Soc 63993, Homework #10 Answer Key:
Year in Review Special Edition

Richard Williams, University of Notre Dame, http://www3.nd.edu/~rwilliam/

Last revised April 6, 2015

This homework is a mini-review of much of the material we have covered throughout the year (including last semester). In a few instances, the "best" answer requires advanced techniques we have only briefly or recently mentioned. Skimming through the syllabi from both semesters and the advanced section of the readings packet (or at least my summary of those readings) may help you.

I. For each of the following circumstances describe the statistical technique you would use for revealing the relationship between the dependent and independent variables. Write a few sentences explaining and justifying your answer. Part of your answer may include an explanation of why other apparent alternatives would not be as good. In some instances more than one technique may be reasonable.

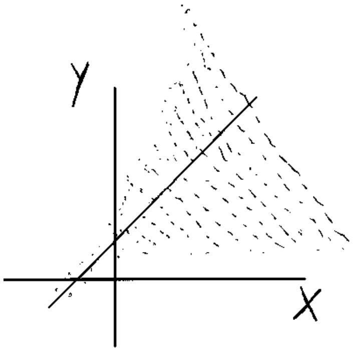
EXAMPLE: The President of Notre Dame wants to see whether anything can be done about student drinking. He intends to have a random sample of 25 students fill out a questionnaire reporting on how much alcohol they drank in the last month. The students will then see a film on alcohol abuse. A month later they will again report on how much alcohol they drank in the last month.

SOLUTION: A matched pairs T-Test is called for. Individuals are interviewed before and after seeing the film, with the goal of seeing whether the amount they drink has changed. An independent samples T-Test would not be appropriate since the same people are measured twice. Such a T-test would be appropriate if, say, half the subjects were shown the film and half were not.

EXAMPLE: For a sample of 10,000 births you know whether the baby died in the first year of life, its birth weight, mother's age and education, and its race. How would you assess the influence of these variables on infant mortality?

SOLUTION: Since the dependent variable is a dichotomy, Logistic regression would probably be best. The logistic regression would show how each variable affected the log odds of a baby dieing. OLS regression could produce impossible estimates of probabilities (greater than 1 or less than 0) and the OLS assumptions of the X's having linear effects on the probabilities are usually not reasonable.

a. A researcher is interested in the relationship between X and Y. A scatterplot of her data reveals the following:



Heteroskedasticity is a problem. You may need to respecify the model, e.g. include interaction terms. If important variables are omitted, your parameter estimates will be biased. If you are confident that the model is correct, use weighted least squares. Otherwise, you will not get the most efficient estimates.

b. A researcher is interested in whether severity of accident and location of accident are related. He plans to draw a sample of 100 accident records. In the accident records, severity is coded as either (1) property damage, (2) injury, or (3) fatality. The three possible locations for accidents are freeway, rural road, or city road.

A simple crosstab with a test for the model of independence will do. You could also consider something a little fancier, like multinomial logit or ordinal regression.

c. A psychologist believes he has come up with a way of motivating students to exercise more. A random sample of 50 students will fill out a questionnaire indicating how many hours a week they exercise. Subjects will then see a series of films which highlight the benefits of exercise and which show how to work exercise into the daily schedule. A month later the students will again report on how many hours a week they exercise. How can he test whether his treatment has caused students to exercise more?

A matched pairs t-test is appropriate. The same subjects are interviewed both before and after they see the films. The researcher is hoping that students will exercise more after seeing the films.

d. A researcher wants to know how educational aspirations are related to educational achievement. Grade school students are asked how much education they want to get, how much education their parents want them to get, and how much education their teachers want them to get. Twenty years later the students are reinterviewed. They are asked retrospective questions about what their educational aspirations were in grade school, what their parents educational aspirations were, and what their teachers' aspirations were. They are also asked how much education they actually achieved. (Assume that educational aspirations and educational achievement are both coded as continuous variables.)

Lisrel or Stata's sem (or some other similar structural equation modeling technique) would probably be very appropriate. The researcher has multiple indicators of educational aspirations, measured both before the event and afterwards. Individually, these measures probably all suffer from at least some degree of random measurement error, hence using them in a conventional OLS regression will likely result in biased estimates and/or inflated standard errors. In Lisrel, she can use these indicators to specify a measurement model in which each type of aspiration has two indicators. If her model works well, her estimates of the structural effects will not be biased because of measurement error.

e. A researcher is interested in the relationship between education and earnings. She suspects that the relationship is non-linear – more education is beneficial up to a point, but after that the effects of education become smaller or even start to go negative.

Plain old OLS with some nonlinear transformations of X will probably do. e.g. include Education^2 in the model, or perhaps do a piecewise regression, or maybe use ln(Education).

f. There is an ongoing debate in demography about the relationship between years of education and a mother's age when she first gives birth. Some say that the more education a woman has (or wants), the longer she will wait before giving birth. Others, however, contend that the mother's age at first birth affects how many years of education she can get, e.g. a woman who gives birth at a young age will often have to curtail her education. A researcher believes that the causality between age at first birth and education flows in both directions and wants to estimate models that reflect that.

Use a nonrecursive model, estimated via 2sls or Lisrel. i.e. specify a model in which education and Mother's age at first birth both affect each other. Regular OLS will provide biased estimates of the effects because OLS assumes that effects only run in one direction. Getting the model identified will be the greatest challenge, of course.

g. The South Bend School Corporation is struggling both with budget deficits and weak student performance in the classroom. An experimental curriculum has been developed that will hopefully make teachers more effective while at the same time making more economical use of limited resources. For one year, this curriculum will be used in half the corporation's schools while the other half continue with traditional methods. The effect of the curriculum on standardized test scores, # of disciplinary problems, and dollars spent per student will be assessed.

Manova (or Lisrel or Stata's sem) would be good. There are multiple dependent variables. The Manova/Lisrel approach will let you see what effect the curriculum has on each of the DVs. It will provide better estimates because, unlike separate regressions, the intercorrelations of the DVs will be taken into account.

h. The Latino Studies Center wants to know more about what affects attitudes toward immigration. Data from a nationwide sample of Americans has been collected. Information has been gathered on years of education, occupational prestige of respondents, political partisanship (measured on a hundred point scale) and various other continuous and dichotomous independent variables. The dependent variable is coded 1 = Favors more immigration, 2 = Level of immigration is about right, 3 = There should be less immigration.

The DV is ordinal, so an ordered logit model is probably best. Trying to use OLS with a 3-category ordinal variable may not work very well. Or, if the assumptions of the ordered logit model are not met, you could use a multinomial logit model that ignored the ordering of the categories. You might also consider dichotomizing the DV and using logistic regression.

i. A researcher has the marital histories for a sample of 1,000 65-year-old women, i.e. she knows when and if each woman got married, when and if she got widowed or divorced, when or if she remarried, etc. She also has information on the women's race, annual income for each of the last 50 years, educational history, and medical history (for both the women and their spouse's, if any). How should she go about examining the determinants of change in marital status?

Event history analysis is called for. The data are censored in that she doesn't know what happened to the women after age 65, e.g. some may have later become widowed or gotten remarried. She has extensive longitudinal data, which will make it possible for her to see how these variables affected change across time.

II. For each of the following, indicate whether the statement is true or false. Explain why.

a. Race has three categories: white, black, and other. A researcher therefore creates two dummy variables from race: BLACK and OTHER. To see whether race affects income, a T-test should be used.

False. A T-test would only work if there were 2 categories for race instead of 3. She should use an incremental F test.

b. Logistic regression coefficients, like ordinary least squares coefficients, tell you the unit increase in Y for a one-unit increase in X.

False. In logistic regression, the coefficients tell you the effect of X on the log odds of Y occurring.

c. When doing multiple-group comparisons (e.g. women vs. men), standardized rather than metric coefficients should be analyzed.

False. Standardized coefficients in multiple-group comparisons are the work of the devil. They can disguise important similarities and differences across groups.

d. The null and alternative hypotheses are

$$H_0 : \beta = 0$$
$$H_A : \beta < 0$$

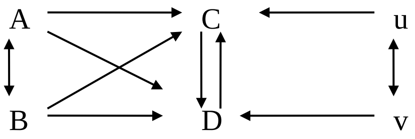
In a sample of 10,000 cases, it is found that $\hat{\beta}$ = 4 with a standard error of 1. The null hypothesis should <u>not</u> be rejected.

True. The estimated sign is in the wrong direction (note that alternative is one-tailed), so the Null wins.

e. Analysis of variance problems (e.g. examining the effect of religion on income) can also usually be addressed via multiple regression techniques.

True. ANOVA is just a variation of regression analysis.

f. A researcher wants to estimate the following model. Variables A, B, C, D have interval-level measurement. u and v are disturbance terms. She should use 2 stage least squares.



False. The model is hopelessly under-identified. Prayer or some sort of alternative model specification is the only answer. If it was legitimate to drop the path from A to D and the path from B to C, then 2sls could be used.