Methods in Epidemiologic Research

Sample Problems

Chapter 15 - Model Building

Preparation

We will use the -mi- dataset and use linear regression models to evaluate how various factors influence the length of hospital stay. Because the data are strongly right skewed we will log transform the outcome variable (-los-). Since it has some values of 0, which would be lost with a log transform, we will first add 0.5 to these values (this assumes that people actually spent a minimum of 1/2 day in hospital.

If you are using Stata, some code to perform this transformation is:

```
gen los_ln=cond(los==0, log(los+0.5), log(los))
```

The variable we will use in this exercise are listed below. The outcome will be -los_ln- created above.

```
Contains data from C:\mer\data\mi_los.dta
  obs:         2,965
 vars:            10                         28 Feb 2012 10:53
 size:        83,020
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
id              float  %9.0g                  patient id
sex             byte   %8.0g                  gender
age             float  %9.0g                  age at admission
white           float  %9.0g                  race=White
mar_c2          float  %9.0g                  married Y/N
bmi             float  %9.0g                  body mass index
prmi            byte   %8.0g                  previous MI
card            byte   %8.0g                  cardiac arrest during hosp.
cabg            byte   %8.0g                  coronay artery bypass surgery
los_ln          float  %9.0g
-------------------------------------------------------------------------------
Sorted by:
```

Note: Because the length of stay was recorded as an integer variable (number of days), these data could also be evaluated using procedures for count data (Chapter 18). The interval from admission to discharge could also be treated as time-to-event data (Chapter 19). We will return to these data in the exercises for those chapters. In addition, it is important to note that, for the purpose of this exercise, we are ignoring the possible clustering of lengths of stay within hospital (ie. some hospitals may have, on average, longer stays than others). We will evaluate the impact of this in the exercises for Chapter 22.

Questions

Your primary interest is how marital status (married vs no-married) (-mar_c2-) and body mass index (-bmi-) influence the length of stay. However, we also investigate the role of other factors.

1. Draw a causal diagram incorporating all of the predictors listed above.

2. Identify which variables are potential confounders for the effects of -mar_c2- and -bmi- on length of stay.
3. Identify any variables that might be considered intervening variables for the main relationships of interest.
4. Identify any exposure-independent variables.
5. Evaluate the functional form of the relationship between the continuous predictors (-age-, and -bmi-) and the outcome -los_ln-. Do either (or both) need to be included in the model in a non-linear fashion? For this evaluation do the following
    (a) generate lowess curves of the relationships
    (b) create quadratic terms and add them to the model
    (c) Optional - use fractional polynomials to determine the best functional form for the relationship. Include all other predictors in the models used in this process.
6. Evaluate:
    (a) descriptive statistics of all variables
    (b) unconditional associations between each of the predictors and the outcome (-los_ln-)
    (c) pairwise correlation analysis among all predictors
    (d) If you were looking to eliminate potential predictors at this stage, are there any likely candidates? (Because we don't have a large number of predictors we wont eliminate any predictors from consideration yet.)
7. What 2-way interactions do you want to examine?
8. Using all predictors, use forward selection, backward elimination and stepwise selection procedures to identify potential models for further investigation. Do the three procedures produce identical models?
9. Evaluate potential confounding effects by forcing all removed predictors that may be confounders, back into the model. Do any of them need to be kept, even though not statistically significant, because they appear to exert a confounding effect?
10. In question 3, you identified intervening variables for the effects of -mar_c2- and -bmi-. What happens to the effects of these two main predictors when intervening variables are removed? Should you leave these intervening variables in or remove them?
11. From here on, we will focus on the effect of body mass index -bmi- on -los_ln-. Identify the model which best evaluates this effect and carry out model diagnostics (as a minimum evaluate: assumptions of homoscedasticity and normality, VIF)
12. Evaluate the reliability of the model. (We will assume that you have already evaluated the validity of the model using the usual regression diagnostics.)