Regression models for one-way analysis of variance

Analyses of variance and covariance can be analyzed in linear regression terms. For example, consider the one-way analysis of variance model

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \begin{cases} i = 1, \ldots, a \\ j = 1, \ldots, n \end{cases}$$

can be recast as a simple linear regression model by defining $X_i = \begin{cases} 1, \text{ if group } i \\ 0, \text{ otherwise} \end{cases} \quad i = 1, \ldots, a-1.$

Thus expressed the one-way ANOVA model becomes $Y_{ij} = \beta_o + \beta_1 X_1 + \cdots + \beta_{a-1} X_{a-1} + \varepsilon_{ij}$ that is,

$$\begin{gathered} \text{Group } 1: Y_{1j} = \beta_o + \beta_1 + \varepsilon_{ij} \\ \text{Group } 2: Y_{2j} = \beta_o + \beta_2 + \varepsilon_{ij} \\ \vdots \\ \text{Group } a-1: Y_{(a-1)j} = \beta_o + \beta_{(a-1)} + \varepsilon_{ij} \\ \text{Group } a: Y_a = \beta_o + \varepsilon_{ij} \end{gathered}$$

Regression models for one-way ANOVA

This regression model is equivalent to the ANOVA model. To see this consider that

$$\mu_i = \begin{cases} \beta_o + \beta_i, \text{ if } i = 1,\ldots,a-1 \\ \beta_o, \text{ if } i = a \end{cases}$$

The usual null hypothesis in regression H_o: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ (with $k=a$-1), means that

$$\beta_1 = \mu_1 - \mu = 0 \Rightarrow \mu_1 = \mu = \mu_a$$
$$\vdots$$
$$\beta_k = \mu_k - \mu = 0 \Rightarrow \mu_k = \mu = \mu_a$$

is thus equivalent to the null hypothesis of the analysis of variance H_o: $=\mu_1 = \mu_2 = \ldots = \mu_a$.

The previous coding scheme is called *reference coding scheme*, since one level of the fixed (categorical) factor is the *reference* level, while the rest are defined as deviations from it. In the model above, we chose level *a* as the reference level but we could have easily chosen level 1 (or 2 or 3). The critical point is that coding a factor with *a* levels requires *a*-1 coding variables[1].

---

[1] This is in the case of a regression model with an intercept. If no intercept exists, then *a* coding variables must be used.

Example: Drug potency data

Consider the example with the potency (dosage at death) of four cardiac substances (Table 17-7 in the text). The usual ANOVA model is given by the following STATA output:

```
. oneway potency sub,tab

            |     Summary of mean Dosage at Death
  Substance |        Mean   Std. Dev.       Freq.
------------+------------------------------------
          1 |        25.9   3.0713732          10
          2 |        22.2   3.4896673          10
          3 |          20   2.9439203          10
          4 |        19.6   2.9514591          10
------------+------------------------------------
      Total |      21.925   3.9248551          40
Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups        249.875      3   83.2916667      8.55     0.0002
 Within groups         350.90     36   9.74722222
------------------------------------------------------------------------
    Total             600.775     39   15.4044872

Bartlett's test for equal variances:  chi2(3) =   0.3439  Prob>chi2 = 0.952
```

The estimates of the coefficients are given by the `regress` option as follows:

```
. anova potency drugid, reg

  Source |       SS       df       MS                  Number of obs =      40
---------+------------------------------               F(  3,    36) =    8.55
   Model |     249.875     3  83.2916667               Prob > F      =  0.0002
Residual |      350.90    36  9.74722222               R-squared     =  0.4159
---------+------------------------------               Adj R-squared =  0.3672
   Total |     600.775    39  15.4044872               Root MSE      =  3.1221

------------------------------------------------------------------------------
 potency        Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
------------------------------------------------------------------------------
_cons            19.6   .9872802     19.853   0.000        17.5977     21.6023
drugid
       1          6.3   1.396225      4.512   0.000       3.468324    9.131676
       2          2.6   1.396225      1.862   0.071      -.2316757    5.431676
       3           .4   1.396225      0.286   0.776      -2.431676    3.231676
       4    (dropped)
------------------------------------------------------------------------------
```

Where level 4 has been “dropped”, it is, in other words, the reference level

The equivalent regression model is given by the following STATA commands:

```
. char drugid[omit] 4

. xi: reg potency i.drugid
i.drugid              Idrugi_1-4     (naturally coded; Idrugi_4 omitted)

  Source |       SS       df       MS                  Number of obs =      40
---------+------------------------------               F(  3,    36) =    8.55
   Model |     249.875     3  83.2916667               Prob > F      =  0.0002
Residual |      350.90    36  9.74722222               R-squared     =  0.4159
---------+------------------------------               Adj R-squared =  0.3672
   Total |     600.775    39  15.4044872               Root MSE      =  3.1221

------------------------------------------------------------------------------
 potency |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
Idrugi_1 |        6.3   1.396225      4.512   0.000       3.468324    9.131676
Idrugi_2 |        2.6   1.396225      1.862   0.071      -.2316757    5.431676
Idrugi_3 |         .4   1.396225      0.286   0.776      -2.431676    3.231676
   _cons |       19.6   .9872802     19.853   0.000        17.5977     21.6023
------------------------------------------------------------------------------
```

Where the reference level is again drugid 4 (due to the `. char drugid[omit] 4` statement) and all other levels are deviations from that level.

Comments:

1. STATA always defines the last level as the reference level by default after the `anova` command.
2. The command `xi varname` by contrast defines the level with the *lowest* numerical value as the the default reference level. We can manipulate which level is the reference level by defining the `omit` variable with the command `char varname[omit] #` where "#" is the numerical value corresponding to the desired reference level. An alternative case is to define as the reference level the most frequent (prevalent) level. To do this we use the command `char _dta[omit] "prevalent"`. Finally, in case of string variables the command becomes `char _dta[omit] "string_literal"` where `string_literal` is the string level that we want to define as reference.
3. The `xi` command defines *a*-1 variables `Ivarname_i`, (*i*=1,…,*a*-1), such that `Ivarname_i=(varname==i)`.
4. The regression can then be carried out by these variables. To invoke them we use the umbrella term `i.varname`.

The previous regression output could have been produced by the following commands:

```
. gen Idrugi_1=(drugid==1)

. gen Idrugi_2=(drugid==2)

. gen Idrugi_3=(drugid==3)

. reg potency I*

  Source |       SS       df       MS                  Number of obs =      40
---------+------------------------------               F(  3,    36) =    8.55
   Model |     249.875     3  83.2916667               Prob > F      =  0.0002
Residual |      350.90    36  9.74722222               R-squared     =  0.4159
---------+------------------------------               Adj R-squared =  0.3672
   Total |     600.775    39  15.4044872               Root MSE      =  3.1221

------------------------------------------------------------------------------
 potency |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
Idrugi_1 |        6.3   1.396225      4.512   0.000       3.468324    9.131676
Idrugi_2 |        2.6   1.396225      1.862   0.071      -.2316757    5.431676
Idrugi_3 |         .4   1.396225      0.286   0.776      -2.431676    3.231676
   _cons |       19.6   .9872802     19.853   0.000        17.5977     21.6023
```

```
.list potency drugid I*

       potency     drugid   Idrugi_1   Idrugi_2   Idrugi_3
  1.        29          1          1          0          0
  2.        28          1          1          0          0
   .
   .
  9.        26          1          1          0          0
 10.        28          1          1          0          0
 11.        17          2          0          1          0
 12.        25          2          0          1          0
 13.        24          2          0          1          0
   .
   .
 27.        20          3          0          0          1
 28.        17          3          0          0          1
 29.        25          3          0          0          1
 30.        21          3          0          0          1
 31.        18          4          0          0          0
   .
   .
 37.        20          4          0          0          0
 38.        17          4          0          0          0
 39.        19          4          0          0          0
 40.        17          4          0          0          0
```

Regression models for general two-way ANOVA

In the two-way and general ANOVA the reference coding scheme is implemented as follows:

$$Y = \mu + \sum_{i=1}^{a-1} \alpha_i X_i + \sum_{j=1}^{b-1} \beta_j Z_j + \sum_{i=1}^{a-1} \sum_{j=1}^{b-1} \gamma_{ij} X_i Z_j \varepsilon_{ij}$$

where $X_i = \begin{cases} 1, \text{ if treatment } i \\ 0, \text{ otherwise} \end{cases}$ $i = 1, \ldots, a-1$ and $Z_j = \begin{cases} 1, \text{ if block } j \\ 0, \text{ otherwise} \end{cases}$ $j = 1, \ldots, b-1$, with a and b the number of treatments and blocks respectively.

The coefficients α, β and γ are linked to the means μ_{ij} by the formulas:

1. $\mu = \mu_{..}$

2. $\alpha_i = \mu_{i.} - \mu_{..},\ i = 1,2,\ldots,a-1$ and $-\sum_{i=1}^{a-1} \alpha_i = \mu_{a.} - \mu_{..}$

3. $\beta_j = \mu_{.j} - \mu_{..},\ j = 1,2,\ldots,b-1$ and $-\sum_{j=1}^{b-1} \beta_i = \mu_{.b} - \mu_{..}$

4. $\gamma_{ij} = \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu_{..},\ i = 1,2,\ldots,a-1;\ j = 1,2,\ldots,b-1$

$$-\sum_{i=1}^{a-1} \gamma_{ij} = \mu_{aj} - \mu_{a.} - \mu_{.j} + \mu_{..},\quad j = 1,2,\ldots,b-1$$

$$-\sum_{j=1}^{b-1} \gamma_{ij} = \mu_{ib} - \mu_{i.} - \mu_{.b} + \mu_{..},\quad i = 1,2,\ldots,a-1$$

On the other hand, the means can be expressed in terms of the coefficients of the regression (and also are helpful for us to be able to interpret the output from statistical packages):

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij},\ i = 1,..,a-1; j = 1,2,...,b-1$$

$$\mu_{aj} = \mu + \beta_j,\ j = 1,...,b-1,\ \mu_{ib} = \mu + \alpha_i,\ i = 1,...,a-1, \text{ and } \mu_{ab} = \mu$$

The treatment and block marginal means are:

$$\mu_{i.} = \mu + \alpha_i + \sum_{j=1}^{b-1}\left\{\frac{\beta_j + \gamma_{ij}}{b}\right\},\ i = 1,..,a-1, \text{ and } \mu_{a.} = \mu + \sum_{i=1}^{b-1}\frac{\beta_j}{b}$$

$$\mu_{.j} = \mu + \beta_j + \sum_{i=1}^{a-1}\left\{\frac{\alpha_i + \gamma_{ij}}{a}\right\},\ j = 1,..,b-1, \text{ and } \mu_{.b} = \mu + \sum_{i=1}^{a-1}\frac{\alpha_i}{a}$$

Example: Toxic substance and industrial plant data

Recall the industrial data on exposure to three toxic substances in three different industrial plants. In this example, we have *a*=3, *b*=3 and *n*=12 and the model can be expressed as

$$Y = \mu + \sum_{i=1}^{2} \alpha_i X_i + \sum_{j=1}^{2} \beta_j Z_j + \sum_{i=1}^{2} \sum_{j=1}^{2} \gamma_{ij} X_i Z_j \varepsilon_{ij}$$

where $X_i = \begin{cases} 1, \text{ if toxic substance } i \\ 0, \text{ otherwise} \end{cases} \quad i = A, B, C$ and $Z_j = \begin{cases} 1, \text{ if industrial plant } j \\ 0, \text{ otherwise} \end{cases} \quad j = 1,2,3$.

The STATA output (using the `xi` command) is as follows:

```
. char plant[omit] 3
. char toxsub[omit] 3
. xi plant toxsub
. xi: reg FEV i.plant i.toxsub i.plant*i.toxsub
i.plant              Iplant_1-3   (naturally coded; Iplant_1 omitted)
i.toxsub             Itoxsu_1-3   (naturally coded; Itoxsu_1 omitted)
i.plant*i.toxsub     IpXt_#-#     (coded as above)

  Source |       SS       df       MS                  Number of obs =     108
---------+------------------------------               F(  8,    99) =   44.10
   Model |  94.6984615     8  11.8373077               Prob > F      =  0.0000
Residual |   26.575859    99   .26844302               R-squared     =  0.7809
---------+------------------------------               Adj R-squared =  0.7632
   Total |   121.27432   107  1.13340486               Root MSE      =  .51811
```

The estimates of the various parameters are given as follows:

FEV	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Iplant_1	-.3891667	.2115195	-1.840	0.069	-.8088673	.0305339
Iplant_2	-.44	.2115195	-2.080	0.040	-.8597006	-.0202994
Itoxsu_1	.9925	.2115195	4.692	0.000	.5727994	1.412201
Itoxsu_2	1.989167	.2115195	9.404	0.000	1.569466	2.408867
Iplant_1	(dropped)					
Iplant_2	(dropped)					
Itoxsu_1	(dropped)					
Itoxsu_2	(dropped)					
IpXt_1_1	1.1575	.2991338	3.870	0.000	.5639538	1.751046
IpXt_1_2	-1.273333	.2991338	-4.257	0.000	-1.86688	-.679787
IpXt_2_1	1.544167	.2991338	5.162	0.000	.9506204	2.137713
IpXt_2_2	-.9091666	.2991338	-3.039	0.003	-1.502713	-.3156203
_cons	3.1375	.1495669	20.977	0.000	2.840727	3.434273

This is equivalent output with that of an analysis of variance (although possibly less easy to reada) as follows:

FEV	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_cons	3.1375	.1495669	20.977	0.000	2.840727	3.434273
plant						
1	-.3891667	.2115195	-1.840	0.069	-.8088673	.0305339
2	-.44	.2115195	-2.080	0.040	-.8597006	-.0202994
3	(dropped)					
toxsub						
1	.9925	.2115195	4.692	0.000	.5727994	1.412201
2	1.989167	.2115195	9.404	0.000	1.569466	2.408867
3	(dropped)					
plant*toxsub						
1 1	1.1575	.2991338	3.870	0.000	.5639538	1.751046
1 2	-1.273333	.2991338	-4.257	0.000	-1.86688	-.679787
1 3	(dropped)					
2 1	1.544167	.2991338	5.162	0.000	.9506204	2.137713
2 2	-.9091666	.2991338	-3.039	0.003	-1.502713	-.3156203
2 3	(dropped)					
3 1	(dropped)					
3 2	(dropped)					
3 3	(dropped)					

Comments:

1. Notice that by default, the first (lowest) level of the variable plant and toxic substance is used as the reference level. Also, all interactions that involve these levels are "dropped" leaving only those factors that can be estimated with the available degrees of freedom. We have however, reparametrized the model leaving out the last (highest) level, to make it identical to a two-way analysis of variance.
2. To derive the equations of the model for each plant and toxic substance we have for all cells. For the three plants and toxic substances we have:

   $\overline{Y}_{11} = 3.1375 + (-0.3892) + 0.9925 + 1.1575 = 4.8983$, $\overline{Y}_{12} = 3.4642$, $\overline{Y}_{13} = 3.1375 + (-0.3892) = 2.7483$

   $\overline{Y}_{21} = 5.2342$, $\overline{Y}_{22} = 3.7775$ and $\overline{Y}_{23} = 2.6975$

   $\overline{Y}_{31} = 4.1300$, $\overline{Y}_{32} = 3.1267$ and $\overline{Y}_{33} = 3.1375 = \mu$.

   Since there is significant interaction, we there is no need to estimate the marginal (`plant/toxsub` means). The cell means can also be estimated using the STATA commands,

```
sort plant toxsub
by plant toxsub: sum FEV
```

Regression models for the analysis of covariance

The analysis of covariance can also be expressed in terms of a linear regression by reparametrizing the fixed effect in the usual way. The complete ANACOVA model (including interaction is as follows:)

$$Y = \beta_o + \beta_1 X + \beta_2 Z + \beta_3 XZ + \varepsilon$$

For example, in the gender and systolic blood pressure example, we define the model above by coding the gender variable $X_i = \begin{cases} 0, \text{ if gender} = \text{female} \\ 1, \text{ if gender} = \text{male} \end{cases}$,Z=age and XZ is the age/gender interaction.

With this parametrization, the model for the males and females are:

$$\text{Males: } Y_M = \underbrace{(\beta_o + \beta_1)}_{\beta_{oM}} + \underbrace{(\beta_2 + \beta_3)}_{\beta_{1M}} Z + \varepsilon$$

$$\text{Females: } Y_F = \beta_o + \beta_2 Z + \varepsilon .$$

The STATA output from the dummy-variable regression of the gender and systolic blood pressure example follows:

```
. char gender[omit] 1
. xi: reg SBP age i.gender i.gender*age
i.gender              Igende_0-1    (naturally coded; Igende_1 omitted)
i.gender*age          IgXage_#      (coded as above)

  Source |       SS       df       MS                  Number of obs =      69
---------+------------------------------               F(  3,    65) =   75.02
   Model |  18010.3287     3   6003.4429               Prob > F      =  0.0000
Residual |   5201.4394    65  80.0221447               R-squared     =  0.7759
---------+------------------------------               Adj R-squared =  0.7656
   Total |  23211.7681    68  341.349531               Root MSE      =  8.9455

------------------------------------------------------------------------------
     SBP |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     age |   .9493225   .1086412      8.738   0.000        .732351    1.166294
Igende_0 |   12.96144   7.011725      1.849   0.069      -1.041936    26.96482
Igende_0 |  (dropped)
     age |  (dropped)
IgXage_0 |   .0120301   .1451933      0.083   0.934      -.2779409    .3020011
   _cons |   97.07708   5.170455     18.775   0.000       86.75097    107.4032
------------------------------------------------------------------------------
```

Consider the equivalent ANACOVA produced by the `anova` command in STATA (which is performing regression since all variables are continuous):

```
. anova SBP age Igende_0 IgXage_0, continuous(age Igende_0 IgXage_0)

                           Number of obs =      69     R-squared     =  0.7759
                           Root MSE      = 8.94551     Adj R-squared =  0.7656

                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  18010.3287     3   6003.4429      75.02     0.0000
                         |
                     age |  6110.10173     1  6110.10173      76.36     0.0000
                Igende_0 |  273.443297     1  273.443297       3.42     0.0691
                IgXage_0 |  .549356192     1  .549356192       0.01     0.9342
                         |
                Residual |   5201.4394    65  80.0221447
              -----------+----------------------------------------------------
                   Total |  23211.7681    68  341.349531
```

```
. reg

  Source |       SS       df       MS                  Number of obs =      69
---------+------------------------------               F(  3,    65) =   75.02
   Model |  18010.3287     3  6003.4429                Prob > F      =  0.0000
Residual |   5201.4394    65  80.0221447               R-squared     =  0.7759
---------+------------------------------               Adj R-squared =  0.7656
   Total |  23211.7681    68  341.349531               Root MSE      =  8.9455

------------------------------------------------------------------------------
     SBP         Coef.   Std. Err.        t     P>|t|       [95% Conf. Interval]
------------------------------------------------------------------------------
_cons         97.07708   5.170455     18.775   0.000       86.75097    107.4032
age           .9493225   .1086412      8.738   0.000        .732351    1.166294
Igende_0      12.96144   7.011725      1.849   0.069      -1.041936    26.96482
IgXage_0      .0120301   .1451933      0.083   0.934      -.2779409    .3020011
------------------------------------------------------------------------------
```

Using the new variables `Igende_0` (≡`gender==0`) and `IgXage_0` (≡`(gender==0)*age`).

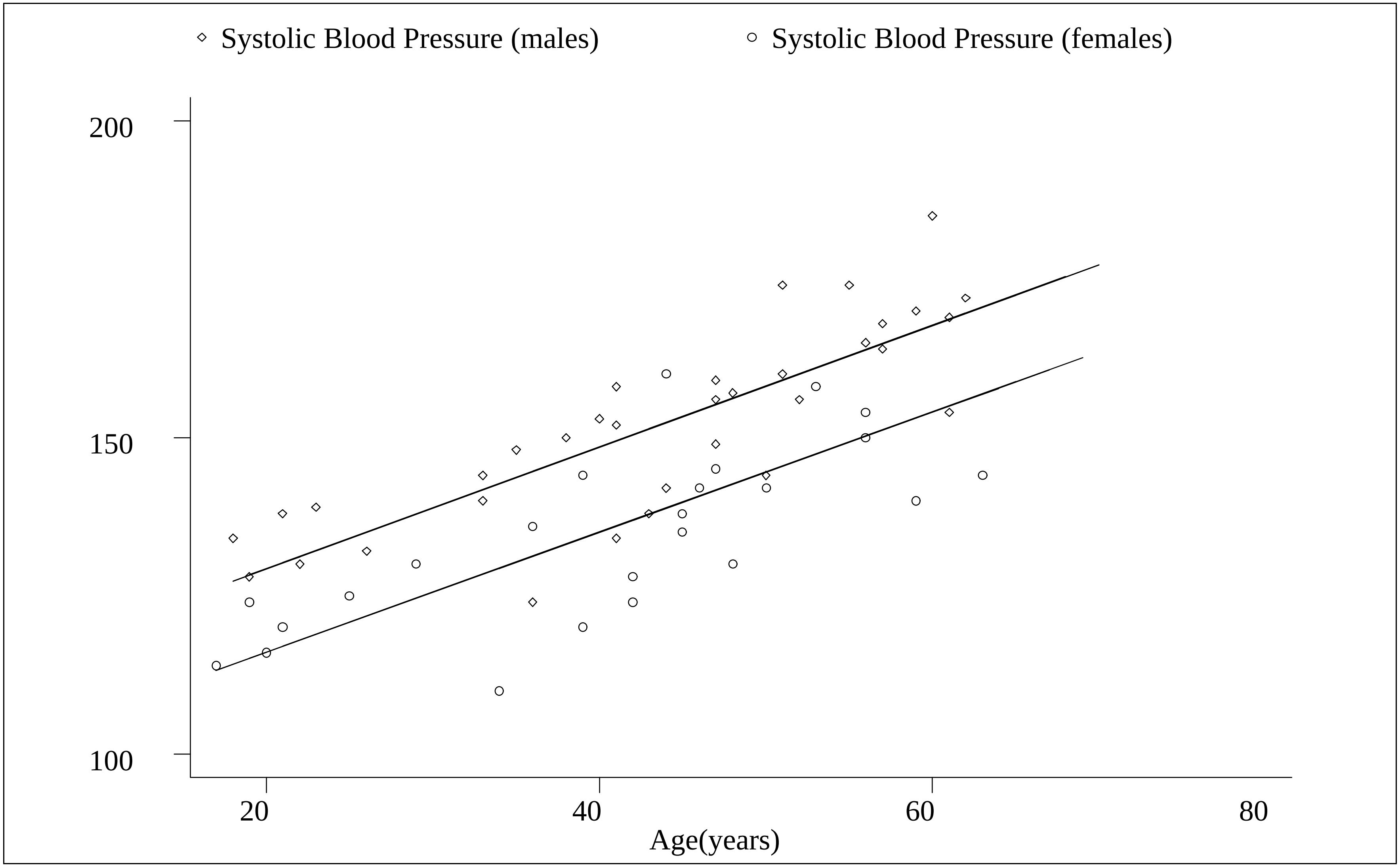
Notice that it would be equivalent in this case to type

`anova age gender age*gender, continuous(age gender)`

We see that the *t* tests in the regression are exactly equivalent to the partial (Type III) *F* tests above.

Comments:

1. The main effect, estimate of the slope of the covariate and the interaction effect are a bit easier to read from the output, since the reparametrization of the model permits direct estimation.
2. From the output, we see that the estimate of the intercept for females is `_cons=`$\hat{\beta}_{oF} = 97.0771$, while that of the slope for age $\hat{\beta}_{age} = 0.9493$ and the estimate of the interaction effect $\hat{\gamma}_{age \times gender} = 0.01203$ (disregard for a moment that the interaction term is not significant)
3. The *t* tests associated with the estimates of the slope and interaction are 8.738 (p-value<0.0001) and 1.849 (p-value 0.069) and if we square them we get 76.353 and 3.419 respectively which are virtually equivalent to the partial (Type III) *F* statistics that were produced in the ANOVA output.
4. The model for males is $Y = (97.0771 + 12.9614) + (0.9493 + 0.0120)\text{age} = 109.9385 + 0.9613 \times \text{age}$, while for the females it is $Y = 97.0771 + 0.9493 \times \text{age}$.


Systolic Blood Pressure (males)
Systolic Blood Pressure (females)
200
150
100
20
40
60
80
Age(years)