

Multiple linear regression 2

Stata 13

Morten Frydenberg ©

Section of Biostatistics, Aarhus Univ, Denmark

Categorical variables in regression models.

The changing reference level

Interaction/effect modification

Interaction between a categorical and continuous variable

Interaction between two categorical variables

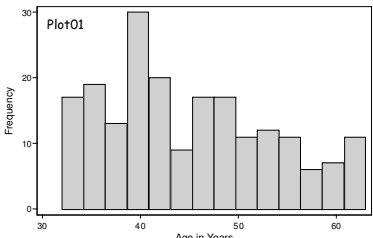
Morten Frydenberg

Linear and Logistic regression - Note 2.2

1

Categorical variable in regression models

The age distribution:



Let us divide age into three agegroups ,

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

and consider the new model

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

Morten Frydenberg

Linear and Logistic regression - Note 2.2

2

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 < 40: $\ln(sbp) = \alpha_0 + \alpha_3 \cdot woman + \alpha_4 \cdot \ln(bmi/25)$

$40 \leq age < 50$: $\ln(sbp) = \alpha_0 + \alpha_1 + \alpha_3 \cdot woman + \alpha_4 \cdot \ln(bmi/25)$

$50 \leq age$: $\ln(sbp) = \alpha_0 + \alpha_2 + \alpha_3 \cdot woman + \alpha_4 \cdot \ln(bmi/25)$

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.

Morten Frydenberg

Linear and Logistic regression - Note 2.2

3

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 is like this:

```
egen agegrp3=cut(age), at(0,40,50,120) tabe7  
regress lnSBP woman i.agegrp3 lnBMI25
```

This is categorical

Morten Frydenberg

Linear and Logistic regression - Note 2.2

4

Categorical variable : age group 0- reference

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

OUTPUT:					
Source	SS	df	MS	Number of obs = 200	
Model	.980169926	4	.245042482	F(4, 195) = 11.20	
Residual	4.26524771	195	.021873065	Prob > F = 0.0000	
				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
agegrp3						
40-	.0715136	.0253373	2.82	0.005	.0215432	.121484
50-	.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

Note 0- is missing: it is base/reference group

Morten Frydenberg

Linear and Logistic regression - Note 2.2

5

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 40- compared to age group 0-

lnSBP	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
woman	.0035403	.0212026	0.17	0.868	-.0382757	.0453562
agegrp3						
40-	.0715136	.0253373	2.82	0.005	.0215432	.121484
50-	.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 50- compared to age group 0

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

Morten Frydenberg

Linear and Logistic regression - Note 2.2

6

The expected values:

$$\text{age} < 40: \ln(sbp) = \alpha_0 + \alpha_3 \cdot \text{woman} + \alpha_4 \cdot \ln(bmi/25)$$
$$40 \leq \text{age} < 50: \ln(sbp) = \alpha_0 + \alpha_1 + \alpha_3 \cdot \text{woman} + \alpha_4 \cdot \ln(bmi/25)$$
$$50 \leq \text{age}: \ln(sbp) = \alpha_0 + \alpha_2 + \alpha_3 \cdot \text{woman} + \alpha_4 \cdot \ln(bmi/25)$$

The estimates

	lnSBP	Coef.	
woman	.003540	3	
40-	.071514	1	
50-	.130465	2	
lnBMI25	.289862	4	
_cons	4.789641	0	

$$\text{age} < 40: \ln(sbp) = 4.789 + 0.004 \cdot \text{woman} + 0.290 \cdot \ln(bmi/25)$$
$$40 \leq \text{age} < 50: \ln(sbp) = 4.789 + 0.072 + 0.004 \cdot \text{woman} + 0.290 \cdot \ln(bmi/25)$$
$$50 \leq \text{age}: \ln(sbp) = 4.789 + 0.130 + 0.004 \cdot \text{woman} + 0.290 \cdot \ln(bmi/25)$$

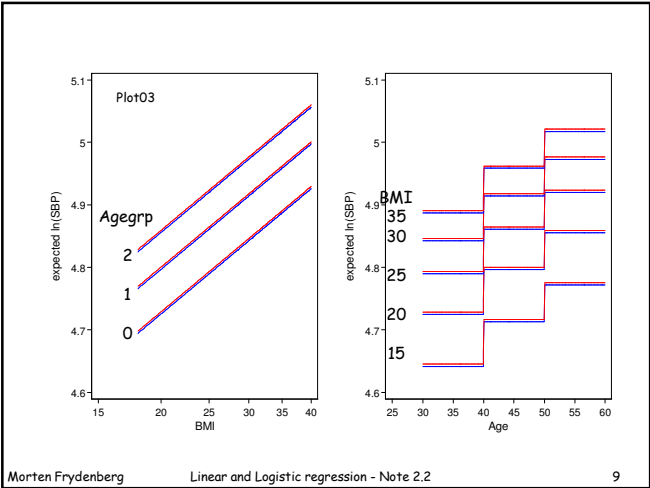
Morten FrydenbergLinear and Logistic regression - Note 2.27

Plot02

Agegroup

	Women	Men
0-:	$4.793 + 0.290 \cdot \ln(bmi/25)$	$4.790 + 0.290 \cdot \ln(bmi/25)$
40-:	$4.865 + 0.290 \cdot \ln(bmi/25)$	$4.861 + 0.290 \cdot \ln(bmi/25)$
50-:	$4.924 + 0.290 \cdot \ln(bmi/25)$	$4.920 + 0.290 \cdot \ln(bmi/25)$

Morten FrydenbergLinear and Logistic regression - Note 2.28



Morten FrydenbergLinear and Logistic regression - Note 2.29

Categorical variable : age group 40- reference

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

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:

$$\text{age} < 40: \ln(sbp) = \gamma_0 + \gamma_1 + \gamma_3 \cdot \text{woman} + \gamma_4 \cdot \ln(bmi/25)$$
$$40 \leq \text{age} < 50: \ln(sbp) = \gamma_0 + \gamma_3 \cdot \text{woman} + \gamma_4 \cdot \ln(bmi/25)$$
$$50 \leq \text{age}: \ln(sbp) = \gamma_0 + \gamma_2 + \gamma_3 \cdot \text{woman} + \gamma_4 \cdot \ln(bmi/25)$$

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**.

Morten FrydenbergLinear and Logistic regression - Note 2.210

Categorical variable : age group 40- reference

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

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 choose the reference group

Since Stata 11 this is done like this:

```
regress lnSBP woman b1.agegrp3 lnBMI25
```

This is a categorical variable and we want 1 to be the base/reference

Morten FrydenbergLinear and Logistic regression - Note 2.211

Categorical variable : age group 40- reference

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

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

	lnSBP	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
woman		.0035403	.0212026	0.17	0.868	-.0382757 .0453562
agegrp3						
0-		-.0715136	.0253373	-2.82	0.005	-.121484 -.0215432
40-		.0589513	.0263496	2.24	0.026	.0069846 .1109181
lnBMI25		.2898622	.0772432	3.75	0.000	.1375229 .4422015
_cons		4.861154	.0207406	234.38	0.000	4.82025 4.902059

Note 40- is missing: it is base/reference group

Morten FrydenbergLinear and Logistic regression - Note 2.212

Categorical variable : age group 40- reference

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

Adjusted difference between for a person in age group 0- compared to age group 40-

	lnSBP	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
woman		.0035403	.0212026	0.17	0.868	-.0382757 .0453562
agegrp3						
0-		-.0715136	.0253373	-2.82	0.005	-.121484 -.0215432
50-		.0589513	.0263496	2.24	0.026	.0069846 .1109181
lnBMI25		.2898622	.0772432	3.75	0.000	.1375229 .4422015
_cons		4.861154	.0207406	234.38	0.000	4.82025 4.902059

Adjusted difference between for a person in age group 50- compared to age group 40-

Expected value for a man in age group 40- with bmi=25.

Morten FrydenbergLinear and Logistic regression - Note 2.213

Categorical variable: Comparing two 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 40-: $\alpha_0 + \alpha_1 = \gamma_0$

age group 50-: $\alpha_0 + \alpha_2 = \gamma_0 + \gamma_2$

$$\alpha_0 = \gamma_0 + \gamma_1$$

$$\alpha_1 = -\gamma_1$$

$$\alpha_2 = \gamma_2 - \gamma_1$$

$$\alpha_3 = \gamma_3$$

$$\alpha_4 = \gamma_4$$

$$\gamma_0 = \alpha_0 + \alpha_1$$

$$\gamma_1 = -\alpha_1$$

$$\gamma_2 = \alpha_2 - \alpha_1$$

$$\gamma_3 = \alpha_3$$

$$\gamma_4 = \alpha_4$$

Morten FrydenbergLinear and Logistic regression - Note 2.214

Categorical variable: Comparing two parameterizations

The estimates:

age group 0- reference

age group 40- reference

Root MSE	=	.1479
lnSBP	Coef.	[95% CI]
woman	.0035	-.0382 .0453
agegrp3		
40-	.0715	.0215 .1214
50-	.1304	.0751 .1857
lnBMI25	.2898	.1375 .4422
_cons	4.7896	4.745 4.833

Root MSE	=	.1479
lnSBP	Coef.	[95% CI]
woman	.0035	-.0382 .0453
agegrp3		
0-	-.0715	-.1214 -.0215
50-	.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 "agegrp3 50- line" and "_cons line" are altered!!!!!!

Always remember: what is the reference group!

Morten FrydenbergLinear and Logistic regression - Note 2.215

Categorical variable : age group 40- reference

	lnSBP	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
woman		.0035403	.0212026	0.17	0.868	-.0382757 .0453562
agegrp3						
0-		-.0715136	.0253373	-2.82	0.005	-.121484 -.0215432
50-		.0589513	.0263496	2.24	0.026	.0069846 .1109181
lnBMI25		.2898622	.0772432	3.75	0.000	.1375229 .4422015
_cons		4.861154	.0207406	234.38	0.000	4.82025 4.902059

Two test:

One testing no difference between age group 0- and 40-.

One testing no difference between age group 50- and 40-.

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

A F-test in Stata: `testparm i.agegrp`

(1) 0.agegrp3 = 0

(2) 2.agegrp3 = 0

F(2, 195) = 10.93

Prob > F = 0.0000

Highly significant

Morten FrydenbergLinear and Logistic regression - Note 2.216

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 the 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 agegrp3 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.

Morten FrydenbergLinear and Logistic regression - Note 2.217

Interactions/effectmodification

$$\ln(sbp) = \omega_0 + \omega_1 \cdot age0 + \omega_2 \cdot age2 + \omega_3 \cdot woman$$

$$+ \omega_4 \cdot \ln\left(\frac{bmi}{25}\right) + \omega_5 \cdot age0 \cdot \ln\left(\frac{bmi}{25}\right) + \omega_6 \cdot age2 \cdot \ln\left(\frac{bmi}{25}\right) + E$$

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

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

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

$$\omega_1$$
 is the difference between the constant for age group 0 and reference group.

$$\omega_5$$
 is the difference between the coefficient to lnBMI25 for age group 0 and reference group.

Morten FrydenbergLinear and Logistic regression - Note 2.218

Interactions/effectmodification

regress lnSBP woman b1.agegrp3#c.lnBMI25

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

	lnSBP	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
woman		.0076438	.0219199	0.35	0.728	-.0355896 .0508772
agegrp3						
0-		-.0708045	.0261198	-2.71	0.007	-.1223213 -.0192877
50-		.0631082	.0270342	2.33	0.021	.0097877 .1164287
lnBMI25		.3155479	.1222905	2.58	0.011	.0743505 .5567453
agegrp3#lnBMI25						
0-		.0429736	.1912373	0.22	0.822	-.3342099 .420157
50-		-.1165375	.1855477	-0.63	0.531	-.4824991 .2494242
_cons		4.859743	.0214122	226.96	0.000	4.817511 4.901975

Morten Frydenberg Linear and Logistic regression - Note 2.2 19

Interactions/effect modification

Ref: constant and 'slope' in reference group

0-: difference in constant and slope compared to reference

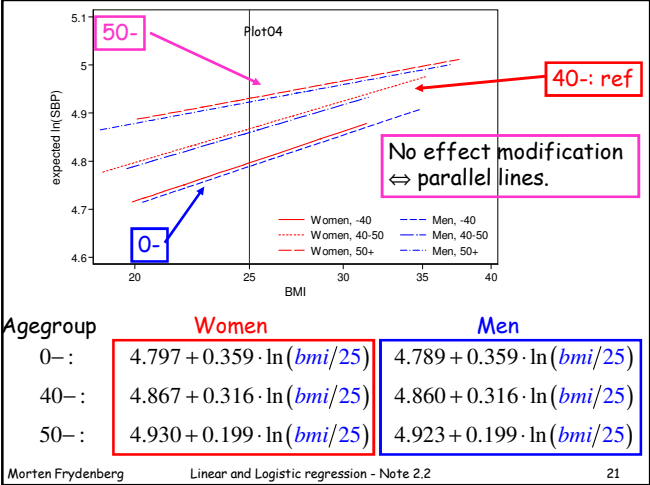
50-: difference in constant and slope compared to reference

	lnSBP	Coef.	Std. E.	t	P> t	[95% Conf. Interval]
woman		.0076438	.0219199	0.35	0.728	-.0355896 .0508772
agegrp3						
0-		-.0708045	.0261198	-2.71	0.007	-.1223213 -.0192877
50-		.0631082	.0270342	2.33	0.021	.0097877 .1164287
lnBMI25		.3155479	.1222905	2.58	0.011	.0743505 .5567453
agegrp3#lnBMI25						
0-		.0429736	.1912373	0.22	0.822	-.3342099 .420157
50-		-.1165375	.1855477	-0.63	0.531	-.4824991 .2494242
_cons		4.859743	.0214122	226.96	0.000	4.817511 4.901975

Note the larger standard errors

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

Morten Frydenberg Linear and Logistic regression - Note 2.2 20



Interactions/effect modification

	lnSBP	Coef.	Std. E.	t	P> t	[95% Conf. Interval]
woman		.0076438	.0219199	0.35	0.728	-.0355896 .0508772
agegrp3						
0-		-.0708045	.0261198	-2.71	0.007	-.1223213 -.0192877
50-		.0631082	.0270342	2.33	0.021	.0097877 .1164287
lnBMI25		.3155479	.1222905	2.58	0.011	.0743505 .5567453
agegrp3#lnBMI25						
0-		.0429736	.1912373	0.22	0.822	-.3342099 .420157
50-		-.1165375	.1855477	-0.63	0.531	-.4824991 .2494242
_cons		4.859743	.0214122	226.96	0.000	4.817511 4.901975

Two tests:

One testing differences between "slope" in group 0- and 40-.

One testing differences between "slope" in group 50- and 40-.

One test testing no difference between age groups!

A F-test in Stata: `testparm i.agegrp3#c.lnBMI25`

(1) 0.agegrp3#c.lnBMI25 = 0

(2) 2.agegrp3#c.lnBMI25 = 0

F(2, 193) = 0.33

Prob > F = 0.7168

Non-significant

Morten Frydenberg Linear and Logistic regression - Note 2.2 22

Interactions/effect modification

The test of no interaction was non-significant.

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

	lnSBP	Coef.	Std. E.	t	P> t	[95% Conf. Interval]
woman		.0076438	.0219199	0.35	0.728	-.0355896 .0508772
agegrp3						
0-		-.0708045	.0261198	-2.71	0.007	-.1223213 -.0192877
50-		.0631082	.0270342	2.33	0.021	.0097877 .1164287
lnBMI25		.3155479	.1222905	2.58	0.011	.0743505 .5567453
agegrp3#lnBMI25						
0-		.0429736	.1912373	0.22	0.822	-.3342099 .420157
50-		-.1165375	.1855477	-0.63	0.531	-.4824991 .2494242
_cons		4.859743	.0214122	226.96	0.000	4.817511 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.

Morten Frydenberg Linear and Logistic regression - Note 2.2 23

Interaction between age group and sex

Male sex==1

regress lnSBP lnBMI25 b1.agegrp3#b1.sex

	lnSBP	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
lnBMI25		.2265018	.0774898	2.92	0.004	.0736662 .3793374
agegrp3						
0-		-.0426734	.0377096	-1.13	0.259	-.1170493 .0317025
50-		-.0025412	.0365457	-0.07	0.945	-.0746215 .0695391
sex						
women		-.0210869	.0322283	-0.65	0.514	-.0846518 .042478
agegrp3#sex						
0-#women		-.0548967	.0501668	-1.09	0.275	-.1538422 .0440488
50-#women		.133379	.0501308	2.66	0.008	.0345043 .2322536
_cons		4.873442	.0251767	193.57	0.000	4.823786 4.923099

testparm i.agegrp3#i.sex

(1) 0.agegrp3#2.sex = 0

(2) 2.agegrp3#2.sex = 0

F(2, 193) = 6.26

Prob > F = 0.0023

Highly significant

Morten Frydenberg Linear and Logistic regression - Note 2.2 24

Interaction between age group and sex

Differences between age groups among men are small

	lnSBP	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
lnBMI25		.2265018	.0774898	2.92	0.004	.0736662 .3793374
agegrp3						
0-		-.0426734	.0377096	-1.13	0.259	-.1170493 .0317025
50-		-.0025412	.0365457	-0.07	0.945	-.0746215 .0695391
sex						
women		-.0210869	.0322283	-0.65	0.514	-.0846518 .042478
agegrp3#sex						
0-#women		-.0548967	.0501668	-1.09	0.275	-.1538422 .0440488
50-#women		.133379	.0501308	2.66	0.008	.0345043 .2322536
_cons		4.873442	.0251767	193.57	0.000	4.823786 4.923099

Women age group 0-: $4.873 - 0.021 = 4.852$
Women age group 0- : $4.873 - 0.021 - 0.042 - 0.055 = 4.755$
Women age group 50-: $4.873 - 0.021 - 0.003 + 0.133 = 4.982$

Large differences in the age groups among women.

Morten Frydenberg Linear and Logistic regression - Note 2.2 25

Interaction between age group and sex

Using women as reference: b2 . sex

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
agegrp3						
0-		-.0975701	.0328469	-2.97	0.003	-.162355 -.0327852
50-		.1308378	.0354804	3.69	0.000	.0608587 .2008168
sex						
Men		.0210869	.0322283	0.65	0.514	-.042478 .0846518
agegrp3#sex						
0-#Men		.0548967	.0501668	1.09	0.275	-.0440488 .1538422
50-#Men		-.133379	.0501308	-2.66	0.008	-.2322536 -.0345043
_cons		4.852355	.0205502	236.12	0.000	4.811824 4.892887

Morten Frydenberg Linear and Logistic regression - Note 2.2 26

Plot05

Agegroup Women Men

0- : $4.755 + 0.227 \cdot \ln(bmi/25)$ $4.831 + 0.227 \cdot \ln(bmi/25)$

40- : $4.852 + 0.227 \cdot \ln(bmi/25)$ $4.873 + 0.227 \cdot \ln(bmi/25)$

50- : $4.982 + 0.227 \cdot \ln(bmi/25)$ $4.871 + 0.227 \cdot \ln(bmi/25)$

Morten Frydenberg Linear and Logistic regression - Note 2.2 27