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:
```

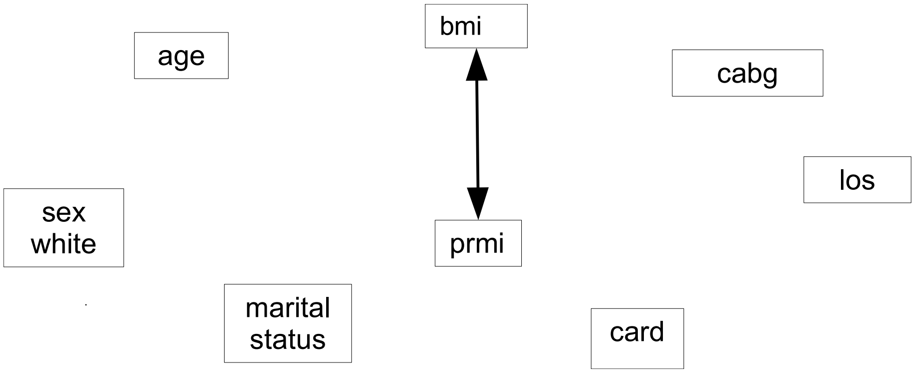
Note: Because the length of stay was recorded as an integer variable (number of days), these data could also be evaluated using procedures for count data (Chapter 18). The interval from admission to discharge could also be treated as time-to-event data (Chapter 19). We will return to these data in the 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.

In the following diagram, lines have been left off for clarity. However it was assumed that factors would influence all factors to the right of them (eg age would influence marital status, bmi, prmi, card, cabg and los). The exception is the relationship between -bmi- and -prmi- in that either may have influenced the other (ie bmi might have influenced the risk of an MI,b ut a previous MI might also influence bmi) ... so a bidirectional arrow was included.



2. Identify which variables are potential confounders for the effects of -mar_c2- and -bmi- on length of stay.

   Gender (-sex-), race (-white-) and -age- would be potential confounders of the effects of -mar_c2- and -bmi- on length of stay (-los_ln-).

3. Identify any variables that might be considered intervening variables for the main relationships of interest.

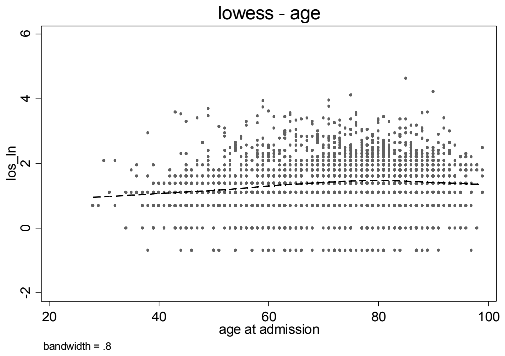
   -card- and -cabg- would definitely be intervening variables. Previous myocardial infarction (-prmi-) may be as well, but this is not as clear, so it will not be treated as an intervening variable.

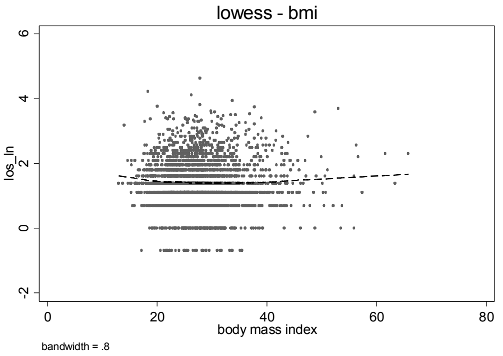
4. Identify any exposure-independent variables.

   None of -age-, -sex-, -white- would be expect to be independent of the main factors of interest.

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





Both variables appear to have some curvature in their relationship with los_ln

(b) create quadratic terms and add them to the model

The code for creating the quadratic terms, and the model including them is shown here.

```
.         sum age bmi

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |      2894    71.56531    14.41251         28         99
         bmi |      2705     27.9263    6.121246   12.98465   65.82788

. gen age_ct=age-72
. gen age_sq=age_ct^2
. gen bmi_ct=bmi-28
. gen bmi_sq=bmi_ct^2
. reg los_ln age_ct age_sq bmi_ct bmi_sq

      Source |       SS       df       MS              Number of obs =    2639
-------------+------------------------------           F(  4,  2634) =   26.19
       Model |  53.9862827     4  13.4965707           Prob > F      =  0.0000
    Residual |  1357.61301  2634  .515418759           R-squared     =  0.0382
-------------+------------------------------           Adj R-squared =  0.0368
       Total |  1411.59929  2638  .535102083           Root MSE      =  .71793

------------------------------------------------------------------------------
      los_ln |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      age_ct |   .0070706   .0011552     6.12   0.000     .0048054    .0093358
      age_sq |  -.0002535   .0000641    -3.96   0.000    -.0003791   -.0001279
      bmi_ct |   .0020034   .0027549     0.73   0.467    -.0033986    .0074053
      bmi_sq |   .0003612    .000199     1.81   0.070     -.000029    .0007514
       _cons |   1.457883   .0201617    72.31   0.000     1.418349    1.497417
------------------------------------------------------------------------------
```

It is clear that the quadratic term for age is significant, but that for bmi is borderline.

(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.

The Stata code for generating the fractional polynomials and the results from these analyses are summarized below (details of models not shown)

fracpoly, compare: reg los_ln age sex white mar_c2 bmi prmi card cabg

** borderline evidence (P=0.035) that a 2-term model (powers 3 3)*

** better than a single term*

fracpoly, compare degree(1): reg los_ln age sex white mar_c2 bmi prmi card cabg

** borderline evidence (P=0.041) that a transfromed single term*

** (power -0.5) better than linear*

** conclusion - 2-term too complex, use transformed single term*

gen age_nsq=age^-0.5

fracpoly, compare : reg los_ln bmi age_nsq sex white mar_c2 prmi card cabg

** no evidence that 2 terms superior (P=0.128)*

fracpoly, compare degree(1): reg los_ln bmi age_nsq sex white mar_c2 prmi card cabg

** minimal evidence (P=0.1) that transformed variable required*

** conclusion - leave -bmi- as linear*

***Note:** in all subsequent models we will use the negative square root transformation on -age- and will leave -bmi- as a linear effect.*

6. Evaluate:

(a) descriptive statistics of all variables

```
. sum los_ln bmi age_nsq sex white mar_c2 prmi card cabg

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      los_ln |      2963    1.384819     .758338  -.6931472   4.624973
         bmi |      2705     27.9263    6.121246   12.98465   65.82788
     age_nsq |      2894    .1203117    .0139225   .1005038   .1889822
         sex |      2965    .5514334    .4974315          0          1
       white |      2965    .8779089    .3274462          0          1
-------------+--------------------------------------------------------
      mar_c2 |      2954    .5118483    .4999442          0          1
        prmi |      2965    .2816189    .4498644          0          1
        card |      2965    .0698145    .2548771          0          1
        cabg |      2965    .0549747    .2279693          0          1
```

important observations are:

- -bmi- has some missing data

- only 12% of subjects are non-white and <7% had -card- or -cabg- so there may be limited power to evaluate these factors

(b) unconditional associations between each of the predictors and the outcome (-los_ln-)

Details are not shown, but the P-values for each variable were:

```
age_nsq   /* P=0.00 */
bmi       /* P=0.73 */
sex       /* P=0.00 */
white     /* P=0.21 */
```

```
mar_c2    /* P=0.00 */
prmi      /* P=0.17 */
card      /* P=0.14 */
cabg      /* P=0.00 */
```

Implications discussed below

(c) pairwise correlation analysis among all predictors

```
.                 pwcorr bmi age_nsq sex white mar_c2 prmi card cabg

             |      bmi  age_nsq      sex    white   mar_c2     prmi     card
-------------+---------------------------------------------------------------
         bmi |   1.0000
     age_nsq |   0.2589   1.0000
         sex |   0.0296   0.2587   1.0000
       white |   0.0119  -0.1104   0.0054   1.0000
      mar_c2 |   0.0466   0.1518   0.3019   0.0388   1.0000
        prmi |  -0.0492  -0.1309  -0.0067  -0.0276  -0.0374   1.0000
        card |   0.0041  -0.0650   0.0076  -0.0029  -0.0131  -0.0068   1.0000
        cabg |   0.0514   0.0410   0.0569   0.0041   0.0300  -0.0293  -0.0196
```

Strictly speaking, it is not appropriate to use correlation coefficients to evaluate relationships involving one or two dichotomous variables. However, it is common practice to do so, with the assumption being that this will identify pairs of variables that are strongly related. There are no strong relationships in these data (max rho=0.302).

(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.)

Ignoring the issue of intervening variables (see question 10), the three variables with P>0.15 are

- bmi - which we will retain because it is one of our primary risk factors of interest

- white - which we will retain because it is a potential confounder

- prmi - which we will retain because it's relationship with bmi is unclear and it is close to borderline significant.

7. What 2-way interactions do you want to examine?

With 8 predictors, there are 28 possible 2-way interactions to consider. Rather than evaluating all possible interactions (and applying a very strict "multiple comparison adjustment", I have chose to evaluate 6 interactions as follows:

- evaluate interactions between -mar_c2- and -bmi- and possible confounders

- evaluate interactions between -mar_c2- and -bmi-

- ignore the possibility of an interaction between -age_nsq- and -bmi- because it is difficult to meet the assumptions required for interactions between 2 continuous variables. The Stata code to evaluate these and the results are:

```
* interactions invovling -mar_c2-reg los_ln i.mar_c2##c.age_nsq sex white bmi
     prmi card cabg/* P=0.014 */
     reg los_ln i.mar_c2##i.sex age_nsq white bmi prmi card cabg   /* P=0.041 */
      los_ln i.mar_c2##i.white age_nsq sex bmi prmi card cabg /* P=0.700 */
* Interactions involving -bmi-
```

```
     reg los_ln i.sex##c.bmi age_nsq white mar_c2 prmi card cabg   /* P=0.861 */
     reg los_ln i.white##c.bmi age_nsq sex mar_c2 prmi card cabg   /* P=0.012 */
* interactions between -mar_c2- and -bmi-
     reg los_ln i.mar_c2##c.bmi age_nsq sex white prmi card cabg   /* P=0.080 */
```

The overall conclusion was that no interactions were significant once a Bonferroni correction (0.05/6=0.008) was applied.

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?

All procedures resulted in the same model (but this is not guaranteed. The table of coefficients from the 4 procedures is as follows.

```
. estimates table fs be sw, star(0.1 0.05 0.01)
----------------------------------------------------------------
    Variable |      fs               be               sw
-------------+--------------------------------------------------
      mar_c2 |  -.0742595***     -.0742595***    -.0742595***
         bmi |  .00347791        .00347791       .00347791
        cabg |  .97976943***     .97976943***    .97976943***
     age_nsq | -9.5778995***    -9.5778995***   -9.5778995***
       white |  -.1007887**      -.1007887**     -.1007887**
         sex | -.06438327**     -.06438327**    -.06438327**
       _cons |  2.5736026***     2.5736026***    2.5736026***
----------------------------------------------------------------
                            legend: * p<.1; ** p<.05; *** p<.01
```

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?

```
. estimates table full dr_sex dr_white dr_age , star(0.1 0.05 0.01)

--------------------------------------------------------------------------------
    Variable |      full           dr_sex         dr_white          dr_age
-------------+------------------------------------------------------------------
     age_nsq |  -9.420652***    -9.9750439***   -9.1120866***
         sex | -.06569266**                     -.06727767**     -.12642949***
       white | -.09946483**     -.10155771**                     -.04543388
      mar_c2 | -.07368327***    -.09085339***   -.07747361***    -.08533847***
         bmi |  .00345778        .00370084*       .0032292       -.00171824
        prmi |  .02049409         .0179612       .02282455        .05439032*
        card |  .05042032        .04760911       .05204125        .06803539
        cabg |  .98136133***     .97524212***    .98110271***     .99218207***
       _cons |  2.5454674***     2.5808407***    2.4294149***     1.5297572***
--------------------------------------------------------------------------------
                                            legend: * p<.1; ** p<.05; *** p<.01
```

Compared with a "full" model, models which drop:

- sex - no substantial changes in other coefficients (or their significance)

- white - no substantial changes in other coefficients (or their significance)

- age- substantial changes in the estimated effects of: -sex-. -white-, and -prmi-. there was also a quite a large change in the coefficient for -bmi-, but it was never significant so this change needs

to be interpreted with caution ("bouncing around zero").

This suggests age was an important confounder.

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?

Comparing the "full" model with one with intervening variables removed ...

```
.estimates table full no_int, star(.1 .05 .01)

----------------------------------------------
    Variable |     full           no_int
-------------+--------------------------------
     age_nsq |  -9.420652***   -9.4528932***
         sex | -.06569266**    -.04059306
       white | -.09946483**    -.09842061**
      mar_c2 | -.07368327***   -.07056093**
         bmi |  .00345778        .0053446**
        prmi |  .02049409       .00754503
        card |  .05042032
        cabg |  .98136133***
       _cons |  2.5454674***    2.5468422***
----------------------------------------------
           legend: * p<.1; ** p<.05; *** p<.01
```

Including intervening variables means that you only pick up the "direct" effect of the variables of interest, The most profound effect is seen in -bmi-. If we (incorrectly) include intervening variables, we would conclude that -bmi- had no significant effect on -los_ln-. However, removal of the intervening variables shows us that -bmi- is a significant predictor with a 1-unit increase in -bmi- resulting in a 0.0053 increase in los_ln.

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)

We won't show all of the output (this was covered in Chapter 14), but the Stata code to carry out the evaluations of major assumptions, and the main results are as follows.

```
reg los_ln age_nsq sex white mar_c2 bmi prmi
     vif
          * no problem with collinearity
     predict rsta, rsta
     rvfplot, name(hetero, replace)
     estat hettest
          * graphic suggest increasing variance ... but test refutes this
     histogram rsta, normal name(normality1, replace)
     qnorm rsta , name(normality2, replace)
     swilk rsta
          * histogram suggest normality, but normal prob. plot
          * and test suggest otherwise
```

The main concern is with the normality of the residuals. However, it is common to find some evidence of departure from normality (especially in Shapiro-Wilks test) even when the assumption is close to being met (as seen in histogram). It is unlikely that this degree of violation of the assumption will have had a major effect on the model, but you might try other *transformations of the outcome to confirm this.*

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.)

Fitting the model to half of the data and predicting the outcome in the other half results in a 6% drop in the R^2 for the model (from 0.0401 to 0.0378). A table of the estimates fo the coefficients from models based on the two halves of the dataset is as follows.

```
. estimates table half otherhalf, star(.1 .05 .01)

---------------------------------------------
    Variable |     half          otherhalf    
-------------+-------------------------------
     age_nsq | -9.3270974***   -9.6591614***  
         sex | -.01399556      -.06648968     
       white |  -.1482086**    -.05036051     
      mar_c2 | -.05577795      -.08364053**   
         bmi |  .00657455**     .00398725     
        prmi |  .02532891      -.01124448     
       _cons |   2.512726***    2.5931788***  
---------------------------------------------
           legend: * p<.1; ** p<.05; *** p<.01
```

The coefficients for:

- sex - varies substantially but is never significant

- white- varies substantially, but this might be expected given the low proportion of "non-white" in the dataset

- mar_c2 - varies substantially - this is a concern given that it was one of our predictors of interest

- bmi - varies substantially - this is a concern given that it was one of our predictors of interest

- prmi - varies substantially but is never significant

Overall, while there is only a small drop in predictive ability when half the data are used to predict the other half, the variability of the estimates (particularly for the 2 main predictors of interest) is of concern.