STA 312f10 Assignment 4

Do this assignment in preparation for the quiz on Friday, Oct. 8th. The non-computer parts are practice for the quiz, and are not to be handed in. But please bring your R printouts to the quiz; they may be handed in. Please do *not* write anything on your printouts before the quiz, except possibly your name and student number. Start by reading Chapter 3.

The full (saturated) log-linear model for a three-dimensional table is

$$\begin{aligned}\log m_{ijk} = \mu &+ \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} \\ &+ \mu_{12(ij)} + \mu_{13(ik)} + \mu_{23(jk)} \\ &+ \mu_{123(ijk)}.\end{aligned}$$

It may be written in a more condensed form (see Table 3-4 in the text) as

$$\log m = \mu \ + \ \mu_1 + \mu_2 + \mu_3 \ + \ \mu_{12} + \mu_{13} + \mu_{23} \ + \ \mu_{123}.$$

We will call this "μ notation." In "bracket notation," the saturated model is written [123].

1. Using μ notation, give two examples of models for a 3-dimensional table that are *not* hierarchical, and briefly indicate why they are not hierarchical.

2. Write each of the following models in μ notation:

   (a) [1] [23]
   (b) [12] [13]
   (c) [1] [2] [3]
   (d) [13]
   (e) [12] [13] [23]

3. Write each of the following models in bracket notation. The bracket notation applies only to hierarchical models, so if the model is not hierarchical just write "Not hierarchical."

   (a) $\log m = \mu + \mu_1 + \mu_2 + \mu_3 + \mu_{12} + \mu_{23}$
   (b) $\log m = \mu + \mu_1 + \mu_2 + \mu_3 + \mu_{13}$
   (c) $\log m = \mu + \mu_1 + \mu_2 + \mu_3 + \mu_{12} + \mu_{13} + \mu_{23} + \mu_{123}$
   (d) $\log m = \mu + \mu_1 + \mu_2 + \mu_3 + \mu_{13} + \mu_{23} + \mu_{123}$
   (e) $\log m = \mu + \mu_1 + \mu_{12} + \mu_{13} + \mu_{23} + \mu_{123}$

4. Let

$$Y = \beta_0 + \beta_2 X_2 + \epsilon,$$

with $E(X_1) = \mu_1$, $E(X_2) = \mu_2$, $Var(X_1) = \sigma_1^2$, $Var(X_2) = \sigma_2^2$, $Cov(X_1, X_2) = \sigma_{12}$, $E(\epsilon) = 0$, and ϵ is independent of X_1 and X_2. Find $Cov(X_1, Y)$. Show your work.

The point of this question is that in regression, Y can be produced by X_2 and not X_1, but still X_1 and Y can be correlated. Something similar happens with categorical data, and it falls under the heading of *conditional independence*.

5. Consider a $2 \times 3 \times 4$ table. Give the degrees of freedom for testing the fit of each of the models below.

   (a) [1] [23]

   (b) [12] [13]

   (c) [1] [2] [3]

   (d) [13]

   (e) [12] [13] [23]

   (f) $\log m = \mu + \mu_1 + \mu_2 + \mu_3 + \mu_{12} + \mu_{23}$

   (g) $\log m = \mu + \mu_1 + \mu_2 + \mu_3 + \mu_{13}$

   (h) $\log m = \mu + \mu_1 + \mu_2 + \mu_3 + \mu_{12} + \mu_{13} + \mu_{23} + \mu_{123}$

   (i) $\log m = \mu + \mu_1 + \mu_2 + \mu_3 + \mu_{13} + \mu_{23} + \mu_{123}$

   (j) $\log m = \mu + \mu_1 + \mu_{12} + \mu_{13} + \mu_{23} + \mu_{123}$

6. Do problem 3.1. Here are some suggestions, comments and one warning.

   - If you create a data frame to get the data into R, please display it by typing its name after you read it. If you put the frequencies directly into a table, show the process on your printout, and be sure to make nice labels.[1]
   - I found getting the data into R to be a bit time consuming. If you finish this problem early, you may be tempted to permit a classmate who is under time pressure to use the data entry part of your work. *Don't do it!* It is academic dishonesty to present any part of someone else's work as your own, or to let yours be used this way. Both people will get in big trouble.
   - After you make your 3-dimensional table, type its name to display it (show this on your printout), and *proofread it.* It's a waste of time to do statistical analyses when the data are wrong. Did you get $N = 1329$?

[1] Each way of getting the data into R has advantages and disadvantages. To me, the data frame method is an easier way to specify the labels of the categories (< 127 and so on), but then when you tabulate the data, R alphabetizes the rows and columns. This can create a mess if you are not aware of the issue when you create the data frame. If you put the frequencies directly into a table you have to play around with `dimnames`, but at least the table looks exactly the way you want it to. I was not aware of the alphabetical order issue until I started doing this problem, and now I appreciate why SAS `proc freq` has an option to create tables using the order in which the categories appear in the data file.

- Part (a) of the Problem 3.1 asks you to test the "model of no second-order interaction." Remember that the models are all hierarchical. So I interpret this as meaning the model of complete independence – a good place to start in general.
  - To test this model, calculate the p-value for both X^2 and G^2. Is your conclusion the same based on the two tests, or do the two tests lead to different conclusions?
  - G^2 is *not* exactly being calculaated using the formula from text and lecture. How do you know this, and how is G^2 being computed?
  - State your conclusion or conclusions in words.
  - From now on in the question, please stick to likelihood ratio tests. For every model you fit, you should be able to give the value of the G^2 statistic (a number), the p-value, and the degrees of freedom.
- For Part (b) of the question, please first add the connection between the explanatory variables. This can be justified by an exploratory model building approach, but none of that is in Chapter 3. Also, later in the course you will see that *when there is a clear distinction between explanatory and response variables, it is common to use models that include all possible interactions among explanatory variables, whether or not they are related.* There is a good justification for this, but it will have to wait. Anyway, for now just start with the model `[cholest bp] [chd]`. Does this model fit the data well enough?

  Now fit three more models:
  - Add `[cholest chd]` but not `[bp chd]`.
  - Add `[bp chd]` but not `[cholest chd]`.
  - Add both

  Base your answer to Part (b) on tests of fit for these three models. Don't forget to calculate the p-values.
- For Part (c) of the question, base your answer on the estimated μ values for the association of blood pressure with coronary heart disease, and for the association of cholesterol with coronary heart disease. I can see the pattern they are talking about.