Homework # 3 Solutions

Biostatistics 210

Question 1:

(a) Using the wide format data, `hwk3 wide.dta` explore the association between missing data at year 1 (missing=1) with the predictors collected at the baseline visit. Summarize what you find. Are baseline factors associated with a missing blood pressure at yr1?

I explore this through logistic regression with missingness at year 1 as the outcome (missing) and baseline diabetes, BMI, and SBP as the predictors.

```
. logistic missing bmi0 sbp0 baseline_dm, nolog

Logistic regression                               Number of obs   =       2607
                                                  LR chi2(3)      =    1924.11
                                                  Prob > chi2     =     0.0000
Log likelihood = -840.00284                       Pseudo R2       =     0.5339

------------------------------------------------------------------------------
     missing | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bmi0 |   1.001015   .0116856     0.09   0.931     .9783716    1.024182
        sbp0 |   1.208036    .009221    24.76   0.000     1.190098    1.226245
 baseline_dm |   .8699221   .1267495    -0.96   0.339     .6538204     1.15745
       _cons |   7.80e-12   8.36e-12   -23.87   0.000     9.54e-13    6.37e-11
------------------------------------------------------------------------------
```

The baseline SBP appears to be the sole predictor of missing data at year 1 -- and it seems like it is a strong predictor.

(b) Does year 1 SBP (`sbp1`) appear to be associated with the indicator of missingness (`missing`)

It does.

```
. logistic missing sbp1

Logistic regression                               Number of obs   =       2607
                                                  LR chi2(1)      =     517.63
                                                  Prob > chi2     =     0.0000
Log likelihood = -1543.2439                       Pseudo R2       =     0.1436

------------------------------------------------------------------------------
     missing | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        sbp1 |   1.055902   .0029136    19.71   0.000     1.050207    1.061628
       _cons |   .0005646   .0002117   -19.95   0.000     .0002707    .0011774
------------------------------------------------------------------------------
```

This mean. Year 1 SBP is different between those with missing and available value. But I would note that Year 1 SBP is not associated with missing data after adjustment for baseline factors -- this is the essence of the missing at random assumption.

```
. logistic missing sbp1 sbp0 bmi0 baseline_dm

Logistic regression                               Number of obs   =       2607
                                                  LR chi2(4)      =    1924.14
                                                  Prob > chi2     =     0.0000
Log likelihood = -839.98794                       Pseudo R2       =     0.5339

------------------------------------------------------------------------------
     missing | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        sbp1 |   1.000695   .0040285     0.17   0.863     .9928306    1.008622
        sbp0 |   1.207511   .0096962    23.48   0.000     1.188656    1.226666
        bmi0 |   1.000975   .0116884     0.08   0.933     .9783266    1.024148
 baseline_dm |   .8684747   .1268182    -0.97   0.334     .6523209    1.156254
       _cons |   7.54e-12   8.22e-12   -23.49   0.000     8.90e-13    6.38e-11
------------------------------------------------------------------------------
```

Question 2: Classify the following scenarios as leading to missingness which is MCAR, MAR or NMAR (and explain your answer).

(a) missingness only depends on baseline SBP

This would be missing at random (MAR) -- missingness only depends on our data though a variable which is available on everyone.

(b) missingness only depends on baseline diabetes

This would be missing at random (MAR) -- missingness only depends on our data though a variable which is available on everyone.

(c) missingness only depends on BMI at year 1

This would be missing at random (MAR) -- missingness only depends on our data though a variable which is available on everyone.

(d) missingness only depends on SBP at year 1

This would be missing at not a random (MNAR) -- missingness depends on our data though the variable which is missing.

Question 3: Fill in the following table. Recall, you have the complete data and can take out the missing values by analyzing the subset with non-missing value (if missing==0)

	Mean SBP
Baseline	134.9
Year One (All Data)	135.2
Year One (Non-Missing Data)	127.5

```
. mean sbp0
Mean estimation                       Number of obs    =    2607

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
        sbp0 |   134.8899   .3716866      134.1611    135.6187
--------------------------------------------------------------

 mean sbp1
Mean estimation                       Number of obs    =    2607

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
        sbp1 |   135.2056   .3761451       134.468    135.9432
--------------------------------------------------------------


. mean sbp1 if missing==0
Mean estimation                       Number of obs    =    1384

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
        sbp1 |   127.5499   .4403621       126.686    128.4137
--------------------------------------------------------------
```

What is the mean change in SBP from baseline to year 1 with the complete data?

+0.3 *mmHg*
What is it using only the non-missing values at year 1?

-*7.3mmHg*

How do you explain the difference?

It appears that the available 1 year SBP is in people with lower baseline SBP (and the people with lower SBP at year 1 as a result of that). This means the apparent lower SBP at year 1 is not from within person changes but a shift in the population

4 . GEE

Question 4: *The model for the complete data is given as*

```
xtgee sbp i.visit , i(pptid) corr(indep) robust

                                (Std. Err. adjusted for clustering on pptid)
------------------------------------------------------------------------------
             |               Robust
         sbp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.visit |   .3156885   .3404026     0.93   0.354    -.3514883    .9828654
       _cons |   134.8899   .3716866   362.91   0.000     134.1614    135.6184
------------------------------------------------------------------------------
```

The model for the non-missing data is

```
GEE population-averaged model                   Number of obs      =      3991
Group variable:                      pptid      Number of groups   =      2607
Link:                             identity      Obs per group: min =         1
Family:                           Gaussian                     avg =       1.5
Correlation:                   independent                     max =         2
                                                Wald chi2(1)       =    202.53
Scale parameter:                  328.1759      Prob > chi2        =    0.0000

Pearson chi2(3991):                1309750      Deviance           =   1309750
Dispersion (Pearson):             328.1759      Dispersion         =  328.1759

                                (Std. Err. adjusted for clustering on pptid)
------------------------------------------------------------------------------
             |               Robust
         sbp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.visit |  -7.340056   .5157637   -14.23   0.000    -8.350935   -6.329178
       _cons |   134.8899   .3716866   362.91   0.000     134.1614    135.6184
------------------------------------------------------------------------------
```

How does that compare with the average change (estimated by GEE) in SBP from baseline to year 1 using the data with non-missing values? Why are they different?

Note, that there is no effect of visit (change from baseline to year 1) in the complete data but a statistically significant one for the incomplete data. In the data with some values missing at year 1, the visit term is -7.3 points or exactly what was derived in question 3. The difference reflects the purely the dropouts with low baseline SBP.

Question 5: Fit first a model to the complete data

```
xtmixed sbp visit|| pptid:, ml

                                                Wald chi2(1)       =      0.86
Log likelihood = -22227.365                     Prob > chi2        =    0.3536

------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       visit |   .3156885   .3403373     0.93   0.354    -.3513603    .9827373
       _cons |   134.8899   .3738508   360.81   0.000     134.1572    135.6226
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
pptid: Identity              |
                   sd(_cons) |    14.6076   .2830671      14.06321    15.17308
-----------------------------+------------------------------------------------
                sd(Residual) |   12.28754   .1701686       11.9585    12.62563
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =  1094.96 Prob >= chibar2 = 0.0000
```

Then fit a model to the non-missing data

```
                                                Wald chi2(1)       =      2.17
Log likelihood = -17052.336                     Prob > chi2        =    0.1409

------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       visit |   .6166597   .4188072     1.47   0.141    -.2041873    1.437507
       _cons |   134.8899   .3709222   363.66   0.000     134.1629    135.6169
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
pptid: Identity              |
                   sd(_cons) |   15.02488   .3668638      14.32277     15.7614
-----------------------------+------------------------------------------------
                sd(Residual) |   11.52964   .2497239      11.05043    12.02963
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =   343.35 Prob >= chibar2 = 0.0000
```

Question 5: How to the mixed model results compare between missing and non-missing data?

For the non-missing dataset, We see that there is an estimated change of 0.32 mmHg in the complete data is identical to the one in the GEE model.

How does that compare with the average change (estimated by the mixed model) in SBP from baseline to year 1 using the data with non-missing values?

We see for the mixed model that there is an estimated change of 0.62 mmHg with missing data -- the effect is non-significant and we should not be surprised. The reason is related to the fact that the data missing at year 1 is MAR with respect to the baseline SBP. The mixed model is fit by maximum likelihood. Maximum likelihood seamlessly handles MAR missing data as long as the variables which define the MAR assumption are included in the model. They need not be include as predictors -- they can be included as an outcome. In this example, the variable which defines the MAR assumption is baseline SBP which ensure s that maximum likelihood will handle the missing data. Note, that baseline SBP is an outcome is this model -- not a predictor.