

Explanatory Variables in Logistic Regression

[explanatory_vars_in_logistic_regr.doc]

The following is based on Hosmer & Lemeshow, Applied Logistic Regression

Part I

1. Graphically & numerically describe the outcome var & potential explanatory vars.
2. List the potential 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; & (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 var/explanatory var relationships discussed in Agresti & Finlay, chapter 10.

Part II

1. Crosscheck the outcome var with potential explanatory vars: at this preliminary stage, select as potential explanatory vars those that have at least a modest association with the outcome var, entering the var names in a table.
2. Check collinearity among the quantitative vars:
 - graph comparative boxplots & a scatterplot matrix for the quantitative vars
 - estimate the model in OLS (including only quantitative explanatory vars on the right-hand side)
 - VIF
3. Explore univariable models with each potential explanatory var: record each coef, se & CI, & p-value. Provisionally select a var if its $p\text{-value} \leq 0.25$.
 - *Estimate a univariable model for each potential explanatory var as follows, & enter the results in a table of results:*

scatlog female wage
logit female wage, or
est st f [abbreviations for 'estimates store full']
logit female, or
lrtest f . [test 'full' versus reduced '.' model]
 - *Preliminarily check the linearity of each potential explanatory quantitative var, indicating the possibly non-linear vars in the table. Here are two of the possible ways of checking linearity.*

sparl female wage, quad | logx
scatter female wage || ffit female wage
fracpoly female wage, compare

4. Estimate a preliminary main-effects multivariable model

- *Compare the coefs & p-values with those in the univariable models*
logistic female wage exper marr
- *Eliminate the explanatory vars that test insignificant*
 - compare the new vs old models via lrtest (or collective Wald-test if using pweights), & compare the coefs & Wald-tests in the new vs old models & vs the univariable models
 - one by one eliminate each explanatory var & re-test the model, comparing the models via lrtest (or collective Wald-test if using pweights) & comparing the coefs & Wald-tests
- *Add to the model any vars that were not selected for the multivariable model, & estimate this model (or estimate logsub: view help logsub)*
 - Eliminate all vars that test insignificant: this represents the "preliminary main effects model"

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

- *Categorical vars: does it make substantive & statistical sense to collapse or otherwise change the categories?*
- *Quantitative vars: check linearity (via, e.g., spar1, fptest, ladder, qladder, spar1, twoway mband, or ksm) & if appropriate, re-estimate the model & compare the coefs & Wald-tests to the original var*
- *Possible interactions: prepare a list of the substantively meaningful interactions*
 - Add each of them one at a time to the model & successively re-estimating it, doing LR-test model comparisons & comparing the p-values; enter the results in a table of results
 - Add all of the statistically significant interactions to the model, re-estimating it, doing a LR-test model comparisons & comparing the Wald-tests; enter the results in the table
 - This is the "preliminary final model"

6. Assess fit of preliminary model

- *Summary measures: obtain & enter the following results in the table*

linktest, nolog

ldev

lfit

lfit, g(10) table

if necessary, collapse rows to increase size of the expected frequencies & to reduce DF

- *Note: these do not indicate "degree of fit" but merely "overall fit" for any given model*

- *Individual diagnostic tests*: obtain & enter the following results in the table

predict p if e(sample)	predicted probabilities
predict db if e(sample), db	approximate standard: db>1
predict dd if e(sample), dd	approximate standard: dd>4
predict dx if e(sample), dx	approximate standard: dx>4
predict h if e(sample), h	approximate standard: (3*k/N)
predict n if e(sample), n	covariate patterns

la var p "predicted probabilities"

hist p, norm
su p-n, detail

For ordinal or multinomial models, do the following for db-n:
predict ... if ycat~=... (see Hosmer & Lemeshow)

- *for each influence-diagnostic test*

scatter db p [w=dx], yline(1) ml(n)
sort db
l id db p female if db>1

Notes:

- h is not linear at pred prob<.1 | >.9.
- db is usually large when 0.1<pred prob<0.3 or 0.7<pred prob<0.9 & both h & dx are least moderately large
- db is usually small when pred prob<0.1 or pred prob>.09 & h is small (while dx is small or large)
- db is a summary measure, so check influential covariate patterns
- db usually must be>1 to exert much effect
- *list influential covariate patterns & check why they are outliers*
l if n==189 | n==26 | n==201
- *re-estimate the preliminary final model without the influential n's, comparing models via LR-tests or Wald-tests, linktest, ldev, lfit, lfit g(10) t*
 - enter the results in the table
 - if results aren't satisfactory, check that (1) model is adequate in its functional form; (2) no important explanatory vars have been omitted; (3) all explanatory vars have been entered in their correct scales (including adjustments for non-linearity & outliers); or (4) there are no clustering effects or other data problems
 - the final iteration is the "final model"

7. Perform model validation on a sub-sample &/or on another data set

8. Obtain post-estimation predictions & graphs (e.g., Long/Freese procedures, postgr; see logistic_regress_example)