Notes for laboratory session 3

1.Single-factor analysis of variance

Consider the effect of gender on levels of retinol in plasma. The one-way ANOVA is given by the following output:

```
. anova  retplasm sex

                           Number of obs =     314     R-squared     =  0.0392
                           Root MSE      = 204.801     Adj R-squared =  0.0361

                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  533837.408     1  533837.408      12.73     0.0004
                         |
                     sex |  533837.408     1  533837.408      12.73     0.0004
                         |
                Residual |  13086344.5   312  41943.4117
              -----------+----------------------------------------------------
                   Total |  13620181.9   313   43514.958
```

a) How can we test if gender has a statistically significant impact on plasma retinol levels? How is the appropriate statistic calculated?

Now do the same using the `regress` command of STATA

```
. reg

      Source |       SS           df       MS      Number of obs   =       314
-------------+----------------------------------   F(1, 312)       =     12.73
       Model |  533837.408         1  533837.408   Prob > F        =    0.0004
    Residual |  13086344.5       312  41943.4117   R-squared       =    0.0392
-------------+----------------------------------   Adj R-squared   =    0.0361
       Total |  13620181.9       313   43514.958   Root MSE        =     204.8

------------------------------------------------------------------------------
    retplasm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
     Female  |  -122.3759   34.30232    -3.57   0.000    -189.8691   -54.88283
       _cons |   710.0976   31.98453    22.20   0.000     647.1649    773.0302
------------------------------------------------------------------------------
```

b) How can we check now if there is a statistical significant gender effect on plasma retinol levels? What is the relation between the statistics used in the `anova` and `regress` commands?
c) How can we calculate the best estimates for mean retinol level for women and men?

Another way of doing this is by using the `xi` STATA command as follows:

```
. xi: reg retplasm i.sex
i.sex                Isex_1-2     (naturally coded; Isex_1 omitted)

  Source |       SS       df       MS              Number of obs =     314
---------+------------------------------           F(  1,   312) =   12.73
   Model |  533837.408     1  533837.408           Prob > F      =  0.0004
Residual |  13086344.5   312  41943.4117           R-squared     =  0.0392
---------+------------------------------           Adj R-squared =  0.0361
   Total |  13620181.9   313   43514.958           Root MSE      =  204.80

------------------------------------------------------------------------------
retplasm |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  Isex_2 |  -122.3759   34.30232     -3.568   0.000      -189.8691   -54.88283
   _cons |   710.0976   31.98453     22.201   0.000       647.1649    773.0302
------------------------------------------------------------------------------
```

Notice that the xi command creates the dummy variables defining the lowest numerical value of the categorical variable as the default reference level. However we can change the reference level as shown below:

```
. char sex[omit] 2

. xi: reg retplasm i.sex
i.sex                Isex_1-2     (naturally coded; Isex_2 omitted)

  Source |       SS       df       MS              Number of obs =     314
---------+------------------------------           F(  1,   312) =   12.73
   Model |  533837.408     1  533837.408           Prob > F      =  0.0004
Residual |  13086344.5   312  41943.4117           R-squared     =  0.0392
---------+------------------------------           Adj R-squared =  0.0361
   Total |  13620181.9   313   43514.958           Root MSE      =  204.80

------------------------------------------------------------------------------
retplasm |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  Isex_1 |   122.3759   34.30232      3.568   0.000       54.88283    189.8691
   _cons |   587.7216   12.39511     47.416   0.000        563.333    612.1102
------------------------------------------------------------------------------
```

d) Calculate the best estimates for mean retinol level for women and men. Check the consistency of the results. (You can check the ANOVA model too, by using the following command: `oneway retplasm sex,tabulate`)

Now using the `glm` command:

```
. char sex[omit] 2

. xi: glm retplasm i.sex
i.sex             _Isex_1-2           (naturally coded; _Isex_2 omitted)

Iteration 0:   log likelihood = -2115.6635

Generalized linear models                          No. of obs      =       314
Optimization     : ML: Newton-Raphson              Residual df     =       312
                                                   Scale parameter =  41943.41
Deviance         =  13086344.45                    (1/df) Deviance =  41943.41
Pearson          =  13086344.45                    (1/df) Pearson  =  41943.41

Variance function: V(u) = 1                        [Gaussian]
Link function    : g(u) = u                        [Identity]
Standard errors  : OIM

Log likelihood   = -2115.663535                    AIC             =   13.4883
BIC              =  13084550.64

------------------------------------------------------------------------------
    retplasm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Isex_1 |   122.3759   34.30232     3.57   0.000     55.14464    189.6073
       _cons |   587.7216   12.39511    47.42   0.000     563.4276    612.0156
------------------------------------------------------------------------------
```

e) Try to notice the similarities between the two approaches.

2.Regression models for general two-way ANOVA

Asses the effect of sex and vitamin use on plasma retinol levels using the glm command with females and no-vitamine-use categories as reference categories.

```
. char sex[omit] 2

. char vituse[omit] 3

. xi: glm retplasm i.sex i.vituse i.sex*i.vituse
i.sex             _Isex_1-2           (naturally coded; _Isex_2 omitted)
i.vituse          _Ivituse_1-3        (naturally coded; _Ivituse_3 omitted)
i.sex*i.vituse    _IsexXvit_#_#       (coded as above)
note: _Isex_1 dropped due to collinearity
note: _Ivituse_1 dropped due to collinearity
note: _Ivituse_2 dropped due to collinearity

Iteration 0:   log likelihood = -2111.9911

Generalized linear models                          No. of obs      =       314
Optimization     : ML: Newton-Raphson              Residual df     =       308
                                                   Scale parameter =  41505.82
Deviance         =  12783793.58                    (1/df) Deviance =  41505.82
Pearson          =  12783793.58                    (1/df) Pearson  =  41505.82

Variance function: V(u) = 1                        [Gaussian]
Link function    : g(u) = u                        [Identity]
Standard errors  : OIM

Log likelihood   = -2111.991142                    AIC             =  13.49039
BIC              =  12782022.77

------------------------------------------------------------------------------
    retplasm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Isex_1 |   166.3468   47.76693     3.48   0.000     72.72537    259.9683
  _Ivituse_1 |   33.46968   29.28935     1.14   0.253    -23.93638    90.87575
  _Ivituse_2 |   39.49589   31.87656     1.24   0.215    -22.98102    101.9728
_IsexXvit_~1 |  -11.72721   76.51943    -0.15   0.878    -161.7025    138.2481
_IsexXvit_~2 |  -255.6611   105.4603    -2.42   0.015    -462.3596   -48.96267
       _cons |   563.2184   21.84213    25.79   0.000     520.4086    606.0282
------------------------------------------------------------------------------
```

a) Calculate the estimates for mean plasma retinol levels for each one of the six categories, which can be created by the combination of gender and vitamin use categories.

The dscriptive statistics of the plasma retinol levels by gender and vitamin use are given in the STATA output below:

```
. tabulate sex vituse, summarize(retplasm)

   Means, Standard Deviations and Frequencies of Plasma retinol (ng/ml)

           |         Vitamine use
       Sex |         1          2          3 |     Total
-----------+---------------------------------+----------
         1 | 751.30769      513.4  729.56522 | 710.09756
           | 329.43269  298.59303   290.0285 | 305.52208
           |        13          5         23 |        41
-----------+---------------------------------+----------
         2 | 596.68807  602.71429  563.21839 | 587.72161
           | 203.71816   184.6959  159.92785 | 185.43069
           |       109         77         87 |       273
-----------+---------------------------------+----------
     Total | 613.16393  597.26829        598 | 603.70064
           | 223.83038  192.02109  204.39088 | 208.60239
           |       122         82        110 |       314
```

b) Compare the results listed above with those calculated in the previous question.

3.Regression models for the analysis of covariance

The analysis of covariance can be expressed in terms of a linear regression. We can assess the effect of gender and age on plasma retinol levels using the following command in STATA (the model includes the gender-age interaction):

```
. xi: glm retplasm i.sex*age
i.sex             _Isex_1-2           (naturally coded; _Isex_2 omitted)
i.sex*age         _IsexXage_#         (coded as above)

Iteration 0:   log likelihood = -2110.4432

Generalized linear models                          No. of obs      =       314
Optimization     : ML: Newton-Raphson              Residual df     =       310
                                                   Scale parameter =  40833.46
Deviance         =  12658374.05                    (1/df) Deviance =  40833.46
Pearson          =  12658374.05                    (1/df) Pearson  =  40833.46

Variance function: V(u) = 1                        [Gaussian]
Link function    : g(u) = u                        [Identity]
Standard errors  : OIM

Log likelihood   = -2110.443238                    AIC             =  13.46779
BIC              =  12656591.74

------------------------------------------------------------------------------
    retplasm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   2.810887   .8693928     3.23   0.001     1.106909    4.514866
      _Isex_1 |   235.3007   151.7706     1.55   0.121    -62.16429    532.7657
  _IsexXage_1 |  -2.421536   2.502083    -0.97   0.333    -7.325528    2.482455
       _cons |   451.2649   43.94161    10.27   0.000     365.1409    537.3888
------------------------------------------------------------------------------
```

From the STATA output above we have that there is no significant interaction between gender and age (Why?) .

a) Check the parallelism by creating an appropriate graph.

Thus we proceed with a more parsimonious model excluding the interaction term.

```
. xi: glm retplasm i.sex age
i.sex             _Isex_1-2           (naturally coded; _Isex_2 omitted)

Iteration 0:   log likelihood = -2110.9169

Generalized linear models                          No. of obs      =       314
Optimization     : ML: Newton-Raphson              Residual df     =       311
                                                   Scale parameter =  40825.15
Deviance         =  12696620.84                    (1/df) Deviance =  40825.15
Pearson          =  12696620.84                    (1/df) Pearson  =  40825.15

Variance function: V(u) = 1                        [Gaussian]
Link function    : g(u) = u                        [Identity]
Standard errors  : OIM

Log likelihood   = -2110.916892                    AIC             =  13.46444
BIC              =  12694832.78

------------------------------------------------------------------------------
    retplasm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Isex_1 |   92.42252   35.20318     2.63   0.009     23.42555    161.4195
         age |   2.518526   .8151396     3.09   0.002      .920882    4.116171
       _cons |   465.4578   41.41804    11.24   0.000     384.2799    546.6356
------------------------------------------------------------------------------
```

Which leads to a significant gender effect (p value 0.009) at the 5% level.