Multiple Regression with Qualitative Information

ECONOMETRICS (ECON 360)

BEN VAN KAMMEN, PHD

Introduction

There is a lot of (relevant) information in data about the elements observed that is not in quantitative form.

This chapter explores how that information can be used to create variables that can be used in a regression.

These methods are powerful because without them one would have to confine his methods to explicitly quantitative variables like age, income, years of schooling, high school GPA, et al.

Outline

Describing Qualitative Information.

Regression with a Single Binary Variable

Using Binary Variables for Multiple Categories.

Interactions Involving Binary Variables.

- Allowing for Different Slopes.

A Binary Dependent Variable: the Linear Probability Model.

Policy Analysis and Program Evaluation.

Interpreting Regression Results with Discrete Dependent Variables.

Describing qualitative information

Qualitative information can be turned into quantitative information in a straightforward way, using binary coding for “yes” and “no”.

For example “is a certain person in the sample female?”

$$\text{yes } female \rightarrow x = 1 \text{ and no not } female \rightarrow x = 0.$$

According to the above example, the variable *x* can be called a binary variable for whether or not each observation is *female*.

- Synonymous terms you will often hear for it are indicator variable, zero-one variable, or (regrettably) dummy variable.

Indicator variables

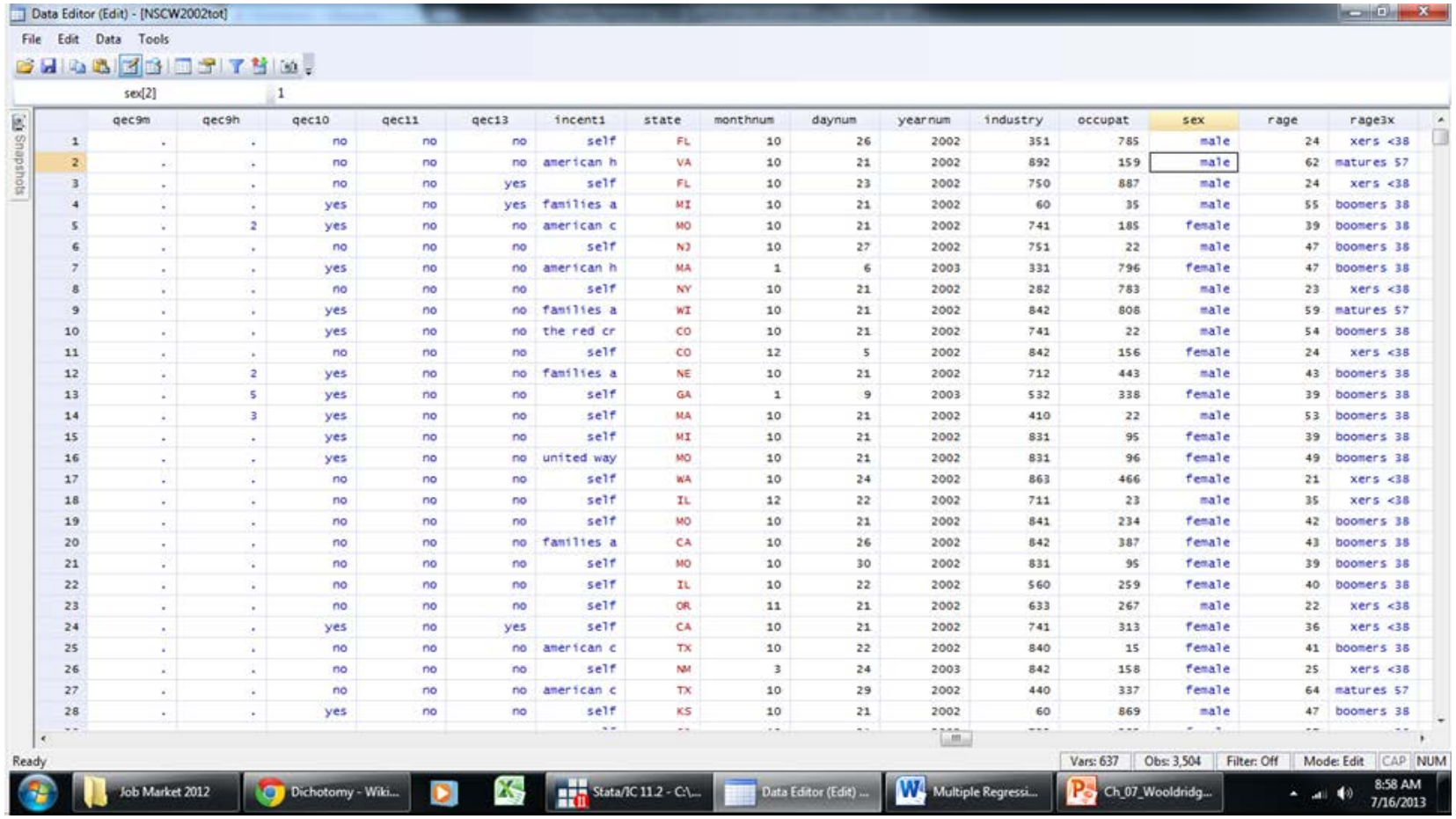
It is fairly simple to assign zeroes and ones to observations, based on dichotomous gender.

For the sake of clarity, though, it is vital to name the variable according to whether $female = 1$ or $female = 0$. The interpretation of the variable (and its estimated regression coefficient) depends on it; call the variable "female" if $female = 1$ and call it "male" if $female = 0$.

Sometimes you will find data with indicator variables already generated.

- Sometimes it will be purely in "string" format, i.e., a column of cells containing the words "male" or "female".
- Sometimes it comes in numerical format that is not binary, to allow for other information, such as instances in which the survey respondent did not answer the question.
- The data set on the next slide (shown in STATA Data Editor view) illustrates these possibilities.

Qualitative information example



Qualitative information example (continued)

state is a string variable; *sex* is a binary variable that has value labels that decode the numbers into words (the blue font).

- A “male” cell is selected, and the formula bar says that the value is “1”.
- For a “female” cell the value would be “2”, so this data does not have a (0,1) indicator for *sex* yet. To generate one in STATA one would merely use the syntax:

quietly tabulate [categorical var], generate(name of indicator).

In this example, it would look like:

quietly tabulate sex, gen(sex01_)

from which STATA would generate 2 new variables, “sex01_1” and “sex01_2”.

- Then rename them, “male” and “female” using STATA’s **rename** command, e.g., **rename sex01_1 male**.

Indicator variables (continued)

Often in data, qualitative information can take more than 2 possible "values," e.g., a sample of Midwesterners may report their *state of residence* as: Wisconsin, Minnesota, Illinois, Iowa, Indiana, Ohio, or Michigan.

Generating indicator variables for *state* will result in one new variable per value, i.e., 7 for the Midwest.

- It would be 50 if you had the whole U.S. (excluding Washington, D.C., and the territories).
- Tabulating the variable "race3" in this data would result in 3 indicators: "white", "black" and "other".

Qualitative information example (2)

	incent1	state	monthnum	daynum	yearnum	industry	occupat	sex	rage	rage3x	rage4	rage6	race3	race2	hispanic
1	self	FL	10	26	2002	351	785	male	24	xers <38	<30 yrs	18-24	white	white	no
2	american h	VA	10	21	2002	892	159	male	62	matures 57	50+ yrs	55-64	white	white	no
3	self	FL	10	23	2002	750	887	male	24	xers <38	<30 yrs	18-24	black	other	no
4	families a	MI	10	21	2002	60	35	male	55	boomers 38	50+ yrs	55-64	white	white	no
5	american c	MO	10	21	2002	741	185	female	39	boomers 38	30-39 yrs	35-44	white	white	no
6	self	NJ	10	27	2002	751	22	male	47	boomers 38	40-49 yrs	45-54	white	white	no
7	american h	MA	1	6	2003	331	796	female	47	boomers 38	40-49 yrs	45-54	white	white	no
8	self	NY	10	21	2002	282	783	male	23	xers <38	<30 yrs	18-24	other	other	no
9	families a	WI	10	21	2002	842	808	male	59	matures 57	50+ yrs	55-64	white	white	no
10	the red cr	CO	10	21	2002	741	22	male	54	boomers 38	50+ yrs	45-54	other	other	yes
11	self	CO	12	5	2002	842	156	female	24	xers <38	<30 yrs	18-24	other	other	yes
12	families a	NE	10	21	2002	712	443	male	43	boomers 38	40-49 yrs	35-44	white	white	no
13	self	GA	1	9	2003	532	338	female	39	boomers 38	30-39 yrs	35-44	white	white	no
14	self	MA	10	21	2002	410	22	male	53	boomers 38	50+ yrs	45-54	white	white	no
15	self	MI	10	21	2002	831	95	female	39	boomers 38	30-39 yrs	35-44	white	white	no
16	united way	MO	10	21	2002	831	96	female	49	boomers 38	40-49 yrs	45-54	white	white	no
17	self	WA	10	24	2002	863	466	female	21	xers <38	<30 yrs	18-24	white	white	no
18	self	IL	12	22	2002	711	23	male	35	xers <38	30-39 yrs	35-44	white	white	no
19	self	MO	10	21	2002	841	234	female	42	boomers 38	40-49 yrs	35-44	white	white	no
20	families a	CA	10	26	2002	842	387	female	43	boomers 38	40-49 yrs	35-44	white	white	no
21	self	MO	10	30	2002	831	95	female	39	boomers 38	30-39 yrs	35-44	black	other	no
22	self	IL	10	22	2002	560	259	female	40	boomers 38	40-49 yrs	35-44	black	other	no
23	self	OR	11	21	2002	633	267	male	22	xers <38	<30 yrs	18-24	white	white	no
24	self	CA	10	21	2002	741	313	female	36	xers <38	30-39 yrs	35-44	white	white	no
25	american c	TX	10	22	2002	840	15	female	41	boomers 38	40-49 yrs	35-44	white	white	no
26	self	NM	3	24	2003	842	158	female	25	xers <38	<30 yrs	25-34	other	other	no
27	american c	TX	10	29	2002	440	337	female	64	matures 57	50+ yrs	55-64	white	white	no
28	self	KS	10	21	2002	60	869	male	47	boomers 38	40-49 yrs	45-54	white	white	no

Indicator variables (concluded)

For a given observation, only 1 of the indicators equals 1; all the others equalzero.

- Indicator variables break the qualitative information into mutually exclusive categories.
- Using “tabulate” to create indicator variables generalizes even to variables that have ordinal significance, such as a *schooling* variable that takes on values such as: “no H.S. diploma”, “H.S. diploma/GED”, “Some college”, “College degree”.
- But the interpretation is more difficult when the variable being tabulated has ordinal significance or takes on many values or both.

We will begin with a very straightforward case in which neither of those complications exists.

Regression with a single binary independent variable

Indicator variables can enter a regression model the same way as continuous (*x*) variables. E.g., the *female* indicator in a wage regression would be modeled:

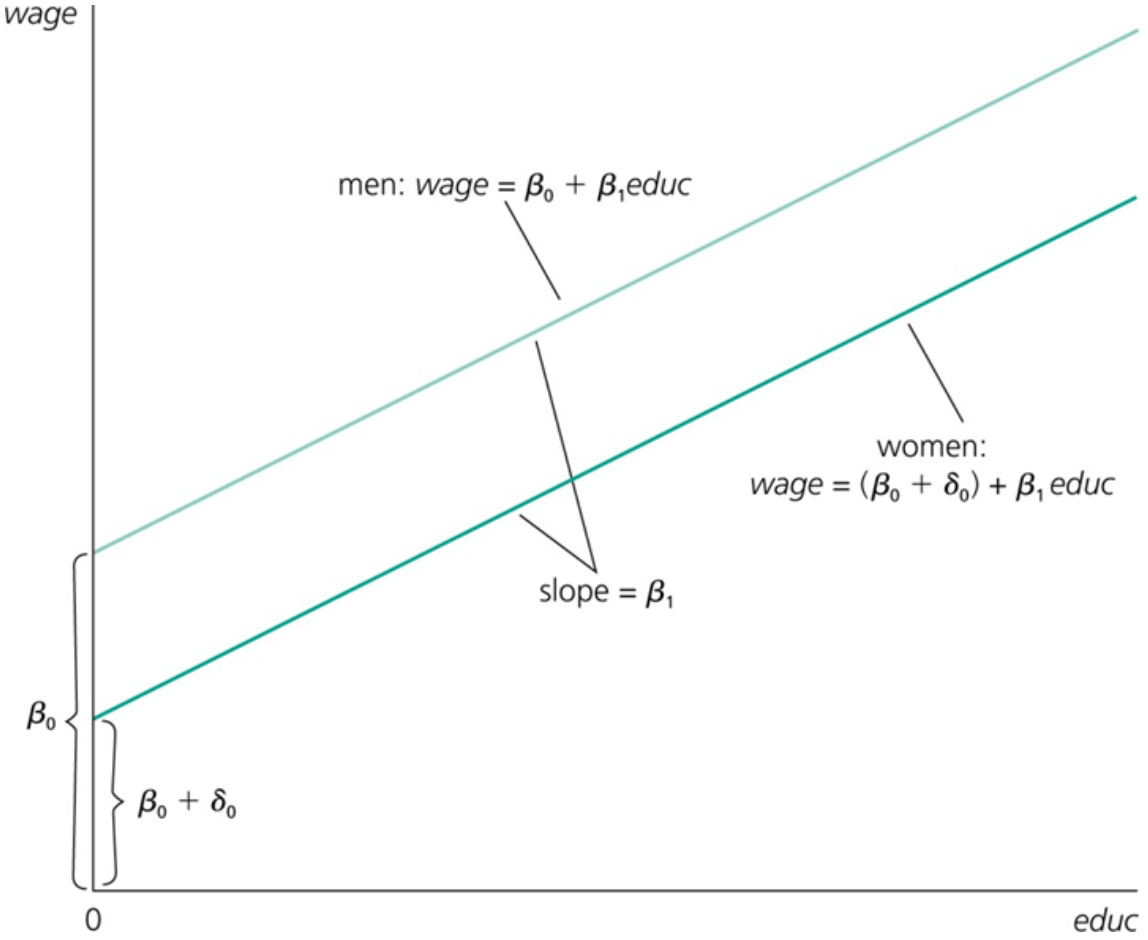
$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u.$$

The coefficient on *female* is interpreted as the difference in conditional expectations between when $female = 1$ (woman) and when $female = 0$ (man).

$$\delta_0 = E(wage|educ, female = 1) - E(wage|educ, female = 0) \text{ or}$$

$$\delta_0 = E(wage|educ, female) - E(wage|educ, male).$$

The effect of a binary indicator variable



A single binary independent variable (continued)

If $\delta_0 < 0$, the results show that, for a given level of education, women earn less than men in the sample.

The common coefficient on *education* for women and men restricts each group to having the same returns to additional schooling.

But a non-zero coefficient on *female* means that women have a different intercept than men.

1. The intercept for observations with $female = 0$ is merely β_0,
2. For observations with $female = 1$, the intercept shifts by δ_0,
3. The slope of the wage-schooling relationship does not change; for both sexes they are parallel.

A single binary independent variable (continued)

In this estimation, men are the base group, i.e., the group that maintains the generic intercept term, β_0.

- It would be redundant to include the indicator for "male" in the regression and attempt to estimate another parameter that β_0 already estimates.

Furthermore it would be impossible because female and male are perfectly collinear, as a result of being mutually exclusive and exhaustive.

$$male + female = 1 \Leftrightarrow male = 1 - female,$$

means *male* is a perfect linear function of *female* and has no independent variation in the sample with which to estimate its coefficient.

A single binary independent variable (continued)

Including additional non-indicator variables in the regression does not alter the above interpretation of the indicator coefficient.

The practical significance of δ_0 is measuring whether comparably productive (!) men and women earn the same wages, or whether discrimination (or something else?) could contribute to wage disparity.

This could be inferred from the test of statistical significance:

$$H_0\colon\ \delta_0 = 0\ (both\ paid\ the\ same), H_1\colon\ \delta_0 \neq 0.$$

Aside: interpreting the gender indicator's coefficient

As with non-binary regressors, an empiricist must ask whether Assumption MLR.4 is realistic: "is the binary variable of interest correlated with the error term?"

In the example of gender in the labor market, the models in the text do not condition on the attributes of the jobs chosen by each gender.

- Compensating Differentials theory predicts that workers are paid more for working jobs that involve disamenities such as unpleasant (noisy, dirty, et al.) conditions, injury risk, or strenuous schedules.
- If men are more likely to select into unpleasant jobs than comparably-skilled women, their higher wages reflect payments for tolerating disamenities, rather than discrimination.

A single binary independent variable (concluded)

Similar considerations should be made in the other textbook examples, as well as any instance in which the assignment to groups is the result of an agent's choice, e.g., owning a personal computer (example 7.2) and participating in a job training program (7.3).

The responsible empiricist ought to ask whether "more ambitious students are more likely to voluntarily purchase a PC than less ambitious students?" and whether "a firm is more likely to pursue a training subsidy if it was already planning to perform a lot of worker training?"

- A PC is probably a valuable input to success in college, and incentivizing training with a grant is likely to cause more training, but when the selection is non-random, it is impossible to know how much of the estimated effect is causal and how much is from the self-selection bias.

The promise of OLS in answering questions like these (so called program evaluations) lies in controlling for enough other factors that the estimated effect ($\hat{\delta}_0$) of program participation (participation=1) can be interpreted as evidence of a causal effect.

Interpreting coefficients on binary variables when the LHS is log(y)

As with non-binary regressors, the coefficient on a binary variable has a percentage change interpretation.

For the example of the *female* indicator in the wage regression, the results (next slide) from estimating [7.9] show a coefficient estimate of $\hat{\delta}_0 = -0.2965$.

Using the approximation to % change, this is a 29.65% wage penalty forwomen.

- Since it is a fairly large change, however, the approximation is likely inappropriate. The % penalty from a discrete change of one unit (0 to 1) in the *female* indicator is:

$$\%\Delta lwage = 100[\exp(-0.296511) - 1] = -25.66\%.$$

Effects when the LHS is log(*y*) example

```
. reg lwage female educ c.exper##c.exper c.tenure##c.tenure
```

Source	SS	df	MS
Model	65.3791009	6	10.8965168
Residual	82.9506505	519	.159827843
Total	148.329751	525	.28253286

Number of obs	= 526
F(6, 519)	= 68.18
Prob > F	= 0.0000
R-squared	= 0.4408
Adj R-squared	= 0.4343
Root MSE	= .39978

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-.296511	.0358055	-8.28	0.000	-.3668524	-.2261696
educ	.0801967	.0067573	11.87	0.000	.0669217	.0934716
exper	.0294324	.0049752	5.92	0.000	.0196585	.0392063
c.exper# c.exper	-.0005827	.0001073	-5.43	0.000	-.0007935	-.0003719
tenure	.0317139	.0068452	4.63	0.000	.0182663	.0451616
c.tenure# c.tenure	-.0005852	.0002347	-2.49	0.013	-.0010463	-.0001241
_cons	.416691	.0989279	4.21	0.000	.2223425	.6110394

Using binary variables for multiple categories

It is not significantly more challenging to include multiple sets of indicator variables in the same regression model.

- On the subject of discrimination, a researcher may also be interested in whether white workers are paid a premium compared to similar (in terms of education, experience, tenure) non-white workers.

An indicator for *nonwhite* is in the data, and it can be added individually to the regression.

It can also be interacted with another indicator, e.g., *female*.

The model would become:

$$lwage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 tenure + \beta_5 tenure^2 + \beta_6 female + \beta_7 nonwhite + \beta_8 female * nonwhite + u.$$

Binary variables for multiple categories (continued)

Now it is possible to estimate 4 different intercepts for each combination of *female* and *nonwhite*, with white males being the base group (β_0).

- The intercept for white females is ($\beta_0 + \beta_6$);
- for non-white males it is ($\beta_0 + \beta_7$), and
- for non-white females it is ($\beta_0 + \beta_6 + \beta_7 + \beta_8$).

Binary variables for multiple categories (continued)

This isn't exactly how the concept is presented in Wooldridge, but it is equivalent to generating indicators for each combination of *female* and *nonwhite* and using any 3 (=4-1) of them in the OLS estimation.

- Even though the interaction of indicators isn't covered until Chapter 7.4, it's easier to explain this as a special case of an interaction instead of thinking about exotically-defined groups.

This is a convenient opportunity to explain how to tell STATA to regress using indicator variables—without generating the indicators in the data set—using factor variables (**fvvarlist**).

To perform the regression described above, the syntax (Wooldridge data file **wage1.dta**) is:

reg lwage i.female##i.nonwhite educ c.exper##c.exper c.tenure##c.tenure.

Interaction of indicator variables example

```
. reg lwage i.female##i.nonwhite educ c.exper##c.exper c.tenure##c.tenure
```

Source	SS	df	MS
Model	65.7314291	8	8.21642864
Residual	82.5983223	517	.159764647
Total	148.329751	525	.28253286

Number of obs	526
F(8, 517)	51.43
Prob > F	0.0000
R-squared	0.4431
Adj R-squared	0.4345
Root MSE	.39971

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
1.female	-.2800874	.0377878	-7.41	0.000	-.3543239	-.2058509
1.nonwhite	.0439158	.0792357	0.55	0.580	-.1117477	.1995793
female# nonwhite						
1 1	-.1607668	.1162072	-1.38	0.167	-.3890631	.0675296
educ	.0804544	.0067983	11.83	0.000	.0670988	.0938101
exper	.0302209	.005005	6.04	0.000	.0203882	.0400536
c.exper# c.exper	-.0005998	.0001079	-5.56	0.000	-.0008118	-.0003878
tenure	.0316791	.0068528	4.62	0.000	.0182163	.0451419
c.tenure# c.tenure	-.0005872	.0002349	-2.50	0.013	-.0010487	-.0001257
_cons	.4037444	.1008726	4.00	0.000	.2055739	.6019149

Interaction of indicator variables example (continued)

The regression output implies estimated wage penalties (compared to white men) of:

$$100[\exp(-0.2800874) - 1] = -24.43\% \text{ for white women,}$$
$$100[\exp(-0.39693838) - 1] = -32.69\% \text{ for non} - \text{white women.}$$

The comparison between white and non-white men is not statistically significant in this sample—and opposite in sign compared to the prediction of discrimination.

Neither is there a statistically significant difference ($|t| = 1.38$) between white and non-white women, but it does run in the predicted direction.

Binary variables for multiple categories (concluded)

The emphasis, in terms of STATA programming, is on the use of the "i." in the factor variables coding, in contrast to the "c." used for interactions between continuous variables.

"i." interacts combinations of values of the variables instead of simply interacting the variables themselves (*exper* and *tenure* in the present example).

Use of "i." will automatically drop one category as the base. In this case, it dropped the "right" one as specified in our model.

- If it drops the "wrong" one, you may specify the base category by adding "**b#**" to the prefix, where # is the category you want to be the base.

Incorporating ordinal information by using binary variables

The variables examined so far have been categorical (merely classifying by group) and has not had ordinal significance, i.e., meaningful order.

- The ratio scale variables used so far have meaningful order, but they also have a meaningful zero value (*educ*=0 implies zero years of education) and differences (12 years is 1 more than 11 years) are meaningful.

In this discussion, the focus is on variables like credit (bond) *rating*, subjective well-being, and the so-called "temperature" scales of approval/disapproval.

- They are ordinal in the sense that a higher rating indicates "more" of something (creditworthiness, well-being, approval), but a one unit increase is difficult to interpret and is unlikely to be uniformly meaningful across the scale of measurement.
- E.g., is the difference between "strongly disapprove" and "disapprove" the same as the difference between "disapprove" and "neither approve nor disapprove"?

Ordinal information (continued)

To ameliorate this shortcoming, indicator variables for each level (excluding one as the base) of an ordinal variable are added to the regression model to capture the non-constant effects of each change.

For concreteness a survey may ask respondents to approve or disapprove of the Federal Reserve's monetary policy according to the scale:

$$FRa = \begin{cases} 5, & strongly\ approve \\ 4, & approve \\ 3, & neither\ disapprove\ nor\ approve \\ 2, & disapprove \\ 1, & strongly\ disapprove \end{cases}$$

Ordinal information (continued)

A regression of the respondents' investment behavior on Fed approval level (and other not explicitly listed factors) would appear as:

$$invest = \beta_0 + \delta_1 FRa1 + \delta_2 FRa2 + \delta_3 FRa3 + \delta_4 FRa4 + other\ factors,$$

where "strongly approve" is the omitted (base) level.

The coefficients are simple to interpret.

- δ_4 is the effect on investment of diminishing one's approval of the Fed from strong to regular.
- δ_3 captures the effect of diminishing it to the point of indifference, and so on.

Ordinal information (continued)

A special case of this method emerges when you reconstruct the indicator variables into the ordinal variable *FRa*.

$$FRa = FRa1 + 2FRa2 + 3FRa3 + 4FRa4 + 5FRa5.$$

Estimating the model with *FRa* as ratio scale,

$$invest = \beta_0 + \gamma_1 FRa + other\ factors,$$

is tantamount to placing restrictions on the coefficients in the more flexible model with indicators.

- The restrictions are: $\delta_2 = -2\gamma_1, \delta_3 = -3\gamma_1$, and so forth.

As usual the validity of those restrictions can be tested using the appropriate *F* statistic.

Ordinal information (concluded)

Occasionally an ordinal variable takes on too many values to include indicators for each one.

When this occurs, they can be broken down into a smaller number of groups that are larger in size.

E.g., if approval is rated on a 1-100 scale, the ratings can be broken down into groups of 10 instead of putting in 99 indicators.

Interactions of binary and ratio variables

Now you have encountered interactions between two ratio scale variables (think*education* and *experience*), two indicator variables (*female* and *nonwhite*).

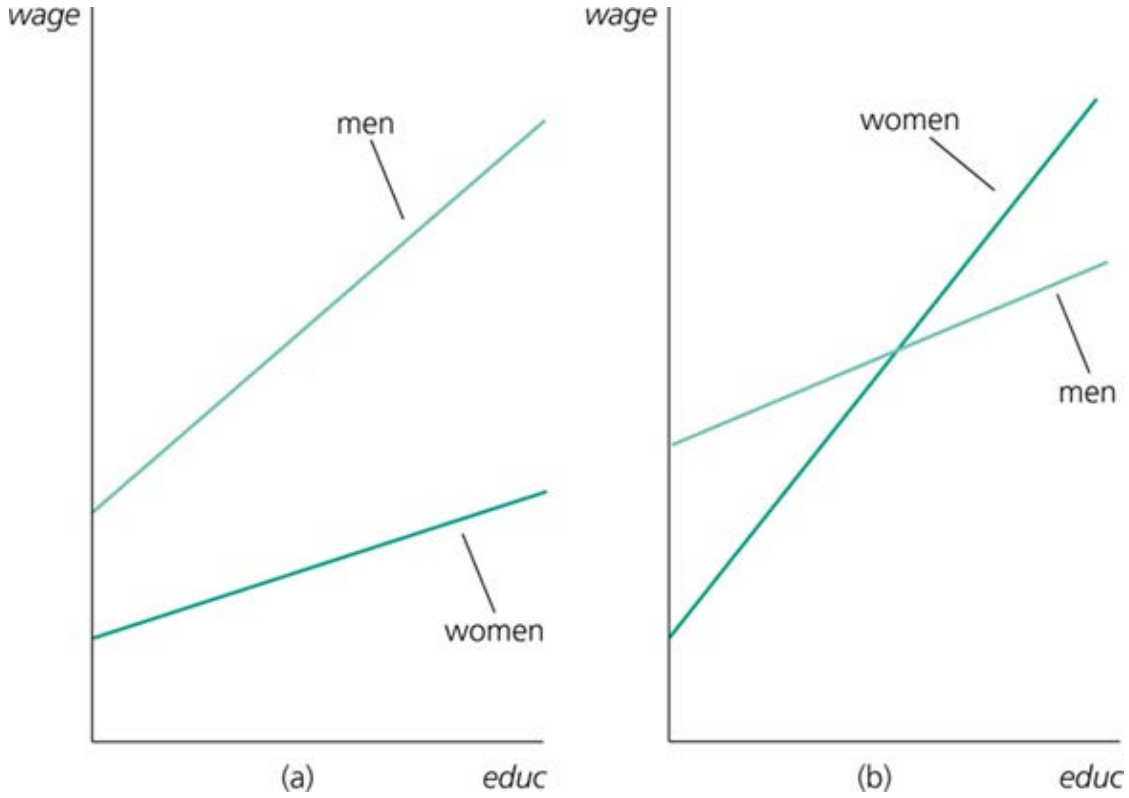
One more possibility is the interaction of a ratio variable with an indicator variable.

- In the wage-by-gender example, such an interaction would enable the researcher to see if there is a difference in the slopes of the two genders' education profiles as well as in the intercepts.
- This can be accomplished by including an interaction term between *female* and *educ*:

$$lwage = \beta_0 + \delta_0 female + \beta_1 educ + \delta_1 female * educ + u.$$

Interactions of binary and ratio variables (continued)

Now $(\beta_0 + \beta_1 educ)$ is the wage-schooling locus for men, and the wage-schooling locus for women is $(\beta_0 + \delta_0 +$



Interactions of binary and ratio variables (continued)

The first panel (a) depicts a case where women's intercept and slope are smaller than men's: $\delta_0 < 0 \; and \; \delta_1 < 0$.

In panel (b), the intercept is lower but the "return to schooling" (slope) is actually greater for women: $\delta_0 < 0 \text{ and } \delta_1 > 0$.

Either of the above hypotheses can be tested individually, or they can be tested jointly to determine if men and women have different wage-schooling loci:

$$H_0: \; \delta_0 = \delta_1 = 0; H_1: \; \text{at least one } delta \text{ coefficient is nonzero.}$$

Binary-ratio interactions in Stata

To perform the regression in STATA, the**fvvarlist** syntax to put in the regression command is:

i.binaryvar##c.nonbinaryvar.

The estimates using the **wage1.dta** data set yield a rejection of H_0 with high confidence, and it is probable that the rejection comes from a difference in intercepts (the difference in slopes is small in magnitude and not even close to significant on its own).

- See results on the next slide for the regression syntax and syntax for the (lazy man's) *F* test.

Binary-ratio interactions in Stata

```
. reg lwage i.female##c.educ c.exper##c.exper c.tenure##c.tenure
```

Source	SS	df	MS
Model	65.4081534	7	9.34402192
Residual	82.921598	518	.160080305
Total	148.329751	525	.28253286

```
Number of obs =     526
F(  7,   518) =   58.37
Prob > F      =  0.0000
R-squared     =  0.4410
Adj R-squared =  0.4334
Root MSE      =   .4001
```

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
1.female	-.2267886	.1675394	-1.35	0.176	-.5559289	.1023517
educ	.0823692	.0084699	9.72	0.000	.0657296	.0990088
female# c.educ 1	-.0055645	.0130618	-0.43	0.670	-.0312252	.0200962
exper	.0293366	.0049842	5.89	0.000	.019545	.0391283
c.exper# c.exper	-.0005804	.0001075	-5.40	0.000	-.0007916	-.0003691
tenure	.0318967	.006864	4.65	0.000	.018412	.0453814
c.tenure# c.tenure	-.00059	.0002352	-2.51	0.012	-.001052	-.000128
_cons	.388806	.1186871	3.28	0.001	.1556388	.6219732

```
. testparm 1.female 1.female#c.educ

 ( 1)  1.female = 0
 ( 2)  1.female#c.educ = 0

       F(  2,   518) =   34.33
            Prob > F =    0.0000
```

Testing for differences in regression functions across groups

To take interactions involving binary variables to their logical conclusion, consider a model in which all the other explanatory variables are interacted with the group indicator.

- Wooldridge describes an example of a regression used to determine whether male and female college students perform the same in college:

$$cumgpa = \beta_0 + \delta_0 female + (\beta_1 + \delta_1 female)sat + (\beta_2 + \delta_2 female)hsperc + (\beta_3 + \delta_3 female)tothrs + u.$$

The hypothesis being tested is whether all the *female* coefficients are simultaneously zero:

$$H_0: \ \delta_0 = \delta_1 = \delta_2 = \delta_3 = 0; H_1: \ \text{at least one } delta \text{ is nonzero.}$$

Performing the test in Stata

```
. reg cumgpa i.female##c.( sat hsperc tothrs) if  spring==1

      Source |       SS       df       MS              Number of obs =     366
-------------+------------------------------           F(  7,   358) =   34.95
       Model |  53.5391808     7  7.6484544             Prob > F      =  0.0000
    Residual |  78.3545052   358  .218867333            R-squared     =  0.4059
-------------+------------------------------           Adj R-squared =  0.3943
       Total |  131.893686   365  .361352564            Root MSE      =  .46783

------------------------------------------------------------------------------
      cumgpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.female |  -.3534862   .4105293    -0.86   0.390    -1.160838    .4538659
         sat |   .0010516   .0001811     5.81   0.000     .0006955    .0014078
      hsperc |  -.0084516   .0013704    -6.17   0.000    -.0111465   -.0057566
     tothrs |   .0023441   .0008624     2.72   0.007     .0006482    .0040401
            |
female#c.sat |
          1  |   .0007506   .0003852     1.95   0.052    -6.88e-06    .0015081
            |
  female#   |
c.hsperc    |
          1  |  -.0005498   .0031617    -0.17   0.862    -.0067676    .0056681
            |
  female#   |
c.tothrs    |
          1  |  -.0001158   .0016277    -0.07   0.943    -.0033169    .0030852
            |
       _cons |   1.480812   .2073336     7.14   0.000     1.073067    1.888557
------------------------------------------------------------------------------

. testparm 1.female 1.female#c.sat 1.female#c.hsperc 1.female#c.tothrs

 ( 1)  1.female = 0
 ( 2)  1.female#c.sat = 0
 ( 3)  1.female#c.hsperc = 0
 ( 4)  1.female#c.tothrs = 0

       F(  4,   358) =    8.18
            Prob > F =    0.0000

. testparm 1.female#c.sat 1.female#c.hsperc 1.female#c.tothrs

 ( 1)  1.female#c.sat = 0
 ( 2)  1.female#c.hsperc = 0
 ( 3)  1.female#c.tothrs = 0

       F(  3,   358) =    1.53
            Prob > F =    0.2054
```

Testing for differences across groups (continued)

The *F* test of this hypothesis rejects the null with a high level (>99.99%) of confidence.

- Testing the joint significance of only the three interaction terms, however, does not reject the null.

This recommends modeling the relationship using an indicator for *female* as the only difference between the genders ($F = 1.53$).

Stata can easily accommodate interactions between an indicator variable and a large set of other explanatory variables using **fvvarlist**.

The syntax just replaces a single interaction variable with a parenthesis that contains a list:

i.binaryvar##c.(list of interaction variables).

A binary dependent variable: the linear probability model

Technically there is nothing that prevents an empiricist from estimating a regression using OLS with a binary dependent variable, i.e., $y \in \{0,1\}$.

There is an entire Chapter in Wooldridge (17) devoted to superior ways to handle a dependent variable with such limitations, but estimating it by OLS yields easy-to-interpret results and is easier to explain to non-economists than other methods such as *probit*.

The linear probability model

Setting up the usual regression model with *k* regressors and a binary outcome, *y*,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + u,$$

could be used to analyze qualitative outcomes like:

- whether or not an individual participates in the labor force,
- whether spouses will dissolve (divorce) their marriage,
- if a firm will innovate (as measured by patent filings) or not.

The linear probability model (continued)

Very little changes about the estimation, but the interpretation of the results requires attention.

- The non-stochastic part of the model represents the expectation of *y*, conditional on all the *x*.

When *y* is binary, though, its conditional expectation equals the probability that $y = 1$:

$$E(y|\boldsymbol{x}) = \Pr(y = 1|\boldsymbol{x}).$$

Coefficients are usually interpreted as marginal effects on the conditional expectation of *y*.

- This is still true when *y* is binary, but the effects now have the "change in probability" interpretation.
- Since the model is linear in parameters and OLS estimates constant rates of change in probability, this is called the linear probability model.

The effect of the j^{th} regressor measures

$$\hat{\beta}_j = \frac{\partial \Pr(y = 1|\boldsymbol{x})}{\partial x_j} \text{ or without calculus notation, } \Delta \Pr(y = 1|\boldsymbol{x}) = \hat{\beta}_j \Delta x_j.$$

Linear probability model example

```
. reg inlf nwifeinc educ c.exper##c.exper age kidslt6 kidsge6
```

Source	SS	df	MS
Model	48.8080578	7	6.97257969
Residual	135.919698	745	.182442547
Total	184.727756	752	.245648611

Number of obs	= 753
F(7, 745)	= 38.22
Prob > F	= 0.0000
R-squared	= 0.2642
Adj R-squared	= 0.2573
Root MSE	= .42713

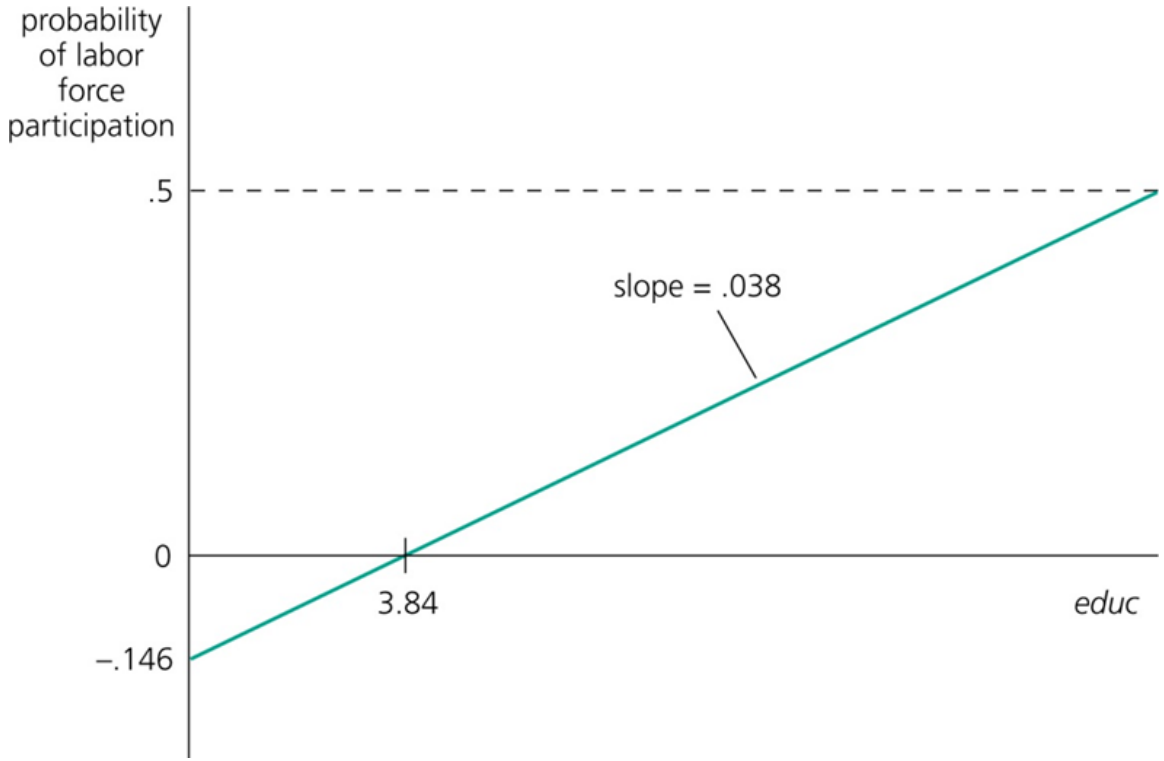
inlf	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
nwifeinc	-.0034052	.0014485	-2.35	0.019	-.0062488	-.0005616
educ	.0379953	.007376	5.15	0.000	.023515	.0524756
exper	.0394924	.0056727	6.96	0.000	.0283561	.0506287
c.exper# c.exper	-.0005963	.0001848	-3.23	0.001	-.0009591	-.0002335
age	-.0160908	.0024847	-6.48	0.000	-.0209686	-.011213
kidslt6	-.2618105	.0335058	-7.81	0.000	-.3275875	-.1960335
kidsge6	.0130122	.013196	0.99	0.324	-.0128935	.0389179
_cons	.5855192	.154178	3.80	0.000	.2828442	.8881943

Linear probability model example (continued)

To visualize the effect, consider the model of labor force participation as a function of education and other factors.

- The estimated relationship is graphed here.

You can see that the slope of the line is constant and equal to the estimated coefficient from the regression results (previous slide).



The linear probability model (continued)

These features, however, are the undoing of the linear probabilitymodel.

- On the graph you will notice that the model predicts a negative (!) probability of participation (for very low levels of education).
- This is less of a problem compared to the other end of the distribution, in which college graduate women (with observed covariates) are predicted to participate in the labor force with probability:

$$\Pr(y = 1|\boldsymbol{x}) = 0.5855 - 16.73 * 0.0034 + 16 * 0.38 + 27 * 0.0395 - 27^2 * 0.0006 - 45 * 0.0161 + 3 * 0.013 = 1.083 > \mathbf{1}.$$

- The covariates are: age 45, 3 children over 6 years old, 27 years of experience, non-wife income of 16.73 (1000s), and 16 years of education.

The linear probability model (continued)

So the linear probability model allows for fitted values that are impossible, but this can still be useful if you don't take the results literally.

- Consider the model's prediction to mean merely that a 45 year old college graduate with no small children is highly likely to work.
- But there are 17 women in the sample for which the model predicts probability greater than 1 (and 16 with less than 0).

The constant marginal effects are equally uncomfortable. The possibility of changing probability by more than 1 in either direction is absurd, but technically possible, in the linear probability model:

- this is what would happen by going from 0 to 4 children under 6.

$$\Delta \Pr(y = 1|\boldsymbol{x}) = -0.2618 * 4 = -1.048.$$

The linear probability model (continued)

Even if this was not a risk, the restriction that the effects are constant over the distribution of*x* is unrealistic.

Going from 0 to 1 child would likely reduce the probability of participation severely, but having a second child would not reduce the probability of working much more,

- i.e., a mother who is going to stay home to care for children is probably doing it already for the first child.

Inference about the linear probability model

The problem of impossible fitted values can be ameliorated using an indicator function that assigns a 0 or 1, based on whether the fitted value is at least 0.5.

$$\tilde{y}_i \equiv 1[\hat{y}_i \geq 0.5] \; s.t. \; \tilde{y}_i = \begin{cases} 0, & \hat{y}_i < 0.5 \\ 1, & \hat{y}_i \geq 0.5 \end{cases}$$

Along with the observed values of *y*, this set of predictions suggests a way of testing the goodness-of-fit: the percentage correctly predicted.

$$\%CP \equiv n^{-1} \sum_{i=1}^{n} [\tilde{y}_i y_i + (1 - \tilde{y}_i)(1 - y_i)] \, .$$

The linear probability model (concluded)

One more problem with the linear probability model is that it automatically violates the homoskedasticity assumption because the variance of the dependent variable, and therefore the error term, depends on *x*:

$$Var(y|\boldsymbol{x}) = \Pr(y = 1|\boldsymbol{x}) * [1 - \Pr(y = 1|\boldsymbol{x})].$$

This presents a problem for inference because the test statistics presented so far have been justified under homoskedasticity only (even asymptotically).

- Since the homoskedasticity assumption often needs to be relaxed in non-binary dependent variable models as well, this is not a huge shortcoming, but it is important at a minimum to interpret the test statistics from a LPM with caution.

More on policy analysis and program evaluation

When natural scientists perform lab experiments to test their theories, they take two groups of subjects ("mice") that are initially the same.

- Then they randomly assign some of the mice to a treatment group and the rest to a control group.
- Afterward they observe both groups and compare them to see if there is any difference.
- If the two groups differ, it can reasonably be inferred that the treatment caused the change.

Example: two identical groups of mice. One group is given a vaccine expected to protect against Anthrax. Then both groups are given a lethal dose of Anthrax.

- If the group given the treatment is in better health ("not dead") afterward than the other group, the scientists conclude that the vaccine protected them from it and caused them to stay alive.

Policy analysis and program evaluation

One wouldn't even need regression analysis to test the hypothesis that survival is independent of the treatment.

- The researcher could just test for a difference in sample survival proportions (like you learn about in introductory statistics class).

If you had a research question in which assignment to treatment ("participation in the program") was random, only conditional on some other factors, you would want to control for the other factors and use regression analysis such as:

$$outcome = \beta_0 + \delta_0 partic + other\ factors + u.$$

The challenge in estimating δ_0 lies in successfully conditioning on ("controlling for") *other factors* that are correlated with participation (*partic*).

This challenge is almost universal in empirical economics. It takes the following forms.

Program evaluation challenges

Participation is voluntary, and there is selection bias resulting from self-selection.

- A voluntary review session is more likely to be attended by "good" students, exaggerating its effect on subsequent test scores.
- Alcohol use is more likely among discouraged individuals, exaggerating its effect (if it has one) on labor market outcomes, e.g., unemployment.

Binary variables that are not self-selected, such as gender and race, can still be correlated with confounding variables.

- Parents' incomes,
- Quality of schooling,
- Classmates' ("peer") effects on school performance,
- Parents' expectations and human capital investment decisions.

Cross-sectional regressions that fail to control for relevant differences between groups will produce biased estimates and misleading evidence about discrimination.

Discrete dependent variables

It should not be surprising that the interpretation of regression coefficients when *y* is binary generalizes to cases in which *y* takes on a small number of integer values. Common examples:

- *number of children*,
- *games won* during a football season,
- *patents filed*, and
- *job offers received*.

It's important to remember that the coefficients estimate the change in the conditional expectation of *y* for a one unit increase in *x*.

In the following (textbook) example, the effect of education on number of children is (-0.079).

- This can be interpreted through the familiar lens by imagining 100 women obtaining a marginal year of schooling each. The fertility would be predicted to fall by about 8 children as a result.

Discrete dependent variables example

```
. reg children age educ electric
```

Source	SS	df	MS
Model	12090.395	3	4030.13167
Residual	9419.6371	4354	2.16344444
Total	21510.0321	4357	4.93689055

Number of obs	= 4358
F(3, 4354)	= 1862.83
Prob > F	= 0.0000
R-squared	= 0.5621
Adj R-squared	= 0.5618
Root MSE	= 1.4709

children	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.1769991	.0027291	64.86	0.000	.1716486	.1823496
educ	-.0787507	.0063195	-12.46	0.000	-.09114	-.0663614
electric	-.3617579	.0680316	-5.32	0.000	-.4951345	-.2283813
_cons	-2.071091	.0947413	-21.86	0.000	-2.256832	-1.88535

Discrete dependent variables (concluded)

Once again there are estimation methods better suited to analyzing discrete (“limited”) dependent variables that are discussed later in the text, but this is an easy-to-interpret extension of regression analysis that can be useful for grasping the relationships between variables.

Conclusion

It's hard to overemphasize how much qualitative information is out there and (potentially) subject to econometric analysis.

Incorporating qualitative variables into OLS as independent variables can be don easily using indicator variables.

- And it presents no significant problems for estimation or inference.

Binary and other discrete (taking on a limited number of values) variables can hypothetically be used as dependent ("*y*") variables, too, but there are better methods than OLS for estimation.

- See Chapter 17 in Wooldridge.
- When a binary variable is used as the *y* variable in OLS, the coefficients have a "change in probability" interpretation, under the Linear Probability Model.

Policy analysis using an indicator for exposure to a program is one of the most important tasks empirical economists do, but

- estimating the effects is fraught with self-selection bias.

(optional) Restrictions when testing ordinal regressors

You want to know whether all the marginal effects are the same in both specifications. So think about comparing them, a là the table.

If all the *delta* coefficients are perfectly proportional to γ_1, then you might as well estimate the model with fewer parameters and just enter the ordinal variable as a regressor.

- If they're not, then the model fits better with indicators and you should reject the null hypothesis.

	Marginal Effect	
Change	Ordinal Regressor	Indicators for Each Category
$FRa = 5 \rightarrow FRa = 4$	$-\gamma_1$	δ_4
$FRa = 5 \rightarrow FRa = 3$	$-2\gamma_1$	δ_3
$FRa = 5 \rightarrow FRa = 2$	$-3\gamma_1$	δ_2
$FRa = 5 \rightarrow FRa = 1$	$-4\gamma_1$	δ_1

Restrictions when testing ordinal regressors (continued)

So your null hypothesis is that the indicators' marginal effects are the same as the marginal effects of 1, 2, 3, ***and*** 4 unit changes in the ordinal measure. I.e.,

$$H_0: \delta_1 = -\gamma_1; \delta_2 = -2\gamma_1; \delta_3 = -3\gamma_1; \delta_4 = -4\gamma_1.$$
$$H_1: \ H_0 \text{ is not true.}$$

- The negative signs are due to the fact that my base category is the highest category.
- The example in Wooldridge's text differs by omitting the lowest category.
  - If my informal theory is correct, γ_1 will be positive, and all the δ_j will be negative, so the negative signs are necessary to make the marginal effects the same sign.

Back.