Social Science Inquiry II

Week 7: Multivariate regression, part I

Molly Offer-Westort

Department of Political Science,
University of Chicago

Winter 2023

Loading packages for this class

```
> library(estimatr)
> library(modelsummary)
```

Housekeeping

- Final project

Banerjee, A., Duflo, E., Glennerster, R., & Kinnan, C. (2015). The miracle of microfinance? Evidence from a randomized evaluation. *American Economic Journal: Applied Economics*, 7(1), 22-53.

Data is available at openICPSR:
`https://www.openicpsr.org/openicpsr/project/113599`

Reading papers

What to get out of reading a research paper:

- What is the main question of the paper?
- What method do the authors use to address the question? For empirical papers:
  - Data (Where does it come from/how is it generated? What is the sample population? What is being measured?)
  - Research design/strategy
  - Statistical tools
- What is the answer that the authors get to the main question?

How would you answer these questions with the Banerjee et al. (2015) paper?

- The miracle of microfinance

- What is the point of microfinance, and why might incurring debt help poor households?
- Downsides?
- Proposed add-on benefits?

Study design

- Neighborhood selection criteria
  These areas were selected based on having no preexisting microfinance presence and on having residents who were desirable potential borrowers: poor, but not "the poorest of the poor." Areas with high concentrations of construction workers were avoided because they move frequently, which makes them undesirable as microfinance clients...Conversely, the largest such areas in Hyderabad were not selected for the study, since Spandana was keen to start operations there.

Study design

- Baseline survey “random” sampling procedures. Problems with this?
- Pairwise randomization: Why?
- Endline survey sampling procedures:
  - households with high likelihood of having borrowed: *those that had resided in the area for at least 3 years and contained at least 1 woman aged 18 to 55.*
  - Oversample Spandana borrowers to search for heterogeneous treatment effects; does this cause problems for inference?

- What is the difference between treatment and control in this setting?
- Reduced-form/intent-to-treat estimates

Too many variables. . .

```
> dat <- read.csv('../data/banerjee-et-al.csv')
> str(dat)
'data.frame':	6863 obs. of  188 variables:
 $ X                       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ hhid                    : int  1 2 3 4 5 6 7 8 9 10 ...
 $ areaid                  : int  1 1 1 1 1 1 1 1 1 1 ...
 $ treatment               : int  1 1 1 1 1 1 1 1 1 1 ...
 $ w                       : num  0.82 1 1 1 1 ...
 $ w1                      : num  0.777 1 1 1 1 ...
 $ w2                      : num  0.82 1 1 1 1 ...
 $ sample1                 : int  1 1 1 1 1 1 1 1 1 1 ...
 $ sample2                 : int  1 1 1 1 1 1 1 1 1 1 ...
 $ old_biz                 : int  0 0 1 1 1 1 0 0 1 1 ...
 $ any_old_biz             : int  0 0 1 1 1 1 0 0 1 1 ...
 $ area_pop_base           : int  272 272 272 272 272 272 272 272 272 272 ...
 $ area_debt_total_base    : num  81050 81050 81050 81050 81050 ...
 $ area_business_total_base : int  11 11 11 11 11 11 11 11 11 11 ...
 $ area_exp_pc_mean_base   : num  1335 1335 1335 1335 1335 ...
 $ area_literate_head_base  : num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
 $ area_literate_base      : num  0.534 0.534 0.534 0.534 0.534 ...
 $ visitday_1              : int  22 22 23 22 22 23 23 22 22 22 ...
 $ visitmonth_1            : int  8 8 8 8 8 8 8 8 8 8 ...
 $ visityear_1             : int  2007 2007 2007 2007 2007 2007 2007 2007 2007 2007 ...
 $ visitday_2              : int  16 16 16 16 17 13 16 16 16 17 ...
 $ visitmonth_2            : int  12 12 12 12 12 5 12 12 12 12 ...
 $ visityear_2             : int  2009 2009 2009 2009 2009 2010 2009 2009 2009 2009 ...
 $ hhsize_1                : int  3 4 5 5 6 6 4 4 7 6 ...
 $ hhsize_adj_1            : num  2.8 3.24 4.18 4.03 5.41 ...
 $ adults_1                : int  3 2 2 2 4 3 4 2 7 4 ...
 $ children_1              : int  0 2 3 3 2 3 0 2 0 2 ...
 $ male_head_1             : int  1 1 1 1 1 1 1 1 1 1 ...
 $ head_age_1              : int  20 34 40 37 32 40 43 31 62 64 ...
 $ head_noeduc_1           : int  1 0 0 0 0 1 1 0 0 1 ...
 $ women1845_1             : int  2 1 1 1 1 1 2 1 2 1 ...
 $ anychild1318_1          : int  0 0 1 1 1 1 1 0 1 0 ...
 $ hhsize_2                : int  3 4 6 7 6 7 6 4 7 6 ...
                                  2.42 3.51 5.35 6.08 5.41
```

Table 2, Panel A

	Spandana	Other MFI	Any MFI	Other bank	Informal	Total	Ever Late	Num Cycles
treatment	0.127***	−0.012	0.083**	0.003	−0.052*	−0.022	−0.060*	0.084*
	(0.020)	(0.025)	(0.028)	(0.012)	(0.022)	(0.014)	(0.027)	(0.043)
Num.Obs.	6811	6657	6811	6811	6811	6862	6475	6816
R2	0.052	0.014	0.019	0.003	0.010	0.007	0.015	0.011
R2 Adj.	0.051	0.012	0.018	0.002	0.009	0.006	0.014	0.010
AIC								13 826.9
BIC								13 888.3
RMSE								0.67
Std.Errors	by: areaid	by: areaid	by: areaid	by: areaid	by: areaid	by: areaid	by: areaid	by: areaid
Control mean	0.051	0.149	0.183	0.079	0.761	0.867	0.616	0.33

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Interpreting multiple regression

For relationships with a single independent variable, we proposed the following framework:

- We would like to describe a conditional relationship in the data

$$\mathrm{E}\left[Y|X=x\right]=h(x)$$

where a simple version is

$$g(x)=\beta_0+\beta_1 x$$

- In other words,

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

where

$$\mathrm{E}\left[\epsilon_i|X_i\right]=0$$

and

$$\mathrm{Var}\left[\epsilon_i|X_i\right]=\sigma^2$$

- For a multivariate generalization to K variables, we can consider:

$$\mathrm{E}\left[Y|X_1 = x_1, X_2 = x_2, \ldots, X_K = x_K\right] = h(x_1, x_2, \ldots, x_K)$$

  where the we propose that $g(x_1, x_2, \ldots, x_K)$ is

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_K x_K$$

- This is a *model* of the conditional relationship we're interested in.
- It produces a *linear approximation* of the conditional expectation, but if the true relationship is not linear, we are just approximating it.

- We can then define *residuals* in the same way we did for the univariate model:

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_K X_{Ki}\right).$$

- In the same way as with univariate regression, we calculate estimates of $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, ..., $\hat{\beta}_K$ as the values that minimize the residual sums of squares

$$\text{RSS} = \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

- This tells us how we get the least squares regression estimates for coefficients, but how do we interpret them?

- Consider a simple multiple regression model,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

- Including more independent variables in our model can give us more predictive power for Y; if Y varies with X_1 *and* X_2, and X_1 and X_2 are not perfectly correlated, we are going to be able to explain more of the variation in Y by including both X_1 and X_2.
- We interpret the coefficient on β_1 as the amount that our prediction for Y changes with a one unit change in X_1, *holding the value of* X_2 *constant*. This is what we mean when we say we *control* for additional variables in a model.

- This model is not inherently causal. It just describes a relationship.
- Consider a *causal* model, where W_i is a manipulated binary treatment variable:

$$Y = \gamma_0 + \gamma_1 W + \epsilon$$

- The model is not different from our univariate regression model, but our interpretation differs based on what we know about the relationships between the variables.
- We can also have a causal model, where we have one causal variable, and we control for additional explanatory variables that are not directly causal.

$$Y = \gamma_0 + \gamma_1 W + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

- If we already believe that we can get a causal interpretation from the first model (for example, because of *independence* of treatment and outcomes), why would we use the second model?

Returning to Table 2, Panel A

	Spandana	Other MFI	Any MFI	Other bank	Informal	Total	Ever Late	Num Cycles
treatment	0.127***	−0.012	0.083**	0.003	−0.052*	−0.022	−0.060*	0.084*
	(0.020)	(0.025)	(0.028)	(0.012)	(0.022)	(0.014)	(0.027)	(0.043)
Num.Obs.	6811	6657	6811	6811	6811	6862	6475	6816
R2	0.052	0.014	0.019	0.003	0.010	0.007	0.015	0.011
R2 Adj.	0.051	0.012	0.018	0.002	0.009	0.006	0.014	0.010
AIC								13 826.9
BIC								13 888.3
RMSE								0.67
Std.Errors	by: areaid	by: areaid	by: areaid	by: areaid	by: areaid	by: areaid	by: areaid	by: areaid
Control mean	0.051	0.149	0.183	0.079	0.761	0.867	0.616	0.33

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

```
> formula1 <- formula(paste0('spandana_1 ~ treatment',
+                            '+ area_pop_base',
+                            '+ area_debt_total_base',
+                            '+ area_business_total_base',
+                            '+ area_exp_pc_mean_base',
+                            '+ area_literate_head_base',
+                            '+ area_literate_base'))
> lm_robust(formula1, data = dat, clusters = areaid, weights = w1)
                              Estimate   Std. Error     t value     Pr(>|t|)
(Intercept)               3.190415e-02 8.686962e-02  0.36726470 7.161097e-01
treatment                 1.273842e-01 2.048632e-02  6.21801345 2.601529e-08
area_pop_base            -5.274797e-05 7.088902e-05 -0.74409225 4.605530e-01
area_debt_total_base      1.195938e-06 5.611348e-07  2.13128540 1.158952e-01
area_business_total_base -1.850362e-03 1.921429e-03 -0.96301363 3.435519e-01
area_exp_pc_mean_base    -1.057161e-06 6.256874e-05 -0.01689599 9.866699e-01
area_literate_head_base  -7.989478e-02 1.373983e-01 -0.58148313 5.653486e-01
area_literate_base        7.785476e-02 2.031835e-01  0.38317461 7.041409e-01
                              CI Lower     CI Upper        DF
(Intercept)              -1.458268e-01 2.096351e-01 28.766786
treatment                 8.657366e-02 1.681948e-01 75.030695
area_pop_base            -1.953871e-04 8.989115e-05 46.643193
area_debt_total_base     -5.146508e-07 2.906527e-06  3.249778
area_business_total_base -5.781191e-03 2.080468e-03 28.819932
area_exp_pc_mean_base    -1.307158e-04 1.286014e-04 22.299643
area_literate_head_base  -3.607444e-01 2.009548e-01 29.388940
area_literate_base       -3.361247e-01 4.918342e-01 31.788222
```

- Banerjee et al. account for oversampling of Spandana borrowers, and also cluster standard errors at the area level (why?)

If we just use the regression estimate without controls, is it meaningfully different from the version with controls?

```
> formula2 <- formula('spandana_1 ~ treatment')
> lm_robust(formula2, data = dat, clusters = areaid, weights = w1)
```

```
              Estimate Std. Error  t value     Pr(>|t|)   CI Lower  CI Upper
(Intercept) 0.04801341  0.0137500 3.491883 1.129800e-03 0.02027712 0.0757497
treatment   0.12325966  0.0213796 5.765294 1.245994e-07 0.08075896 0.1657604
                  DF
(Intercept) 42.64019
treatment   86.07593
```

Table 2, Panel A, with controls

	Spandana	Other MFI	Any MFI	Other bank	Informal	Total	Ever Late	Num Cycles
treatment	0.127***	−0.012	0.083**	0.003	−0.052*	−0.022	−0.060*	0.084*
	(0.020)	(0.025)	(0.028)	(0.012)	(0.022)	(0.014)	(0.027)	(0.043)
Num.Obs.	6811	6657	6811	6811	6811	6862	6475	6816
R2	0.052	0.014	0.019	0.003	0.010	0.007	0.015	0.011
R2 Adj.	0.051	0.012	0.018	0.002	0.009	0.006	0.014	0.010
AIC								13 826.9
BIC								13 888.3
RMSE								0.67
Std.Errors	by: areaid	by: areaid	by: areaid	by: areaid	by: areaid	by: areaid	by: areaid	by: areaid
Control mean	0.051	0.149	0.183	0.079	0.761	0.867	0.616	0.33

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2, Panel A, without controls

	Spandana	Other MFI	Any MFI	Other bank	Informal	Total	Ever Late	Num Cycles
treatment	0.123***	−0.019	0.074*	0.005	−0.053*	−0.025+	−0.061*	0.079+
	(0.021)	(0.025)	(0.029)	(0.011)	(0.022)	(0.015)	(0.029)	(0.044)
Num.Obs.	6811	6657	6811	6811	6811	6862	6475	6816
R2	0.038	0.0008	0.008	0.000 07	0.004	0.001	0.004	0.003
R2 Adj.	0.038	0.0006	0.008	−0.000 07	0.003	0.001	0.004	0.003
AIC								13 864.4
BIC								13 884.9
RMSE								0.67
Std.Errors	by: areaid	by: areaid	by: areaid	by: areaid	by: areaid	by: areaid	by: areaid	by: areaid
Control mean	0.051	0.149	0.183	0.079	0.761	0.867	0.616	0.33

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

- Is there a design-based reason why we might not see much difference with covariate adjustment?

Where do we see effects of microfinance?

- Credit?
- Small business development? (among which businesses?)
- Poverty reduction?
- Household consumption?
- Social welfare: education, health, female empowerment?

Big picture take-away

- Take-aways from this study?
- How much can we generalize about effects of microfinance?

References I

Banerjee, A., Duflo, E., Glennerster, R., and Kinnan, C. (2015). The miracle of microfinance? Evidence from a randomized evaluation. American Economic Journal: Applied Economics, 7(1):22–53.