Multiple regression I

Johan A. Elkink

University College Dublin

18 April 2013

Outline

Causation

Slightly simplified, for T to be a cause of Y, we generally require:

1. T to precede Y

Causation

Slightly simplified, for T to be a cause of Y, we generally require:

1. T to precede Y
2. T to correlate with Y (either positively or negatively)

Causation

Slightly simplified, for T to be a cause of Y, we generally require:

1. T to precede Y
2. T to correlate with Y (either positively or negatively)
3. no other factor to explain the correlation between T and Y (no confounding factor)

Causation: terminology

If T causes Y,

- Y is called the dependent variable, or outcome variable, or response, or . . . ;
- T is called the independent variable, or explanatory variable, or factor, or treatment, or . . . .

Causation: terminology

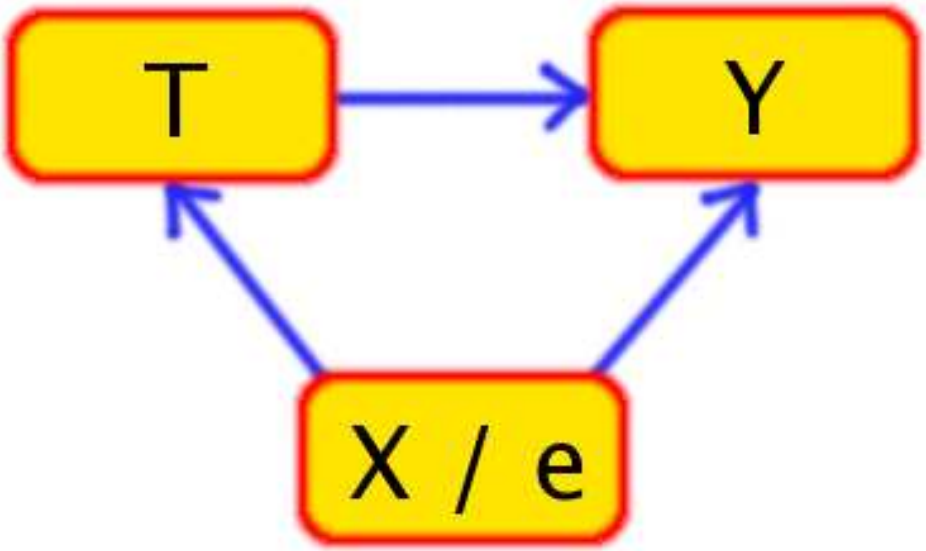
If T causes Y,

- Y is called the dependent variable, or outcome variable, or response, or . . . ;
- T is called the independent variable, or explanatory variable, or factor, or treatment, or . . . .

In political science, most common (unfortunately) is the usage of the terms independent and dependent variables.

Confounding

Confounding refers to having a third variable that explains the relationship between two variables.



Outline

When to control?

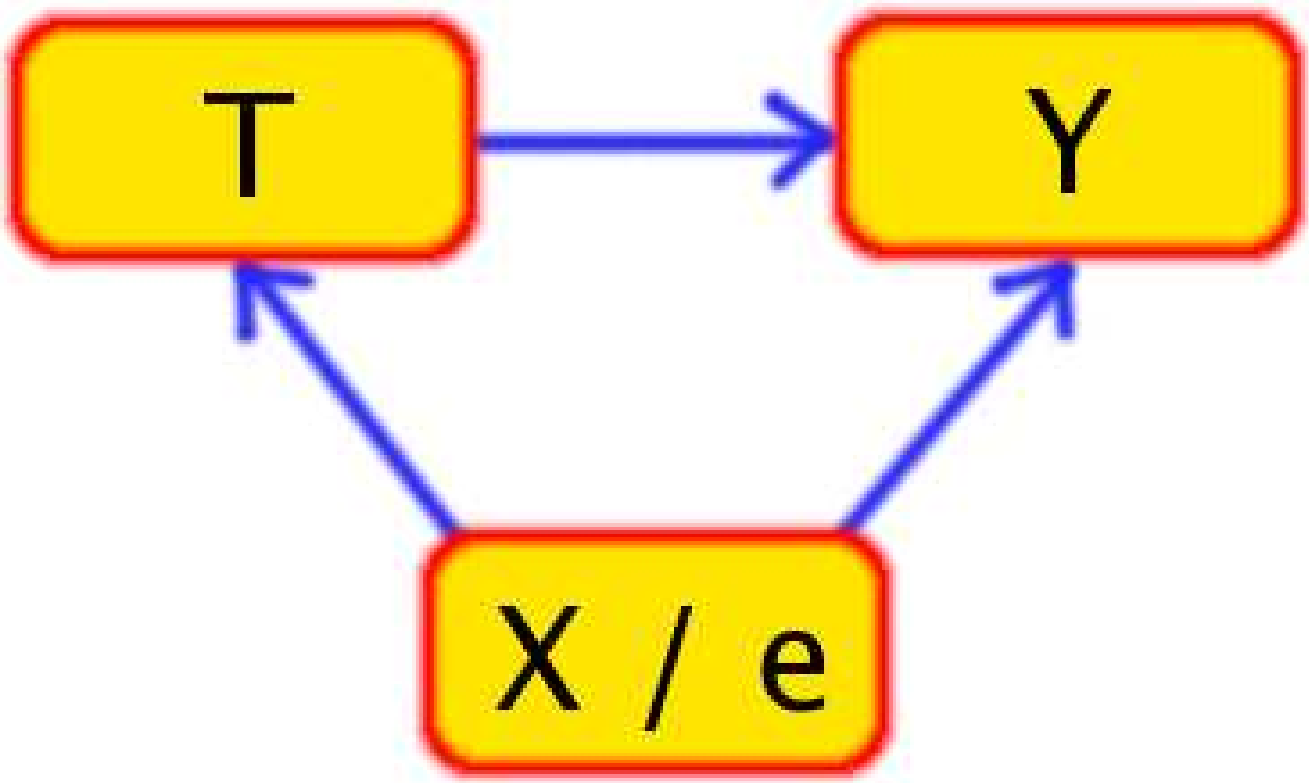
- X affects both T and $Y \Longrightarrow$ control

(Lee 2005: 43-48)

Do control

This is the typical case of a confounding factor, and hence should be eliminated through controlling.

Do control

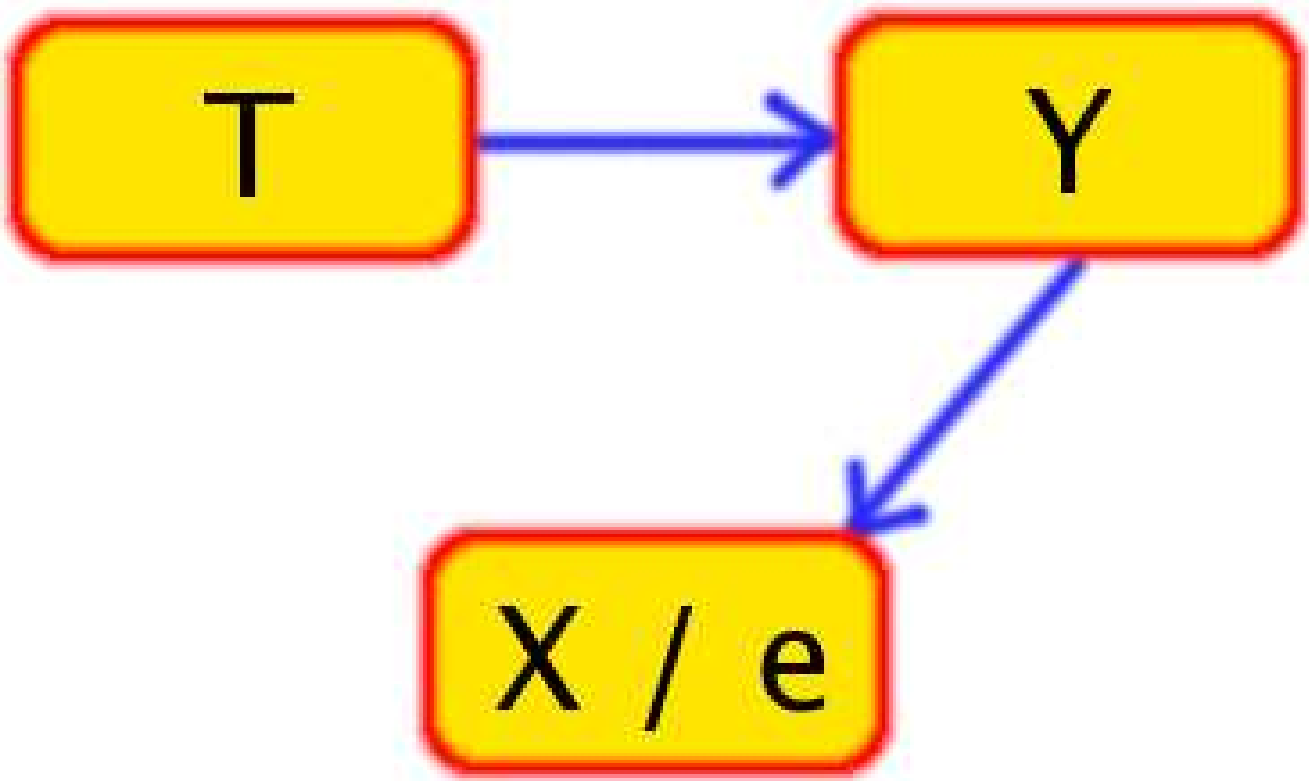


When to control?

- X affects both T and $Y \Longrightarrow$ control
- T affects Y, which in turn affects $X \Longrightarrow$ do not control

(Lee 2005: 43-48)

Don't control



Don't control

In this case, X is an effect of Y. By controlling for X, you can severily *underestimate* the effect of T on Y.

Don't control

In this case, X is an effect of Y. By controlling for X, you can severily *underestimate* the effect of T on Y.

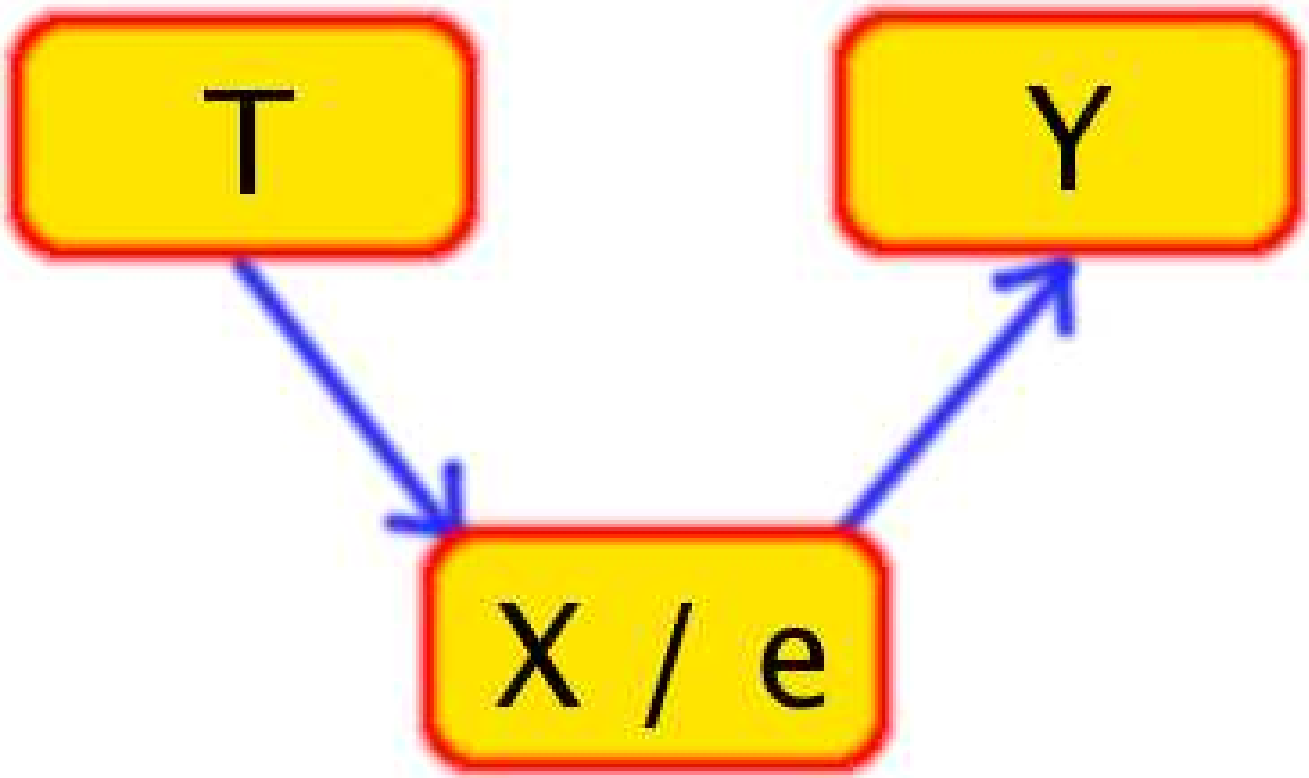
Imagine that a college degree leads to a better income leads to a nicer car. Controlling for the price of the car in estimating the effect of having a college degree on income might cancel the effect.

When to control?

- X affects both T and $Y \Longrightarrow$ control
- T affects Y, which in turn affects $X \Longrightarrow$ do not control
- T affects X, which in turn affects $Y \Longrightarrow$ do not control ...

(Lee 2005: 43-48)

Don't control



Don't control

To get the overall effect of T on Y, you want to include the effect through X.

Don't control

To get the overall effect of T on Y, you want to include the effect through X.

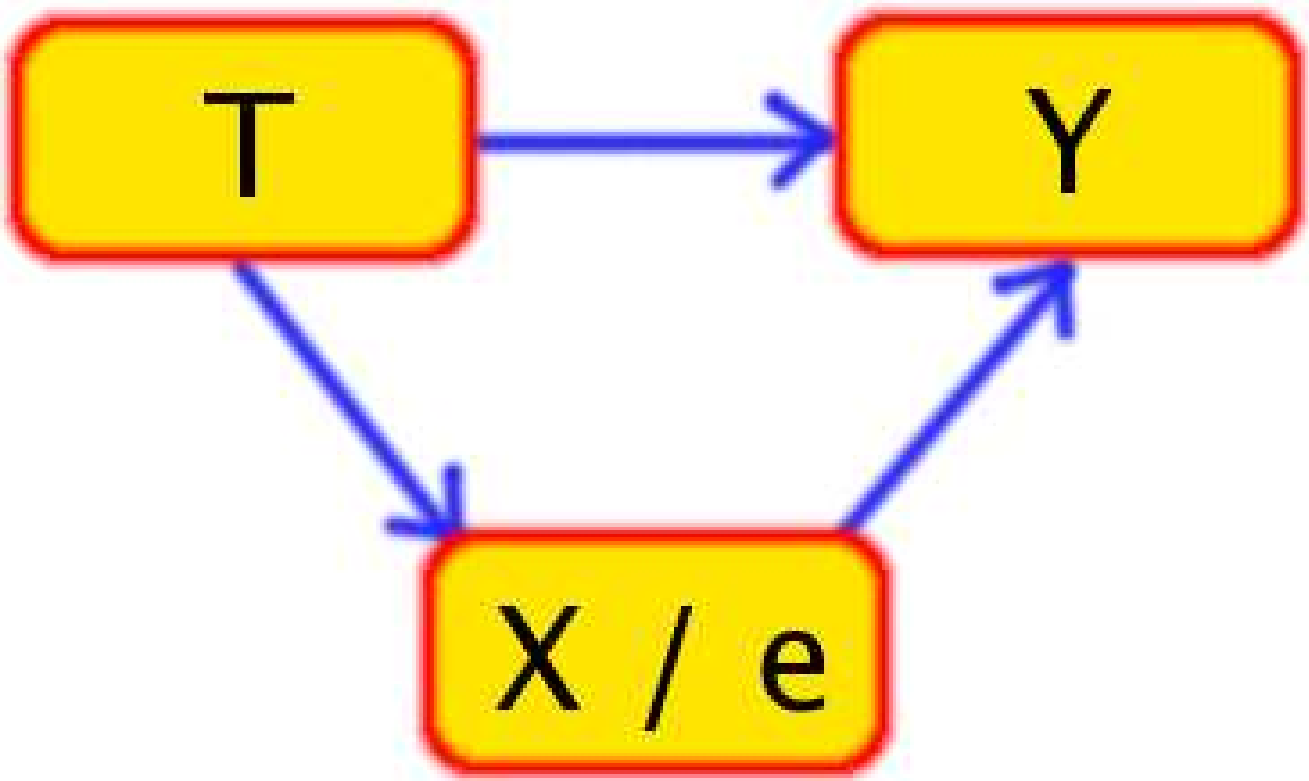
E.g. if you want to know the effect of changing the policy regarding smoking in pubs on the amount of smoking in general, you do not care through what mechanism this happened (through peer pressure, laziness, etc.), but only about the overall effect.

When to control?

- X affects both T and $Y \Longrightarrow$ control
- T affects Y, which in turn affects $X \Longrightarrow$ do not control
- T affects X, which in turn affects $Y \Longrightarrow$ do not control ...
- ... unless you explicitly want only the direct effect

(Lee 2005: 43-48)

Maybe control



Maybe control

Example: A scholarship for poorer students might help them to get a college degree, which in turn might help them to earn more money later in life. Having a scholarship on your CV, however, might also further your career, independent of the effect of having a college degree.

Maybe control

Example: A scholarship for poorer students might help them to get a college degree, which in turn might help them to earn more money later in life. Having a scholarship on your CV, however, might also further your career, independent of the effect of having a college degree.

To see the overall effect of the scholarship, don't control on having a college degree.

Maybe control

Example: A scholarship for poorer students might help them to get a college degree, which in turn might help them to earn more money later in life. Having a scholarship on your CV, however, might also further your career, independent of the effect of having a college degree.

To see the overall effect of the scholarship, don't control on having a college degree.

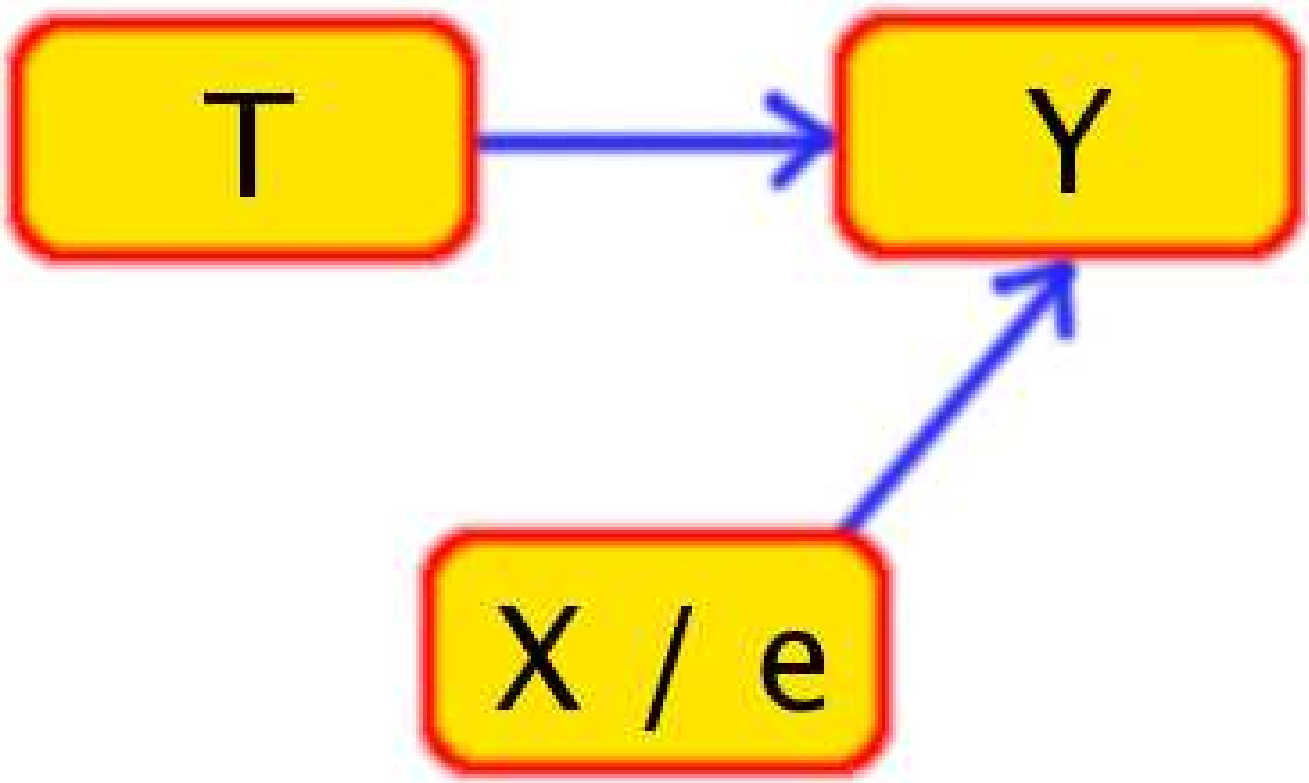
To see the effect of having a scholarship, independent of the effect of getting a college degree, do control for college degree.

When to control?

- X affects both T and $Y \Longrightarrow$ control
- T affects Y, which in turn affects $X \Longrightarrow$ do not control
- T affects X, which in turn affects $Y \Longrightarrow$ do not control ...
- ... unless you explicitly want only the direct effect
- X affects Y, but not T, nor the effect of T on Y

(Lee 2005: 43-48)

Maybe control



Maybe control

When X affects Y, but not T, there is no confounding issue and the estimates for the effect of T on Y should not be affected by inclusion of X. However, including X in the model can still help for efficiency.

(Gelman & Hill 2007: 177)

When to control?

- X affects both T and $Y \Longrightarrow$ control
- T affects Y, which in turn affects $X \Longrightarrow$ do not control
- T affects X, which in turn affects $Y \Longrightarrow$ do not control ...
- ... unless you explicitly want only the direct effect
- X affects Y, but not T, nor the effect of T on Y
- X affects Y, not T, but it does affect of effect of T on Y (*interaction*)

(Lee 2005: 43-48)

Maybe control

Here including the interaction in your model can highlight how the effect is different for different groups.

(next lecture)

Outline

Adjusted R^2

One of the problems with looking at R^2 is that the more independent variables, the higher R^2, which discourages parsimony. One solution for this the adjusted R^2:

$$adjR^2 = 1 - \frac{n-1}{n-k}(1 - R^2)$$

So this R^2 has a penalty for having many parameters (high k).

Outline

Exercise

We will study the relationship between having a degree and future earnings (`education.dta`).

1. Regress `earnings` on `degree`.
2. Repeat, but control for `ability`.
3. Repeat, but control also for `schooling`.

What do you conclude?

Exercise

Let's take the silly example of movie description lengths again (`films.dta`).

1. Regress `desclength` on `year` and `length`. What do you conclude about the relation between the duration of a movie and the number of lines used in the review?
2. Repeat, controlling for `castsize`. Does this revise your conclusion?

Table presentation

% Year	0.05	*
	(0.015)	
Duration	0.02	
	(0.016)	
Cast size	0.52	*
	(0.125)	
intercept	-91.04	*
	(29.49)	
Observations	100	
Adjusted R^2	0.31	
F	15.87	*

Regression coefficients explaining the number of lines devoted to a movie review in Leonard Maltin's Movie and Video Guide, 1996. Standard errors in parentheses.