

HW 3: due Monday, 27 February

1. **Hispanic Americans and dietary cholesterol.** The Hispanic Health and Nutrition Examination Survey (HHANES) was conducted throughout the US from 1982 to 1984. Researchers used the data to compare dietary components among adult Mexican Americans, Cuban Americans, and mainland Puerto Ricans (*Am J Public Health*, 1995; 85:684–689). One dietary component examined was cholesterol, presented in the table below.

TABLE 4—Daily Cholesterol Intakes (Mean ± Standard Error, in Milligrams), by Sex, Age Group, and Ethnicity (HHANES, 1982 to 84)

Sex and Age Group	Mexican American	Cuban American	Puerto Rican
Males			
20–39 y	505.0 ± 19.8	442.9 ± 19.7	427.1 ± 18.4
40–59 y	458.9 ± 17.4 ^{ca}	366.5 ± 11.6 ^{ma}	399.6 ± 16.6
60–74 y	440.9 ± 25.2 ^{ca,pr}	279.4 ± 13.2 ^{ma}	305.5 ± 22.7 ^{ma}
All ^a	478.8 ± 11.2	389.3 ± 16.8	396.7 ± 14.0
Females			
20–39 y	332.0 ± 10.7	282.7 ± 13.1	311.0 ± 10.3
40–59 y	300.3 ± 10.0	279.5 ± 25.4	263.2 ± 8.5
60–74 y	302.3 ± 18.2 ^{ca,pr}	210.9 ± 9.9 ^{ma}	199.5 ± 12.4 ^{ma}
All ^a	316.6 ± 6.4	268.9 ± 17.4	275.8 ± 5.7

Note. Cell mean differs from mean of ma = Mexican American, ca = Cuban American, pr = Puerto Rican, nhw = non-Hispanic white, or nhb = non-Hispanic black ($P < .01$).

^aAge-adjusted by the direct method to the total 1980 US population.

- Identify the three factors in this table and list the levels for each factor.
- Make two interaction plots, one for men and one for women, from this table using means only (no SE bars). Put age groups on the horizontal axis and use the same vertical axis (cholesterol, from 200 to 525 mg) for both plots, so they can be compared side by side. Do not include the means labelled “All.”
- The rows in the table labeled “All” are basically averages across the age groups. Assume that each group has the same number of participants and use the rows labeled “All” to estimate the gender main effect.
- The footnotes in the table indicate significant differences at the 1% level. Were the comparisons done between rows or between columns?
- The paper summarized the table above by noting significant differences at the 1% level:

Average cholesterol intakes for men were similar among Cuban Americans and Puerto Ricans and lower than intakes among Mexican Americans. Mexican American women 60 to 74 years of age had higher intakes than did their Cuban American and Puerto Rican counterparts, but no differences were found between younger women in those three groups.

Effects of two factors were not summarized in this paragraph. Identify the factors and describe their main effects in one or two sentences.

2. **Child IQ.** This example is taken from a text by Gelman and Hill, who cite the National Longitudinal Survey of Youth (<http://www.bls.gov/nls/home.htm>). We will look at the difference in IQ score in 4-year-old children between those whose mother graduated from high school and those whose mother did not.

The data are given on the course website in the file `childIQ.csv`: `kid_score` = child's IQ score; `mom_hs` = 1 if mother graduated from high school, = 0 if not; `mom_iq` = mother's IQ score. Also download and run the program `hw3.sas`.

- a. The first model gives the unadjusted difference:

```
Proc GLM;
  class mom_hs;
  model kid_score = mom_hs;
```

The second model gives the difference adjusted for mother's IQ:

```
Proc GLM;
  class mom_hs;
  model kid_score = mom_hs mom_iq;
```

Make a table with two rows, unadjusted and adjusted results. In each row, report the mean \pm SE for each group, the mean difference and its 95% confidence interval.

- b. Which of the two means moves farther with the adjustment? Does the adjustment move the means closer together? Is the adjusted difference outside the confidence interval for the unadjusted difference?
- c. The last output from the program is a graph of child's IQ against mother's IQ. You can find the plot in the Results panel on the left, by clicking on SGPlot, the clicking on the plot icon. Is the association positive or negative?

The plot uses different marks to indicate mothers who graduated from high school. Can you use the graph to explain why adjusting for mother's IQ has such a large effect on one of the group means?

3. **Smoking and birthweight.** This problem refers to the following paper (from E-Journals): Eskenazi, Prehn, and Christianson (1995) Passive and active maternal smoking as measured by serum cotinine: the effect on birthweight, *American Journal of Public Health*, 85:395–398.

The primary aim for this study was to estimate the effects on birthweight when pregnant women were exposed to second-hand ("environmental") smoke. Exposure to environmental tobacco smoke during pregnancy is now considered hazardous, so it is hard to find cases to study. This group went back almost 50 years to records and samples from the Child Health and Development Studies, done in Oakland from 1964 to 1967, when many women smoked during their pregnancy. They used interview records from these mothers and serum samples frozen since 1967.

The frozen sera were tested for cotinine, which is a by-product of exposure to nicotine. The idea was to identify women who said they did not smoke but whose serum contained cotinine, because these were nonsmoking women exposed to environmental tobacco smoke during pregnancy. Then they looked for a negative relationship between exposure and birthweight.

- a. The first paragraph (§1) of *Methods* says that $n = 3896$ women were in the sample. The total in Table 1 is only 3529. What happened to the missing 367?

- b. Table 1 compares three study groups with respect to the predictors (covariates). Although Table 1 has no p -values, the first sentence of the *Results* lists the significant differences between the groups. Which covariates are significantly different?
- c. What is the relationship between the three study groups compared in Table 1 and the five study groups in Table 2? Which study group is the most interesting to the authors and why? What kind of ANOVA was used to compare the birthweights in Table 2?
- d. Birthweight depends on many things besides exposure to tobacco smoke. Table 3 compares the five study groups after adjusting for other covariates. Table 3 is a list of the regression coefficients from one model in Proc GLM. Which covariates are class variables?
- e. How can you tell from Table 3 that the regression coefficients for *Age* are not significant? Why did they keep this variable in their model anyway?
- f. ¶3 of the Results section states:
Nonsmokers exposed to environmental tobacco smoke had infants that were, on average, 45 g lighter than those of other nonsmokers; however, this difference was not statistically significant.
Where does the adjusted estimate "45 g lighter" come from?
- g. What is the *unadjusted* estimate of the effect on birthweight of exposure to environmental tobacco smoke?
- h. ¶5 of the Results section describes a "multivariate analysis of cotinine as a continuous variable." What do you change in the model for Table 3 to make cotinine a continuous variable?
- i. Why do you think that the authors failed to find a significant effect on birthweight for the nonsmokers exposed to smoke? *Hints*: think about study group sample sizes; misclassification of non-smokers based on cotinine.