Wooldridge, Introductory Econometrics, 5th ed.

Chapter 7: Multiple regression analysis with qualitative information: Binary (or dummy) variables

We often consider relationships between observed outcomes and qualitative factors: models in which a continuous dependent variable is related to a number of explanatory factors, some of which are quantitative, and some of which are qualitative. In econometrics, we also consider models of qualitative dependent variables, but we will not explore those models in this course due to time constraints. But we can readily evaluate the use of qualitative information in standard regression models with continuous dependent variables.

Qualitative information often arises in terms of some coding, or index, which takes on a

number of values: for instance, we may know in which one of the six New England states each of the individuals in our sample resides. The data themselves may be coded with the biliteral "MA", "RI", "ME", etc. How can we use this factor in a regression equation? In the data, `state` takes on six distinct values. We must create six **binary variables**, or **dummy variables**, each of which will refer to one state–that is, that variable will be 1 if the individual comes from that state, and 0 otherwise. We can generate this set of 6 variables easily in Stata with the command `tab state, gen(st)`, which will create 6 new variables in our dataset: `st1, st2, ... st6`. Each of these variables are dummies–that is, they only contain 0 or 1 values. If we add up these variables, we get–exactly–a vector of 1's, suggesting that we will never want to use all 6 variables in a regression (since by knowing the values of any 5...) We may also find the proportions of each state's citizens in our sample

very easily: `summ st*` will give the descriptive statistics of all 6 variables, and the mean of each `st` dummy is the sample proportion living in that state.

In Stata 11+, we actually do not have to create these variables explicitly; we can make use of *factor variables*, which will automatically create the dummies.

How can we use these dummy variables? Say that we wanted to know whether incomes differed significantly across the 6-state region. What if we regressed `income` on **any five** of these `st` dummies? We could do this with explicit variables as

```
regress income st2-st6
```

or with factor variables as

```
regress income i.state
```

In either case, we are estimating the equation

$$income = \beta_0 + \beta_2 st_2 + \beta_3 st_3 + \beta_4 st_4 + \beta_5 st_5 + \beta_6 st_6 + u \tag{1}$$

where I have suppressed the observation subscripts. What are the regression coefficients in this case? β_0 is the average income in the 1^{st} state–the dummy for which is excluded from the regression. β_2 is the difference between the income in state 2 and the income in state 1. β_3 is the difference between the income in state 3 and the income in state 1, and so on. What is the ordinary "ANOVA F" in this context–the test that all the slopes are equal to zero? Precisely the test of the null hypothesis:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 \tag{2}$$

versus the alternative that not all six of the state means are the same value. It turns out

that we can test this same hypothesis by excluding any one of the dummies, and including the remaining five in the regression. The coefficients will differ, but the $p-$value of the ANOVA F will be identical for any of these regressions. In fact, this regression is an example of "classical one-way ANOVA"–testing whether a qualitative factor (in this case, state of residence) explains a significant fraction of the variation in income.

What if we wanted to generate point and interval estimates of the state means of income? Then it would be most convenient to reformulate (1) by including all 6 dummies, and removing the constant term. This is, algebraically, the same regression:

```
regress income st1-st6, noconstant
```

or with factor variables as

```
regress income ibn.state, noconstant
```

The coefficient on the now-included st_1 will be precisely that reported above as β_0. The coefficient reported for st_2 will be precisely $(\beta_0 + \beta_2)$ from the previous model, and so on. But now those coefficients will be reported with confidence intervals around the state means. Those statistics could all be calculated if you only estimated (1), but to do so you would have to use `lincom` for each coefficient. Running this alternative form of the model is much more convenient for estimating the state means in point and interval form. But to test the hypothesis (2), it is most convenient to run the original regression–since then the ANOVA F performs the appropriate test with no further ado.

What if we fail to reject the ANOVA F null? Then it appears that the qualitative factor "state"

does not explain a significant fraction of the variation in income. Perhaps the relevant classification is between northern, more rural New England states (NEN) and southern, more populated New England states (NES). Given the nature of dummy variables, we may generate these dummies two ways. We can express the Boolean condition in terms of the `state` variable: `gen nen = (state==‘‘VT’’ | state==‘‘NH’’ | state==‘‘ME’’)`. This expression, with parens on the right hand side of the `generate` statement, evaluates that expression and returns true (1) or false (0). The vertical bar (|) is Stata's OR operator; since every person in the sample lives in one and only one state, we must use OR to phrase the condition that they live in northern New England. But there is another way to generate this `nen` dummy, given that we have `st1...st6` defined for the regression above. Let's say that Vermont, New Hampshire and Maine have been coded as `st6, st4`

and `st3`, respectively. We may just `gen nen = st3+st4+st6`, since the sum of mutually exclusive and exhaustive dummies must be another dummy. To check, the resulting `nen` will have a mean equal to the percentage of the sample that live in northern New England; the equivalent `nes` dummy will have a mean for southern New England residents; and the sum of those two means will of course be 1. We can then run a simplified form of our model as `regress inc nen`; the ANOVA F statistic for that regression tests the null hypothesis that incomes in northern and southern New England do not differ significantly. Since we have excluded nes, the "slope" coefficient on `nen` measures the amount by which northern New England income differs from southern New England income; the mean income for southern New England is the constant term. If we want point and interval estimates for those means, we should `regress inc nen nes, noc`.

Regression with continuous and dummy variables

In the above examples, we have estimated "pure ANOVA" models–regression models in which all of the explanatory variables are dummies. In econometric research, we often want to combine quantitative and qualitative information, including some regressors that are measurable and others that are dummies. Consder the simplest example: we have data on individuals' wages, years of education, and their gender. We could create two gender dummies, male and female, but we will only need one in the analysis: say, female. We create this variable as `gen female = (gender==''F'')`. We can then estimate the model:

$$wage = \beta_0 + \beta_1 educ + \beta_2 female + u \qquad (3)$$

The constant term in this model now becomes the wage for a male with zero years of education. Male wages are predicted as $b_0 +$

$b_1 educ$, while female wages are predicted as $b_0 + b_1 educ + b_2$. The gender differential is thus b_2. How would we test for the existence of "statistical discrimination"–that, say, females with the same qualifications are paid a lower wage? This would be $H_0 : \beta_2 < 0$. The $t-$statistic for b_2 will provide us with this hypothesis test. What is this model saying about wage structure? Wages are a linear function of the years of education. If b_2 is significantly different than zero, then there are two "wage profiles"– parallel lines in $\{educ, wage\}$ space, each with a slope of b_1, with their intercepts differing by b_2.

What if we wanted to expand this model to consider the possibility that wages differ by both gender and race? Say that each worker is classified as race=white or race=black. Then we could `gen black = (race==''black'')` to create the dummy variable, and add it to (3).

What, now, is the constant term? The wage for a white male with zero years of education. Is there a significant race differential in wages? If so, the coefficient b_3, which measures the difference between white and black wages, ceteris paribus, will be significantly different from zero. In $\{educ, wage\}$ space, the model can be represented as four parallel lines, with each intercept labelled by a combination of gender and race.

What if our racial data classified each worker as white, Black or Asian? Then we would run the regression:

$$wage = \beta_0 + \beta_1 educ + \beta_2 female + \beta_3 Black + \beta_4 Asian + u \quad (4)$$

or, with factor variables,

```
regress wage educ female i.race
```

where the constant term still refers to a white male. In this model, b_3 measures the difference between black and white wages, ceteris paribus, while b_4 measures the difference between Asian and white wages. Each can be examined for significance. But how can we determine whether the qualitative factor, `race`, affects wages? That is a joint test, that both $\beta_3 = 0$ and $\beta_4 = 0$, and should be conducted as such. If factor variables were used, we could do this with

```
testparm i.race
```

No matter how the equation is estimated, we should not make judgments based on the individual dummies' coefficients, but should rather include both race variables if the null is rejected, or remove them both if it is not. When

we examine a qualitative factor, which may give rise to a number of dummy variables, they should be treated as a group. For instance, we might want to modify (3) to consider the effect of state of residence:

$$wage = \beta_0 + \beta_1 educ + \beta_2 female + \sum_{j=2}^{6} \gamma_j st_j + u \tag{5}$$

where we include any 5 of the 6 `st` variables designating the New England states. The test that wage levels differ significantly due to state of residence is the joint test that $\gamma_j = 0, j = 2, ..., 6$ (or, if factor variables are used, `testparm i.state`). A judgment concerning the relevance of state of residence should be made on the basis of this joint test (an F-test with 5 numerator degrees of freedom).

Note that if the dependent variable was measured in log form, the coefficients on dummies

would be interpreted as percentage changes; if (5) was respecified to place log($wage$) as the dependent variable, the coefficient b_1 would measure the percentage return to education (how many percent does the wage change for each additional year of education), while the coefficient b_2 would measure the (approximate) percentage difference in wage levels between females and males, ceteris paribus. The state dummies would, likewise, measure the percentage difference in wage levels between that state and the excluded state (state 1).

We must be careful when working with variables that have an ordinal interpretation, and are thus coded in numeric form, to treat them as ordinal. For instance, if we model the interest rate corporations must pay to borrow ($corprt$) as a function of their credit rating, we consider that Moody's and Standard and Poor's assign credit ratings somewhat like grades:

AAA, AA, A, BAA, BA, B, C, et cetera. Those could be coded as 1,2,...,7. Just as we can agree that an "A" grade is better than a "B", a triple-A bond rating results in a lower borrowing cost than a double-A rating. But while GPAs are measured on a clear four-point scale, the bond ratings are merely ordinal, or ordered: everyone agrees on the rating scale, but the differential between AA borrowers' rates and A borrowers' rates might be much smaller than that between B and C borrowers' rates: especially the case if C denotes "below investment grade", which will reduce the market for such bonds. Thus, although we might have a numeric index corresponding to $AAA...C$, we should not assume that $\partial corprt/\partial index$ is constant; we should not treat $index$ as a cardinal measure. Clearly, the appropriate way to proceed is to create dummy variables for each rating class, and include all but one of those variables in a regression of $corprt$ on bond rating and other relevant factors. For instance, if

we leave out the AAA dummy, all of the ratings class dummies' coefficients will then measure the degree to which those borrowers' bonds bear higher rates than those of AAA borrowers. But we could just as well leave out the C rating class dummy, and measure the effects of ratings classes relative to the worst credits' cost of borrowing.

Interactions involving dummy variables

Just as continuous variables may be interacted in regression equations, so can dummy variables. We might, for instance, have one set of dummies indicating the gender of respondents ($female$) and another set indicating their marital status ($married$). We could regress $lwage$ on these two dummies:

$$lwage = b_0 + b_1 female + b_2 married + u$$

which gives rise to the following classification of mean wages, conditional on the two factors (which is thus a classic "two-way ANOVA" setup):

	$male$	$female$
$unmarried$	b_0	$b_0 + b_1$
$married$	$b_0 + b_2$	$b_0 + b_1 + b_2$

We assume that the two effects, gender and marital status, have independent effects on the dependent variable. Why? Because this joint distribution is modelled as the product of the marginals. What is the difference between male and female wages? b_1, irrespective of marital status. What is the difference between unmarried and married wages? b_2, irrespective of gender.

If we were to relax the assumption that gender and marital status had independent effects

on wages, we would want to consider their **interaction**. Since there are only two categories of each variable, we only need one interaction term, fm, to capture the possible effects. As above, that term could be generated as a Boolean (noting that `&` is Stata's AND operator): `gen fm=(female==1) & (married==1)`, or we could generate it algebraically, as `gen fm=female*married`. In either case, it represents the intersection of the sets. We then add a term, $b_3 fm$, to the equation, which then appears as an additive constant in the lower right cell of the table. Now, if the coefficient on fm is significantly nonzero, the effect of being female on the wage differs, depending on marital status, and vice versa. Are the interaction effects important–that is, does the joint distribution differ from the product of the marginals? That is easily discerned, since if that is so b_3 will be significantly nonzero.

Using explicit variables, this would be estimated as

```
regress wage female married fm
```

or, with factor variables, we can make use of the *factorial interaction* operator:

```
regress wage female married i.female#i.married
```

or, in an even simpler form,

```
regress wage i.female##i.married
```

where the double hash mark indicates the *full factorial* interaction, including both the main effects of each factor and their interaction.

Two extensions of this framework come to mind. Sticking with two-way ANOVA (considering two factors' effects), imagine that instead of marital status we consider $race =$

$\{white, Black, Asian\}$. To run the model without interactions, we would include two of these dummies in the regression–say, $Black$ and $Asian$; the constant term would be the mean wage of a white male (the excluded class). What if we wanted to include interactions? Then we would define f_Black and f_Asian, and include those two regressors as well. The test for the significance of interactions is now a joint test that these two coefficients are jointly zero.

With factor variables, we can just say

```
regress wage i.female##i.race
```

where the factorial interaction includes all race categories, both in levels and interacted with the female dummy.

A second extension of the interaction concept is far more important: what if we want to consider a regular regression, on quantitative variables, but want to allow for different slopes

for different categories of observations? Then we create interaction effects between the dummies that define those categories and the measured variables. For instance,

$$lwage = b_0 + b_1 female + b_2 educ + b_3 (female \times educ) + u$$

Here, we are in essence estimating two separate regressions in one: a regression for males, with an intercept of b_0 and a slope of b_2, and a regression for females, with an intercept of $(b_0 + b_1)$ and a slope of $(b_2 + b_3)$. Why would we want to do this? We could clearly estimate the two separate regressions, but if we did that, we could not conduct any tests (e.g. do males and females have the same intercept? The same slope?). If we use interacted dummies, we can run one regression, and test all of the special cases of this model which are nested within: that the slopes are the same, that

the intercepts are the same, and the "pooled" case in which we need not distinguish between males and females. Since each of these special cases merely involves restrictions on this general form, we can run this equation and then just conduct the appropriate tests.

This can be done with factor variables as

```
regress wage i.female##c.educ
```

where we must use the `c.` operator to tell Stata that `educ` is to be treated as a continuous variable, rather than considering all possible levels of that variable in the dataset.

If we extended this logic to include $race$, as defined above, as an additional factor, we would include two of the race dummies (say, $Black$ and $Asian$) and interact each with $educ$. This would be a model without interactions, where

the effects of gender and race are considered to be independent, but it would allow us to estimate different regression lines for each combination of gender and race, and test for the importance of each factor. These interaction methods are often used to test hypotheses about the importance of a qualitative factor–for instance, in a sample of companies from which we are estimating their profitability, we may want to distinguish between companies in different industries, or companies that underwent a significant merger, or companies that were formed within the last decade, and evaluate whether their expenditures on R&D or advertising have the same effects across those categories.

All of the necessary tests involving dummy variables and interacted dummy variables may be easily specified and computed, since models without interacted dummies (or without certain dummies in any form) are merely restricted

forms of more general models in which they appear. Thus, the standard "subset F" testing strategy that we have discussed for the testing of joint hypotheses on the coefficient vector may be readily applied in this context. The text describes how a "Chow test" may be formulated by running the general regression, running a restricted form in which certain constraints are imposed, and performing a computation using their sums of squared errors; this computation is precisely that done with Stata's `test` command. The advantage of setting up the problem for the `test` command is that any number of tests (e.g. above, for the importance of gender, or for the importance of race) may be conducted after estimating a single regression; it is not necessary to estimate additional regressions to compute any possible "subset F" test statistic, which is what the "Chow test" is doing.