Problem Set 10

The data set rehosp.csv has 48,470 observations on newly born infants. The following variables are available:

1) mage = mother's age
2) mom_dropout = 1 if mother has < HS education
3) mom_hs =1 if mother has HS and no college
4) mom_somecoll =1 if mother has 1-3 years of college, no BA or higher
5) mom_college = 1 if mother has BA or higher degree
6) lmean_govins = mean fraction of mothers from same zip code who are covered by Medicaid (a low income insurance program)
7) bmi = mother's BMI prior to pregnancy (bmi = weight/height-squared)
8) infant's 5 minute apgar score
9) bweight = infant's birthweight in grams
10) hospital = 1 if infant returns to hospital in year after birth

Your job in this assignment is to develop a predictive model for the probability that a new infant returns to hospital using the 9 other variables. *You will develop the model using the rehosp.csv data set and then test it on the "holdback" data set called test_rehosp.csv.* The holdback data set has approximately the same size and similar means etc for all the variables, BUT you will see that a model developed on rehosp does not necessarily work as well on test_rehosp.

We will award points for creative ways to use the data. For example, you may want to look at bins of birthweight, or bins of apgar5. You may also want to look at interactions of variables (like lmean_govins and bweight, for example). You may also want to consider *splines* or other kinds of basis functions for certain variables.

You can use OLS, lasso, or any other method for predicting readmission. The only restrictions are these:

a) you MUST show how you selected your model **using the rehosp data**
b) you MUST show RMSE for your final selected model on the **test_rehosp** data set

A prize will be awarded for the lowest RMSE model. In addition, your grade in this problem set will be used to award "bonus points" to help top up your midterm grades. The students with the top 10 models will get +10 points; those with models in the 11-20 range will get +6 points. Everyone who submits a problem set graded 5/10 or better will get +3 points.

```
. use rehospitalization

. desc

Contains data from rehospitalization.dta
  obs:        48,871
 vars:            10                          16 Apr 2019 15:53
 size:     1,661,614                          (_dta has notes)
-------------------------------------------------------------------------------
-
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
-
apgar5          byte    %10.0g                Apgar 5
mage            byte    %8.0g                 Mother's Age
lmean_govins    float   %9.0g
mom_dropout     float   %9.0g
mom_hs          float   %9.0g
mom_somecoll    float   %9.0g
mom_college     float   %9.0g
bweight         float   %9.0g
hospital        float   %9.0g
bmi             float   %9.0g
-------------------------------------------------------------------------------
-
Sorted by:

. sum

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      apgar5 |     48,470    8.912482    .4852047          0         10
       mage |     48,871    25.62929    4.983087         18         35
lmean_govins |     48,871    .4514875     .227822          0          1
 mom_dropout |     48,871    .1421088    .3491653          0          1
      mom_hs |     48,871     .270017    .4439728          0          1
-------------+---------------------------------------------------------
mom_somecoll |     48,871    .2590289     .438106          0          1
 mom_college |     48,871    .3288453    .4697985          0          1
     bweight |     48,871    3347.035    434.3749       1840       4441
    hospital |     48,871    .0815412    .2736673          0          1
         bmi |     48,871    23.61631    3.852591   11.34509   33.83044

. tab mage

   Mother's |
        Age |      Freq.     Percent        Cum.
------------+-----------------------------------
         18 |      2,900        5.93        5.93
         19 |      3,690        7.55       13.48
         20 |      3,600        7.37       20.85
         21 |      3,155        6.46       27.31
         22 |      2,966        6.07       33.38
         23 |      2,749        5.63       39.00
         24 |      2,680        5.48       44.48
         25 |      2,790        5.71       50.19
         26 |      2,689        5.50       55.70
         27 |      2,810        5.75       61.45
         28 |      3,044        6.23       67.67
         29 |      2,885        5.90       73.58
         30 |      2,794        5.72       79.29
         31 |      2,649        5.42       84.71
```

```
         32 |      2,304        4.71       89.43
         33 |      1,992        4.08       93.51
         34 |      1,759        3.60       97.10
         35 |      1,415        2.90      100.00
------------+-----------------------------------
      Total |     48,871      100.00

. reg hospital mage mom_* lmean_govins bmi apgar5 bweight
note: mom_college omitted because of collinearity

      Source |       SS           df       MS      Number of obs   =    48,470
-------------+----------------------------------   F(8, 48461)     =     26.56
       Model |  15.8234522         8  1.97793153   Prob > F        =    0.0000
    Residual |  3608.92802    48,461  .074470771   R-squared       =    0.0044
-------------+----------------------------------   Adj R-squared   =    0.0042
       Total |  3624.75148    48,469  .074784945   Root MSE        =    .27289

------------------------------------------------------------------------------
    hospital |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        mage |  -.0012384   .0003213    -3.85   0.000    -.0018682   -.0006087
 mom_dropout |   .0204028    .004823     4.23   0.000     .0109495     .029856
      mom_hs |   .0075903   .0040441     1.88   0.061    -.0003362    .0155168
mom_somecoll |  -.0002199   .0036667    -0.06   0.952    -.0074066    .0069669
 mom_college |          0  (omitted)
lmean_govins |   .0295518   .0063116     4.68   0.000      .017181    .0419225
         bmi |   .0002332   .0003296     0.71   0.479    -.0004129    .0008793
      apgar5 |  -.0097088   .0025554    -3.80   0.000    -.0147174   -.0047001
     bweight |  -.0000163   2.89e-06    -5.64   0.000    -.0000219   -.0000106
       _cons |   .2304142   .0273996     8.41   0.000     .1767107    .2841178
------------------------------------------------------------------------------
```