Lecture 10: The Multiple Linear Regression Model

Thomas Chadefaux

Quantitative Methods I

Today

1. Causal Inference
2. Multiple Regression

Causal Inference

Statistical inference and causal inference

- Statistical inference: How likely are we to observe this relationship given the null hypothesis that there is no relationship (Bayesian version: how much do we update our beliefs given the new evidence)
- Causal inference: is the IV a *cause* of the DV?

Causation

- A causal claim is a statement about what did not happen: counterfactual
- "X causes Y" means that Y would not have happened in the absence of X
- There need not be a single cause for an outcome Y.
  - E.g., NRA: "guns don't kill people, people kill people". That makes little sense in a counterfactual framework.
- Causation does not mean that there is a "causal path" between X and Y.
  - e.g., A wants to do Y. B wants to stop her, but C stops B from stopping A. In the end, A completes Y, without even knowing B and C existed. But C still caused Y

The fundamental problem of causal inference

- Causal effects are statements about the difference between what happened and what would have happened
- But of course, we can never observe what would have happened.
    - You can give a treatment or not give a treatment, but not both

Causation

What we *can* do, is look at the average treatment effect

$$E[\tau_i] = E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)]$$

where τ_i is the treatment effect, E is the expectations operators, and $Y(1)$ is the outcome for those who receive the treatment.

Causation

We require, loosely, 3 conditions:

- X precedes Y
- X and Y are correlated
- No other confounding factor can explain that correlation

Confounding variables

- A confounder is a variable that influences both T (the treatment) and Y.

Examples

- Red shoes $\rightarrow$ Height

Examples

- We are interested in the effect of education on income. We estimate

$$\text{income}_i = \beta_0 + \beta_1 \text{education}_i + \varepsilon_i$$

and probably find β_1 to be significantly greater than 0. Can we conclude that education increases revenue?

Examples

- We are interested in the effect of education on income. We estimate

$$\text{income}_i = \beta_0 + \beta_1 \text{education}_i + \varepsilon_i$$

  and probably find β_1 to be significantly greater than 0. Can we conclude that education increases revenue?
- We are missing *ability*

Examples

- You are in the forest and observe ducks flying. Suddenly you hear a loud bang, and the duck falls to the ground. Another duck, another bang, again this one falls to the ground. You observe this over and over, and most of the time a loud bang precedes the death of the duck (sometimes, but rarely, the duck escapes). In fact, you find that 90% of the time, a bang is followed by the death of the duck.

Examples

- You are in the forest and observe ducks flying. Suddenly you hear a loud bang, and the duck falls to the ground. Another duck, another bang, again this one falls to the ground. You observe this over and over, and most of the time a loud bang precedes the death of the duck (sometimes, but rarely, the duck escapes). In fact, you find that 90% of the time, a bang is followed by the death of the duck.
- You therefore write a letter to the Prime Minister, requesting that all ducks be equipped with noise-cancelling earphones. The problem here, of course, is again one of omitted variable. The sound is not causing the death, but rather the gun is.

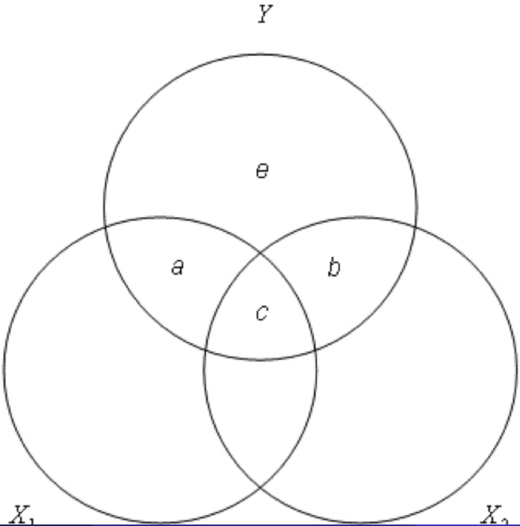
Why multiple regression?

- The multiple linear regression model is used to study the relationship between a dependent variable y and one or more independent variables $x_2, x_3, \ldots, x_K$.
- Why not just use the correlation coefficient? The correlation coefficient gives a measure of whether two variables are associated with one another, but not of the kind of relationship involved. It is also limited to two variables.

Why multiple regression?

- The effect of each variable could be studied in isolation.
- But the results may be misleading if the explanatory variables are mutually related.
- For example, voting might be affected by age and by wealth. But if we regress `vote'  on`age' and then vote on wealth, we are likely to overestimate the importance of both age and wealth, because both are correlated. In other words, some of the effect of age on vote is accounted for by wealth (e.g., older people might simply be wealthier). Attributing the entire effect to age will lead to a biased coefficient.

Intuition for multiple regression (VERY IMPORTANT!)



When to Control?

When to control?

Control when: X affects both T and Y
Do not control when:

- T affects Y, which in turn affects X
  - e.g., college degree -> higher income -> expensive car. Do not control for car
- T affects X, which in turn affects Y *AND* you don't care about the effect of X.
  - e.g., effect of war onset (T) on bond yields (Y) and central bank rates (X)
  - E.g., scholarship -> college -> higher earnings, but also: scholarship -> prestige -> higher earnings
    - To see the overall effect of the scholarship (T), do not control
    - To see the effect of the scholarship, independently of having a college degree, control
- X affects Y, but not T

Kitchen sink

Avoid adding every variable you can think of - Loses precision (i.e., increases S.E. of the estimates) - Misleading results

How to control?

- Experiment
  - Lab
- Field Experiment
  - Canvassing
- Natural experiment with true randomization
  - Military draft (Angrist 1990)
  - Windfall on long-term welfare
- Natural experiments with as-if randomization
  - Bombing by drunk soldiers in Chechnya (Lyall 2009)
  - Regression discontinuity
- Matching
- Multiple regression

Multiple Regression

Regression model in matrix form

Our model is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i.$$

In matrix form

This model can be expressed in matrix form as:

$$y = X\beta + \varepsilon$$

where

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{12} & \dots & x_{1k} \\ 1 & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \dots & x_{nk} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

and the estimated model as

$$y = Xb + e$$

In matrix form

Alternatively, we can also write it as:

$$y_i = \mathbf{x}_i'\mathbf{b} + e_i,$$

where $\mathbf{x}_i' = \begin{pmatrix} 1 & x_{1i} & x_{2i} & \dots & x_{ki} \end{pmatrix}$.
The **disturbance** associated with the i^{th} data point is:

$$\varepsilon_i = y_i - \mathbf{x}_i'\boldsymbol{\beta},$$

and the **residual** is

$$e_i = y_i - \mathbf{x}_i'\mathbf{b},$$

Another way to put it:

$$y_i = \mathbf{x}_i'\beta + \varepsilon_i = \mathbf{x}_i'\mathbf{b} + e_i = \hat{y}_i + e_i,$$

K > 2, matrix notation

More generally, in matrix form, we have:

$$e = y - Xb,$$

and so

$$\begin{aligned} S(b) =& \sum_i e_i^2 \\ =& e'e \\ =& (y - Xb)'(y - Xb) \\ =& y'y - y'Xb - b'X'y + b'X'Xb \\ =& y'y - 2b'X'y + b'X'Xb \end{aligned} \qquad (1)$$

(Why the last step? Since $y'Xb$ is a scalar, it is equal to its transpose and so we can write $y'Xb = (y'Xb)' = b'X'y$)

K $>$ 2, matrix notation

Now we want to minimise (1) with respect to b. Remember that b is a vector of the form

$$b = \begin{pmatrix} b_1 & b_2 & \dots & b_K \end{pmatrix}'.$$

So when we calculate $\frac{\partial S(b)}{\partial b}$, we'll get K first order conditions. So let's calculate $\frac{\partial S(b)}{\partial b}$:

$$\begin{aligned} \frac{\partial S(b)}{\partial b} =& 0 - 2X'y + 2X'Xb = 0 \\ & \rightarrow X'Xb = X'y \\ & \boxed{\rightarrow b = (X'X)^{-1}X'y} \end{aligned} \tag{2}$$

Back to the example in lecture 9

We have the following sample data:

L/R scale (y)	constant	Income (x)
2	1	10
3	1	42
7	1	39
1	1	30
9	1	80

Calculate b

Let us calculate b using the matrix notation formula:

$$
\begin{aligned}
b &= (X'X)^{-1}X'y \\
&= \left(\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 10 & 42 & 39 & 30 & 80 \end{pmatrix} \begin{pmatrix} 1 & 10 \\ 1 & 42 \\ 1 & 39 \\ 1 & 30 \\ 1 & 80 \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 10 & 42 & 39 & 30 & 80 \end{pmatrix} \begin{pmatrix} 2 \\ 3 \\ 7 \\ 1 \\ 9 \end{pmatrix} \\
&= \begin{pmatrix} 5 & 201 \\ 201 & 10685 \end{pmatrix}^{-1} \begin{pmatrix} 22 \\ 1169 \end{pmatrix} \\
&= \begin{pmatrix} 0.82040848 & -0.0154330467 \\ -0.01543305 & 0.0003839066 \end{pmatrix} \begin{pmatrix} 22 \\ 1169 \end{pmatrix} \\
&= \begin{pmatrix} 0.007755 \\ 0.109260 \end{pmatrix} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}
\end{aligned}
$$

Check our results using R (matrix version):

```
#--- Input data
x0 = c(1,1,1,1,1)
x1 = c(10, 42, 39, 30,80)
X <- cbind(x0,x1)
y = c(2,3,7,1,9)
```

Check our results using R (matrix version):

```
# Calculate (X'X)^{-1}X'y
XprimeX <- t(X)%*%X
XprimeXinverse <- solve(XprimeX)
Xprimey <- t(X)%*%y
XprimeXinverse %*% Xprimey
```

```
##           [,1]
## x0 0.007754914
## x1 0.109259828
```

```
# or the equivalent, in one line:
solve(t(X)%*%X) %*% t(X)%*%y
```

```
##           [,1]
## x0 0.007754914
## x1 0.109259828
```

Geometric Interpretation

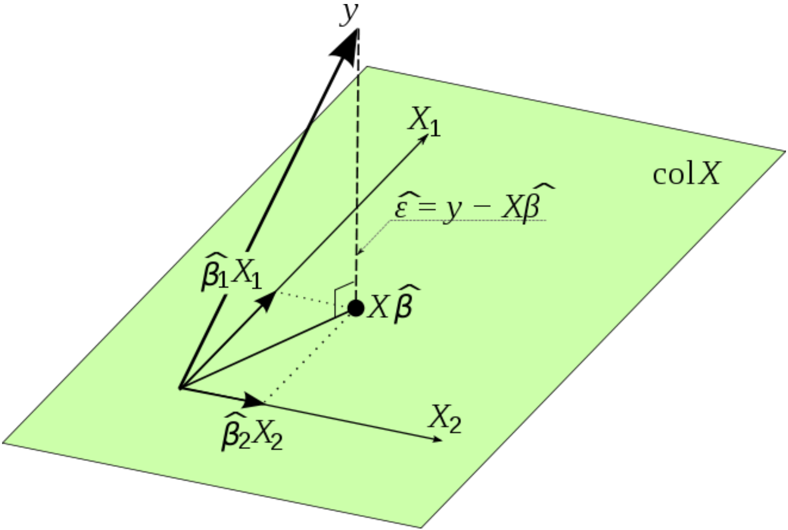


Figure 9

Graphical intuition: what does it mean to “control for”, or to “holding other variables constant”?

See plot3d.R

table presentation

```
n <- 100
x1 <- c(1:n)/100
x2 <- rnorm(n)
y <- x1*2 + x2*1/2 + rnorm(n)
library(scatterplot3d)
scatterplot3d(x1,x2,y)
```

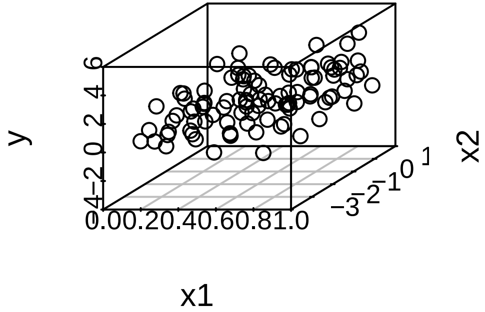


table presentation

```r
lm1 <- lm(y~x1+x2)
library(texreg)
texreg(lm1)
```

	Model 1
(Intercept)	0.06
	(0.20)
x1	2.17***
	(0.35)
x2	0.59***
	(0.11)
R^2	0.41
Adj. R^2	0.40
Num. obs.	100
RMSE	1.01

table presentation

Table 4: Regression Results

	Dependent variable:		
	Overall Rating		High Rating
	OLS		*probit*
	(1)	(2)	(3)
Handling of Complaints	0.692***	0.682***	
	(0.149)	(0.129)	
No Special Privileges	−0.104	−0.103	
	(0.135)	(0.129)	
Opportunity to Learn	0.249	0.238*	0.164***
	(0.160)	(0.139)	(0.053)
Performance-Based Raises	−0.033		
	(0.202)		
Too Critical	0.015		−0.001
	(0.147)		(0.044)
Advancement			−0.062
			(0.042)
Constant	11.011	11.258	−7.476**
	(11.704)	(7.318)	(3.570)
Observations	30	30	30
R^2	0.715	0.715	
Adjusted R^2	0.656	0.682	
Akaike Inf. Crit.			26.175
Note:			*p<0.1; **p<0.05; ***p<0.01

Figure 3