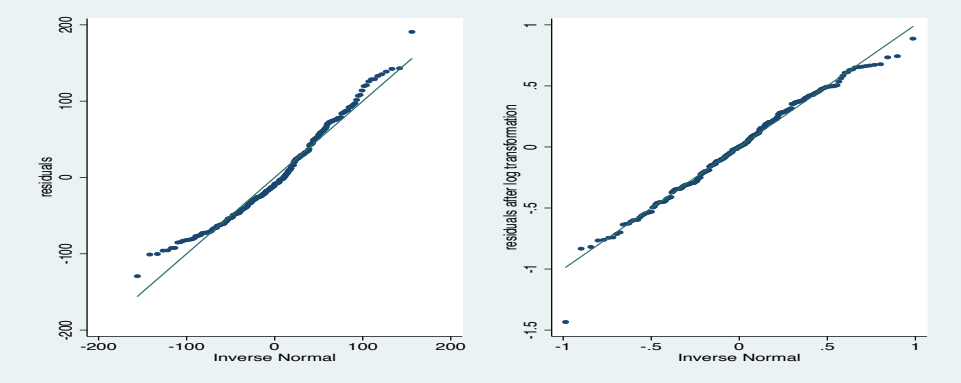
Biostatistics 208 **Winter 2018**

Comments on Homework 3

1 Checking assumptions

1. *TGL is often log-transformed because it 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.*

   The QQ-plots show that the log-tranformation gets rid of the slight right skewness in the residuals, but with a sample with 276 observations, the skewness is not bad enough to cause trouble.



   For what it is worth, here is the code to produce the side-by-side plots.

```
set graphics off
qnorm resid, msize(small) ytitle("residuals") name(qq1, replace)
qnorm lnresid, msize(small) ytitle("residuals after log transformation") name(qq2, replace)
set graphics on
graph combine qq1 qq2
```

   (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.*

   Using the rescaled variable `whr10=whr/0.1`, the model estimates that for a 0.1 unit increase in WHR, average TGL increases by 12.5 mg/dL. With the log-transformed TGL, I found a 1.09-fold (8.7%) increase in mean TGL per 0.1 increase in WHR. Starting from the mean TGL value of 164 mg/dL this would equivalent to an increase 14.2 mg/dL. The next page shows how I got the results with log-transformed TGL, using the `eform("exp(beta)")` option in fitting the model.

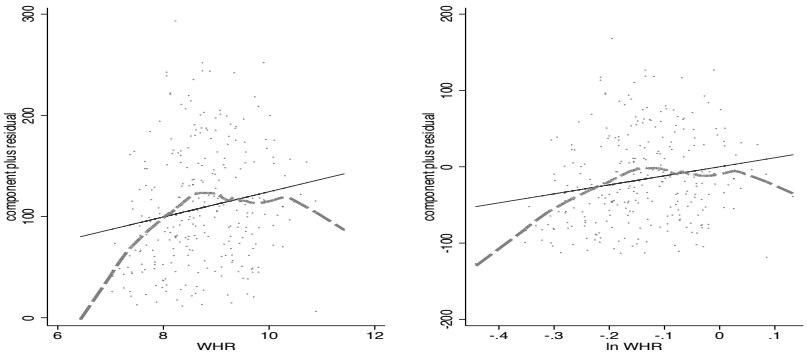
```
. reg lntgl age i.raceth exercise drinkany whr10 i.bmicat, eform("exp(beta)")

      Source |       SS       df       MS              Number of obs =     274
-------------+------------------------------           F(  9,   264) =    4.24
       Model |  5.32159755     9  .591288617            Prob > F      =  0.0000
    Residual |  36.8573086   264  .139611017            R-squared     =  0.1262
-------------+------------------------------           Adj R-squared =  0.0964
       Total |  42.1789061   273  .154501488            Root MSE      =  .37365

------------------------------------------------------------------------------
       lntgl |  exp(beta)   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  ....
       whr10 |   1.086568   .0311233     2.90   0.004     1.026982    1.149611
  ....
------------------------------------------------------------------------------
. * percent increase in TGL for a 0.1 increase in WHR
. nlcom 100*(exp(_b[whr10])-1)
------------------------------------------------------------------------------
       lntgl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |   8.656793   3.112332     2.78   0.006      2.52864    14.78495
------------------------------------------------------------------------------
. * absolute increase in TGL starting from the mean of 164
. nlcom 164*(exp(_b[whr10])-1)
------------------------------------------------------------------------------
       lntgl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |   14.19714   5.104225     2.78   0.006      4.14697    24.24731
------------------------------------------------------------------------------
```

(b) *Are your substantive conclusions affected by the outcome transformation?*
Both models show that the adjusted association of WHR with TGL is highly statistically significant, and of similar magnitude. Clearly, the models can't mean exactly the same thing, because the first posits the same *absolute* increase in mean TGL for every unit increase in each predictor, while the second model, using the log-transformed outcome, posits the same *relative* increase. That being said, my take is that the substantive conclusions are unchanged by the transformation.

(c) *Which version of the outcome you would choose for the final analysis, and why?*
From the point of view of normality of the residuals and performance of p-values and CIs, both versions are fine. R^2 is a bit higher for the log-transformed outcome, and log-transformation might have more face validity, because TGL is usually right skewed, though not in this case. On the other hand, interpretation of the coefficient estimates is more straightforward using the untransformed outcome. So either version of the outome would be reasonable to use.

2. *Check the linearity of the associations of WHR with TGL in the model adjusting for age, race/ethnicity, alcohol use, and exercise.*

   (a) *Does log transformation fix the linearity problem with WHR?*



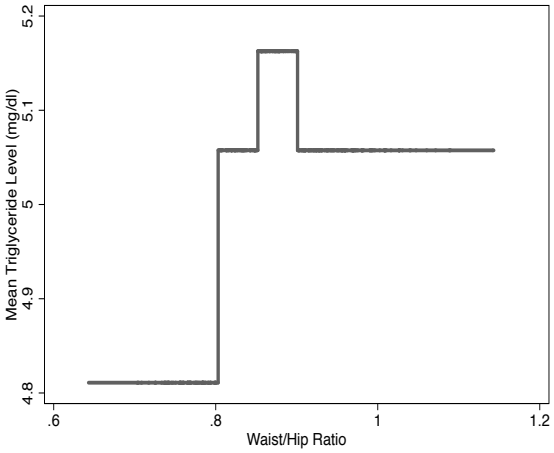
   My interpretation of the CPR plots is that log transformation does not adequately fix the problem with WHR. The dashed LOWESS line in the right hand panel for log-transformed WHR retains considerable curvature. A similar problem arises with BMI (data not shown), which was the motivation for categorizing it.

   (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?*

   The regression output for WHR shows that after adjusting for BMI and the other covariates in the model, mean TGL is lowest in women with WHR less than 0.8 (the reference group in the model), substantially elevated in women with WHR between 0.80 and 0.85 (category 2), highest among women with WHR between 0.85 and 0.90 (category 3), and about the same among women with WHR above 0.90 (category 4) as among those with WHR between 0.80 and 0.85. The regression coefficients for variables `2.whrcat`, `3.whrcat`, and `4.whrcat`, give the adjusted differences in mean TGL in the second, third, and fourth categories with respect to the reference category. The `testparm` result shows that the heterogeneity across the four categories is highly statistically significant, and the p-values for the three indicators show that each of the higher WHR categories differs from the reference category. The `lincom` results on the next page show that pairwise differences between categories 2 and 3 and categories 3 and 4 are borderline statistically significant. If you did this analysis using log-TGL as the outcome, which is perfectly legitimate, the evidence for overall heterogeneity across the four levels remains strong, but evidence for heterogeneity among the upper three levels in pairwise comparisons is weaker. That probably results from downweighting some high TGL values in the third WHR category.

```
. recode whr min/.8 = 1 .80001/.85 = 2 .85001/.9 = 3 .90001/max = 4, gen(whrcat)
. reg tgl age i.raceth exercise drinkany i.whrcat i.bmicat
 .....
 ------------------------------------------------------------------------------
         tgl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  .....
      whrcat |
          2  |   37.46069   10.94904     3.42   0.001     15.90138       59.02
          3  |   56.70995   11.24566     5.04   0.000     34.56658    78.85332
          4  |   37.16786    10.2245     3.64   0.000     17.03522    57.30051
 .....
------------------------------------------------------------------------------
. testparm i.whrcat
  ....
       F(  3,   262) =    8.81
            Prob > F =    0.0000

. lincom 3.whrcat - 2.whrcat
------------------------------------------------------------------------------
         tgl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   19.24925   10.76934     1.79   0.075    -1.956217    40.45473
------------------------------------------------------------------------------

. lincom 4.whrcat - 3.whrcat
------------------------------------------------------------------------------
         tgl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -19.54208   9.666057    -2.02   0.044    -38.57513   -.5090403
------------------------------------------------------------------------------

. * Obtain and plot the adjusted means by WHR category
. margins whrcat
Predictive margins                              Number of obs   =        274
Model VCE    : OLS
Expression   : Linear prediction, predict()
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      whrcat |
          1  |   130.7279   8.029682    16.28   0.000       114.99    146.4658
          2  |   168.1886   7.574942    22.20   0.000      153.342    183.0352
          3  |   187.4378   7.626962    24.58   0.000     172.4893    202.3864
          4  |   167.8957   5.888192    28.51   0.000     156.3551    179.4364
------------------------------------------------------------------------------
. qui gen catfit =  .
. forvalues i = 1/4 {
  2.         qui replace catfit = el(r(b), 1, `i') if whrcat==`i'
  3.         }
. twoway (line catfit whr, c(J) sort lw(thick)), ///
>     ytitle("Mean Triglyceride Level (mg/dl)") ///
>     xtitle("Waist/Hip Ratio") legend(off)
```

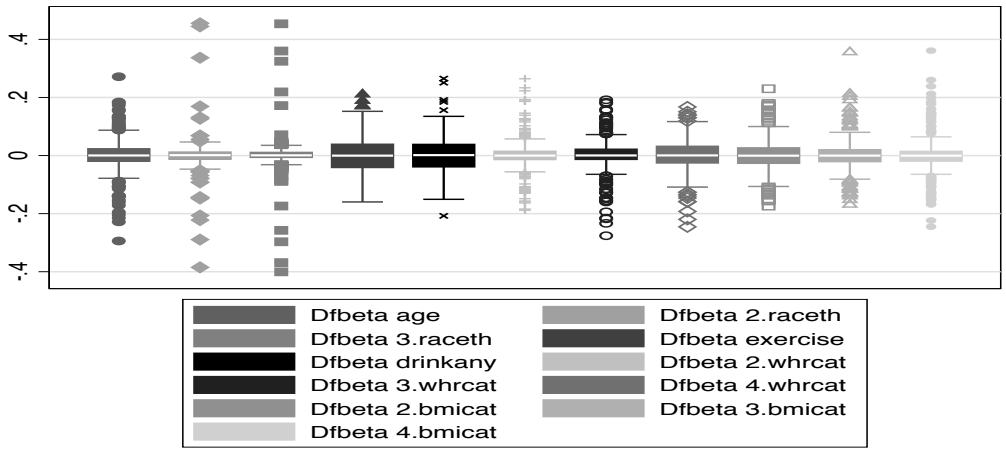
The `margins` command provides adjusted estimates of mean TGL in each of the four WHR categories. The estimates are returned by the `margins command` in the matrix `r(b)`, which has one row and four columns, one for each of the four means, in the same order as they appear in the output. (To see what else `margins` returns in usable form, type `return list`.) We can extract the elements of this matrix using the `el` matrix operator, obtaining the adjusted means for each category of `whrcat`. Finally, we plot the fitted means against the observed WHR values. In the Extra Credit question, we'll see how to do this for the effect of a continuous predictor using the `adjust` command instead. Here is the resulting plot of the adjusted mean TGL levels in the four categories.



Note that the downturn in mean TGL among women in the highest WHR category can't be modeled using the log transformation of WHR, but can be modeled using the quadratic, categorical, or linear spline model used for the extra credit problem. More later about evaluation and interpretation of this non-linear pattern.

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?*

   I chose a cutoff of 0.35 on the basis of the boxplot of the `dfbeta`s for all variables below, although more conservative cutoffs would also be reasonable. The sensitivity analysis omitting the 9 observations with `dfbeta`s larger than 0.35 in absolute value gives very similar regression coefficient estimates for WHR, as shown below. If the more conservative cutoff of 0.3 is used, an additional two observations are dropped from the sensitivity analysis, but the conclusions remain unchanged.



Here are the regression results, first with all the data:

```
. reg tgl age i.raceth exercise drinkany i.whrcat i.bmicat
      Source |       SS       df       MS              Number of obs =     274
-------------+------------------------------           F( 11,   262) =    5.08
       Model |  183439.096    11  16676.2814           Prob > F      =  0.0000
    Residual |  860110.087   262  3282.86293           R-squared     =  0.1758
-------------+------------------------------           Adj R-squared =  0.1412
       Total |  1043549.18   273  3822.52448           Root MSE      =  57.296
------------------------------------------------------------------------------
         tgl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .4970569   .5547607     0.90   0.371    -.5952999    1.589414
      raceth |
           2 |  -29.23307   12.71086    -2.30   0.022    -54.26151   -4.204641
           3 |   29.84161     16.008     1.86   0.063    -1.679102    61.36231
    exercise |  -11.22049   7.257886    -1.55   0.123     -25.5117    3.070719
    drinkany |   3.309081   7.404534     0.45   0.655    -11.27089    17.88905
      whrcat |
           2 |   37.46069   10.94904     3.42   0.001     15.90138       59.02
           3 |   56.70995   11.24566     5.04   0.000     34.56658    78.85332
           4 |   37.16786    10.2245     3.64   0.000     17.03522    57.30051
      bmicat |
           2 |   13.53463   9.038234     1.50   0.135    -4.262193    31.33145
           3 |   10.65932   10.54745     1.01   0.313    -10.10924    31.42787
           4 |   24.01831   12.01034     2.00   0.047     .3692297    47.66739
       _cons |   91.25666   39.35489     2.32   0.021     13.76453    168.7488
-------------+----------------------------------------------------------------
```

Here is some code for computing the largest DFbeta for any of the variables,which is then used to omit the observations with big `dfbetas` from the sensitivity analysis (if you try this, please recall the misrepresentation of the single quotes with this word processor).

```
gen maxDF = 0
forvalues i = 1/11 {
    replace maxDF = abs(_dfbeta_‘i’) if abs(_dfbeta_‘i’) > maxDF & _dfbeta_‘i’ ~= .
    }
```

Next the regression results omitting the nine influential observations

```
. reg tgl age i.raceth exercise drinkany i.whrcat i.bmicat if maxDF < 0.35

      Source |       SS       df       MS              Number of obs =     265
-------------+------------------------------           F( 11,   253) =    5.70
       Model |  182561.774    11  16596.5249           Prob > F      =  0.0000
    Residual |   736922.43   253  2912.73688           R-squared     =  0.1985
-------------+------------------------------           Adj R-squared =  0.1637
       Total |  919484.204   264  3482.89471           Root MSE      =   53.97

------------------------------------------------------------------------------
         tgl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .7487803   .5290437     1.42   0.158    -.2931102    1.790671
      raceth |
           2 |  -33.65153   12.80885    -2.63   0.009    -58.87707   -8.425978
           3 |   31.41343   17.68853     1.78   0.077    -3.422096    66.24895
    exercise |  -11.25907   6.939074    -1.62   0.106    -24.92477    2.406638
    drinkany |  -3.806741   7.136157    -0.53   0.594    -17.86058     10.2471
      whrcat |
           2 |   35.43876   10.47368     3.38   0.001     14.81206    56.06546
           3 |   59.10742   10.70219     5.52   0.000      38.0307    80.18415
           4 |   38.79429   9.706993     4.00   0.000     19.67748    57.91109
      bmicat |
           2 |   13.04661   8.595198     1.52   0.130    -3.880647    29.97386
           3 |   8.412434   10.08702     0.83   0.405    -11.45278    28.27765
           4 |   15.05868   11.51722     1.31   0.192    -7.623164    37.74053
       _cons |   77.07882   37.38671     2.06   0.040     3.450004    150.7076
------------------------------------------------------------------------------
```

Next, a comparison of the regression coefficients tabulated in a friendlier format.

```
              all data  omitting IPs    difference  % difference
      age          0.5           0.7           0.3          50.6
1b.raceth          0.0           0.0           0.0             .
 2.raceth        -29.2         -33.7          -4.4          15.1
 3.raceth         29.8          31.4           1.6           5.3
 exercise        -11.2         -11.3          -0.0           0.3
 drinkany          3.3          -3.8          -7.1        -215.0
1b.whrcat          0.0           0.0           0.0             .
 2.whrcat         37.5          35.4          -2.0          -5.4
 3.whrcat         56.7          59.1           2.4           4.2
 4.whrcat         37.2          38.8           1.6           4.4
1b.bmicat          0.0           0.0           0.0             .
 2.bmicat         13.5          13.0          -0.5          -3.6
 3.bmicat         10.7           8.4          -2.2         -21.1
 4.bmicat         24.0          15.1          -9.0         -37.3
```

The columns of the table below show the coefficient estimates for each variable using all the data, omitting the influential points (IPs) with at least one `dfbeta` > 0.35, and then lists the absolute and percentage difference between the two estimates. The changes for the WHR indicators are all pretty minor. The changes for BMI are larger, and weaken the evidence for a direct effect after adjustment for WHR. The coefficients for `age` and

`drinkany` undergo big percentage changes, but both are close to zero in the full-data model, and statistically significant in neither model. Stata code to produce the table is in the posted do-file.

It would also be reasonable just to focus on points that are influential for WHR and possibly BMI, or only WHR, since they are the focus of the research question. If we just omit observations with a `dfbeta` > 0.25 for either of those two factors, then four observations are omitted; here are the tabular results:

	all data	omitting IPs	difference	% difference
1b.whrcat	0.0	0.0	0.0	.
2.whrcat	37.5	30.8	-6.7	-17.8
3.whrcat	56.7	62.0	5.3	9.4
4.whrcat	37.2	39.3	2.1	5.7
1b.bmicat	0.0	0.0	0.0	.
2.bmicat	13.5	10.6	-2.9	-21.6
3.bmicat	10.7	4.3	-6.3	-59.4
4.bmicat	24.0	12.6	-11.5	-47.7

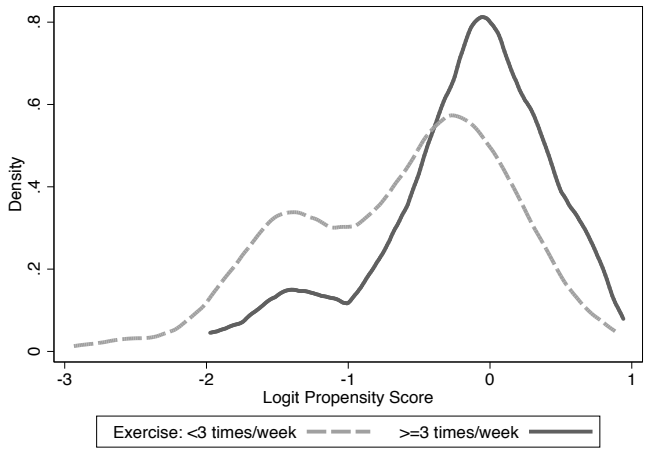
Although the percentage changes in the BMI coefficients are larger in this case, and the coefficient for the highest category is no longer statistically significant, most of the adjusted BMI effects are small and not significant before omitting the influential points. My take is that the results for WHR are not qualitatively changed but that evidence for a direct effect of BMI is weakened in this model, as in the sensitivity analysis using the `dfbeta`s for all included variables.

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.*

   In single variable checks (data not shown), average age and BMI differed by exercise, but their IQRs and ranges were similar, indicating decent overlap. Likewise, while race/ethnicity and prevalence of alcohol use and smoking also differed, African Americans and Latinas, as well as women reporting alcohol use and smoking, were represented in both groups. Next, an analysis of overlap using logit propensity scores.

```
. qui logistic exercise age i.raceth smoking drinkany i.whrcat i.bmicat
. predict logit_pscore, xb
. twoway (kdensity logit_pscore if exercise==1, lp(solid) lw(thick)) ///
>         (kdensity logit_pscore if exercise==0, lp(longdash) lw(thick)), ///
>         ytitle("Density") xtitle("Logit Propensity Score") ///
>         legend(order(2 "Exercise: <3 times/week" 1 ">=3 times/week") textfirst)
```

   The density plot of propensity scores indicates considerable lack of overlap on the left. So caution would be required in estimating the causal effect of exercise, which might be on solider ground if the analysis was run excluding women with very low propensity to exercise – say, with logit propensity scores less than -1.8. This restriction of inference would need to be made explicit in discussing the results.



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.*

```
. mkspline whr1 8 whr2 8.5 whr3 9 whr4 = whr10
. mkspline bmi1 20 bmi2 25 bmi3 30 bmi4 35 bmi5 = bmi
. xi: reg tgl age i.raceth exercise drinkany whr1-whr4 bmi1-bmi5
      Source |       SS       df       MS              Number of obs =     274
-------------+------------------------------           F( 14,   259) =    3.98
       Model |  184751.018    14  13196.5013           Prob > F      =  0.0000
    Residual |  858798.164   259  3315.82303           R-squared     =  0.1770
-------------+------------------------------           Adj R-squared =  0.1326
       Total |  1043549.18   273  3822.52448           Root MSE      =  57.583
------------------------------------------------------------------------------
         tgl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 .....
        whr1 |   53.29476   22.61797     2.36   0.019     8.756225    97.83329
        whr2 |   39.42009   31.33313     1.26   0.209    -22.28002    101.1202
        whr3 |   -2.04113   27.17024    -0.08   0.940    -55.54383    51.46157
        whr4 |  -12.90673   11.36346    -1.14   0.257    -35.28327    9.469818
  .....
------------------------------------------------------------------------------
```

In the regression output for the linear spline model, the coefficients for the spline variables `whr1-whr4` give the slope of the spline fit (i.e., the increase in average TGL for each increase of 0.1 in WHR) within each of the four regions defined by WHR: < 0.80, 0.80-0.85, 0.85-0.90, and > 0.90. I rescaled WHR before making the splines, so that the coefficient estimates would be directly interpretable. Note that the slopes in the regions above WHR of 0.8 are all NS (recall that the p-values are unaffected by the rescaling).

We can infer from the `testparm` results below that while the slope in the region below 0.8 differs from the rest, the slopes in the region above 0.8 are indistinguishable from each other. My interpretation is that TGL levels increase with WHR up to 0.80, and there is little evidence for any further increase above 0.85, but it is unclear exactly where in the region from 0.8 to 0.85 the increases cease.

```
. testparm whr1-whr4, equal
 ( 1)  - whr1 + whr2 = 0
 ( 2)  - whr1 + whr3 = 0
 ( 3)  - whr1 + whr4 = 0
       F(  3,   259) =    5.27
            Prob > F =    0.0015

. testparm whr2-whr4, equal
 ( 1)  - whr2 + whr3 = 0
 ( 2)  - whr2 + whr4 = 0
       F(  2,   259) =    1.56
            Prob > F =    0.2120
```

This is in contrast to the categorical model, which suggested differences between the three groups with WHR >0.8. So this is an example of findings being model dependent – not the last you are likely to encounter. Which result do you find more interpretable?

```
. capture drop lsfit
. qui xi: reg tgl age i.raceth exercise drinkany whr1-whr4 bmi1-bmi5
. capture drop lsfit
. qui adjust age _Iraceth* exercise drinkany bmi1-bmi5, gen(lsfit)
. twoway (scatter tgl whr, msize(vsmall)) ///
>     (line lsfit whr, sort lpattern(dash_dot) lwidth(medthick)) ///
>     (lowess tgl whr, sort lpattern(longdash) lwidth(medthick)), ///
>     ytitle("Triglyceride levels (mg/dL)") ///
>     legend(order(2 "Linear spline" 3 "LOWESS") rows(1))
```

As for the plot, the `adjust` command computes the expected value of the outcome for each observation, holding each of the predictors listed in the command constant at its sample mean. Because the spline variables `whr1-whr4` are *not* included in the `adjust` command, the expected values are calculated at the observed values of WHR – exactly what we want to plot. The expected values are saved as `lsfit` for each observation by the `gen()` option. Note we had to use `xi:` with `i.raceth`, because the `adjust` command does not work with the so-called `factor` variables that would have been made in Stata Versions 11-15 if I just included `i.raceth` in the `reg` command. The resulting plot shows that the linear spline and LOWESS fits track closely.