_model,
  newdata = data.frame(
    plantype = c("Catastrophic" = "Catastrophic",
                 "Coinsurance" = "Coinsurance",
                 "Deductible" = "Deductible",
                 "Free" = "Free")))
```

```
## Catastrophic  Coinsurance   Deductible         Free
##     3.407016     3.451128     3.397108     3.330572
```

Non-binary treatments (factor)

```
# Treatment effects for different plans
predict(
  categorical_treatment_model,
  newdata = data.frame(
    plantype = c("Catastrophic" = "Catastrophic",
                 "Coinsurance" = "Coinsurance",
                 "Deductible" = "Deductible",
                 "Free" = "Free")))
```

```
## Catastrophic  Coinsurance   Deductible         Free
##     3.407016     3.451128     3.397108     3.330572
```

Implication: The average outcome is similar for all three treatment groups, which are similar to the control group.

Randomization and regression

Using regression to analyse the RAND health experiment, we have seen that:

1. The average causal effect of health insurance on health outcomes is very small
2. The causal effect is only positive for low income individuals, though it is still small
3. There is little difference in the estimated causal effects of different types of insurance plans

Sidenote: Despite these modest effects, the RAND experiment showed much larger effects in terms of use of health care services. Self-reported health may not be the most important health care outcome!

Regression and observational data

What happens when we don't have experimental data?

So far, we have focused on how we can use regression to also analyse experimental data.

But most data we have in political science is not experimental, it is observational.

Regression also provides a framework for controlling for confounding factors that arise in observational data.

Confounding recap

Confounding

Confounding exists when there are differences **other than the treatment** between treatment and control groups (i.e., observations taking on different values of our independent variable of interest), **and** which may affect the outcome.

Confounding recap

Confounding

Confounding exists when there are differences **other than the treatment** between treatment and control groups (i.e., observations taking on different values of our independent variable of interest), **and** which may affect the outcome.

Imagine that we observe a positive (bivariate) relationship between the following variables:

- National chocolate consumption → Number of Nobel prizes

Question: Can you think of any potentially confounding variables?

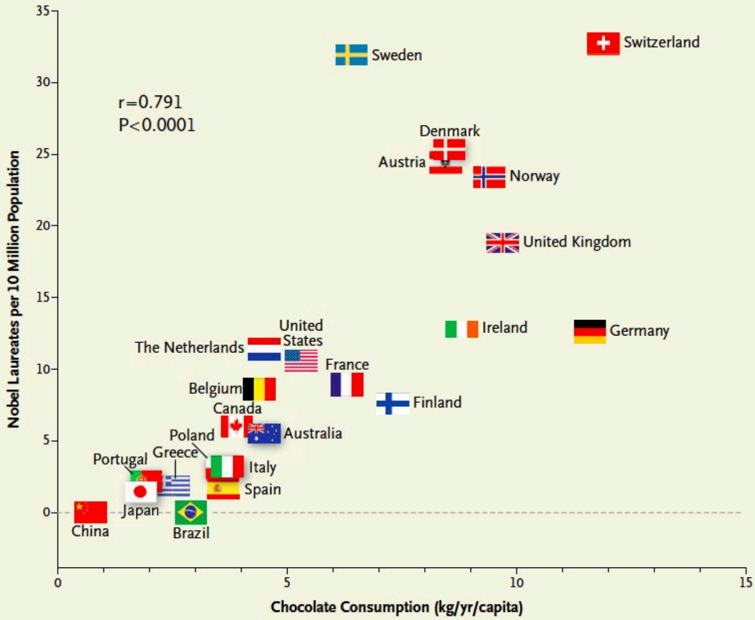


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Multiple Linear Regression & OVB

In the context of regression, failing to adjust for potentially confounding factors is usually known as omitted variable bias (OVB).

Multiple Linear Regression & OVB

In the context of regression, failing to adjust for potentially confounding factors is usually known as omitted variable bias (OVB).

- OVB occurs when **two** conditions are met:
  1. When our X variable (X_1) is correlated with another X variable (X_2) that has not been included in the analysis
  2. When the omitted variable (X_2) is also correlated with our Y variable
- We call an omitted variable that leads to such bias a confounding variable

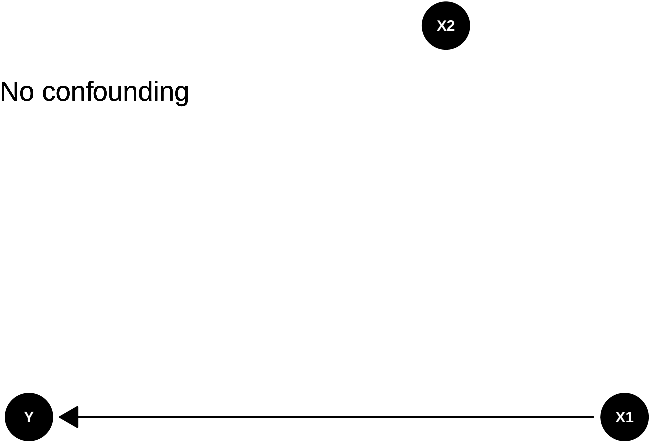
Multiple Linear Regression & OVB

In the context of regression, failing to adjust for potentially confounding factors is usually known as omitted variable bias (OVB).

- OVB occurs when **two** conditions are met:
  1. When our X variable (X_1) is correlated with another X variable (X_2) that has not been included in the analysis
  2. When the omitted variable (X_2) is also correlated with our Y variable
- We call an omitted variable that leads to such bias a confounding variable

Intuition: The association that we observe between X_1 and Y might be more meaningfully attributed to X_2

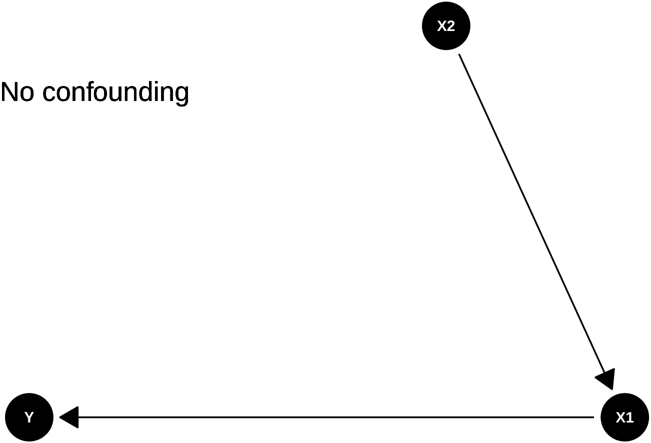
Multiple Linear Regression & OVB



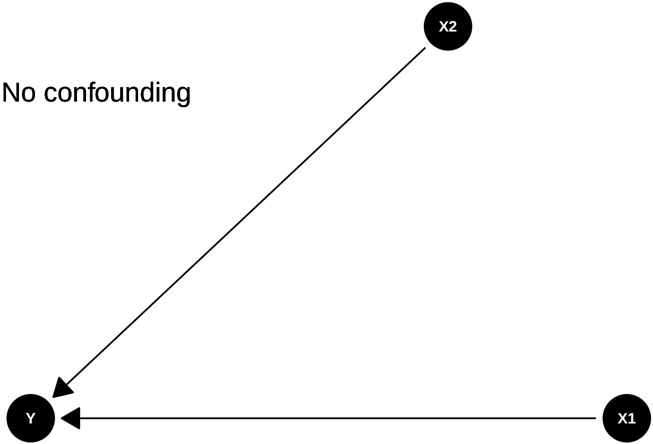
No confounding

Multiple Linear Regression & OVB

No confounding

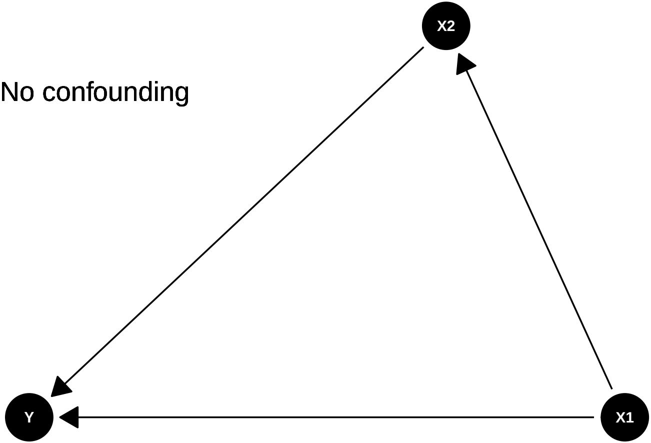


Multiple Linear Regression & OVB

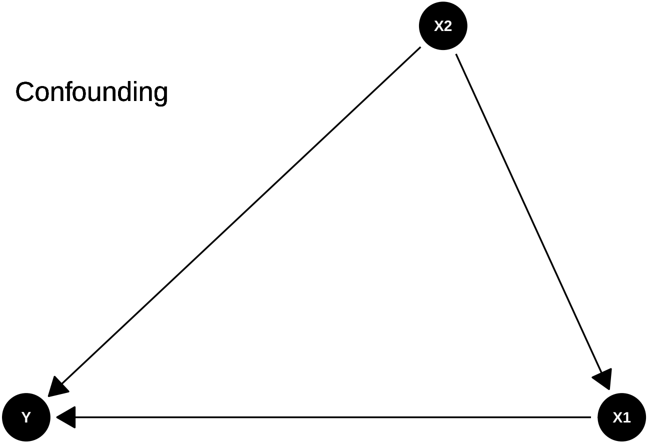


Multiple Linear Regression & OVB

No confounding

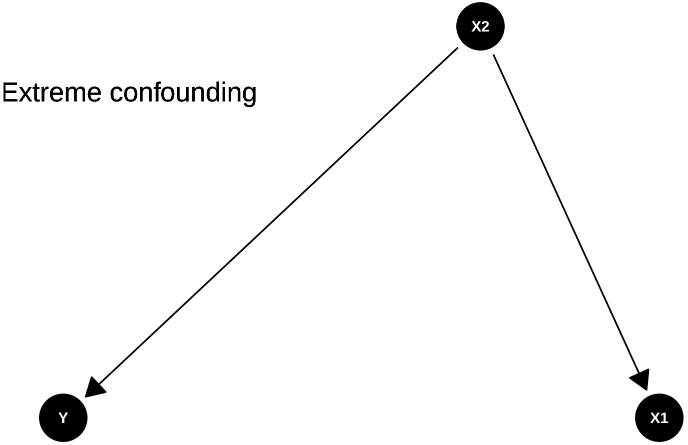


Multiple Linear Regression & OVB



Confounding

Multiple Linear Regression & OVB



Extreme confounding

Multiple Linear Regression & OVB

If OVB is present, and we do nothing to address it, then our estimate of β_1 will be biased (wrong)

	$cor(X_1, X_2) > 0$	$cor(X_1, X_2) < 0$
$cor(X_2, Y) > 0$	$\hat{\beta}_1$ too big	$\hat{\beta}_1$ too small
$cor(X_2, Y) < 0$	$\hat{\beta}_1$ too small	$\hat{\beta}_1$ too big

- $cor(X_1, X_2)$ is the correlation between X_1 and X_2
- $cor(X_2, Y)$ is the correlation between X_2 and Y

Multiple Linear Regression & OVB

If OVB is present, and we do nothing to address it, then our estimate of β_1 will be biased (wrong)

	$cor(X_1, X_2) > 0$	$cor(X_1, X_2) < 0$
$cor(X_2, Y) > 0$	$\hat{\beta}_1$ too big	$\hat{\beta}_1$ too small
$cor(X_2, Y) < 0$	$\hat{\beta}_1$ too small	$\hat{\beta}_1$ too big

- $cor(X_1, X_2)$ is the correlation between X_1 and X_2
- $cor(X_2, Y)$ is the correlation between X_2 and Y

Implication: Depending on the relationship that X_2 has with X_1 and Y, $\hat{\beta}_1$ could be either too big or too small!

Controlling for confounders of health insurance

Omitted variable bias is a problem when...

1. The independent variable is correlated with other factors

and

2. The dependent variable is also correlated with those factors

Controlling for confounders of health insurance

Omitted variable bias is a problem when...

1. The independent variable is correlated with other factors

and

2. The dependent variable is also correlated with those factors

Let's look into whether confounding could be an issue in the NHIS data set:

Controlling for confounders of health insurance

Omitted variable bias is a problem when...

1. The independent variable is correlated with other factors

and

2. The dependent variable is also correlated with those factors

Let's look into whether confounding could be an issue in the NHIS data set:

- Is the **independent** variable correlated with other factors?

insured	health	age	female	years_educ	income
Insured	3.9	43.3	50.2	14.1	101.3
Uninsured	3.6	40.9	49.0	11.3	42.9
Difference	0.3	2.4	1.2	2.9	58.4

Controlling for confounders of health insurance

Omitted variable bias is a problem when...

1. The independent variable is correlated with other factors

and

2. The dependent variable is also correlated with those factors

Let's look into whether confounding could be an issue in the NHIS data set:

- Is the **dependent** variable correlated with those factors?

	age	female	years_educ	income
$corr$(health,X)	-0.162	-0.001	0.255	0.269

Controlling for confounders of health insurance

Omitted variable bias is a problem when...

1. The independent variable is correlated with other factors

and

2. The dependent variable is also correlated with those factors

Both conditions are met in this example.

Question: What can we do about it?

"Controlling for" confounders

One strategy for dealing with confounding is to compare average outcomes between treatment and control units while **controlling** for potential confounders.

Question: How can we ***control*** for potential confounders?

“Controlling for” confounders

One strategy for dealing with confounding is to compare average outcomes between treatment and control units while **controlling** for potential confounders.

Question: How can we ***control*** for potential confounders?

Answer 1: By using subclassification

“Controlling for” confounders

One strategy for dealing with confounding is to compare average outcomes between treatment and control units while **controlling** for potential confounders.

Question: How can we ***control*** for potential confounders?

Answer 1: By using subclassification

Answer 2: By using regression

Subclassification

What is subclassification and when can we use it?

- Cross-sectional data: One observation per unit (i.e. individuals, countries, firms, etc), many units
- Approach: Compare average outcomes between treatment and control units, "controlling" for potential confounders
- Assumption required: No *unmeasured* confounding between treated and control units
- Example: Insurance status is imbalanced with respect to income. We can control for income by using **subclassification**: calculating the differences between treatment and control *within levels of a confounding variable* (e.g., income levels).

Subclassification (example)

Our observational **difference-in-means** estimate is relatively large:

```r
mean(nhis$health[nhis$insured == T]) -
  mean(nhis$health[nhis$insured == F])
```

```
## [1] 0.3262003
```

But **insurance status is imbalanced** with respect to income:

```r
mean(nhis$income[nhis$insured == T]) -
  mean(nhis$income[nhis$insured == F])
```

```
## [1] 58.42287
```

How can we **control** for income?

Subclassification (example)

Subclassification: calculate differences between treatment and control *within levels of a confounding variable*.

Imagine that we have just 3 levels of income (low, mid and high).

Calculate the average health of insured and uninsured for each level (using the NHIS data set):

```r
insured_low_mean <- mean(nhis$health[nhis$insured == T &
                              nhis$income_cat == "Low"])
uninsured_low_mean <- mean(nhis$health[nhis$insured == F &
                              nhis$income_cat == "Low"])

insured_low_mean - uninsured_low_mean
```

```
## [1] 0.1010293
```

Subclassification

Subclassification: calculate differences between treatment and control *within levels of a confounding variable*.

Imagine that we have just 3 levels of income (low, mid and high).

Calculate the average health of insured and uninsured for each level (using the NHIS data set):

```
insured_mid_mean <- mean(nhis$health[nhis$insured == T &
                               nhis$income_cat == "Mid"])
uninsured_mid_mean <- mean(nhis$health[nhis$insured == F &
                               nhis$income_cat == "Mid"])

insured_mid_mean - uninsured_mid_mean
```

```
## [1] 0.04954519
```

Subclassification

Subclassification: calculate differences between treatment and control *within levels of a confounding variable*.

Imagine that we have just 3 levels of income (low, mid and high).

Calculate the average health of insured and uninsured for each level (using the NHIS data set):

```r
insured_high_mean <- mean(nhis$health[nhis$insured == T &
                              nhis$income_cat == "High"])
uninsured_high_mean <- mean(nhis$health[nhis$insured == F &
                              nhis$income_cat == "High"])

insured_high_mean - uninsured_high_mean
```

```
## [1] 0.1542666
```

Subclassification

```
insured_low_mean - uninsured_low_mean
```

```
## [1] 0.1010293
```

```
insured_mid_mean - uninsured_mid_mean
```

```
## [1] 0.04954519
```

```
insured_high_mean - uninsured_high_mean
```

```
## [1] 0.1542666
```

Subclassification

```r
insured_low_mean - uninsured_low_mean
```

```
## [1] 0.1010293
```

```r
insured_mid_mean - uninsured_mid_mean
```

```
## [1] 0.04954519
```

```r
insured_high_mean - uninsured_high_mean
```

```
## [1] 0.1542666
```

Consequence: Once we control for income, the effects of insurance on health are much smaller than in the naive difference-in-means comparison.

```r
mean(nhis$health[nhis$insured == T]) -
  mean(nhis$health[nhis$insured == F])
```

“Controlling for” confounders with regression

Regression also allows us to control for other variables, and is more flexible than subclassification:

- If we believe that the association between X_1 and Y is confounded by X_2, and we are able to measure X_2, we can control for X_2
  - We can hold the value of X_2 constant while estimating the association between X_1 and Y
- If we believe that the association between X_1 and Y is confounded by other variables, we can also control for those variables!

“Controlling for” confounders with regression

Regression also allows us to control for other variables, and is more flexible than subclassification:

- If we believe that the association between X_1 and Y is confounded by X_2, and we are able to measure X_2, we can control for X_2
  - We can hold the value of X_2 constant while estimating the association between X_1 and Y
- If we believe that the association between X_1 and Y is confounded by other variables, we can also control for those variables!

Intuition: The idea is the same as for subclassification, but regression allows us to calculate differences between treatment and control while *holding multiple variables constant*.

Controlling for confounders of health insurance

```r
# Naive model
nhis_model <- lm(health ~ insured, data = nhis)

# Model controlling for income only
nhis_model_with_income <- lm(health ~ insured + income,
                             data = nhis)

# Model controlling for many variables
nhis_model_with_covariates <- lm(health ~ insured + income +
                                   age + female +
                                   years_educ,
                                 data = nhis)
```

Controlling for confounders of health insurance

	health		
	(1)	(2)	(3)
insured	0.33	0.06	0.02
income		0.004	0.004
age			−0.02
female			−0.05
years_educ			0.05
Constant	3.62	3.43	3.80
Observations	19,996	19,996	19,996
R^2	0.02	0.07	0.13

- In this example, the $\widehat{\beta}$ coefficient on `insured` *decreases* when controlling for other variables
- $\widehat{\beta}_{\text{insured}}$ in model 3 is much closer to the experimental estimate

Controlling for confounders of health insurance

	health		
	(1)	(2)	(3)
insured	0.33	0.06	0.02
income		0.004	0.004
age			−0.02
female			−0.05
years_educ			0.05
Constant	3.62	3.43	3.80
Observations	19,996	19,996	19,996
R^2	0.02	0.07	0.13

Question: Does $\widehat{\beta}_{\text{insured}}$ represent the *causal* effect of insurance on self-reported health?

Controlling for confounders of health insurance

	health		
	(1)	(2)	(3)
insured	0.33	0.06	0.02
income		0.004	0.004
age			−0.02
female			−0.05
years_educ			0.05
Constant	3.62	3.43	3.80
Observations	19,996	19,996	19,996
R^2	0.02	0.07	0.13

Question: Does $\widehat{\beta}_{\text{insured}}$ represent the *causal* effect of insurance on self-reported health?

Answer: *Only if we are willing to assume that we are controlling for* ***all*** *confounders/omitted variables.*

Task

Is the assumption that you controlled for **all** possible confounding variables plausible in the case of the multiple linear regression using (cross-sectional) NHIS data?

Speak to your neighbour about this question and which variables you would should control for in your regression, if you wanted to make the case that we can use multiple linear regression to estimate the causal effect of health insurance on self-reported health using the NHIS data set.

Controlling for confounders of health insurance

	health		
	(1)	(2)	(3)
insured	0.33	0.06	0.02
income		0.004	0.004
age			−0.02
female			−0.05
years_educ			0.05
Constant	3.62	3.43	3.80
Observations	19,996	19,996	19,996
R^2	0.02	0.07	0.13

Example: One potential missing control variable here is baseline health levels. People who were less healthy in the past may be less healthy now *and* more likely to take out insurance!

Controlling for confounders of health insurance (experimental data)

Note that the same does not apply with experimental data!

	health		
	(1)	(2)	(3)
insured	−0.02	−0.01	−0.01
income		0.004	0.003
age			−0.01
female			−0.03
years_educ			0.05
Constant	3.41	3.29	3.19
Observations	2,702	2,702	2,702
R^2	0.0001	0.01	0.09

- The coefficient for `insured` is nearly the same in models 1, 2 and 3. Why?

Controlling for confounders of health insurance (experimental data)

Note that the same does not apply with experimental data!

	health		
	(1)	(2)	(3)
insured	−0.02	−0.01	−0.01
income		0.004	0.003
age			−0.01
female			−0.03
years_educ			0.05
Constant	3.41	3.29	3.19
Observations	2,702	2,702	2,702
R^2	0.0001	0.01	0.09

- The coefficient for `insured` is nearly the same in models 1, 2 and 3. Why?
- OVB is present when omitted variables are correlated with **both** independent and dependent variables.
- Insurance status is randomly assigned, so cannot be correlated with other factors.

Controlling for confounders with regression

Confounders, control, and causal inference

In order to claim that estimates based on multiple linear regressions using observational data represent **causal** differences, you have to argue that you have controlled for **all** possible confounding variables.

This is difficult because you may not:

- *know* what all the confounders are
- be able to *measure* some confounders
- be able to *observe* some confounders

Selection on observables

Where the assumption that you can control for **all** possible confounding variables is plausible, you can use multiple linear regression to estimate causal effects from observational data.

- We can refer to such cases as cases where “**selection on observables**” is possible
- That is, because in these cases, the treatment and control groups only differ by a set of *observable* characteristics.

Conclusion

Regression and causality

When can we interpret a regression coefficient causally?

1. Randomized experiments
   - Coefficient on a binary treatment is estimate of the average treatment effect

Regression and causality

When can we interpret a regression coefficient causally?

1. Randomized experiments
   - Coefficient on a binary treatment is estimate of the average treatment effect

2. Observational studies
   - Confounders: variables that cause both the treatment and the outcome
   - We can only interpret coefficients causally when we have controlled for **all confounders** as additional X variables (cross-sectional design, selection on observables, "controlling for all potential confounders")

Regression and causality

When can we interpret a regression coefficient causally?

1. Randomized experiments
   - Coefficient on a binary treatment is estimate of the average treatment effect

2. Observational studies
   - Confounders: variables that cause both the treatment and the outcome
   - We can only interpret coefficients causally when we have controlled for **all confounders** as additional X variables (cross-sectional design, selection on observables, "controlling for all potential confounders")
   - or when we employ alternative designs, as we'll see next week

Seminar

In seminars this week, you will learn how to ...

1. ...implement more regressions.
2. ...argue about whether regression coefficients are good estimates of causal effects.