Multiple linear regression 2

Morten Frydenberg ©
Institut for Biostatistik

Categorical variables in regression models.

The changing reference level

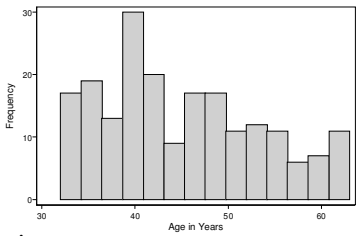
Interaction/effect modification

Interaction between a categorical and continuous variable

Interaction between two categorical variables

Categorical variable in regression models

The age distribution:



Let us divide *age* into **three** agegroups ,

0: $age \le 40$, 1: $40 < age \le 50$, 2: $50 < age$

and consider the new model

$$\ln(sbp) = \alpha_0 + \alpha_1 \cdot age1 + \alpha_2 \cdot age2 + \alpha_3 \cdot woman + \alpha_4 \cdot \ln\left(\frac{bmi}{25}\right) + E$$

Categorical variable : age group 0 reference

$$\ln(sbp) = \alpha_0 + \alpha_1 \cdot age1 + \alpha_2 \cdot age2 + \alpha_3 \cdot woman + \alpha_4 \cdot \ln\left(\frac{bmi}{25}\right) + E$$

age1 is one if a person is in age group **1** and zero otherwise
age2 is one if a person is in age group **2** and zero otherwise

The expected $\ln(sbp)$ in the three age groups will be:

$age \le 40$: $\ln(sbp) = \alpha_0 + \alpha_3 \cdot woman + \alpha_4 \cdot \ln(bmi)$

$40 < age \le 50$: $\ln(sbp) = \alpha_0 + \alpha_1 + \alpha_3 \cdot woman + \alpha_4 \cdot \ln(bmi)$

$50 < age$: $\ln(sbp) = \alpha_0 + \alpha_2 + \alpha_3 \cdot woman + \alpha_4 \cdot \ln(bmi)$

We see that α_1 is the adjusted difference in $\ln(sbp)$ when comparing a person in the **second group** with one in the **first group**.

And α_2 is the adjusted difference in $\ln(sbp)$ when comparing a person in the **third group** with one in the **first group**.

Categorical variable : age group 0 reference

$$\ln(sbp) = \alpha_0 + \alpha_1 \cdot age1 + \alpha_2 \cdot age2 + \alpha_3 \cdot woman + \alpha_4 \cdot \ln\left(\frac{bmi}{25}\right) + E$$

Finally we see that α_0 is the expected $\ln(sbp)$ for a man in the **first** (reference) age group, with bmi=25.

In most programs the model is fitted by first generating the grouping variable and then making the regression telling the program which variables are categorical.

In STATA this done like this:

```
egen agegrp3=cut(age), at(0,40,50,120) label
xi: regress lnSBP woman i.agegrp3 lnBMI25
```

Categorical variables included

This is categorical

Categorical variable : age group 0 reference

$$\ln(sbp) = \alpha_0 + \alpha_1 \cdot age1 + \alpha_2 \cdot age2 + \alpha_3 \cdot woman + \alpha_4 \cdot \ln\left(\frac{bmi}{25}\right) + E$$

`agegrp3` is treated as categorical

The value 0 is chosen as reference

OUTPUT:

```
i.agegrp3         _Iagegrp3_0-2       (naturally coded; _Iagegrp3_0 omitted)

  Source |       SS       df       MS              Number of obs =     200
---------+------------------------------           F(  4,   195) =   11.20
   Model |  .980169926     4  .245042482           Prob > F      =  0.0000
Residual |  4.26524771   195  .021873065           R-squared     =  0.1869
---------+------------------------------           Adj R-squared =  0.1702
   Total |  5.24541764   199  .026358883           Root MSE      =  .1479

------------------------------------------------------------------------------
       lnSBP |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       woman |   .0035403   .0212026     0.17   0.868    -.0382757    .0453562
_Iagegrp3_1 |   .0715136   .0253373     2.82   0.005     .0215432     .121484
_Iagegrp3_2 |    .130465   .0280521     4.65   0.000     .0751404    .1857895
    lnBMI25 |   .2898622   .0772432     3.75   0.000     .1375229    .4422015
      _cons |   4.789641   .0224814   213.05   0.000     4.745303    4.833979
------------------------------------------------------------------------------
```


Categorical variable : age group 0 reference

$$\ln(sbp) = \alpha_0 + \alpha_1 \cdot age1 + \alpha_2 \cdot age2 + \alpha_3 \cdot woman + \alpha_4 \cdot \ln\left(\frac{bmi}{25}\right) + E$$

Adjusted difference between for a person in **age group 1** compared to **age group 0**

```
------------------------------------------------------------------------------
       lnSBP |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       woman |   .0035403   .0212026     0.17   0.868    -.0382757    .0453562
_Iagegrp3_1 |   .0715136   .0253373     2.82   0.005     .0215432     .121484
_Iagegrp3_2 |    .130465   .0280521     4.65   0.000     .0751404    .1857895
    lnBMI25 |   .2898622   .0772432     3.75   0.000     .1375229    .4422015
      _cons |   4.789641   .0224814   213.05   0.000     4.745303    4.833979
------------------------------------------------------------------------------
```

Adjusted difference between for a person in **age group 2** compared to **age group 0**

Expected value for a man in **age group 0** with bmi=25.

Categorical variable : age group 1 reference

$$\ln(sbp) = \gamma_0 + \gamma_1 \cdot age0 + \gamma_2 \cdot age2 + \gamma_3 \cdot woman + \gamma_4 \cdot \ln\left(\frac{bmi}{25}\right) + E$$

age0 is one if a person is in age group **0** and zero otherwise
age2 is one if a person is in age group **2** and zero otherwise

The expected $\ln(sbp)$ in the three age groups will be:

$age \le 40$: $\ln(sbp) = \gamma_0 + \gamma_1 + \gamma_3 \cdot woman + \gamma_4 \cdot \ln(bmi)$

$40 < age \le 50$: $\ln(sbp) = \gamma_0 + \gamma_3 \cdot woman + \gamma_4 \cdot \ln(bmi)$

$50 < age$: $\ln(sbp) = \gamma_0 + \gamma_2 + \gamma_3 \cdot woman + \gamma_4 \cdot \ln(bmi)$

We see that γ_1 is the adjusted difference in $\ln(sbp)$ when comparing a person in the **first group** with one in the **second group**.

And γ_2 is the adjusted difference in $\ln(sbp)$ when comparing a person in the **third group** with one in the **second group**.

Categorical variable : age group 0 reference

$$\ln(sbp) = \gamma_0 + \gamma_1 \cdot age0 + \gamma_2 \cdot age2 + \gamma_3 \cdot woman + \gamma_4 \cdot \ln\left(\frac{bmi}{25}\right) + E$$

Finally we see that γ_0 is the expected $\ln(sbp)$ for a man in the **second** (reference) age group, with bmi=25.

Many programs (but regression in SPSS) let you chose the reference group

In STATA this done like this:

```
char agegrp3[omit] 1
```

The group label 1 is reference

```
xi: regress lnSBP woman i.agegrp3 lnBMI25
```

Categorical variables included

This is categorical

Categorical variable : age group 1 reference

$$\ln(sbp) = \gamma_0 + \gamma_1 \cdot age0 + \gamma_2 \cdot age2 + \gamma_3 \cdot woman + \gamma_4 \cdot \ln\left(\frac{bmi}{25}\right) + E$$

agegrp3 is treated as categorical

The value 1 is chosen as reference

```
i.agegrp3         _Iagegrp3_0-2     (naturally coded; _Iagegrp3_1 omitted)

  Source |       SS       df       MS              Number of obs =     200
---------+------------------------------           F(  4,   195) =   11.20
   Model |  .980169926     4  .245042482           Prob > F      =  0.0000
Residual |  4.26524771   195  .021873065           R-squared     =  0.1869
---------+------------------------------           Adj R-squared =  0.1702
   Total |  5.24541764   199  .026358883           Root MSE      =   .1479

------------------------------------------------------------------------------
       lnSBP |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       woman |   .0035403   .0212026     0.17   0.868    -.0382757    .0453562
 _Iagegrp3_0 |  -.0715136   .0253373    -2.82   0.005    -.121484    -.0215432
 _Iagegrp3_2 |   .0589513   .0263496     2.24   0.026     .0069846    .1109181
     lnBMI25 |   .2898622   .0772432     3.75   0.000     .1375229    .4422015
       _cons |   4.861154   .0207406   234.4   0.000     4.82025     4.902059
------------------------------------------------------------------------------
```


Categorical variable : age group 1 reference

$$\ln(sbp) = \alpha_0 + \alpha_1 \cdot age1 + \alpha_2 \cdot age2 + \alpha_3 \cdot woman + \alpha_4 \cdot \ln\left(\frac{bmi}{25}\right) + E$$

Adjusted difference between for a person in **age group 0** compared to **age group 1**

```
------------------------------------------------------------------------------
       lnSBP |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       woman |   .0035403   .0212026     0.17   0.868    -.0382757    .0453562
 _Iagegrp3_0 |  -.0715136   .0253373    -2.82   0.005    -.121484    -.0215432
 _Iagegrp3_2 |   .0589513   .0263496     2.24   0.026     .0069846    .1109181
     lnBMI25 |   .2898622   .0772432     3.75   0.000     .1375229    .4422015
       _cons |   4.861154   .0207406   234.4   0.000     4.82025     4.902059
------------------------------------------------------------------------------
```

Adjusted difference between for a person in **age group 2** compared to **age group 1**

Expected value for a man in **age group 1** with bmi=25.

Categorical variable: Comparing to parameterizations

$$\ln(sbp) = \alpha_0 + \alpha_1 \cdot age1 + \alpha_2 \cdot age2 + \alpha_3 \cdot woman + \alpha_4 \cdot \ln\left(\frac{bmi}{25}\right) + E$$

$$\ln(sbp) = \gamma_0 + \gamma_1 \cdot age0 + \gamma_2 \cdot age2 + \gamma_3 \cdot woman + \gamma_4 \cdot \ln\left(\frac{bmi}{25}\right) + E$$

age group 0: $\alpha_0 = \gamma_0 + \gamma_1$

age group 1: $\alpha_0 + \alpha_1 = \gamma_0$

age group 2: $\alpha_0 + \alpha_2 = \gamma_0 + \gamma_2$

$$\alpha_0 = \gamma_0 + \gamma_1 \qquad \gamma_0 = \alpha_0 + \alpha_1$$
$$\alpha_1 = -\gamma_1 \qquad \gamma_1 = -\alpha_1$$
$$\alpha_2 = \gamma_2 - \gamma_1 \qquad \gamma_2 = \alpha_2 - \alpha_1$$
$$\alpha_3 = \gamma_3 \qquad \gamma_3 = \alpha_3$$
$$\alpha_4 = \gamma_4 \qquad \gamma_4 = \alpha_4$$

Categorical variable: Comparing two parameterizations

The estimates:

age group 0 reference

```
Root MSE      =   .1479
------------------------------------------------------------------------------
       lnSBP |      Coef.   [95% CI]
-------------+----------------------------------------------------------------
       woman |   .0035 -.0382   .0453
 _Iagegrp3_1 |   .0715   .0215   .1214
 _Iagegrp3_2 |   .1304   .0751   .1857
     lnBMI25 |   .2898   .1375   .4422
       _cons |  4.7896   4.745   4.833
------------------------------------------------------------------------------
```

age group 1 reference

```
Root MSE      =   .1479
------------------------------------------------------------------------------
       lnSBP |      Coef.   [95% CI]
-------------+----------------------------------------------------------------
       woman |   .0035   -.0382   .0453
 _Iagegrp3_0 |  -.0715   -.1214  -.0215
 _Iagegrp3_2 |   .0589    .0069   .1109
     lnBMI25 |   .2898    .1375   .4422
       _cons |  4.8611    4.820   4.902
------------------------------------------------------------------------------
```

Note, the estimates fulfil the same equations.

The interpretation of the " _Iagegrp3_2 line" and " _cons line" are altered!!!!!!!!

Always remember: what is the reference group!

Categorical variable : age group 1 reference

```
------------------------------------------------------------------------------
       lnSBP |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       woman |   .0035403   .0212026     0.17   0.868    -.0382757    .0453562
_Iagegrp3_0 |  -.0715136   .0253373    -2.82   0.005    -.121484    -.0215432
_Iagegrp3_2 |   .0589513   .0263496     2.24   0.026     .0069846    .1109181
    lnBMI25 |   .2898622   .0772432     3.75   0.000     .1375229    .4422015
      _cons |  4.861154    .0207406   234.4   0.000    4.82025      4.902059
------------------------------------------------------------------------------
```

Two test:

One testing no difference between age group **0** and **1**.
One testing no difference between age group **2** and **1**.

Can we get **one test** testing no difference between age groups?

A F-test in STATA: `testparm _Iagegrp3*`

Highly significant

```
 ( 1)  _Iagegrp3_0 = 0
 ( 2)  _Iagegrp3_2 = 0
       F(  2,   195) =   10.93
            Prob > F =    0.0000
```


Interactions/effectmodification

$$\ln(sbp) = \gamma_0 + \gamma_1 \cdot age0 + \gamma_2 \cdot age2 + \gamma_3 \cdot woman + \gamma_4 \cdot \ln\left(\frac{bmi}{25}\right) + E$$

One of the **central** assumptions was "no effect modification".

E.g. in model above the "effect" of **age**, **sex** and **bmi** did not depend on the value of each other.

One can **introduce** effect modification between a categorical variable and another variable.

Here we first will look at **agegr3** and **lnBMI25**.

The effect modification will be that the coefficient to **lnBMI25** depend on **age group**.

That is, we will allow different effect of bmi in the different age groups.

Interactions/effectmodification

$$\ln(sbp) = \omega_0 + \omega_1 \cdot age0 + \omega_2 \cdot age2 + \omega_3 \cdot woman + \omega_4 \cdot \ln\left(\frac{bmi}{25}\right) + \omega_5 \cdot age0 \cdot \ln\left(\frac{bmi}{25}\right) + \omega_6 \cdot age1 \cdot \ln\left(\frac{bmi}{25}\right) + E$$

$age \le 40$: $\ln(sbp) = (\omega_0 + \omega_1) + (\omega_4 + \omega_5) \cdot \ln\left(\frac{bmi}{25}\right) + \omega_3 \cdot woman$

$40 < age \le 50$: $\ln(sbp) = \omega_0 + \omega_4 \cdot \ln\left(\frac{bmi}{25}\right) + \omega_3 \cdot woman$

$50 < age$: $\ln(sbp) = (\omega_0 + \omega_2) + (\omega_4 + \omega_6) \cdot \ln\left(\frac{bmi}{25}\right) + \omega_3 \cdot woman$

ω_0 is the **difference** between the constant for age **group 0** and **reference** group.
ω_5 is the **difference** between the coefficient to `lnBMI25` for age **group 0** and **reference** group.

Interactions/effectmodification

```
. xi: regress lnSBP woman i.agegrp3*lnBMI25
i.agegrp3         _Iagegrp3_0-2     (naturally coded; _Iagegrp3_1 omitted)
i.ageg~3*lnB~25   _IageXlnBMI_#     (coded as above)

  Source |       SS       df       MS              Number of obs =     200
---------+------------------------------           F(  6,   193) =    7.53
   Model |  .994860827     6  .165810138           Prob > F      =  0.0000
Residual |  4.25055681   193  .02202361            R-squared     =  0.1897
---------+------------------------------           Adj R-squared =  0.1645
   Total |  5.24541764   199  .026358883           Root MSE      =  .1484

------------------------------------------------------------------------------
       lnSBP |      Coef.   Std. E.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       woman |   .0076438   .02191     0.35   0.728    -.03558      .0508772
 _Iagegrp3_0 |  -.0708045   .02611    -2.71   0.007    -.1223213    -.0192877
 _Iagegrp3_2 |   .0631082   .02703     2.33   0.021     .009787     .1164287
    lnBMI25 |   .3155479   .12229     2.58   0.011     .074350     .5567453
_IageXlnBM~0 |   .0429736   .19123     0.22   0.822    -.334209     .420157
_IageXlnBM~2 |  -.1165375   .18554    -0.63   0.531    -.482499     .2494241
       _cons |   4.859743   .02141   226.9   0.000     4.81751      4.901975
------------------------------------------------------------------------------
```


Interactions/effect modification

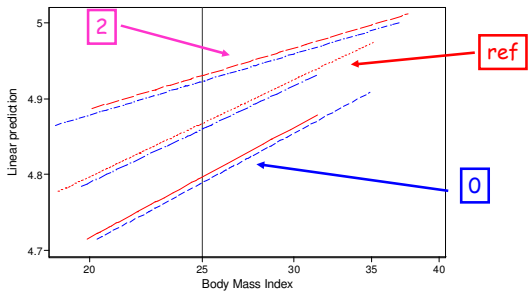
Ref: constant and 'slope' in reference group
0: difference in constant and slope compared to reference
2: difference in constant and slope compared to reference

```
       lnSBP |      Coef.   Std. E.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       woman |   .0076438   .02191     0.35   0.728    -.03558      .0508772
 _Iagegrp3_0 |  -.0708045   .02611    -2.71   0.007    -.1223213    -.0192877
 _Iagegrp3_2 |   .0631082   .02703     2.33   0.021     .009787     .1164287
    lnBMI25 |   .3155479   .12229     2.58   0.011     .074350     .5567453
_IageXlnBM~0 |   .0429736   .19123     0.22   0.822    -.334209     .420157
_IageXlnBM~2 |  -.1165375   .18554    -0.63   0.531    -.482499     .2494241
       _cons |   4.859743   .02141   226.9   0.000     4.81751      4.901975
------------------------------------------------------------------------------
```

0 2 Ref 0 2 Ref

Note the larger standard errors

Based on the estimates one can find the six "dose-response" curves:



log scale

No effect modification ⇔ parallel lines.

Interactions/effect modification

lnSBP	Coef.	Std. E.	t	P>|t|	[95% Conf.	Interval]
woman	.0076438	.02191	0.35	0.728	-.03558	.0508772
_Iagegrp3_0	-.0708045	.02611	-2.71	0.007	-.1223213	-.0192877
_Iagegrp3_2	.0631082	.02703	2.33	0.021	.009787	.1164287
lnBMI25	.3155479	.12229	2.58	0.011	.074350	.5567453
_IageXlnBM~0	.0429736	.19123	0.22	0.822	-.334209	.420157
_IageXlnBM~2	-.1165375	.18554	-0.63	0.531	-.482499	.2494241
_cons	4.859743	.02141	226.9	0.000	4.81751	4.901975

Two test:
One testing differences between "slope" in age group **0** and **1**.
One testing differences between "slope" age group **2** and **1**.

Can we get **one test** testing no difference between age groups?

A F-test in STATA: `testparm _IageX*`

```
( 1)  _IageXlnBMI_0 = 0
( 2)  _IageXlnBMI_2 = 0
      F(  2,   193) =    0.33
           Prob > F =    0.7168
```

Non-significant

Interactions/effect modification

The test of no interaction was non-significant.

But look at the **confidence interval** for the difference in slope for age group **2** and group **1** !

lnSBP	Coef.	Std. E.	t	P>|t|	[95% Conf.	Interval]
woman	.0076438	.02191	0.35	0.728	-.03558	.0508772
_Iagegrp3_0	-.0708045	.02611	-2.71	0.007	-.1223213	-.0192877
_Iagegrp3_2	.0631082	.02703	2.33	0.021	.009787	.1164287
lnBMI25	.3155479	.12229	2.58	0.011	.074350	.5567453
_IageXlnBM~0	.0429736	.19123	0.22	0.822	-.334209	.420157
_IageXlnBM~2	-.1165375	.18554	-0.63	0.531	-.482499	.2494241
_cons	4.859743	.02141	226.9	0.000	4.81751	4.901975

It is very wide!!! We know very little about this difference!

The test for no interaction has **very low power**!!!

The data have very **little information** on whether there is effect modification.

Interaction between age group and sex

`xi: regress lnSBP lnBMI25 i.agegrp3*i.sex`

```
i.agegrp3         _Iagegrp3_0-2       (naturally coded; _Iagegrp3_1 omitted)
i.sex             _Isex_1-2           (naturally coded; _Isex_1 omitted)
i.age~3*i.sex     _IageXsex_#_#       (coded as above)

      Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  6,   193) =    9.96
       Model |  1.24006476     6   .20667746           Prob > F      =  0.0000
    Residual |  4.00535288   193  .020753124           R-squared     =  0.2364
-------------+------------------------------           Adj R-squared =  0.2127
       Total |  5.24541764   199  .026358883           Root MSE      =  .14406
```

lnSBP	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnBMI25	.2265018	.0774898	2.92	0.004	.0736662	.3793374
_Iagegrp3_0	-.0426734	.0377096	-1.13	0.259	-.1170493	.0317025
_Iagegrp3_2	-.0025412	.0365457	-0.07	0.945	-.0746215	.0695391
_Isex_2	-.0210869	.0322283	-0.65	0.514	-.0846518	.042478
_IageXse~0_2	-.0548967	.0501668	-1.09	0.275	-.1538422	.0440488
_IageXse~2_2	.133379	.0501308	2.66	0.008	.0345043	.2322536
_cons	4.873442	.0251767	193.57	0.000	4.823786	4.923099

```
. testparm _IageX*
 ( 1)  _IageXsex_0_2 = 0
 ( 2)  _IageXsex_2_2 = 0
       F(  2,   193) =    6.26
            Prob > F =    0.0023
```

Highly significant

Interaction between age group and sex

Differences between age groups among men is small

lnSBP	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnBMI25	.2265018	.0774898	2.92	0.004	.0736662	.3793374
_Iagegrp3_0	-.0426734	.0377096	-1.13	0.259	-.1170493	.0317025
_Iagegrp3_2	-.0025412	.0365457	-0.07	0.945	-.0746215	.0695391
_Isex_2	-.0210869	.0322283	-0.65	0.514	-.0846518	.042478
_IageXse~0_2	-.0548967	.0501668	-1.09	0.275	-.1538422	.0440488
_IageXse~2_2	.133379	.0501308	2.66	0.008	.0345043	.2322536
_cons	4.873442	.0251767	193.57	0.000	4.823786	4.923099

Women age group 1: 4.873–0.021=4.894
Women age group 0: 4.873–0.021+0.042–0.055=4.755
Women age group 2: 4.873–0.021+0.003+0.133=4.982

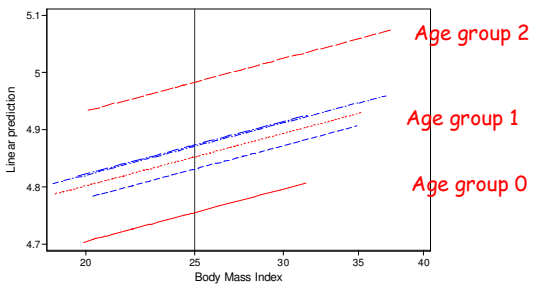
0.097

0.130

Large difference in the age groups among women.

Interaction between age group and sex

Using women as reference: `char sex[omit]2`

Large differences between age groups among women

lnSBP	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnBMI25	.2265018	.0774898	2.92	0.004	.0736662	.3793374
_Iagegrp3_0	-.0975701	.0328469	-2.97	0.003	-.162355	-.0327852
_Iagegrp3_2	.1308378	.0354804	3.69	0.000	.0608587	.2008168
_Isex_1	.0210869	.0322283	0.65	0.514	-.042478	.0846518
_IageXse~0_1	.0548967	.0501668	1.09	0.275	-.0440488	.1538422
_IageXse~2_1	-.133379	.0501308	-2.66	0.008	-.2322536	-.0345043
_cons	4.852355	.0205502	236.12	0.000	4.811824	4.892887