

Comments on Biostatistics 208 Lab #3 1/23/20

The comments below refer to the accompanying Stata log for the lab on the course website. A do-file for the lab is posted as well.

From the results for the first, unadjusted model, the `lincom` commands run using the `for-loop` or the `margins` command reproduce what we got using `tab, sum`. There is strong evidence for heterogeneity in mean BMI across levels of physical activity ($F(4, 2753) = 49.01, p < 0.0001$), either from the first `testparm` result or the model F -test results (since the initial model includes no predictors). The declining pattern in the group-specific means constitutes a statistically significant linear trend ($F = 134.1, p < 0.0001$ from the first `contrast` command result). However, the fact that the observed decrease is more pronounced at higher levels of activity also indicates some evidence for a departure from linearity in the pattern. This is born out by the results of the second `contrast` command.

Although age and smoking are also statistically significant predictors of BMI, the adjusted results for `physact` are fairly similar to the unadjusted results, suggesting that these variables don't pose a strong confounding influence.

- The intercept is an estimate of mean BMI for “much less active” women of age zero who do not smoke or drink – not directly interpretable, but also not of much interest. The coefficient for the 3rd level of `physact` estimates the adjusted difference in mean BMI between the “about as active” and “much less active” (reference) groups, holding the other predictors constant (i.e. at any fixed value). The interpretation of the coefficients for the other levels is similar. The coefficient for `age10` estimates the increase in mean BMI for every 10 years increase in age, while that for `drnkspwk` the change in mean BMI for a one-drink per-week increase in consumption – in both cases, accounting for the effects of the other variables.

- We still find the non-monotone pattern in the adjusted means, and strong evidence for heterogeneity and trend. Note that the adjusted means obtained by the two different `margins` commands are estimated by the model predictions of the average BMI for groups defined by physical activity level. For the first version of the `margins` command, the marginal means are obtained for the entire sample, assuming that each individual was “assigned” a specified activity level, and leaving all the other predictors at their observed values. This represents an adjusted marginal mean. For the second (with the `atmeans` option), the other predictors used for these estimates are set to their respective average values, so the resulting group means reflect average BMI for each activity level for individuals with average values of the remaining predictors. You'll notice that the results are the same. This is expected in strictly linear models, but we'll see when we look at inherently nonlinear models like logistic regression that `margins` will behave somewhat differently.

- You can easily verify that the regression coefficients for the four included `physact` indicator variables are exactly equal to the differences in the adjusted means between the “much less active” reference group and the other four physical activity groups given in the either version of the `margins` command. You can check this by cutting and pasting means using the `display` command, or use an option to `margins` as follows:

```
margins, dydx(physact)
```

A number of questions were raised about the `margins` command in lab. We used it as a way of getting estimated means for `bmi` in the different physical activity groups. If you want to see what `margins` is doing, try the following two sets of commands. The first uses the coefficients extracted from the fitted regression model including the model intercept, and the coefficients for `physact`, `age10`, `smoking` and `drnkspwk` as predictors.

```
gen pr1 = _b[_cons] + _b[age10]*age10 + _b[smoking]*smoking + _b[drnkspwk]*drnkspwk
gen pr2 = _b[_cons] + _b[2.physact] + _b[age10]*age10 + _b[smoking]*smoking + _b[drnkspwk]*drnkspwk
gen pr3 = _b[_cons] + _b[3.physact] + _b[age10]*age10 + _b[smoking]*smoking + _b[drnkspwk]*drnkspwk
gen pr4 = _b[_cons] + _b[4.physact] + _b[age10]*age10 + _b[smoking]*smoking + _b[drnkspwk]*drnkspwk
gen pr5 = _b[_cons] + _b[5.physact] + _b[age10]*age10 + _b[smoking]*smoking + _b[drnkspwk]*drnkspwk
summ pr1-pr5
```

These commands generate predicted mean BMI for each individual in the dataset, for each of the 5 physical activity groups. Note that each command fixes the level of physical activity. The other predictors are allowed to take on their observed values in the sample. In essence, each of the variables `pr1-pr5` created represents a “population” of individual BMI values, where the level of physical activity is fixed, and the other variables are allowed to vary. The average of these new variables represent the marginal means. The last command asks for the marginal means for the five groups. You’ll notice that these results correspond to those from the `margins` command.

A feature of the linear model is that if we do the same exercise, but fix the values of the other predictors at their mean level, we get the same results (i.e. the averaging across observations yields the mean value for each of the predictors allowed to vary). The following commands illustrate this, using the `egen` command to first generate new variables containing the mean values for each of the additional predictors.

```
egen m_age10 = mean(age10) if e(sample)
egen m_smoking = mean(smoking) if e(sample)
egen m_drnkspwk = mean(drnkspwk) if e(sample)
gen z1 = _b[_cons] + _b[1.physact]*1 + _b[age10]*m_age10 + _b[smoking]*m_smoking + _b[drnkspwk]*m_drnkspwk
if e(sample)
gen z2 = _b[_cons] + _b[2.physact]*1 + _b[age10]*m_age10 + _b[smoking]*m_smoking + _b[drnkspwk]*m_drnkspwk
if e(sample)
gen z3 = _b[_cons] + _b[3.physact]*1 + _b[age10]*m_age10 + _b[smoking]*m_smoking + _b[drnkspwk]*m_drnkspwk
if e(sample)
gen z4 = _b[_cons] + _b[4.physact]*1 + _b[age10]*m_age10 + _b[smoking]*m_smoking + _b[drnkspwk]*m_drnkspwk
if e(sample)
gen z5 = _b[_cons] + _b[5.physact]*1 + _b[age10]*m_age10 + _b[smoking]*m_smoking + _b[drnkspwk]*m_drnkspwk
if e(sample)
summ z1 z2 z3 z4 z5
```

The results are the same when we fix the predictors at their mean values.

To conclude, the `margins` command allows us to obtain marginal group-specific means for a selected predictor, averaging over the other predictors. These group-specific averages are *adjusted* for the presence of the other predictors in the model. Although not required to get the adjusted means in this example, this command is very useful in interpreting interactions and in estimating marginal effects in binary regression models.