Biostatistics 208 HW #3, Due 2/28/17

Please email your completed assignment to biostat208hw@gmail.com by 10:30 AM on Tuesday, Feb 28, 2017.

The exercise concerns the association between central obesity, as measured by waist-hip ratio (WHR), and triglyceride (TGL) levels. Both are components of the so-called metabolic syndrome and precursors of diabetes and cardiovascular disease. Prepare a Word file with answers to each of the following questions. Paste relevant Stata output and graphs into the file to illustrate your points. In particular, the regression output and diagnostic plots on which you base your conclusions should be included. For this homework you will need `hwk3.dta`, available on the course web site. The dataset includes observations for a 10% random sub-sample of the HERS cohort.

Checking model assumptions

Fit a model regressing TGL on age, race/ethnicity, alcohol use, exercise, WHR, and BMI, and obtain the residuals. Treat BMI as categorical, using categories ≤25, 25^{+}-30, 30^{+}-35, and >35.

1. TGL is often log-transformed because it usually has a long right tail. That may not hold in the HERS data, because women with baseline TGL > 300 *mg/dL* were excluded. Check the residuals from the initial model for TGL for normality, then re-run your multivariable analysis using the natural log of TGL as the outcome, and recheck the normality of the residuals from the new model.

(a) Interpret the coefficients for WHR both before and after log-transformation of TGL. Note that the range of WHR is only 0.64 to 1.14, so a 1-unit change is uninterpretable and involves extrapolation way beyond the range of the data. So compare the predicted increases in TGL for an increase in WHR of only 0.05 or 0.1, rather than an increase of 1 unit.

(b) Are your substantive conclusions affected by the transformation of the outcome?

(c) Which version of the outcome you would choose for the final analysis, and why?

2. Check the linearity of the association of WHR with TGL in the model adjusting for age, race/ethnicity, alcohol use, exercise, and BMI, categorized as above.

(a) Does log transforming WHR fix the linearity problem?

(b) Categorization of WHR would almost surely work, and also be relatively easy to interpret. Re-fit the model controlling for age, race/ethnicity, alcohol use, exercise, and BMI (again categorized), with WHR categorized ≤0.8, 0.8^{+}-0.85, 0.85^{+}-0.90, and >0.90. Interpret the coefficients for the WHR variables. Is there heterogeneity in mean TGL across the four levels of WHR? (To look at the adjusted mean TGL levels by WHR category (`i.whrcat`), use the command "`margins whrcat`" after fitting the model). Are the second and third categories statistically significantly different from each other? What about the third and fourth?

3. Check for influential points in the model for TGL treating WHR and BMI as categorical, using a boxplot of the `dfbeta` statistics. What cutoff would you use, and do you find sensitivity of the regression results to the removal of data points that exceed it?

4. Suppose we were assessing the potential effects of exercise on TGL levels. Check for covariate overlap between participants who do and do not exercise, and summarize your findings.

5. *Extra credit*: Run the final model replacing the categorical versions of WHR and BMI with linear splines, using the same cutpoints, and interpret the regression coefficients for the WHR variables. Plot the regression line for WHR along with a LOWESS smooth.