GLM ANCOVA

Caspar J. van Lissa

2023-08-11

Today: ANCOVA

> **ANCOVA:** ANalysis of COVAriance; regression with a categorical predictor and continuous control variable(s).

- Nothing complicated!
- Building upon:
  - Regression with dummies
  - Multiple regression
  - Causality (multiple regression lecture)

ANCOVA is just regression with one categorical predictor of interest, and one or more continuous *covariates*

What are Covariates?

Covariates: variables that are related to the dependent variable, but not of primary interest to the researcher. Examples: *age* , *gender* , *education level* , …

- Unavoidable to discuss causal assumptions at this point
- Remember: Control for *confounders*
  - Variables that cause the predictor of interest and/or outcome
- Never control for *colliders*
  - Variables that are influenced by both the predictor of interest and the outcome
  - Your results will be meaningless!
- Don’t blindly control for “everything but the kitchen sink”
- Read Pearl (Book of Why) for a sensible approach to covariates

Why Control at all?

Covariates can reduce the residual variance

- This increases power to find an effect for your predictor of interest

Covariates are essential for causal inference

- E.g., in quasi-experiments
- Proper selection of covariates that make causal inference possible is beyond the scope of this course!

Cinelli, C., Forney, A., & Pearl, J. (2022). A crash course in good and bad controls. Sociological Methods & Research, 00491241221099552.
https://ftp.cs.ucla.edu/pub/stat_ser/r493.pdf

Good Controls

Assume that A is your factor and represents an intervention through a *natural experiment*

- Natural experiment: No random assignment, people volunteer or choose to receive each treatment
- A is the predictor (a factor)
- B is the outcome (continuous)
- C is a third variable

Good Control: Confounder

For example:

- A is taking a homeopathic supplement (1) or not (0)
- B is perceived health improvement
- C is belief in the efficacy of the supplement
- C would cause a spurious effect of A on B, so we should control!

Neutral Control: Covariate

- C is unrelated to A
- Controlling for C reduces error variance of outcome B
- This increases power to detect an effect of A on B

Randomized Controlled Experiment

- The experimental method breaks connections between confounders C and treatment A:

Neutral Control: Cause of X

- In a natural experiment:
  - This reduces variance in A
  - Which may reduce the precision of the effect of A on B
- In a randomized controlled experiment:
  - Differences between groups should be due to random chance
  - Effects on A are prevented by random assignment
  - Controlling for “causes of” A thus introduces bias

Bad Control: Collider

- E.g.: A is smoking
- B infant mortality
- C is low birthweight
- Controlling for low birthweight creates a (spurious!) negative effect of smoking on infant mortality!
- Tobacco industry used this as evidence that smoking is good for babies…

Bad Control: Case Control Bias

- Controlling for an outcome of the outcome B

Bad Control: Overcontrol Bias

- Controlling for a mediator (process variable) of the effect of A on B

Take Home Message

Think about causal assumptions before you control!
Draw the presumed causal diagram on a piece of paper

Example: Red Bull study 1

Natural experiment at the train station

- Participants choose a drink
  - Factor, 2 level: Red bull VS herbal tea
- Participants perform a memory task:
  - DV: # words remembered
- Covariate: participants' age

Example: Red Bull study

ANOVA:

	B	SE	t	p
(Intercept)	22.75	0.80	28.39	0
drinkRed Bull	-10.73	1.16	-9.22	0

Does drinking Red Bull harm memory? Why? Does it make people distractable?

ANCOVA:

	B	SE	t	p
(Intercept)	30.26	2.25	13.43	0.0
drinkRed Bull	-3.09	2.42	-1.28	0.2
age	-0.26	0.07	-3.53	0.0

No: Age is a confounder; old people's memory is slightly worse, and they tend to prefer tea over red bull.

Example: Red Bull study 2

Randomized controlled experiment at the train station

- Participants are assigned a drink
  - Factor, 2 level: Red bull VS herbal tea
- Participants perform a memory task:
  - DV: # words remembered
- Covariate: participants' age

Example: Red Bull study 2

ANOVA:

	B	SE	t	p
(Intercept)	14.65	0.91	16.09	0.00
drinkRed Bull	0.49	1.38	0.36	0.72

Correct answer: No effect of Red Bull

ANCOVA:

	B	SE	t	p
(Intercept)	25.74	1.82	14.16	0.0
drinkRed Bull	0.75	1.10	0.68	0.5
age	-0.25	0.04	-6.66	0.0

Still correct answer (age = neutral control)

Adjusted Means

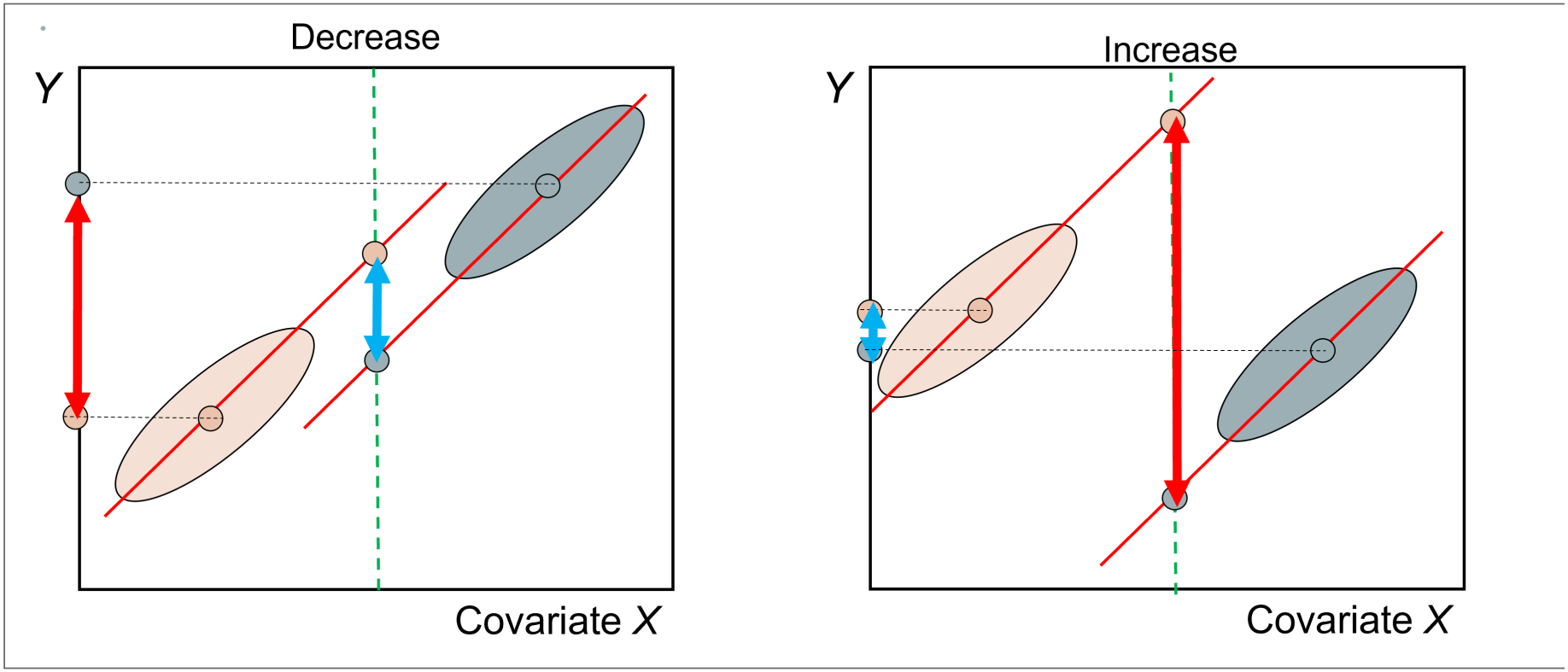
Controlling for Covariate

One way to think of controlling for the covariate is to account for differences between groups

- Recall: in multiple regression, we get the effect of each predictor while controlling for all other predictors

We can use this property to reconstruct the “adjusted means” we would have observed if both groups had scored equal on the covariate

Visualization



Using Regression Coefficients

- The regression coefficients give us the mean of the reference group, for a covariate value of 0
- The slope of the dummy variable(s) allows us to calculate the mean for the other group(s), for a covariate value of 0
- We can use the regression formula to calculate adjusted means for other covariate values

Example Adjusted Means

The unadjusted mean is $M = 22.75$ for tea drinkers, and $M = 12.02$ for red bull drinkers.

	B	SE	t	p
(Intercept)	30.26	2.25	13.43	0.0
drinkRed Bull	-3.09	2.42	-1.28	0.2
age	-0.26	0.07	-3.53	0.0

The adjusted means for 20-year old participants are:

- 20yo tea drinkers: $30.26 + (20 * -0.26) = 25.00$
- 20yo red bull drinkers: $30.26 + -3.09 + (20 * -0.26) = 21.90$

Method 2: Using Group Means

$$\bar{Y}_g^{adj} = \bar{Y}_g - b(\bar{X}_g - \bar{X})$$

Where:

- $\bar{Y}_g^{adj}$: Adjusted mean of the outcome for group g
- $\bar{Y}_g$: Unadjusted mean of the outcome for group g
- b: Regression coefficient of the covariate
- $\bar{X}_g$: Group mean of covariate X
- $\bar{X}$: Overall mean of covariate X

Error ×