Lecture 4: Multivariate OLS

Dr. Rachel Blum

November 5, 2019

The multivariate model

Shortcoming of Bivariate OLS

- It's rare that a social science or policy outcome is explained by a single factor—plenty of other factors might determine the outcome.
- This is called **omitted variable bias**. We can reduce it (and endogeneity) by adding more variables to the core model.

Multivariate OLS to the rescue!

- **Multivariate OLS** is the same as our initial OLS model—but with multiple independent variables.
- Two major benefits over Bivariate OLS:
    1. Reduces the bias that was created by endogeneity problems)
    2. Improves the precision of our estimates

The core model for Multivariate OLS

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \epsilon_i$$

β_0 Still our intercept (expected value for Y when all IVs equal zero)

β_1 Slope for X_1; *All else equal*, expected change in Y for one-unit increase in X_1.

$\vdots$

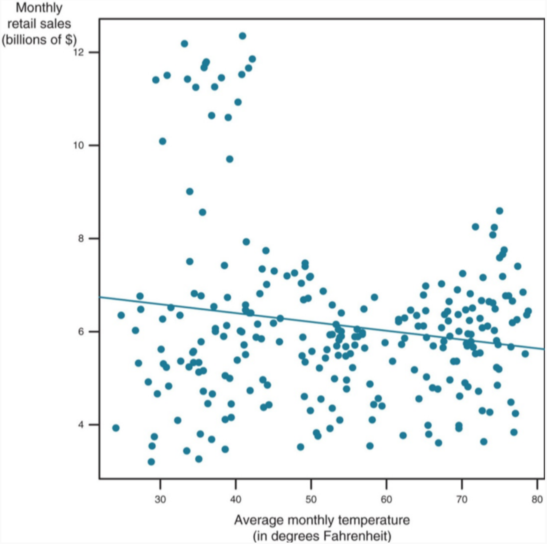
β_k Slope for X_k; *all else equal*, expected change in Y for one-unit increase in X_k. Here k is the number of independent variables in our analysis.

Multivariate OLS with *two* independent variables

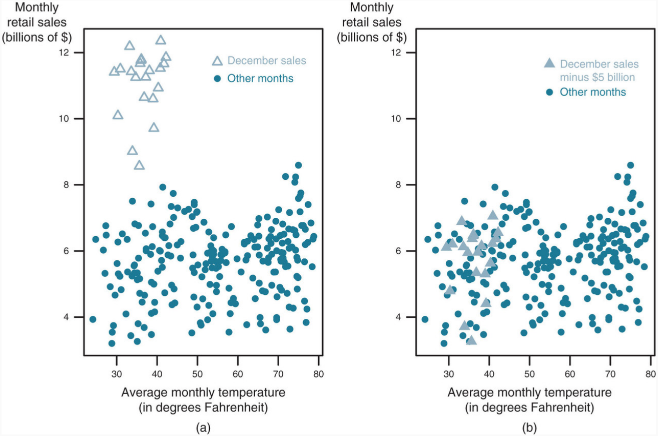
Control variables account for factors that could affect the dependent variable and/or be correlated with the independent variable.

- Allow us to net out the effect of the control variable and *then* look at the relationship between our outcome and key IVs.

Example: Retail and temperature



Example: Retail and temperature



Example: Retail and temperature

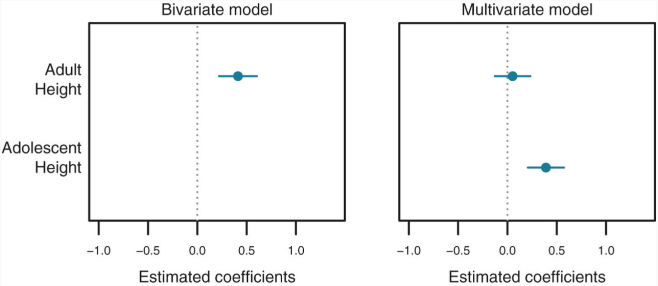
TABLE 5.1 Bivariate and Multivariate Results for Retail Sales Data

	Bivariate	Multivariate
Temperature	−0.019*	0.014*
	(0.007)	(0.005)
	[$t = 2.59$]	[$t = 3.02$]
December		5.63*
		(0.26)
		[$t = 21.76$]
Constant	7.16*	4.94*
	(0.41)	(0.26)
	[$t = 17.54$]	[$t = 18.86$]
N	256	256
$\hat{\sigma}$	1.82	1.07
R^2	0.026	0.661

Standard errors in parentheses.

* *indicates significance at* $p < 0.05$*, two-tailed.*

Example: Height and wages



Example: Height and wages

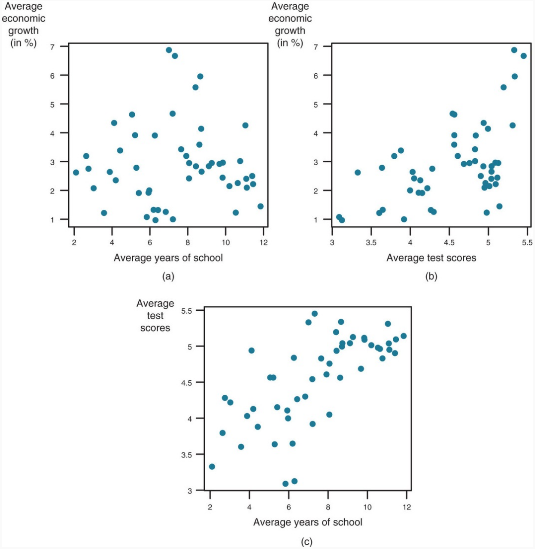
TABLE 5.2 Bivariate and Multiple Multivariate Results for Height and Wages Data

	Bivariate	Multivariate (a)	Multivariate (b)
Adult height	0.41*	0.003	0.03
	(0.10)	(0.20)	(0.20)
	[$t = 4.23$]	[$t = 0.02$]	[$t = 0.17$]
Adolescent height		0.48*	0.35
		(0.19)	(0.19)
		[$t = 2.49$]	[$t = 1.82$]
Athletics			3.02*
			(0.56)
			[$t = 5.36$]
Clubs			1.88*
			(0.28)
			[$t = 6.69$]
Constant	−13.09	−18.14*	−13.57
	(6.90)	(7.14)	(7.05)
	[$t = 1.90$]	[$t = 2.54$]	[$t = 1.92$]
N	1,910	1,870	1,851
$\hat{\sigma}$	11.9	12.0	11.7
R^2	0.01	0.01	0.06

Standard errors in parentheses.

* *indicates significance at $p < 0.05$, two-tailed.*

Example: Education and salary



Example: Education and salary

TABLE 5.3 Using Multiple Measures of Education to Study Economic Growth and Education

	Without math/science test scores	With math/science test scores
Avg. years of school	0.44*	0.02
	(0.10)	(0.08)
	[$t = 4.22$]	[$t = 0.28$]
Math/science test scores		1.97*
		(0.24)
		[$t = 8.28$]
GDP in 1960	−0.39*	−0.30*
	(0.08)	(0.05)
	[$t = 5.19$]	[$t = 6.02$]
Constant	1.59*	−4.76*
	(0.54)	(0.84)
	[$t = 2.93$]	[$t = 5.66$]
N	50	50
$\hat{\sigma}$	1.13	0.72
R^2	0.36	0.74

Standard errors in parentheses.

* *indicates significance at $p < 0.05$, two-tailed.*

Control variable types

Review: Variable types

1. **Categorical** (aka *nominal*): can be put in categories. Order is arbitrary.
    - **Ordinal** (aka *ranked*): categorical, but with a clear order.
    - **Binary** (aka *dummy* or *indicator*): two categories, reference and comparison. Order is arbitrary.
2. **Continuous** (aka *interval*): potentially infinite number of values. Order matters.
    - **Ratio**: similar to continuous/interval variables, but with meaningful zero. Order matters.

Dummy variables and OLS

Dummy variables let us compare differences in average outcome (**difference in means**) across two groups.

- Bivariate OLS is one way to test the difference in means.
- Multivariate OLS allows us to go further and test differences in means, *controlling for other factors* that might matter for our study.

Difference of means tests, way 1

Comparing the mean of Y for one group in our sample against the mean of Y for a different group in our sample using a *t-test*:

$$\bar{Y}_1 - \bar{Y}_0 \sim t_{df}$$

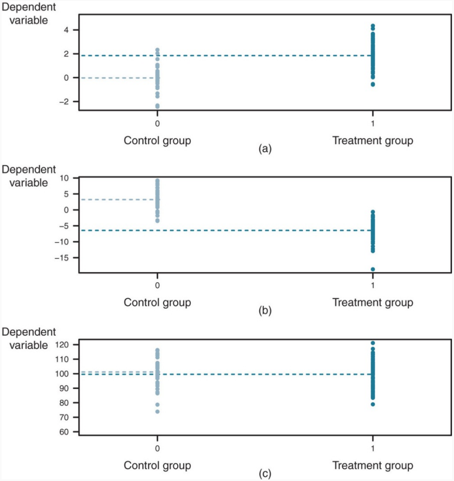
Difference of means tests, way 2

We can also do this using OLS (and get the same *t-statistic*)

$$Y_i = \beta_0 + \beta_1 Dummy_i + \epsilon_i$$

- β_0 (intercept) is the mean of Y when the dummy variable is zero (e.g., the control group)
- β_1 (slope) is the difference in the mean of Y when the dummy variable is one (e.g., the treatment group)

Visualizing this



When the dummy variable equals zero...

The mean for the control group is given by the intercept.

$$\begin{aligned}\hat{Y}_i &= \hat{\beta}_0 + \hat{\beta}_1 Dummy_i \\ &= \hat{\beta}_0 + \hat{\beta}_1 \times 0 \\ &= \hat{\beta}_0\end{aligned}$$

When the dummy variable equals one...

The mean for the treatment group is given by the intercept **plus** the slope.

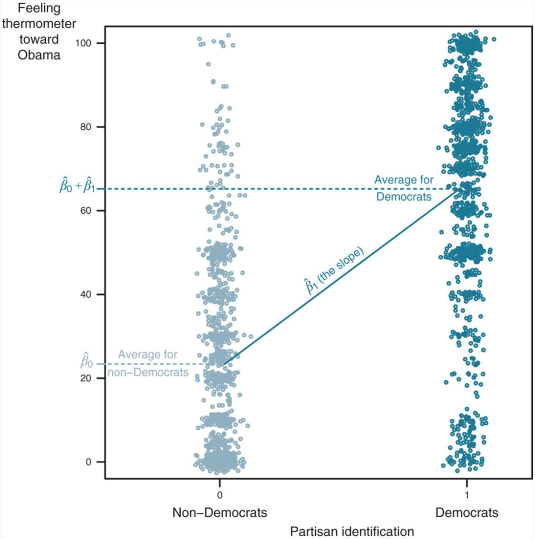
$$\begin{aligned}\hat{Y}_i &= \hat{\beta}_0 + \hat{\beta}_1 Dummy_i \\ &= \hat{\beta}_0 + \hat{\beta}_1 \times 1 \\ &= \hat{\beta}_0 + \hat{\beta}_1\end{aligned}$$

Interpreting coefficients of dummy variables

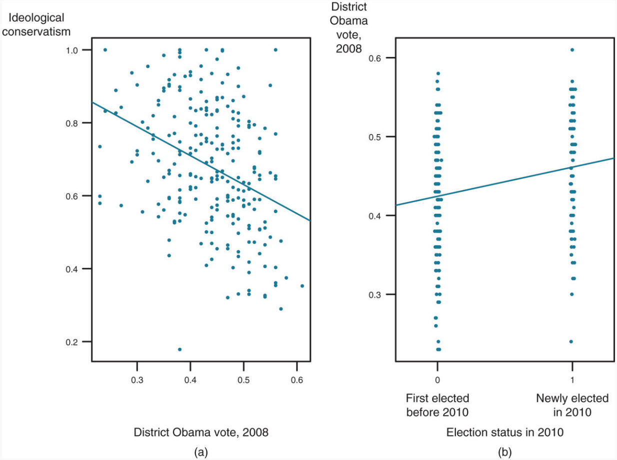
So how to interpret $\hat{\beta}_1$?

- If $\hat{\beta}_0$ is the mean of Y_0 and $\hat{\beta}_0 + \hat{\beta}_1$ is the mean of Y_1, then $\hat{\beta}_1$ is the **difference in means between the two groups**.
- Standard error still tells us how much uncertainty comes from sample size, variance of X, and variance of the regression $\hat{\sigma}^2$.
- Confidence interval stills tells us we are 95% confident that the difference in means between groups is between $\hat{\beta}_1 - t_{critical} \times se(\hat{\beta}_1)$ and $\hat{\beta}_1 + t_{critical} \times se(\hat{\beta}_1)$.

Example: Party ID and Obama



Example: Party ID and Obama



Example: Party ID and Obama

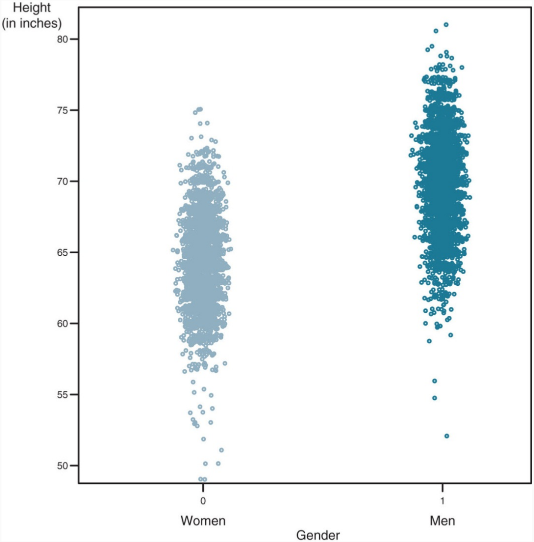
TABLE 6.1 Feeling Thermometer Toward Barack Obama

	Treatment = Democrat	Treatment = Not Democrat
Democrat	41.82*	
	(1.09)	
	[$t = 38.51$]	
Not Democrat		−41.82*
		(1.09)
		[$t = 38.51$]
Constant	23.38*	65.20*
	(0.78)	(0.76)
	[$t = 30.17$]	[$t = 85.72$]
N	2,183	2,183
R^2	0.40	0.40

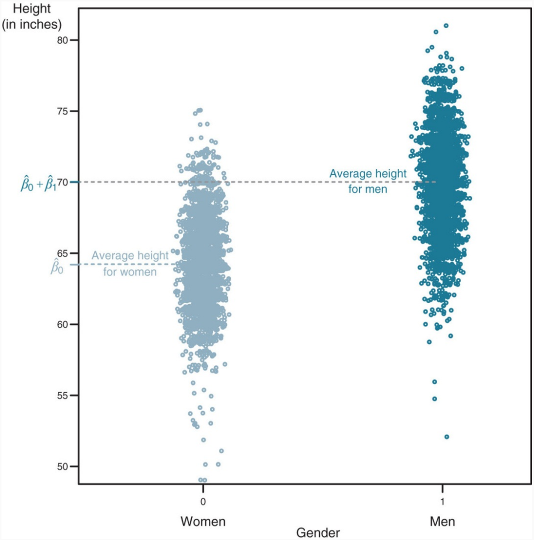
Standard errors in parentheses.

* *indicates significance at* $p < 0.05$, *two-tailed.*

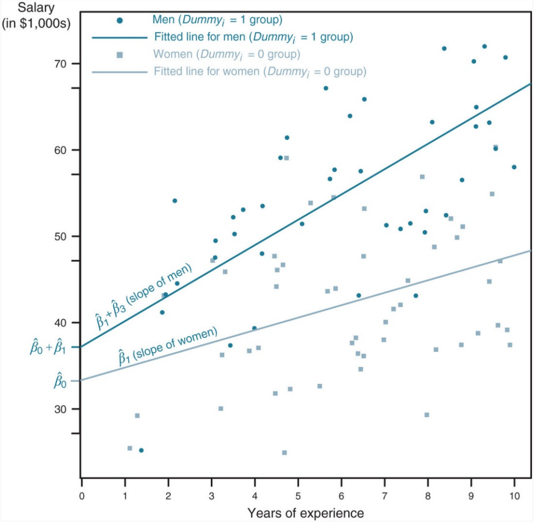
Example: Height and wages



Example: Height and wages



Example: Height and wages



Categorical variables

Categorical variables take ≥ 3 categories; categories have no intrinsic ordering.
Examples:

- *Categorical* region: 1 = Northeast, 2 = Midwest, 3 = South, 4 = West
- *Ordinal* survey preferences: 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = strongly agree.

Categorical variables in multivariate OLS

- How do we incorporate these variables into a regression model?
- Do we add them in as one variable to the model?

Transforming categorical variables to multiple dummy variables

- No, we cannot simply add categorical variables to the model because a one-unit change has no proper meaning as we shift from 1 to, say, 4.
- Instead, we break them apart into multiple dummy variables:

$$Y_i = \beta_0 + \beta_1 Northeast_i + \beta_2 Midwest_i + \beta_3 South_i + \epsilon_i$$

- The catch is that we can't include all categories. Why not?

Reference category

- If we add dummy variables for every category, we end up with **perfect multicollinearity**.
- Exclude one dummy, treat as **reference category**.
- Coefficients on each dummy are differences of means in reference to the reference category.

Example: region

TABLE 6.5 Using Different Excluded Categories for Wages and Region

	(a) Exclude West	(b) Exclude South	(c) Exclude Midwest	(d) Exclude Northeast
Northeast	2.02*	4.15*	3.61*	
	(0.59)	(0.506)	(0.56)	
	[$t = 3.42$]	[$t = 8.19$]	[$t = 6.44$]	
Midwest	−1.59*	0.54		−3.61*
	(0.534)	(0.44)		(0.56)
	[$t = 2.97$]	[$t = 1.23$]		[$t = 6.44$]
South	−2.13*		−0.54	−4.15*
	(0.48)		(0.44)	(0.51)
	[$t = 4.47$]		[$t = 1.23$]	[$t = 8.19$]
West		2.13*	1.59*	−2.02*
		(0.48)	(0.53)	(0.59)
		[$t = 4.47$]	[$t = 2.97$]	[$t = 3.42$]
Constant	12.50*	10.37*	10.91*	14.52*
	(0.40)	(0.26)	(0.36)	(0.43)
	[$t = 31.34$]	[$t = 39.50$]	[$t = 30.69$]	[$t = 33.53$]
N	3,223	3,223	3,223	3,223
R^2	0.023	0.023	0.023	0.023

Standard errors in parentheses.

* *indicates significance at $p < 0.05$, two-tailed.*

Example: region

TABLE 6.6 Hypothetical Results for Wages and Region When Different Categories Are Excluded

	Exclude West	Exclude South	Exclude Midwest	Exclude Northeast
Constant	125.0	95.0	(d)	(g)
	(0.9)	(1.1)	(1.0)	(0.9)
Northeast	−5.0	(a)	(e)	
	(1.3)	(1.4)	(1.3)	
Midwest	−10.0	(b)		(h)
	(1.4)	(1.5)		(1.3)
South	−30.0		(f)	(i)
	(1.4)		(1.5)	(1.4)
West		(c)	10.0	(j)
		(1.4)	(1.4)	(1.3)
N	1,000	1,000	1,000	1,000
R^2	0.3	0.3	0.3	0.3

Standard errors in parentheses.

Making model choices

What variables should I include in my model?

A variable X_2 must be included in a model if:

- X_2 is correlated with X_1
- X_2 is independently associated with Y
- Driven by theory.

Focus on a variable of interest

We have a theory that says X_1 causes Y.

- We control for X_2 (. . . and X_3, etc.) because the relationship might be spurious.
- Choose your controls to get the best estimate of X_1 on Y.
- Don't focus on the best estimate of X_2, except as it will help with X_1.

Can you include too much?

- You may run out of degrees of freedom.
- Hard to visualize the data.
- High leverage cases can emerge where they are not expected.

Two approaches

1. Control for everything
   - If it might belong, include it.
   - Motivated by ruling out rival hypotheses.
   - Motivated by responding to critics.
   - Kitchen sink regression/Garbage Can Model

2. Rule of three
   - Prioritize interpretability.
   - Include more than three explanatory (independent) variables.
   - Use limited samples to control for other things.

What NOT to do

1. Try to make your R_2 big.
2. Try to minimize noise at all costs.
3. Data mine.

What NOT to do

- Don't search blindly until you find a model that looks good.
- Especially dangerous if you start focusing too much on statistical significance (p-values) over substantive significance.
- By chance alone, at some point 1/20 variables will be statistically significant.

On deck

1. PS3
2. Make sure to catch up on readings if you haven't already
3. I'll post PS3 and grades/answer keys later today.