How to identify which predictors should be included in a multiple regression?

I am not a statistician, but a medical researcher and I have 5 outcomes that I want to identify independent predictor(s) for each using multiple regression. I have many potential variables that could be included in the multiple regression as independent variables (IV).

One colleague advises to run Spearman correlation matrix between all IV and DV, then to include only the significantly correlated IV in the multiple regression.

Questions

- Is it appropriate to include only significant predictors with significant bivariate spearman correlations with the outcome?
- Alternatively, what is a good way to determine inclusion of predictors in a multiple regression?

correlation | multiple-regression

edited Jul 17 '13 at 4:15

Jeromy Anglim
26.5k 11 87 189

asked Jul 16 '13 at 22:08

Emaddin Kidher
47 6

This approach has been called "double dipping". See my answer below. – Frank Harrell Jul 16 '13 at 22:25

3 Answers

The model should be formulated by subject matter expertise. It is not a good idea to use the data to tell you which data to use. The data are not information-rich enough to be able to reliably do this. Should you have too many events per variable (one rule of thumb is to have at least 15 subjects per parameter in the model), strongly consider data reduction methods that are blinded to Y. These include principal components, variable clustering, and redundancy analysis. Examples are in my course notes at http://biostat.mc.vanderbilt.edu/CourseBios330.

answered Jul 16 '13 at 22:25

Frank Harrell
35.6k 1 62 138

1 For many of the data sets I work with, subject-matter expertise is not well-developed either. There are many studies and guesses, but no consensus it seems. So although the advice is sound to me, I find it hard to use in many situations I encounter. – julieth Jul 17 '13 at 3:01

1 Then be clear that the project is worth doing. Modeling for the same of modeling is sometimes not a fruitful endeavor. Most interesting analyses are driven by interesting questions. Barring that, you can still use data reduction and penalization (lasso, elastic net) methods to find predictive signals. – Frank Harrell Jul 17 '13 at 11:44

There are lots of methods that can be used for variable selection. LASSO is one of the better data driven variable selection models. Do not, whatever you do, use forward stepwise. You'll be glad you didn't:

http://www.nesug.org/proceedings/nesug07/sa/sa07.pdf

answered Jul 17 '13 at 2:29

user2589635
11 1

It is probably important to not let the analysis drive the theory. Which variables are the **best** predictors should be based on previous research, or as a minimum, on a consensus of the opinions of subject matter experts. Some of the decision will rest on how large is your sample size. If the size is sufficiently large, you could take a subgroup and check for associations

between the independent variables and the dependent variables. When you run multiple regression, you do risk an error with each step of the analysis, so it is important to not just throw everything you have into the regression. If you are able to work with a subgroup, you can then verify what you think you have found with a different group for confirmation. Could you tell us a little more about your sample?

answered Jul 16 '13 at 22:20

doug.numbers
763 4 15