Modern Methods in Biostatistics and Epidemiology
Missing data in observational and randomized studies
Lab 2 Sample Solution

Nicholas J. Horton
Amherst College

May 30, 2014

Part A: Describing missingness

Before we start to account for missing data, we need to first describe it in a clear and comprehensible manner, then fit a complete case model. We will undertake these preliminary steps using the Health Services (`routine`) dataset.

We will focus on predictors of routine discharge (yes/no) for these pediatric inpatients. Key covariates include: the length of stay (in days, `los`), age (in years), weekend admission (`aweekend`), gender (`female`), number of medical diagnoses (`ndx`) and total charges (`totchg`). The latter variable is partially observed.

We begin by reading in the dataset and keeping only these 6 variables.

```
. use https://www.amherst.edu/~nhorton/data/routine
. keep routine age aweekend female los ndx totchg
```

1. Provide a short but comprehensive summary of each of these seven variables. For continuous variables, include a graphical display of your choice as well as appropriate numerical summaries. For the categorical variables `aweekend` and `female` provide a description of the percentage in each group.

```
. summarize

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |     13477    16.32196    2.709657         10         20
    aweekend |     13477    .1964087    .3972959          0          1
      female |     13477    .5362469    .4987029          0          1
         los |     13477    6.459375    11.89629          0        339
         ndx |     13477    3.452697    1.994336          1         16
-------------+--------------------------------------------------------
      totchg |     13004    9242.434    16714.29         26     459786
     routine |     13477    .8645841    .3421799          0          1

. summarize los, detail

                  length of stay (cleaned)
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            1              0
```

```
10%            1              0       Obs               13477
25%            2              0       Sum of Wgt.       13477

50%            4                      Mean           6.459375
                        Largest       Std. Dev.      11.89629
75%            7            218
90%           13            249       Variance       141.5217
95%           19            285       Skewness       9.681498
99%           49            339       Kurtosis       155.1679
```

```
. summarize ndx, detail

             number of diagnoses on this record
-------------------------------------------------------------
      Percentiles      Smallest
 1%            1              1
 5%            1              1
10%            1              1       Obs               13477
25%            2              1       Sum of Wgt.       13477

50%            3                      Mean           3.452697
                        Largest       Std. Dev.      1.994336
75%            4             15
90%            6             15       Variance       3.977374
95%            7             15       Skewness       1.182669
99%           10             16       Kurtosis       4.833281
```

```
. summarize totchg, detail

                   total charges (cleaned)
-------------------------------------------------------------
      Percentiles      Smallest
 1%          730             26
 5%         1353             30
10%         1821             31       Obs               13004
25%         2991             36       Sum of Wgt.       13004

50%         5218                      Mean           9242.434
                        Largest       Std. Dev.      16714.29
75%       9619.5         305223
90%        18078         348279       Variance       2.79e+08
95%        26899         385820       Skewness       9.639567
99%        73272         459786       Kurtosis       151.1259
```

The mean age is 16.3 years (sd 2.7 years), with a range from 10–20. The mean length of stay was 6.5 days (sd=11.9 days, indicating dramatic skew). The length of stay ranged from 0 to 339 days, with the 99th percentile at 49 days. The number of diagnoses was also slightly skewed (mean 3.5, median 3, range 1–16, 99th percentile 10). The total charges were dramatically skewed, with a mean of $9,242 and a median of $5,218 and 99th percentile at $73,272.

Approximately 20% of the admissions were during the weekend, 54% of the sample was female, and 86% of the discharges were routine (aka not AMA or transfers or deaths).

Figure 1 displays the histogram of age for this sample, Figure 2 displays the histogram of length of

Figure 1: Histogram of age (in years)

```
. histogram age
(bin=41, start=10, width=.24390244)
```

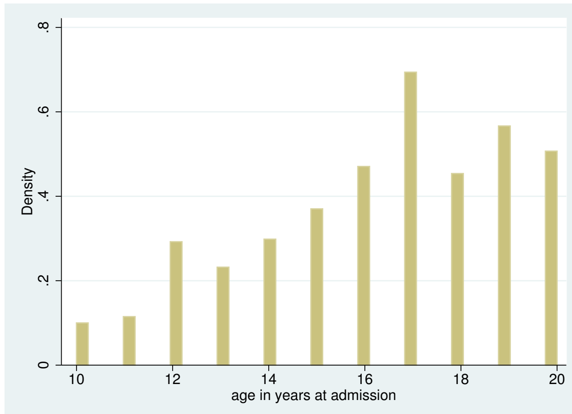


Figure 2: Histogram of length of stay (in days, pruned to include only those < 60)

```
. histogram los if los < 60
(bin=41, start=0, width=1.4390244)
```

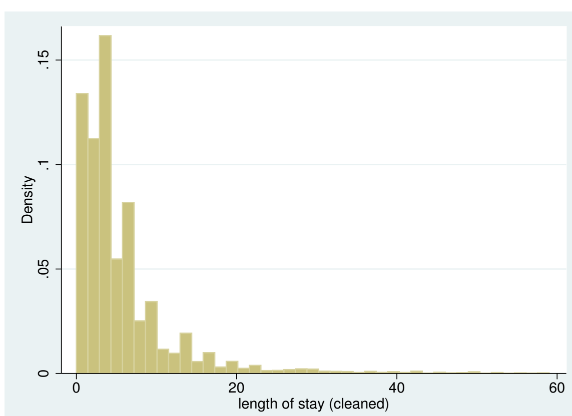
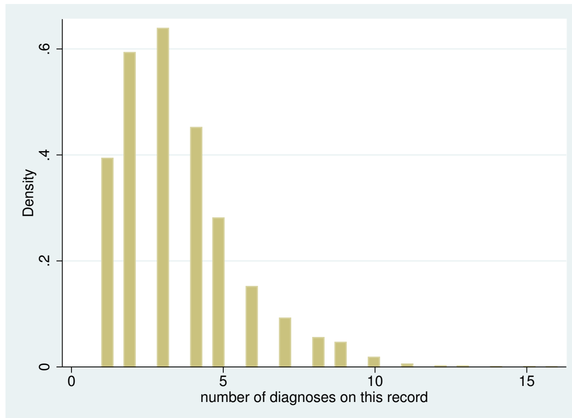


Figure 3: Histogram of number of medical diagnoses

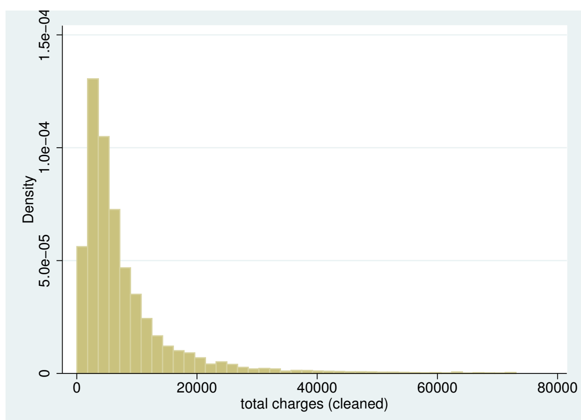
```
. histogram ndx
(bin=41, start=1, width=.36585366)
```



stay in the hospital while Figure 3 displays the histogram of number of medical diagnoses. Finally, Figure 4 displays the histogram of total charges.

Figure 4: Histogram of total charges (less than 99th percentile value of $73,272)

```
. histogram totchg if totchg < 73272
(bin=41, start=26, width=1781.439)
```



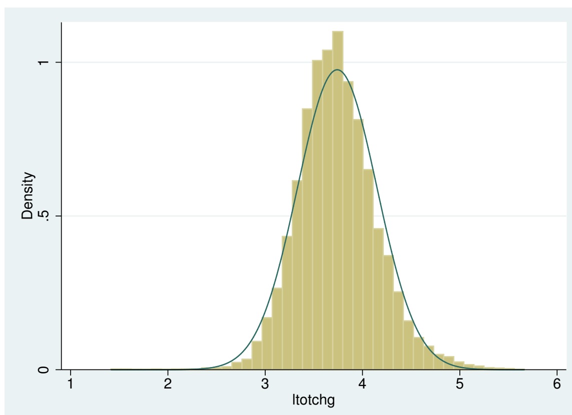
2. The total charges (`totchg`) are dramatically skewed (no excuses offered for the state of the United States health care system). Create a new variable called `ltotchg` which is the log base 10 of the total charge variable (hint: see the `log10()` function). Describe the shape, center and spread of the transformed variable as well as generating a histogram with superimposed normal density.

```
. gen ltotchg = log10(totchg)
. summarize ltotchg

(473 missing values generated)

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     ltotchg |     13004    3.740443    .4088494   1.414973   5.662556
```

The transformed total charge variable has a mean of 3.7 (less than $10,000 since `log10(10000)=4`), sd of 0.4 and shape that is approximately normal. Figure 5 displays the histogram of log total charges with superimposed normal density. Note that the distribution of predictors of a logistic (or linear) regression are not required to be normally distributed, but transforming our model may clarify the form of the associations with routine discharge.

Figure 5: Histogram of log total charges

```
. histogram ltotchg, normal

(bin=41, start=1.4149734, width=.10359957)
```



3. Fit and interpret the regression coefficients for the complete case model: `logistic routine age aweekend female los ndx ltotchg`.

```
. logistic routine age aweekend female los ndx ltotchg

Logistic regression                               Number of obs   =      13004
                                                  LR chi2(6)      =     175.46
                                                  Prob > chi2     =     0.0000
Log likelihood = -5085.8101                       Pseudo R2       =     0.0170

------------------------------------------------------------------------------
     routine | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
```

```
-------------+----------------------------------------------------------------
         age |   .9597191   .0094862    -4.16   0.000     .9413054     .978493
    aweekend |   1.054727   .0689272     0.82   0.415      .927926    1.198855
      female |   1.279471    .066214     4.76   0.000     1.156059    1.416059
         los |   .9874725     .00206    -6.04   0.000     .9834432    .9915184
         ndx |   .8867657   .0106406   -10.02   0.000     .8661538    .9078682
     ltotchg |   1.251245    .091979     3.05   0.002     1.083354    1.445154
       _cons |    7.93592   2.533765     6.49   0.000     4.244507    14.83772
------------------------------------------------------------------------------
```

Older age ($p < 0.001$) is statistically significantly associated with higher probability of routine discharge (OR=0.96, 95% CI=0.94 to 0.98), as is gender ($p < 0.001$), shorter length of stay ($p < 0.001$), fewer diagnoses ($p < 0.001$) and increased log total charges ($p = 0.002$). After controlling for these other factors, weekend admission was not statistically significant (OR=1.05, 95% CI=0.93 to 1.20).

4. Save the results from the logistic regression using the command:

```
. estimates store cc
```

5. Generate an indicator of missingness for `ltotchg` (hint: the command `misstable summarize, generate(miss_)` will generate a new variable `miss_ltotchg` which is set to 1 for observations missing log of total charges, and 0 for those that are fully observed.

```
. drop totchg
. misstable summarize, generate(miss_)

                                                               Obs<.
                                                +------------------------------
               |                                | Unique
      Variable |     Obs=.     Obs>.     Obs<.  | values        Min         Max
  -------------+--------------------------------+------------------------------
       ltotchg |       473              13,004  |   >500   1.414973    5.662556
  -----------------------------------------------------------------------------

. describe miss_*

              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
miss_ltotchg    byte    %8.0g                 (ltotchg>=.)
```

6. What variables are associated with missingness? (Hint: fit a logistic regression model predicting the outcome `miss_ltotchg`).

```
. logistic miss_ltotchg routine age aweekend female los ndx

Logistic regression                               Number of obs   =      13477
                                                  LR chi2(6)      =      23.03
                                                  Prob > chi2     =     0.0008
Log likelihood =  -2037.469                       Pseudo R2       =     0.0056

------------------------------------------------------------------------------
miss_ltotchg | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

```
     routine |   1.178531   .1727279     1.12   0.262     .8842744    1.570705
         age |   .9720152   .0169578    -1.63   0.104     .9393404    1.005827
    aweekend |   .8585185   .1057357    -1.24   0.215     .6743962    1.092909
      female |   .9914006   .0936312    -0.09   0.927     .8238703    1.192998
         los |    .969689   .0084456    -3.53   0.000     .9532764    .9863842
         ndx |   1.023069   .0242752     0.96   0.336     .9765799    1.071771
       _cons |   .0564077   .0186651    -8.69   0.000     .0294903    .1078941
------------------------------------------------------------------------------
```

```
. di 1 - .969689
```

```
.030311
```

We observe that little is predictive of missing the total charges: only length of stay is statistically significant ($p < 0.001$). We predict that for every additional day of stay, the odds of observing the total charges decreases by $100.00 - 96.97 = 3.03\%$ (95% CI from 1.4% to 4.7%)

We can also try to assess what is predictive of log of total charges through a linear regression model among the complete cases:

```
. regress ltotchg routine age aweekend female los ndx
```

```
      Source |       SS       df       MS              Number of obs =   13004
-------------+------------------------------           F(  6, 12997) =  991.93
       Model |  682.692007     6  113.782001           Prob > F      =  0.0000
    Residual |  1490.86089 12997  .114708078           R-squared     =  0.3141
-------------+------------------------------           Adj R-squared =  0.3138
       Total |   2173.5529 13003  .167157802           Root MSE      =  .33869

------------------------------------------------------------------------------
     ltotchg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     routine |   .0321034   .0087223     3.68   0.000     .0150065    .0492003
         age |  -.0065459   .0011173    -5.86   0.000     -.008736   -.0043558
    aweekend |   .0033279   .0075108     0.44   0.658    -.0113944    .0180502
      female |  -.0291525    .005978    -4.88   0.000    -.0408702   -.0174347
         los |   .0184093   .0002494    73.83   0.000     .0179205     .018898
         ndx |   .0200639   .0015052    13.33   0.000     .0171135    .0230144
       _cons |   3.645295   .0208296   175.01   0.000     3.604466    3.686124
------------------------------------------------------------------------------
```

With the exception of weekend admission ($p = 0.66$), all of the other variables are statistically significant predictors of log total charges.

Here I would include all of the variables in the imputation model (and potentially others in the dataset), though I would not argue if you dropped `aweekend` from future consideration.

7. Set up Stata to undertake the imputations using the following commands:

```
. mi set wide
. mi register imputed ltotchg
. mi register regular routine age aweekend female los ndx
. mi describe
```

```
  Style:  wide
          last mi update 30may2014 09:50:09, 0 seconds ago

  Obs.:   complete       13,004
          incomplete        473  (M = 0 imputations)
          ---------------------
          total          13,477

  Vars.:  imputed:  1; ltotchg(473)

          passive:  0

          regular:  6; routine age aweekend female los ndx

          system:   1; _mi_miss

         (there are 2 unregistered variables; _est_cc miss_ltotchg)
```

8. Fit an imputation model to fill in the missing `ltotchg` values using the `regress` command. Generate 25 imputations, and use a random seed value of 1964.

```
. mi impute regress ltotchg routine age aweekend female los ndx, add(25) rseed(1964)

Univariate imputation                        Imputations =       25
Linear regression                                  added =       25
Imputed: m=1 through m=25                        updated =        0

------------------------------------------------------------------
                   |               Observations per m
                   |----------------------------------------------
          Variable |   Complete   Incomplete   Imputed |     Total
-------------------+-----------------------------------+----------
           ltotchg |      13004          473       473 |     13477
------------------------------------------------------------------
(complete + incomplete = total; imputed is the minimum across m
 of the number of filled-in observations.)
```

9. Fit the logistic regression model using these imputed values and store the results.

```
. mi estimate, post: logistic routine age aweekend female los ndx ltotchg
. estimates store mireg

Multiple-imputation estimates                     Imputations     =         25
Logistic regression                               Number of obs   =      13477
                                                  Average RVI     =     0.0024
                                                  Largest FMI     =     0.0167
DF adjustment:   Large sample                     DF:     min     =   86265.60
                                                          avg     =   1.39e+10
                                                          max     =   7.10e+10
Model F test:       Equal FMI                     F(   6, 1.7e+07)=      30.19
Within VCE type:          OIM                     Prob > F        =     0.0000

------------------------------------------------------------------------------
```

```
     routine |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0438771   .0097091    -4.52   0.000    -.0629066   -.0248476
    aweekend |   .0456617   .0642702     0.71   0.477    -.0803056     .171629
      female |   .2519158   .0509513     4.94   0.000     .1520531    .3517784
         los |  -.0124888   .0020711    -6.03   0.000    -.0165481   -.0084295
         ndx |  -.1182045   .0118176   -10.00   0.000    -.1413666   -.0950424
     ltotchg |   .2260574   .0729831     3.10   0.002     .0830112    .3691036
       _cons |   2.105795   .3159207     6.67   0.000     1.486597    2.724992
------------------------------------------------------------------------------
```

10. Calculate and interpret the fraction of missing information for each of the parameters. The following Stata code will be helpful:

```
. matrix list e(fmi_mi)

e(fmi_mi)[1,7]
      routine:   routine:   routine:   routine:   routine:   routine:
          age   aweekend     female        los        ndx    ltotchg
r1  .00004714  .00001838  .00003986  .00517454  .00039635  .01670244

      routine:
        _cons
r1  .01175051
```

How do these values relate to the fraction of the sample that are missing total charges?

```
. di 473/(13004 + 473)

.03509683
```

The highest fraction of missing information is for the `ltotchg` variable (value is 0.0167). We note that this is considerably smaller than the observed proportion of subjects that are missing this variable, indicating that we are able to recover a considerable amount of information regarding the distribution among the unobserved from relationships amongst the observed subjects. (Note that this extrapolation relies directly on the MAR [“missing at random"] assumption).

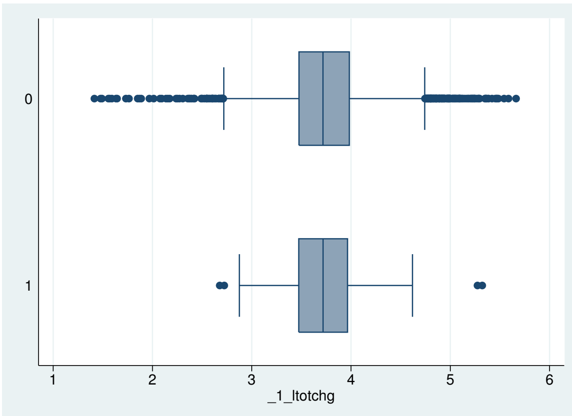
A related concept is the relative variance increase (RVI), which is also available from the return values from `mi estimate`:

```
. matrix list e(rvi_mi)

e(rvi_mi)[1,7]
      routine:   routine:   routine:   routine:   routine:   routine:
          age   aweekend     female        los        ndx    ltotchg
r1  .00004714  .00001838  .00003987  .00519921  .00039649  .01696257

      routine:
        _cons
r1   .0118786
```

With the exception of log total charges (1.7%) and the constant (1.2%), all of the parameter specific RVI’s are no more than (0.5%).

11. Compare the distribution of the imputed values from the first imputation for `ltotchg` to the observed values using a boxplot:

```
. graph hbox _1_ltotchg, over(_mi_miss)
```



The imputations seem to be less variable in the tails than the observed values (though the middle 50% have similar center and spread).

12. Replace the imputations with a set of 25 more that use predictive mean matching rather than the regression method (use a seed of 1965).

```
. mi impute pmm ltotchg routine age aweekend female los ndx, replace rseed(1965)
. mi estimate, post: logistic routine age aweekend female los ndx ltotchg
. estimates store mipmm
```

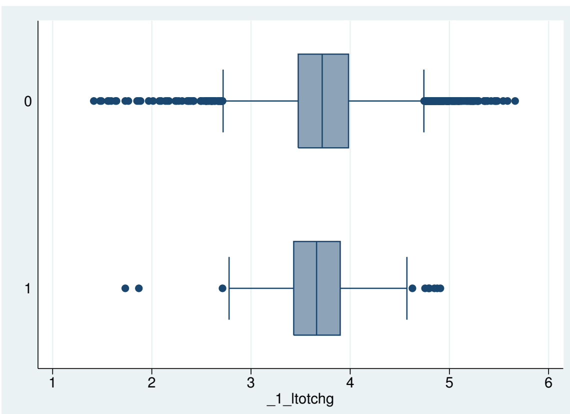
```
Univariate imputation                           Imputations =       25
Predictive mean matching                              added =        0
Imputed: m=1 through m=25                           updated =       25

                                          Nearest neighbors =        1

------------------------------------------------------------------
                   |               Observations per m
                   |----------------------------------------------
          Variable |   Complete   Incomplete   Imputed |     Total
-------------------+-----------------------------------+----------
           ltotchg |      13004          473       473 |     13477
------------------------------------------------------------------
(complete + incomplete = total; imputed is the minimum across m
 of the number of filled-in observations.)

Multiple-imputation estimates                     Imputations     =         25
Logistic regression                               Number of obs   =      13477
                                                  Average RVI     =     0.0040
                                                  Largest FMI     =     0.0272
DF adjustment:   Large sample                     DF:     min     =   32496.78
                                                          avg     =   4.66e+10
                                                          max     =   3.17e+11
Model F test:       Equal FMI                     F(   6, 6.3e+06)=      30.41
```

```
Within VCE type:          OIM                        Prob > F        =     0.0000

------------------------------------------------------------------------------
     routine |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0436202   .0097129    -4.49   0.000    -.0626571   -.0245833
    aweekend |   .0452641   .0642734     0.70   0.481    -.0807094    .1712376
      female |   .2529341   .0509615     4.96   0.000     .1530514    .3528169
         los |  -.0128611   .0020862    -6.16   0.000    -.0169501   -.0087721
         ndx |  -.1187721   .0118246   -10.04   0.000    -.1419478   -.0955963
     ltotchg |   .2487348   .0734925     3.38   0.001     .1046868    .3927827
       _cons |    2.02127    .317667     6.36   0.000     1.398642    2.643898
------------------------------------------------------------------------------
```

Note that to specify a different number of donors for the PMM, you would add the `kmm()` option.

13. Compare the distribution of the imputed values from the first imputation for `ltotchg` to the observed values using a boxplot:

```
. graph hbox _1_ltotchg, over(_mi_miss)
```



The distributions appear to be slightly shifted, with similar variance.

14. Display the estimates for the three models. How do they compare? How do they differ? Which one do you prefer?

```
. estimates table cc mireg mipmm, b se

----------------------------------------------------
    Variable |     cc          mireg        mipmm
-------------+--------------------------------------
         age | -.04111468   -.04387713   -.04362018
             |  .00988436    .00970911    .00971287
    aweekend |  .05328187    .04566175    .04526408
             |  .06535079    .06427021    .06427338
      female |  .24644706    .25191577    .25293414
             |  .05175109    .05095128    .05096153
         los | -.01260658   -.01248882   -.01286111
             |  .00208616     .0020711    .00208624
         ndx | -.12017444   -.11820451   -.11877206
```

```
             |   .01199938    .01181763    .01182458
     ltotchg |   .22413913    .22605742    .24873478
             |      .07351    .07298306    .07349247
       _cons |   2.0713992    2.1057949      2.02127
             |   .31927804    .31592067    .31766698
-----------------------------------------------------
                                       legend: b/se
```

The two imputation methods have very similar results in terms of the parameter estimates and standard errors. The complete case estimator has slightly different parameter estimates as well as slightly larger standard errors (though all three methods have similar standard errors for the variable with incomplete data).

This is not surprising since there was a small fraction of missing values, and little was predictive of missingness (except length of stay).

We will return to this example later in the course when we consider multivariate imputation models.