

Biostatistics 208 Lab #3 1/21/10 – Stata 10 version

The purpose of this lab is to give you practice in dealing with categorical predictors in linear regression in Stata. If you have Stata 11, use the other posted version of this lab. Note that most of the commands covered will carry over to applications in logistic and Cox regression without change. We will use baseline data for a 20% random sub-sample of the 2763 post-menopausal women with heart disease who participated in the HERS trial. You can download `lab3.dta` from the course website. The variables in the dataset are as follows:

1. `bmi` – body mass index (BMI) in kg/m^2
2. `physact` – relative physical activity by self-report, coded 1-5
3. `age10` – age in units of 10 years
4. `smoking` – an indicator of current smoking (1 = yes, 0 = no).
5. `drinkany` – an indicator of any current alcohol use, coded like smoking.

We will focus on `physact`, the 5-level self-reported physical activity variable discussed in lecture, as an ordinal categorical predictor of BMI. First, tabulate the physical activity variable to see how many women there are in each category, and at the same time obtain the mean and SD of BMI in each group:

```
tab physact, sum(bmi)
```

Now use the `xi:` and `i.` command prefixes to regress `bmi` on `physact` as a five-level categorical predictor. This prompts Stata to create the indicator variables needed to represent the distinct levels of `physact`. As described in the last lecture, only four of the five are needed (due to the inclusion of the intercept term); Stata omits the indicator for the first level by default.

```
xi: reg bmi i.physact
```

You can easily see that the coefficient estimate for the intercept (`_cons`) is the sample mean for BMI among “much less active” women (`physact=1`).

The following Stata code runs a so-called “do-loop” to efficiently compute the means in the other four groups, using the regression results. The commands must be typed exactly as written, on three lines, including the brackets; the opening single quote in the second line is a back quote (also an accent grave), and must be typed using the key located on the upper left of the keyboard between the tab and escape keys. The closing single quote is an apostrophe (the key to the left of return key).

```
forvalues i = 2/5 {  
  lincom _cons + _Iphysact_`i'  
}
```

Stata help for the `forvalues` command (and the related command `foreach`) has other examples of how to run do-loops in STATA, which can make some repetitive data management tasks much easier. We could also get the group-specific means more simply using the command

```
adjust, by(physact)
```

Now, use the following commands to test for heterogeneity, trend, and departure from trend in the unadjusted association of physical activity with BMI. In the following commands, you don’t need to type comments, which begin with asterisks:

```
* test for heterogeneity
testparm _Iphys*
```

```
* test for linear trend
test - _Iphysact_2 + _Iphysact_4 + 2*_Iphysact_5 = 0
```

Recall that rejecting the hypothesis of heterogeneity implies that the average BMI varies across levels of physical activity. Is there convincing evidence for heterogeneity in mean BMI across levels of physical activity?

A linear trend in the average BMI over the physical activity groups is a particular type of heterogeneity that is of interest here, since `physact` represents an ordinal variable. The procedure above is especially appropriate for predictors that can be considered as ordinal, but where the values don't have a meaningful numerical interpretation (as is the case with `physact`). Is there a statistically significant linear trend?

As discussed in section 4.3.5 of the book, the finding of evidence for a linear trend in means leaves open the possibility that the trend is actually nonlinear (e.g. quadratic). The book provides a procedure to evaluate evidence for a departure from linearity that is appropriate for ordinal categorical predictors. Please look over this section when you have the chance. We'll see in later lectures that alternate procedures are available to detect nonlinearity in the relationship between regression outcomes and continuous predictors.

Next, we will re-do the above analysis adjusting for age, smoking, and alcohol use, possible confounders of the association between physical activity and BMI. In this model, estimated mean BMI in each of the 5 groups must reflect the adjustment for any between-group differences in age, smoking, and alcohol use, whose values we need to hold constant. This can be done using the `adjust` command, which computes adjusted mean BMI in each of the 5 groups, holding the other predictors fixed at their sample mean values.

```
* repeat analysis adjusting for age, smoking, and alcohol use
xi: reg bmi i.physact age10 smoking drinkany
```

```
* compute adjusted mean BMI at each level of physical activity
adjust age10 smoking drinkany, by(physact)
```

```
* test for heterogeneity
testparm _Iphys*
```

```
* test for linear trend
test _Iphysact_2 = _Iphysact_4 + 2*_Iphysact_5
```

Using the output, make sure you can you can:

- Interpret the regression coefficients for variables `_cons`, `physact` (levels 2-5), `age10`, and `drinkany`.
- Interpret the results of the tests for heterogeneity and trend in the light of the pattern in the adjusted means produced by the `margins` command.
- Relate the differences between the adjusted group means given by the `margins` command to the coefficient estimates for variables `_cons` and `physact` (levels 2-5). Recall that

the coefficients for the four `physact` indicators estimate adjusted differences in mean BMI with respect to the reference group (level 1 of `physact`).