

Explanatory Variables in OLS Regression
[explanatory_vars_in_OLS_regr.doc]

Part I.

Options in conceptualizing the relation of the outcome variable to an explanatory variable, holding the other explanatory variables constant

1. ***Linear, independent*** (e.g., relationship of outcome variable “income” to education—i.e. slope—is the same at all levels of “education” & for “females-males”, holding the other explanatory variables constant; & the explanatory variables do not interact): first-order quantitative variable (e.g., income measured in thousands of dollars; the dummy variable for “gender” will test insignificant in this example).
 - a. Check the sample size at each level of the quantitative explanatory variable.

2. ***Same slope coefficient but unequal y-intercept*** (e.g., same slope relationship of outcome variable “income” to “years of education” for females-males but females earn less than males): dummy variable (e.g., dummy variable “female” [or “gender”]: male=0 female=1).
 - a. The sample size must be adequate in each category. If there are more than two categories, the largest category is commonly used as the “base” (or “reference” or “comparison”) category.

3. ***Critical thresholds*** (e.g., education is associated with a major increase in outcome variable income only at “one or more years of college” & higher, or only at “one or more years of college” [see *increasing or decreasing effect*]): categorize “years of education” into appropriate categories for level of education (i.e. turn quantitative “years of education” into a categorical ordinal variable) (see also *increasing or decreasing effect*: in some such cases, squaring the explanatory variable or doing a log transformation may be an alternative option, though the former requires interval data & the latter ratio data with positive values).
 - a. The sample size must be adequate in each categorical level.
 - b. The interpretation of *critical thresholds* is typically more intuitive than using either squared or log transformations (see *increasing or decreasing effect*).
 - c. Check the critical threshold: Does it make substantive or theoretical sense?
 - d. But creating a categorical ordinal variable uses more degrees of freedom than either squared or log transformations; thus the sample size must be adequate.

4. ***Increasing or decreasing effect*** (e.g., “years of education” is associated with increasing outcome variable “income” until six years of college, decreasing from that point onward [see second aspect of *critical thresholds*]): square “years of education”; or perhaps log “years of education.”

- a. Squaring a quantitative variable requires that it be measured at the interval level; taking the log of a quantitative variable requires the more stringent criteria that the variable be measured at the ratio level & that all of its values be positive – e.g., if necessary do the following: $\text{gen leduc} = \log(\text{educ} + 1)$.
 - b. The interpretation of especially a log-transformed variable but also a squared variable is typically less intuitive than that of a categorical ordinal variable.
 - c. Check at what level the effect starts increasing or decreasing: Does it make substantive or theoretical sense?
5. ***Increasing or decreasing effect for one group or some particular groups*** (e.g., education is associated with increasing outcome variable income for males but not for females; i.e the rate of change is not the same for each group but rather varies by group): interact dummy variable “gender” & quantitative variable “years of education” (e.g., femaleXeduc_yrs [the reverse is equivalent]).
- a. The sample size must be adequate at each of the interacting levels.
 - b. Graph each level of the interaction (e.g., females at years corresponding to “less than high school,” “high school,” etc.; males at the same levels): Do these make substantive or theoretical sense?
6. ***Increasing or decreasing effect of a quantitative variable varies by level of another quantitative variable*** (e.g., association between outcome variable income and “years of education” varies with “workweek hours”; i.e. the rate of change varies by level of another quantitative variable): interact “years of education” with “workweek hours” (e.g., educ_yrsXwork_hrs [the reverse is equivalent]).
- a. The sample size must be adequate at each of the interacting levels.
 - b. Graph particular, illustrative levels of one or both variables (e.g., “educ_yrs” at the levels for less than high school, high school, etc.): Do these make substantive or theoretical sense?

Part II

1. List the explanatory variables in order of their anticipated, conceptual importance in relation to the outcome variable, briefly noting (a) the dimensions of such importance; (b) the anticipated type of relationship to the outcome variable according to the criteria in Part I; & (c) how each explanatory variable relates to each of the other explanatory variables.
2. In delineating the variables in these ways, consider the kinds of outcome variable/explanatory variable relationships discussed in Agresti & Finlay, chapter 10.

Part III

Are the vars conceptualized & measured correctly?

Are the measurement premises met?

Are there curvilinearities that need to be assessed/adjusted?

Are there outliers that need to be assessed/adjusted?

Should robust se's, weighted least squares regression, or survey-stats be used?

Should a selection model (Heckman, etc.) be used?

1. Estimate a preliminary main-effects multivariable model, including all the provisionally selected explanatory vars.

- *For now, eliminate the vars that test insignificant:*
 - compare the new vs old model via nested F-test, adjR2, p-values, coefs, se's & CI's
 - one by one drop each outcome var & re-estimate the model, comparing models via nested F-tests, adjR2, p-values, coefs, se's & CI's
 - For now, eliminate all the explanatory vars that test insignificant: this represents the "preliminary main effects model"

2. In the preliminary main effects model, examine each explanatory var more closely

- *Quantitative explanatory vars:* use any of the following to check the linearity of the y/x relationship, & if appropriate, transform the variable, re-estimate the model & compare the coefs & p-values to the original var. [Note: consider also & explore if it makes sense to transform the outcome var]
 - ladder x1 & qladder x1
 - spar1 y x1, logy|logx|pow|quad
 - locpoly y x1
 - scatter mband y x1, band(8)
 - lowess y x1, bw(.2)
 - fracpoly regress y x1, com
 - fracplot x1
- *Categorical explanatory vars:* consider & explore whether it make substantive & statistical sense to collapse or otherwise change the categories of multilevel categorical vars
- *Possible interactions:* prepare a list of the substantively meaningful interactions
 - Add each of the interactions one at a time to the model, checking the p-values of both the interaction & the other vars; make a table of the results
 - Add all of the statistically significant interactions to the model, doing nested F-tests, adjR2 comparison & checking the p-values; enter the results in the table

- Add all of the variables that had previously been eliminated, doing nested F-tests, adjR2 comparison & checking the p-values; enter the results in the table
- Select the best model on the basis of both theoretical-practical significance & statistical significance
- This is the *"preliminary final model"*

3. Assess the *"preliminary final model"* via diagnostic tests, then modify accordingly.