Additional Topics - Dummy Variables, Adjusted R-Squared & Heteroskedasticity

Caio Vigo

The University of Kansas
Department of Economics

Summer 2019

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 do 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
```

Describing Qualitative Information

- 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 |
##           |-----------|-----------|
```

```
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

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

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

A Single Dummy Independent Variable

- Another way to write δ_0:

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

where *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).

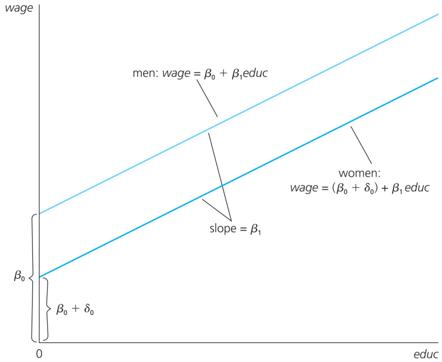
- The model imposes a common slope on *educ* for men and women, β_1, estimated to be $.506$ in this example.

- 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.

Interpreting Dummy Variable Coefficients with $\log(y)$ as the Dependent Variable

- Consider the following regression:

$$log(y) = \beta_0 + \beta_1 x_{dummy} + \beta_2 x_2 + u$$

- When $log(y)$ is the dependent variable in a model, the coefficient on a dummy variable, when multiplied by 100, is interpreted as the percentage difference in y, holding all other factors fixed.

Interpreting Dummy Variable Coefficients with $\log(y)$ as the Dependent Variable

- When the coefficient on a dummy variable suggests a large proportionate change in y, the exact percentage difference can be obtained exactly as with the semi-elasticity calculation.

Recall,

Model	Dependent Variable	Independent Variable	Interpretation of β_1
Level-Level	y	x	$\Delta y = \beta_1 \Delta x$
Level-Log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
Log-Level	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
Log-Log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

KU

Interpreting Coefficients on Dummy Explanatory Variables when the Dependent Variable is $\log(y)$

	Dependent variable:
	lwage
female	-0.397***
	(0.043)
Constant	1.814***
	(0.030)
Observations	526
R2	0.140
Adjusted R2	0.138
Residual Std. Error	0.494 (df = 524)
F Statistic	85.044*** (df = 1; 524)
Note:	*p<0.1; **p<0.05; ***p<0.01

Interpreting Coefficients on Dummy Explanatory Variables when the Dependent Variable is $\log(y)$

$$\begin{aligned} \widehat{lwage} &= \underset{(.030)}{1.814} - \underset{(.043)}{.397}\, female \\ n &= 526,\ R^2 = .138 \end{aligned}$$

- A rough estimate is that in the population of working, high school graduates, the average wage for women is below that of men by 39.7%.

Interpreting Dummy Variable Coefficients with $\log(y)$ as the Dependent Variable

- Thus, for the following regression:

$$log(y) = \beta_0 + \beta_1 x_{dummy} + \beta_2 x_2 + u$$

for the dummy variable x_{dummy}, the exact percentage difference in the predicted y when $x_{dummy} = 1$ versus when $x_{dummy} = 0$ is:

$$100 \cdot [exp(\hat{\beta}_1) - 1]$$

Interpreting Coefficients on Dummy Explanatory Variables when the Dependent Variable is $\log(y)$


```
===============================================
                       Dependent variable:
                    ---------------------------
                               lwage
-----------------------------------------------
female                       -0.397***
                              (0.043)

Constant                      1.814***
                              (0.030)

-----------------------------------------------
Observations                    526
R2                             0.140
Adjusted R2                    0.138
Residual Std. Error       0.494 (df = 524)
F Statistic           85.044*** (df = 1; 524)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

Interpreting Coefficients on Dummy Explanatory Variables when the Dependent Variable is $\log(y)$

Exact Percentage Difference

Using,

- **Men as the base (reference) group:**,

precise estimate in wage difference: $\exp(-.397) - 1 \approx -.328$, or 32.8% lower for women.

- **Women as the base (reference) group:**,

precise estimate in wage difference: $\exp(.397) - 1 \approx -.487$, or 48.7% higher for men.

Interpreting Coefficients on Dummy Explanatory Variables when the Dependent Variable is $\log(y)$

	Dependent variable: lwage
female	-0.361***
	(0.039)
educ	0.077***
	(0.007)
Constant	0.826***
	(0.094)
Observations	526
R2	0.300
Adjusted R2	0.298
Residual Std. Error	0.445 (df = 523)
F Statistic	112.189*** (df = 2; 523)
Note:	*p<0.1; **p<0.05; ***p<0.01

	Dependent variable: lwage
female	-0.344***
	(0.038)
educ	0.091***
	(0.007)
exper	0.009***
	(0.001)
Constant	0.481***
	(0.105)
Observations	526
R2	0.353
Adjusted R2	0.349
Residual Std. Error	0.429 (df = 522)
F Statistic	94.747*** (df = 3; 522)
Note:	*p<0.1; **p<0.05; ***p<0.01

Interpreting Coefficients on Dummy Explanatory Variables when the Dependent Variable is $\log(y)$

- The gap shrinks, but is still substantial.

- If we control for workforce experience and education, the difference is approximately 34.4% lower for women. The precise estimate in wage difference: $\exp(-.344) - 1 \approx -.291$, or 29.1% lower for women.

- That is, at any given levels of experience and education, a woman is predicted to earn about 29% less than a man.

Dummy Variables for Multiple Categories

- Suppose in the wage example we have two qualitative variables, gender and marital status. Call these $female$ and $married$.

- We can define four exhaustive and mutually exclusive groups. These are married males $(marrmale)$, married females $(marrfem)$, single males (*singmale*), and single females (*singfem*).

- Note that we can define each of these dummy variables in terms of $female$ and $married$:

Dummy Variables for Multiple Categories

$$
\begin{aligned}
marrmale &= married \cdot (1 - female) \\
marrfem &= married \cdot female \\
singmale &= (1 - married) \cdot (1 - female) \\
singfem &= (1 - married) \cdot female
\end{aligned}
$$

Dummy Variables for Multiple Categories

- We can allow each of the four groups to have a different intercept by choosing a base group and then including dummies for the other three groups.

- So, if we choose single males as the base group, we include $marrmale$, $marrfem$, and $singfem$ in the regression. The coefficients on these variabels are relative to single men.

- With $lwage$ as the dependent variable, we can give them a percentage change interpretation.

Interpreting Coefficients on Dummy Explanatory Variables when the Dependent Variable is $\log(y)$

	Dependent variable:
	lwage
marrmale	0.292***
	(0.055)
marrfem	-0.120**
	(0.058)
singfem	-0.097*
	(0.057)
educ	0.084***
	(0.007)
exper	0.003*
	(0.002)
tenure	0.016***
	(0.003)
Constant	0.388***
	(0.102)
Observations	526
R2	0.424
Adjusted R2	0.417
Residual Std. Error	0.406 (df = 519)
F Statistic	63.626*** (df = 6; 519)
Note:	*p<0.1; **p<0.05; ***p<0.01

Dummy Variables for Multiple Categories

- Using the usual approximation based on differences in logarithms – and holding fixed education, experience, and tenure – a married man is estimated to earn about 29.2% more than a single man.

- Remember, this compares two men with the same level of schooling, general workforce experience, and tenure with the current employer.

Interpreting Coefficients on Dummy Explanatory Variables when the Dependent Variable is $\log(y)$

- What if we want to compare married women and single women? Just plug in the correct set of zeros and ones.

$$\begin{aligned}
\text{intercept for married women} &= .388 - .120 \\
\text{intercept for single women} &= .388 - .097 \\
\text{difference} &= -0.268 - (-0.291) = -.023
\end{aligned}$$

so married women earn about 2.3% less than single women (controlling for other factors).

- We cannot tell from the previous output whether this difference is statistically significant.
- Note how the intercept for single men gets differenced away.

KU

Topics

Adjusted R-Squared

Recall that,

- How do we decide whether to include a single new independent variable: t **test**.
- How do we decide whether to include a group of new variables: F **test**.

Adjusted R-Squared

Motivation: R^2 can never go down (usually increases) when one or more variables is added to a regression.

- We use the **adjusted R-squared** to compare across models that have different numbers of explanatory variables but where one is not a special case of the other (nonnested models).

- The **adjusted R-squared** imposes a penalty for adding additional explanatory variables.

Adjusted R-Squared

- As usual, start with

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u$$

- Now we need to be more careful with variance labels:

$$\begin{aligned} \sigma_y^2 &= Var(y) \\ \sigma_u^2 &= Var(u) \end{aligned}$$

Define

$$\rho^2 = 1 - \frac{\sigma_u^2}{\sigma_y^2}$$

This is the **population** ***R*-squared** – the amount of population variation in y explained by $x_1, ..., x_k$.

Adjusted R-Squared

- The formula for the R^2 can be written as

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{(SSR/n)}{(SST/n)},$$

which shows we can think of R^2 as using SSR/n to estimate σ_u^2 and SST/n to estimate σ_y^2. These are consistent but not unbiased estimators.

- Instead, use

$$SSR/(n-k-1)$$
$$SST/(n-1)$$

as the unbiased estimators.

Adjusted R-Squared

- Plugging in gives the **adjusted *R*-squared**, also called "R-bar-squared":

$$\begin{aligned} \bar{R}^2 &= 1 - \frac{[SSR/(n-k-1)]}{[SST/(n-1)]} \\ &= 1 - \frac{\hat{\sigma}^2}{[SST/(n-1)]} \end{aligned}$$

where $\hat{\sigma}^2$ is the usual variance parameter estimator.

- $\bar{R}^2$ **imposes a penalty:** When more regressors are added, SSR falls, but so does $df = n - k - 1$. $\bar{R}^2$ can increase or decrease.
- For $k \geq 1$, $\bar{R}^2 < R^2$ unless $SSR = 0$ (not an interesting case).
- It is possible that $\bar{R}^2 < 0$, especially if df is small. Remember that $R^2 \geq 0$ always.

Adjusted R-Squared

Algebraic Facts:

1. If a single variable is added to a regression, $\bar{R}^2$ increases if and only if the absolute t statistic of the new variable is greater than one.

2. If two or more variables are added to a regression, $\bar{R}^2$ increases if and only if the F statistic for joint significance of the new variables is greater than one.

- **Important:** In the R-squared form of the F statistic that we covered, it is the usual R-squared, not the adjusted R-squared, that appears.

- Sometimes $\bar{R}^2$ is called the "corrected R-squared".

Topics

Heteroskedasticity

- Recall the five **Gauss-Markov** Assumptions for OLS regression:

Gauss-Markov Assumptions

MLR.1: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$
MLR.2: random sampling from the population
MLR.3: no perfect collinearity in the sample
MLR.4: $E(u|x_1, ..., x_k) = E(u) = 0$ (exogenous explanatory variables)
MLR.5: $Var(u|x_1, ..., x_k) = Var(u) = \sigma^2$ (homoskedasticity)

KU

Heteroskedasticity

- Under these five assumptions, OLS has lots of nice properties.
  - OLS is BLUE.
  - OLS is (asymptotically) efficient

Consequences of adding/removing assumption MLR.6

- With normality **(MLR.6)**, the tests and confidence intervals are exact given any sample size.

- Without normality **(MLR.6)**, the usual OLS test statistics and CIs are only asymptotically justified $\Rightarrow$ you need to have a large sample to use them.

Heteroskedasticity

Consequences of adding/removing assumption MLR.5

- If we do not impose or assume homoskedastic errors, i.e., if we drop **MLR.5** and act as if we know nothing about $Var(u|x_1, ..., x_k) = ?$
- Since, **heteroskedasticity** does not cause bias in the $\hat{\beta}_j$, OLS is still unbiased under **MLR.1** to **MLR.4**.
- OLS is no longer **BLUE**.
- It is possible to find **unbiased estimators** that have smaller variances than the OLS estimators.
- **Important:** standard errors are no longer valid.

Heteroskedasticity

- This means the t statistics and confidence intervals that use these standard errors cannot be trusted.

- This is true even in large samples.

- Joint hypotheses tests using the usual F statistic are no longer valid in the presence of heteroskedasticity.

Heteroskedasticity

- Standard errors and all test statistics can be modified to be valid in the presence of **heteroskedasticity of unknown form**.

Heteroskedasticity-Robust Standard Errors

- We need to compute **heteroskedasticity-robust standard errors**.
  - Which produces **heteroskedasticity-robust** t **statistics** and **heteroskedasticity-robust confidence intervals**.
  - The **heteroskedasticity-robust** test statistics and CIs only have asymptotic justification, even if the full set of CLM assumptions hold.
  - With smaller sample sizes, the **heteroskedasticity-robust** statistics need not be well behaved.

Example:

$$
\begin{array}{rcrcrlrlrl}
\widehat{lwage} & = & 1.6492 & - & .2202 & female + & .0521 & exper + & .0762 & coll \\
 & & (.0720) & & (.0318) & & (.0058) & & (.0066) & \\
 & & [.0754] & & [.0325] & & [.0060] & & [.0068] & \\
n & = & \multicolumn{8}{l}{750,\ R^2 = .302,\ \bar{R}^2 = .299}
\end{array}
$$

- The robust statistics are virtually always different from the usual statistics, regardless of which set of assumptions holds in the population.

- **In this example:** The robust standard errors (between square brackets) are all slightly larger than the usual standard errors.

- **In this example:** CIs are slightly wider, t statistics slightly lower.

Heteroskedasticity

Tests of Heteroskedasticity:

Assuming **MLR.1** to **MLR.4** holds:

- **Breusch-Pagan test for heteroskedasticity**
- **White test for heteroskedasticity**

Steps in Computing the Breusch-Pagan (and White) Test

1. Estimate the equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$ by OLS, saving the OLS residuals, $\hat{u}_i$.

2. Compute the squared residuals, $\hat{u}_i^2$.

3. Regress $\hat{u}_i^2$ on all explanatory variables (**for White:** ... on all explanatory variables and also the nonredundant squares and interactions of all explanatory variables) and compute the usual F test of joint significance of the explanatory variables.

4. If the p-value of the test is sufficiently small, reject the null of homoskedasticity and conclude that the homoskedasticity assumption **(MLR.5)** fails.