When I run a regression, why do some of the variables disappear from the output?

Multivariate linear regression assumes that your independent variables are not highly correlated. If the variables are collinear (i.e., the assumption is violated), this is called multicollinearity or singularity (perfect correlation). You don’t want multicollinearity or singularity because they basically mean that your IVs are redundant; if you have these redundant variables in your regression, you aren’t adding any predictive value and you are losing a degree of freedom. In other words, multicollinearity and singularity weaken your regression model. If Stata detects that variables in a regression are collinear (0.9 or greater), it will remove one or more of the variables from the regression so that the assumption isn’t violated.

This happens a lot with dummy variables. Let’s say you have six dummies. With any five, you can perfectly predict the last column. Thus, STATA always drops one dummy if they are fully saturated.

Try running correlations among your IVs to detect high bivariate correlations (e.g., correlate height weight). In general, you don’t want to include two variables that correlate with one another at >= 0.70. If you do have high bivariate correlations, your problem is easily solved by removing one of the two variables from your regression equation. However, before dropping variables you should check your programming because correlations between IVs are often due to a mistake when you created the variable(s).

When I run a regression, why are there so few observations included in the regression? My dataset has a lot more observations than what was used in the regression!

Let’s say your original dataset includes 1000 observations, but only 500 are being used in your regression equation. Your regression includes 1 DV and 5 IV. If any of the IVs are missing values for each observation, that observation will be dropped from the regression. Check your IVs and see how many have missing values. If a lot of observations are missing values for an IV, you may want to reconsider whether to include the variable in your regression. If you created the IV (e.g., a dummy variable) that has many missing values, go back and check your code to make sure that it was created correctly.

For more help with regression and its assumptions, check out Princeton’s introductory guide: http://dss.princeton.edu/online_help/analysis/regression_intro.htm