

**HOMEWORK 3****Due in class on 2/23/10**

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. Using size 8 Monaco font for the STATA output keeps it relatively small and properly aligned. 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.

**1 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  $\leq 25$ , 25-30, 30-35, and  $> 35$ . The rationale for controlling for BMI is explored in Part 2.

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 rerun 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 transformation fix any linearity problem with WHR?
  - (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  $\leq 0.8$ , 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? Do the second and third categories differ from each other? Do 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.
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.

## 2 Selecting a model

Suppose we want to understand the causal connections of central obesity, as measured by WHR, with TGL levels. We can think of this as an instance of the second inferential goal – assessing WHR as a predictor of primary interest.

1. Consider first evaluating the *overall independent influence* of WHR on TGL levels.
  - (a) Distinguish potential confounders from mediators in terms of their causal relationship to central obesity, the predictor of primary interest in this analysis.
  - (b) Is it appropriate to include confounders of WHR in the model evaluating its overall independent effect on TGL?
  - (c) Is it appropriate to include mediators of WHR in the model evaluating its overall independent effect on TGL?
  - (d) If mediators of WHR are added to the model, what is the causal interpretation of the resulting coefficient estimates for WHR?
2. Suppose we regard BMI as a confounder of WHR. What is your interpretation of the BMI results in the model we estimated in Section 1, using categorical versions of both WHR and BMI?
3. Reviewers might ask whether BMI and WHR are co-linear, calling into question the validity of your findings for WHR. How would you respond?
4. The model with categorical WHR and BMI from Section 1 showed that fitted mean levels of TGL follow a somewhat complicated pattern across the four categories of WHR.
  - (a) Is there compelling evidence that the means in the upper three categories are unequal? (Hint: use `testparm _Iwhr*, equal` in Version 10 or `testparm i(2/4).whrcat, equal` in Version 11).
  - (b) Can you think of a biological explanation for this pattern?
  - (c) Suppose you had proposed this 4-level categorization *a priori*. On the basis of these results, would you simplify the model by combining the upper three categories? Why or why not?
5. This is a moderately small ( $N = 276$ ) dataset. I included age, alcohol use, and exercise in the initial model for “face validity.” Should any of these be dropped, in view of the sample size and model results?
6. Should the model also control for diabetes? Motivate your argument using a DAG.
7. Suppose our final model violated the rule of thumb of 10 observations per predictor (or 10 events per predictor with binary or survival outcomes). What are some diagnostics you could use to check whether you have too many predictors in the model?