Multiple regression

Johan A. Elkink
School of Politics & International Relations
University College Dublin

9 November 2015

Outline

Causation

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

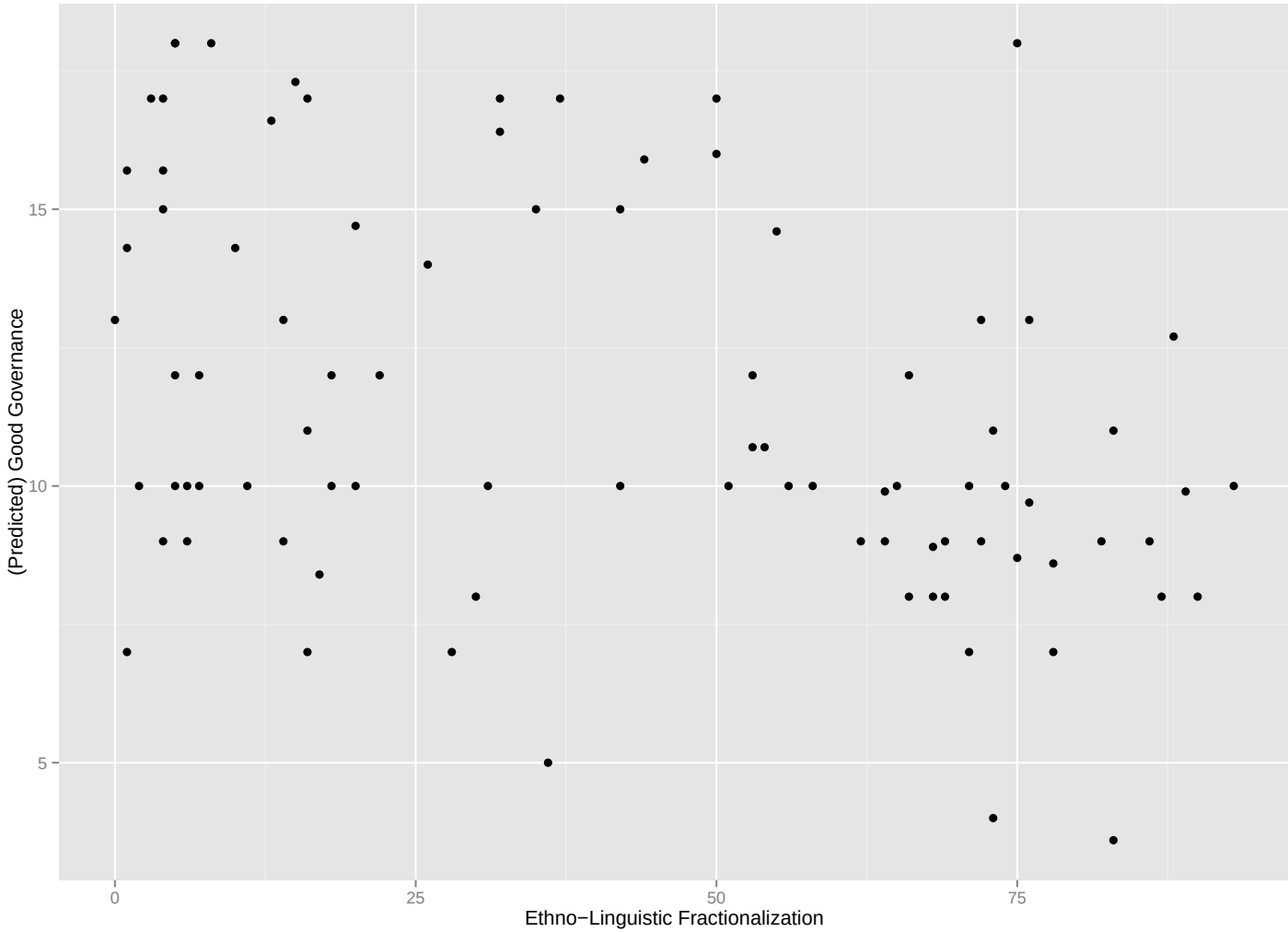
1. X to precede Y
2. X to correlate with Y (either positively or negatively)
3. no other factor to explain the correlation between X and Y (no **confounding factor**)

If X causes Y,

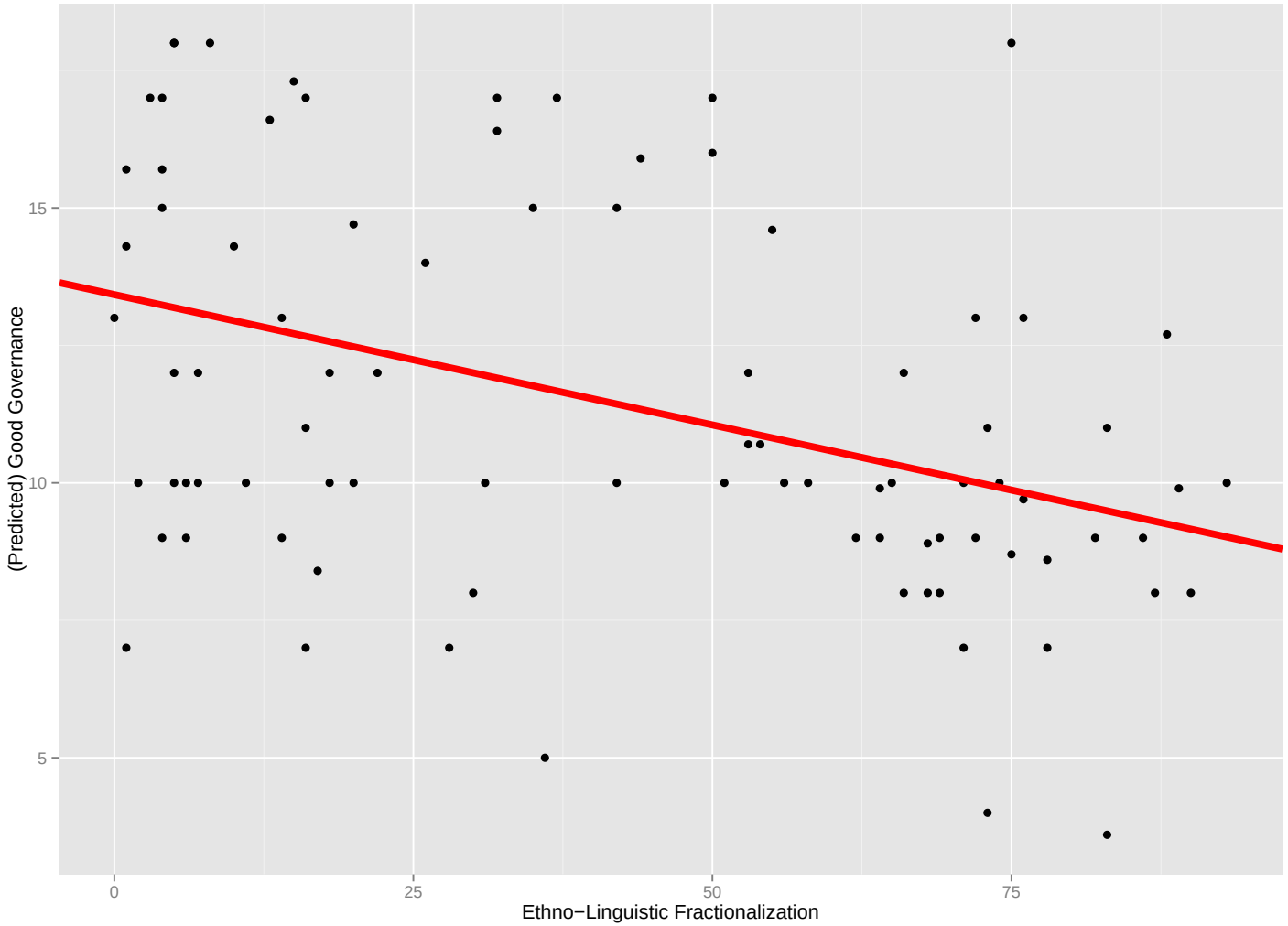
- Y is called the **dependent variable**, or **outcome variable**, or **response**, or . . . ;
- X is called the **independent variable**, or **explanatory variable**, or **factor**, or . . . .

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

Linear regression



Linear regression



Notation

y_i	Value on the dependent variable for case i
x_i	Value on the independent variable for case i
$\bar{x}$	Mean value on the independent variable for case i
ε_i	The error for case i: $\varepsilon_i = y_i - \hat{y}_i$
β_k	True coefficient for variable k
$\hat{\beta}_k$	Estimated coefficient for variable k
$\hat{y}_i$	Predicted value on the dependent variable for case i

Ordinary Least Squares

"Quickly put, the regression line is chosen to minimize the RSS; it has slope $\hat{\beta}_1$, intercept $\hat{\beta}_0$, and goes through the point $(\bar{x}, \bar{y})$. Furthermore, the estimate for σ^2 is $\hat{\sigma}^2 = RSS/(n-2)$" (Verzani, 2005, 280).

BLUE

An unbiased estimator of a coefficient β is an estimator where the **mean of the sampling distribution** is identical to the true β. I.e. $E(\hat{\beta}) = \beta$.

The bias of an estimator is thus $E(\hat{\beta}) - \beta$.

Often, many estimators could be defined whereby $E(\hat{\beta}) - \beta = 0$, i.e. that are unbiased. The **best unbiased** estimator is the estimator that leads to an unbiased estimated with the **smallest variance**.

Another term for this is **efficiency** – a smaller standard error means a more efficient estimator.

If all assumptions underlying OLS hold, OLS is BLUE, i.e. the **best linear unbiased estimator** (BLUE) of β.

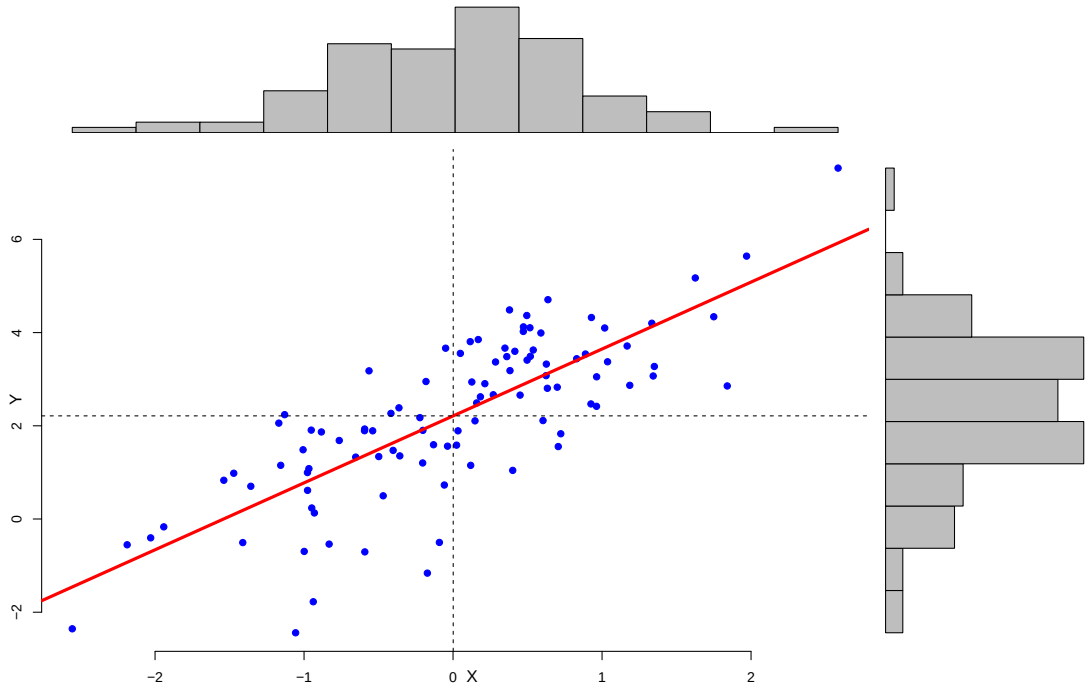
Breakdown of variance

Total Sum of Squares (TSS): $\sum_{i=1}^{N}(y_i - \bar{y})^2$

Explained Sum of Squares (ESS): $\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2$

Residual Sum of Squares (RSS): $\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{N}\varepsilon_i^2$

$$TSS = ESS + RSS$$



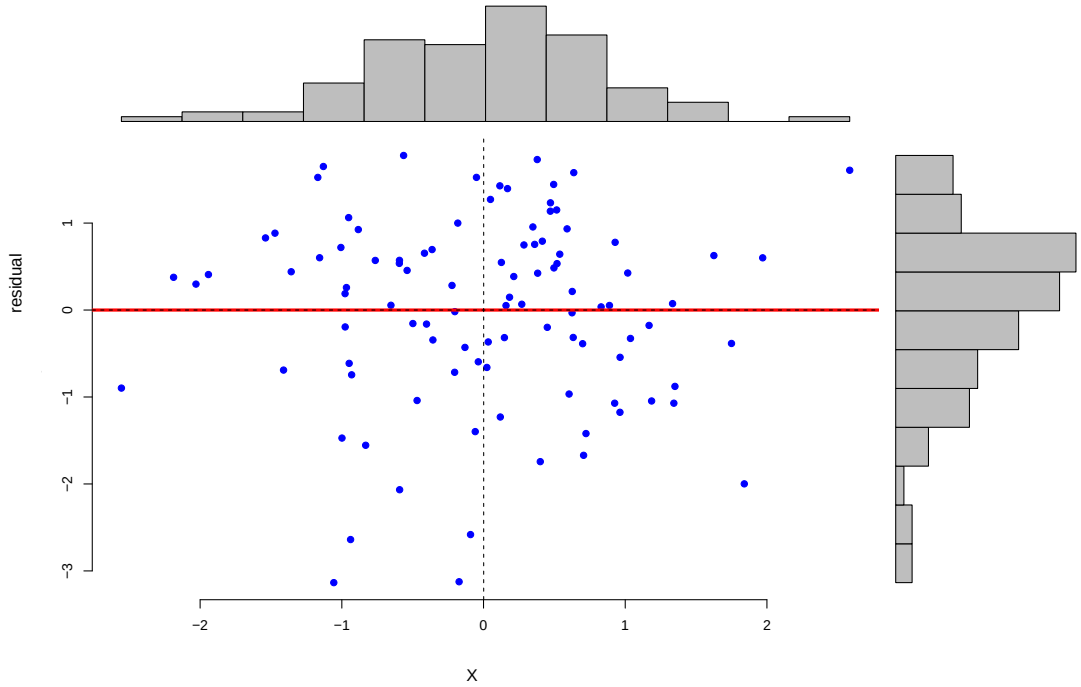
Breakdown of variance

Total Sum of Squares (TSS): $\sum_{i=1}^{N}(y_i - \bar{y})^2$

Explained Sum of Squares (ESS): $\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2$

Residual Sum of Squares (RSS): $\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{N}\varepsilon_i^2$

$$TSS = ESS + RSS$$



Breakdown of variance

Total Sum of Squares (TSS): $\sum_{i=1}^{N}(y_i - \bar{y})^2$

Explained Sum of Squares (ESS): $\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2$

Residual Sum of Squares (RSS): $\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{N}\varepsilon_i^2$

$$TSS = ESS + RSS$$

Sometimes the second is called "regression sum of squares" (RSS) and the third "errors sum of squares" (ESS), which might in fact be more accurate, since ε really represents errors, not residuals, in this specification. Beware the confusion!

R^2

How much of the variance did we explain?

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

Can be interpreted as the *proportion of total variance explained by the model*. For this interpretation, the model must include an intercept.

For simple linear regression (i.e. one independent variable), R^2 is the same as the correlation coefficient, Pearson's r, squared.

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).

Table presentation

	Dependent variable:
	goodgovt
elf	-0.047^{***}
	(0.012)
Constant	13.423^{***}
	(0.597)
Observations	87
R^2	0.162
Adjusted R^2	0.153
Residual Std. Error	3.247 (df = 85)
F Statistic	16.484^{***} (df = 1; 85)
Note:	*p<0.1; **p<0.05; ***p<0.01

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.

* Significant at $\alpha = .05$.

Outline

Inference from regression

In linear regression, the **sampling distribution** of the coefficient estimates form a normal distribution, which is approximated by a ***t*-distribution** due to approximating σ by s.

Thus we can calculate a **confidence interval** for each estimated coefficient.

Or perform a hypothesis test along the lines of:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Inference from regression

To calculate the confidence interval, we need to calculate the **standard error** of the coefficient.

Rule of thumb to get the 95% confidence interval:

$$\beta - 2SE < \beta < \beta + 2SE$$

Thus if β is positive, we are 95% certain it is different from zero when $\beta - 2SE > 0$. (Or when the ***t*-value** is greater than 2 or less than -2.)

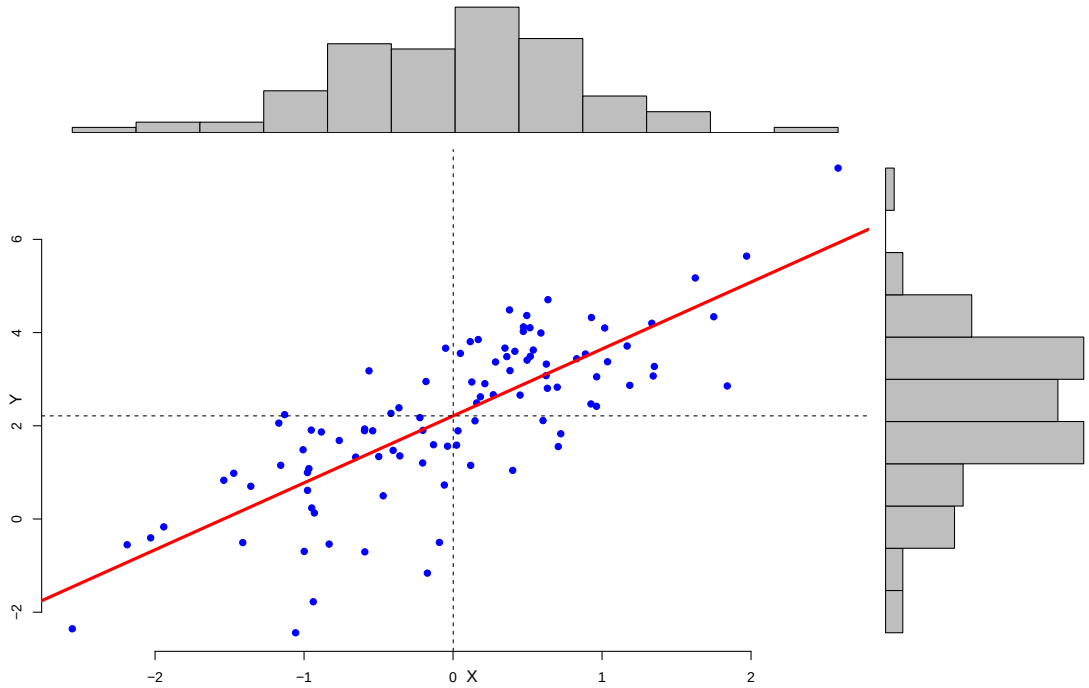
Breakdown of variance

Total Sum of Squares (TSS): $\sum_{i=1}^{N}(y_i - \bar{y})^2$

Explained Sum of Squares (ESS): $\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2$

Residual Sum of Squares (RSS): $\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{N}\varepsilon_i^2$

$$TSS = ESS + RSS$$



F-test

In simple linear regression, we can do an F-test:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

$$F = \frac{ESS/1}{RSS/(n-2)} = \frac{ESS}{\hat{\sigma}^2} \sim F_{1,n-2}$$

with 1 and $n-2$ degrees of freedom.

For multiple regression, this would generalize to:

$$F = \frac{ESS/(k-1)}{RSS/(n-k)} \sim F_{k-1,n-k}$$

Outline

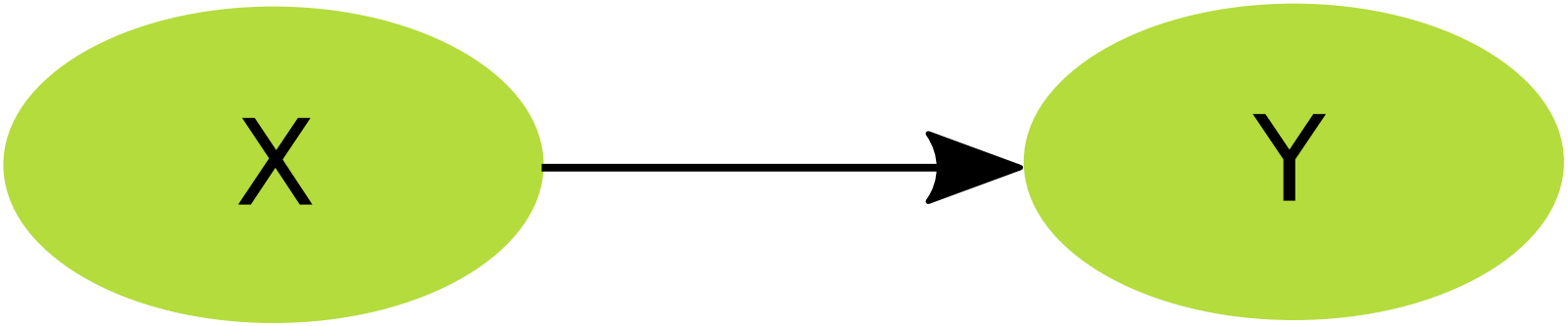
Causation

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**)

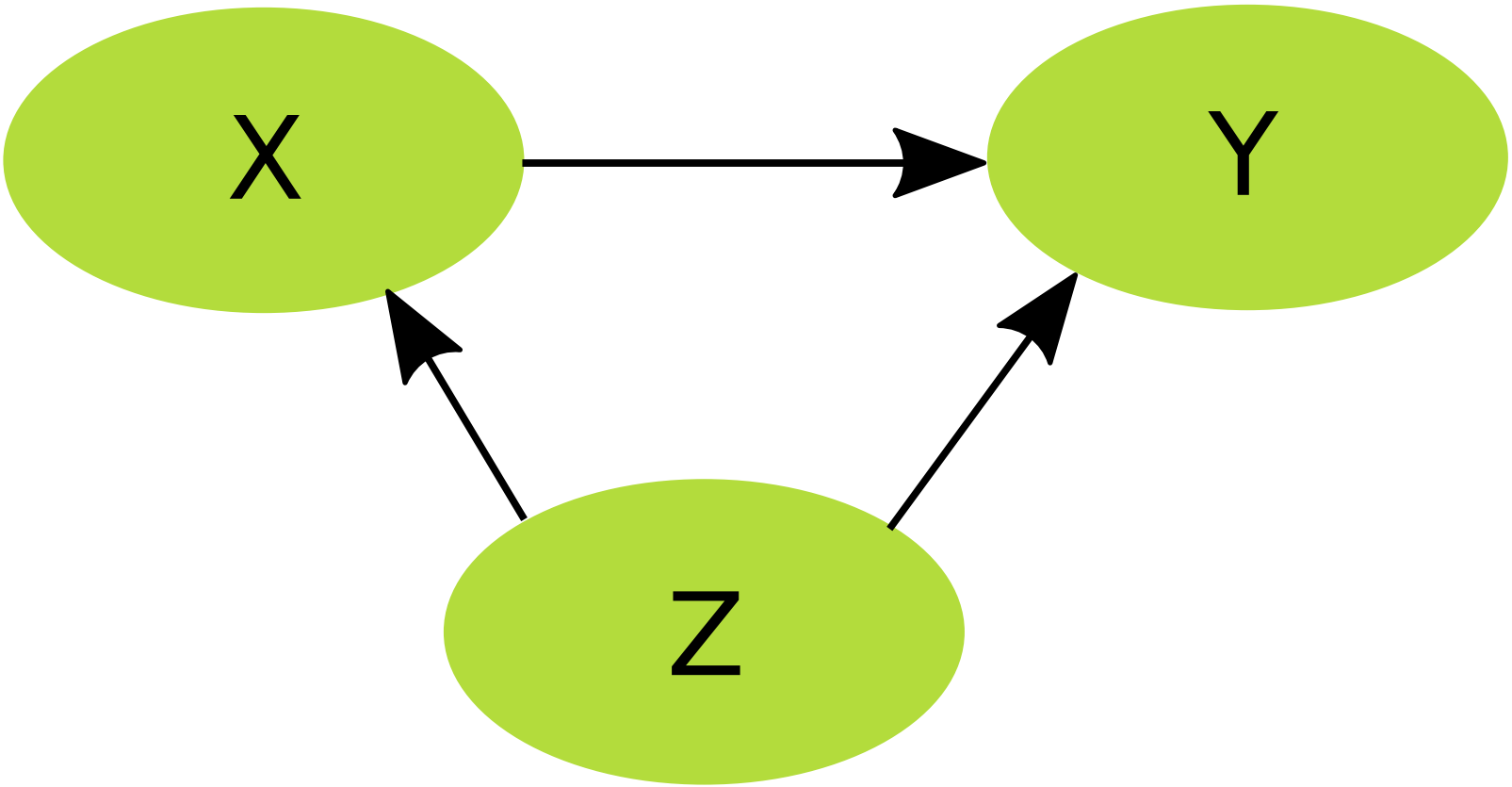
Confounding

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



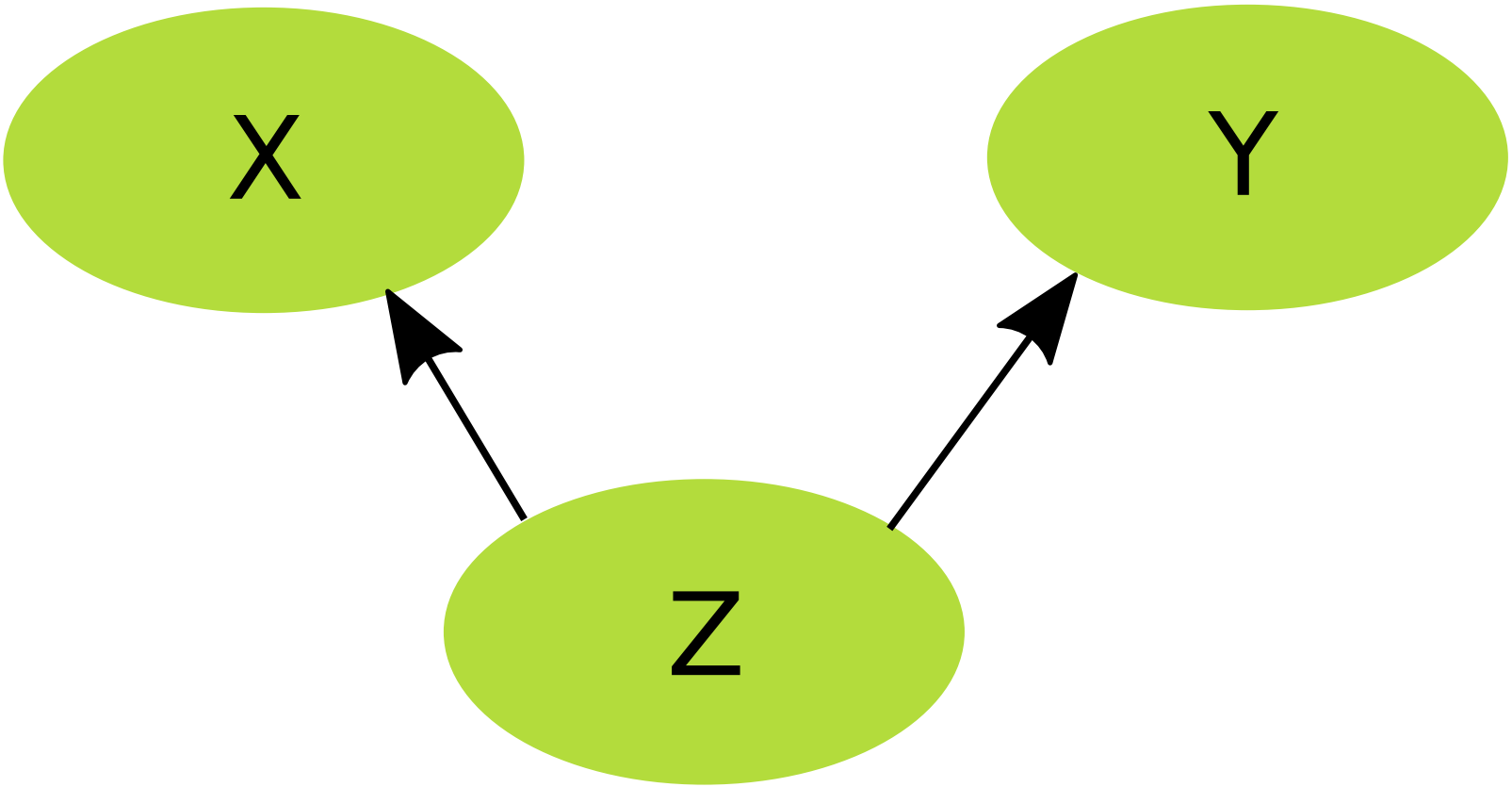
Confounding

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



Confounding

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



When to control?

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

(Lee, 2005, 43–48)

Do control

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

When to control?

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

(Lee, 2005, 43–48)

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 $Y \implies$ control
- T affects Y, which in turn affects $X \implies$ do not control
- T affects X, which in turn affects $Y \implies$ do not control ...

(Lee, 2005, 43–48)

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

(Lee, 2005, 43–48)

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 $Y \implies$ control
- T affects Y, which in turn affects $X \implies$ do not control
- T affects X, which in turn affects $Y \implies$ 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

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 and Hill, 2007, 177)

When to control?

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

Gelman, Andrew and Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. Analytical Methods for Social Research Cambridge: Cambridge University Press.

Lee, Myoung Jae. 2005. *Micro-econometrics for policy, program, and treatment effects*. Oxford: Oxford University Press.

Verzani, John. 2005. *Using R for introductory statistics*. Boca Raton, FL: Chapman & Hall/CRC.