STA305/1004 Class Notes - Week 4

Nathan Taback

January 28, 2016

- Introduction to Causal Inference
    - Hypothetical example of zero causal effect but positive predictive comparison
    - Hypothetical example of positive causal effect but zero positive predictive comparison
    - Adding regression predictors; “omitted” or “lurking” variables
    - Formula for omitted variable bias
- The fundemental Problem of Causal Inference
    - Ways to get around the problem (pg. 171-172, Gelman and Hill, 2007)
- Randomized Experiments
    - Average causal effects and randomized experiments
    - Stable Unit Treatment Value Assumption (SUTVA)
    - Electric company study
- Observational Studies
    - Ignorable Treatment Assignment
- Causal Inference in Observational Studies
    - Defining a treatment variable
    - Cochran’s Basic Advice

Introduction to Causal Inference

Three basic concepts are used to define causal effects (Rubin, 2007).

> A **unit** is a physical object, for example, a patient, at a particular place and point of time, say time t.

> A **treatment** is an action or intervention that can be initiated or withheld from that unit at t (e.g., an anti-hypertensive drug, a statin); if the active treatment is withheld, we will say that the unit has been exposed to the control treatment.

Associated with that unit are two **potential outcomes** at a future point in time, say, $t^* > t$: the value of some outcome measurements Y (e.g., blood pressure, cholesterol level) if the active treatment is given at t and the value of Y at the same future point in time if the control treatment is given at t. The causal effect of the treatment on that unit is defined to be the comparison of the treatment and control potential outcomes at t$_*$ (e.g., their difference, their ratio, the ratio of their squares). The times t can vary from unit to unit in a population of N units, but typically the intervals, $t^* - t$, are constant across the N units.

Gelman and Hill (2007) compare causal inference to predictive inference.

Causal inference 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

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

In the usual regression context, predictive inference relates to comparisons between units, whereas causal inference addresses comparisons of different treatments if applied to the same units. More generally, causal inference can be viewed as a special case of prediction in which the goal is to predict what would have happened under different treatment options. (pg. 167, Gelman and Hill, 2007)

Hypothetical example of zero causal effect but positive predictive comparison

Consider a hypothetical medical experiment in which 100 patients receive the treatment and 100 receive the control condition. In this scenario, the causal effect represents a comparison between what would have happened to a given patient had he or she received the treatment compared to what would have happened under control. We first suppose that the treatment would have no effect on the health status of any given patient, compared with what would have happened under the control. That is, the causal effect of the treatment is zero. However, let us further suppose that treated and control groups systematically differ, with healthier patients receiving the treatment and sicker patients receiving the control. This scenario leads to a positive predictive comparison between the treatment and control groups, even though the causal effect is zero. This sort of discrepancy between the predictive comparison and the causal effect is sometimes called self-selection bias, or simply selection bias, because participants are selecting themselves into different treatments.

Hypothetical example of positive causal effect but zero positive predictive comparison

Conversely, it is possible for a truly nonzero treatment effect to not show up in the predictive comparison. In this scenario, the treatment has a positive effect for all patients, whatever their previous health status. So, for any given unit, we would expect the outcome to be better under treatment than control. However, suppose that this time, sicker patients are given the treatment and healthier patients are assigned to the control condition. It is then possible to see equal average outcomes of patients in the two groups, with sick patients who received the treatment canceling out healthy patients who received the control. Previous health status plays an important role in both these scenarios because it is related both to treatment assignment and future health status. If a causal estimate is desired, simple comparisons of average outcomes across groups that ignore this variable will be misleading because the effect of the treatment will be “confounded” with the effect of previous health status. For this reason, such predictors are sometimes called confounding covariates.

The preceding theoretical examples illustrate how a simple predictive comparison is not necessarily an appropriate estimate of a causal effect. In these simple examples, however, there is a simple solution, which is to compare treated and control units conditional on previous health status. Intuitively, the simplest way to do this is to compare the averages of the current health status measurements across treatment groups only within each previous health status category; we discuss this kind of subclassification strategy later. Another way to estimate the causal effect in this scenario is to regress the outcome on two inputs: the treatment indicator and previous health status. If health status is the only confounding covariate—that is, the only variable that predicts both the treatment and the outcome—and if the regression model is properly specified, then the coefficient of the treatment indicator corresponds to the average causal effect in the sample. In this example a simple way to avoid possible misspecification would be to discretize health status using indicator variables rather than including it as a single continuous predictor. In general, then, causal effects can be estimated using regression if the model includes all confounding covariates (predictors that can affect treatment assignment or the outcome) and if the model is correct. If the confounding covariates are all observed (as in this example), then accurate estimation comes down to proper modeling and the extent to which the model is forced to extrapolate beyond the support of the data. If the confounding covariates are not observed (for example, if we suspect that healthier patients received the treatment, but no accurate measure of previous health status is included in the model), then they are “omitted” or “lurking” variables that complicate the quest to estimate causal effects. We consider these issues in more detail in the rest of this chapter and the next, but first we will provide some intuition in the form of an algebraic formula. (pg. 169, Gelman and Hill, 2007)

Adding regression predictors; “omitted” or “lurking” variables

The preceding theoretical examples illustrate how a simple predictive comparison is not necessarily an appropriate estimate of a causal effect. In these simple examples, however, there is a simple solution, which is to compare treated and control units conditional on previous health status. Intuitively, the simplest way to do this is to compare the averages of the current health status measurements across treatment groups only within each previous health status category; we discuss this kind of subclassification strategy later. Another way to estimate the causal effect in this scenario is to regress the outcome on two inputs: the treatment indicator and previous health status. If health status is the only confounding covariate—that is, the only variable that predicts both the treatment and the outcome—and if the regression model is properly specified, then the coefficient of the treatment indicator corresponds to the average causal effect in the sample. In this example a simple way to avoid possible misspecification would be to discretize health status using indicator variables rather than including it as a single continuous predictor. In general, then, causal effects can be estimated using regression if the model includes all confounding covariates (predictors that can affect treatment assignment or the outcome) and if the model is correct. If the confounding covariates are all observed (as in this example), then accurate estimation comes down to proper modeling and the extent to which the model is forced to extrapolate beyond the support of the data. If the confounding covariates are not observed (for example, if we suspect that healthier patients received the treatment, but no accurate measure of previous health status is included in the model), then they are “omitted” or “lurking” variables that complicate the quest to estimate causal effects. (pg. 169, Gelman and Hill, 2007)

Formula for omitted variable bias

We can quantify the bias incurred by excluding a confounding covariate in the context where a simple linear regression model is appropriate and there is only one confounding covariate. First define the “correct” specification as

$$y_i = \beta_0 + \beta_1 T_i + \beta_2 x_i + \epsilon_i$$

where T_i is the treatment and x_i is the covariate for unit i. If instead the confounding covariate, x_i, is ignored, one can fit the model

$$y_i = \beta_0^* + \beta_1^* T_i + \varepsilon_i^*$$

What is the relation between these models? To understand, it helps to define a third regression,

$$x_i = y_0 + y_1 T_i + \nu_i$$

If we substitute this representation of x into the original, correct, equation, and rearrange terms, we get

$$y_i = \beta_0 + \beta_2 y_0 + (\beta_1 + \beta_2 \gamma_1) T_i + \epsilon_i + \beta_2 \nu_i$$

Equating the coefficients of T yields

$$\beta_1^* = \beta_1 + \beta_2^* \gamma_1$$

This correspondence helps demonstrate the definition of a confounding covariate. If there is no association between the treatment and the purported confounder (that is, $\gamma_1 = 0$) or if there is no association between the outcome and the confounder (that is, $\beta_2 = 0$) then the variable is not a confounder because there will be no bias ($\beta_2 * \gamma_1 = 0$) . (pg. 170, Gelman and Hill, 2007)

The fundemental Problem of Causal Inference

The following is adapted from Imbens and Rubin, 2007 (pages 5-7)

Suppose Bob, at a particular point in time, is contemplating whether or not to take an aspirin for a headache. There are two treatment levels, taking an aspirin, and not taking an aspirin. If Bob takes the aspirin, his headache may be gone, or it may remain, say, an hour later; we denote this outcome, which can be either “Headache” or “No Headache,” by $Y(\text{Aspirin})$. (We could use a finer measure of the sta- tus of my headache an hour later, for example, rating my headache on a ten-point scale, but that does not alter the fundamental issues involved here.) Similarly, if Bob does not take the aspirin, his headache may remain an hour later, or it may not; we denote this potential outcome by $Y(\text{No Aspirin})$, which also can be either “Headache,” or “No Headache.” There are therefore two potential outcomes, $Y(\text{Aspirin})$ and $Y(\text{No Aspirin})$, one for each level of the treatment. The causal effect of the treatment involves the comparison of these two potential outcomes.

Because in this example each potential outcome can take on only two values, the unit- level causal effect – the comparison of these two outcomes for the same unit – involves one of four (two by two) possibilities:

1. Headache gone only with aspirin: Y(Aspirin) = No Headache, Y(No Aspirin) = Headache
2. No effect of aspirin, with a headache in both cases: Y(Aspirin) = Headache, Y(No Aspirin) = Headache
3. No effect of aspirin, with the headache gone in both cases: Y(Aspirin) = No Headache, Y(No Aspirin) = No Headache
4. Headache gone only without aspirin: Y(Aspirin) = Headache, Y(No Aspirin) = No Headache

There are two important aspects of this definition of a causal effect. 1. The definition of the causal effect depends on the potential outcomes, but it does not depend on which outcome is actually observed. Specifically, whether Bob takes an aspirin (and am therefore unable to observe the state of my headache with no aspirin) or do not take an aspirin (and am thus unable to observe the outcome with an aspirin) does not affect the definition of the causal effect. 2. The causal effect is the comparison of potential outcomes, for the same unit, at the same moment in time post-treatment. In particular, the causal effect is not defined in terms of comparisons of outcomes at different times, as in a before-and-after comparison of my headache before and after deciding to take or not to take the aspirin. “The fundamental problem of causal inference” (Holland, 1986, p. 947) is therefore the problem that at most one of the potential outcomes can be realized and thus observed. If the action you take is Aspirin, you observe Y(Aspirin) and will never know the value of Y(No Aspirin) because you cannot go back in time. Similarly, if your action is No Aspirin, you observe Y(No Aspirin) but cannot know the value of Y(Aspirin). In general, therefore, even though the unit-level causal effect (the comparison of the two potential outcomes) may be well defined, by definition we cannot learn its value from just the single realized potential outcome.

The outcomes that would be observed under control and treatment conditions are often called counterfactuals or potential outcomes.

If Bob took asprin for his headache then he would be assigned to the treatment condition so $T_i = 1$. Then $Y(\text{Aspirin})$ is observed and $Y(\text{No Aspirin})$ is the unobserved counterfactual outcome—it represents what would have happened to Bob if he had no taken aspirin. Conversely, if Bob had not taken aspirin then $Y(\text{No Aspirin})$ is observed and $Y(\text{Aspirin})$ is counterfactual. In either case, a simple treatment effect for Bob can be defined as

$$\text{treatment effect for Bob } = Y(\text{Aspirin}) - Y(\text{No Aspirin}).$$

The table below shows hypothetical data for an experiment with 100 units (200 potential outcomes). The table shows what the data that is required to determine causal effects for each person in the dataset—that is, it includes both potential outcomes for each person.

Unit	X_i	T_i	y_i^0	y_i^1	$y_i^0 - y_i^1$
1	2	0	**69**	75	-6
2	6	1	80	**76**	4
3	3	1	71	**69**	1
. . .					

100	9	0	**81**	78	3

The observed data is actually

Unit	X_i	T_i	y_i^0	y_i^1	$y_i^0 - y_i^1$
1	2	0	**69**	?	?
2	6	1	?	**76**	?
3	3	1	?	**69**	?
. . .					
100	9	0	**81**	?	?

The fundamental problem of causal inference is that at most one of these two potential outcomes, y_i^0 and y_i^1 , can be observed for each unit i. The bottom table displays the data that can actually be observed. The y_i^1 values are “missing” for those in the control group and the y_i^0 values are “missing” for those in the treatment group. (pg. 170-171, Gelman and Hill, 2007)

We cannot observe both what happens to an individual after taking the treatment (at a particular point in time) and what happens to that same individual after not taking the treatment (at the same point in time). Thus we can never measure a causal effect directly.

Ways to get around the problem (pg. 171-172, Gelman and Hill, 2007)

Close substitutes. One might object to the formulation of the fundamental problem of causal inference by noting situations where it appears one can actually measure both yi0 and yi1 on the same unit. Consider, for example drinking tea one evening and milk another evening, and then measuring the amount of sleep each time. A careful consideration of this example reveals the implicit assumption that there are no systematic differences between days that could also affect sleep. An additional assumption is that applying the treatment on one day has no effect on the outcome on another day. More pristine examples can generally be found in the natural and physical sciences. For instance, imagine dividing a piece of plastic into two parts and then exposing each piece to a corrosive chemical. In this case, the hidden assumption is that pieces are identical in how they would respond with and without treatment, that is, $y_1^0 = y_2^0$ and $y_1^1 = y_2^1$. As a third example, suppose you want to measure the effect of a new diet by comparing your weight before the diet and your weight after. The hidden assumption here is that the pre-treatment measure can act as a substitute for the potential outcome under control, that is, $y_i^0 = x_i$. It is not unusual to see studies that attempt to make causal inferences by substi- tuting values in this way. It is important to keep in mind the strong assumptions often implicit in such strategies.

Randomization and experimentation. A different approach to causal inference is the “statistical” idea of using the outcomes observed on a sample of units to learn about the distribution of outcomes in the population. The basic idea is that since we cannot compare treatment and control outcomes for the same units, we try to compare them on similar units. Similarity can be attained by using randomization to decide which units are assigned to the treat- ment group and which units are assigned to the control group. We will discuss this strategy in depth in the next section.

Statistical adjustment. For a variety of reasons, it is not always possible to achieve close similarity between the treated and control groups in a causal study. In obser- vational studies, units often end up treated or not based on characteristics that are predictive of the outcome of interest (for example, men enter a job training program because they have low earnings and future earnings is the outcome of interest). Ran- domized experiments, however, can be impractical or unethical, and even in this context imbalance can arise from small-sample variation or from unwillingness or inability of subjects to follow the assigned treatment. When treatment and control groups are not similar, modeling or other forms of statistical adjustment can be used to fill in the gap. For instance, by fitting a regression (or more complicated model), we may be able to estimate what would have happened to the treated units had they received the control, and vice versa. Alternately, one can attempt to divide the sample into subsets within which the treatment/control allocation mimics an experimental allocation of subjects. We discuss regression approaches in this chapter. We discuss imbalance and related issues more thoroughly in Chapter 10 along with a description of ways to help observational studies mimic randomized experiments.

Randomized Experiments

Suppose that an experiment with units randomly assigned to receive treatment and control, and also that the units are a random sample from a population of interest. The random sampling and random treatment assignment allow us to estimate the average causal effect of the treatment in the population, and regression modeling can be used to refine this estimate.

Average causal effects and randomized experiments

In most practical situations it's impossible to estimate individual-level causal effects but, we can design studies to estimate the population average treatment effect:

$$\text{average treatment effect} = \bar{y}^1 - \bar{y}^0,$$

where $\bar{y}^1 = \sum_{i=1}^{n_1} y_i^1/n_1$ and $\bar{y}^0 = \sum_{i=1}^{n_0} y_i^0/n_0$.

Considered more broadly, we can think of the control group as a group of units that could just as well have ended up in the treatment group, they just happened not to get the treatment. Therefore, on average, their outcomes represent what would have happened to the treated units had they not been treated; similarly, the treatment group outcomes represent what might have happened to the control group had they been treated.

if n_0 units are selected at random from the population and given the control, and n_1 other units are randomly selected and given the treatment, then the observed sample averages of y for the treated and control units can be used to estimate the corresponding population quantities, $\bar{y}^1$ and $\bar{y}^0$, with their difference estimating the average treatment effect (and with standard error $\sqrt{S_0^2/n_0 + S_1^2/n_1}$). This works because the y_i^0 for the control group are a random sample of the values of y_i^0 in the entire population. Similarly, the y_i^1 for the treatment group are a random sample of the y_i^1's in the population. Equivalently, if we select $n_0 + n_1$ units at random from the population, and then randomly assign n_0 of them to the control and n_1 to the treatment, we can think of each of the sample groups as representing the corresponding population of control or treated units. Therefore the control group mean can act as a counterfactual for the treatment group (and vice versa).

What if the $n_0 + n_1$ units are selected nonrandomly from the population but then the treatment is assigned at random within this sample? This is common practice, for example, in experiments involving human subjects. Experiments in medicine, for instance, are conducted on volunteers with specified medical conditions who are willing to participate in such a study, and experiments in psychology are often conducted on university students taking introductory psychology courses. In this case, causal inferences are still justified, but inferences no longer generalize to the entire population. It is usual instead to consider the inference to be appropriate to a hypothetical superpopulation from which the experimental subjects were drawn. Further modeling is needed to generalize to any other population.

Stable Unit Treatment Value Assumption (SUTVA)

The assumption. The potential outcomes for any unit do not vary with the treatments assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.

The stable unit treatment value assumption (Rubin, 1980) involves assuming that treatments applied to one unit do not affect the outcome for another unit. For example, if Adam and Oliver are in different locations and have no contact with each other, it would appear reasonable to assume that if Oliver takes an an aspirin for his headache then his behaviour has no effect on the status of Adam's headache. This assumption might not hold if Adam and Oliver are in the same location, and Adam's behavior, affects Oliver's behaviour. SUTVA incorporates the idea that Adam Oliver do not interfere with one another and the idea that for each unit there is only a single version of each treatment level (e.g., there is only one dose of aspirin). (Imbens and Rubin, 2015)

The causal effect of aspirin on headaches can be estimated if we could exclude the possibility that your taking or not taking aspirin has any effect on my headache, and the possibility that the aspirin tablets available to me are of different strengths. These are assumptions that rely on previously acquired knowledge of the subject matter for their justification. Causal inference is generally impossible without such assumptions, and thus it is critical to be explicit about their content and their justifications. (Imbens and Rubin, 2015)

Electric company study

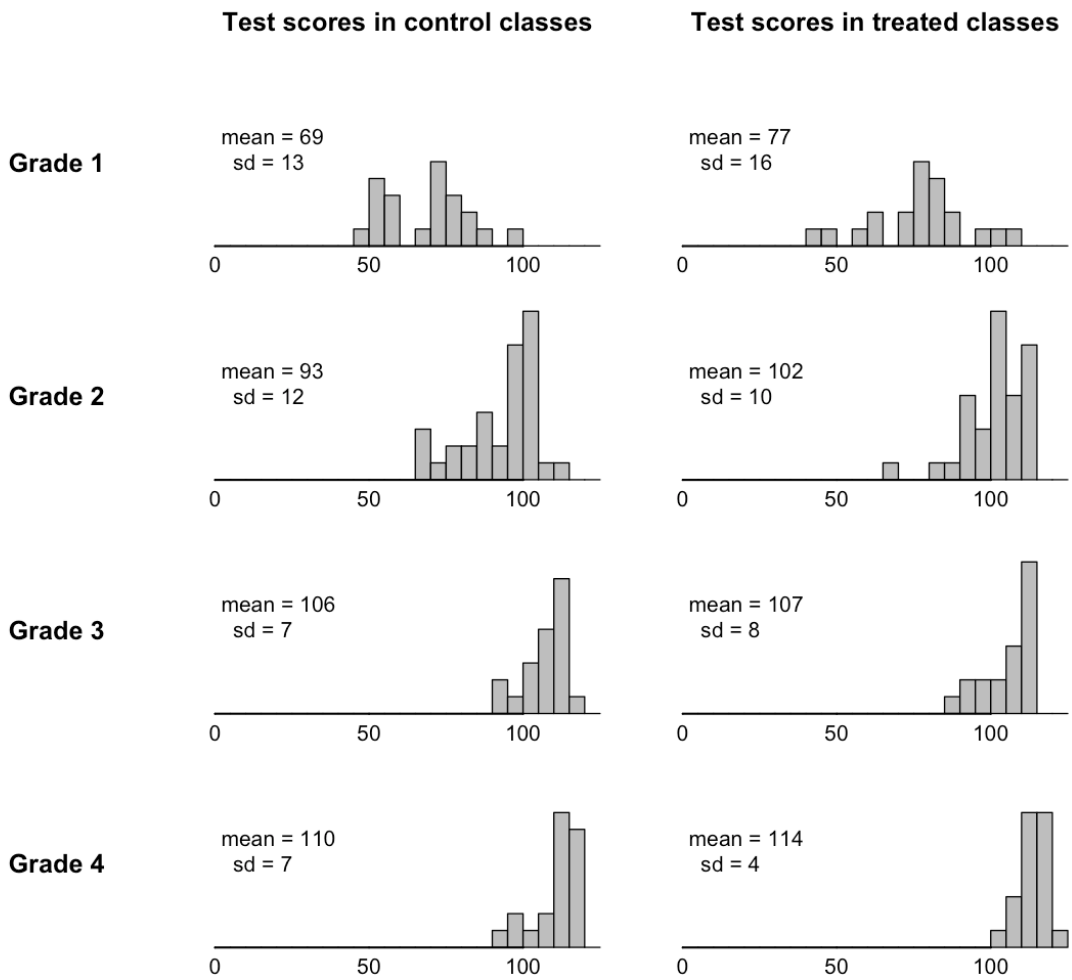
The following example is from Gelman and Hill (2007).

An educational experiment performed on four elementary school classes (grades 1-4) was conducted.

- **Treatment:** Exposure to a new educational television show called The Electric Company.
- **Treatment assignment:** Classes were randomized into treated and control groups.
- **Primary study outcome:** At the end of the school year, students in all the classes were given a reading test, and the average test score within each class was recorded.

A pre-test was given at the beginning of the school year before the treatment. For grade 1 the pre-test was a subset of the longer post-test and in grades 2-4 the pre-test was the same as the post-test.

The distribution of scores are shown for each treatment and control groups for each grade.



The average causal effect of the treatment is the coefficient in the regression

$$y_i = \beta_0 + \beta_1 T_i + \epsilon_i,$$

for each grade.

```
attach(electric)
post.test <- c (treated.Posttest, control.Posttest)
grade <- rep (Grade, 2)
treatment <- rep (c(1,0), rep(length(treated.Posttest),2))

for (k in 1:4){
  cat("Grade:",k, "\n")
  print(summary((lm (post.test ~ treatment, subset=(grade==k)))))
  print(confint(lm (post.test ~ treatment, subset=(grade==k))))
}
```

```
Grade: 1

Call:
lm(formula = post.test ~ treatment, subset = (grade == k))

Residuals:
    Min      1Q  Median      3Q     Max
-32.890 -13.190   2.060   7.685  31.510

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   68.790      3.268  21.047   <2e-16 ***
treatment      8.300      4.622   1.796   0.0801 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.98 on 40 degrees of freedom
Multiple R-squared:  0.0746,    Adjusted R-squared:  0.05146
F-statistic: 3.224 on 1 and 40 DF,  p-value: 0.0801

                2.5 %  97.5 %
(Intercept) 62.184754 75.3962
treatment   -1.041902 17.6419
Grade: 2

Call:
lm(formula = post.test ~ treatment, subset = (grade == k))

Residuals:
    Min      1Q  Median      3Q     Max
-35.171  -6.796   2.509   9.299  17.088

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   93.212      1.907   48.88  < 2e-16 ***
treatment      8.359      2.697    3.10  0.00285 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.12 on 66 degrees of freedom
Multiple R-squared:  0.1271,    Adjusted R-squared:  0.1138
F-statistic: 9.607 on 1 and 66 DF,  p-value: 0.002848

               2.5 %   97.5 %
(Intercept) 89.40453 97.01900
treatment    2.97458 13.74307
Grade: 3

Call:
lm(formula = post.test ~ treatment, subset = (grade == k))

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-17.610  -3.525   2.740   4.900   9.125

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  106.175      1.663  63.858   <2e-16 ***
treatment      0.335      2.351   0.142    0.887
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.436 on 38 degrees of freedom
Multiple R-squared:  0.0005339, Adjusted R-squared:  -0.02577
F-statistic: 0.0203 on 1 and 38 DF,  p-value: 0.8875

                 2.5 %     97.5 %
(Intercept) 102.809071 109.540929
treatment    -4.425143   5.095143
Grade: 4

Call:
lm(formula = post.test ~ treatment, subset = (grade == k))

Residuals:
    Min      1Q  Median      3Q     Max
-16.357  -1.489   1.093   3.918   7.933

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  110.357      1.299   84.98   <2e-16 ***
treatment      3.710      1.837    2.02   0.0501 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.951 on 40 degrees of freedom
Multiple R-squared:  0.09255,   Adjusted R-squared:  0.06987
F-statistic:  4.08 on 1 and 40 DF,  p-value: 0.05014

                    2.5 %     97.5 %
(Intercept) 107.732487230 112.981798
treatment    -0.002299775   7.421347
```

For example, the average causal effect of the television show for grade 1 is 8.3 (95% confidence interval [-1.04, 17.64]).

The pre-test can be controlled by fitting the model:

$$y_i = \beta_0 + \beta_1 T_i + \beta_2 x_i + \epsilon_i,$$

where x_i is the pre-test.

```
attach(electric)
post.test <- c (treated.Posttest, control.Posttest)
pre.test <- c (treated.Pretest, control.Pretest)
grade <- rep (Grade, 2)
treatment <- rep (c(1,0), rep(length(treated.Posttest),2))

for (k in 1:4){
  cat("Grade:",k, "\n")
  print(summary((lm (post.test ~ treatment+pre.test, subset=(grade==k)))))
  print(confint(lm (post.test ~ treatment+pre.test, subset=(grade==k))))
}
```

```
Grade: 1

Call:
lm(formula = post.test ~ treatment + pre.test, subset = (grade ==
    k))

Residuals:
    Min      1Q  Median      3Q     Max
-19.360  -5.059   0.445   5.640  15.349

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.0229     8.7860  -1.255  0.21709
treatment     8.7865     2.6118   3.364  0.00173 **
pre.test      5.1084     0.5498   9.292 1.96e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.461 on 39 degrees of freedom
Multiple R-squared:  0.712, Adjusted R-squared:  0.6973
F-statistic: 48.22 on 2 and 39 DF,  p-value: 2.863e-11

                2.5 %    97.5 %
(Intercept) -28.794124  6.748422
treatment     3.503741 14.069296
pre.test      3.996396  6.220489
Grade: 2

Call:
lm(formula = post.test ~ treatment + pre.test, subset = (grade ==
    k))

Residuals:
     Min       1Q   Median       3Q      Max
-15.8446  -3.4414   0.3449   3.8631  11.2716

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.42877    3.99098   9.378 1.07e-13 ***
treatment    4.26577    1.35896   3.139  0.00255 **
pre.test     0.78891    0.05486  14.382  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.479 on 65 degrees of freedom
Multiple R-squared:  0.7913,    Adjusted R-squared:  0.7848
F-statistic: 123.2 on 2 and 65 DF,  p-value: < 2.2e-16

                2.5 %     97.5 %
(Intercept) 29.4582320 45.3993009
treatment    1.5517446  6.9797863
pre.test     0.6793567  0.8984661
Grade: 3
```

```
Call:
lm(formula = post.test ~ treatment + pre.test, subset = (grade ==
    k))

Residuals:
    Min      1Q  Median      3Q     Max
-5.2063 -1.7614  0.3153  1.7005  6.9502

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 40.58424    3.72776   10.89  4.3e-13 ***
treatment    1.90973    0.77616    2.46   0.0187 *
pre.test     0.68466    0.03849   17.79  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.438 on 37 degrees of freedom
Multiple R-squared:  0.8953,    Adjusted R-squared:  0.8897
F-statistic: 158.3 on 2 and 37 DF,  p-value: < 2.2e-16

                2.5 %     97.5 %
(Intercept) 33.0310748 48.1374132
treatment    0.3370718  3.4823799
pre.test     0.6066683  0.7626586
Grade: 4

Call:
lm(formula = post.test ~ treatment + pre.test, subset = (grade ==
    k))

Residuals:
    Min      1Q  Median      3Q     Max
-5.3504 -1.0094  0.0801  0.7166  7.0962

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 41.99473    4.28162   9.808 4.42e-12 ***
treatment    1.70144    0.68535   2.483   0.0175 *
pre.test     0.65583    0.04082  16.066  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.184 on 39 degrees of freedom
Multiple R-squared:  0.8809,    Adjusted R-squared:  0.8748
F-statistic: 144.2 on 2 and 39 DF,  p-value: < 2.2e-16

                2.5 %     97.5 %
(Intercept) 33.3343230 50.6551288
treatment    0.3151918  3.0876810
pre.test     0.5732627  0.7383963
```

The coefficient for the treatment indicator still represents the average treatment effect, but controlling for pre-test score.

The pre-test score is not affected by the treatment. Therefore, it's appropriate to control for this covariate since it will improve the efficiency of the coefficient for treatment (average causal effect). The standard error of the treatment coefficient for grade 1 is 4.6 without adjusting for the pre-test and 2.6 after adjusting for pre-test.

Observational Studies

- A solution to the fundamental problem of causal inference is, as we have described, to randomly sample a different set of units for each treatment group assignment from a common population, and then apply the appropriate treatments to each group; or randomly assign the treatment conditions among a selected set of units. Both approaches are supposed to guaruntee that, on average, the different treatment groups are balanced. In other words, $\bar{y}^0$ and $\bar{y}^1$ from the sample are estimating the average outcomes under control and treatment for the same population.
- In observational studies treatments are observed rather than assigned (e.g., comparison of lung cancer rates in smokers to nonsmokers), and the observed data under different treatments is usually not a random sample from a common population.
- In an observational study there might be systematic differences between groups of units, outside the control of the experimenter, that receive different treatments that can affect the outcome, y.

> The educational experiment described above had an embedded observational study. Once the treatments had been assigned, the teacher for each class assigned to the Electric Company treatment chose to either replace or sup- plement the regular reading program with the Electric Company television show. That is, all the classes in the treatment group watched the show, but some watched it instead of the regular reading program and others got it in addition. The simplest starting point to analyzing these observational data (now limited to the randomized treatment group) is to consider the choice between the two treatment options —"replace" or "supplement"—to be randomly assigned conditional on pre-test scores. This is a strong assumption but we use it simply as a starting point. We can then estimate the treatment effect by regression, as with an actual experiment.

```
attach(electric)
supp <- c(as.numeric(electric[,"Supplement."])-1, rep(NA,nrow(electric)))
# supp=0 for replace, 1 for supplement, NA for control

est1 <- rep(NA,4)
se1 <- rep(NA,4)
for (k in 1:4){
  ok <- (grade==k) & (!is.na(supp))
  cat("Grade:",k, "\n")
  print(summary(lm (post.test ~ supp + pre.test, subset=ok)))
  print(confint(lm (post.test ~ supp + pre.test, subset=ok)))
}
```

```
Grade: 1

Call:
lm(formula = post.test ~ supp + pre.test, subset = ok)

Residuals:
    Min      1Q  Median      3Q     Max
-21.999  -6.582  -0.485   6.908  14.055

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.2160    12.1456  -0.018    0.986
supp          5.9230     4.4441   1.333    0.199
pre.test      4.7241     0.7846   6.021 1.08e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.42 on 18 degrees of freedom
Multiple R-squared:  0.7035,    Adjusted R-squared:  0.6706
F-statistic: 21.36 on 2 and 18 DF,  p-value: 1.769e-05

                2.5 %    97.5 %
(Intercept) -25.732934 25.301026
supp         -3.413773 15.259775
pre.test      3.075675  6.372427
Grade: 2

Call:
lm(formula = post.test ~ supp + pre.test, subset = ok)

Residuals:
     Min       1Q   Median       3Q      Max
-13.2646  -2.8259  -0.8626   3.1002  12.7672

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.48608    6.52677   6.050 1.06e-06 ***
supp         4.68689    1.83402   2.556   0.0157 *
pre.test     0.78168    0.08466   9.233 2.08e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.257 on 31 degrees of freedom
Multiple R-squared:  0.7525,    Adjusted R-squared:  0.7366
F-statistic: 47.13 on 2 and 31 DF,  p-value: 3.98e-10

                2.5 %     97.5 %
(Intercept) 26.1746504 52.7975109
supp         0.9463903  8.4273952
pre.test     0.6090171  0.9543507
Grade: 3

Call:
```

```
lm(formula = post.test ~ supp + pre.test, subset = ok)

Residuals:
   Min     1Q Median     3Q    Max
-6.180 -1.131  0.290  1.647  5.977

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 43.46910    6.23642   6.970 2.26e-06 ***
supp        -1.29788    1.51307  -0.858    0.403
pre.test     0.68465    0.06245  10.963 3.96e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.848 on 17 degrees of freedom
Multiple R-squared:  0.8864,    Adjusted R-squared:  0.873
F-statistic: 66.32 on 2 and 17 DF,  p-value: 9.357e-09

                2.5 %     97.5 %
(Intercept) 30.3113901 56.6268005
supp        -4.4901749  1.8944099
pre.test     0.5528908  0.8163994
Grade: 4

Call:
lm(formula = post.test ~ supp + pre.test, subset = ok)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0600 -0.6256  0.3921  1.1380  6.2540

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 54.72928   10.11638   5.410 3.86e-05 ***
supp        -0.32384    1.17112  -0.277    0.785
pre.test     0.55473    0.09541   5.814 1.65e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.581 on 18 degrees of freedom
Multiple R-squared:  0.6609,    Adjusted R-squared:  0.6232
F-statistic: 17.54 on 2 and 18 DF,  p-value: 5.936e-05

                2.5 %     97.5 %
(Intercept) 33.4755555 75.9830111
supp        -2.7842683  2.1365878
pre.test     0.3542762  0.7551821
```

There is a significant difference at the 5% level in grade 2. The other grades don't show evidence of a difference between replace or supplement.

Ignorable Treatment Assignment

The observational study of replace or supplement was treated as a completely randomized experiment and an estimate of the treatment effect was calculated. What has been assumed? The distribution of classes across the treatment variable (replace/supplement) is random with respect to the potential outcomes. This is almost correct, since the regression adjusts for pre-test scores. So, the distribution of classes with respect to potential outcomes is random across the treatment groups *conditional* on the confounding covariate pre-test.

If treatment assignment T is conditionally independent of y^0, y^1 given confounding covariates X then the treatment assignment is said to be **ignorable**. Symbollically this is represented as

$$y^0, y^1 \perp T|X.$$

This means that the conditional distribution of potential responses is the same across levels of the treatment variable once we condition on the confounding covariates X.

Another way to view ignorability is:

$$P(T = 1|y^0, y^1, X) = P(T = 1|X).$$

If treatments were assigned by the flip of a coin then the coin flips may depend on the covariates but not on the potential responses.

Imagine that two grade 1 classes have the same average pre-test scores. For example, a coin toss might determine which of these two classes would get replacement or supplement. This doesn't mean that any two classes will have the same probability of receiving the suplement. The ignorability assumption means that if we want to interpret the regression coefficient for treatment as an average causal effect then all the counfounding covariates should be controlled for in the regression model.

In a randomized experiments the treatment assignment is ignorable. In a completely randomized experiment conditioning on pre-tretment variables is not required. This is why we can you use the difference in means to estimate the causal effects. In a paired randomized experiment the ignorability condition is true conditional on the design variable used to pair.

Examples of non-ignorable treatment assignments in the Electric Company supplement/replacement example.

1. If supplementing was more likely to be chosen by a teacher if she or he believed that it would be more effective for that particular class based on if the teacher considered the class to be intelligent, and intelligence is related to their subsequent test scores.
2. If supplementing was less likely to be chosen by more experienced teachers, and teaching experience was also associated with the students' future test scores.

> In an observational study, the “treatment assignment” is not under the control of the statistician, but one can aim for ignorability by conditioning in the analysis stage on as much pre-treatment information in the regression model as possible. For example, if teachers’ motivation might affect treatment assignment, it would be advisable to have a pre-treatment measure of teacher motivation and include this as an input in the regression model. This would increase the plausibility of the ignorability assumption. Realistically, this may be a difficult characteristic to measure, but other teacher characteristics such as years of experience and schooling might act as partial proxies. In general, one can never prove that the treatment assignment process in an observational study is ignorable—it is always possible that the choice of treatment depends on relevant information that has not been recorded. In an educational study this information could be characteristics of the teacher or school that are related both to treatment assignment and to post-treatment test scores. Thus, if we interpret the estimates in Figure 9.9 as causal effects, we do so with the understanding that we would prefer to have further pre-treatment information, especially on the teachers, in order to be more confident in ignorability. (pg. 183-184, Gelman and Hill, 2007)

If the treatment assignments depends on information not included in the model then caution should be exercised in teh interpretation of treatment effects. Different analysis strategies can be chosen (see Chapter 10, Gelman and Hill, 2007).

Causal Inference in Observational Studies

Good observational studies are designed and not found lying around. It may seem tempting since there are so many data sources available where “discoveries” are just waiting to be made. When we are designing experiments we don’t have any outcome data, but we plan the collection, organization, and analysis of the data to improve our chances of obtaining causal answers. The same should be done in an observational study, even if the outcome data are available at the design stage.

Observational studies can also refer to survey data settings where no intervention has been performed. One consideration in this setting is defining the treatment variable, and another is identifying the effects of multiple treatment factors.

Defining a treatment variable

In order to define a treatmemt variable each unit has to be able to potentially experience each level of the treatment variable for which causal effects will be defined for the unit. For example, the “effect” of eye colour on happiness is ill-defined unless there is a treatment that could change eye colour (e.g.,

coloured contact lenses). What does it mean to define a potential happiness outcome for a person that would occur if he or she had a different eye colour?

Consider the effect of homelessness on children's health outcomes. We might envision a policy intervention that could change homelessness before or after a child's health outcome. Defining the treatment can often be a difficult exercise, but has important implications for the design, analysis, and interpretation of the results.

Cochran's Basic Advice

Rosenbaum (2010) discusses this stand alone quote from William Cochran.

The planner of an observational study should always ask himself the question, "How would the study be conducted if it were possible to do it by controlled experimentation?" Cochran (1965).