Modern Methods in Biostatistics and Epidemiology
Missing data in observational and randomized studies
Lab 1 Sample Solution

Nicholas J. Horton
Amherst College

June 8, 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 a subset of the Health Services (`routine`) dataset.

We will focus on the question of whether it is possible to predict the length of stay (in days, `los`) for these subjects as a function of whether it was a routine discharge (`routine`), age (in years), weekend admission (`aweekend`), gender (`female`), number of medical diagnoses (`ndx`) and subject race (partially observed, `race`, where 1=white, 2=black, 3=hispanic, 4=other). 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 race

. 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
-------------+--------------------------------------------------------
        race |     11268    1.523518    .8767465          1          4
     routine |     13477    .8645841    .3421799          0          1
```

1. Add labels to ensure that the `race` variable is clearer (hint: use the `label define` and `label values` commands).

```
. label define racegrp 1 "white" 2 "black" 3 "hispanic" 4 "other"
. label values race racegrp
. tabulate race
```

```
       race |
  (uniform) |      Freq.     Percent        Cum.
------------+-----------------------------------
      white |      7,706       68.39       68.39
      black |      1,813       16.09       84.48
   hispanic |      1,161       10.30       94.78
      other |        588        5.22      100.00
------------+-----------------------------------
      Total |     11,268      100.00
```

2. 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`, `female`, `race` and `routine` provide a description of the percentage in each level of the factor.

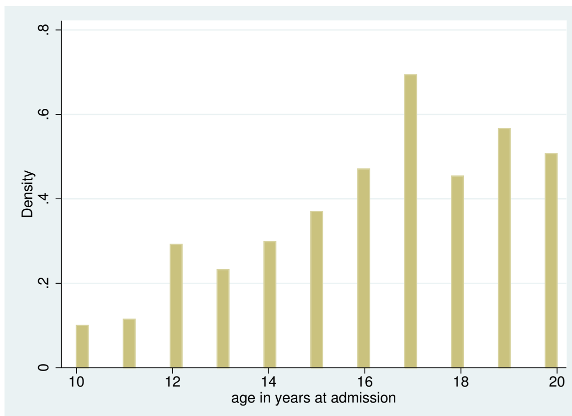
```
. tabulate aweekend
. tabulate female
. tabulate routine

  admission |
   day is a |
    weekend |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |     10,830       80.36       80.36
          1 |      2,647       19.64      100.00
------------+-----------------------------------
      Total |     13,477      100.00

  indicator |
     of sex |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      6,250       46.38       46.38
          1 |      7,227       53.62      100.00
------------+-----------------------------------
      Total |     13,477      100.00

    routine |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,825       13.54       13.54
          1 |     11,652       86.46      100.00
------------+-----------------------------------
      Total |     13,477      100.00

. summarize age los ndx

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
```

```
         age |     13477    16.32196    2.709657         10         20
         los |     13477    6.459375    11.89629          0        339
         ndx |     13477    3.452697    1.994336          1         16
```

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

Figure 1: Histogram of age (in years)

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



length of stay in the hospital while Figure 3 displays the histogram of number of medical diagnoses.

Of the 13,477 observations, approximately 20% of the admissions are on a weekend, while 54% of the subjects are female. Most (86.5%) of the discharges were routine. Subjects ranged from 10 to 20 years old, with a mean age of 16.3 years and standard deviation of 2.7 years. The length of stay and number of diagnoses were both skewed with long right tails (mean 6.5 for length of stay [in days] and 3.5 for number of diagnoses, with sd 11.9 and 2.0, respectively).

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)
```

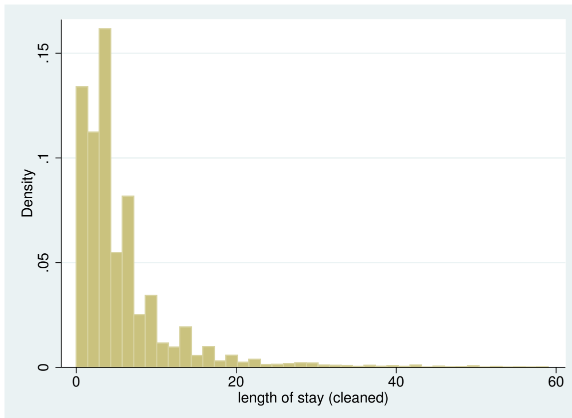
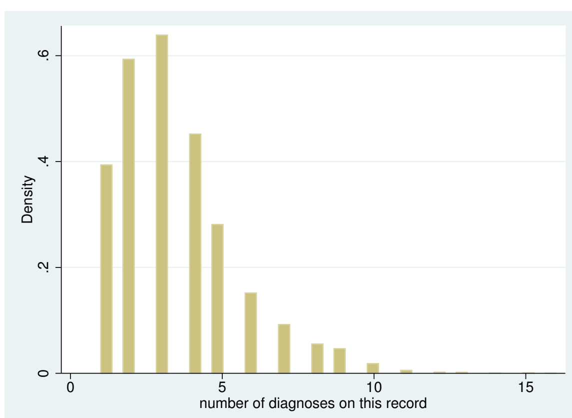


Figure 3: Histogram of number of medical diagnoses

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



3. Fit and interpret the regression coefficients for the complete case model: `regress los routine age female ndx i.race`.

```
. regress los routine age female ndx i.race
. test (2.race=0) (3.race=0) (4.race=0)

      Source |       SS       df       MS              Number of obs =   11268
-------------+------------------------------           F(  7, 11260) =   43.78
       Model |  46082.4603     7  6583.20861           Prob > F      =  0.0000
    Residual |  1693284.39 11260  150.380496           R-squared     =  0.0265
-------------+------------------------------           Adj R-squared =  0.0259
       Total |  1739366.85 11267  154.377106           Root MSE      =  12.263

------------------------------------------------------------------------------
         los |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     routine |  -1.880481   .3366137    -5.59   0.000    -2.540303    -1.22066
         age |  -.4676682   .0426833   -10.96   0.000     -.551335   -.3840014
      female |  -1.030675   .2322176    -4.44   0.000    -1.485862   -.5754881
         ndx |   .2994255   .0583796     5.13   0.000     .1849912    .4138597
             |
        race |
       black  |   3.103983   .3209864     9.67   0.000     2.474794    3.733173
    hispanic  |   1.000233   .3862419     2.59   0.010     .2431309    1.757334
       other  |   2.567519    .525042     4.89   0.000     1.538345    3.596693
             |
       _cons |   14.67111   .7963618    18.42   0.000     13.11011    16.23212
------------------------------------------------------------------------------

 ( 1)  2.race = 0
 ( 2)  3.race = 0
 ( 3)  4.race = 0

       F(  3, 11260) =   36.13
            Prob > F =    0.0000
```

Note that this model includes only the 11,268 subjects with complete data (since some subjects are missing race). The overall model is highly significant ($F(7, 11260) = 43.78, p < 0.0001$), though the R^2 value of 0.0265 is modest, indicating that the large sample size may yield statistically significant results that may not necessarily be clinically significant. All of the individual predictors are statistically significant ($p < 0.001$), including the overall test of race ($F(3, 11260 = 36.13, p < 0.0001$). After controlling for other factors, we see that routine discharge is associated with a length of stay that is 1.9 days shorter (95% CI -2.5 to -1.2 days), while length of stay tends to be shorter for older subjects (predicted decrease of 0.47 days, 95% CI=-0.55 to -0.38 days). Women tend to have a shorter length of stay (estimate=-1.03, 95% CI=-1.49 to -.58 days) while longer length of stay is associated with more diagnoses (estimate=0.30, 95% CI=0.18 to 0.41 days). Race/ethnicity is also associated with length of stay, with all non-white groups having longer stays (black 3.1 additional days, hispanic 1.0 additional days and other 2.6 additional days on average).

4. We can and should take a look at the residuals. These are straightforward to generate using

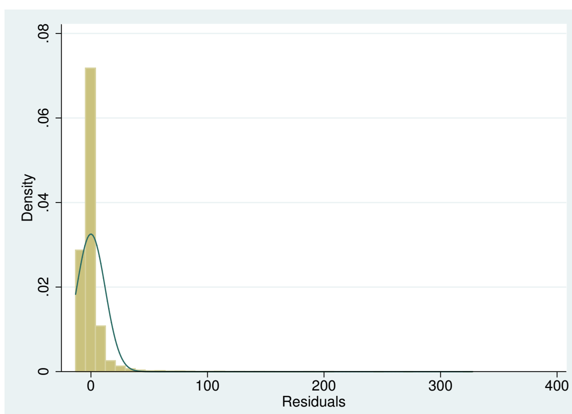
```
. predict resid, resid
. predict yhat, xb
. summarize resid

(2209 missing values generated)
(2209 missing values generated)

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       resid |     11268    2.26e-09    12.25916   -13.1608   327.5682
```

Undertake a residual analysis of this model, and include one graphic which sheds light on the goodness of fit.

Figure 4 displays the empirical density of residuals, while Figure 5 displays the empirical

Figure 4: Empirical density of residuals

```
. histogram resid, norm

(bin=40, start=-13.160802, width=8.5182245)
```

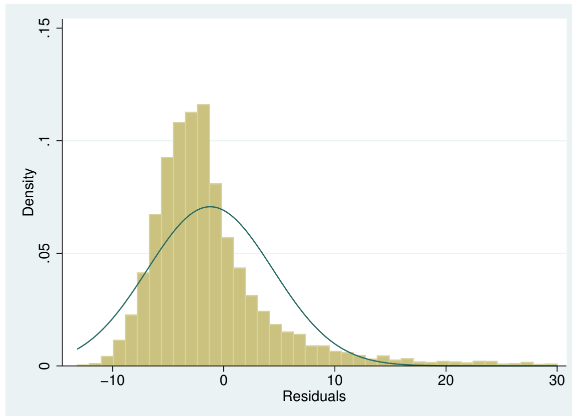


density of residuals less than 30 days. There is one dramatic outlier:

```
. list age female race los yhat resid if los > 300

          +---------------------------------------------------+
          | age    female    race    los       yhat     resid |
          |---------------------------------------------------|
```

Figure 5: Empirical density of residuals

```
. histogram resid if resid < 30, norm

(bin=40, start=-13.160802, width=1.078884)
```




```
10203. |  12        1   black   339   11.43183   327.5682 |
       +--------------------------------------------------+
```

Overall, we note that the residuals are moderately skewed.

Figure 6 displays the scatterplot of normalized residuals by predicted values, while Figure 7 displays the scatterplot of normalized residuals by number of diagnoses.

Figure 6: Scatterplot of residual by predicted value

```
. twoway scatter resid yhat || (lowess resid yhat)
```

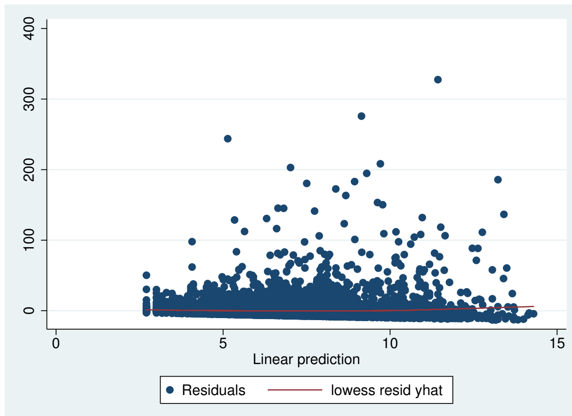
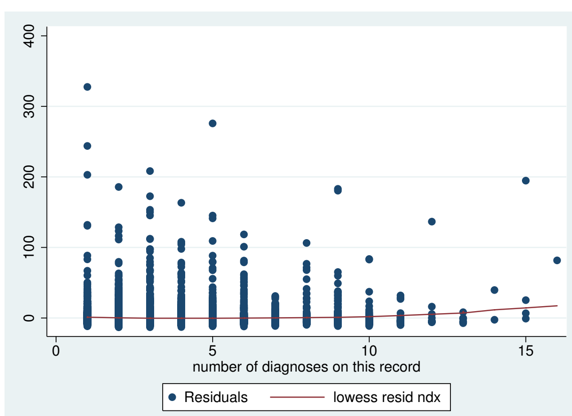


Figure 7: Scatterplot of residual by number of medical diagnoses

```
. twoway scatter resid ndx || (lowess resid ndx)
```



5. How might you improve the model?

   There are lots of possible improvements to consider. A transformation of the outcome variable (perhaps log base 10?) may help address the normality assumption of the residuals. A linear model for number of diagnoses seems plausible (but this might be allowed to vary in some fashion). In addition, possible interactions may merit investigation.

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

```
. misstable summarize, generate(miss_)

                                                               Obs<.
                                                +------------------------------
               |                                | Unique
      Variable |     Obs=.     Obs>.     Obs<.  | values        Min         Max
  -------------+--------------------------------+------------------------------
          race |     2,209              11,268  |      4          1           4
         resid |     2,209              11,268  |   >500   -13.1608    327.5682
          yhat |     2,209              11,268  |   >500   2.706018    14.29612
  -----------------------------------------------------------------------------

. describe miss_*

              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
miss_race       byte    %8.0g                 (race>=.)
miss_resid      byte    %8.0g                 (resid>=.)
miss_yhat       byte    %8.0g                 (yhat>=.)
```

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

```
. logistic miss_race los routine age female ndx

Logistic regression                               Number of obs   =      13477
                                                  LR chi2(5)      =      51.68
                                                  Prob > chi2     =     0.0000
Log likelihood = -5986.1863                       Pseudo R2       =     0.0043

------------------------------------------------------------------------------
   miss_race | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         los |   .9888999   .0028817    -3.83   0.000     .9832679    .9945642
     routine |   1.179211   .0840154     2.31   0.021     1.025524     1.35593
         age |   1.022221   .0090569     2.48   0.013     1.004623    1.040127
```

```
      female |   1.064992   .0501031     1.34   0.181     .9711831    1.167862
         ndx |   1.046204   .0120297     3.93   0.000      1.02289     1.07005
       _cons |   .1044303   .0175713   -13.43   0.000     .0750939    .1452275
------------------------------------------------------------------------------
```

With the exception of gender ($p = 0.18$), all of the other predictors are associated with missingness of the race/ethnicity variable. Subjects with shorter length of stay ($p < 0.001$), routine discharge ($p = 0.021$), older age ($p = 0.013$) and more diagnoses ($p < 0.001$) are more likely to be missing race.

To verify these results using a bivariate analysis, consider the 2x2 table defined by missing race and routine discharge:

```
. tabulate routine miss_race, row

+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+

           |      (race>=.)
   routine |         0          1 |     Total
-----------+----------------------+----------
         0 |     1,557        268 |     1,825
           |     85.32      14.68 |    100.00
-----------+----------------------+----------
         1 |     9,711      1,941 |    11,652
           |     83.34      16.66 |    100.00
-----------+----------------------+----------
     Total |    11,268      2,209 |    13,477
           |     83.61      16.39 |    100.00
```

In the bivariate analysis we observe that 15% of the subjects with non-routine discharge were missing race/ethnicity, as opposed to the nearly 17% with routine discharge.

```
. gen older = 0
. replace older = 1 if age >= 16
. mean miss_race, over(older)

(8848 real changes made)

Mean estimation                    Number of obs    =   13477

            0: older = 0
            1: older = 1
```

```
------------------------------------------------------------
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------
miss_race    |
           0 |   .1566213   .0053424      .1461494    .1670932
           1 |   .1677215   .0039722      .1599355    .1755076
------------------------------------------------------------
```

Similar results hold for age: 16% of subjects less than 16 are missing race, while 17% of subjects greater than or equal to 16 are unobserved.

Note that subject matter and specific study knowledge may help to explain much of this missingness and needs to be incorporated into any substantive analysis. For this health services dataset, it is likely that some states masked race/ethnicity for disclosure avoidance or it was not routinely collected. Inclusion of factors associated with missingness or the unobserved variable will be helpful in improving imputations that we will undertake in the future.

Part B: Reporting practice

What is the state of the art for missing data methods in your field? Take a sample of three quantitative articles that appeared within the last 5 years in the best electronically accessible journal in your field. For each of the papers, report:

1. Did the authors report missing values?
2. Was there (likely) missing data?
3. Was it clear how missing data were handled?
4. Were the appropriate approaches for missing data used?
5. Would Burton and Altman be pleased with your results? What is missing from their guidelines?
6. Would the journal be interested in your findings?

Results here will vary: I look forward to seeing what you found!