Analysis of Economics Data
Chapter 14: Indicator Variables

© A. Colin Cameron
Univ. of Calif. Davis

November 2022

CHAPTER 14: Indicator Variables

- Regression when some regressors are **indicator variables**
  - also called **categorical variables** or **dummy variables**.
- An indicator variable records whether or not an observation falls in a particular category
  - e.g. whether or not person is female
  - e.g. whether or not a person is unemployed
  - coded $d = 1$ if in the category and $d = 0$ if not in the category.
- Warrants its own chapter as
  - frequently used in economics
  - interpretation is not always straightforward
    - interactions of indicators with regressors
    - for sets of mutually exclusive indicators omit one indicator
    - do joint F tests for statistical significance.

Outline

1. Example: Earnings, Gender, Education and Type of Worker
2. Regression on just a Single Indicator Variable
3. Regression on an Indicator Variable and Additional Regressors
4. Regression with Sets of Indicator Variables

Datasets: EARNINGS_COMPLETE

Example: Earnings, Education and Type of Worker

- Dataset EARNINGS_COMPLETE
  - 872 female and male full-time workers aged 25-65 years in 2000
  - indicators `Gender`, `d1`, `d2`, `d3` and interactions such as `Genderbyeduc`.

			Standard		
Variable	**Definition**	**Mean**	**Deviation**	**Min**	**Max**
Earnings	Annual earnings in $	56369	51516	4000	504000
Age	Age in years	43.31	10.68	25	65
Gender	= 1 if female	0.433	0.496	0	1
Education	Years of schooling	13.85	2.88	0	20
Genderbyeduc	*Gender* times *Education*	6.08	7.17	0	20
Age	Age in years	43.31	10.68	25	65
Genderbyage	*Gender* times *Age*	19.04	22.87	0	65
Hours	Usual hours worked per week	44.34	8.50	35	99
Genderbyhours	*Gender* times *Hours*	18.56	21.76	0	80
d1 or dself	= 1 if self-employed	0.089	0.286	0	1
d2 or dpriv	=1 if private sector employee	0.760	0.427	0	1
d3 or dgovt	=1 if govt. sector employee	0.149	0.356	0	1

14.2 Regression on a Single Indicator Variable

- **Indicator variable** takes just two values, for simplicity 0 and 1

$$d = \begin{cases} 1 & \text{if in the category} \\ 0 & \text{otherwise.} \end{cases}$$

- Regress y on just an intercept and the indicator variable

$$\widehat{y} = b + ad.$$

- Then $\widehat{y}_i$ takes one of only two possible values

$$\widehat{y}_i = \begin{cases} b + a & \text{if } d_i = 1 \\ b & \text{if } d_i = 0. \end{cases}$$

- For OLS regression it can be shown that

$$\begin{array}{rl} b = \bar{y}_0 & \text{where } \bar{y}_0 \text{ is mean of } y \text{ when } d = 0 \\ a = \bar{y}_1 - \bar{y}_0 & \text{where } \bar{y}_1 \text{ is mean of } y \text{ when } d = 1 \end{array}$$

- So the slope is the difference in means across the two categories.
- Inference is based on the population model $y = \beta + \alpha d + u$.

Example: Earnings and Gender

- Earnings by gender (-1 if female)

Gender	**Sample size**	**Mean**	**Standard Deviation**	**Min**	**Max**
Male Earnings (Gender=0)	494	63476	61713	5000	504000
Female earnings (Gender=1)	378	47080	31596	4000	322000

- OLS regression with heteroskedastic-robust standard errors

$$\widehat{Earnings} = \underset{(2290)}{63476} - \underset{(3478)}{16396} \times Gender \qquad R^2 = 0.025.$$

- Intercept $=$ mean male earnings $(d = 0) = 63,476$.
- Slope $=$ Difference in means $= 47080 - 47080 = -16396$
  - women earn $16,396 less on average
  - statistically significant at 5% as $t = -16396/3478 = -4.71$.

Difference in Means Specialized Methods

- Many areas of statistics avoid regression
  - instead use a specialized method for difference in means
  - e.g. Stata: `ttest earnings, by(gender) unequal`
  - yields same estimate but slightly different standard error.
- Two samples
  - $d = 1$ has mean $\bar{y}_1$, variance s_1^2 and $se(\bar{y}_1) = s_1 / \sqrt{n_1}$
  - $d = 0$ has mean $\bar{y}_0$, variance s_0^2 and $se(\bar{y}_0) = s_0 / \sqrt{n_0}$
- Estimate is $\bar{y}_1 - \bar{y}_0 = -16396$.
- Given independence of samples
  - $\text{Var}[\bar{y}_1 - \bar{y}_0] = \text{Var}[\bar{y}_1] + \text{Var}[\bar{y}_0]$
  - $se^2(\bar{y}_1 - \bar{y}_0) = se^2(\bar{y}_0) + se^2(\bar{y}_0)$
  - $se(\bar{y}_1 - \bar{y}_0) = \sqrt{se^2(\bar{y}_0) + se^2(\bar{y}_0)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_0^2}{n0}}$
- Here $se(\bar{y}_1 - \bar{y}_0) = \sqrt{(31596^2 / 378) + (61713^2 / 494)} = 3217$.

14.3 Regression on an Indicator Variable and Additional Regressors

- The difference in y across the two categories may be partly explained by other variables
  - e.g. earnings difference by gender is partly due to hours worked.
- Now bring in additional regressors (for simplicity just one)

$$y = \beta_1 + \beta_2 x + \alpha d + u.$$

- In the fitted model $\widehat{y} = b_1 + b_2 x + ad$

$$\widehat{y}_i = \begin{cases} b_1 + b_2 x_i + a & \text{if } d_i = 1 \\ b_1 + b_2 x_i & \text{if } d_i = 0. \end{cases}$$

- Now a measures the **difference** in y across categories **after controlling for the additional variables**.

Interacted Indicator Variables

- An **interacted indicator variable** is a regressor that is the product of an indicator variable and another regressor.
- Consider the model that adds the term $d \times x$.

$$y = \beta_1 + \beta_2 x + \alpha_1 d + \alpha_2 d \times x + u$$

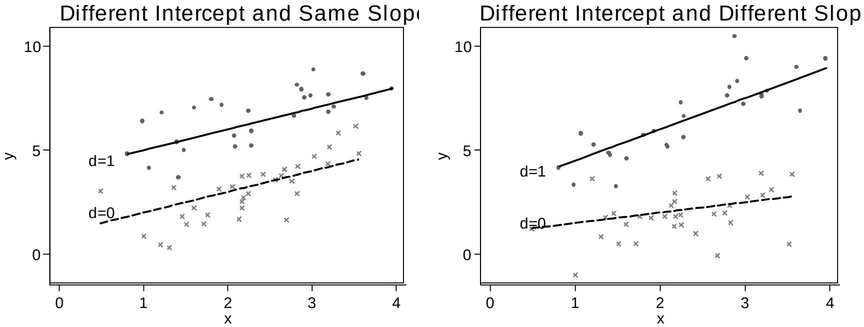
- In the fitted model $\widehat{y} = b_1 + b_2 x + a_1 d + a_2 d \times x$

$$\widehat{y} = \begin{cases} (b_1 + a_1) + (b_2 + a_2)x & \text{if } d = 1 \\ b_1 + b_2 x & \text{if } d = 0. \end{cases}$$

- An interacted indicator variable is a regressor that is the product of an indicator variable and another regressor.
- This enables slope coefficients to vary according to the value of the indicator variable.

Indicator Variable versus Indicator plus Interaction

- First panel: $\widehat{y} = b_1 + b_2 x + ad$
  - indicator variable shifts intercept
- Second panel: $\widehat{y} = b_1 + b_2 x + a_1 d + a_2 d \times x$
  - additional interacted regressor $d \times x$ additionally shifts slope.



Example

- Earnings on gender and education (with heteroskedastic-robust t statistics)

$$\widehat{Earnings} = \underset{(-2.20)}{-17552} - \underset{(-5.82)}{18258} \times Gender + \underset{(-5.82)}{5907} \times Education, \ R^2 = .134$$

- Add interaction between gender and education

$$\begin{aligned} \widehat{Earnings} &= \underset{(-2.66)}{-31451} + \underset{(1.32)}{20219} \times Gender + \underset{(7.31)}{6921} \times Education \\ &\quad - \underset{(-2.37)}{2765} \times Gender \times Education, \ R^2 = .140. \end{aligned}$$

- To test whether gender is statistically significant need a joint test
  - $H_0 : \beta_{gender} = 0, \ \beta_{genderxeducation} = 0$ versus H_a : at least one $\neq 0$
  - here $F = 31.92$ with $p = 0.000$ so statistically significant at 5%.

Indicator Variable is a Dependent Variable

- For example model employment decision
  - $y = 1$ if person works and $y = 0$ if does not work.
- Can still do OLS
  - but use heteroskedastic-robust standard errors.
- It is better to use models specific to such data
  - logit model or probit model.

14.4 Regression with Sets of Indicator Variables

- A set of indicator variables is **mutually exclusive** if any individual in the sample falls into exactly one of the categories.
  - then for any individual observation only one indicator variables takes value 1
  - while the remaining indicator variables take value 0
  - so the indicator variables sum to one: $d1 + d2 + d3 = 1$.
- Indicators could be formed from categorical data that is
  - unordered: such as blue, red or orange
  - ordered: such as small, medium, large.

Example: Type of Worker

- Type of worker that has three categories – self-employed, employed in the private sector and employed in the government sector.
- Then three mutually exclusive indicator variables are defined as

$$
\begin{aligned}
d1 &= \begin{cases} 1 & \text{if self-employed} \\ 0 & \text{otherwise,} \end{cases} \\
d2 &= \begin{cases} 1 & \text{if employed in private sector} \\ 0 & \text{otherwise.} \end{cases} \\
d3 &= \begin{cases} 1 & \text{if employed in government sector} \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}
$$

Dummy Variable Trap

- Not all three indicators and an intercept can be included in the regression
    - erroneous inclusion of all is called the **dummy variable trap.**
- Since $d1 + d2 + d3 = 1$ we have $d1 = 1 - d2 - d3$, so

$$\begin{aligned} y &= \beta_1 + \beta_2 x + \alpha_1 d1 + \alpha_2 d2 + \alpha_3 d3 + u \\ &= \beta_1 + \beta_2 x + \alpha_1 (1 - d2 - d3) + \alpha_2 d2 + \alpha_3 d3 + u \\ &= (\beta_1 + \alpha_1) + \beta_2 x + (\alpha_2 - \alpha_1) d2 + (\alpha_3 - \alpha_1) d3 + u. \end{aligned}$$

- We can only identify four coefficients (of intercept, x, $d2$, $d3$)
    - but have five parameters to estimate $(\beta_1,\beta_2,\alpha_1,\alpha_2$ and $\alpha_2)$.
- Solution: drop one of the indicator variables or the intercept.

Base Category

- In current example $d1$ is dropped
  - coefficient $(\alpha_2 - \alpha_1)$ of $d2$ measures difference between earnings for a private sector worker $(d2 = 1)$ and a self-employed worker $(d1 = 1)$ after controlling for the other regressors.
- Suppose a categorical variable has C categories
  - Form a set of C mutually exclusive indicator variables d1, d2,..., dC.
  - To avoid the dummy variable trap drop one of the indicator variables
    - ★ called the **omitted category** or **base category**.
- The coefficient of an included indicator variable measures the marginal effect of being in that category **compared to the base category**, after controlling for the other regressors.

Hypothesis Testing

- Care is needed in interpreting hypothesis tests.
  - e.g. when $d1$ is the omitted category the coefficient of $d2$ measures $(\alpha_2 - \alpha_1)$, so a test of statistical significance of $d2$ is a test of $H_0 : \alpha_2 = \alpha_1$ against $H_a : \alpha_2 \neq \alpha_1$
    - ★ it is not a test of $H_0 : \alpha_2 = 0$ against $H_a : \alpha_2 \neq 0$.
- A t test of the statistical significance of a single indicator variable tests whether the ME of that category differs from that for the base category.
  - It is not a test of whether the ME effect of that category is zero.
- An F test of the joint statistical significance of the C-1 included indicator variables tests whether the set of indicator variables is statistically significant.
  - This joint F test leads to the same result regardless of the category that is dropped.

Example: Earnings and Type of Worker

- Regress earnings on all 3 categorical variables for type of worker and excluding the constant;

$$\widehat{y} = \underset{(9636)}{72306}d1 + \underset{(1897)}{54521}d2 - \underset{(2825)}{56105}d3 \qquad R^2 = 0.550,$$

  where heteroskedastic-robust standard errors are given in parentheses.

- For OLS estimation with all three mutually exclusive categories
  - the coefficients are just the sample averages for each category
  - e.g. average earnings for the self-employed are $72,306 with a standard error of $9,636.

Example: Earnings and Type of Worker

- Same results for F test and coefficients of *Age* and *Education* regardless of what is dropped (heteroskedastic-robust $t's$ in [ ])

Variable	**No**	**Drop**	**Drop**	**Drop**	**Drop**
	Indicators	**d1**	**d2**	**d3**	**intercept**
Age	525	488	488	488	488
	[3.47]	[3.26]	[3.26]	[3.26]	[3.26]
Education	5811	5865	5865	5865	5865
	[9.06]	[8.99]	[8.99]	[8.99]	[8.99]
d1 (self-employed)	-	-	17098	19123	−30151
	-		[1.83]	[1.99]	[-2.29]
d2 (private sector)	-	-17098	-	2025	-47249
	-	[-2.99]	-	[0.65]	[-4.15]
d3 (government sector)	-	-19123	-2025	-	-49274
	-	[-1.99]	[-0.65]	-	[-4.00]
Intercept	-46875	-30151	-47249	-49274	-
	[-4.40]	[-2.29]	[-4.15]	[-4.00]	-
F(2,n-k) for indicators	-	2.01	2.01	2.01	2.01
R^2	.115	.125	.125	.125	.601 (!)
Overall F	42.85	22.12	22.12	22.12	313.06

Difference in Means

- Test whether earnings vary across the type of worker, without inclusion of any controls.
- In areas of applied statistics that do not use regression
  - test using **analysis of variance (ANOVA) methods** that extend t test for difference in two means.
- Equivalently test by regress earnings on an intercept, $d2$ and $d3$

$$\widehat{y} = \underset{(7.50)}{72306} - \underset{(-1.81)}{17785}\, d2 - \underset{(-1.61)}{16201}\, d3 \qquad R^2 = 0.010,$$

where heteroskedastic-robust t statistics are given in parentheses.

- Then earnings
  - \$72,306 for self-employed workers, the omitted category
  - \$17,785 less than this for private sector workers
  - \$16,201 lower for government sector workers.
- F-statistic for joint statistical significance of $d2$ and $d3$ equals 1.68
  - since $p = 0.188$ there is not a statistically significant difference in earnings across the three types of workers at significance level 0.05.

Key Stata Commands

```
use AED_EARNINGS_COMPLETE.DTA, clear
regress earnings
mean earnings
sum earnings if gender == 1
sum earnings if gender == 0
regress earnings gender, vce(robust)
ttest earnings, by(gender) unequal
regress earnings education gender, vce(robust)
regress earnings education gender genderbyeduc, vce(robust)
regress price size bedrooms bathroom lotsize age monthsold,
vce(robust)
regress earnings dself dprivate dgovt, noconstant
vce(robust)
```

Some in-class Exercises

1. OLS regression using all data yields $\widehat{y} = 3 + 5d$. Give $\bar{y}$ for the subsample with $d = 0$ and $\bar{y}$ for the subsample with $d = 1$.
2. Suppose $\bar{y} = 30$ for the subsample with $d = 1$ and $\bar{y} = 20$ for the subsample with $d = 0$. Give the fitted model from OLS regression of y on an intercept and d using the full sample.
3. Suppose we have three mutually exclusive indicator variables $d1$, $d2$ and $d3$. OLS yields $\widehat{y} = 1 + 3d2 + 5d3$. What is the estimated difference between y those in category 2 $(d2 = 1)$ and those in category 1 $(d1 = 1)$.
4. For the preceding fitted model, give the coefficient estimates if instead we regressed y on an intercept, $d1$ and $d2$.