

## Biostat 208 Lab #5, 2/04/10

The purpose of this lab is to give you practice in using the `regress` and `lincom` commands in STATA to look at interaction. Download the dataset `lab5.dta` from the website. *Please note:* the optional sections are really optional – only do them if you would like to learn a bit more STATA programming.

### 1 Interaction of hormone therapy and statin use

In lecture we considered the interaction of HT and baseline statin use on LDL cholesterol levels at the first annual post-randomization HERS visit. In this section we will examine whether HT interacts with baseline statin use in its effect on *change* in LDL cholesterol from baseline to the first annual visit. This will substantially change the results for statins but only trivially affect the results for HT, with no effect on interpretation (why?). First, we will tabulate the mean changes in LDL by treatment group, stratified by statin use, using the command

```
table ht statins, contents(mean ldlch)
```

This may help to interpret some of the regression results. Now compute the LDL changes and the product term involving the treatment and statin use indicators, then run the regression.

```
gen ldlch = ldl1 - ldl0
gen htstat = ht * statins
reg ldlch statins ht htstat
```

1. *Is there evidence for differential effects of HT on change in LDL according to use of statins at baseline?*

Use `lincom` statements as necessary to estimate the effect of HT in the two subgroups defined by baseline statin use. Here is the regression equation:

$$E[\Delta\text{LDL}|\mathbf{x}] = \beta_0 + \beta_1\text{statins} + \beta_2\text{ht} + \beta_3\text{htstat}$$

Here is the problem worked out for non-users of statins

Statins	HT	_cons	statins	ht	htstat
No	Yes	1	0	1	0
No	No	1	0	0	0
	Difference	0	0	1	0

So the effect of HT in non-users of statins is given by the coefficient for `ht`, which we can read from the `regress` output. Now here is the table for statin users:

Statins	HT	_cons	statins	ht	htstat
Yes	Yes	1	1	1	1
Yes	No	1	1	0	0
	Difference	0	0	1	1

So to evaluate the effect of HT among users of statins, use the command

```
lincom ht + htstat
```

2. *Summarize the HT effects by baseline statin use.*

Now use the following two tables to compute the associations of being a statin user with change in LDL, first in the HT group.

Statins	HT	_cons	statins	ht	htstat
Yes	Yes				
No	Yes				
	Difference				

Use `lincom` if necessary to compute the difference in LDL changes within the HT group, and interpret. Now do the computation for the placebo group.

Statins	HT	_cons	statins	ht	htstat
Yes	No				
No	No				
	Difference				

Again use `lincom` if necessary to compute the statin effect.

3. Summarize the differences between users and non-users of statins in each treatment group, and comment on the results for the placebo group.

4. How do these estimates of the statin effects stratified by assignment to HT differ from the estimated effects of HT stratified by statin use?

5. Relate the numbers in the table of LDL changes by treatment assignment and baseline statin use to estimates you got from `regress` and `lincom`.

## 2 Interaction between glucose metabolism and BMI

Using the tools you have developed here, we will check whether the association of BMI with systolic blood pressure (SBP) is modified by glucose metabolism group, a three-level variable with categories normal (fasting glucose (FG) < 100), impaired fasting glucose (IFG; FG 100-125), and diabetes (by self-report or FG ≥ 126). Here is the code to make the variable and run the model. In this part of the lab, we will use Stata's built-in ability to code interactions. Note that to get an overall test for interaction, we need to use an *F*-test, since we are in effect comparing the effect of BMI across three, not just two, subgroups. As in the lecture, we will center BMI to make the estimates for metabolic group in the regression output more interpretable.

```

recode glucose min/99=1 100/125=2 126/max=3, gen(dmgrp)
replace dmgrp = 3 if diabetes==1
label define dmgrpvals 1 "normal" 2 "IFG" 3 "diabetes"
label values dmgrp dmgrpvals
gen cbmi = bmi-28.6

```

```

* version 11 code
reg sbp i.dmgrp##c.cbmi
* overall 2-df test for interaction
testparm dmgrp#c.cbmi

```

```

* Version 10 and earlier code
xi: reg sbp i.dmgrp*c.bmi
* overall 2-df test for interaction
testparm *X*

```

*STATA notes:* The new variables automatically made by STATA are keyed to the codes of the underlying `dmgrp` variable. Here is a table giving the Version 11 and Version 10 and earlier variable names:

Variable definition	Version 10	Version 11
Indicator for IFG	<code>_Idmgrp_2</code>	<code>2.dmgrp</code>
Indicator for Diabetes	<code>_Idmgrp_3</code>	<code>3.dmgrp</code>
IFG-cbmi product term	<code>_IdmgXcbmi_2</code>	<code>2.dmgrp#c.cbmi</code>
Diabetes-cbmi product term	<code>_IdmgXcbmi_3</code>	<code>3.dmgrp#c.cbmi</code>

The reference group is normals, because they have the lowest code, and no indicator is made for them. The indicator `2.dmgrp = _Idmgrp_2 = 1` for the IFG group and 0 for the others, while `3.dmgrp = _Idmgrp_3 = 1` for the diabetes group and 0 for the others. In Version 11 these new variables are not retained as part of the dataset, in contrast to earlier versions. Within the `reg` command, the `##` or `*` operators instruct STATA to set up the interaction product terms automatically. The Version 10 `testparm` command takes advantage of the fact that the product terms made by `xi:` are the only predictors in the model with names including a capital X.

1. *Is there statistically significant variability in the associations of BMI with SBP across the three groups?*

We will first evaluate the effects of a 5-unit increase BMI in the different glucose metabolism groups. Here is the regression equation using Version 11 variable names

$$ESBP|\mathbf{x}] = \beta_0 + \beta_1 2.dmgrp + \beta_2 3.dmgrp + \beta_3 cbmi + \beta_4 2.dmgrp\#c.cbmi + \beta_5 3.dmgrp\#c.cbmi$$

Now the regression equation using Version 10 and earlier variable names:

$$ESBP|\mathbf{x}] = \beta_0 + \beta_1\_Idmgrp\_2 + \beta_2\_Idmgrp\_3 + \beta_3 cbmi + \beta_4\_IdmgXcbmi\_2 + \beta_5\_IdmgXcbmi\_3$$

Recall that `2.dmgrp#c.cbmi = _IdmgXcbmi_2 = cbmi` for the IFG group (`2.dmgrp = _Idmgrp_2 = 1`) and 0 for the other two groups; likewise `3.dmgrp#c.cbmi = _IdmgXcbmi_3 = cbmi` for the Diabetes group and 0 for others. Now we can set up our table, first for the Normal group. I will use the Version 11 variable names.

Group	BMI	<code>_cons</code>	<code>2.dmgrp</code>	<code>3.dmgrp</code>	<code>cbmi</code>	<code>2.dmgrp#c.cbmi</code>	<code>3.dmgrp#c.cbmi</code>
Normal	$k + 5$	1	0	0	$k - 28.6 + 5$	0	0
Normal	$k$	1	0	0	$k - 28.6$	0	0
	Difference	0	0	0	5	0	0

So we can use `lincom 5*cbmi` to estimate what we want. Now for the IFG group:

Group	BMI	<code>_cons</code>	<code>2.dmgrp</code>	<code>3.dmgrp</code>	<code>cbmi</code>	<code>2.dmgrp#c.cbmi</code>	<code>3.dmgrp#c.cbmi</code>
IFG	$k + 5$	1	1	0	$k - 28.6 + 5$	$k - 28.6 + 5$	0
IFG	$k$	1	1	0	$k - 28.6$	$k - 28.6$	0
	Difference	0	0	0	5	5	0

So in this case, we can use `lincom 5*(cbmi+2.dmgrp#c.cbmi)` or `lincom 5*(cbmi+_IdmgXcbmi_2)` (Version 10).

2. *Do this for the women with diabetes on your own.*

Group	BMI	<code>_cons</code>	<code>2.dmgrp</code>	<code>3.dmgrp</code>	<code>cbmi</code>	<code>2.dmgrp#c.cbmi</code>	<code>3.dmgrp#c.cbmi</code>
Diabetes	$k + 5$						
Diabetes	$k$						
	Difference						

3. Interpret the coefficient estimates for `2.dmgrp` (`_Idmgrp_2`) and `3.dmgrp` (`_Idmgrp_3`).
4. Now compute the difference between the diabetes and IFG groups if  $BMI = 30$  (i.e., centered  $BMI = 1.4$ ).

Group	BMI	_cons	2.dmgrp	3.dmgrp	cbmi	2.dmgrp#c.cbmi	3.dmgrp#c.cbmi
Diabetic	30						
IFG	30						
	Difference						

5. Are the differences between the glucose metabolism groups interpretable as average causal effects?

### 3 Optional: plotting adjusted regression lines

You can use the following Version 10 code to get a plot of the regression lines for the three glucose metabolism groups. (I haven't figured out how to do this in Version 11). To make the problem more interesting, we'll adjust for age, race/ethnicity, and poor/fair self-reported health. In the following, the lengthy `twoway` command has to be typed without carriage returns – I have just broken it up to make it more legible. You could paste all of this into a do-file and edit it until you get it de-bugged. In that case, the command can be on separate lines separated by carriage returns, but you have to add a space and then the line separator `///` at the end of each individual line making up the `twoway` command. However, that line separator will not work in interactive mode at the command line.

```
capture drop adjusted
xi: reg sbp i.dmgrp*cbmi age i.raceth poorfair
quietly adjust age _Iraceth_2 _Iraceth_3 poorfair, by(dmgrp cbmi) gen(adjusted)
twoway (line adjusted bmi if dmgrp == 1, sort lpattern(solid))
      (line adjusted bmi if dmgrp == 2, sort lpattern(longdash))
      (line adjusted bmi if dmgrp == 3, sort lpattern(shortdash)),
      legend(label(1 "Normal") label(2 "IFG") label(3 "DM") row(1))
      yscale(range(120 160)) ylabel(120(10)160) ytitle("Average SBP")
      xtitle("Body Mass Index (kg/m^2)")
```

*Stata notes:* The pre-command `capture` is a handy bit of programming. Suppose we implement this in a do-file, it takes several tries to de-bug the code, and so we have to keep dropping the new variable `adjusted` so we can try running it again. The problem is that Stata will stop if you try to drop a variable that is not there. The `capture` pre-command instructs Stata to ignore the resulting error message and keep on going. The `adjust` command makes a new variable `adjusted` that gives the predicted SBP for each possible combination of `dmgrp` and `bmi` (that is, one for each observation), but holding age, race/ethnicity, and poor/fair self-reported health at their sample average values. In effect, `adjust` fiddles with the intercept for each of the three group-specific lines to remove confounding by these three factors. Note that we use `cbmi` in the `reg` and `adjust` commands, but `bmi` in the plot commands; this is done to make the x-axis interpretable, and works because both versions of BMI are on every record in the dataset. The `quietly` pre-command keeps it from sending page after page of output to the screen. The `twoway` command plots the predicted SBP values against BMI for each of the three glucose metabolism groups, then controls what the legend, y-axis, the axis titles look like. Please check help or the manual for more details.

## 4 Optional: alternative coding of interaction

Use this code to create new variables and run the model. *Note: This word processor doesn't accurately represent the single quotes in the forvalues loop below. To type the opening single quote, use the key at the upper left of the keyboard, just below the escape key, and for the closing single quote use the key just to the left of the return key.* If you are running Version 10, add `xi:` to the beginning of the `reg` command.

```
* alternative coding of model
forvalues i = 1/3 {
    gen cbmidmgrp_`i' = cbmi*(dmgrp=='i')
}
reg sbp i.dmgrp cbmidmgrp*
testparm cbmidmgrp*, equal
```

*STATA Notes:* The `forvalues` loop evaluates the code between the brackets for values of `i` from 1 to 3, sequentially plugging in 1, 2, or 3 at each appearance of `'i'` in the code. The logical expression within parentheses (`dmgrp=='i'`) is evaluated to 1 if it is true and to 0 otherwise. Thus `cbmidmgrp_1` is set equal to `cbmi` for participants in the normal group (`dmgrp == 1`) and to zero for the other two groups, and similarly for `cbmidmgrp_2` and `cbmidmgrp_3` and the IFG and diabetes groups respectively. Note that STATA doesn't allow us to name variables using the Version 11 conventions for the indicators and interactions it sets up automatically within regression commands (i.e., no leading digits or periods allowed).

1. Interpret the coefficient estimates for `cbmidmgrp_1`, `cbmidmgrp_2`, and `cbmidmgrp_3`.
2. Is the interaction test result the same? Why do we need the `equal` option in the `testparm` statement? This was not part of the `testparm` command used with the other version of the model to evaluate interaction.

Here is the regression equation for the model

$$ESBP|\mathbf{x}] = \beta_0 + \beta_1 2.dmgrp + \beta_2 3.dmgrp + \beta_3 cbmidmgrp_1 + \beta_4 cbmidmgrp_2 + \beta_5 cbmidmgrp_3$$

3. Use the following table to figure out the `lincom` command to compute the effect of a 5-unit increase in BMI in the diabetes group.

Group	BMI	_cons	2.dmgrp	3.dmgrp	cbmidmgrp_1	cbmidmgrp_2	cbmidmgrp_3
Diabetic	$k + 5$						
Diabetic	$k$						
	Difference						

4. Figure out the `lincom` command to compare diabetes to IFG when  $BMI = 30$ , recalling that the group-specific slope terms are defined in terms of centered BMI.

Group	BMI	_cons	2.dmgrp	3.dmgrp	cbmidmgrp_1	cbmidmgrp_2	cbmidmgrp_3
Diabetic	30						
IFG	30						
	Difference						