Week 14 Live Session - Selected Answers

w203 Instructional Team

April 17, 2017

Announcements

Causal analysis

So far, we have been interpreting regressions predictively: given the values of several inputs, the fitted model allows us to predict y, considering the n data points as a simple random sample from a hypothetical infinite "superpopulation" or probability distribution. Then we can make comparisons across different combinations of values for these inputs.

In the context of regressions, predictive inference relates to comparisons between units, whereas causal inference compares different treatments if applied to the same units. Causal interpretations of regression coefficients can only be justified by relying on much stricter assumptions than are needed for predictive inference.

Recall (intuitively) the way we think about causal models.

We believe that the outcome we care about really has causes. If we could just write down all of these (correctly), we would have a causal model:

$$wage = \beta_0 + \beta_1 educ + \beta_2 ability + u$$

Here, u is a truly random error. This model is causal in the sense that if we manipulate educ, increasing it by 1 year, wage will actually go up by β_1.

The central problem of causal inference is that, even though these causes exist, we can't measure all of them. Perhaps we can only measure wage and educ:

$$wage = \beta_0 + \beta_1 educ + w$$

What's the problem with this model? Solving for w, we have,

$$w = \beta_2 ability + u$$

We know u is uncorrelated with educ, but what about ability? People with more ability tend to have more education as well.

$$\begin{aligned} cov(educ, w) &= cov(educ, \beta_2 ability + u) \\ &= \beta_2 cov(educ, ability) > 0 \end{aligned}$$

This is *endogeneity.* OLS regression cannot identify β_1, because ols can only find the line of best fit - but that's not the line we want!

Omitted variable bias

To see what line ols actually gives us, we can start by writing the regression of ability on educ:

$$ability = \gamma_0 + \gamma_1 educ + v,$$

where $E(v) = 0$ and $cov(educ, v) = 0$.

Now we substitute in:

$$\begin{aligned} wage &= \beta_0 + \beta_1 educ + \beta_2 ability + u && (1)\\ &= \beta_0 + \beta_1 educ + \beta_2(\gamma_0 + \gamma_1 educ + v) + u && (2)\\ &= (\beta_0 + \beta_2\gamma_0) + (\beta_1 + \beta_2\gamma_1)educ + (\beta_2 v + u) && (3)\\ & && (4) \end{aligned}$$

Now notice that we fulfilled our moment conditions:

$$E(\beta_2 v + u) = \beta_2 E(v) + E(u) = 0$$

$$cov(educ, \beta_2 v + u) = \beta_2 cov(educ, v) + cov(educ, u) = 0$$

So ols will consistently estimate this new slope, $\beta_1 + \beta_2\gamma_1$. This is the causal effect we want, β_1, plus an extra term, $\beta_2\gamma_1$, which we call *omitted variable bias.*

R Exercise

The file htv.RData contains data from the 1991 National Longitudinal Survey of Youth, provided by Wooldridge. All people in the sample are males age 26 to 34. The data is interesting here, because it includes education (educ), but also a score on an ability test (abil).

We will assume that the true model is,

$$wage = \beta_0 + \beta_1 educ + \beta_2 abil + u$$

Note: One problem with this analysis is that we're not really measuring ability. *abil* is a *proxy* for ability, not ability itself. And there is a lot of evidence to suggest that standardized tests are not a very good proxy. But for now, let's pretend that we really are measuing ability.

Q1. Using R, estimate (1) the true model, and (2) the regression of abil on educ. Write down the expression for what omitted variable bias would be if you couldn't measure abil. Add this omitted variable bias to the coefficient for educ to see what it would be.

Q2. Now confirm your previous result by fitting the model,

$$wage = \alpha_0 + \alpha_1 educ + w$$

. Make sure your coefficient for *educ* corresponds to what you computed in Q1.

Q3. What does the direction of omitted variable bias suggest about ols estimates of returns to education? What does this suggest about the reported statistical significance of education?

```r
#setwd("/Users/js/Google Drive/_UCB_W203/W203_Fall2016/LiveSession/week_14")
load("htv.RData")

model_true = lm(wage ~ educ + abil, data = data)
model_true
```

```
##
## Call:
## lm(formula = wage ~ educ + abil, data = data)
##
## Coefficients:
## (Intercept)         educ         abil
##     -2.5226       1.1530       0.4333
```

```r
first_stage = lm(abil ~ educ, data = data)
first_stage
```

```
##
## Call:
## lm(formula = abil ~ educ, data = data)
##
## Coefficients:
## (Intercept)         educ
##     -5.3890       0.5512
```

```r
# Omitted variable bias
0.4333 *  0.5512
```

```
## [1] 0.238835
```

```r
# From the above coefficients, we can predict the following
# coefficient for educ:
0.4333 *  0.5512 + 1.1530
```

```
## [1] 1.391835
```

```r
# This is confirmed below:
model_rest = lm(wage ~ educ, data = data)
model_rest
```

```
##
## Call:
## lm(formula = wage ~ educ, data = data)
##
## Coefficients:
## (Intercept)         educ
##      -4.857        1.392
```

The Direction of Omitted Variable Bias

Q4. For each of the following regressions, use your background knowledge to estimate whether omitted variable bias will drive your slope coefficient towards zero or away from zero.

1. $grade = \beta_0 + \beta_1 attendance + u$, omitted: time studying
2. $lifespan = \beta_0 + \beta_1 cigarettes + u$, omitted: exercise
3. $lifespan = \beta_0 + \beta_1 cigarettes + u$, omitted: time socializing

4. $wage = \beta_0 + \beta_1 grad\_education + u$, omitted: experience
5. $wage = \beta_0 + \beta_1 grad\_education + u$, omitted: desire to effect social good

The importance of Causal Inference

A common story is that managers will say that an associative model is fine, then make decisions based on regression coefficients.

A story from David Brockman, one of the instructors for *w241 Causal Inference*:

> A friend who helps write one of the important apps for a Very Large Tech Firm told me that last year his team figured out that a great predictor of how many times you would use the app in the future is how much time you spent in the app the first time. So now they've been using "how many minutes spent during first open" as a key goal metric. I asked him what kind of things increase that metric, and he said a lot of is counterintuitive - things that, to his mind, actually make the app worse and more confusing, so it takes users longer to figure them out. But, he said, we know from the data that this is the kind of thing that users like and keeps them coming back. I asked if he had ever examined the long-term goals of keeping people around, and he said no, the data science team assured them the correlation between that and time-in-app-first-time was still there.

Q5. Provide an example from your industry or area of expertise, in which the causal effect matters, and associative modeling is not enough.

Review: Adding Variables to a Regression.

Suppose you are modeling returns to education. Thinking about (1) causal pathways and (2) the equation for variance of ols coefficients,

$$var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$

Explain how each of the following scenarios could occur:

1. You add a new variable to your regression and the standard error for educ goes down.

The new variable explains variation, decreasing the numerator of the variance equation.

2. You add a new variable to your regression and the standard error for educ goes up.

The new variable is colinear with educ, increasing its VIF. (increasing R_j^2 so the denominator decreases)

3. You add a new variable to your regression and your coefficient gets closer to the true causal effect.

You add an omitted var, or a good proxy for an omitted var, like ability.

4. You add a new variable to your regression, and your coefficient gets further from the true causal effect.

You add a variable that's really an effect of education, like starting wage.

Recognizing Different Types of Studies

It's important to be aware of the sheer variety of scientific evidence. Often, different study types will be depicted as a hierarchy of evidence. One example is provided below. As you move upwards towards randomized control trials, you can think of each study type as controlling for more omitted variables, uncovering the true causal effect.

Next week, we will take the final step to discuss meta analysis. As you will see, this relates more to publication bias and other issues that arise when we look at larger bodies of research.

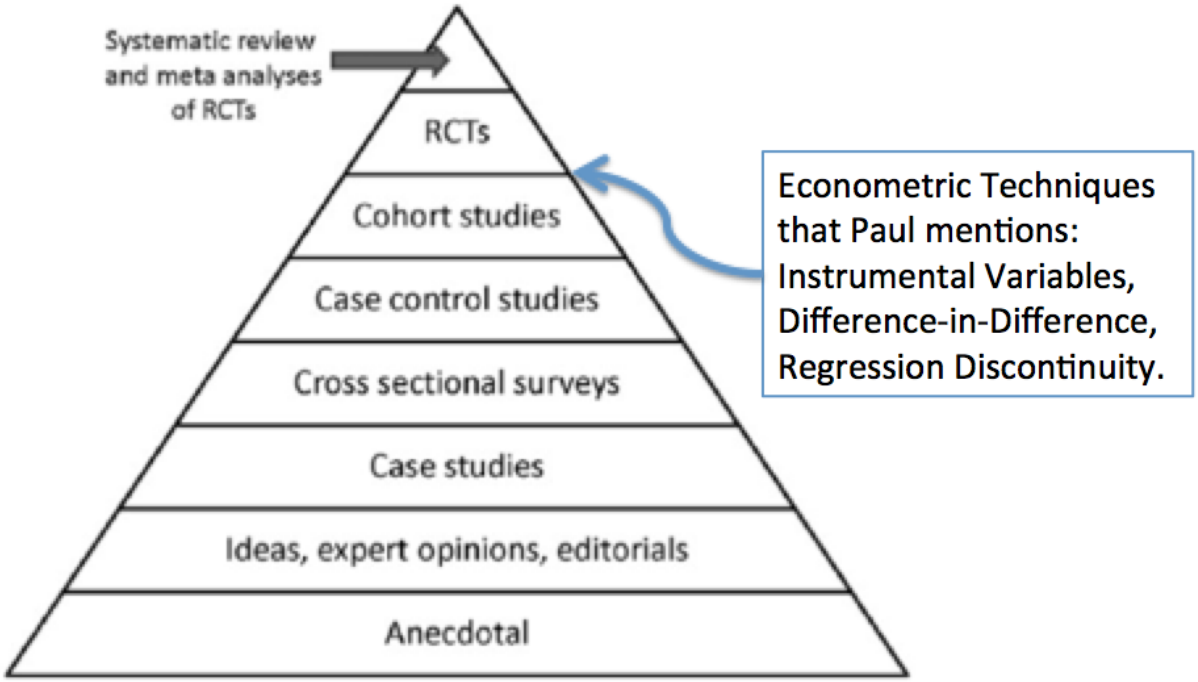

Systematic review and meta analyses of RCTs
RCTs
Cohort studies
Case control studies
Cross sectional surveys
Case studies
Ideas, expert opinions, editorials
Anecdotal
Econometric Techniques that Paul mentions: Instrumental Variables, Difference-in-Difference, Regression Discontinuity.

Figure 1: Hierarchy of Evidence.

Randomized Controlled Trials

Causal inference, which concerns what would happen to an outcome y as a result of a hypothesized “treatment” or intervention, in a regression framework, the treatment can be written as a variable T:

$$T = \begin{cases} 1, & \text{if unit i receives the "treatment"} \\ 0, & \text{if unit i receives the "control,"} \end{cases}$$

or, for a continuous treatment, Ti = level of the “treatment” assigned to unit i.

Recall from your reading, i.e. Angrist and Pischke “Mastering Metrics” Chapter 1 that in the investigation of the average causal effect of insurance naturally begins by comparing the average health of groups of insured and uninsured people, Difference in group means = Average causal effect + Selection bias.

Experimental random assignment is one method of eliminating selection bias. The logistics of a randomized experiment, sometimes called a randomized trial, may seem simple, but using it effectively can be a complex process. In a randomized experiment, participants in various groups should not differ in any systematic way. In a clinical trial, if treatment groups are systematically different, trial results will be biased.

Case-control study

Case-control studies start with the identification of a group of cases (individuals with a particular health outcome) in a given population and a group of controls (individuals without the health outcome) to be included in the study.A case-control study involves the identification of individuals with (‘cases’) and without ('controls') a particular disease or condition. The prevalence (or level) of exposure to a factor is then measured in each group. If the prevalence of exposure among cases and controls is different, it is possible to infer that the exposure may be associated with an increased or decreased occurrence of the outcome of interest.

In a case-control study, researchers typically know the proportion of cases and controls who agree to participate and the exposure prevalence of participating cases and controls is known from the data gathered on participants, but the exposure prevalence of cases and controls who did not participate is unknown.

In practice, controls must fulfil all the eligibility criteria defined for the cases apart from those relating to diagnosis of the disease. In case-control studies, controls should represent the population from which the cases are drawn, i.e., they should provide an estimate of the exposure prevalence in the population from which the cases arise. If not, the results of the study are likely to be distorted because of selection bias.

Selection Bias

With true randomization, all participants in the study are given the same opportunity to be allocated or assigned to each of the study groups. But even a perfectly randomized method to allocate participants to the study groups does not protect against selection bias, which can occur both in the way that individuals are accepted or rejected for participation in a trial, and in the way that the interventions are assigned to individuals once they have been accepted into a trial.

Selection bias occurs when the two variables whose association is under study, usually an exposure and a disease outcome, both affect participation in the study. Selection bias can arise when participants enroll in a study, if the variables affect initial participation rates, and it can also arise when participants withdraw from the study, if the variables affect follow-up rates. The association between the exposure and outcome must be measured among participants, so is effectively conditioned on participation.

Conditioning an estimate of association on participation can induce an association between exposure and disease, when no association exists among all those eligible to participate. Researchers can adjust estimates of association measured among participants to account for the bias introduced by conditioning on participation, when participation is affected by both exposure and disease. This adjustment is often difficult since it ideally requires an assessment of the participation proportion among each of the four combinations of exposed (when exposure is dichotomous) and diseased. Often the exposure status and disease status of the nonparticipants will be unknown - their participation is required to ascertain this information.

Does Mobile Phone Use and Risk of Uveal Melanoma?

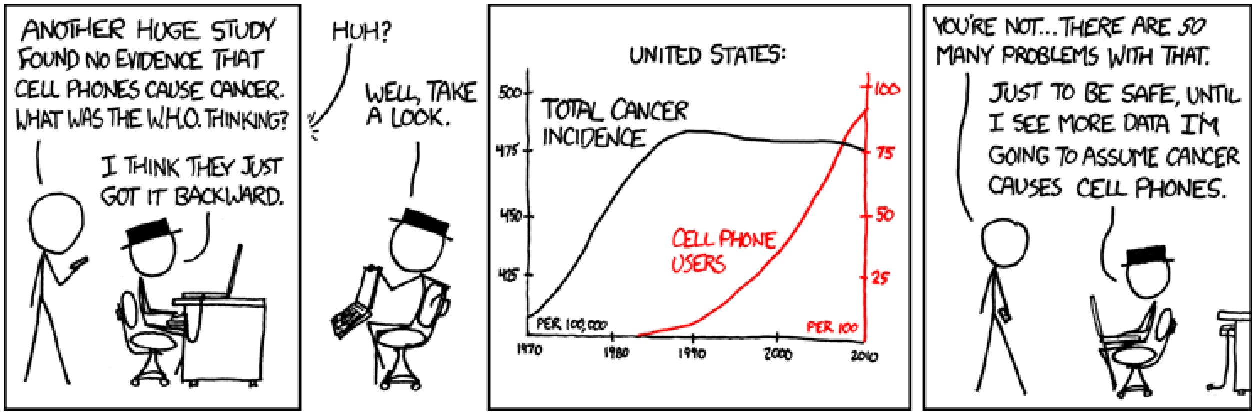

ANOTHER HUGE STUDY FOUND NO EVIDENCE THAT CELL PHONES CAUSE CANCER. WHAT WAS THE W.H.O. THINKING?
I THINK THEY JUST GOT IT BACKWARD.
HUH?
WELL, TAKE A LOOK.
UNITED STATES:
TOTAL CANCER INCIDENCE
CELL PHONE USERS
500
475
450
425
PER 100,000
100
75
50
25
PER 100
1970
1980
1990
2000
2010
YOU'RE NOT... THERE ARE SO MANY PROBLEMS WITH THAT.
JUST TO BE SAFE, UNTIL I SEE MORE DATA I'M GOING TO ASSUME CANCER CAUSES CELL PHONES.

Figure 2: Correlation does not imply causation

In 2009, Stang et al published their findings that an increased risk of uveal melanoma among mobile phone users based on a case-control study that assessed the association between mobile phone use and risk of uveal melanoma (http://jnci.oxfordjournals.org/content/101/2/120.full). In the study, 459 uveal melanoma case patients were recruited at the University of Duisburg-Essen and matched 455 case patients with 827 population control subjects.

The flow chart below illustrates the process of participant enrollment into the study, including the participation frequencies, which are given as the number of cases divided by the number of controls.

Q1.What bias might exist in this study?

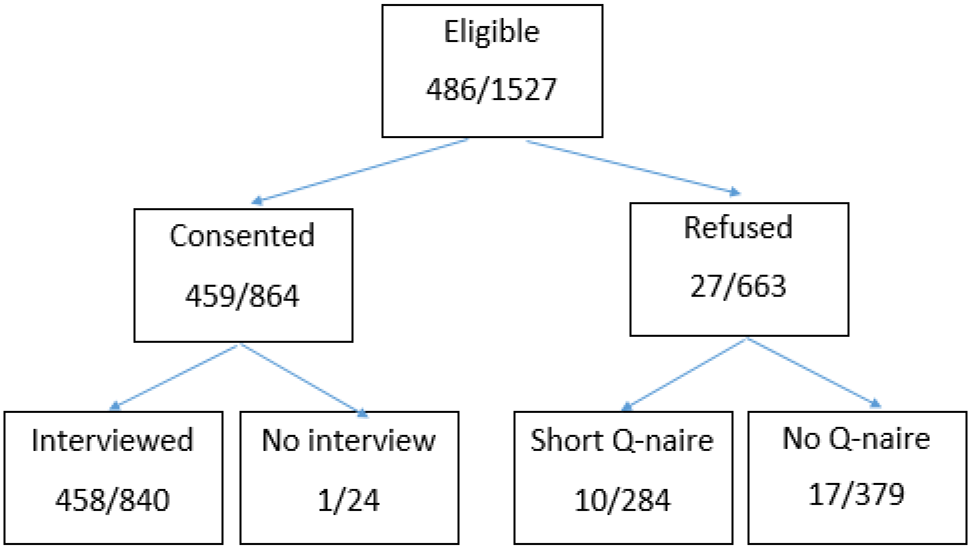
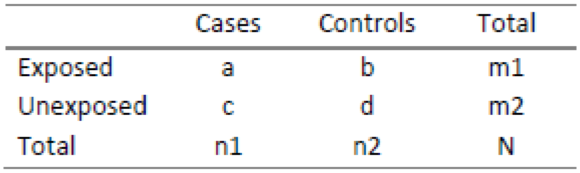


Figure 3: Flow Chart

Odds Ratio

The odds ratio (OR) is used in case-control studies to estimate the strength of the association between exposure and outcome. Note that it is not possible to estimate the incidence of disease from a case control study unless the study is population based and all cases in a defined population are obtained.

The numbers of cases and controls found to have been exposed and not exposed to the factor under investigation can be arranged in a 2 ? 2 table:



	Cases	Controls	Total
Exposed	a	b	m1
Unexposed	c	d	m2
Total	n1	n2	N

Figure 4: 2x2 Table

In case-control studies, it is not possible to directly estimate disease incidence in those exposed and those unexposed, since people are selected on the basis of having or not having the condition of interest, not on the basis of their exposure status. It is, however, possible to calculate the odds of exposure in the cases and in the controls: Odds of exposure:$\frac{Exposed}{Unexposed}$

Odds of exposure in the cases = a/b

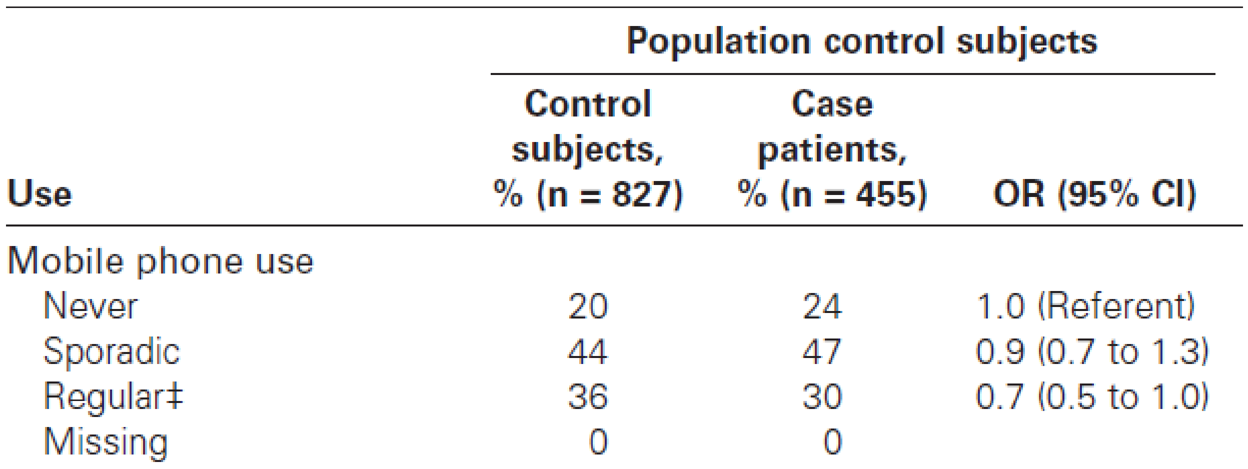
Odds of exposure in the controls = c/d

Exercise

Of these 486 cases, 458 (94%) agreed to participate in the case-control study and completed the interview. There were 1,527 eligible population-based controls, of which 840 (55%) agreed to participate and were interviewed.

The substantial difference in participation rates between cases and controls (94% vs 55%, respectively) suggests taht selection bias may impact the estimate of association. Hence, we need to check whether the ratio estimate of association has to be corrected for selection bias induced by differential participation,

Question: Using the following table, populate the following 2x2 table of the participants from the population control subjects with regular mobile phones or no mobile phone use:



	Population control subjects		
Use	Control subjects, % (n = 827)	Case patients, % (n = 455)	OR (95% CI)
Mobile phone use			
Never	20	24	1.0 (Referent)
Sporadic	44	47	0.9 (0.7 to 1.3)
Regular‡	36	30	0.7 (0.5 to 1.0)
Missing	0	0	

Figure 5: Table 2

```r
#Participants
tab <- as.table(rbind(c("a","b"), c("c","d")))
dimnames(tab) <- list(c("Cases", "Controls"),
 Participants = c("Regular Mobile Use", "No Mobile Use"))
tab
```

```
##          Participants
##           Regular Mobile Use No Mobile Use
##   Cases    a                  b
##   Controls c                  d
```

In this study, the observed odds ratio for the association between regular mobile phone use and non mobile phone use with uveal melanoma incidence is 0.71 [95% CI 0.51-0.97]. Calculate the observed odds ratio for the association between regular mobile phone use and non mobile phone use with uveal melanoma incidence and compare against the reported number.

```r
OR_mobile <- pop_case[1]/pop_control[1]
OR_nomobile <- pop_case[2]/pop_control[2]
OR_mobile
```

```
## [1] 0.4584845
```

```
OR_nomobile
```

```
## [1] 0.6602177
```

```
OR_mobile/OR_nomobile
```

```
## [1] 0.6944444
```

Correcting for Selection Bias

Stang et al. (2009) asked those who refused to participate whether they would answer a short-questionnaire to estimate the prevalence of mobile phone use among nonparticipants. As shown in the flow chart (Figure 2), of the 27 nonparticipating cases, 10 completed the short questionnaire, and 3 of the 10 (30%) reported regular mobile phone use. Of the 663 nonparticipating controls, 284 completed the short questionnaire and 72 (25%) reported regular mobile phone use. Only two categories were available on the short questionnaire, so those who did not report regular mobile phone use were categorized as nonusers.

Just as we created a table for participants, we can create a 2x2 table for non participants and calculate the odds ratio.

```
#Nonparticipants/short questionnaire
#opt_case_no<-27
#opt_control_no<-663
opt_case <- c(3,7)
opt_control <- c(72,212)

tab <- as.table(rbind(opt_case, opt_control))
dimnames(tab) <- list(c("Cases", "Control"),
 NonParticipants = c("Regular Mobile Use", "No Mobile Use"))
tab
```

```
##          NonParticipants
##           Regular Mobile Use No Mobile Use
##   Cases                    3             7
##   Control                 72           212
```

```
OR_opt_mobile <- opt_case[1]/opt_control[1]
OR_opt_nomobile <- opt_case[2]/opt_control[2]
OR_opt_mobile/OR_opt_nomobile
```

```
## [1] 1.261905
```

Among nonparticipants who answered the short questionnaire, the odds ratio equals 1.26, which is in the opposite direction from the odds ratio observed among participants.

This difference illustrates the potential impact of selection bias.