Additional Topics

Caio Vigo

The University of Kansas
Department of Economics

Fall 2018

These slides were based on *Introductory Econometrics* by Jeffrey M. Wooldridge (2015)

Topics

Describing Qualitative Information

- We have been studying variables (dependent and independent) with **quantitative** meaning.
- Now we need to study how to incorporate **qualitative** information in our framework (Multiple Regression Analysis).
- How to we describe binary qualitative information? Examples:
  - A person is either male or female. binary or dummy variable
  - A worker belongs to a union or does not. binary or dummy variable
  - A firm offers a 401(k) pension plan or it does not. binary or dummy variable
  - the race of an individual. multiple categories variable
  - the region where a firm is located (N, S, W, E). multiple categories variable

Describing Qualitative Information

- We will discuss only **binary variables**.

- **Binary variable** (or **dummy variable**) are also called a **zero-one** variable to emphasize the two values it takes on.

- Therefore, we must decide which outcome is assigned zero, which is one.

- Good practice: to choose the variable name to be descriptive.

- For example, to indicate gender, *female*, which is one if the person is female, zero if the person is male, is a better name than *gender* or *sex* (unclear what $gender = 1$ corresponds to).

Describing Qualitative Information

- Consider the following dataset:

```r
head(wage1_dummy)
```

```
##   wage    lwage educ exper tenure female married
## 1 3.10 1.131402   11     2      0      1       0
## 2 3.24 1.175573   12    22      2      1       1
## 3 3.00 1.098612   11     2      0      0       0
## 4 6.00 1.791759    8    44     28      0       1
## 5 5.30 1.667707   12     7      2      0       1
## 6 8.75 2.169054   16     9      8      0       1
```

```r
tail(wage1_dummy)
```

```
##      wage     lwage educ exper tenure female married
## 521  5.65 1.7316556   12     2      0      0       0
## 522 15.00 2.7080503   16    14      2      1       1
## 523  2.27 0.8197798   10     2      0      1       0
## 524  4.67 1.5411590   15    13     18      0       1
## 525 11.56 2.4475510   16     5      1      0       1
## 526  3.50 1.2527629   14     5      4      1       0
```

- For distinguishing different categories, any two different values would work.
Example: 5 or 6

- 0 and 1 make the interpretation in regression analysis much easier.

Topics

A Single Dummy Independent Variable

- What would it mean to specify a simple regression model where the explanatory variable is binary? Consider

$$wage = \beta_0 + \delta_0 female + u$$

where we assume SLR.4 holds:

$$E(u|female) = 0$$

- Therefore,

$$E(wage|female) = \beta_0 + \delta_0 female$$

A Single Dummy Independent Variable

- There are only two values of $female$, 0 and 1.

$$
\begin{aligned}
E(wage|female &= 0) = \beta_0 + \delta_0 \cdot 0 = \beta_0 \\
E(wage|female &= 1) = \beta_0 + \delta_0 \cdot 1 = \beta_0 + \delta_0
\end{aligned}
$$

In other words, the average $wage$ for men is β_0 and the average $wage$ for women is $\beta_0 + \delta_0$.

A Single Dummy Independent Variable

- We can write

$$\delta_0 = E(wage|female = 1) - E(wage|female = 0)$$

as the difference in average $wage$ between women and men.

- So δ_0 is not really a slope.

It is just a difference in average outcomes between the two groups.

A Single Dummy Independent Variable

- The population relationship is mimicked in the simple regression estimates.

$$\begin{aligned} \hat{\beta}_0 &= \overline{wage}_m \\ \hat{\beta}_0 + \hat{\delta}_0 &= \overline{wage}_f \\ \hat{\delta}_0 &= \overline{wage}_f - \overline{wage}_m \end{aligned}$$

where $\overline{wage}_m$ is the average wage for men in the sample and $\overline{wage}_f$ is the average wage for women in the sample.

A Single Dummy Independent Variable


```
## Total Observations in Table:  526
##
##
##            |         0 |         1 |
##            |-----------|-----------|
##            |       274 |       252 |
##            |     0.521 |     0.479 |
##            |-----------|-----------|
```

```r
stargazer(wage1_dummy, type='text')
```

```
## ================================================================
## Statistic  N   Mean  St. Dev.  Min   Pctl(25) Pctl(75)  Max
## ----------------------------------------------------------------
## wage      526 5.896   3.693   0.530   3.330    6.880   24.980
## lwage     526 1.623   0.532   -0.635  1.203    1.929   3.218
## educ      526 12.563  2.769     0      12       14       18
## exper     526 17.017  13.572    1      5        26       51
## tenure    526 5.105   7.224     0      0        7        44
## female    526 0.479   0.500     0      0        1        1
## married   526 0.608   0.489     0      0        1        1
## ----------------------------------------------------------------
```

A Single Dummy Independent Variable

	Dependent variable:
	wage
female	-2.512***
	(0.303)
Constant	7.099***
	(0.210)
Observations	526
R2	0.116
Adjusted R2	0.114
Residual Std. Error	3.476 (df = 524)
F Statistic	68.537*** (df = 1; 524)
Note:	*p<0.1; **p<0.05; ***p<0.01

A Single Dummy Independent Variable

- The estimated difference is very large. Women earn about \$2.51 less than men per hour, on average.

- Of course, there are some women who earn more than some men; this is a difference in averages.

A Single Dummy Independent Variable

• This simple regression allows us to do a simple **comparison of means test**. The null is

$$H_0 : \mu_f = \mu_m$$

where μ_f is the population average $wage$ for women and μ_m is the population average $wage$ for men.

• Under MLR.1 to MLR.5, we can use the usual t statistic as approximately valid (or exactly under MLR.6):

$$t_{female} = -8.28$$

which is a very strong rejection of H_0.

A Single Dummy Independent Variable

- The estimate $\hat{\delta}_0 = -2.51$ does not control for factors that should affect wage, such as workforce experience and schooling.

- If women have, on average, less education, that could explain the difference in average wages.

- If we just control for education, the model written in expected value form is

$$E(wage|female, educ) = \beta_0 + \delta_0 female + \beta_1 educ$$

where now δ_0 measures the gender difference when we hold fixed *exper*.

A Single Dummy Independent Variable

- Another way to write δ_0:

$$\delta_0 = E(wage|female, \textsf{educ}0) - E(wage|male, \textsf{educ}_0)$$

where $\textit{educ}er_0$ is any level of experience that is the same for the woman and man.

A Single Dummy Independent Variable

	Dependent variable:
	wage
female	-2.273***
	(0.279)
educ	0.506***
	(0.050)
Constant	0.623
	(0.673)
Observations	526
R2	0.259
Adjusted R2	0.256
Residual Std. Error	3.186 (df = 523)
F Statistic	91.315*** (df = 2; 523)
Note:	*p<0.1; **p<0.05; ***p<0.01

A Single Dummy Independent Variable

- Notice that there is still a difference of about $\$2.27$ (now it's smaller, but still large and statistically significant).

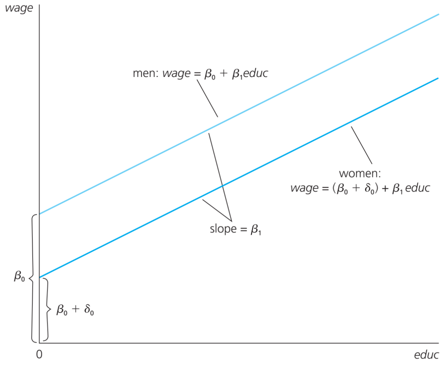
- The model imposes a common slope on *educ* for men and women, β_1, estimated to be $.506$ in this example.

- Recall, that the **intercept** is the only number that differ both categories (men and women).

- The estimated difference in average wages is the same at all levels of experience: $2.27.

A Single Dummy Independent Variable

Figure: Graph of $wage = \beta_0 + \delta_0 female + \beta_1 educ$ for $\delta_0 < 0$



A Single Dummy Independent Variable

- Notice that we can add other variables.

	Dependent variable:
	wage
female	-2.156***
	(0.270)
educ	0.603***
	(0.051)
exper	0.064***
	(0.010)
Constant	-1.734**
	(0.754)
Observations	526
R2	0.309
Adjusted R2	0.305
Residual Std. Error	3.078 (df = 522)
F Statistic	77.920*** (df = 3; 522)
Note:	*p<0.1; **p<0.05; ***p<0.01

- Note that if we also control for *exper*, the gap declines to $2.16 (still large and statistically significant).

A Single Dummy Independent Variable

- The previous regressions use males as the **base group** (or **benchmark group** or **reference group**). The coefficient -2.16 on $female$ tells us how women do compared with men.

- Of course, we get the same answer if we women as the base group, which means using a dummy variable for males rather than females.

- Because $male = 1 - female$, the coefficient on the dummy changes sign but must remain the same magnitude.

- The intercept changes because now the base (or reference) group is females.

A Single Dummy Independent Variable

- Putting $female$ and $male$ both in the equation is redundant. We have two groups so need only two intercepts.

- This is the simplest example of the so-called **dummy variable trap**, which results from putting in too many dummy variables to represent the given number of groups (two in this case).

- Because an intercept is estimated for the base group, we need only one dummy variable that distinguishes the two groups.