Empirical Methods II (API-202 B)
Miguel Angel Santos

Kennedy School of Government
Harvard University

HANDOUT 3 – MULTIVARIATE REGRESSION ANALYSIS (I)

AGENDA

- Basic concepts
- Example 1: California School Districts with two covariates
- Understanding "holding constant"
- Vocabulary
- Takeaways

BIBLIOGRAPHY FOR TODAY´S CLASS

- Bilmes and Stier (2010). Freeze on federal jobs won't reduce spending. (*)
- Stock and Watson, 4.1,4.2, Appendix 4.1, 5.1, 5.2, 6.2, 6.3 (**)
- Angrist and Pischke, 1 (pp. 39-46)

BASIC CONCEPTS

Our ultimate goal continues to be to learn about causal effect of one variable (X) on another variable (Y)

With observational data, many factors might be changing along with (X) that have influence in (Y), isolating (X)'s causal impact is challenging

Multiple regression estimates the effect of X on Y, while holding constant other factors that may be responsible for the observed association between X and Y

Vocabulary for "holding constant":

- Holding fixed
- Controlling for
- Conditional on
- Ceteris paribus

We can write a **population regression function (PRF)** with two explanatory variables, (X_1) and (X_2):

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

Ordinary least squares can be used to estimate all three coefficients. The sample regression function (SFR) then looks like:

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\varepsilon}$$

or

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

What we need to understand from each of these coefficients is:

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

The constant or Y-intercept β_0 tells us the predicted value of Y in those cases or individuals with X_1 =0 and X_2=0

The slope of coefficient β_1 tells us the average change in Y associated with a one-unit change in X_1, holding X_2 constant.

The slope of coefficient β_2 tells us the average change in Y associated with a one-unit change in X_2, holding X_1 constant.

EXAMPLE 1: CALIFORNIA SCHOOL DISTRICTS WITH TWO COVARIATES

Let us add income to the equation we used to estimate the impact of class size (student-teacher ratio) on test scores:

Y = *testscr*	district's mean math/reading 5th grade test score
X_1 = *str*	district's student-teacher ratio (class size)
X_2 = *avginc*	district's mean family income (in 000's of $)

```
. regress testscr str, robust

Linear regression                                      Number of obs =      420
------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -2.279808   .5194892    -4.39   0.000    -3.300945   -1.258671
       _cons |    698.933   10.36436    67.44   0.000     678.5602    719.3057
------------------------------------------------------------------------------


. regress testscr str avginc, robust

Linear regression                                      Number of obs =      420
------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -.6487401   .3533403    -1.84   0.067     -1.34329      .04581
      avginc |   1.839112    .114733    16.03   0.000     1.613585    2.064639
       _cons |   638.7292   7.301234    87.48   0.000     624.3773     653.081
------------------------------------------------------------------------------
```

Note that:

- An increase in the STR by one student, is associated "on average" to a significant decrease of 2.28 points in average test scores of the schools on that district
- Two students: 4.56 (the larger the delta, the higher the error we are making, as we risk the function not being linear: Concept of small changes=slope)
- We can also use equations to predict test scores: What is the test score expected in a classroom with STR 20 = 698.9 - 2.28*20 = 653.3 ("absent those other factors")
- Is that significant? (Interpret also the intervals)

What is the SRF being estimated by each model?

- Model 1:
- Model 2:

Interpret the class size coefficients from each model:

- Model 1:
- Model 2:

Are coefficients statistically significant at the 5% level?

- Model 1:
- Model 2:

Why does the class size coefficient change from model 1 to model 2?

UNDERSTANDING "HOLDING CONSTANT"

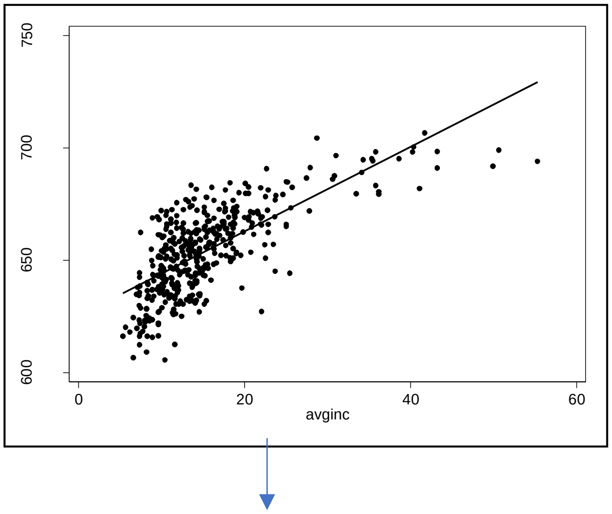
The Frisch-Waugh theorem guarantees that the β_1 that we get from out multivariate regression is the same coefficient we would get from the last of these three bivariate regressions:

1. Regress Y on $\mathbf{X}_2$, save the residuals
2. Regress $\mathbf{X}_1$ on $\mathbf{X}_2$, save the residuals
3. Regress the first set of residuals on the second set of residuals

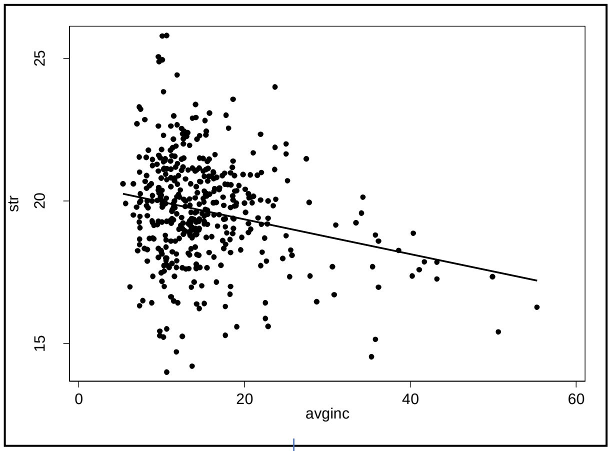
In other words, β_1 truly measures the relationship between "the part of X1 unexplained by X2, and that part of Y unexplained by X2".

Here is a way to visualize those first two generations:

Y (test scores) on X_2 (income)



X_1 (test scores) on X_2 (income)



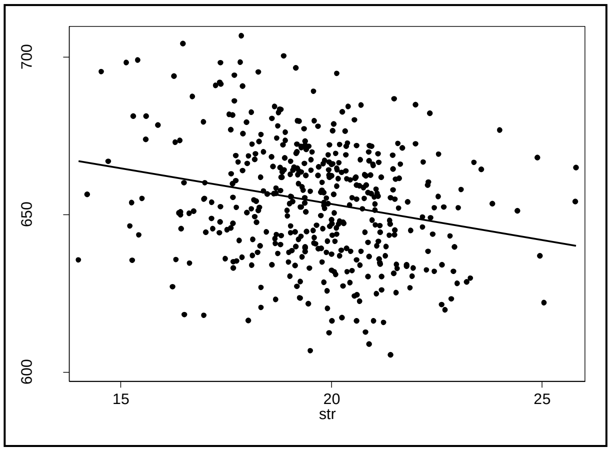
Higher income districts have higher test scores

Higher income districts have lower STRs

Some of the association we have noted in the bivariate regression between class size and the text scores is driven by income.

Original bivariate regression: Test scores and STR

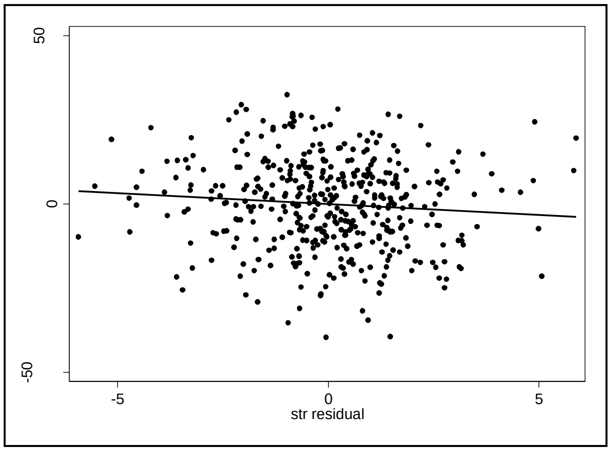
Y (test scores) on X_1 (STR)



```
Linear regression                                      Number of obs =     420
------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -2.279808   .5194892    -4.39   0.000    -3.300945   -1.258671
       _cons |    698.933   10.36436    67.44   0.000     678.5602    719.3057
------------------------------------------------------------------------------
```

Bivariate regression using residuals of (1) and (2)

Scores residuals and STR residuals



```
Linear regression                                      Number of obs =     420
------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -.6487401   .3533403    -1.84   0.067     -1.34329      .04581
      avginc |   1.839112    .114733    16.03   0.000     1.613585    2.064639
       _cons |   638.7292   7.301234    87.48   0.000     624.3773     653.081
------------------------------------------------------------------------------
```

TAKEAWAYS

- Multivariate regression helps us measure the magnitude of the relationship between two variables, while holding other variables constant
- As with bivariate regression, hypothesis testing and confidence help us measure how certain we are about that magnitude
- Holding other variables constant gets us closer to true causal estimates
- Multivariate regression is usually more convincing than bivariate regression
- The key question is whether the researcher has controlled for all relevant variables that substantially affect the story being told
- Deciding on the optimal number of variables we should control for is key (more on that coming)

VOCABULARY:

- Covariates/ Regressors (try not to use "independent variables"): a variable appearing on the right-hand side of a regression, that we assume is "causing" the other.
- Holding constant/holding fixed/conditional on/ controlling for/partial effects/ceteris paribus: the effect on (Y) of changing one of the regressors, holding the other regressors constant.
- Population regression function (PRF): It tells how the mean value of (Y) varies with (X), for the entire population. THE PRF is what we are trying to estimate in a linear regression.
- Sample regression function (SRF, as it is frequently difficult to obtain data for the entire population): the results of sampling procedures to obtain data from a sample and then estimate for the population.

- Constant/ Y-intercept: the regression intercept, the value of Y in the event that are Xs are zero.
- Confidence intervals: an interval that contains the true value of a population parameter with a pre-specified probability when computed over repeated samples.
- Coefficients: describe the relationship between each predictor variable and the response. The coefficient value represents the mean change in the response given a one-unit increase in the predictor.