
API-202B Empirical Methods II

Session #4: Multivariate regression (II) Dummy variables

miguel_santos@hks.harvard.edu
@miguelsantos12

Today's class: Multivariate regression (II) + Dummy variables

- Multiple regression with more than two covariates
- Reading regression tables
- Introduction to dummy variables
- Defining dummy variables
- Using dummy variables
- Program evaluation
- More than two categories
- Vocabulary
- Takeaways

Today's class: Multivariate regression (II) + Dummy variables

- Multiple regression with more than two covariates
- **Reading regression tables**
- Introduction to dummy variables
- Defining dummy variables
- Using dummy variables
- Program evaluation
- More than two categories
- Vocabulary
- Takeaways

California School Districts with four covariates

- Let us add two more variables to the equation we used to estimate the impact of class size (STR) controlling for income on test scores: a) computers per student, and b) fraction of students that have English as a second language:

$Y = testscr$ 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 \$)

$X_3 = comp_stu$ computers per student

$X_4 = el_pct$ fraction of students with English as second language

- PRF: $testscr = \beta_0 + \beta_1 str + \beta_2 avginc + \beta_3 comp_stu + \beta_4 el_pct + \varepsilon$

- SRF: $testscr = \hat{\beta}_0 + \hat{\beta}_1 str + \hat{\beta}_2 avginc + \hat{\beta}_3 comp_stu + \hat{\beta}_4 el_pct + \hat{\varepsilon}$

or

$$\widehat{testscr} = \hat{\beta}_0 + \hat{\beta}_1 str + \hat{\beta}_2 avginc + \hat{\beta}_3 comp_stu + \hat{\beta}_4 el_pct$$

Example 2: California School Districts with four covariates

```
. regress testscr str avginc comp_stu el_pct, robust
```

Linear regression

Number of obs = 420

testscr	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
str	.052132	.3016771	0.17	0.863	-.5408737	.6451377
avginc	1.48448	.0923195	16.08	0.000	1.303008	1.665952
comp_stu	13.8179	9.161063	1.51	0.132	-4.189974	31.82577
el_pct	-.4795127	.0285596	-16.79	0.000	-.5356522	-.4233733
_cons	636.0783	6.707339	94.83	0.000	622.8937	649.2629

- What is the predicted test score for a district with student-teacher ratio=20, average income=\$40,000, computers per student=0.5, and fraction of students that have English as a second language=20%?
- Interpret the student-teacher ratio coefficient? Does it have the expected sign? Is it important from a practical standpoint? Is it statistically significant?

Reading regression tables

- Most research articles display regression results in nice tables, not Stata!
- In the typical table, each column contains a single regression, with explanatory variables listed at left: Easy comparison of different regression specifications.
- Coefficients are reported, usually with standard errors below in parentheses so that you can divide to compute a t-statistic (sometimes they display the t-statistic directly)
- Statistically significant coefficients are often noted by stars, as specified in table notes. A typical example: * $p < .05$, ** $p < .01$.

Reading regression tables

- That table for all the regressions we have analyzed in class would look like this:

Dependent var.: Test score	(1)	(2)	(3)
Student-teacher ratio	-2.280** (0.519)	-0.649 (0.353)	0.0521 (0.302)
Average income		1.839** (0.115)	1.484** (0.092)
Computers per student			13.82 (9.161)
Fraction English learners			-0.480** (0.029)
Constant	698.9** (10.36)	638.7** (7.301)	636.1** (6.707)
N	420	420	420

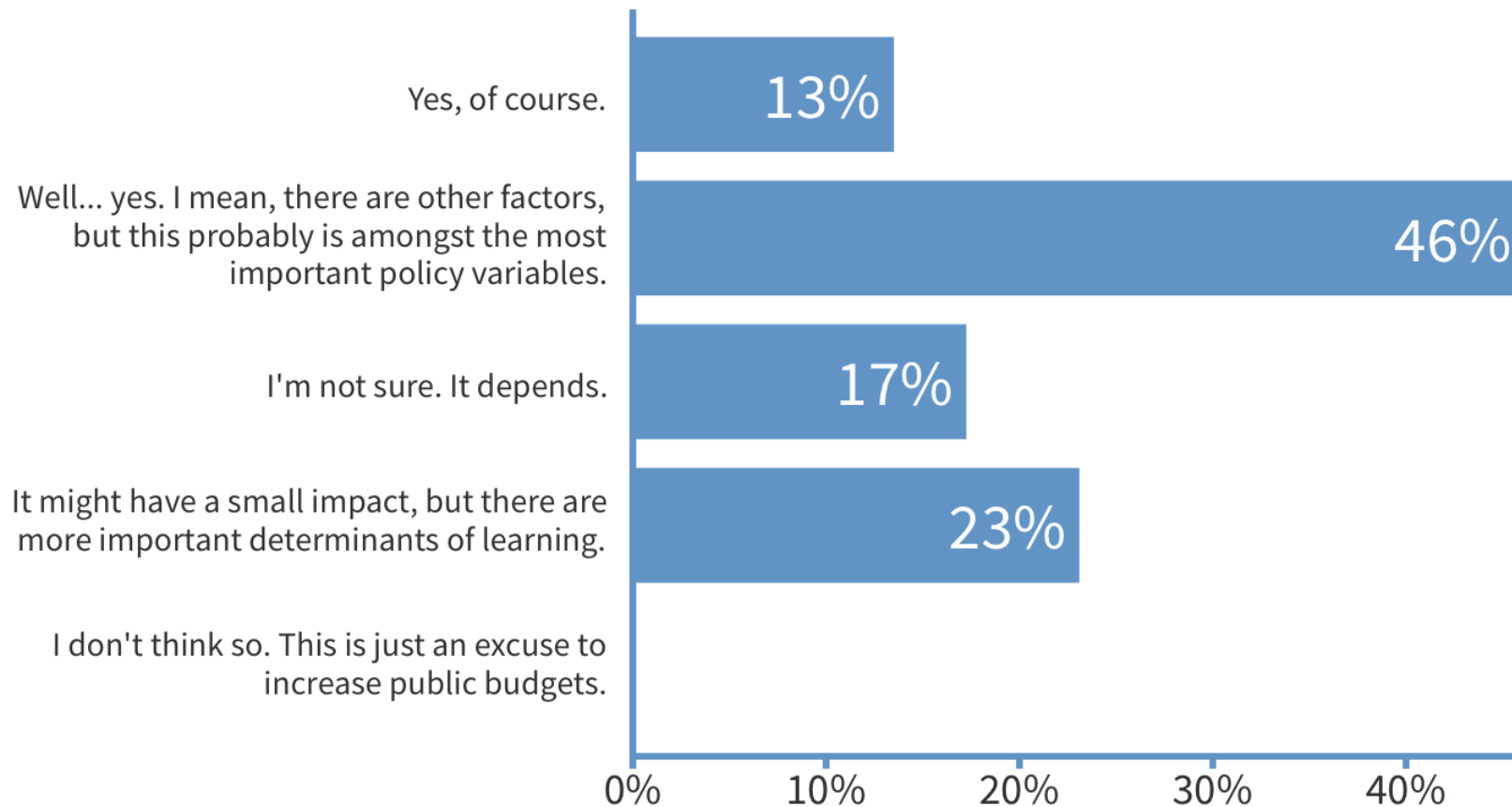
Do you believe that smaller class sizes enhance learning and have a lasting (positive) impact on students?



When poll is active, respond at PollEv.com/miguelsantos



Text **MIGUELSANTOS** to **37607** once to join



Introduction to Dummy variables

- Our regression models have used only continuous variables, which take a range of values (test scores, class sizes, income)
- We will often be interested in variables indicating whether an observation falls into a particular category. Is the individual:
 - Male or female?
 - Of indigenous origins?
 - Foreign worker?
 - Urban or rural?
- Dummy variables divide data into categories
- Researchers often have to assume simple categories for research purposes (two genders, four racial/ethnic groups, etc.)

BUSINESS DAY

How to Win the Battle of the Sexes Over Pay (Hint: It Isn't Simple.)

Economic View

By CLAUDIA GOLDIN NOV. 10, 2017

Defining dummy variables

- Dummy variables take only the values zero or one, and indicate whether an observation belongs to a category (0/1 are arbitrary values but make regressions easier to interpret)

Dummy variable
Binary variable
Indicator variable

- For example, we can define female as:
 - $female = 1$ if individual is female
 - $female = 0$ if individual is male
- Alternatively, we can define male as:
 - $male = 1$ if individual is male
 - $male = 0$ if individual is female

Defining dummy variables

- Here is data from the 2009 American Community Survey on MA residents 30-50 years old with positive earnings:

income: earnings in 2008 (in 000s dollars)

age: age in years

educ: years of schooling completed

male: = 1 if male (= 0 otherwise)

female: = 1 if female (= 0 otherwise)

person	income	age	educ	male	female
1	50	38	16	0	1
2	62	38	13	1	0
3	45	47	16	0	1
4	25	33	17	0	1

1432	73	43	17	1	0

Defining dummy variables

- What does male + female equal?
- When a set of variables always adds to a fixed number, we call those variables **multicollinear** or **collinear**
- **Multicollinearity** causes the dummy variable **trap**
- More generally: Multicollinearity occurs when one of the variables in a multiple regression model can be linearly predicted from one or various of the others

Using dummy variables

- We can also do this with a bivariate regression, using this PRF:

$$income_i = \beta_0 + \beta_1 female_i + \varepsilon_i$$

- The SRF for predicted earnings is thus:

$$\widehat{income}_i = \hat{\beta}_0 + \hat{\beta}_1 female_i$$

- Given this PRF and SRF, what are the predicted earnings of:
 - Men?
 - Women?

Using dummy variables

- Now look at the regression results:

```
. regress income female, robust noheader
```

income	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
female	-22.38741	2.30123	-9.73	0.000	-26.90156	-17.87326
_cons	66.75743	1.927088	34.64	0.000	62.97721	70.53766

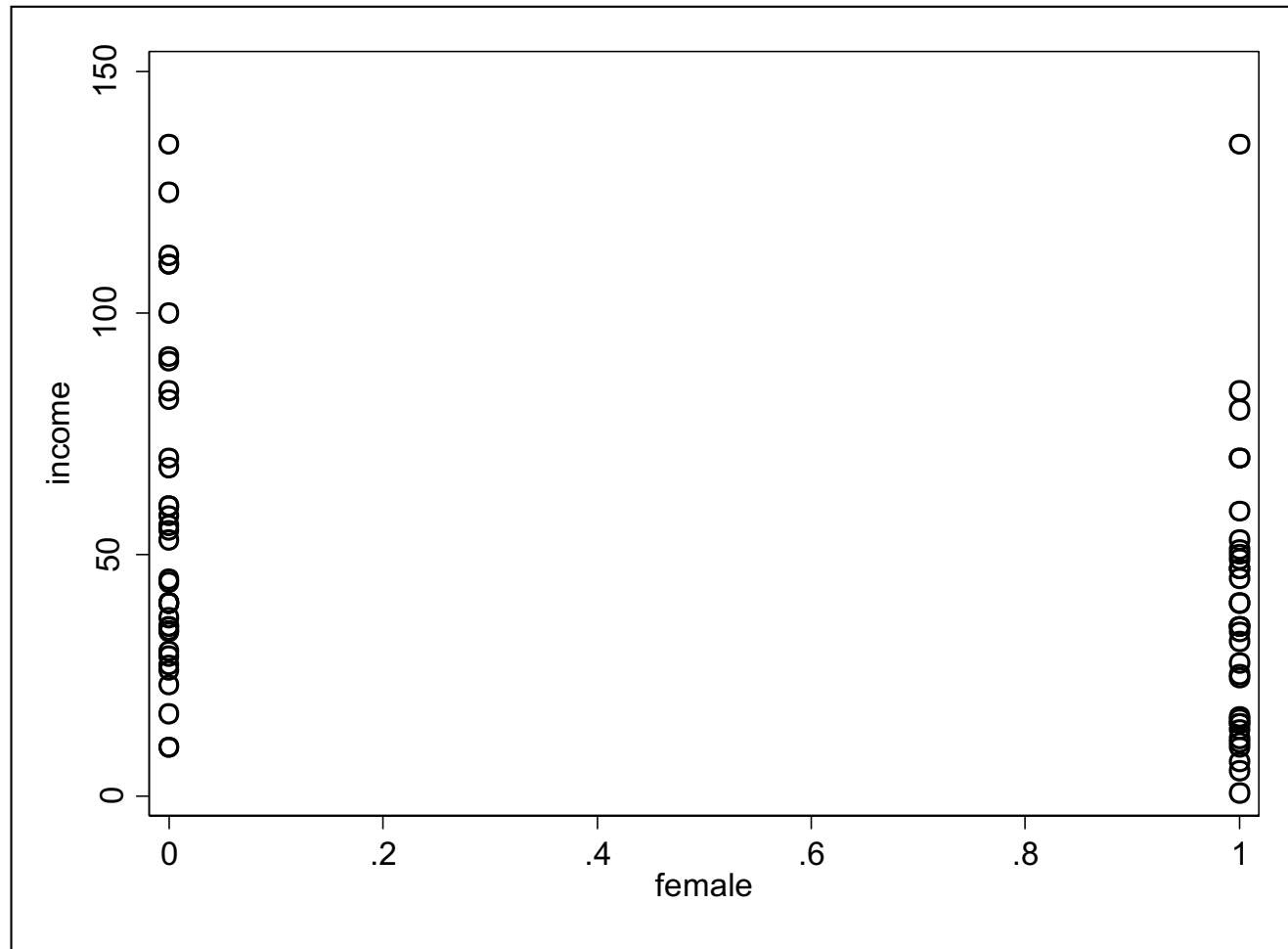
- What is the difference between women and men?

Using dummy variables

- What is the sample average income for men?
- What is the sample average income for women?
- What is the t-statistic for the test of the null hypothesis of no income difference in this population?
- Should we reject or fail to reject that null hypothesis?

Using dummy variables

- Here is a scatter plot to visualize this dummy variable regression:



- A bivariate regression with a dummy variable can be used for a comparison-of-means test between two groups

Using dummy variables

- Interpreting dummy variable coefficients **requires knowing the omitted group** against which comparisons are made
- We used men as the omitted group and compared women to them
- If we had instead used this PRF, our results would look like:

$$income_i = \alpha_0 + \alpha_1 male_i + \varepsilon_i$$

```
. regress income male, robust noheader
```

income	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
male	22.38741	2.30123	9.73	0.000	17.87326	26.90156
_cons	44.37003	1.257773	35.28	0.000	41.90275	46.8373

Using dummy variables

- What if we include both dummy variables in the regression?

$$income_i = \beta_0 + \beta_1 female_i + \beta_2 male_i + \varepsilon_i$$

```
. regress income female male, robust noheader
note: male omitted because of collinearity
```

income	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
female	-22.38741	2.30123	-9.73	0.000	-26.90156	-17.87326
male	0	(omitted)				
_cons	66.75743	1.927088	34.64	0.000	62.97721	70.53766

- ... that would lead to the **dummy variable trap**
- The trap is driven by our inclusion of at least one collinear variable, a variable that can be expressed as a linear combination of other variables
- Mind the omitted group in the interpretation of the coefficients

Using dummy variables in program evaluation

- Randomized experiments assign individuals to two categories, a treatment group and a control group. Bivariate regression with a dummy variable can thus estimate the impact of the treatment:
 - Define a dummy variable *treat* equal to 1 for those in treatment group and 0 for those in control group
 - Write out the PRF: $Y_i = \beta_0 + \beta_1 \textit{treat}_i + \varepsilon_i$
 - Estimate β_1 , the outcome difference between the two groups
- If you reject the null hypothesis $H_0: \beta_1 = 0$, we have found evidence of different outcomes between the two groups
- Because treatment is randomly assigned, no other factors should differ on average between the two groups. $\hat{\beta}_1$ will therefore give a causal estimate of the effect of the treatment on the outcome

Using dummy variables in program evaluation

- We previously used a comparison-of-means test to estimate the impact of the Tennessee STAR experiment on class size reduction
 - Students randomly assigned to a small class scored 13.7 points higher than students assigned to a regular class
 - We rejected the null hypothesis that the two groups had equal test scores (t-statistic of 6.38)
- Let's re-do this analysis with a bivariate regression approach
- Define a dummy variable *small*, equal to 1 for those assigned to a small class and 0 for those assigned to a regular class.
- Then run: $testscr_i = \beta_0 + \beta_1 small_i + \varepsilon_i$

Using dummy variables in program evaluation

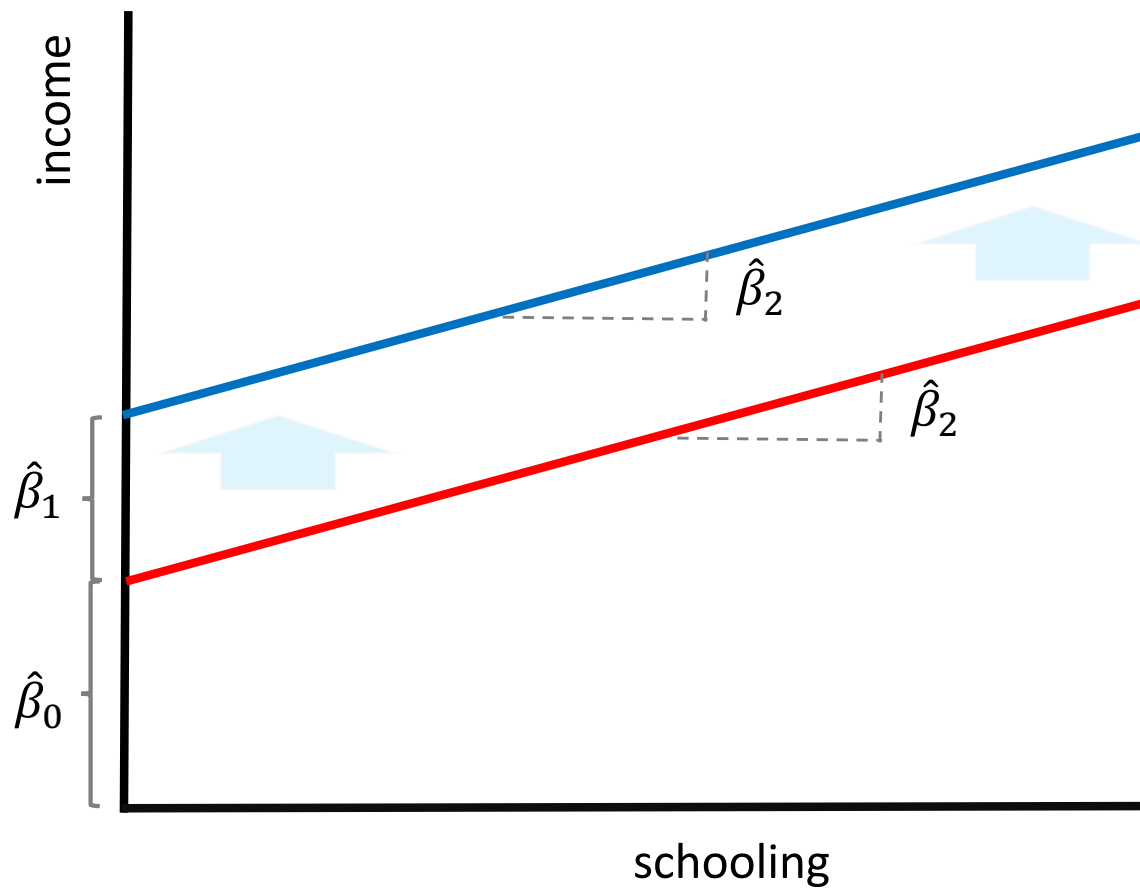
```
. regress testscr small, robust noheader
```

testscr	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
small	13.74055	2.154628	6.38	0.000	9.516677	17.96443
_cons	918.2013	1.135073	808.94	0.000	915.9762	920.4265

- Interpret $\hat{\beta}_0$
- Interpret $\hat{\beta}_1$
- Is the difference between the groups statistically significant?

Dummy variables in multiple regression models

- We can have a dummy variable within a multivariate regression:
- SRF: $\widehat{income} = \hat{\beta}_0 + \hat{\beta}_1 male + \hat{\beta}_2 schooling$



More than two categories

- We can use dummy variables with more than two categories
- For example, do different racial/ethnic groups in Massachusetts have different earnings?
- We will create one dummy variable for each of three groups (assume each person must be in exactly one of these categories):

white: = 1 if white (= 0 otherwise)

black: = 1 if black (= 0 otherwise)

hispanic: = 1 if Hispanic (= 0 otherwise)

person	income	age	educ	male	female	white	black	hispanic
1	50	38	16	0	1	1	0	0
2	62	38	13	1	0	1	0	0
3	45	47	16	0	1	0	0	1
4	25	33	17	0	1	0	1	0

1432	73	43	17	1	0	1	0	0

More than two categories

```
. regress income black hispanic, robust noheader
```

income	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
black	-15.03508	3.993734	-3.76	0.000	-22.86929	-7.200867
hispanic	-24.87962	2.46249	-10.10	0.000	-29.7101	-20.04914
_cons	57.05987	1.263512	45.16	0.000	54.58133	59.53841

- Interpret $\hat{\beta}_0$
- Interpret $\hat{\beta}_1$, $\hat{\beta}_2$
- Why not include a variable race, equal to 1 for whites, 2 for blacks, and 3 for Hispanics?

Takeaways

- Dummy variables take only the values 0 and 1: They are used to describe data that fall into 2 or more categories.
- Inclusion of variables that can be expressed as linear combination of other variables already in the regression (multicollinearity) results in leads to the dummy trap.
- Dummy variables can be used to estimate treatment effects in randomized experiments, comparing outcomes in treatment (dummy=1) and control groups (zero).
- The key to interpreting dummy variables is to be clear about which is the omitted category that the group of interest is being compared to.

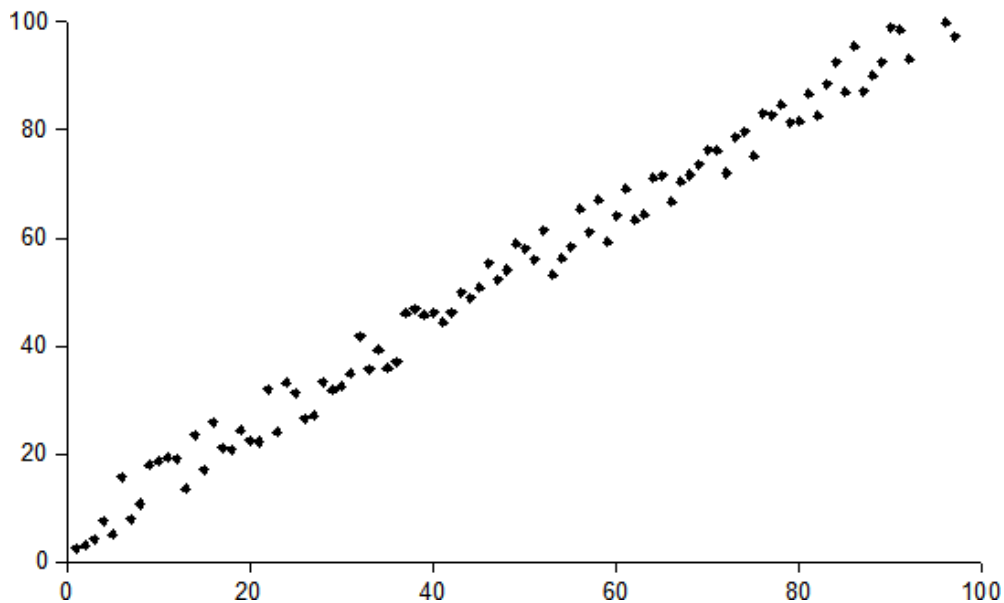
Vocabulary

- Dummy variable / Binary variable / Indicator variable
- Coefficient multiplying the dummy variable (as opposed to slope)
- Multicollinearity / Multicollinear variables / Collinear variables
- Dummy variable trap
- Omitted group

Optional (Multivariate regression): Robust standard errors

- The default formula used for computing standard errors of OLS regression coefficients assumes that the data are **homoskedastic** (variance around the regression line is the same for all values of the predictor variable X).
- Most data is at least somewhat **heteroskedastic** (variance around the regression line varies by value of the predictor variable X).
- The “robust” option for standard errors alters the formula to account for this.

Homoscedasticity



Heteroscedasticity

