```
#### Biost 518: Applied Biostatistics II
#### Emerson, Winter 2006

#### Homework #6 Key
#### Annotated Stata Log File
#### March 12, 2006

#### The following output was used to generate the numbers that I wanted to present
#### in tables, as well as the plots I wanted to present as figures in the paper.
#### I note that Stata does not present its output in a form suitable for presentation.
#### Numbers need to be rounded to an interesting number of significant digits, and
#### the columns and rows need to facilitate comparison of relevant measures.
#### I used Excel to manipulate this output into the form I wanted, then copied the
#### resulting tables to the MS-Word document.

#### Comments edited into the log file produced by Stata are
#### on the lines that start with the four '#' signs and are
#### printed in italics.

#### The Stata commands are put in bold face.

#### Stata output is displayed in regular typeface in blue.

#### Read in data. I use a dataset previously created in "long" format.
. use SEPlong

#### Creating variables to model interactions
. g ha= height * age
. g hm= height * sex
. g am= age * sex
. g ham= height * am

####################
#### Problem 1a
#### Regressing p60 for the right leg on height, age, and the height age
#### interaction for females
####################
. regress p60r height age ha if sex==0
      Source |       SS       df       MS              Number of obs =     137
-------------+------------------------------           F(  3,   133) =   24.92
       Model |  1100.94149     3  366.980495           Prob > F      =  0.0000
    Residual |  1958.87791   133  14.7284053           R-squared     =  0.3598
-------------+------------------------------           Adj R-squared =  0.3454
       Total |   3059.8194   136   22.498672           Root MSE      =  3.8378

------------------------------------------------------------------------------
        p60r |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   1.415643   .3852599     3.67   0.000     .6536138    2.177672
         age |   1.182864   .4514123     2.62   0.010     .2899883     2.07574
          ha |  -.0157824   .0070526    -2.24   0.027    -.0297321   -.0018327
       _cons |  -39.00772   24.95136    -1.56   0.120    -88.36054     10.3451
------------------------------------------------------------------------------

#### Saving residuals to investigate the impact of modeling averages below
. predict rsdr, resid
(1750 missing values generated)

####################
#### Problem 1b
```

```
#### Regressing p60 for the left leg on height, age, and the height age
#### interaction for females
####################
. regress p60l height age ha if sex==0

      Source |       SS       df       MS              Number of obs =     137
-------------+------------------------------           F(  3,   133) =   23.14
       Model |  1022.78875     3  340.929582           Prob > F      =  0.0000
    Residual |  1959.20482   133  14.7308633           R-squared     =  0.3430
-------------+------------------------------           Adj R-squared =  0.3282
       Total |  2981.99357   136  21.9264233           Root MSE      =  3.8381

------------------------------------------------------------------------------
        p60l |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   1.344906    .385292     3.49   0.001     .5828136    2.106999
         age |   1.075982     .45145     2.38   0.019     .1830315    1.968933
          ha |  -.0142147   .0070532    -2.02   0.046    -.0281656   -.0002639
       _cons |  -33.87798   24.95344    -1.36   0.177    -83.23492    15.47896
------------------------------------------------------------------------------

#### Saving residuals to investigate the impact of modeling averages below
. predict rsdl, resid
(1750 missing values generated)

####################
#### Problem 1c
#### Regressing p60 for the average of left and right leg on height, age, and the
#### height age interaction for females
####################
. g p60= (p60r + p60l) /2
(1750 missing values generated)

. regress p60 height age ha if sex==0

      Source |       SS       df       MS              Number of obs =     137
-------------+------------------------------           F(  3,   133) =   27.28
       Model |  1061.10064     3  353.700213           Prob > F      =  0.0000
    Residual |  1724.28955   133  12.9645831           R-squared     =  0.3810
-------------+------------------------------           Adj R-squared =  0.3670
       Total |  2785.39019   136  20.4808102           Root MSE      =  3.6006

------------------------------------------------------------------------------
         p60 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   1.380275   .3614558     3.82   0.000     .6653291     2.09522
         age |   1.129423   .4235208     2.67   0.009     .2917155    1.967131
          ha |  -.0149985   .0066168    -2.27   0.025    -.0280863   -.0019108
       _cons |  -36.44286   23.40969    -1.56   0.122    -82.74631    9.860598
------------------------------------------------------------------------------

#### Note that the root MSE is less for the averages than it is for left or right
#### by themselves. We know that
```

$$p60 = \frac{(p60R + p60L)}{2} \quad \Rightarrow \quad Var(p60) = \frac{Var(p60R) + Var(p60L) + 2\rho\sqrt{Var(p60R)Var(p60L)}}{4}$$

```
#### So we should be able to predict the root MSE for the analysis based on the average.
#### Now the value of the correlation between the p60R and p60L measurements should be
```

```
#### the correlation of those measurements ADJUSTED for height, age, and the
#### interaction. That is, we are interested in the correlation of the residuals. We
#### have modeled any similarity between measurements on the same subject due to the
#### height and age. The correlation of the residuals should be less extreme than
#### the correlation between the raw measurements:

#### Correlation between unadjusted measurements
. corr p60r p60l if sex==0
(obs=137)

             |     p60r     p60l
-------------+------------------
        p60r |   1.0000
        p60l |   0.8441   1.0000


#### Correlation between residuals (so adjusted for height, age, and interaction)
. corr rsdr rsdl if sex==0
(obs=137)

             |     rsdr     rsdl
-------------+------------------
        rsdr |   1.0000
        rsdl |   0.7603   1.0000


#### Using root MSE from the models on right and left and correlation of residuals
#### to estimate the root MSE for the model of the averaged right and left
#### measurements. Note the near perfect agreement.
. disp sqrt(3.8378^2 + 3.8381^2 + 2 * 3.8378 * 3.8381 * 0.7603) / 2
3.6006231

####################
#### Problem 2a
#### Regressing p60 for the right leg on height, age, and the height age
#### interaction for males
#### (I used Excel to format the tables and compute the differences in
#### parameter estimates for females and males.)
####################
. regress p60r height age ha if sex==1

      Source |       SS       df       MS              Number of obs =     113
-------------+------------------------------           F(  3,   109) =   16.51
       Model |   758.05846     3  252.686153           Prob > F      =  0.0000
    Residual |  1667.78719   109     15.3008           R-squared     =  0.3125
-------------+------------------------------           Adj R-squared =  0.2936
       Total |  2425.84565   112  21.6593362           Root MSE      =  3.9116

------------------------------------------------------------------------------
        p60r |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   .2384118   .3446893     0.69   0.491    -.4447511    .9215748
         age |  -.1026585    .430216    -0.24   0.812    -.9553327    .7500157
          ha |   .0035149   .0062066     0.57   0.572    -.0087864    .0158162
       _cons |   39.41305   24.02054     1.64   0.104    -8.194868    87.02098
------------------------------------------------------------------------------

####################
#### Problem 3a
```

```
#### Regressing p60 for the right leg on height, age, sex, the height-age,
#### height-sex, and age-sex two-way interactions, and the three-way interaction for
#### both males and females.
####################
. regress p60r height age sex ha hm am ham

      Source |       SS       df       MS              Number of obs =     250
-------------+------------------------------           F(  7,   242) =   20.21
       Model |  2119.60382     7  302.800545           Prob > F      =  0.0000
    Residual |  3626.66511   242  14.9862194           R-squared     =  0.3689
-------------+------------------------------           Adj R-squared =  0.3506
       Total |  5746.26892   249  23.0773852           Root MSE      =  3.8712

------------------------------------------------------------------------------
        p60r |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   1.415643   .3886172     3.64   0.000     .6501389    2.181147
         age |   1.182864    .455346     2.60   0.010     .2859168    2.079812
         sex |   78.42078   34.62068     2.27   0.024     10.22443    146.6171
          ha |  -.0157824    .007114    -2.22   0.027    -.0297957    -.001769
          hm |  -1.177231   .5170989    -2.28   0.024     -2.19582   -.1586418
          am |  -1.285523   .6233944    -2.06   0.040    -2.513494   -.0575512
         ham |   .0192972   .0093989     2.05   0.041     .0007831    .0378113
       _cons |  -39.00772   25.16879    -1.55   0.122    -88.58559    10.57015
------------------------------------------------------------------------------

####################
#### Problem 3d
#### Multiple partial F test that no interactions (either two-way or three-way) exist.
####################
. test ha hm am ham

 ( 1)  ha = 0
 ( 2)  hm = 0
 ( 3)  am = 0
 ( 4)  ham = 0

       F(  4,   242) =    1.76
            Prob > F =    0.1385

####################
#### Just for fun:
#### We found a statistically significant three-way interaction when considered alone.
#### But we did not find a test of all interactions statistically significant.
#### This seeming contradiction can happen (obviously: It did.). If, in truth, only
#### one of the interactions were really important, throwing extraneous predictors
#### into the hypothesis test will lessen power. Alternatively, examining the
#### interaction terms singly is subject to a multiple comparison issue. (It is
#### interesting (but relatively irrelevant due to their hierarchical nature) to
#### note that each interaction term was significant by itself, but jointly they
#### were not. Again: This can happen.)
####
#### Below I explore statistical significance of all interactions when using the
#### average of right and left measurements (it doesn't really matter—hardly
#### surprising given the very slight improvement in root MSE with the average),
#### as well as the possibility that the influential observation identified in
#### class (ptid==140) could have markedly changed our inference.
####################

#### Regression on males, females using average of right and left
```

```
. regress p60 height age sex ha hm am ham

      Source |       SS       df       MS              Number of obs =     250
-------------+------------------------------           F(  7,   242) =   22.29
       Model |  2035.84797     7  290.835424           Prob > F      =  0.0000
    Residual |  3158.14276   242  13.0501767           R-squared     =  0.3920
-------------+------------------------------           Adj R-squared =  0.3744
       Total |  5193.99073   249  20.8594005           Root MSE      =  3.6125

------------------------------------------------------------------------------
         p60 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   1.380275    .362647     3.81   0.000     .6659271    2.094622
         age |   1.129423   .4249166     2.66   0.008     .2924161     1.96643
         sex |   74.95773   32.30708     2.32   0.021     11.31875    138.5967
          ha |  -.0149985   .0066386    -2.26   0.025    -.0280754   -.0019217
          hm |  -1.127006   .4825427    -2.34   0.020    -2.077526   -.1764858
          am |  -1.162866   .5817348    -2.00   0.047    -2.308776   -.0169558
         ham |   .0175005   .0087708     2.00   0.047     .0002236    .0347773
       _cons |  -36.44286   23.48684    -1.55   0.122    -82.70758     9.82187
------------------------------------------------------------------------------

#### Multiple partial F test that no interactions (either two-way or three-way) exist.
. test ha hm am ham

 ( 1)  ha = 0
 ( 2)  hm = 0
 ( 3)  am = 0
 ( 4)  ham = 0

       F(  4,   242) =    1.95
            Prob > F =    0.1034

#### Now omitting case 140
. regress p60 height age sex ha hm am ham if ptid!=140

      Source |       SS       df       MS              Number of obs =     249
-------------+------------------------------           F(  7,   241) =   25.23
       Model |  2187.21418     7  312.459168           Prob > F      =  0.0000
    Residual |  2984.46553   241  12.3836744           R-squared     =  0.4229
-------------+------------------------------           Adj R-squared =  0.4062
       Total |   5171.6797   248  20.8535472           Root MSE      =   3.519

------------------------------------------------------------------------------
         p60 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |    1.68736   .3626571     4.65   0.000     .9729774    2.401742
         age |   1.385432   .4195307     3.30   0.001     .5590165    2.211847
         sex |   95.34379    31.9386     2.99   0.003     32.42934    158.2582
          ha |  -.0188283   .0065472    -2.88   0.004    -.0317254   -.0059312
          hm |  -1.434091   .4771576    -3.01   0.003    -2.374023    -.494159
          am |  -1.418874   .5707932    -2.49   0.014    -2.543255   -.2944936
         ham |   .0213302   .0086049     2.48   0.014     .0043798    .0382806
       _cons |  -56.82891   23.51789    -2.42   0.016    -103.1558   -10.50205
------------------------------------------------------------------------------

#### Multiple partial F test for some interaction is now significant
```

```
. test ha hm am ham

 ( 1)  ha = 0
 ( 2)  hm = 0
 ( 3)  am = 0
 ( 4)  ham = 0

       F(  4,   241) =    3.36
            Prob > F =    0.0106

#### Multiple partial F test for effect of height in this model involving interactions
. test height ha hm ham

 ( 1)  height = 0
 ( 2)  ha = 0
 ( 3)  hm = 0
 ( 4)  ham = 0

       F(  4,   241) =   14.97
            Prob > F =    0.0000

#### Multiple partial F test for effect of age in this model involving interactions
. test age ha am ham

 ( 1)  age = 0
 ( 2)  ha = 0
 ( 3)  am = 0
 ( 4)  ham = 0

       F(  4,   241) =   35.91
            Prob > F =    0.0000

#### Multiple partial F test for effect of sex in this model involving interactions
. test sex hm am ham

 ( 1)  sex = 0
 ( 2)  hm = 0
 ( 3)  am = 0
 ( 4)  ham = 0

       F(  4,   241) =    2.69
            Prob > F =    0.0318

#### Note that the fitted values shown in class suggested a clear height-age interaction
#### among females, but the fitted lines looked pretty parallel for males. One could
#### of course consider whether future studies might not fit a height age interaction
#### for males, while still fitting one for females. This more parsimonious model could
#### be parameterized by reversing the coding for males and females:
. g female = 1 - sex
. g hf= height * female
. g af= age * female
. g haf= ha * female

#### Note that in this model, the height age interaction is interpretable as the
#### height age interaction for males. It is not statistically significant. If we
#### were brave, we could go into the future studies fitting a model without that
#### height age two-way interaction. As a rule, however, it is better to keep in
#### all the lower order terms, even if they are not significant. There is a slight
```

```
#### loss of power (see the key to homework #5 for a situation where this might
#### be evident). But generally the interpretation of the individual parameters
#### is so difficult (and quite often way outside the range of our data) that
#### assuming we are sure that a choice of zero is appropriate is fraught with
#### peril. Thus, I would still keep all two-way interactions in the model.
. regress p60 height age female ha hf af haf

      Source |       SS       df       MS              Number of obs =     250
-------------+------------------------------           F(  7,   242) =   22.29
       Model |  2035.84797     7  290.835424           Prob > F      =  0.0000
    Residual |  3158.14276   242  13.0501767           R-squared     =  0.3920
-------------+------------------------------           Adj R-squared =  0.3744
       Total |  5193.99073   249  20.8594005           Root MSE      =  3.6125

------------------------------------------------------------------------------
         p60 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   .2532689    .318331     0.80   0.427    -.3737843    .8803221
         age |  -.0334425   .3973175    -0.08   0.933    -.8160846    .7491996
      female |  -74.95773   32.30708    -2.32   0.021    -138.5967   -11.31875
          ha |   .0025019    .005732     0.44   0.663    -.0087891    .0137929
          hf |   1.127006   .4825427     2.34   0.020     .1764858    2.077526
          af |   1.162866   .5817348     2.00   0.047     .0169558    2.308776
         haf |  -.0175005   .0087708    -2.00   0.047    -.0347773   -.0002236
       _cons |   38.51487   22.18369     1.74   0.084      -5.1829    82.21264
------------------------------------------------------------------------------


####################
#### Problem 4a
#### Regressing p60 for the both legs on height, age, sex, the height-age,
#### height-sex, and age-sex two-way interactions, and the three-way interaction for
#### both males and females.
#### Note that I use the cluster() option to properly adjust for the correlated
#### response within individuals.
####################
. regress sep height age sex ha hm am ham if peak=="p60", cluster(ptid)

Linear regression                                      Number of obs =     500
                                                       F(  7,   249) =   19.04
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.3606
Number of clusters (ptid) = 250                        Root MSE      =   3.831

------------------------------------------------------------------------------
             |               Robust
         sep |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   1.380275   .4278169     3.23   0.001     .5376735    2.222876
         age |   1.129423   .4427753     2.55   0.011      .257361    2.001485
         sex |   74.95771   34.60971     2.17   0.031      6.79261    143.1228
          ha |  -.0149985   .0067958    -2.21   0.028    -.0283832   -.0016139
          hm |  -1.127006   .5170952    -2.18   0.030    -2.145444   -.1085676
          am |  -1.162865   .5682818    -2.05   0.042    -2.282117   -.0436135
         ham |   .0175004   .0085333     2.05   0.041     .0006937    .0343072
       _cons |  -36.44285   28.10098    -1.30   0.196    -91.78877    18.90307
------------------------------------------------------------------------------
```

```
####################
#### Problem 4b
#### Predicting the mean p60 measurement in height, age, sex groups, along with
#### 95% CI for those group means.
####################
. predict fit
(option xb assumed; fitted values)

. predict sefit, stdp
. g cifitlo= fit - invttail(249,.025) * sefit
. g cifithi= fit + invttail(249,.025) * sefit

####################
#### Problem 4c
#### Predicting the "normal range" of p60 measurements in height, age, sex groups
#### using 95% prediction intervals (forecast intervals in Stata terminology).
#### Because we used robust SE, Stata assumes there might be heteroscedasticity
#### and thus refuses to compute these intervals. However, we used the robust SE
#### only to account for the correlation, so we will proceed with the methods
#### described in class based on combining the SE of the fitted values and
#### the root MSE.
####################
. g sefore= sqrt(sefit^2 + 3.831^2)
. g ciforelo= fit - invttail(249,.025) * sefore
. g ciforehi= fit + invttail(249,.025) * sefore

. list p60r p60l p60 fit cifitlo cifithi ciforelo ciforehi if age==60 &
height==69 & sex==1

      +-------------------------------------------------------------------------+
      | p60r p60l  p60        fit    cifitlo    cifithi   ciforelo   ciforehi |
      |-------------------------------------------------------------------------|
 113. | 59.4   57 58.2   64.34175   63.48837   65.19513   56.74835   71.93515 |
 363. |    .    .    .   64.34175   63.48837   65.19513   56.74835   71.93515 |
 613. |    .    .    .   64.34175   63.48837   65.19513   56.74835   71.93515 |
 863. |    .    .    .   64.34175   63.48837   65.19513   56.74835   71.93515 |
1113. |    .    .    .   64.34175   63.48837   65.19513   56.74835   71.93515 |
      |-------------------------------------------------------------------------|
1363. |    .    .    .   64.34175   63.48837   65.19513   56.74835   71.93515 |
1613. |    .    .    .   64.34175   63.48837   65.19513   56.74835   71.93515 |
1863. |    .    .    .   64.34175   63.48837   65.19513   56.74835   71.93515 |
      +-------------------------------------------------------------------------+


####################
#### Just for fun:
#### Examining the coverage probabilities for the prediction intervals in this data set.
#### Of course, the PI were computed to agree with these data, so this does not really
#### tell us anything: As I pointed out in class, the true coverage probability can
#### be substantially less than the desired 95%, but will on average (across repeated
#### experiments) be 95% providing 1) our model is correct (straight line relationship
#### in main effects and interactions), and 2) the data is normally distributed within
#### height-age-sex groups.
####################
. g cvrgR= 0
. replace cvrgR=1 if p60r > ciforelo & p60r < ciforehi
(236 real changes made)

. g cvrgL= 0
```

```
. replace cvrgL=1 if p60l > ciforelo & p60l < ciforehi
(236 real changes made)

. summ cvrgR cvrgL in 1/250

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       cvrgR |       250        .944     .230383          0          1
       cvrgL |       250        .944     .230383          0          1


. table cvrgL cvrgR in 1/250

----------------------
          |   cvrgR
    cvrgL |    0     1
----------+-----------
        0 |    5     9
        1 |    9   227
----------------------
```