

5. ESTIMATING CAUSAL EFFECTS WITH OBSERVATIONAL DATA

In chapter 2, we learned how to estimate average causal effects using data from randomized experiments. Here, we learn how to estimate them when we cannot randomly assign the treatment and instead have to rely on observational data. As an illustration, we estimate the causal effects of Russian TV reception on the 2014 Ukrainian parliamentary election.

5.1 RUSSIAN STATE-CONTROLLED TV COVERAGE OF 2014 UKRAINIAN AFFAIRS

Ukraine became independent from the Soviet Union in 1991. Since then, attitudes toward Russia have often been a point of contention. For a long time, the Ukrainian population and political parties were divided into pro-Russian and anti-Russian.

Leading up to the 2014 Ukrainian elections, Russia and Ukraine (which, at the time, was governed by a party with an “anti-Russian” agenda) were in fierce political and military conflict. Russian state-controlled TV coverage of the conflict, and of the issues at stake in the Ukrainian elections, was intense and one-sided. For instance, the coverage deemed the Ukrainian government illegitimate and claimed that the revolution that brought it to power had been organized by foreign countries. Such coverage was aired not only in Russian territory but also in parts of Ukraine. Some Ukrainians living close to the border received the signal, and thus, could be exposed to pro-Russia propaganda.

In this chapter, we estimate the effect of Russian TV reception on Ukraine’s 2014 parliamentary election. We do so at two levels. First, we analyze individual-level survey data to estimate the impact on an individual’s propensity to vote for a pro-Russian party. Second, we analyze aggregate-level data to estimate the effect on the vote share of pro-Russian parties at the precinct level. In both cases, we focus on areas close to the Russian border.

Based on Leonid Peisakhin and Arturas Rozenas, “Electoral Effects of Biased Media: Russian Television in Ukraine,” *American Journal of Political Science* 62, no. 3 (2018): 535–50. To simplify the analyses, we consider that a signal strength of 50 dBuV or above provides reception, and we limit the number of potential confounders.



5.2 CHALLENGES OF ESTIMATING CAUSAL EFFECTS WITH OBSERVATIONAL DATA

RECALL: The fundamental problem of causal inference is that we can never observe the counterfactual outcome. Yet to infer causal effects, we need to compare the factual outcome with the counterfactual outcome.

RECALL: Observational data are data collected about naturally occurring events, where researchers do not assign treatment.

As we discussed in chapter 2, to estimate causal effects, we must find or create a situation in which the treatment and control groups are comparable with respect to all the variables that might affect the outcome other than the treatment variable itself. Only when this assumption is satisfied can we use the average factual or observed outcome of one group as a good estimate of the average counterfactual outcome of the other group.

As we have already seen, in randomized experiments, we can rely on the random assignment of the treatment to make treatment and control groups, on average, identical to each other in terms of all observed and unobserved pre-treatment characteristics. But what happens when we cannot conduct a randomized experiment and have to analyze observational data instead? We can no longer assume that treatment and control groups are comparable. To estimate causal effects using observational data, we have to first identify any relevant differences between treatment and control groups—known as confounding variables or confounders—and then statistically control for them so that we can make the two groups as comparable to each other as possible.

We begin this section by defining confounding variables. Then, we explore why their presence poses a problem when estimating causal effects and discuss how the randomization of treatment assignment eliminates all potential confounders in randomized experiments.

5.2.1 CONFOUNDING VARIABLES

A confounding variable, or confounder, denoted as Z , is a variable that affects both (i) the likelihood of receiving the treatment X and (ii) the outcome Y .

A confounding variable, also known as a confounder, is a variable that affects both (i) the likelihood of receiving the treatment X and (ii) the outcome Y .

In mathematical notation, just as we represent the treatment variable as X and the outcome variable as Y , we represent a potential confounding variable as Z . The diagram in figure 5.1 shows the causal relationships between these variables. Note that the arrows between Z and X and between Z and Y both originate from Z , indicating that changes in Z affect the values of X and Y but not the other way around.

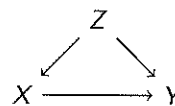


FIGURE 5.1. Representation of the causal relationships between the confounding variable, Z , the treatment variable, X , and the outcome variable, Y . (Recall, we represent a causal relationship with an arrow; the direction of the arrow indicates which one of the variables affects the other.)

Let's look at a simple hypothetical example to get a better sense of how this works. Suppose we are interested in the average causal effect of attending a private school, as opposed to a public one, on student performance. Given the goal of our research:

- the treatment variable, X , is a binary variable indicating whether the student attended a private school (call it *private school*)
- the outcome variable, Y , is student performance on a standardized test such as the SAT (call it *test scores*).

If we are collecting data from the real world, where children attend the school their parents choose, can we think of any variable that affects both (i) the likelihood of attending a private school and (ii) student performance on a test? In other words, can we think of a confounding variable, Z ?

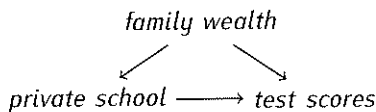
One potential confounding variable is *family wealth*. Given that private schools require that students pay tuition, private school students are likely to come from wealthier families than public school students. Thus, family wealth affects the likelihood that a student attends a private school.

family wealth \rightarrow *private school*

Family wealth also affects the likelihood that a student receives after-school help such as one-on-one tutoring, which, in turn, will improve performance on standardized tests.

family wealth \rightarrow *tutoring* \rightarrow *test scores*

Thus, since *family wealth* affects both *private school* and *test scores*, it is a confounding variable.



5.2.2 WHY ARE CONFOUNDERS A PROBLEM?

Why does the presence of a confounder pose a problem when estimating causal effects? Because confounders obscure the causal relationship between X and Y .

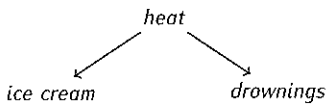
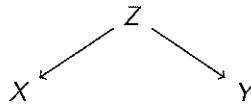
Returning to the example above, if we observed that, on average, private school students perform better on tests than public school students, we would not know whether it is because they attended a private school or because they came from wealthier families

RECALL: When we speak of a “high degree of correlation,” we mean that the correlation coefficient is high in absolute terms, regardless of its sign.

that could afford to provide them with after-school help. In other words, if we were to calculate the difference in average test scores between the two groups (the difference-in-means estimator), we would not know what portion of this difference, if any, could be attributed to the treatment (attending a private school) and what portion was the result of the confounding variable (coming from a wealthier family).

In the presence of confounders, correlation does not necessarily imply causation. Just because we observe that two variables are highly correlated with each other does not automatically mean that one causes the other. There could be a third variable—a confounder—that affects both variables.

In the extreme, by affecting both X and Y at the same time, confounding variables might create a completely spurious relationship between X and Y , misleading us into thinking that X and Y are causally related to each other when, in fact, there is no direct causal link between the two.



For example, ice cream sales and the number of drownings are positively correlated with each other. When we observe a larger number of ice cream sales, we usually also observe a larger number of drownings. That does not mean that eating ice cream causes one to drown. There is an obvious confounder: heat.

When it is hot, people are more likely to eat ice cream, and they are also more likely to go swimming, which might sadly lead to some drownings. The presence of the confounder, heat, then, makes ice cream sales and number of drownings positively correlated with each other. As far as we know, however, there is no direct causal link between them. Eating ice cream does not make one more likely to drown. (Note the lack of a causal link/arrow between *ice cream* and *drownings* in the diagram in the margin.)

Not all cases are this extreme. Typically, there is a causal link between the treatment and the outcome, but the presence of a confounder makes it difficult for us to estimate the causal effect of X on Y accurately (as we saw in the example of the effect of attending a private school on student test scores).

In short, when there is a confounding variable Z affecting X and Y , we should not trust correlation as a measure of causation, and thus, we cannot use the difference-in-means estimator to estimate average causal effects.

IN THE PRESENCE OF CONFOUNDING VARIABLES:

Treatment and control groups are not comparable, correlation does not necessarily imply causation, and the difference-in-means estimator does *not* provide a valid estimate of the average treatment effect.

Note that in order for a variable to be considered a confounder, it has to affect both (i) the likelihood of being treated and (ii) the outcome. If it affects only one, it is not a confounding variable, and therefore, its presence does not complicate the estimation of causal effects. (See scenarios I and II in figure 5.2.)

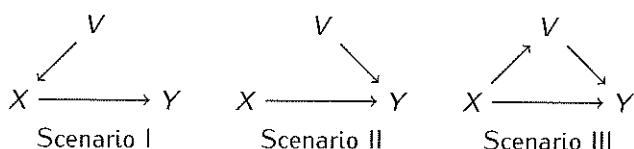
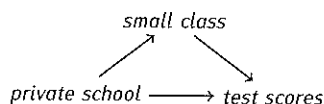
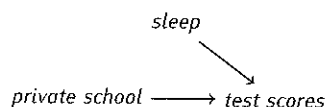
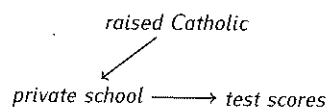


FIGURE 5.2. Representation of scenarios where the variable V , despite its being causally linked to X , or Y , or both, is not a confounding variable.

For example, perhaps students who are raised Catholic are more likely to attend a private school. As long as being raised Catholic does not also affect test performance, it does not constitute a confounder (Scenario I).

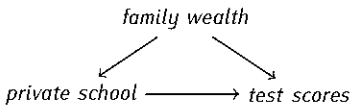
Similarly, perhaps students who get more sleep perform better academically, but if sleeping more doesn't affect the likelihood of attending a private school, then it is not a confounding variable (Scenario II).

Also, mechanisms by which the treatment affects the outcome are not confounders (Scenario III). For example, private schools might have smaller classes than public schools, and smaller classes may improve student performance. The use of smaller classes in private schools is not a confounder but may be one of the mechanisms by which private schools improve student performance. One easy way of seeing this distinction is by thinking about the direction of the causal relationships. A confounder causally affects the treatment and outcome rather than the other way around.



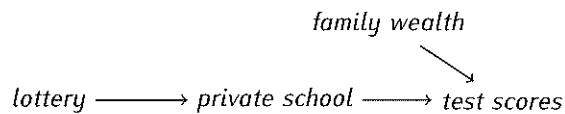
5.2.3 CONFOUNDERS IN RANDOMIZED EXPERIMENTS

Why don't we have to worry about confounders in randomized experiments? Randomization of treatment assignment eliminates all potential confounders. It ensures that treatment and control groups are comparable by breaking the link between any potential confounder and the treatment.



Let's return to the example above, where we were interested in the causal effect of attending a private school on student performance. As discussed, if parents choose the school their children attend, a potential confounding variable is *family wealth*.

If we are designing our study and want to ensure that there are no confounders, how should we decide who attends and does not attend a private school? We can flip a coin (or use any other method of random assignment) to determine which students attend a private school and which attend a public school. If, for example, there were more applicants than open seats in a private school voucher program, we could ensure that there would be no confounders by allocating the vouchers through a method of random assignment such as a lottery.



Now, students from non-wealthy families would be as likely as students from wealthy families to receive the voucher, and thus, attend a private school. In other words, by assigning students to attend a private school with the flip of a coin, we break the link between *family wealth* and *private school*. As a result, *family wealth* is no longer a confounder, since it no longer affects the probability of receiving the treatment (although it continues to affect the outcome).

In general, by assigning the treatment at random, we ensure that nothing related to the outcome is also related to the likelihood of receiving the treatment, including factors that we cannot observe such as student aptitude or motivation. Random assignment of treatment, then, eliminates any potential confounders. This is why in chapter 2 we stated that by randomly assigning treatment, we ensure that treatment and control groups have identical pre-treatment characteristics, on average.

WHY ARE THERE NO CONFOUNDING VARIABLES IN RANDOMIZED EXPERIMENTS? By randomly assigning treatment, we break the link between any potential confounders and the treatment variable, thereby eliminating all potential confounding variables.

This is the reason randomized experiments are regarded as the gold standard for establishing causal relationships in many scientific disciplines. Randomization of treatment assignment makes the estimation of valid causal effects relatively straightforward. All we need to do is compute the difference-in-means estimator.

5.3 THE EFFECT OF RUSSIAN TV ON UKRAINIANS' VOTING BEHAVIOR

In this section, we learn how to estimate average treatment effects using observational, as opposed to experimental, data. As our running example, we study the effects of receiving Russian TV on the voting behavior of Ukrainians in the 2014 parliamentary election. In particular, we analyze data from a survey conducted a few months after the election on a random sample of Ukrainians living in precincts within 50 kilometers (about 31 miles) of the Russian border. (See figure 5.3.)

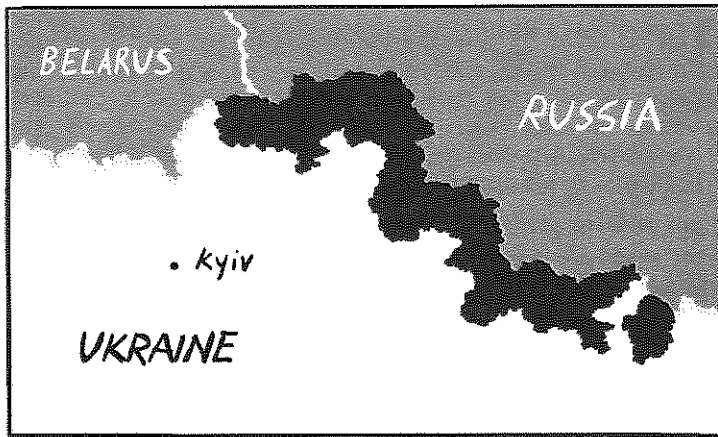
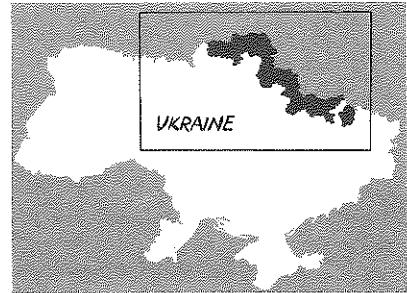


FIGURE 5.3. The precincts studied are within 50 kilometers of the border with Russia (shown in black).



The dataset is provided in the file "UA_survey.csv". Table 5.1 shows the names and descriptions of the variables included.

variable	description
<i>russian_tv</i>	identifies whether the respondent's precinct receives Russian TV: 1=there is reception or 0=there is no reception
<i>pro_russian_vote</i>	identifies respondents who reported having voted for a pro-Russian party in the 2014 parliamentary election: 1=voted for a pro-Russian party or 0=did not
<i>within_25km</i>	identifies whether the respondent's precinct is within 25 kilometers of the Ukraine-Russia border: 1=it is within 25 kilometers of the border or 0=it is not

TABLE 5.1. Description of the variables in the UA_survey dataset, where the unit of observation is respondents.

The code for this chapter's analysis can be found in the "Observational.R" file. As always, we begin by reading and storing the data (assuming we have already set the working directory):

RECALL: If the DSS folder is saved directly on your Desktop, to set the working directory, you must run `setwd("~/Desktop/DSS")` if you have a Mac and `setwd("C:/user/Desktop/DSS")` if you have a Windows computer (where *user* is your own username). If the DSS folder is saved elsewhere, please see subsection 1.7.1 for instructions on how to set the working directory.

```
uas <- read.csv("UA_survey.csv") # reads and stores data
```

To get a sense of the dataset, we look at the first few observations:

```
head(uas) # shows first observations
##  russian_tv  pro_russian_vote  within_25km
##  1          1                0            1
##  2          1                1            1
##  3          0                0            0
##  4          0                0            1
##  5          0                0            1
##  6          1                0            0
```

Based on table 5.1 and the output above, we learn that each observation in the dataset represents a respondent, and that the dataset contains three variables:

- *russian_tv* is a binary variable that identifies whether the respondent's precinct received Russian TV
- *pro_russian_vote* is a binary variable that identifies whether the respondent reported having voted for a pro-Russian party in the 2014 Ukrainian parliamentary election
- *within_25km* is a binary variable that identifies whether the respondent's precinct is very close to the border with Russia (defined as within 25 kilometers).

We interpret the first observation as representing a respondent who lived in a precinct that received Russian TV, did not vote for a pro-Russian party, and lived in a precinct within 25 kilometers (km) of the border.

To find the total number of observations in the dataset, we run:

```
dim(uas) # provides dimensions of dataframe: rows, columns
## [1] 358  3
```

The dataset contains information for 358 survey respondents.

5.3.1 USING THE SIMPLE LINEAR MODEL TO COMPUTE THE DIFFERENCE-IN-MEANS ESTIMATOR

RECALL: Simple linear models use only one *X* variable to predict *Y*.

In this subsection, we learn to fit a simple linear model that produces an estimated coefficient that is equivalent to the difference-in-means estimator. This procedure is a stepping stone toward fitting a more complex model in which we estimate an average causal effect while statistically controlling for confounders.

While we use the same statistical method as in the previous chapter, we do so with a different goal in mind. In chapter 4, we fitted a linear model to *predict* a quantity of interest, that is, to predict

the outcome Y given a value of the predictor X . In this chapter, we fit a linear model to *explain* a quantity of interest, that is, to estimate the causal relationship between the treatment X and the outcome Y . (Recall, X denotes the predictor when we are making predictions, but it denotes the treatment variable when we are estimating causal effects.) As we will soon see, the goal of the analysis does not affect the mathematical underpinnings of the model (the method used to fit the line and the mathematical definitions of the coefficients remain the same), but it does affect the substantive interpretations of the coefficients.

Let's analyze the `UA_survey` dataset as an example. Here, we are interested in estimating the average causal effect that receiving Russian TV had on a respondent's probability of voting for a pro-Russian party in the 2014 Ukrainian parliamentary election. In other words, we are interested in the causal link between *russian_tv* and *pro_russian_vote*, where *russian_tv* is the treatment variable and *pro_russian_vote* is the outcome variable.

Russian TV reception → *pro-Russian vote*

Can we use the difference-in-means estimator to estimate this average treatment effect? The information contained in this dataset does not come from a randomized experiment, but rather from naturally occurring events. The reception of Russian TV was not randomly assigned to different precincts. Instead, Russian TV reception was determined by factors such as the terrain and distance between the precinct where the respondent lived and the Russian TV transmitters. The data we are analyzing are therefore observational, not experimental.

Having said that, while the factors that determined Russian TV reception were outside the researchers' control, one could argue that they produced an "as-if-random" variation of treatment that had nothing to do with the determinants of individual voting behavior. For example, small differences in terrain affected Russian TV reception but probably did not affect voting behavior directly. For now, then, we assume that the respondents who received Russian TV were similar in all relevant characteristics to those who did not and use the difference-in-means estimator to estimate the average treatment effect. (Later, we will see what happens when we relax this assumption.)

In the running example, to compute the difference-in-means estimator (just as we did in subsection 2.5.3), we run:

```
## calculate the difference-in-means estimator
mean(uas$pro_russian_vote[uas$russian_tv==1]) -
  mean(uas$pro_russian_vote[uas$russian_tv==0])
## [1] 0.1191139
```

RECALL: The difference-in-means estimator is defined as the average outcome for the treatment group minus the average outcome for the control group:

$$\bar{Y}_{\text{treatment group}} - \bar{Y}_{\text{control group}}$$

It produces a valid estimate of the average treatment effect when treatment and control groups are comparable, that is, when there are no confounders present.

RECALL: In R, `mean()` calculates the mean of a variable and `[]` is the operator used to extract a selection of observations from a variable. Example: `mean(data$var1[data$var2==1])` calculates the mean of the observations of the variable `var1` for which the variable `var2` equals 1.

RECALL: The difference-in-means estimator is measured in:

- the same unit of measurement as Y , if Y is non-binary
- percentage points (after multiplying the output by 100), if Y is binary.

Here, since *pro_russian_vote* is binary, the estimator is measured in percentage points (after multiplying the output by 100).

Based on this output, we would write the following conclusion statement: Assuming that respondents who received Russian TV were comparable to those who did not, we estimate that receiving Russian TV increased a respondent's probability of voting for a pro-Russian party by 12 percentage points, on average.

As we will see next, we can arrive at the same estimate by fitting a line where X is our treatment variable and Y is our outcome variable of interest. Then, the estimated slope coefficient ($\hat{\beta}$) is numerically equivalent to the difference-in-means estimator.

TO COMPUTE THE DIFFERENCE-IN-MEANS ESTIMATOR: We can either

- (a) calculate it directly, or
- (b) fit a simple linear model where Y is our outcome variable of interest and X is the treatment variable. In this case, the estimated slope coefficient ($\hat{\beta}$) is equivalent to the difference-in-means estimator.

Recall that the formula of the fitted line is:

$$\hat{Y} = \hat{\alpha} + \hat{\beta}X$$

where the estimated slope coefficient ($\hat{\beta}$) equals the change in the predicted outcome associated with a one-unit increase in X .

RECALL: In this book, we define a treatment variable, X , as binary and identifying receipt of treatment:

$$X_i = \begin{cases} 1 & \text{if individual } i \text{ received} \\ & \text{the treatment} \\ 0 & \text{if individual } i \text{ did not receive} \\ & \text{the treatment} \end{cases}$$

When X is the treatment variable, a one-unit increase in X occurs when X changes from 0 to 1, since those are the only two values that the treatment variable can take. This increase in X is equivalent to changing from not receiving the treatment ($X=0$) to receiving the treatment ($X=1$). The value of $\hat{\beta}$ is, therefore, the estimated average change in the outcome variable ($\Delta\hat{Y}$) associated with the change from the control condition to the treatment condition, also known as the difference-in-means estimator. (See the formula in detail below for a step-by-step explanation.)

FORMULA IN DETAIL

As we learned in chapter 4, the estimated slope coefficient equals the change in \hat{Y} associated with a one-unit increase in X :

$$\hat{\beta} = \Delta\hat{Y} \quad (\text{if } \Delta X=1)$$

The change in \hat{Y} can be calculated as $\hat{Y}_{\text{final}} - \hat{Y}_{\text{initial}}$:

$$\hat{\beta} = \hat{Y}_{\text{final}} - \hat{Y}_{\text{initial}} \quad (\text{if } \Delta X = 1)$$

When the X is the treatment variable, a one-unit increase in X is equivalent to changing from the control group ($X=0$) to the treatment group ($X=1$). This makes the control group the initial state, and the treatment group the final state:

$$\hat{\beta} = \hat{Y}_{\text{treatment group}} - \hat{Y}_{\text{control group}}$$

Finally, recall that \hat{Y} is an *average* predicted value. In this case, it turns out that the \hat{Y} s are exactly equal to the \bar{Y} s for their respective groups. The estimated slope coefficient is, then:

$$\hat{\beta} = \bar{Y}_{\text{treatment group}} - \bar{Y}_{\text{control group}}$$

When in a fitted linear model the X variable is the treatment variable, then the estimated slope coefficient $\hat{\beta}$ is numerically equivalent to the difference-in-means estimator.

Now, let's take a moment to figure out the substantive interpretation of $\hat{\beta}$ in this model. As we just saw, $\hat{\beta}$ is equivalent to the difference-in-means estimator, which, under certain conditions, produces a valid estimate of the average treatment effect, defined as the average change in the outcome variable *caused by* a change in the treatment variable. As a result, when interpreting $\hat{\beta}$ in a linear model where X is the treatment variable, we use causal as opposed to predictive language. We interpret the value of $\hat{\beta}$ as the estimated change in the outcome variable *caused by*, not just *associated with*, the treatment. The validity of this causal interpretation depends on the extent to which the treatment and control groups are comparable, that is, on the absence of confounding variables.

INTERPRETATION OF THE ESTIMATED SLOPE COEFFICIENT IN THE SIMPLE LINEAR MODEL:

- By default, we interpret $\hat{\beta}$ using predictive language: It is the $\Delta \hat{Y}$ *associated with* $\Delta X = 1$.
- When X is the treatment variable, then $\hat{\beta}$ is equivalent to the difference-in-means estimator, and thus, we interpret $\hat{\beta}$ using causal language: It is the $\Delta \hat{Y}$ *caused by* $\Delta X = 1$ (the presence of the treatment). This causal interpretation is valid if there are no confounding variables present, and thus, the treatment and control groups are comparable.

RECALL: This model uses the true values of α , β , and ϵ_i (that is, without the hats) because it is the theoretical model that we assume reflects the true relationship between X and Y . Since we do not know these values, we have to estimate them by fitting the model to the data.

Turning back to the running example, given that our treatment variable is *russian_tv* and our outcome variable is *pro_russian_vote*, the linear model we are interested in is:

$$pro_russian_vote_i = \alpha + \beta russian_tv_i + \epsilon_i \quad (i=\text{respondents})$$

where:

- *pro_russian_vote_i* is the binary variable that identifies whether respondent *i* voted for a pro-Russian party in the 2014 Ukrainian parliamentary election
- *russian_tv_i* is the treatment variable, which indicates whether the precinct where respondent *i* lives received Russian TV
- ϵ_i is the error term for respondent *i*.

RECALL: `lm()` fits a linear model. It requires a formula of the type $Y \sim X$. To specify the object where the dataframe is stored, we can use the optional argument `data` or the `%>%` character. Examples: `lm(y_var ~ x_var, data=data)` or `lm(data%>%y_var ~ data%x_var)`.

To fit the linear model to the data, we use the `lm()` function:

```
lm(pro_russian_vote ~ russian_tv,
   data=uas) # fits linear model
##
## Call:
## lm(formula = pro_russian_vote ~ russian_tv, data = uas)
##
## Coefficients:
## (Intercept)    russian_tv
##      0.1709         0.1191
```

RECALL: The fitted model uses the estimated coefficients, $\hat{\alpha}$ and $\hat{\beta}$, but it does not include ϵ_i (the residuals or error terms). For every value of X , the fitted model provides an average value of Y , that is, the value of \hat{Y} on the line.

Based on the output, the fitted linear model is:

$$\widehat{pro_russian_vote} = 0.17 + 0.12 russian_tv$$

In this type of analysis, we typically go straight to the interpretation of $\hat{\beta}$, since that is the coefficient that helps us estimate the average treatment effect.

RECALL: The estimated slope coefficient, $\hat{\beta}$, is measured in:

- the same unit of measurement as Y , if Y is non-binary
- percentage points (after multiplying the output by 100), if Y is binary.

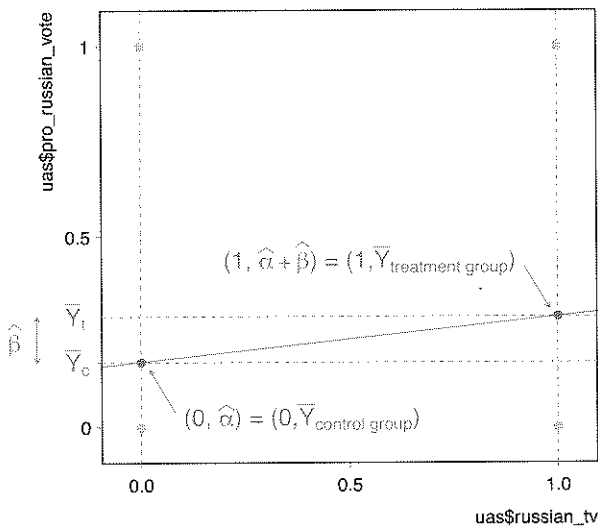
Here, since *pro_russian_vote* is binary, $\hat{\beta}$ is measured in percentage points (after multiplying the output by 100).

How should we interpret $\hat{\beta}=0.12$? The value of $\hat{\beta}$ equals the $\Delta\hat{Y}$ associated with $\Delta X=1$, and because here *russian_tv* (the X variable in the model) is the treatment variable, $\hat{\beta}$ is also equivalent to the difference-in-means estimator. (Note that the value of $\hat{\beta}$ is indeed the same value we arrived at above, when we calculated the difference-in-means estimator directly.) As a result, we interpret the value of $\hat{\beta}$ as estimating that receiving Russian TV (as compared to not receiving it) increased a respondent's probability of voting for a pro-Russian party by 12 percentage points, on average. This causal interpretation would be valid if respondents who received Russian TV were comparable to those who did not. (In the formula in detail below, we show how the fitted line on the scatter plot relates to the substantive interpretation of the two coefficients in this model.)

FORMULA IN DETAIL

As shown in the scatter plot, if X is the treatment variable:

- $\hat{\alpha} + \hat{\beta}$, which is the height of the point on the line that corresponds to $X=1$, can be interpreted as the average outcome for the treatment group ($\bar{Y}_{\text{treatment group}}$)
- $\hat{\alpha}$, which is the height of the point on the line that corresponds to $X=0$, can be interpreted as the average outcome for the control group ($\bar{Y}_{\text{control group}}$)
- $\hat{\beta}$, which is the difference between these two heights, is then equivalent to the difference-in-means estimator ($\bar{Y}_{\text{treatment group}} - \bar{Y}_{\text{control group}}$).



TIP: When X and Y are both binary, the scatter plot will show at most four dots representing all observations in the dataset. These correspond to the only four possible combinations of 0s and 1s: (0,1), (1,1), (1,0), and (0,0). In this case, we will not be able to discern how many observations in the dataset have the same combination of values because the dots that represent them will be displayed on top of each other.

In the running example:

- $\hat{\alpha} + \hat{\beta} = 0.29$; indicates that 29 percent of the respondents who lived in a precinct with Russian TV reception ($russian_tv=1$) voted for a pro-Russian party
- $\hat{\alpha} = 0.17$; indicates that 17 percent of the respondents who lived in a precinct without Russian TV reception ($russian_tv=0$) voted for a pro-Russian party
- $\hat{\beta} = 0.12$; estimates that receiving Russian TV increased the probability of voting for a pro-Russian party by 12 percentage points, on average ($29\% - 17\% = 12$ p.p.).

RECALL: The predicted outcome, \hat{Y} , and the estimated intercept coefficient, $\hat{\alpha}$, are measured in:

- the same unit of measurement as Y , if Y is non-binary
- percentages (after multiplying the output by 100), if Y is binary.

Here, since $pro_russian_vote$ is binary, both \hat{Y} and $\hat{\alpha}$ are measured in percentages (after multiplying the output by 100).

Had the UA_survey dataset come from a randomized experiment, we could interpret the difference-in-means estimator as a valid estimate of the average treatment effect. Here, we are working with observational data, however, and so we need to worry about potential confounders.

5.3.2 CONTROLLING FOR CONFOUNDERS USING A MULTIPLE LINEAR REGRESSION MODEL

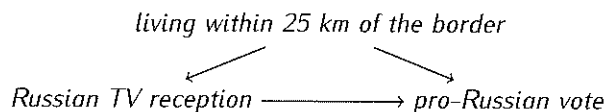
When dealing with observational data, our first step should be to identify every potential confounding variable in the relationship between X and Y . In the case at hand, we might worry about whether living *very* close to the Russian border affected both (i) the likelihood of receiving Russian TV and (ii) respondents' attitudes toward pro-Russian parties.

TIP: A confounder can affect the likelihood of receiving the treatment and the outcome in opposite directions. In our example, living very close to the border might increase the chances of receiving Russian TV but decrease the probability of voting for a pro-Russian party.

On the one hand, residents living very close to the border should be more likely to receive Russian TV, given their geographical proximity to Russian TV transmitters ($Z \rightarrow X$). On the other hand, given the military fortifications along the border during this time period, residents living very close to the border were probably less likely to vote for a pro-Russian party ($Z \rightarrow Y$).

In the months leading up to the 2014 election, Ukraine prepared to defend itself from a possible Russian invasion by deploying its army to the border. The Ukrainian army built military fortifications (trenches and defensive walls) at a distance of up to 10 km from the border, depending on local terrain and road access. Within that buffer zone, the army positioned tanks and troops in strategic locations and set up military checkpoints. Residents of a precinct located very close to the border (such as within 25 km of it) were either in immediate proximity of a military fortification or, at the very least, aware of its existence, making them especially cognizant of the threat of a Russian invasion, and therefore, more fearful of Russian influence.

In summary, living very close to the border may affect both the treatment and outcome variables and is, therefore, a potential confounding variable. (See the diagram below, which represents the causal relationships between the three variables of interest.)



In the `UA_survey` dataset, the variable `within_25km` identifies whether a respondent lived in a precinct within 25 km of the border, and thus, measures our confounder. We can confirm that the confounding variable, `within_25km`, and the treatment variable, `russian_tv`, are related to each other by computing their correlation coefficient:

```

## compute correlation
cor(uas$within_25km, uas$russian_tv)
## [1] 0.8127747

```

Based on the output above, *within_25km* and *russian_tv* are highly correlated with each other. As we know, this does not necessarily mean that changes in one variable cause changes in the other. The positive correlation, however, does mean that a higher value of *within_25km* is associated with a higher value of *russian_tv*, on average. Since both variables are binary, *russian_tv* is more likely to equal 1 when *within_25km* also equals 1. To confirm this, we can create the two-way table of frequencies by running:

```
## create two-way table of frequencies
table(uas$within_25km, uas$russian_tv)
##      0  1
### 0 139 14
### 1  19 186
```

As shown in the table above, among respondents living within 25 km of the border, about 91% are in a precinct that receives Russian TV ($186 \div (19 + 186) = 0.91$). In contrast, among respondents living more than 25 km away from the border, about 9% are in a precinct that receives Russian TV ($14 \div (139 + 14) = 0.09$). Compared to Ukrainians living farther away from the border, then, those living very close to it (i) are more likely to receive Russian TV and (ii) might have many different observed and unobserved characteristics that affect their propensity to vote for a pro-Russian party, including being more aware of the threat of a Russian invasion.

Once we have identified the potential confounders, the next step is to statistically control for them by fitting a multiple linear regression model. In contrast to simple linear regression models, multiple linear regression models have more than one *X* variable ("multi" means more than one). The multiple linear regression model is defined as:

$$Y_i = \alpha + \beta_1 X_{i1} + \dots + \beta_p X_{ip} + \epsilon_i$$

where:

- Y_i is the outcome for observation i
- α is the intercept coefficient
- each β_j is the coefficient for variable X_j —we use j as a stand-in for all the different subscripts from 1 to p ($j=1, \dots, p$)
- each X_{ij} is the observed value of the variable X_j for observation i ($j=1, \dots, p$)
- p is the total number of *X* variables in the model
- ϵ_i is the error term for observation i .

Just as the simple linear model, this is a theoretical model that is assumed to reflect the true relationship between all the *X* variables and *Y*. Because we do not know the values of any of

RECALL: `table()` creates a two-way frequency table when two variables are specified as required arguments. Example: `table(data$variable1, data$variable2)`. In the output, the values of the variable specified as the first argument in the function are shown in the rows; the values of the second variable are shown in the columns.

TIP: In this two-way table of frequencies, very few observations are in the off-diagonal (the diagonal running from the upper right to the lower left). There are only 14 respondents living more than 25 km away from the border who receive Russian TV, and there are only 19 respondents living within 25 km of the border who do not receive Russian TV. This suggests that *within_25km* is a strong confounder and that our estimate of the average treatment effect will rest on this small number of observations.

Multiple linear regression models are linear models with more than one *X* variable.

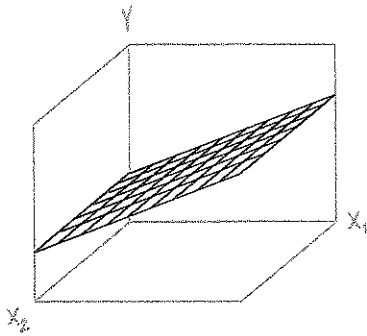
the coefficients ($\alpha, \beta_1, \beta_2, \dots, \beta_p$) or of the error terms (ϵ_i), we have to estimate them by fitting the model to the data.

In this case, the fitted model can be written as:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1 X_{i1} + \dots + \hat{\beta}_p X_{ip}$$

where:

- \hat{Y}_i is the predicted value of Y for observation i
- $\hat{\alpha}$ is the estimated intercept coefficient
- each $\hat{\beta}_j$ (pronounced beta hat sub j) is the estimated coefficient for variable X_j ($j=1, \dots, p$)
- each X_{ij} is the observed value of the variable X_j for observation i ($j=1, \dots, p$)
- p is the total number of X variables in the model.



Note that the simple regression linear model is a special case of the multiple linear regression model (the case in which p equals 1). When there is only one X variable, the fitted model is a line, and we are back in the simple linear regression model. For any p other than 1, the fitted model is not a line. If p equals 2, for instance, the fitted model is a plane in a three-dimensional space. (See example plane in the margin.)

Table 5.2 provides the mathematical definitions of each of the coefficients in the multiple linear regression model. As we can see there, the definitions of the coefficients in the simple linear regression model can be derived from those in the multiple linear regression model by setting the number of X variables to one.

TABLE 5.2. Mathematical definition of coefficients in the multiple and simple linear regression models. (Note: The Latin expression *ceteris paribus* here means holding all other X variables constant.)

In the multiple linear regression model:

- the value of $\hat{\alpha}$ equals the \hat{Y} when all X variables equal 0
- the value of each $\hat{\beta}_j$ equals the $\Delta\hat{Y}$ associated with $\Delta X_j=1$, while holding all other X variables constant.

multiple regression	simple regression
$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_p X_p$	$\hat{Y} = \hat{\alpha} + \hat{\beta} X$
$\hat{\alpha}$: \hat{Y} when all $X_j=0$ ($j=1, \dots, p$)	$\hat{\alpha}$: \hat{Y} when $X=0$
each $\hat{\beta}_j$: $\Delta\hat{Y}$ associated with $\Delta X_j=1$, while holding all other X variables constant or <i>ceteris paribus</i>	$\hat{\beta}$: $\Delta\hat{Y}$ associated with $\Delta X=1$

Let's look at the definition of each coefficient in turn:

- When there are multiple X variables, the value of $\hat{\alpha}$ equals the predicted value of Y when all X variables equal zero. When there is only one X variable, the value of $\hat{\alpha}$ equals the predicted value of Y when that one X variable equals zero.

- When there are multiple X variables, there will be multiple $\hat{\beta}$ coefficients (one for each X variable). The value of each $\hat{\beta}_j$ equals the predicted change in Y associated with a one-unit increase in X_j (the X variable affected by $\hat{\beta}_j$), *while holding all other X variables constant*. When there is only one X variable, there will be only one $\hat{\beta}$ coefficient. The value of $\hat{\beta}$ equals the predicted change in Y associated with a one-unit increase in the one X variable included in the model. (Since there are no other X variables here, there is no need to hold them constant.)

How can the multiple linear regression model help us estimate average causal effects when confounders are present?

Let's assume the first X variable (X_1) is the treatment variable. The value of the corresponding estimated coefficient ($\hat{\beta}_1$) equals the change in \hat{Y} *associated with* the presence of the treatment, while holding all the other X variables constant.

Now, if the model includes each potential confounding variable we are worried about as an additional X variable (that is, as a "control variable"), then the value of $\hat{\beta}_1$ equals the change in \hat{Y} *caused by* the presence of the treatment, while holding the values of all confounding variables constant. In other words, now we can interpret $\hat{\beta}_1$ using causal language because statistically controlling for all confounding variables in the estimation process makes the treatment and control groups comparable.

To better understand this, let's look at the diagram shown in figure 5.4, which represents the causal relationships between a confounding variable, Z , the treatment variable, X , and the outcome variable, Y . Intuitively, by adding Z as a control variable in the model, we statistically hold the values of Z constant, blocking the path shown with a gray dashed line, which links X and Y through Z . With this path blocked, no changes in Y can be attributed to changes in Z . Since the value of Z is being held constant, the only remaining source of change in Y is a change in X .



TIP: In this model, only the estimated coefficient that affects the treatment variable, $\hat{\beta}_1$, can be interpreted using causal language; all others should continue to be interpreted using predictive language.

FIGURE 5.4. Representation of the causal relationships between a confounding variable, Z , the treatment variable, X , and the outcome variable, Y . The path blocked by adding Z as a control variable in the model is shown with a gray dashed line.

In other words, the difference in the average outcomes between the treatment and control groups that remains after holding all confounding variables constant can now be directly attributed to their difference with respect to the treatment (treated vs. untreated); no other differences between the two groups are in play.

Post-treatment variables are variables affected by the treatment:

$X \rightarrow \text{post-treatment variable}$

Example: if the treatment is attending a private school and private schools have smaller classes than public schools, then *small class* is a post-treatment variable because it is affected by the value of *private school*.

$\text{private school} \rightarrow \text{small class}$

FIGURE 5.5. Representation of the potential causal relationships between a post-treatment variable, V , the treatment variable, X , and the outcome variable, Y . The path blocked by adding V as a control variable in the model is shown with a gray dashed line.



Does this mean that we should add to the model as many control variables as possible? No. For example, we should make sure *not* to control for post-treatment variables, which are variables affected by the treatment. Adding a post-treatment variable to the model would render our causal estimates invalid because we would be controlling for a consequence of the treatment when trying to estimate its total effect.

To illustrate this, consider the causal diagram in figure 5.5. Suppose that we control for the post-treatment variable V when estimating the causal effect of X on Y . Doing so would block the causal path going from X to Y through V , which is one of the ways by which changes in X cause changes in Y , and therefore represents a portion of the total causal effect of X on Y .

In our current analysis, for example, we would not want to add to the model a variable capturing the average number of hours a respondent spent watching Russian TV each week. This is a post-treatment variable since it is causally affected by the treatment; its value directly depends on whether a respondent received Russian TV to begin with. Thus, controlling for this variable would soak up part of the causal effect we are interested in estimating.

ESTIMATING AVERAGE CAUSAL EFFECTS USING OBSERVATIONAL DATA AND MULTIPLE LINEAR REGRESSION MODELS. If, in the multiple linear regression model where X_1 is the treatment variable, we control for *all* potential confounders by including them in the model as additional X variables, then we can interpret $\hat{\beta}_1$ as a valid estimate of the average causal effect of X on Y .

Now that we know how to estimate an average treatment effect in the presence of confounders, let's return to our example. Given that our treatment variable is *russian_tv*, our outcome variable is *pro_russian_vote*, and our confounding variable is *within_25km*, the linear model we are interested in is:

$$\text{pro_russian_vote}_i = \alpha + \beta_1 \text{russian_tv}_i + \beta_2 \text{within_25km}_i + \epsilon_i \quad (i=\text{respondents})$$

To fit a multiple linear regression model in R, we also use the `lm()` function. As you may recall, this function requires a formula of the type $Y \sim X$ when there is only one X variable. It requires a formula of the type $Y \sim X_1 + \dots + X_p$ when there are multiple X variables. For example, to fit the linear model above, we run:

```
lm(pro_russian_vote ~ russian_tv + within_25km,
   data=uas) # fits linear model
##
## Call:
## lm(formula = pro_russian_vote ~ russian_tv +
##   within_25km, data=uas)
##
## Coefficients:
## (Intercept)  russian_tv  within_25km
##    0.1959      0.2876     -0.2081
```

Based on the output, the new fitted linear model is:

$$\widehat{pro_russian_vote} = 0.2 + 0.29 \text{ russian_tv} - 0.21 \text{ within_25km}$$

How should we interpret $\widehat{\beta}_1=0.29$? The value of $\widehat{\beta}_1$ equals the $\Delta\widehat{Y}$ associated with $\Delta X_1=1$, while holding all other variables constant. In addition, because the variable affecting this coefficient is the treatment variable, *russian_tv*, and the confounder we are worried about, *within_25*, is included in the model as a control variable, we can interpret $\widehat{\beta}_1$ using causal language. Thus, we interpret the value of $\widehat{\beta}_1$ as estimating that, when we hold living very close to the border constant, receiving Russian TV (as compared to not receiving it) increased a respondent's probability of voting for a pro-Russian party by 29 percentage points, on average. The validity of this causal interpretation depends on whether living very close to the border is the only confounding variable. If there are other confounders, this estimate of the average treatment effect would not be valid.

5.4 THE EFFECT OF RUSSIAN TV ON UKRAINIAN ELECTORAL OUTCOMES

In the prior section, we found that Russian TV reception was estimated to increase a respondent's probability of voting for a pro-Russian party, suggesting that the propaganda aired by Russian TV in the months leading up to the 2014 Ukrainian parliamentary election may have helped parties with a pro-Russian agenda garner more votes. In this section, we examine whether we can find a similar causal relationship at the aggregate level. This analysis is particularly appropriate since the treatment variable itself (Russian TV reception) is measured at the precinct level.

`lm()` fits a linear model. It requires a formula of the type $Y \sim X_1 + \dots + X_p$. Note that when there is only one X variable, this formula becomes $Y \sim X$. To specify the object where the dataframe is stored, we can use the optional argument `data` or the `$` character. Examples: `lm(y_var ~ x_var1 + x_var2, data=data)` or `lm(data$y_var ~ data$x_var1 + data$x_var2)`.

TIP: The unit of measurement of $\widehat{\beta}_1$ in the multiple linear regression model follows the same rules as the unit of measurement of $\widehat{\beta}$ in the simple linear regression model. Here, since *pro_russian_vote* is binary, $\widehat{\beta}_1$ is measured in percentage points (after multiplying the output by 100).

RECALL: In an individual-level analysis, the unit of observation is individuals. By contrast, in an aggregate-level analysis, the unit of observation is collections of individuals. For example, here our unit of observation is precincts, and therefore, each observation represents the residents of a particular precinct.

Here, we use aggregate-level data from all the precincts in three provinces in northeastern Ukraine: Chernihiv, Sumy, and Kharkiv. Among the Ukrainian provinces bordering Russia, only these three did not close their polling stations as a result of the ongoing conflict. These are the same provinces where the respondents to the survey analyzed above lived.

The dataset is provided in the file "UA_precincts.csv". Table 5.3 shows the names and descriptions of the variables included.

TABLE 5.3. Description of the variables in the UA_precincts dataset, where the unit of observation is precincts.

variable	description
<i>russian_tv</i>	identifies precincts that receive Russian TV: 1=there is reception or 0=there is no reception
<i>pro_russian</i>	vote share received in the precinct by pro-Russian parties in the 2014 Ukrainian parliamentary election (in percentages)
<i>prior_pro_russian</i>	vote share received in the precinct by pro-Russian parties in the 2012 Ukrainian parliamentary election (in percentages)
<i>within_25km</i>	identifies precincts that are within 25 kilometers of the Russian border: 1=it is within 25 kilometers of the border or 0=it is not within 25 kilometers of the border

RECALL: If the DSS folder is saved directly on your Desktop, to set the working directory, you must run `setwd("~/Desktop/DSS")` if you have a Mac and `setwd("C:/user/Desktop/DSS")` if you have a Windows computer (where *user* is your own username). If the DSS folder is saved elsewhere, please see subsection 1.7.1 for instructions on how to set the working directory.

As always, we start by reading and storing the data (assuming we have already set the working directory):

```
uap <- read.csv("UA_precincts.csv") # reads and stores data
```

To get a sense of the dataset, we look at the first few observations:

```
head(uap) # shows first observations
##   russian_tv pro_russian prior_pro_russian within_25km
## 1         0  2.7210884      25.14286         1
## 2         0  0.8928571      35.34483         0
## 3         1  1.6949153      20.53232         1
## 4         0  72.2689076     84.47761         1
## 5         0  1.2820513      28.99408         0
## 6         1  1.4285714      45.58824         0
```

Based on table 5.3 and the output of above, we learn that each observation in the dataset represents a precinct, and that the dataset contains four variables:

- *russian_tv* is a binary variable that identifies whether the precinct received Russian TV
- *pro_russian* and *prior_pro_russian* are the vote shares received by pro-Russian parties in the precinct in the parliamentary

elections of 2014 and 2012, respectively (both variables are measured in percentages)

- *within_25km* is a binary variable that identifies whether the precinct is within 25 km of the border.

We interpret the first observation as representing a precinct in Ukraine that does not receive Russian TV, where pro-Russian parties received about 3% and 25% of the votes in the parliamentary elections of 2014 and 2012, and that is within 25 km of the border with Russia.

To find the total number of observations in the dataset, we run:

```
dim(uap) # provides dimensions of dataframe: rows, columns
## [1] 3589 4
```

The dataset contains information about 3,589 precincts.

5.4.1 USING THE SIMPLE LINEAR MODEL TO COMPUTE THE DIFFERENCE-IN-MEANS ESTIMATOR

In this analysis, we are interested in estimating the effect that the intense, one-sided Russian TV coverage of Ukrainian politics had on the electoral performance of pro-Russian parties in the 2014 Ukrainian parliamentary election at the precinct level. Since the treatment took place between the 2012 and 2014 elections, we define our outcome variable as the change in the vote share received by pro-Russian parties between these two elections.

The causal link we are interested in is, then, between *russian_tv* and *pro_russian_change*, where *russian_tv* is the treatment variable and *pro_russian_change* is the outcome variable.

Russian TV reception → pro-Russian vote share change

Since we do not have our outcome variable of interest readily available in the dataset, we start the analysis by creating it. The change in the precinct-level vote share received by pro-Russian parties between 2012 and 2014 is defined as:

$$pro_russian_change = pro_russian - prior_pro_russian$$

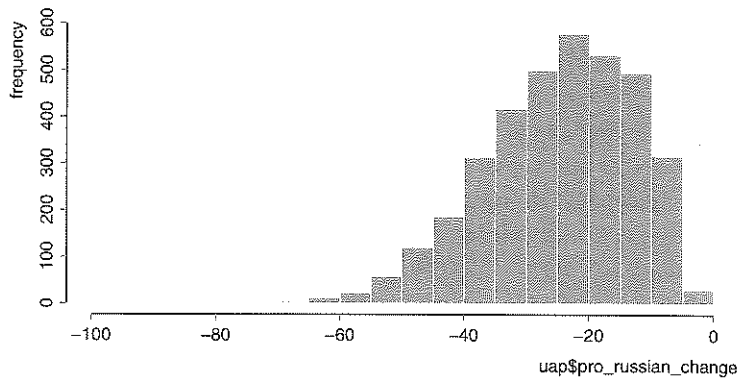
To create this variable, we run:

```
## create pro-russian change variable
uap$pro_russian_change <-
  uap$pro_russian - uap$prior_pro_russian
```

The new variable, *pro_russian_change*, is measured in percentage points because it is the difference between two percentages. For example, it equals -20 p.p. when a precinct's pro-Russian vote share dropped to 40% from 60% (40%–60%=-20 p.p.).

To get a sense of the contents of *pro_russian_change*, we can create its histogram by running:

```
## create histogram
hist(uap$pro_russian_change)
```



Note that all the values of *pro_russian_change* are negative, which means that in all the precincts under study, the vote share received by pro-Russian parties decreased between these two elections. As a result of the conflict leading up to the 2014 election, pro-Russian political parties lost support across the country, even in their traditional strongholds in eastern and southern Ukraine. Our question is, then, whether Russian TV reception caused the precinct-level vote share for pro-Russian parties to decline by a smaller amount.

To calculate the difference-in-means estimator, we can fit a simple linear model without any controls, just as we did in the previous section. The linear model we are interested in here is:

$$pro_russian_change_i = \alpha + \beta russian_tv_i + \epsilon_i \quad (i=\text{precincts})$$

where:

- *pro_russian_change_i* is the percentage-point change in the vote share received by pro-Russian parties in precinct *i* between the 2012 and 2014 Ukrainian parliamentary elections
- *russian_tv_i* is the treatment variable, which indicates whether precinct *i* received Russian TV
- ϵ_i is the error term for precinct *i*.

To fit the linear model, we run:

```
lm(pro_russian_change ~ russian_tv,
   data=uap) # fits linear model
##
## Call:
## lm(formula = pro_russian_change ~ russian_tv, data=uap)
##
## Coefficients:
## (Intercept)    russian_tv
##    -25.146         1.783
```

Based on the output, the fitted linear model is:

$$\widehat{\text{pro_russian_change}} = -25.15 + 1.78 \text{ russian_tv}$$

How should we interpret $\hat{\beta}=1.78$? The value of $\hat{\beta}$ equals the $\Delta\hat{Y}$ associated with $\Delta X=1$, and because *russian_tv* (the X variable in the model) is the treatment variable, $\hat{\beta}$ is also equivalent to the difference-in-means estimator. As a result, we interpret the value of $\hat{\beta}$ as estimating that receiving Russian TV (as compared to not receiving it) increased the change in the precinct-level vote share received by pro-Russian parties by 1.78 percentage points, on average. Note that the positive sign of $\hat{\beta}$ is consistent with our expectation regarding the effect of Russian TV propaganda. It indicates that pro-Russian parties experienced smaller vote share losses in precincts with Russian TV reception. The validity of this causal effect estimate depends on whether the precincts that received Russian TV were comparable to the precincts that did not; that is, it depends on the absence of confounding variables.

RECALL: The estimated slope coefficient, $\hat{\beta}$, is measured in:

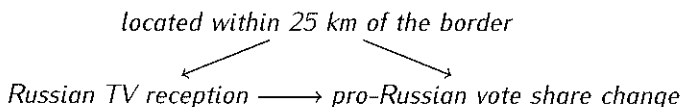
- the same unit of measurement as Y , if Y is non-binary
- percentage points (after multiplying the output by 100), if Y is binary.

Here, since *pro_russian_change* is non-binary and measured in percentage points, $\hat{\beta}$ is also measured in percentage points.

5.4.2 CONTROLLING FOR CONFOUNDERS USING A MULTIPLE LINEAR REGRESSION MODEL

A confounding variable we might worry about, again, is close proximity to the border. On the one hand, precincts very close to the border should be more likely to receive Russian TV ($Z \rightarrow X$). On the other hand, given the military deployments along the border, we might expect that pro-Russian parties experienced larger vote share losses in precincts very close to the border ($Z \rightarrow Y$).

Given that close proximity to the border (defined here as being within 25 km) affects both (i) the likelihood of receiving the treatment and (ii) the outcome, it constitutes a confounder.



In the `UA_precincts` dataset, the variable `within_25km` captures whether a precinct is within 25 km of the border, and thus, measures our confounder. We can confirm that the confounding variable, `within_25km`, and the treatment variable, `russian_tv`, are related to each other by computing their correlation coefficient:

```
## compute correlation
cor(uap$within_25km, uap$russian_tv)
## [1] 0.5317845
```

Based on the output above, `within_25km` and `russian_tv` are moderately correlated with each other.

Now that we have identified the confounding variable, we are ready to fit a multiple linear regression model to estimate the average treatment effect. Here, since the treatment variable is `russian_tv` and the potential confounding variable is `within_25km`, the linear model we are interested in is:

$$\begin{aligned} \text{pro_russian_change}_i &= \alpha + \beta_1 \text{russian_tv}_i \\ &+ \beta_2 \text{within_25km}_i + \epsilon_i \quad (i=\text{precincts}) \end{aligned}$$

To fit the multiple linear regression model above, we run:

```
lm(pro_russian_change ~ russian_tv + within_25km,
    data=uap) # fits linear model
##
## Call:
## lm(formula = pro_russian_change ~ russian_tv +
##   within_25km, data = uap)
##
## Coefficients:
##   (Intercept)    russian_tv    within_25km
##   -24.302         8.822         -14.614
```

Based on the output, the new fitted linear regression model is:

$$\widehat{\text{pro_russian_change}} = -24.3 + 8.82 \text{russian_tv} - 14.61 \text{within_25km}$$

How should we interpret $\hat{\beta}_1=8.82$? The value of $\hat{\beta}_1$ equals the $\Delta \hat{Y}$ associated with $\Delta X_1=1$, while holding all other variables constant. In addition, because the variable affecting this coefficient is the treatment variable, `russian_tv`, and the confounder we are worried about, `within_25`, is included in the model as a control variable, we can interpret $\hat{\beta}_1$ using causal language. Thus, we interpret the value of $\hat{\beta}_1$ as estimating that, when we hold close proximity to the border constant, receiving Russian TV (as compared to not receiving it) increased the change in

the precinct-level vote share received by pro-Russian parties by 8.82 percentage points, on average. If the close proximity of the precincts to the border successfully captures the only confounding variable in the relationship between our two main variables of interest, then this is a valid estimate of the average treatment effect.

5.5 INTERNAL AND EXTERNAL VALIDITY

We have already learned how to estimate the average change in the outcome caused by the treatment. (In chapter 2, we saw how to estimate the average treatment effect using data from a randomized experiment, and in this chapter, we have seen how to estimate it using observational data.) There are more issues we must consider when conducting or evaluating a scientific causal study, including the following two properties: (i) internal validity and (ii) external validity.

The *internal validity* of a study refers to the extent to which the causal assumptions are satisfied. In other words, it reflects the confidence we have in our causal estimates. It asks, is the estimated causal effect valid for the sample of observations in the study? The answer depends on whether we have successfully eliminated or controlled for all potential confounders, that is, on whether the treatment and control groups used for the estimation can be considered comparable, after statistical controls are applied (if any are).

The *external validity* of a study refers to the extent to which the conclusions can be generalized. It asks, is the estimated causal effect valid beyond this particular study? The answer depends on (i) whether the sample of observations in the study is representative of the population to which we want to generalize the results, and (ii) whether the treatment used in the study is representative of the treatment for which we want to generalize the results.

The internal validity of a study refers to the extent to which its causal conclusions are valid for the sample of observations in the study. The external validity of a study refers to the extent to which its causal conclusions can be generalized.

RECALL: In a representative sample, characteristics appear at similar rates as in the population as a whole.

5.5.1 RANDOMIZED EXPERIMENTS VS. OBSERVATIONAL STUDIES

How do studies based on experimental data compare to those based on observational data along these two dimensions?

When it comes to internal validity, randomized experiments have a significant advantage over observational studies. In experiments, the use of random treatment assignment eliminates all potential confounding variables. By contrast, in observational studies, while we can statistically control for observed confounders, there is always the possibility that we fail to account for unobserved confounders.

When it comes to external validity, randomized experiments can suffer from limitations that put them at a disadvantage compared to observational studies. First, for ethical and logistical reasons, randomized experiments are often done using a convenient sample of subjects who are willing to participate in the study. (For example, you have probably seen ads recruiting subjects for experiments in exchange for money.) In some cases, then, volunteers come from a particular segment of the population; they may be low-income and/or underemployed. In such cases, the sample of individuals would likely be non-representative of the whole population of interest. By contrast, in observational studies, we can usually analyze data from either the entire population or a random selection of observations from that population.

Second, randomized experiments are often conducted in artificial environments such as laboratories, making the treatments less realistic, and therefore, less comparable to real-world treatments. For example, it is not the same to watch a TV program in a laboratory as in the comfort of your own home, where many other things compete for your attention (phone calls, visits to the fridge, and TV programs on other channels). By contrast, in observational studies, we usually observe the treatment in the environment in which we are interested.

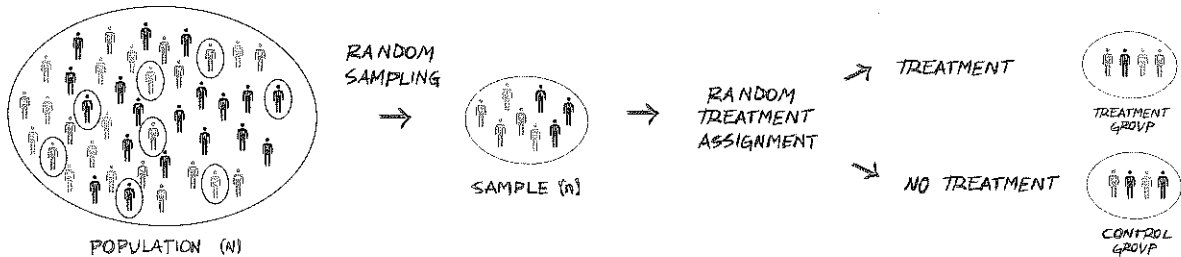
In summary, an advantage in internal validity often comes with a compromise in external validity, and vice versa. Studies based on randomized experiments tend to have strong internal validity but relatively weak external validity. Observational studies tend to have relatively weak internal validity but strong external validity. This dynamic explains why scholars use both types of studies to estimate causal effects; they often have complementary strengths. Nonetheless, some studies based on experimental data have strong external validity, and some studies based on observational data have strong internal validity. We should pay attention to the study details when evaluating them.

5.5.2 THE ROLE OF RANDOMIZATION

The ideal research design for estimating average treatment effects would make use of the two kinds of randomization we have seen. It would not only randomly select its observations from the population, but it would also randomly assign treatment among those observations. (See figure 5.6 on the next page.)

Assuming that we were also able to make the treatment as realistic as possible, this design would create a study with strong external and internal validity. As discussed, random sampling is the best way to make the sample representative of the population and, thereby, ensure strong external validity (enabling us to generalize the results to the target population). Similarly, ran-

dom treatment assignment is the best way to make treatment and control groups comparable and, thereby, ensure strong internal validity (enabling us to draw valid causal inferences).



As discussed in chapter 2 and above, for ethical, logistical, and financial reasons, few studies include both types of randomization. It is useful, however, to know what the ideal research design would look like; it serves as a benchmark when designing or evaluating causal studies.

FIGURE 5.6. The ideal research design would make use of the two kinds of randomization we have seen: random sampling and random treatment assignment.

5.5.3 HOW GOOD ARE THE TWO CAUSAL ANALYSES IN THIS CHAPTER?

Let's evaluate the internal and external validity of the two data analyses in this chapter: (i) the individual-level analysis and (ii) the precinct-level analysis.

How strong is their internal validity? In both analyses, receipt of the treatment (Russian TV reception) was determined by factors outside of the control of the researchers, such as the terrain and distance of the precincts to the Russian TV transmitters. Neither study is, therefore, a randomized experiment. Despite the fact that we cannot rely on the randomization of treatment assignment to eliminate all potential confounders, we can argue that both analyses have relatively strong internal validity.

First, once we focus on areas close to the Ukraine–Russia border (as we do in both cases), the variation in the reception of Russian TV plausibly yields an “as-if-random” assignment of the treatment; it is influenced by terrain and other factors that are likely to be unrelated to the level of support for pro–Russian parties. Second, we arguably remove any remaining differences between the treatment and control groups by statistically controlling for potential confounders. In both instances, we control for being very close to the border. If this is the only confounding variable present, then the internal validity of the analyses is strong.

How strong is their external validity? In the individual-level analysis, we use data from a random sample of individuals living in the precincts in which we are interested. In the aggregate-level analysis, we use data from all the Ukrainian precincts in which we are interested. In addition, in both studies we observe the treatment (that is, Russian TV reception) in its real-world environment. As a result, if we were interested in generalizing our results to the region from which our observations come, then, the external validity of both studies is strong. If we were interested in generalizing the conclusions to a different type of one-sided televised coverage of a political event in a different region of the world, we would have to assess to what degree the treatment and the observations in the analyses here are representative of the actual treatment and population of interest.

5.5.4 HOW GOOD WAS THE CAUSAL ANALYSIS IN CHAPTER 2?

As you may recall, in chapter 2 we analyzed the STAR dataset to estimate the effects of attending a small class on student performance. The data came from a randomized experiment conducted in Tennessee, where students were randomly assigned to attend either a small class or a regular-size class.

How strong is its internal validity? Since the treatment was assigned at random, all potential confounding variables should have been eliminated, making the group of students who attended a small class similar in all aspects to the group of students who attended a regular-size class. Thanks to random treatment assignment, then, the causal assumption is satisfied, and we can be confident that the causal estimates we arrived at are valid for the group of students who participated in the experiment. We can conclude that this analysis has strong internal validity.

See Diane Whitmore Schanzenbach, "What Have Researchers Learned from Project STAR?" *Brookings Papers on Education Policy*, no. 9 (2006): 205–28.

How strong is its external validity? Given the characteristics of the study, only students from large schools in Tennessee were able to participate in the experiment. As a result, the sample of participating students was not perfectly representative of all students in Tennessee. The sample was also not representative of students in the United States. For example, according to Schanzenbach (2006), the proportion of African Americans was larger in the sample than in the state overall, and the proportion of Hispanics and Asians was smaller in the sample than in the country as a whole. Consequently, we can conclude that, although we do get to observe the treatment of interest in the real world, the analysis has relatively weak external validity, especially if one wishes to generalize the study's conclusions to all schools and students in Tennessee or in the entire United States.

5.5.5 THE COEFFICIENT OF DETERMINATION, R^2

Note that at no point during our evaluation of the causal analyses did we mention any of the models' coefficient of determination, or R^2 . This statistic is of no direct relevance when estimating average treatment effects. A model with a small R^2 might do a fine job estimating a valid causal effect, especially when the effect is small and there are few (or no) confounders we need to control for statistically. Alternatively, a model with a large R^2 might estimate an invalid causal effect, especially if the confounders we control for explain a large variation of the outcome variable, yet controlling for them fails to make treatment and control groups comparable.

RECALL: R^2 , also known as the coefficient of determination, ranges from 0 to 1 and measures the proportion of the variation of the outcome variable explained by the model. The higher the R^2 , the better the model fits the data.

5.6 SUMMARY

In this chapter, we returned to estimating causal effects but, this time, using observational data. We learned about confounding variables and why their presence complicates the estimation of causal effects. We saw how to fit a simple linear model to compute the difference-in-means estimator and how to fit a multiple linear model to control for confounders. Finally, we discussed how to evaluate causal studies based on their internal and external validity.

The statistical method used in this chapter, fitting a linear regression model, is the same as the one we used in the previous chapter. (Although we did not see an example of it, social scientists often use multiple linear regression models to make predictions, and not just simple linear regression models.) The goals of the analyses, however, differ. In chapter 4, we aimed to *predict* a quantity of interest, while in this chapter we aimed to *explain* a quantity of interest (that is, to estimate a causal effect).

Even though the mathematical models are the same, the role the X variable plays in the research question, the substantive interpretations of the estimated coefficients, and what we pay attention to in the analysis depend on whether we are analyzing data to make predictions or to estimate causal effects.

For example, when fitting a simple linear regression model to make predictions:

- X is a predictor.
- We interpret $\hat{\beta}$ as the change in \hat{Y} associated with a one-unit increase in X .
- Since the goal is to make predictions with the smallest possible errors, we seek predictors that are highly correlated with the outcome variable of interest. The stronger the linear association between X and Y , the higher the R^2 and the better the fitted linear model will usually be at predicting Y using X .

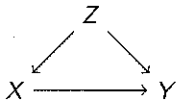
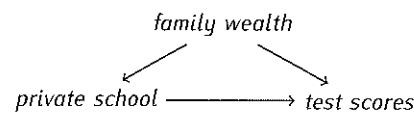
By contrast, when fitting a simple linear regression model to estimate causal effects:

- X is the treatment variable.
- We interpret $\hat{\beta}$ as the change in \hat{Y} *caused by* the presence of the treatment X .
- Since the goal is to arrive at valid estimates of causal effects, we seek to find or create situations in which the treatment and control groups used in the analysis can be considered comparable, after statistical controls are applied (if any are). In other words, we seek to eliminate or control for all potential confounding variables.

Thus, whenever we conduct a regression analysis or evaluate one conducted by someone else, we should keep the goal in mind.

5.7 CHEATSHEETS

5.7.1 CONCEPTS AND NOTATION

concept/notation	description	example(s)
confounding variable or confounder (Z)	<p>also known as an omitted variable or a control variable; variable that affects both (i) the likelihood of receiving the treatment X and (ii) the outcome Y</p>  <p> confounders obscure the causal relationship between X and Y; just because we observe that two variables are highly correlated with each other does not automatically mean that one causes the other; there could be a third variable—a confounder—that affects both variables </p> <p>in the presence of confounding variables, correlation does not necessarily imply causation, and the difference-in-means estimator does <i>not</i> provide a valid estimate of the average treatment effect</p> <p>in randomized experiments, the randomization of treatment assignment eliminates all potential confounding variables</p>	<p><i>family wealth</i> is a confounder in the causal relationship between attending a <i>private school</i> and <i>test scores</i></p>  <p>students from wealthy families are more likely to attend a private school (<i>family wealth</i> → <i>private school</i>); students from wealthy families are more likely to receive after-school help such as one-on-one tutoring, which, in turn, will improve performance on tests (<i>family wealth</i> → <i>tutoring</i> → <i>test scores</i>)</p> <p>in the presence of the confounder, <i>family wealth</i>, we do not know what portion (if any) of the observed difference in average test performance between private and public school students is due to the schools the students attend and what portion is due to their differing levels of family wealth</p>
fitted simple linear regression model where X is the treatment variable	<p>if, in the fitted simple linear regression model, X is the treatment variable, $\hat{\beta}$ is equivalent to the difference-in-means estimator, and thus, we interpret it using causal, not predictive, language</p> <p>this causal interpretation is valid if there are no confounding variables present</p>	<p>if X is the treatment variable and the fitted model is $\hat{Y} = 2 - 3X$:</p> <p>we interpret $\hat{\beta}$ as estimating that receiving the treatment decreases the outcome by 3 units, on average (in the same unit of measurement as the difference-in-means estimator)</p>
multiple linear regression model	<p>linear model with more than one X variable; theoretical model that we assume reflects the true relationship between Y and multiple X variables</p> $Y_i = \alpha + \beta_1 X_{i1} + \dots + \beta_p X_{ip} + \epsilon_i$ <p>where:</p> <ul style="list-style-type: none"> - Y_i is the outcome for observation i - α is the intercept coefficient - each β_j is the coefficient for variable X_j ($j=1, \dots, p$) - each X_{ij} is the observed value of the variable X_j for observation i ($j=1, \dots, p$) - p is the total number of X variables - ϵ_i is the error term for observation i 	$Y_i = 1 + 3X_{i1} + 5X_{i2} + \epsilon_i$

continues on next page...

5.7.1 CONCEPTS AND NOTATION (CONTINUED)

concept/notation	description	example(s)
fitted multiple linear regression model	<p>linear model fitted to the data to describe the relationship between Y and multiple X variables</p> $\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1 X_{i1} + \dots + \hat{\beta}_p X_{ip}$ <p>where:</p> <ul style="list-style-type: none"> - \hat{Y}_i is the predicted outcome for observation i - $\hat{\alpha}$ is the estimated intercept coefficient - each $\hat{\beta}_j$ is the estimated coefficient for variable X_j ($j=1, \dots, p$) - each X_{ij} is the observed value of the variable X_j for observation i ($j=1, \dots, p$) <p>interpretation of $\hat{\alpha}$: the \hat{Y} when all X variables equal 0</p> <p>interpretation of each $\hat{\beta}_j$: the $\Delta \hat{Y}$ associated with $\Delta X_j=1$, while holding all other X variables constant</p>	$\hat{Y}_i = 1 + 3X_{i1} - 5X_{i2}$ <p>in this fitted multiple linear regression model:</p> <p>$\hat{\alpha}=1$; when both X_1 and X_2 equal 0, we predict that Y will equal 1 unit, on average</p> <p>$\hat{\beta}_1=3$; when X_1 increases by 1 and X_2 remains constant, we predict an associated increase in Y of 3 units, on average</p> <p>$\hat{\beta}_2=-5$; when X_2 increases by 1 and X_1 remains constant, we predict an associated decrease in Y of 5 units, on average</p>
fitted multiple linear regression model where X_1 is the treatment variable	<p>if, in the fitted multiple linear regression model where X_1 is the treatment variable, we control for <i>all</i> potential confounders by including them in the model as additional X variables (that is, as control variables), then we can interpret $\hat{\beta}_1$ as a valid estimate of the average causal effect of X on Y</p>	<p>if X_1 is the treatment variable, X_2 is the only potential confounder, and the fitted model is $\hat{Y} = 1 + 3X_1 + 5X_2$:</p> <p>we interpret $\hat{\beta}_1$ as estimating that, while holding X_2 constant, receiving the treatment increases the outcome by 3 units, on average (in the same unit of measurement as the difference-in-means estimator)</p>
post-treatment variables	<p>variables affected by the treatment:</p> $X \rightarrow \text{post-treatment variable}$ <p>post-treatment variables should not be added as control variables; adding a post-treatment variable to the regression model would render our causal estimates invalid because we would be controlling for a consequence of the treatment when trying to estimate its total effect</p>	<p>if the treatment is attending a private school and private schools have smaller classes than public schools, then <i>small class</i> is a post-treatment variable because it is affected by the value of <i>private school</i></p> $\text{private school} \rightarrow \text{small class}$ <p>when estimating the causal effect of attending a <i>private school</i>, we should not control for class size</p>
internal validity	<p>refers to the extent to which the causal conclusions of a study are valid for the sample of observations in the study; it depends on whether the treatment and control groups used for the estimation can be considered comparable, after statistical controls are applied (if any are)</p>	<p>randomized experiments have strong internal validity because the randomization of the treatment assignment eliminates all potential confounders; observational studies may also have strong internal validity if the analysis controls for all potential confounders</p>
external validity	<p>refers to the extent to which the causal conclusions of a study can be generalized; it depends on (i) whether the sample of observations is representative of the population to which we want to generalize the results, and (ii) whether the treatment used in the study is representative of the treatment for which we want to generalize the results</p>	<p>observational studies typically have strong external validity because they often analyze the entire target population and observe the treatment in the environment in which we are interested; randomized experiments may also have strong external validity if they manage to use a representative sample of subjects and make the treatment comparable to the real-world one</p>

5.7.2 R FUNCTIONS

function	description	required argument(s)	example(s)
lm()	fits a linear model	<p>when there is only one X variable: $Y \sim X$;</p> <p>when there are multiple X variables: $Y \sim X_1 + \dots + X_p$</p> <p>optional argument <code>data</code>: specifies the object where the dataframe is stored; alternative to using <code>\$</code> for each variable</p>	<p>## both of these pieces of code fit the same simple linear regression model:</p> <pre>lm(data\$y_var ~ data\$x_var)</pre> <pre>lm(y_var ~ x_var, data=data)</pre> <p>## both of these pieces of code fit the same multiple linear regression model:</p> <pre>lm(data\$y_var ~ data\$x_var1 + data\$x_var2)</pre> <pre>lm(y_var ~ x_var1 + x_var2, data=data)</pre>