Continuous Data: Linear Models

Jan van den Broek

2024

Suppose we want to measure something that is continuous. For instance a sample of 12 persons from whom we measure the blood pressure. This measure is called the dependent variable. If we were able to measure the blood pressure from every person from the population from which we took the sample we could make a histogram of all these blood pressures. If this histogram looks like a normal distribution we say that the sample of blood pressures come from a normal distribution. Often we want to relate this continuous dependent variable to another one, the so called independent variable. Linear models can be divided in 4 groups according to the nature of this independent variable:

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

1 Anova

1.1 The experiment

Suppose we have a sample of 12 individuals with high blood pressure and divide this in two groups of 6 persons each. One group gets a placebo and the other gets a treatment.

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

The question of this research was: does the treatment work. Is there a difference between the groups in mean blood pressure?

The observations are denoted by 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 in the sample are represented by $\bar{y}_i$, so for $i = 1$ the mean of the first group is $\bar{y}_1 = 87.67$ and tge mean for the second group ($i = 2$) is $\bar{y}_1 = 85.42$. The overall mean is $\bar{y}$ without an index attached to it.

1.2 The population: a model for the data generating process

We have a sample from some population. If we come up with a description of this population we can understand where the data comes from. This description is called a model. So we first give a model for the population that gave rise to the data. To put it differently: We model the data generating process. Of course this model is only an approximation of that, often complicated, process.

1. The population is the group of individuals from which we draw the sample of 12 persons. But that's not all. It is the population we would have had if the experiment from the sample was done in the whole population. So we would then divide all people in the population randomly in two groups, one group gets the placebo and the other one gets the treatment. Then we measure the blood pressure.

2. 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. If we make a histogram of the data we would see something close to a normal distribution. This is a description of the population that gave rise to the data we have in the sample. This population does not exist. It's a theoretical construct.

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

(a) The group means are approximately the same : $\mu_1 \approx \mu_2$. The two group means are close to each other so they are close to the overall mean. The difference between the group means and the overall mean is small so the deviation $(\mu_i - \mu)$ is small.
In this case the best summary of the data is the general mean μ. Some of the observations will be above μ and some will be below μ. How much an observation is above or below the mean is called a residual denoted with ϵ_{ij}. Now a model for the observations in this case is : observation = best summary + residual or $y_{ij} = \mu + \epsilon_{ij}$.

(b) The group means are not equal so they will be far apart. So they don't look like the overall mean. The deviation $(\mu_i - \mu)$ is large (positive or negative).
In this case the best summary of the data are the group means μ_i. Also here the observation will be above or below its own group mean. How much this is, is called the residual. Thus a model in this case can be $y_{ij} = \mu_i + \epsilon_{ij}$.

3. These two models can be written in one line, using that the deviations $(\mu_i - \mu)$, which are called group effects, will show whether or not there are differences between the groups:

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

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

In short:

The linear model for the anova case is

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

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

1.3 The sample

In the sample we have to estimate the model. This can be simply done by replacing the population means in the model by the sample means

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

where $e_{ij} = y_{ij} - \bar{y}_i$. Now write this model in terms of deviations by bringing $\bar{y}$ to the left of the = sign:

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

$(y_{ij} - \bar{y})$ is the deviation between an observation and the overall mean. It is called the total deviation.

$(\bar{y}_i - \bar{y})$ is the deviation between the group mean and the overall mean. These are small if there are no differences between the groups and large if there are. These deviation show whether or not there are differences between the groups so they are called the between groups deviations.

$e_{ij} = (y_{ij} - \bar{y}_i)$ are the residual deviations, in this case also called the within group deviations since they look at the deviation of an observation with respect to its own group mean. So

$$Total\ deviation = between\ groups\ deviation + residual\ deviation$$

Now square the deviations and sum over all observations:

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

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}. Similarly the between group deviations squared and summed is called the between groups sum of squares or the sum of squares for the groups SS_{Group}. The sum of the squared residuals is called the residual sum of squares SS_{Res}. Then

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

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: $df_{res} = df_{Total} - df_{group}$. Divide the sums of squares by their 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!)

Suppose there are no differences between the groups so $\mu_1 = \mu_2$. The observations are not all the same, they differ. The data generating mechanism is such that the observation within the groups are different. If there are no systematic differences between the groups (population means are the same)

then the data generating mechanism that causes differences between the observations within the group also causes the differences between the groups (between the group means). 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.

To test the hypothesis $H_0 : \mu_1 = \mu_2$ 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. One can put everything in a table, the so called anova table.

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 their degrees of freedom to get MS_{Total}, MS_{Group} and MS_{res}. See how many times larger the between group variance is as compared to the residual variance (F).
6. 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 than the residual. This can only be if there is a systematic difference between the groups. (The p-value is 0.0074)

If there are more than 2 groups, the analysis does not change so points 1-6 stay the same except that the degrees of freedom for the groups are now different according to: $df_{Group} = number\ of\ groups - 1$

1.4 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

An observation can be represented by y_{ijk}. This is the k^{th}observations from treatment i and dose j. The means in the population for the treatment groups are μ_i and for the dose groups they are μ_j. The treatment effects are then $\mu_i - \mu$ and the dose effects are $\mu_j - \mu$. The analysis causes no extra problems since it goes along the same lines as in the one factor case:

1. Write down the model in the population: $y_{ijk} = \mu + (\mu_i - \mu) + (\mu_j - \mu) + \epsilon_{ijk}$. The y_{ijk} are normally distributed with mean $\mu + (\mu_i - \mu) + (\mu_j - \mu)$ and with variance σ^2

2. Estimate the model in the sample: $y_{ijk} = \bar{y} - (\bar{y}_i - \bar{y}) + (\bar{y}_j - \bar{y}) + e_{ij}$

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 their degrees of freedom to get MS_{Total}, $MS_{treatment}$, MS_{dose} and MS_{res}. The degrees of freedom for the treatment sum of squares is number of treatment groups minus 1, for the dose sum of squares this is the number of dose groups minus 1. See how many times larger the between group variance is as compared to the residual variance (F).

6. 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}		

To test whether or not there are differences between the 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.

This analysis can be extended to the case where there are more than two factors. The group effects for all factors are added to the model and to the anova table.

2 Regression analysis

2.1 The experiment

Suppose we have a sample of 12 individuals with high blood pressure. They are all given a medicine but in different doses. The independent variable is continuous. After a while the blood pressure is measured. This is called a dose-response study. Every dose is given to 2 persons:

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.

2.2 The population: a model for the data generating process

1. The population is again the group of persons from which the sample was taken. But also here: it's a theoretical population. It does not exist for real. It is the population we would have had if the experiment from the sample was done in the whole population. So if all people in the population were given a medicine in a certain dose. (se figure 1)

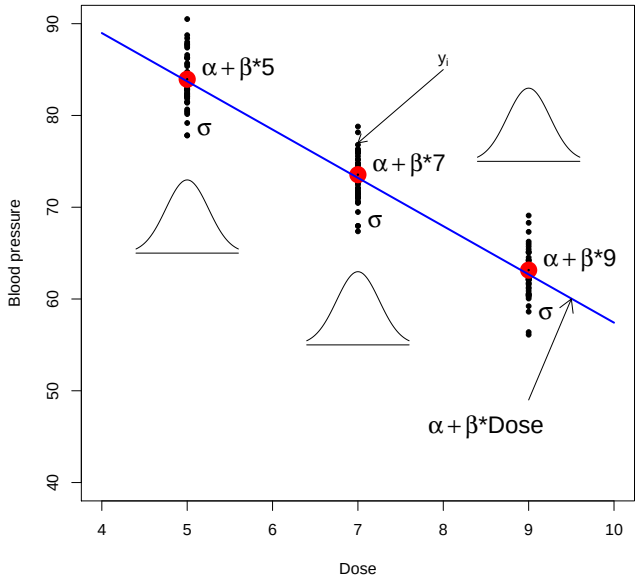

Figure 1: The regression model in the population

2. The general form of a straight line is

$$\alpha + \beta x$$

y is the dependent variable and x is the independent one. α is the intercept of the line, it's the point were the line cuts the y-axis. β is the regression coefficient. It is the amount of change in the dependent variable if the independent variable changes by one.

Let's look at the blood pressure of all the individuals in the population that had dose x_0. We can calculate the mean of this blood pressures. This mean is a point on the straight line: the mean blood pressure of all individuals in the population who had dose x_0 is thus $\alpha + \beta x_0$. The variance of these blood pressures is σ^2, so σ^2 is the variance of the points around the line. A histogram of all the blood pressures of all people with dose x_0 in the population would very much look like a normal distribution. The same is true for those people who had dose x_1: the mean of their blood pressures is $\alpha + \beta x_1$, the variance is σ^2 and these blood pressures follow a normal distribution. Of course this holds for all doses.

3. So a model for the population is obtained by describing an observation

as a point on the line plus a deviation from the line, the residual: $y_i = \alpha + \beta x_i + \epsilon_i$

In short:

> The linear model for the regression case is
>
> $$y_i = \alpha + \beta x_i + \epsilon_i$$
>
> where the y_i are normally distributed with mean $\alpha + \beta x_i$ and variance σ^2

2.3 The sample

In the sample we have to estimate the model. The line is estimated in such a way that the obtained line fits the observed points best. This line is called

$$y_i = a + bx_i + e_i.$$

a is the intercept in the sample and b is the sample slope. That line fits the data best for which the vertical distances from the points to the line is smallest. This distance is the residual. So that line fits the data best that has the smallest residuals. 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}.

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

If the residuals should be smallest to get the best fitted line then this also holds for the residuals sum of squares. So, that line fits the data best that has the smallest residual sum of squares. Those values for a and b, for which SS_{res} is smallest, 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 in the sample is smallest. That is why these estimates are called least square estimates.
The estimated regression model now is:

$$y_i = a + bx_i + e_i$$

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:

$$Total\ deviation = regression\ deviation + residual\ deviation$$

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 degrees of freedom for the total sums of squares is $n-1$, for the regression sum of squares 1 and for the residual sum of squares what is left, so $n-2$. 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 F-value can be calculated as $F = \frac{MS_{regres}}{MS_{res}}$ to test the hypothesis $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$. If this value is much larger than 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 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: $y_i = a + bx_i + e_i$
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 their 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).

6. 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}		

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 linearly related to the blood pressure.

3 Analysis of covariance

3.1 The experiment

A sample of 12 people with high blood pressure was randomly divided in two groups. One group was given a placebo, the other a treatment. Every one of the six persons within a group was given a different dose. After a while the blood pressure was measured. Blood pressure is the dependent variable. The independent variables are the grouping variable treatment and the continuous variable dose. So this is a combination of analysis of variance and regression analysis. The continuous independent variable is called a covariate and the analysis of this kind of data is called an analysis of covariance (ancova). The data is:

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

It is important to think carefully about what to take as a covariate in the analysis. The covariate may not be influenced by the treatment. As an example suppose we want to test a diet. We get 2 groups of people, one

group gets a placebo diet and the other one the real diet. We are interested in the effect of the diet on a certain blood parameter. It was decided to take weight of the individual as a covariate in the analysis since it was known that the blood parameter was related to weight. The research was conducted and the analysis done with the result that there was no effect of the diet on the blood parameter. This was found very surprising. What happened? The people from the placebo group had higher weight and higher values for the blood parameter. The people from the treatment group had lower weight and a lower value for the blood parameter because it was a good diet. Now using weight in the model and correcting for this removes the treatment effect because of the differences in weight between the 2 groups.

3.2 The population: a model for the data generating process

1. The population is again the group of persons from which the sample was taken. But also here: it's a theoretical one. It is the population we would have had if the experiment from the sample was done in the whole population. So if all individuals in the population were divided in 2 groups and were given given a medicine or a placebo in a certain dose.

2. The population consists of 2 groups each with a regression line. At a value x_0 the mean is given by a point on the line for that group. The variance at this value for the dose in the population is σ^2 for both groups. The data is again normally distributed. This holds for all values of the dose. The model is

$$y_{ij} = \alpha_i + \beta x_{ij} + \epsilon_{ij}$$

or

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

In group 1 this becomes $y_{1j} = \alpha_1 + \beta x_{1j} + \epsilon_{1j}$ and in group 2 this is: $y_{2j} = \alpha_2 + \beta x_{2j} + \epsilon_{2j}$. So the models have a different intercept but the same regression coefficient.

In short:

> The linear model for the ancova case is
>
> $$y_{ij} = \alpha_i + \beta x_{ij} + \epsilon_{ij}$$
>
> where the y_i are normally distributed with mean $\alpha_i + \beta x_{ij}$ and variance σ^2

3. It is sometimes convenient to give a different representation of the model. Consider the situation where there are only two groups and no covariates. In the data set one must have a column that indicates the group from which the individual came. This usually is a number, for instance the placebo group gets a 1 and the treatment group a 2. Suppose we don't use a 1 and a 2 but we make a column which contains a 0 if the observation came from the placebo group and a 1 if it came from the treatment group. So this column is a group 2 (treatment) indicator. It indicates whether or not the observation came from group 2, whether or not the treatment was applied. 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. If there are more groups we need more group indicator variables. If there are e.g. 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

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.

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 intercepts.

3.3 The sample

In the sample we can 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 for these estimates nor for the sums of squares. Nevertheless, in general the same method is followed as with anova or regression. One obtains an anova table similar as with anova and regression:

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

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 nulhypothesis cannot be rejected, one might just as well take the model with $\alpha_1 = 0$. One then gets the ordinary regression model. So this hypothesis tests whether or not the intercepts in both groups are the same.

4 General Linear Model

In model discussed above, the 2 groups have the same regression coefficient since there is only one regression coefficient in the model. The groups differ only in intercepts. This means that the lines are parallel. This is not always justified. Let's look at the example:

In the placebo group there is not much happening whatever 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.

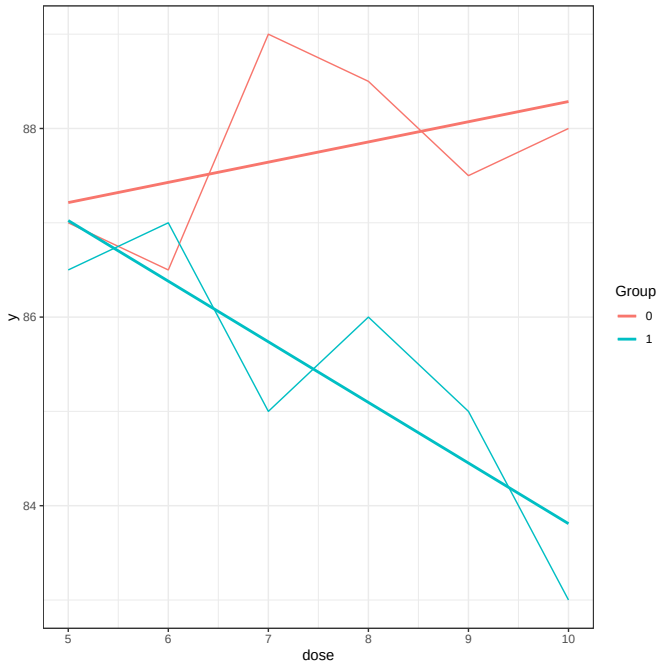

Figure 2: The relation between dose and blood pressure for the two groups.

There can also be an interaction effect between 2 grouping variables. The interpretation is the same: the effect of one grouping variable depends on the levels of the other. Or differently : the differences between the levels of on grouping variable depend on the levels of the other.

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 putting this product in the model:

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

The $\beta_2 group_{ij} \times x_{ij}$ equals $\beta_2 x_{ij}$ only in group 2, so its an extra regression effect only in group 2.
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

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

or

$$y_{ij} = (\alpha_0 + \alpha_1) + (\beta_1 + \beta_2) x_{ij} + \epsilon_{ij}$$

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

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

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

The anova table becomes:

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 are also found by a product: multiply the degrees of freedom of the independent variables that form the interaction. So in the example multiply the degrees of freedom of the group, which is 1, with the degrees of freedom for the dose, which is also one.
To test whether 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 r-output for this model is:

```
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: $y_{2j} = (86.1429+4.0952)+(0.2143-0.8571)dose_{2j} = 90.2381-0.6429\ dose_{2j}$

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 *
---
```

Exercises

R commands
First read in the data:

```
D <- data.frame(y=c(87,86.5,89,88.5,87.5,88,86.5,87,85,86,85,83),
                dose=c(5,6,7,8,9,10,5,6,7,8,9,10),
                group=c(0,0,0,0,0,0,1,1,1,1,1,1))
```

If there is an exposure value in the data file that is a "grouping" variable, then you should tell this to R by using the function: factor(). So factor(group) tells R that group is not a numeric variable but that its numbers should be used as group labels and the computer then makes the appropriate indicator variables.
To fit an anova model use:

```
model.an <- glm(y ~factor(group),family=gaussian,data=D).
```

The family=gaussian part may be left out since the gaussian (normal) is the default. The results wil be stored in an object called model. It will be an object of glm-type because you used glm to create it. The name "model.an" is completely arbitrary. To see what is in it use names(model.an) and if you want to see something specific use e.g. model.an$coefficients or coef(model.an). To get the table with the estimates : summary(model.an).
The function drop1(fit,test="F") in general looks at the terms in the fit, then leaves the terms out one by one and calculates the F-test for every term left out. Of course in model.an there is only one variable so you just get one test.
To fit the ancova model without the interaction:

```
model.anc <- glm(y ~ factor(group)+dose,
          family=gaussian,data=D)
```

The table with the estimates : summary(model.anc)
drop1(model.anc,test="F") gives

```
Single term deletions
Model:
y ~ group + dose
       Df Deviance    AIC F value    Pr(F)
<none>      12.976 42.993
group   1   29.310 50.771 11.3284 0.008313 **
dose    1   14.583 42.394  1.1147 0.318583
--
```

In the column "Deviance" are the residual sums of squares for different models. The first line gives the residual sums of squares if none of the terms is left out so for the model with group and dose in it. The second line gives the residual sum of squares for the model without group so for the model with only dose in it. The difference in these residual sum of squares gives the sum of squares for the group: $29.310 - 12.976 = 16.334$. In the same way the sum of squares for dose can be obtained. For dose and group the F-values and the p-values are shown. With this information an anova table can be constructed.
To get the plot from the text (Figure 1) using the ggplot2 library (https://ggplot2.tidyverse.org/):

```
library(ggplot2)
ggplot(D,aes(x=dose,y=y,color=factor(grp)))+
  geom_line()+
  geom_smooth(method = lm,se=FALSE)+
  labs(color = "Group")+
  theme_bw()
```

R uses ":" to denote an interaction so group:dose is the R interaction term in a model. The model with an interaction can be fitted as:

```
model.anc <- glm(y ~ factor(group)+dose+
     factor(group):dose,family=gaussian,data=D)
```

or the exact same model can be given by:

```
model.anc <- glm(y ~ factor(group)*dose,
          family=gaussian,data=D)
```

So R uses "*" to denote main effects + interaction(s). To make a Normal probability plot:

```
D <- data.frame(D,res=residuals(model.anc))
ggplot(D,aes(sample=res))+
  stat_qq()+
  stat_qq_line(color="red",size=1.1)+
  labs(y="Residuals",
       title="Normal Probability plot")
theme_bw()
```

Exercises

1. To study the reproductive cycle of a specific species of starfish, individuals from two different locations were observed. To check whether the populations of starfish at the two places differ in the mean metabole two random samples were compared:

Location A	Location B
173	185
162	164
176	177
181	175
164	172
169	168
170	

   (a) Read in the data and make a dataframe as in the above example. Make a boxplot of the data.

   `ggplot(D,aes(x=factor(group),y=y))+geom_boxplot()).`

   (b) Fit the anova model with R.

   (c) Use the output from R to make the anova table.

   (d) Test whether or not there are differences between the groups. Which hypothesis are you testing?

2. To study the effect of a hormone treatment on the blood calcium concentration of female and male birds an experiment was conducted resulting in the following data:

No hormone treatment Female	Male	hormone treatment Female	Male
17.0	16.5	18.6	17.1
18.9	14.3	16.2	14.7
13.2	10.9	12.5	15.3
14.6	15.6	15.1	14.2
13.3	8.9	16.2	12.8

(a) Make a boxplot of the data.

(b) Fit the anova model using R.

(c) Make the anova table.

(d) Test whether or not there is a difference between females and males. Also test if there is a difference between hormone treatment or not.

(e) Give the estimate of the difference in means between the no hormone and the hormone group.

3. Many wildlife populations are monitored by taking aerial photographs. Information about the number of animals and their whereabouts is important to protect certain species and to ensure the safety of surrounding human populations. In addition, it is sometimes possible to monitor certain characteristics of the animals. The length of an alligator can be estimated quite accurately from aerial photographs or from a boat. The alligator's weight is much more difficult to determine. In a research it was determined what the length (in inches) and the weight (in pounds) of alligators captured in central Florida was. It was the aim of the research to develop a model to predict the weight from the length.

weight	length	weight	length	weight	length
130	74	83	86	38	72
51	94	70	83	66	128
640	147	61	72	84	85
28	58	54	74	80	82
80	86	44	61	42	76
110	94	106	90	197	114
33	63	84	89	102	90
90	86	39	68	57	78
36	69				

(a) Make a scatter plot of length and weight.

(b) Do the same for ln(length) and ln(weight).

(c) Which seems to give a better fit?

(d) Fit the model for this. Also give the equation of best fitting line.

(e) Give the anova table and draw your conclusion from this.

4. Use the blood pressure data from the lecture.

   (a) Create the following variables in R: fgroup $<-$ factor(group) and prod $<-$ group*dose. Give an interpretation of these new variables. (Remember that group is coded as a zero and a 1)

   (b) Fit the following models with glm:

      i. y$\sim$ group+dose+group:dose
      ii. y$\sim$ fgroup+dose+fgroup:dose
      iii. y$\sim$ group+group:dose
      iv. y$\sim$ group+prod
      v. y$\sim$ fgroup+fgroup:dose

   (c) Give for each model a summary and carefully compare the differences and discuss them.

   (d) Give for each model the estimated regression lines in both groups.

5. First the data file lowbirth.dat needs to be read in. This data set is about birth weights of children. The first line contains the column names. This file can be read in with the command:
   lb $<-$ read.table(file="lowbirth.dat", header=TRUE).
   The header=TRUE states that the first line contains the column names.
   We are only going to use the following variables:

`Age of the Mother in Years`	`age`
`Smoking Status During Pregnancy (1 = Yes, 0 = No)`	`smoke`
`History of Hypertension (1 = Yes, 0 = No)`	`ht`
`Birth Weight in Grams`	`bwt`

   (a) Fit a model with bwt as dependent and with ht, smoke and age as independent variables. Also include the interaction between smoke and ht, and between age and ht in the model.

   (b) Give an interpretation of these interaction terms.

   (c) Check whether you need the interactions in the model.

(d) Find out if the model can be reduced further. Give the final model (the one that cannot be reduced) and interpret this model.