EPIB-609 (2012): **Assignment on Effect Modification** (v2012.11.18)

Questions A and B, plus Q3new and Q5 from the 2003 course (below)
See also the 678 Notes on the 609 Website.

The purposes of this exercise are to
(i) stress the "effect modification" and "product term" terminology over the "interaction" terminology (in the Extra Notes. See in particular the dictionary definition of interaction, and Miettinen's reasons why it does not do as good a job at conveying the number of variables involved (3). See also his example about love and time; and learn to speak about 'need for product terms' rather than interactions.
(ii) familiarize you with regression techniques for representing effect modification
(iii) provide opportunities to practice 'translating' the fitted equations back into words and pictures
(iv) distinguish 'removable' from 'non-removable' product terms, and the role of graphs to see where lines cross.

A. Read the newspaper item on Ideal Weight, but for the formula for men, use a 'starting' value (or as Clayton and Hills would say, a 'corner' value), of 110 pounds rather than 106. The reason for taking this poetic license is to ensure that all of the coefficients in the model have different values, and therefore are not confused one with the other.
* Write out the 2 sex-specific equations separately, and plot them on a single graph,
* Convert the 2 sex-specific equations into one master equation (M) that accommodates both sexes [1]
* Calculate where (at what height) the lines cross.
* Cf. Website for Miettinen panels for a covariate as a modifier/confounder/both. Which applies here?
* Suggest a (somewhat less accurate, but still reasonably adequate) unisex equation
* Rewrite M to highlight (i) the sex-specific slopes for height (ii) the height-specific sex difference
* Rewrite M so the 'corner' for height is located at 0 inches (i.e., at time of conception)
* Over the range 48-72 inches, how closely would a multiplicative model approximate M?
* Do you think ideal weight should be linear in height? Or some other function of height?
* How might you (a) model the pattern of average earnings by level of Education/Age/Sex?
  (b) simplify the equation linking windchill with temperature and windspeed? (cf. Website)

B. Scan the JAMA article on Broad-Spectrum Sunscreen Use and the Development of New Nevi in White Children, and then focus on Table 4 ("Parameter Estimates for Variables Predicting Number of New Nevi in Vancouver Schoolchildren")
* Overlay onto Figure 2 your own plot of the predicted (fitted) values for persons with 0%, 5%, 15%, 25%, 50%, and 75% coverage of freckles on the face, under the "sunscreen" and "control" scenarios. Since you are not given an intercept, 'reverse-engineer' it using the mix of ages, grades, sunlight exposure and sunburn score to age 5 in those studied. Also, for the fitted values, use the average profile with respect to these variates, or else a specified individual profile. Compare it with the ones plotted.
* Write a sentence or two explaining to the authors how they might have improved the presentation and user-friendliness of the information in Table 2.

Q3 new, and **Q5** : see below

E-mail preliminary answers before the beginning of class on November 26, 2012.

---

[1] Use the variate '`i.male`' ('*indicator*' of male, not a `*dummy* variable' for male) rather than the variate 'gender' or 'sex' whose coding you will have to write down somewhere and have to look up later. In some circumstances, and by preference, you might wish to have female be the index category, and male the reference category, but in this case, if you prefer to positive coefficients, you can see the convenience of making females the reference category, and male the index category. Either way, its good to get in the habit of naming the variate by the descriptor of the index category.

2012: 3 new, and 5

1 **β-blocker or β-stimulant?**

- Problem 3.7 G&S pages 106-107 [109-110 in edition 2]
(means and individual data -- "GS D8" on web page)

• Part A (using the 8 means as 8 datapoints)

• Part B (using the 64 datapoints from individuals)

In each part do the formal fitting using *one equation that yields 2 lines*.

• Also verify that you get the same two lines if you split the data into the treated and untreated and fit lines separately. [designate drug as a group variable before using **Fit**(Y|X)]

*For plotting purposes, identify the observations for the treated by selecting (or **Find**ing) them, then using **Edit->Window->Tools** to colour them. Unfortunately in INSIGHT you cannot show the fitted lines superimposed on the plot of the raw data -- the best you can do is click both the fitted (predicted) and actual data SBP points into the "Y" box in the Scatter Plot (Y X) dialog box, and norepinephrine into the X box, to get side by side plots.*

2 **Read through** Problems 3.8 to 3.17 in G&S and **identify which ones seem to call for models with product (interaction) terms i.e. assessment of effect modification.**

3 **new** *[three non-parallel lines .. and their interpretation .... question not as long as it seems! but questions were constructed in a hurry.. so e-mail me if parts are unclear]]*

2012

Water fluoridation, tooth decay in 5 year olds, and social deprivation
[Jones CM et al. BMJ 1997;315:514-517, 30 August: under Resources for session 6]

on 609 site

Figure 1 shows 3 separate lines, each with a different slope and different intercept

i Divide up the dataset and use your favourite regression software to fit 3 separate lines (data points measured by hand from figure by assistant, but not fully checked by JH)

In *INSIGHT* you can do this by designating [in the leftmost of the two rectangles above the variable name] district as a Group variable.. then Fit t_decay versus deprivation...

now called JMP, free for students

In the *Program Editor* window, you would do this using the statements

```
proc sort data=fl_depr; by district;

proc reg data=fl_depr; by district;
  model t_decay = deprived ;
run;
```

Rewrite the three separately fitted equations as one 'master equation', using deprivation, indicator variables for ALREADY and ADDED fluoride, and 2 products, i.e. deprivation with each of these two indicator variables, as 'regressor' variables (your equation should, including the b_0, comprise 6 terms.. in effect, 2 per separate line x 3 lines)

Water fluoridation, tooth decay in 5 year olds, and social deprivation .....

ii Create these indicator terms in the dataset

In ***INSIGHT*** you can do this by twice applying the logical transformation (a ≤ Y ≤ b) found under **Edit ->Variables ->Other..** to the variable district
[you could also use it with the existing variable NATURAL, but because of the way this variable has been coded, it is trickier..]

use a=b=1 to create an indicator for district 1 and a=b=2 for district 2 .. but give the two new variables sensible names like 'ALREADY' and 'ADDED".
[careful if you use the existing variable NATURAL. since it has missing values in the absence of fluoride]

NB *before running the regression from INSIGHT, make sure to turn off district as a Group variable i.e. allow all the datapoints to contribute to the single equation...*

In the ***Program Editor*** window, you would do this by inserting the statements, shown in red below, into the indicated place in the data step, before running it

```
data fl_depr;
INPUT fluoride natural deprived t_decay;
district = fluoride;
if natural eq 1 then district = district + 1;
already = 0; if district = 1 then already = 1;
added   = 0; if district = 2 then added   = 1;
LINES;
0  .  -23   1.4
0  .  -15   0.3
etc..
1  1   40   1
;
run;
```

and fit the single 'master' equation.

In ***INSIGHT***, you can use the Cross button in the Fit dialog box to designate the product terms, or (if you don't trust what SAS programmers may have done) you can create the two products yourselves from the Edit->Variables->Other menu

In the ***Program Editor,*** to physically create products, you would add -- after the statements that created 'already' and 'added' -- these two statements to the data step

```
dep_alr = deprived * already ;
dep_add = deprived * added   ;
```

Then,

(a) double check that from the master equation you get you can extract the 3 separate equations you obtained in (i)

(b) formally test if 3 non-parallel lines are a significantly better fit than 3 parallel lines. Be explicit about the hypothesis you are testing (think Reduced vs. Full!)

Water fluoridation, tooth decay in 5 year olds, and social deprivation .....

iii Make district a 'categorical' variable and use it (instead of your own indicator variables) to fit the master equation again

In ***INSIGHT***, designate district as a single *nominal* variable [in the rightmost of the two rectangles above the variable name]; you can again use the Cross button in the Fit dialog box to designate the interaction terms.

In the ***Program Editor,*** to avoid having to make your own indicator variables, you can take advantage of the *CLASS* statement in PROC GLM (PROC REG does not 'do' categorical variables) [i.e. use the CLASS statement to designate a variable as nominal]

```
proc glm data=fl_depr;
  class district;
  model t_decay = deprived district district*deprived /solution;
```

Why do you not get the same equation as before? (hint: notice which value of district is used as the 'reference' value against which the other two levels are compared). Try to convert the equation obtained with the categorical variable into the one you obtained with your user-constructed indicator variables. And comment on the downside of the 'take the easy route and leave it to SAS' strategy.

3 old **Effect of maternal smoking on birthweight** *[revisited]*

i Previously, we fitted parallel lines for weights of children of smokers and nonsmokers: the vertical distance between them was the "age-adjusted" difference in birthweight.

If the intrauterine growth rate differs between smokers and nonsmokers, then the growth curves should not be parallel. However if we only observe the weights in the narrow age range when the children are born (in this dataset between 35 and 42 weeks), and if we do not have a large sample size, this non-parallelism may not "shine though".

• Assume that the top ends of the growth curves for children of smokers and non-smokers can be approximated by two non-parallel lines; fit two such lines to the data -- using a single equation to do so.

• Formally test if two non-parallel lines are a significantly better fit to these data than two parallel lines. Be explicit about the hypotheses (models) you are testing/comparing.

• Suppose you argue that -- regardless of statistical tests and p-values -- non-parallel lines make more biological sense. In this case, you can still calculate age-adjusted smoker-nonsmoker differences in birthweight ***except now you have to make them age-specific***. Calculate age-adjusted difference at (a) 36 weeks (b) 40 weeks.

ii These data were collected in the pre-ultrasound era, and so the recorded gestational ages may contain "measurement" errors.. **If** the age errors are randomly (and symmetrically) distributed around the true values, what impact does this have on the estimates of the slope of weight on age?

4 **Effect of alcohol on smooth pursuit velocity of the eye** *.(data on web page)*

a Fit the model Decrease = Alcohol Gender

b Interpret the coefficients ... and draw a diagram -- for father-in-law!

(You might find it easier if you code gender as 0 and 1, rather than 1 and 2... you can alter the values right in the data window if you wish, or you can use the method described in the child injury example below for subtracting a constant from a variable)

c Fit the model Decrease = Alcohol Gender Gender*Alcohol

(select both Alcohol and Gender& click "CROSS")

d Interpret the coefficients ... and draw a diagram.

(You can get INSIGHT to plot the predicted (fitted) responses against alcohol. and if you wish, you can colour the observations from one of the two genders using EditMenu -> Windows -> Tools then click on a color and select say gender = 1)

e Rewrite the 4-parameter equation as

Decrease $= (b_0 + b_2.\text{Gender}) + (b_1 + b_3\text{Gender})$ Alcohol

f In this 4-parameter model, if we were to judge by the 4 t-ratios and their associated p-values, none of the 4 parameters is statistically significant

Is this a correct interpretation? Explain.

g For each of the 2 genders separately, regress the decrease on alcohol.

(in INSIGHT, you can do this by first making gender a group variable .. click on the rectangle above the column label)

Match up the coefficients of these two separate equations with the equations implied by the 1 "master" equation with 4 coefficients.

h The sample size is small; the slope is statistically significant at the 0.05 level. only in males.

Compare:

	females	males	
n	6	6	
slope	0.20	0.42	
SE(slope)	0.41	0.12	
RMSE ($s_{y\|x}$)	16.8	13.2	("average" residuals)

Can you explain why the SE(slope) is so much smaller for males?

[hint: look at the numerator and the denominator of formula 2.8. In the data, the numerators are only slightly different, so the denominators must be quite different. You might find it easier to examine the form of the denominator using the version of formula 2.8 that I show on page 8 of my notes on G&S chapter 2.

The widths of the SE's have something to do with differences in the within-gender variability in persons' perceptions of when they felt too drunk to drive]

i Fit the model Decrease = Alcohol Gender Gender*Alcohol

- but with the "intercept" turned off.

Draw the lines implied by the fitted equation.

How about the model Decrease = Alcohol Gender Gender*Alcohol

Does this model make sense in this particular example? (see discussion on bottom of p 89 of G&S) or are there good reasons why the lines might not go through the origin?

5 **Focus on the coefficient of a product term**

2012

For the following analyses, use the data on GIRLS in the intervention and 4 BORDER MUNICIPALITIES in the article "The Lidkoping Accident Prevention Programme -- a community approach to preventing childhood injuries in Sweden" by Svanstrom L et al ; Injury Prevention 1995 1: 169-172; ~~- under datasets on 626 class web page~~ on 609 site

One way to select these particular observations in INSIGHT is to sort the data by gender and area, then extract the observations on girls in areas 0 and 1

a For each of the 2 areas separately, regress the rate on year.

(in INSIGHT, you can do this by first making area a group variable .. click on the rectangle above the column label)

Refer to Table 2 of the article.

- Verify the "beta's" of -0.3 and 0.2 for the two areas.

- Divide them by the average rates to get a "%change per year".

b Interpret the "INTERCEPT" values in your two regressions, when using year as "Annum Domini"

c Change year from "Annum Domini" to "year of program"

```
Edit menu -> Variables ->
   Click YEAR into the Y box
   Click on the "a + b*Y" transformation
   Set a to -1983 and b to 1
     Click on OK
```

This should produce a new year ("A_YEAR") that starts at 1983 (If you like, you can double click on the name A_YEAR and change it to something more meaningful, like Pgm_Year, short for "Program_Year")

Re-run the 2 regressions using the new "year", and re-interpret the coefficients. Compare, and comment on, the SE's of the intercept's in the models using the "Y-almost-2K" and the "Y-starting in 1983" versions of year.

d Why switch to the new variable, year_pgm = year - 1983 ?

e Use a t-test to formally test the between-area difference in the annual change in incidence (to save you time: sqrt[0.29*0.29 + 0.32*0.32] = 0.43)

f Remove the "Group" designation from area

Fit a single regression equation to all 18 observations...

```
Y          X

RATE     Pgm_Year
         Area
         Area*Pgm_Year
```

(you put in the product by entering Pgm_Year and Area , then *selecting both and clicking "CROSS")*

g Use the 4 parameter estimates to recreate the equations of the 2 fitted lines

Draw them in on the scatterplot of rate vs Pgm_Year, or...
have INSIGHT make a scatterplot of the predicted rates versus the Pgm_Year

If you wish, colour the observations from one of the areas ..
use EditMenu -> Windows -> Tools ; Click on a color and select say area = 1

h Interpret the coefficient associated with the product-term Area*Pgm_Year. Show that it agrees exactly with the results obtained by fitting two separate lines in question b. Use the SE associated with the Area*Pgm_Year coefficient to formally test the observed between-area difference in the annual change in incidence.