Continuous data : linear models

Jan van den Broek

Utrecht University

2020

Universiteit Utrecht

Independent variable	Type of linear model
Grouping variable f.i. one placebo group and 2 medication groups; coding 1,2 3	**Anova**
Continuous f.i dose or weight	**Regression**
Both grouping and continuous f.i one grouping variable medicine and a variable dose	**Analysis of covariance**
Extensions with grouping and continuous (interaction)	**General linear model**

The experiment

Placebo	Treatment
87	86.5
86.5	87
89	85
88.5	86
87.5	85
88	83

observations: y_{ij}, the first index stands for the group and the second for the observation number within the group. So y_{25} is the fifth observation from the second group. The group means are denoted with $\bar{y}_i$ and the overall mean with $\bar{y}$

The anova model in the population.

The population is the population we would have had if the experiment from the sample was done in the whole population. It's a theoretical construct.

The overall mean in the population is denoted by μ. The population mean of the first group is μ_1 and that of the second group is μ_2. In general the mean of the i^{th} group is μ_i. The variance in each group is σ^2. The observations are normally distributed.

The anova model in the population.

We want to know if the group means differ from each other. Let's look at 2 situations:

1. $\mu_1 \approx \mu_2$. The deviation $(\mu_i - \mu)$ is small.
2. The group means are not equal. The deviation $(\mu_i - \mu)$ is large (positive or negative).

The anova model in the population.

$(\mu_i - \mu)$ will show whether or not there are differences between the groups . They are called the **group effects**.

The anova model in the population.

Model the data according to the group they com from:

$$observation = constant + groupeffect + residual$$

$$y_{ij} = \mu + (\mu_i - \mu) + \epsilon_{ij}$$

$$\epsilon_{ij} = y_{ij} - \mu - (\mu_i - \mu) = y_{ij} - \mu_i$$

The anova model in the population.

In short:

The linear model for the anova case is

$$y_{ij} = \mu + (\mu_i - \mu) + \epsilon_{ij}$$

were the y_{ij} are normally distributed with mean μ_i and variance σ^2

The anova model in the sample

Estimate the model: Replace the population means in the model by the sample means

$$y_{ij} = \bar{y} + (\bar{y}_i - \bar{y}) + e_{ij}$$

were $e_{ij} = y_{ij} - \bar{y}_i$.
Now write this model in terms of **deviations** :

$$y_{ij} - \bar{y} = (\bar{y}_i - \bar{y}) + e_{ij}$$

$(y_{ij} - \bar{y})$ It is called the total deviation.
$(\bar{y}_i - \bar{y})$ the between groups deviations.
$e_{ij} = (y_{ij} - \bar{y}_i)$ are the residual deviations (the within group deviation)

$$\textit{Total deviation} = \textit{between groups deviation} + \textit{residual deviation}$$

The anova model in the sample

Now square the deviations and sum over all observations:

$$\sum_{\textit{all observations}} (y_{ij} - \bar{y})^2 = \sum_{\textit{all observations}} (\bar{y}_i - \bar{y})^2 + \sum_{\textit{all observations}} e_{ij}^2$$

The deviations squared and summed are called **sums of squares**.

The total deviation squared and summed is called the **total sum of squares** SS_{Total}.

The between group deviations squared and summed is called the **between groups sum of squares** or shortly the **sum of squares for the groups** SS_{Group}.

The sum of the squared residuals is called the **residual sum of squares** SS_{Res}.

The anova model in the sample

Then

$$SS_{Total} = SS_{Group} + SS_{Res}$$

The anova model in the sample

Sums of squares are based on a number of informative observations called **degrees of freedom**.

The total sums of squares has $df_{total} = n - 1$ degrees of freedom.

Divide the total sum of squares by this degrees of freedom to obtain the variance of the 12 observations.

The group sum of squares only contain the group means so the degrees of freedom is $df_{Group} = number\ of\ groups - 1$.

The degrees of freedom for the residual sum of squares is what is left: $df_{res} = df_{Total} - df_{group}$.

The anova model in the sample

Divide the sums of squares by there degrees of freedom to get the variances also called **mean sum of squares (MS)**: MS_{Total}, MS_{Group} and MS_{res}. (These don't add up!)

The anova model in the sample

No differences between the groups so $\mu_1 = \mu_2$ The variance within the groups is the same as the variance between the groups or $F = \frac{MS_{Group}}{MS_{res}} \approx 1$.

If the group means are unequal then the variance between the groups will be much larger than the variance within the groups.

$H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ one can use the result that F has a so called Fisher distribution with df_{group} and df_{res} degrees of freedom. One can use this to calculate p-values.

The anova model in the sample

The general lines of the analysis are

1. Write down the model in the population: $y_{ij} = \mu + (\mu_i - \mu) + \epsilon_{ij}$
2. Estimate the model in the sample: $y_{ij} = \bar{y} + (\bar{y}_i - \bar{y}) + e_{ij}$
3. Write the model in terms of deviations: $y_{ij} - \bar{y} = (\bar{y}_i - \bar{y}) + e_{ij}$
4. Square the deviations and sum over all observations to get the sum of squares: $SS_{Total} = SS_{Group} + SS_{Res}$
5. Divide the sum of squares by there degrees of freedom to get MS_{Total}, MS_{Group} and MS_{res}. See howmany times larger the between group variance is as compared to the residual variance (F).

The anova model in the sample

Put everything in a table (anova-table)

Name	SS	df	MS	F
groups	SS_{group}	df_{group}	MS_{group}	$\frac{MS_{Group}}{MS_{res}}$
Residual	SS_{res}	df_{res}	MS_{res}	
Total	SS_{total}	df_{Total}		

In the blood pressure example the anova table is:

Name	SS	df	MS	F
Treatment	16.331	1	16.331	11.2
Residual	14.583	10	1.458	
Total	30.914	11		

The variance between groups is about 11 times larger as the residual. This can only be if there is a systematic difference between the groups. (The p-value is 0.0074)

Model with two factors

Suppose the first 3 observations were from individuals who got dose 1 and the last 3 are from individuals who got dose 2. The same goes for the treatment group. In that case we have a two factor experiment, one factor is treatment with 2 levels, placebo and treatment and the other factor is dose with two levels dose 1 and dose 2.

	Placebo	Treatment
dose 1	87	86.5
	86.5	87
	89	85
dose 2	88.5	86
	87.5	85
	88	83

Model with two factors

An observation can be represented by y_{ijk}. This is the k^{th} observation from treatment i and dose j.
Treatment groups means are μ_i
Dose groups group they are μ_j.
The treatment effects are then $\mu_i - \mu$
dose effects are $\mu_j - \mu$.

The analysis causes now extra problems since it goes along the same lines as with one factor:

Model with two factors

1. Write down the model in the population:
   $y_{ijk} = \mu + (\mu_i - \mu) + (\mu_j - \mu) + \epsilon_{ijk}$.
2. Estimate the model in the sample:
   $y_{ijk} = \bar{y} + (\bar{y}_i - \bar{y}) + (\bar{y}_j - \bar{y}) + \epsilon_{ijk}$
3. Write the model in terms of deviations:
   $y_{ijk} - \bar{y} = (\bar{y}_i - \bar{y}) + (\bar{y}_j - \bar{y}) + e_{ijk}$
4. Square the deviations and sum over all observations to get the sum of squares: $SS_{Total} = SS_{treatment} + SS_{dose} + SS_{Res}$
5. Divide the sum of squares by there degrees of freedom to get MS_{Total}, $MS_{treatment}$, MS_{dose} and MS_{res}.

Model with two factors

Put everything in a table (anova-table)

Name	SS	df	MS	F
treatment	$SS_{treatment}$	$df_{treatment}$	$MS_{treatment}$	$\frac{MS_{treatment}}{MS_{res}}$
dose	SS_{dose}	df_{dose}	MS_{dose}	$\frac{MS_{dose}}{MS_{res}}$
Residual	SS_{res}	df_{res}	MS_{res}	
Total	SS_{total}	df_{Total}		

Model with two factors

To test whether or not there are differences between then dose groups use the result that $F = \frac{MS_{dose}}{MS_{res}}$ has a Fisher distribution with df_{dose} and df_{res} degrees of freedom.

To test whether or not there are differences between then treatment groups use the result that $F = \frac{MS_{treatment}}{MS_{res}}$ has a Fisher distribution with $df_{treatment}$ and df_{res} degrees of freedom.

Regression analysis. The experiment

Dose	Blood pressure
5	87;86.5
6	86.5;87
7	89;85
8	88.5;86
9	87.5;85
10	88;83

We can make a scatter plot of this. The dose is on the x-axis and the blood pressure is on the y-axis.

The regression model in the population.

The population is again the group of persons from which the sample was taken. But also here: it's a theoretical population.

The general form of a straight line is

$$\alpha + \beta x$$

y is the **dependent variable** and x is the **independent one**.
α s the **intercept** of the line.
β is the **regression coefficient**.

The regression model in the population.

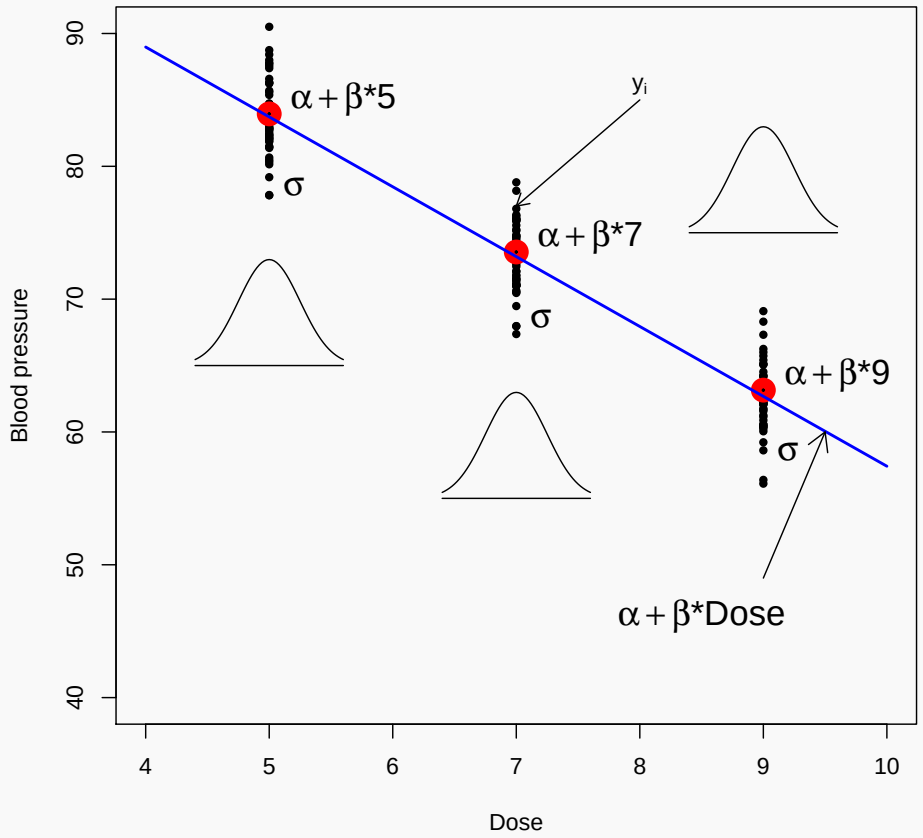

Figure: A population model for the relation between dose and blood pressure.

The regression model in the sample

The sample values for α and β are a, the sample intercept, and b the sample slope. It are those values that give the **best fitted line** in the sample
A **residual** is given by $y_i - (a + bx_i)$.

This residual squared and summed over all data points is called the **residual sum of squares** SS_{res}.

$$SS_{res} = \sum_i [y_i - (a + bx_i)]^2$$

best fitted line has **smallest residuals** thus has smallest residuals sum of squares.

The regression model in the sample

a and b can be calculated as :

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

If a and b are calculated as above then the residual sum of squares is smallest.

That is why these estimates are called **least square estimates**.

The estimated regression model now is:

$$y_i = a + bx_i + e_i$$

The regression model in the sample

Now, plug in $a = \bar{y} - b\bar{x}$ and write in terms of deviations:

$$y_i = \bar{y} - b\bar{x} + bx_i + e_i = \bar{y} + b(x_i - \bar{x}) + e_i$$

so

$$y_i - \bar{y} = b(x_i - \bar{x}) + e_i$$

which shows that:

$$\textit{Total deviation} = \textit{regression deviation} + \textit{residual deviation}$$

square the deviations and sum to get:

$$\sum_i (y_i - \bar{y})^2 = b^2 \sum_i (x_i - \bar{x})^2 + \sum_i e_i^2$$

so

$$SS_{Total} = SS_{regres} + SS_{Res}$$

The regression model in the sample

The **degrees of freedom** for the **total sums of squares** is $n - 1$

The **degrees of freedom** for the **regression sum of squares** is **1**

The **degrees of freedom** for the the **residual sum of squares** is **what is left thus** $n - 2$.

Divide the sum of squares by the degrees of freedom to get the variance or the mean sum of squares: MS_{total}, MS_{regres} and MS_{res}.

Then put everything in a table, the anova table.

The regression model in the sample

The $F - value$ can be calculated as $F = \frac{MS_{regres}}{MS_{res}}$

Can be used to test the hypothesis $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$.

If this value is much larger then 1 this means that the data is best described with the regression line. If the F-value is approximately 1 then the data does not show a linear relationship and the data can just as well be summarized by calculating the mean.

To calculate a p-value use the result that F has a Fisher distribution with 1 and $n - 2$ degrees of freedom.

The regression model in the sample

The general lines of the analysis are the same as in the anova case:

1. Write down the model in the population: $y_i = \alpha + \beta x_i + \epsilon_i$
2. Estimate the model in the sample: $y_i = a + b x_i + e_i$
3. Write the model in terms of deviations: $y_i - \bar{y} = b(x_i - \bar{x}) + e_i$
4. Square the deviations and sum over all observations to get the sum of squares: $\sum_i (y_i - \bar{y})^2 = b^2 \sum_i (x_i - \bar{x})^2 + \sum_i e_i^2$ so $SS_{Total} = SS_{regres} + SS_{Res}$
5. Divide the sum of squares by there degrees of freedom to get MS_{Total}, MS_{regres} and MS_{res}. See how many times larger the regression variance is as compared to the residual variance (F).

The regression model in the sample

Put everything in a table (anova-table)

Name	SS	df	MS	F
Regression	SS_{regres}	df_{regres}	MS_{regres}	$\frac{MS_{regres}}{MS_{res}}$
Residual	SS_{res}	df_{res}	MS_{res}	
Total	SS_{total}	df_{Total}		

The regression model in the sample

In the blood pressure example the anova table is:

Name	SS	df	MS	F
regression	1.6	1	1.6	0.58
Residual	29.3	10	2.9	
Total	30.9	11		

So there is no evidence that the dose is linear related to the blood pressure.

Analysis of covariance. The experiment

Placebo dose	blood pressure	Treatment dose	blood pressure
5	87	5	86.5
6	86.5	6	87
7	89	7	85
8	88.5	8	86
9	87.5	9	85
10	88	10	83

Think carefully about what to take as a covariate. The covariate may not be influenced by the treatment.

The ancova model in the population.

The population is again the group of persons from which the sample was taken. But also here: it's a theoretical one.

The model is

$$y_{ij} = \alpha_i + \beta x_{ij} + \epsilon_{ij} = \alpha + (\alpha_i - \alpha) + \beta x_{ij} + \epsilon_{ij}$$

In group 1 this becomes: $y_{1j} = \alpha_1 + \beta x_{1j} + \epsilon_{1j}$

In group 2 this is: $y_{2j} = \alpha_2 + \beta x_{2j} + \epsilon_{2j}$.

*The **linear model for the ancova** case is* $y_{ij} = \alpha_i + \beta x_{ij} + \epsilon_{ij}$,

y_i *normally distributed with **mean*** $\alpha_i + \beta x_{ij}$ *and **variance*** σ^2

Different representation of the model

Consider the situation were there are only two groups and no covariates. Use a column with zero's and one's. So this column is a group 2 (treatment) **indicator**.

Call this column group and write the anova model as:

$$y_{ij} = \alpha_0 + \alpha_1 group_{ij} + \epsilon_{ij}$$

For the observations of group 1 this is $y_{1j} = \alpha_0 + \epsilon_{1j}$ so α_0 is the mean of group 1.

For group 2 the model is $y_{2j} = \alpha_0 + \alpha_1 + \epsilon_{2j}$. The mean of the model for group 2 minus the mean of the model for group 1 is $\alpha_0 + \alpha_1 - \alpha_0 = \alpha_1$. So α_1 is the difference in group means.

Different representation of the model

If there are f.i. 4 groups we need 3 group indicator variables: one indicating group 2, one indicating group 3 and one indicating group 4.

Group	gr2	gr3	gr4
1	0	0	0
2	1	0	0
3	0	1	0
4	0	0	1

Different representation of the model

The model then is

$$y_{ij} = \alpha_0 + \alpha_1 gr2_{ij} + \alpha_2 gr3_{ij} + \alpha_3 gr4_{ij} + \epsilon_{ij}$$

For group 1 the mean is α_0, for group 2 that is $\alpha_0 + \alpha_1$ so α_1 is the difference in means between group 1 and 2. For group 3 the mean is $\alpha_0 + \alpha_2$ so α_2 is the difference in group means between group 3 and 1. With 4 groups you need 3 such group indicating variables. Each variable has a 1 indicating its own group. The group that is not indicated (group 1 here) is the one with which the comparison is made.

Different representation of the model

The ancova model can now be written as :

$$y_{ij} = \alpha_0 + \alpha_1 group_{ij} + \beta x_{ij} + \epsilon_{ij}$$

For group 1 this is $y_{ij} = \alpha_0 + \beta x_{ij} + \epsilon_{ij}$, α_0 is the intercept in group 1.

For group 2 the model is $y_{ij} = \alpha_0 + \alpha_1 + \beta x_{ij} + \epsilon_{ij}$. So $\alpha_0 + \alpha_1$ is the intercept in group 1, thus α_1 is the difference in intercept.

The sample

Estimate the model as:

$$y_{ij} = a_0 + a_1 group_{ij} + bx_{ij} + e_{ij}$$

The a_0, a_1 and b are estimates of α_0, α_1, and β such that the residual sums of squares is smallest.

There are now no easy formulas

The sample

Name	SS	df	MS	F
group	SS_{group}	df_{group}	MS_{group}	$\frac{MS_{group}}{MS_{res}}$
X	SS_{regres}	df_{regres}	MS_{regres}	$\frac{MS_{regres}}{MS_{res}}$
Residual	SS_{res}	df_{res}	MS_{res}	
Total	SS_{total}	df_{Total}		

Here is $df_{group} = number\ of\ groups - 1$ and $df_{res} = n - 1 - 1 - df_{group}$

The sample

To test the hypothesis: $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$ use the result that $F = \frac{MS_{regres}}{MS_{res}}$ has a Fisher distribution with 1 and df_{res} degrees of freedom.

To test the hypothesis:$H_0 : \alpha_1 = 0$ versus $H_1 : \alpha_1 \neq 0$ use the result that $F = \frac{MS_{group}}{MS_{res}}$ has a Fisher distribution with df_{group} and df_{res} degrees of freedom. If this last null hypothesis can not be rejected, one might just as well take the model with $\alpha_1 = 0$. One then gets the ordinary regression model.

So these hypothesis test **whether or not the intercepts in both groups are the same**.

The general linear model in the population

In this previous model the 2 groups have the same regression coefficient. They differ only in intercepts. This means that the lines are parallel. This is not always justified. Let's look at the example:

The general linear model in the population

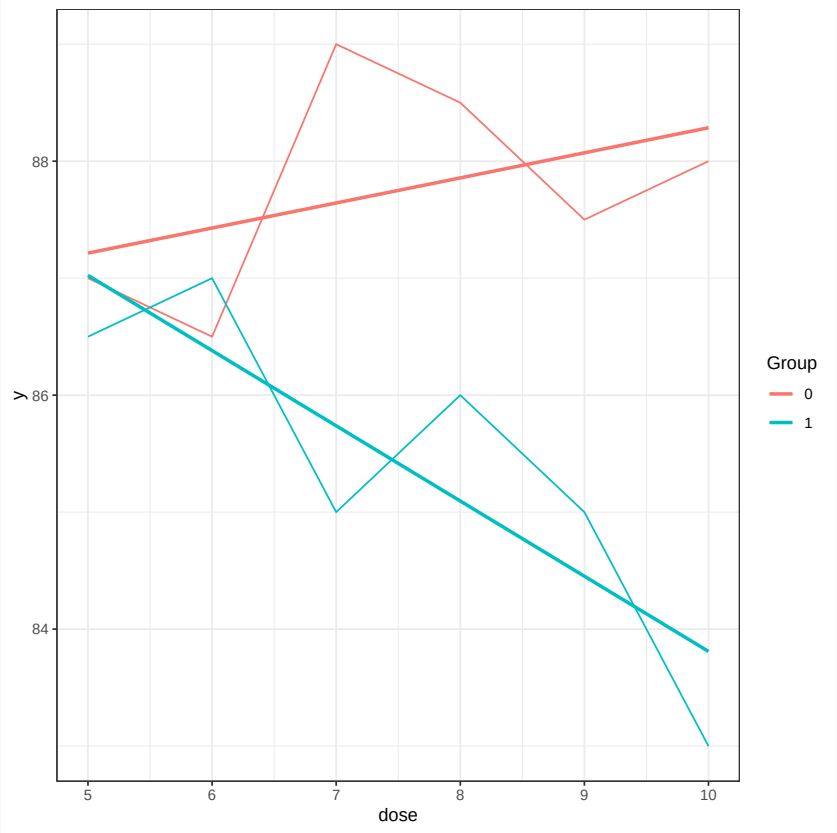

Figure: The relation between dose and blood pressure for the two groups.

The general linear model in the population

In the placebo group there is not much happening what ever the dose is. In the treatment group the blood pressure is going down if the dose is increased. So the effect of the dose depends on the group. This is called an **interaction effect**.

Interaction effect: **The effect of one independent variable on the dependent variable, depends on the outcome of an other independent variable**.

The general linear model in the population

An interaction effect is denoted by a product: a group times dose effect.

This is because one can model the interaction effect by multiplying the two independent variables and put this product in the model:

$$y_{ij} = \alpha_0 + \alpha_1 group_{ij} + \beta_1 x_{ij} + \beta_2 group_{ij} \times x_{ij} + \epsilon_{ij}$$

An extra slope only when group equals 1.

The general linear model in the population

In group 1 this model is:$y_{ij} = \alpha_0 + \beta_1 x_{ij} + \epsilon_{ij}$

The intercept here is α_0 and the regression coefficient is β_1.

In group 2 the model is
$y_{ij} = \alpha_0 + \alpha_1 + \beta_1 x_{ij} + \beta_2 x_{ij} + \epsilon_{ij} = (\alpha_0 + \alpha_1) + (\beta_1 + \beta_2) x_{ij} + \epsilon_{ij}$

The intercept here is $(\alpha_0 + \alpha_1)$ and the regresion coefficient is $(\beta_1 + \beta_2)$ So this model has 2 different intercepts and 2 different regression coefficients.

The general linear model in the population

Name	SS	df	MS	F
group	SS_{group}	df_{group}	MS_{group}	$\frac{MS_{group}}{MS_{res}}$
x	SS_{regres}	df_{regres}	MS_{regres}	$\frac{MS_{regres}}{MS_{res}}$
group*x	$SS_{interaction}$	$df_{interaction}$	$MS_{interaction}$	$\frac{MS_{interaction}}{MS_{res}}$
Residual	SS_{res}	df_{res}	MS_{res}	
Total	SS_{total}	df_{Total}		

The degrees of freedom for the interaction effect : multiply the degrees of freedom of the independent variable that make the inter action.

To test wether or not the lines are parallel: $H_0 : \beta_2 = 0$ vs $H_1 : \beta_2 \neq 0$. One can use the result that $F = \frac{MS_{interaction}}{MS_{res}}$ has a Fisher distribution with $df_{interaction}$ and df_{res} degrees of freedom.

The general linear model in the population

To estimate the model in the sample one has to estimate what the parameters are and the are called a_0, a_1, b_1, b_2 and then the model becomes

$$y_{ij} = a_0 + a_1 group_{ij} + b_1 x_{ij} + b_2 group_{ij} \times x_{ij} + \epsilon_{ij}$$

The r-output for this model is:

The general linear model in the population

```
glm(formula = y ~ factor(group) + dose +
factor(group):dose)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9286  -0.6131  -0.2500   0.6250   1.3571

Coefficients:
                  Estimate Std.Error t value Pr(>|t|)
(Intercept)         86.1429  1.6635  51.785 2.14e-11 ***
factor(group)1       4.0952  2.3525   1.741   0.1199
dose                 0.2143  0.2163   0.991   0.3508
factor(group)1:dose -0.8571  0.3058  -2.803   0.0231 *
---
```

So $a_0 = 86.1429, a_1 = 4.0952, b_1 = 0.2143, b_2 = -0.8571$
Model group 1: $y_{1j} = 86.1429 + 0.2143\ dose_1 j$
Model group 2: $y_{2j} = (86.1429 + 4.0952) + (0.2143 - 0.8571)\ dose_{2j} = 90.2381 - 0.6429\ dose_{2j}$

The general linear model in the population

nested version of the model: In order to see the dose effect (the slope of the dose) within group 0 and within group 0, one can fit the nested version of this model.
If, in general one wants the effect of B only within the A-groups one uses the nested version: A+A:B. That is done below with group and dose:

```
glm(formula = y ~ factor(group) + factor(group):dose)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9286  -0.6131  -0.2500   0.6250   1.3571

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          86.1429     1.6635  51.785 2.14e-11 ***
factor(group)1        4.0952     2.3525   1.741   0.1199
factor(group)0:dose   0.2143     0.2163   0.991   0.3508
factor(group)1:dose  -0.6429     0.2163  -2.973   0.0178 *
---
```