

## Biostat 208 Lab #4, 1/28/10

The purpose of this lab is to give you practice in using the `regress` command in STATA to look at confounding and mediation, as well as interpreting regression results with log-transformed outcomes and predictors. *Although we didn't cover log-transformation on Tuesday, all the code you need is given below, the comments for this lab interpret the results, and we will go over this in lecture next week.*

Mike Shlipak at the VA published a paper on renal insufficiency as a predictor of cardiovascular events in HERS (Shlipak MG, *et al.* Renal insufficiency and cardiovascular events in postmenopausal women with coronary heart disease, *J Amer Coll Cardiol*, 2001;**38**:705-11). In that paper, creatinine was classified in three categories (< 1.2 mg/dl, 1.2-1.4 mg/dl, and > 1.4 mg/dl), because the distribution has a long right tail. We will deal with the right skewness by log-transforming the creatinine measurement (more about this in Session 6). We will look at the association of BMI with log-creatinine levels, first unadjusted, then adjusting for several demographic and lifestyle confounders, and finally adjusting for triglyceride (TG) levels, which might be seen as a mediator of the association of BMI with creatinine levels.

Download the dataset `lab4.dta` from the course website. The variables and values in the Stata dataset are already labeled. Some new variables will be needed. Here is the code to make them.

```
* indicator for being relatively inactive
recode physact 2=1 3/5=0, gen(lessactive)
label variable lessactive "less active than peers"
label values lessactive noyes
tab physact lessactive
* three-level alcohol use variable
recode drnkspwk 0.2/4.99999 = 1 5/max = 2, gen(drinkamt)
label variable drinkamt "alcohol consumption"
label define drinklabel 0 "none" 1 "<5 drinks/week" 2 ">= 5 drinks/week"
label values drinkamt drinklabel
tab drnkspwk drinkamt
* natural log of triglyceride level
gen lntg = log(tgl)
```

*STATA Notes:* The `gen` option in the `recode` command is a convenient way to make a new categorical variable from an existing one. The `tab` commands are for checking that the new variables are correct. The value label `noyes`, suitable for indicator variables, is already part of the dataset; you can see all of them using the command `label list`. We'll use the log-transformation of TG to give you practice with a log-transformed predictor; TG over 300 mg/dL was an exclusion criterion in HERS so it is not too badly right skewed. The function `log()` computes the natural log; you have to use `log10()` to get the base-10 log.

## 1 Confounding

We will use two regression models to estimate the unadjusted and adjusted associations of BMI with log-transformed creatinine. When the outcome is natural log-transformed, the exponentiated regression coefficients can be interpreted as the approximate relative increase in the average value of the *untransformed* outcome per unit increase in the predictor. We can get the exponentiated coefficients using the option `eform("exp(beta)")`; the quoted string tells STATA how to label the column of coefficients in the `regress` output. The code to run the two models follows. Also included is code to compute the relative and percent increases in creatinine associated with 5-unit increases in

BMI. If you are running STATA Version 11, you can omit the `xi:` prefix at the beginning of regression commands with a categorical predictor of the form `i.varname`.

```
* unadjusted model for crude association of BMI with log-creatinine levels
reg lncreat bmi, eform("exp(beta)")
* Relative increase in creatinine associated with 5-unit increase in BMI
lincom bmi*5, eform
* Percent increase in creatinine associated with 5-unit increase in BMI
nlcom 100*(exp(_b[bmi]*5)-1)
```

Now fit the adjusted model. This is Model 2 in the sequence of models for evaluating mediation below.

```
* Model 2
xi: regress lncreat bmi age i.raceth educyrs i.drinkamt lessactive, eform("exp(beta)")
* Store estimates for testing using suest command
estimates store m2
* Store adjusted coefficient for BMI for computing PTE
scalar b_overall = _b[bmi]
* Relative increase in creatinine associated with 5-unit increase in BMI
lincom bmi*5, eform
* Percent increase in creatinine associated with 5-unit increase in BMI
nlcom 100*(exp(_b[bmi]*5)-1)
```

*STATA Notes:* After running a regression model, the coefficient estimates can be stored using the `estimates store` command, and can be accessed using the syntax `_b[varname]`. In addition, they can be stored for later use using the `scalar` pre-command, as in the code `scalar b_overall = _b[bmi]`. Commands like these operate on the coefficient estimates from the most recently estimated model.

1. *What is the interpretation of the unadjusted and adjusted exponentiated regression coefficients for BMI, also the `lincom` and `nlcom` results?*

## 2 Mediation

Does TG mediate part or all of the effect of BMI on creatinine levels? To assess this, we fit three models

1. regress mediator on primary predictor and confounders
2. regress outcome on primary predictor and confounders
3. regress outcome on primary predictor, mediator, and confounders

and finally compare adjusted coefficient for primary predictor in Models 2 and 3. We already estimated Model 2 in the first part of the lab. In Model 1, we regress log-transformed TG levels on BMI as well as the demographic and lifestyle confounders.

```
* Model 1
xi: regress lntg bmi age i.raceth educyrs i.drinkamt lessactive, eform("exp(beta)")
* Percent increase in TG for a 1-kg/m2 increase in BMI
nlcom 100*(exp(_b[bmi])-1)
```

1. *Is BMI an independent predictor of TG levels? Or in other words, is there evidence that the primary predictor affects the proposed mediator?*

Now we will estimate Model 3 by adding log TG to Model 2. Then we will compute the relative and percent increases in creatinine for a 25% increase in triglycerides, rather than the unrealistically large 2.7-fold increase that is displayed by default in the `regress` output. Finally we will test for the difference between the estimates of the overall and direct BMI effects, and assess the attenuation of the coefficient for BMI as a measure of the proportion of its adjusted effect that is mediated by TG levels.

```
* Model 3
xi: reg lncr bmi age i.raceth educ i.drinkamt lessact lntg, eform("exp(beta)")
* Store estimates for testing for mediation of BMI effects
estimates store m3
* store coefficient for BMI adjusted for confounders and mediator
scalar b_direct = _b[bmi]
* Relative increase in creatinine for a 25% increase in triglycerides
display log(1.25)
lincom lntg*.22314355, eform
* Percent increase in creatinine for a 25% increase in triglycerides
nlcom 100*(exp(_b[lntg]*log(1.25))-1)
* Check for equality of BMI effects before and after adjustment for TG levels
suest m2 m3
test [m2_mean]bmi = [m3_mean]bmi
* Percentage of adjusted association of BMI with creatinine explained by triglycerides
scalar pte = (b_overall-b_direct)/b_overall*100
display round(pte, .1)
```

*STATA Notes:* I truncated some of the variable names in the `regress` commands, which is allowed in STATA, so the `regress` command for Model 3 would fit on a single line. *In a do-file*, you can break a long line of code using `///`, as in the following

```
xi: regress lncreat bmi age i.raceth educ i.drinkamt lessactive lntg, ///
    eform("exp(beta)")
```

but this will not work on the command line.

2. Are TG levels an independent predictor of creatinine levels? What is the interpretation of the estimate for `lntg` in the `regress` output? And in the `lincom` and `nlcom` results?

3. What is the interpretation of the BMI effect after adjustment for TG? Can we rule out chance as an explanation for the difference between the overall and direct effects of BMI on creatinine? What percentage of the BMI effect on creatinine is explained by TG levels?