Analysis of variance - ANOVA

Based on a book by Julian J. Faraway

University of Iceland

Anova

- In ANOVAs all predictors are categorical/qualitative.
- The original thinking was to try to partition the overall variance in the response to that due to each of the factors and the error.
- Predictors are typically called *factors* which have some number of levels.
- The parameters are often called *effects*.
- We will only consider only models where the parameters are considered fixed but unknown, called fixed-effects models.
- Random-effects models are used where parameters are taken to be random variables.

Where are we...

One sided ANOVA - example of application

A pharmacutical company is testing new blood pressure medicine and conducts a little experiment. Eighteen individuals participated in the experiment and they were randomly allocated to three groups. Group one got drug 1, group two drug 2 and group three drug 3. The blood pressure was measured before and after the intake of the drug. The variable of interest is the difference in blood pressure before and after the drug intake. The mean difference blood pressure in the three groups was calculated. In all cases the blood pressure had decreased on average.

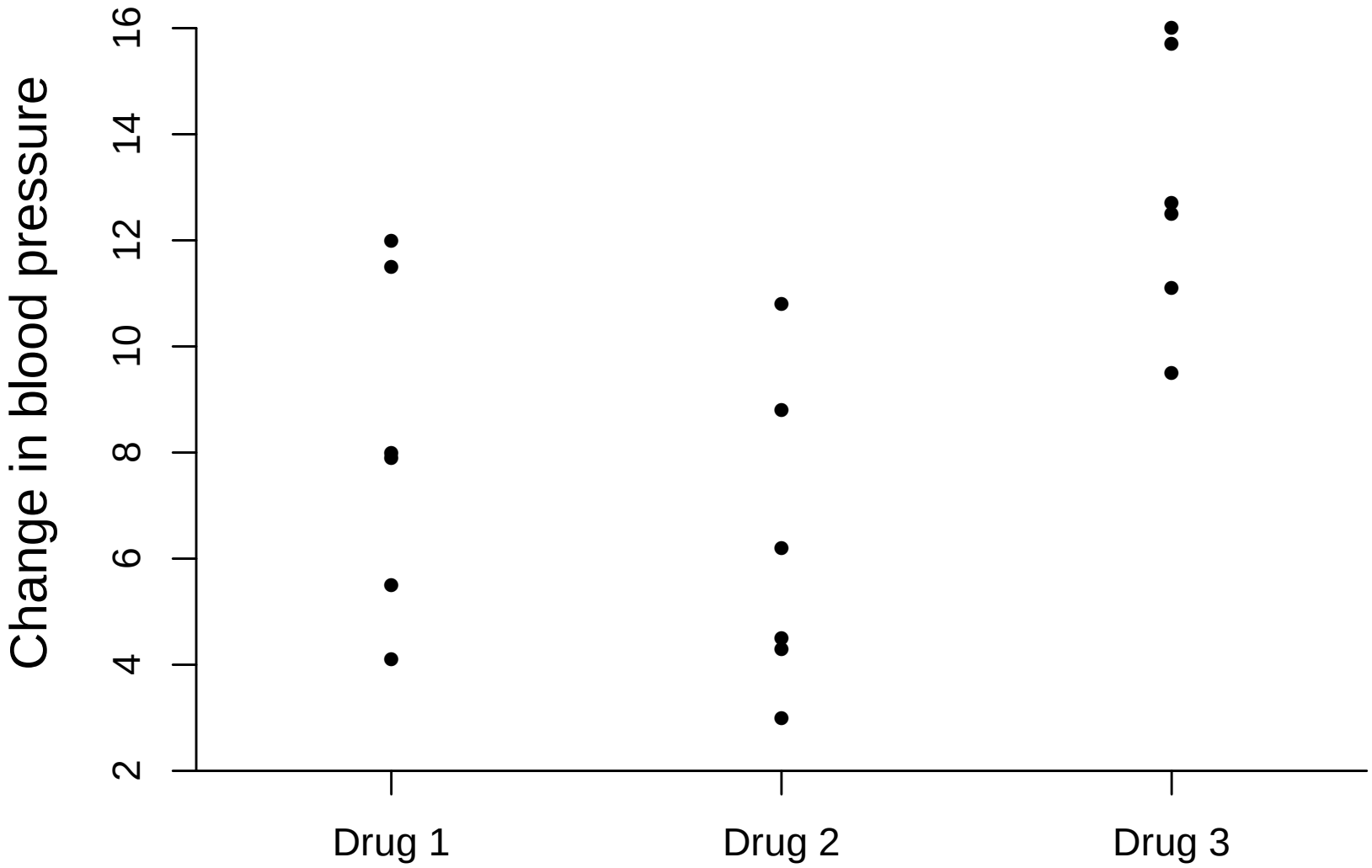
Average change group 1: $\bar{y}_{1.} = 8.14$

Average change group 2: $\bar{y}_{2.} = 6.28$

Average change group 3: $\bar{y}_{3.} = 13.01$

The question is, do the drug decrease the blood pressure equally or not?

The data



Syntax

Syntax in ANOVA

The following syntax is common in textbooks and papers discussing ANOVA.

y_{ij} : i denotes the number of the group and j denotes the number of a measurement within a group. y_{ij} is the j-th measurement in group i.

I : We denote the number of groups with I (often with a).

J_i : We denote the number of measurements in group i with J_i (often with n_i).

N : The total number of measurements is denoted with N.

$$N = J_1 + J_2 + ... + J_I.$$

$\bar{y}_{i.}$: $\bar{y}_{i.}$ denotes the mean of group i

$$\bar{y}_{i.} = \frac{\sum_{j=1}^{J_i} y_{ij}}{J_i}.$$

$\bar{y}_{..}$: $\bar{y}_{..}$ denotes the overall mean of all measurements (in all groups).

$$\bar{y}_{..} = \frac{\sum_{i=1}^{I} \sum_{j=1}^{J_i} y_{ij}}{N}.$$

Sums of squares

- We need to calculate three sums of squares, denoted with SS_T, SS_{Tr} and SS_E.
- SS_T is the total sums of squares and is a measure of the total variation.
- SS_{Tr} is a measure of the variation between groups (or treatments), that is, how much to the means of the groups vary.
- SS_E is a measure of the variability within groups (or treatments) and is therefore a measure of the error. It shows how much the measurements deviate from the mean of the group.

Sums of squares

Sums of squares in one sided ANOVA

The Sums of squares are calculated with

$$SS_T = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (y_{ij} - \bar{y}_{..})^2$$

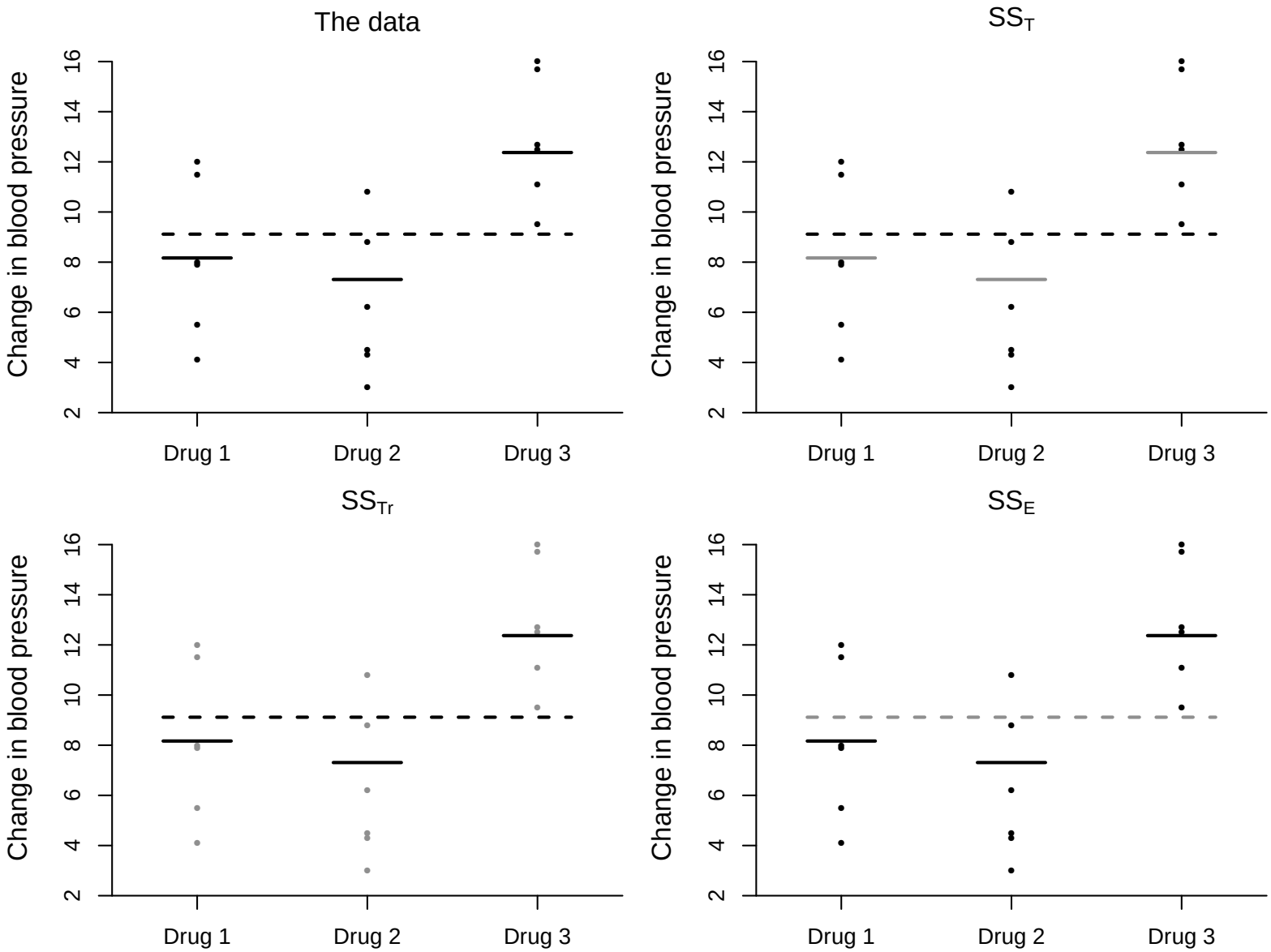
$$SS_{Tr} = \sum_{i=1}^{I} J_i (\bar{y}_{i.} - \bar{y}_{..})^2$$

$$SS_E = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (y_{ij} - \bar{y}_{i.})^2$$

The total variation can be divided into the variation between groups on one hand and the variation within groups on the other hand or

$$SS_T = SS_{Tr} + SS_E.$$

Sums of squares



ANOVA table - drug example

- It is common to visualize the sums of squares in a so-called *ANOVA table*.
- For one-way anova the table consist of three columns and three lines.
- The first column contains the sums of squares, the second one contains the number of *degrees of freedom* for each sum of squares and the third column contains so-called mean squares.
- Mean squares are calculated by dividing the corresponding sum of squares with the number of corresponding degrees of freedom (in the same line).

ANOVA table - one-way anova

Sums of squares	Degrees of freedom	Mean sum of squares
SS_{Tr}	$I-1$	$MS_{Tr} = \frac{SS_{Tr}}{I-1}$
SS_E	$N-I$	$MS_E = \frac{SS_E}{N-I}$
SS_T	$N-1$	

Hypothesis testing with ANOVA

Hypothesis testing with ANOVA

The hypothesis we want to test is generally

$$H_0 : \mu_1 = \mu_2 = ... = \mu_I$$

against the alternative hypothesis

$$H_1 : \text{At least one of the means differs from the other means.}$$

The test statistic is

$$F = \frac{SS_{Tr}/(I-1)}{SS_E/(N-I)} = \frac{MS_{Tr}}{MS_E}.$$

If the null hypothesis is true, the test statistic follows the F-distribution with $a-1$ and $N-a$ degrees of freedom, or $F \sim F_{(I-1,N-I)}$, where I is the number of groups and N is the total number of measurements.
H_0 is rejected if F $> F_{1-\alpha,(I-1,N-a)}$

One-way anova model

Given a factor α occurring at $i = 1, ..., I$ levels, with $j = 1, ..., J_i$ observationsns per level. We can write the anova model in different ways. The *means model* can be written as

$$y_{ij} = \mu_i + \varepsilon_{ij}$$

where μ_i is the mean of the ith factor level treatment.
An alternative way is to write the model as an *effect model*:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

and apply some restrictions:

1. Set $\alpha_1 = 0$ - this corresponds to the treatment contrasts in R.
2. Set $\sum_i \alpha_i = 0$ - this corresponds to the sum contrast in R.

In this form μ is a parameter common to all treatments and α_i is a parameter unique to the ith treatment.

One-way anova

- The second method is the most commonly recommended for manual calculation in older textbooks.
- The first one is the default in R.
- As usual, some preliminary graphical analysis is appropriate before fitting. A side-by-side boxplot is often the most useful plot. Look for equality of variance, possible transformations, outliers, ...

One-way anova

On the boxplot we are hoping *not* to see:

- Outliers - these will be apparent as separated points on the boxplots. The default is to extend the whiskers of the boxplot no more than one and half times the interquartiles range from the quartiles. Any points further away than this are plotted separately.
- Skewness - this will be apparent from an asymmetrical form for the boxes.
- Unequal variance - this will be apparent from clearly unequal box sizes. Some care is required because often there is very little data be used in the construction of the boxplots and so even when the variances truly are equal in the groups, we can expect a great deal of variability.

Testing for homogeneity of variance

- Most common are the Bartlett'as and the Levene's test.
- The Levene's test is insensitive to non-normality so it is often preferred.
- Rejecting the null hypothesis would indicate non constant variance.
- Most tests and CI's are relatively insensitive to non-constant variance so there is no need to take action unless the Levene test is significant at the 1% level.
- We can perform the Levene test using the `levelneTest()` in the `car` package.

One-way anova - testing

The first test of interest is whether there is a difference in the levels of the factor. The following hypotheses are equivalent to the ones showed before:

$$\begin{aligned} H_0 &: \quad \alpha_i = 0 \quad \forall i \\ H_1 &: \quad \text{At least one } \alpha_i \text{ is nonzero.} \end{aligned}$$

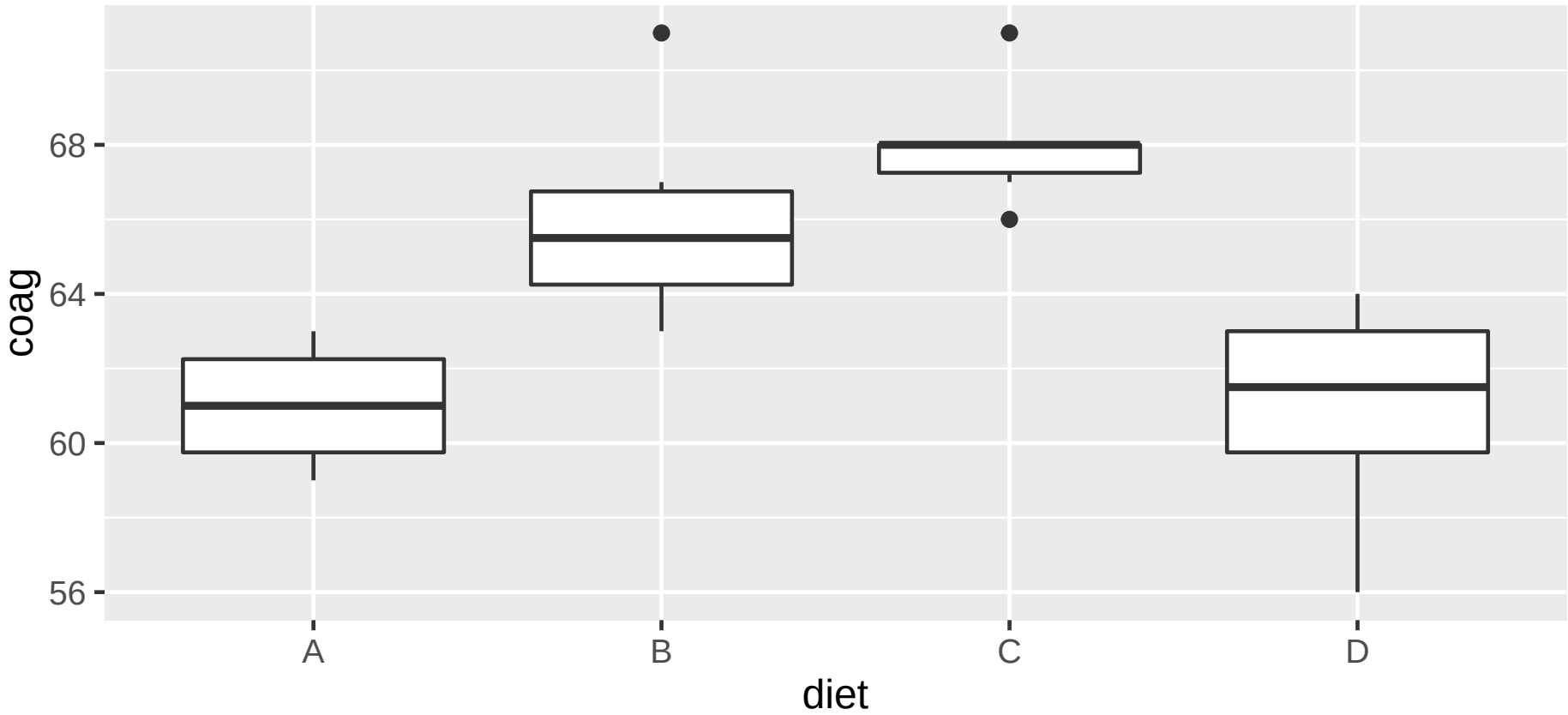
- We use the same F-test as we have used for regression.
- The outcome of this test will be the same no matter what coding/restriction we use.
- If we cannot reject the null we are done (subject to an investigation of transformation and outliers).
- If we reject the null, we must investigate which levels differ.

An example

```
library(faraway)
data(coagulation)
str(coagulation)

## 'data.frame': 24 obs. of  2 variables:
##  $ coag: num  62 60 63 59 63 67 71 64 65 66 ...
##  $ diet: Factor w/ 4 levels "A","B","C","D": 1 1 1 1 2 2 2 2 2 2 ...
```

An example

```
library(ggplot2)
ggplot(coagulation, aes(x=diet,y=coag)) + geom_boxplot()
```



Testing

```
library(car)  # need to install first
leveneTest(coag~diet,data=coagulation)

## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  3  0.6492 0.5926
##       20
```

Testing

```
fit.lm<-lm(coag~diet,data=coagulation)
summary(fit.lm)
```

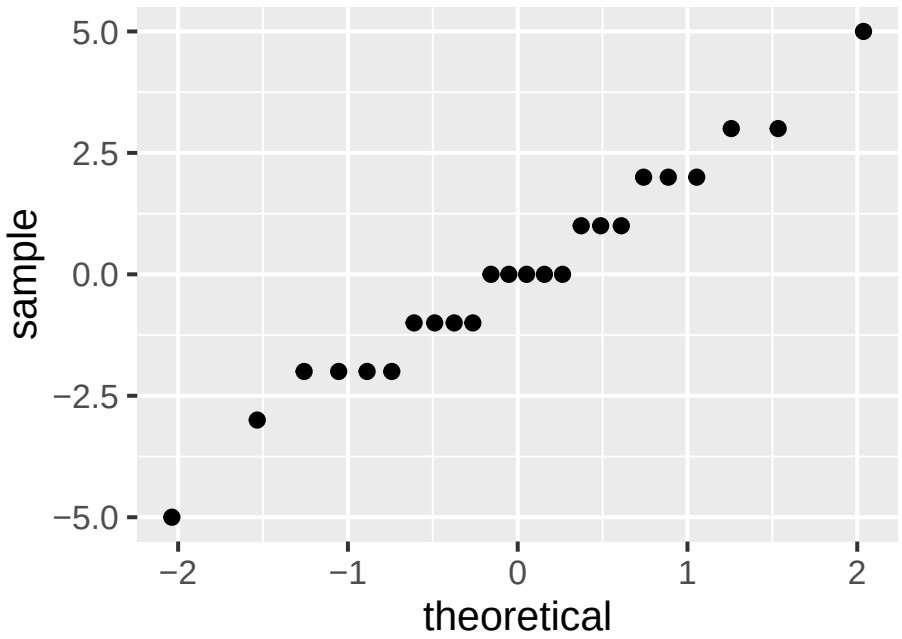
```
##
## Call:
## lm(formula = coag ~ diet, data = coagulation)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5.00  -1.25   0.00   1.25   5.00
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.100e+01  1.183e+00  51.554  < 2e-16 ***
## dietB       5.000e+00  1.528e+00   3.273 0.003803 **
## dietC       7.000e+00  1.528e+00   4.583 0.000181 ***
## dietD       2.991e-15  1.449e+00   0.000 1.000000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.366 on 20 degrees of freedom
## Multiple R-squared:  0.6706, Adjusted R-squared:  0.6212
## F-statistic: 13.57 on 3 and 20 DF,  p-value: 4.658e-05
```

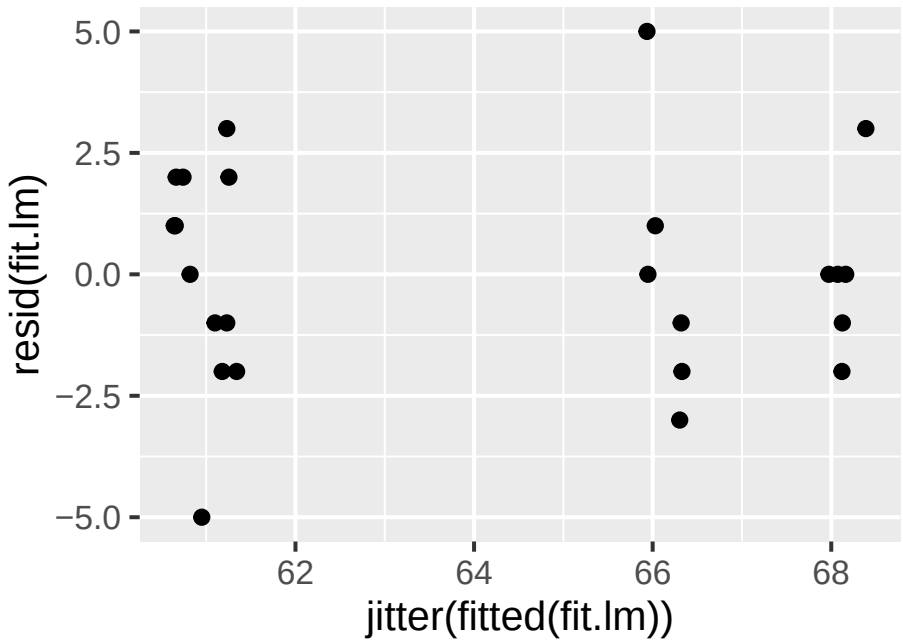
Testing

```
anova(fit.lm)

## Analysis of Variance Table
##
## Response: coag
##           Df Sum Sq Mean Sq F value    Pr(>F)
## diet       3    228    76.0  13.571 4.658e-05 ***
## Residuals 20    112     5.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Diagnostics

```
library(gridExtra)  # need to install first
p1<-ggplot(fit.lm, aes(sample = resid(fit.lm))) + stat_qq()
p2<-ggplot(fit.lm,aes(x=jitter(fitted(fit.lm)),y=resid(fit.lm))) + geom_point()
grid.arrange(p1, p2, ncol=2)
```





Diagnostics

- Because the data are integers and the fitted values turn out to integers also, some discreteness is obvious in the Q-Q plot.
- Discrete data can't be normally distributed - however, here it is approximately normal and so we can go ahead with the inference without any qualms.
- The discreteness in the residuals and fitted values shows up in the residual-fitted plot because we can see fewer points than the sample size. This is because of overplotting of the point symbols.
- There are several ways round this problem. One simple solution is to add a small amount of noise to the data. This is called jittering. Sometimes you have to tune the amount of noise but the default setting is adequate here.

Multiple Comparisons

- After detecting some difference in the levels of the factor, interest centers on which levels or combinations of levels are different.
- Note that it does not make sense to ask whether a particular level is significant since this begs the question, "significantly different from what".
- Any meaningful test must involve a comparison of some kind.
- It is important to ascertain whether the comparison made were decided on before or after examining the data.
- After fitting a model, one might decide to test only those differences that look large - to make such a decision, you also have to examine the small differences. Even if you do not actually test these small differences, it does have an effect on the inference.

Multiple Comparisons

If the comparisons were decided on prior to examining the data, there are three cases:

1. Just one comparison — use the standard t-based confidence intervals that we have used before.
2. Few comparisons — use the Bonferroni adjustment for the t. If there are m comparisons, use α/m for the critical value.
3. Many comparisons — Bonferroni becomes increasingly conservative as m increases. At some point it is better to use the Tukey or Scheffe or related methods

Multiple Comparisons

- It is difficult to be honest and be seen to be honest when using pre-data comparisons.
- Will people really believe that you only planned to make certain comparisons?
- Some might make a distinction between pre and post-data comparisons, Faraway thinks it is best to consider all comparisons as post-data.
- If the comparisons were decided on after examining the data, you must adjust the CI to allow for the possibility of all comparisons of the type to be made.
- There are two important cases:
  - Pairwise comparisons only: use the Tukey method.
  - All contrasts i.e. linear combinations: use the Scheffe method.

Tukey's Honest Significant Difference (HSD)

Tukey's Honest Significant Difference (HSD) is designed for all pairwise comparisons and depends on the studentized range distribution. The Tukey C.I.'s are:

$$\hat{\mu}_i - \hat{\mu}_j \pm q_{1-\alpha,(I,n-I)} \frac{\hat{\sigma}}{\sqrt{2}} \sqrt{\frac{1}{J_i} + \frac{1}{J_j}}.$$

Where q are quantiles from the studentized range distribution. This distribution gives the exact sampling distribution of the largest difference between a set of means originating from the same population.

When the sample sizes are very unequal, Tukey's HSD may be too conservative but in general they are narrower than those produced by Scheffe's theorem.

Tukey's HSD in R

```
fit.aov<-aov(coag~diet,data=coagulation)
TukeyHSD(fit.aov)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = coag ~ diet, data = coagulation)
##
## $diet
##     diff         lwr       upr     p adj
## B-A    5   0.7245544  9.275446 0.0183283
## C-A    7   2.7245544 11.275446 0.0009577
## D-A    0  -4.0560438  4.056044 1.0000000
## C-B    2  -1.8240748  5.824075 0.4766005
## D-B   -5  -8.5770944 -1.422906 0.0044114
## D-C   -7 -10.5770944 -3.422906 0.0001268
```

Where are we...

Two-factor anova

Suppose we have two factors, α at I levels and β at J levels.

Let n_{ij} be the number of observations at level i of α and level j of β and let those observations be $y_{ij1}, y_{ij2}, ...$

A complete layout has $n_{ij} \geq 1$ for all i, j.

The most general model that may be considered is

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}.$$

Two-way anova

The interaction effect $(\alpha\beta)_{ij}$ is interpreted as that part of the mean response not attributable to the additive effect of α_i and β_j.

A balanced layout requires that $n_{ij} = n$.

Not all the parameters are identifiable but if the main effects α and β are coded appropriately and the interaction effects coding is then derived from the product of these codings, then every contrast of parameters can be estimated.

ANOVA table - two-factor anova

For two factors A on I levels and B on J levels with n observations in each combination of levels the anova table becomes:

Sums of squares	Degrees of freedom	Mean sum of squares
SS_A	$I-1$	$MS_A = \frac{SS_A}{I-1}$
SS_B	$J-1$	$MS_B = \frac{SS_B}{J-1}$
SS_{AB}	$(I-1)(J-1)$	$MS_{AB} = \frac{SS_{AB}}{(I-1)(J-1)}$
SS_E	$IJ(n-1)$	$MS_E = \frac{SS_E}{IJ(n-1)}$
SS_T	$nIJ-1$	

One observation per cell

- When $n_{ij} = 1$ we would have as many observation as parameters if we tried to fit the full model as above.
- The parameters could be estimated but no further inference would be possible.
- We can assume $(\alpha\beta)_{ij} = 0$ free up degrees of freedom to make some tests and CI's.
- This assumption can be checked graphically using an interaction plot - plot the cell means on the vertical axis and the factor α on the horizontal. Join points with same level of β. The role of α and β can then be reversed.
- Parallel lines on the plot are a sign of a lack of interaction.

More than one observation per cell

With more than one observation per cell we are now free to fit and test the model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}.$$

The interaction effect may be tested by comparison to the model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}.$$

and computing the usual F-test.

If the interaction effect is found to be significant, do not test the main effects even if they appear not to be significant. The estimation of the main effects and their significance is coding dependent when interactions are included in the model.

Interpreting the interaction effect

No interactions: You can do pairwise comparisons on α without regard to β and vice versa.

Interaction present: A comparison of the levels of α will depend on the level of β - interpretation is not simple.

Interpreting the interaction effect

- When the interaction is significant, the main effects cannot be defined in an obvious and universal way.
- When we have a significant inteaction, we can fit a model

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

and then treat the data as a one-way anova with $I \cdot J$ levels. Obviously this makes for more complex comparisons but this is unavoidable when interactions exist.

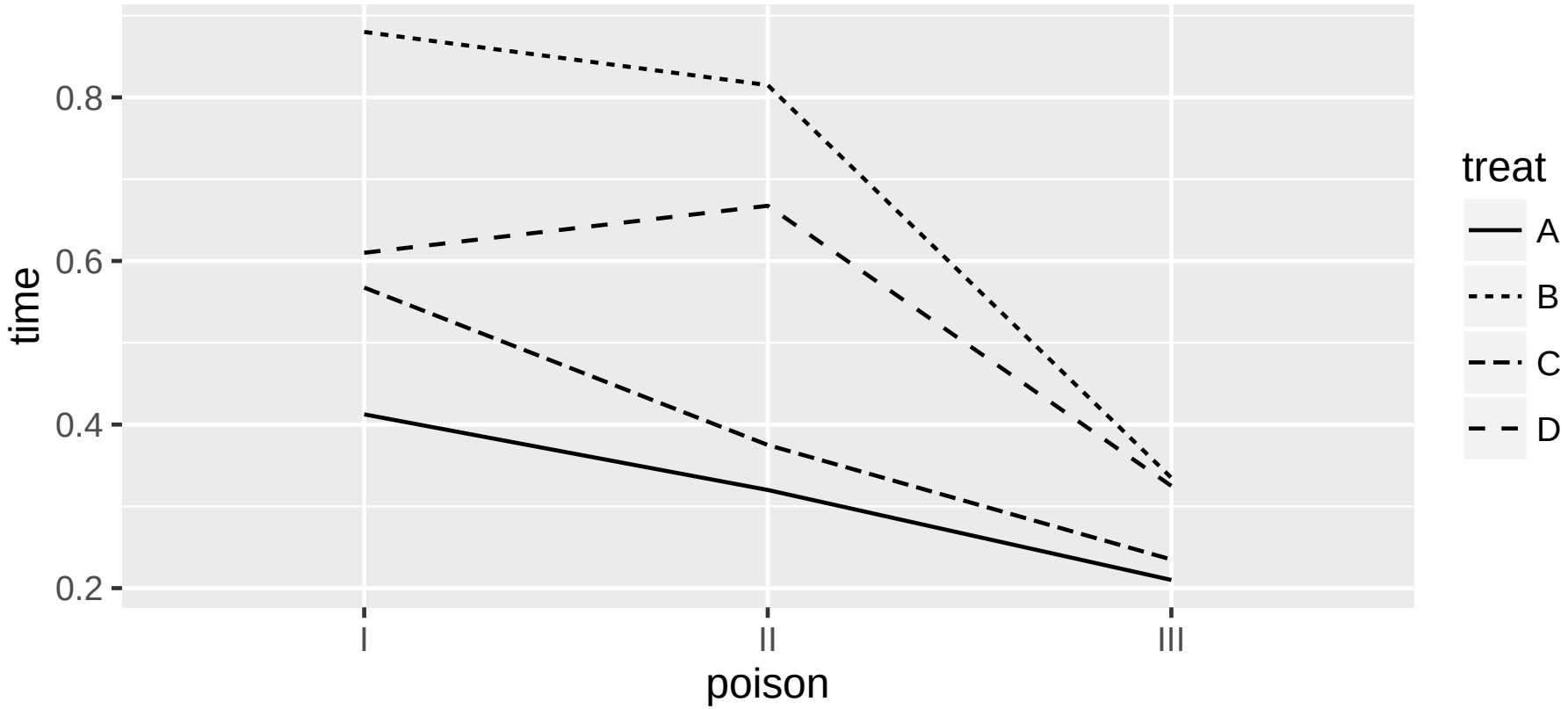
An example

As part of an investigation of toxic agents, 48 rats were allocated to 3 poisons (I,II,III) and 4 treatments (A,B,C,D). The response was survival time in tens of hours. The Data:

```
data(rats)
str(rats)

## 'data.frame': 48 obs. of  3 variables:
##  $ time  : num  0.31 0.82 0.43 0.45 0.45 1.1 0.45 0.71 0.46 0.88 ...
##  $ poison: Factor w/ 3 levels "I","II","III": 1 1 1 1 1 1 1 1 1 1 ...
##  $ treat : Factor w/ 4 levels "A","B","C","D": 1 2 3 4 1 2 3 4 1 2 ...

table(rats$poison,rats$treat)

##
##       A B C D
##   I   4 4 4 4
##   II  4 4 4 4
##   III 4 4 4 4
```

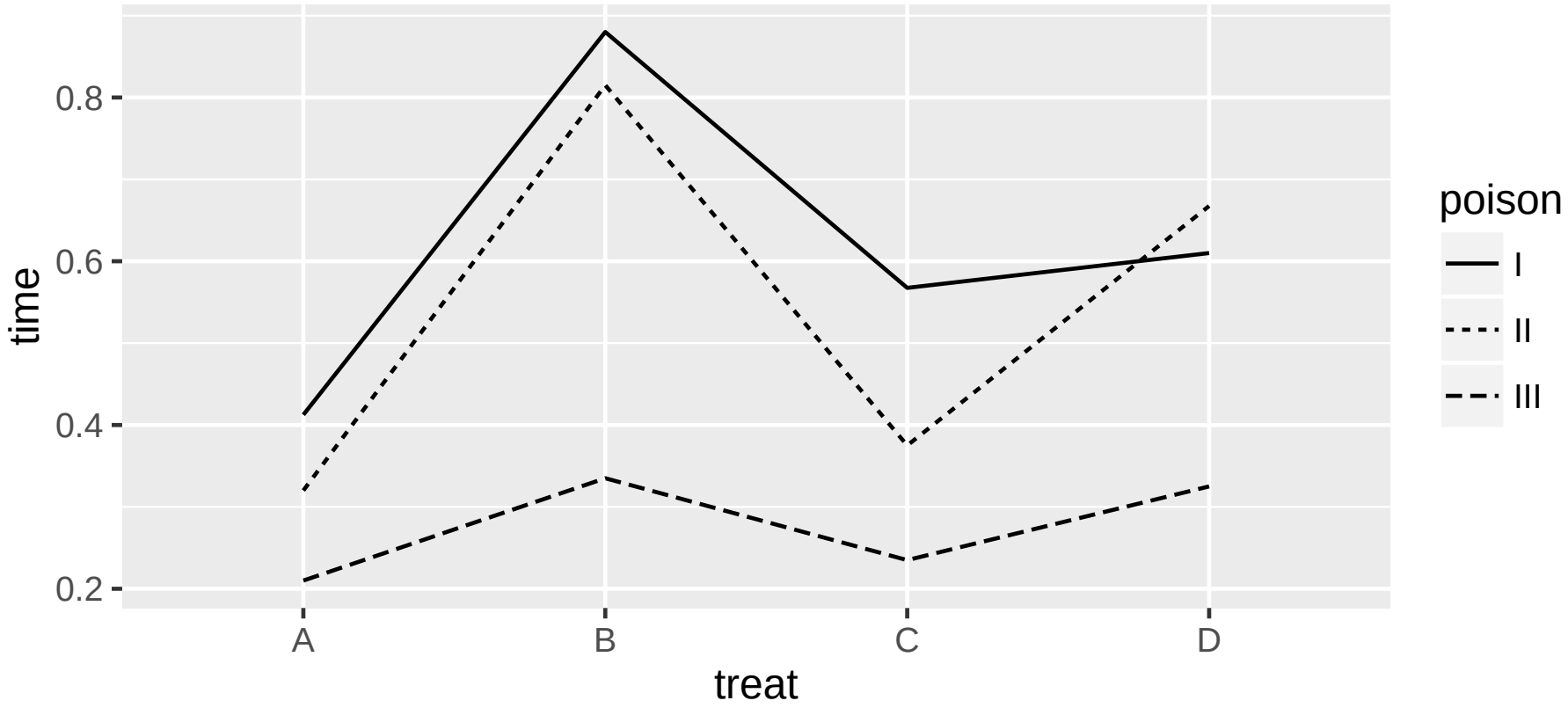
Two-way anova - an example

```
library(ggplot2)
ggplot(rats, aes(x=poison,y=time,lty=treat)) +
stat_summary(aes(group=treat),fun.y=mean,geom='line')
```



Two-way anova - an example

```
library(ggplot2)
ggplot(rats, aes(x=treat,y=time,lty=poison)) +
stat_summary(aes(group=poison),fun.y=mean,geom='line')
```



Two-way anova - an example

```
# fit the full model
fit.1<-lm(time~poison*treat ,data=rats)
anova(fit.1)

## Analysis of Variance Table
##
## Response: time
##               Df  Sum Sq Mean Sq F value    Pr(>F)
## poison         2 1.03301 0.51651 23.2217 3.331e-07 ***
## treat          3 0.92121 0.30707 13.8056 3.777e-06 ***
## poison:treat   6 0.25014 0.04169  1.8743    0.1123
## Residuals     36 0.80073 0.02224
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

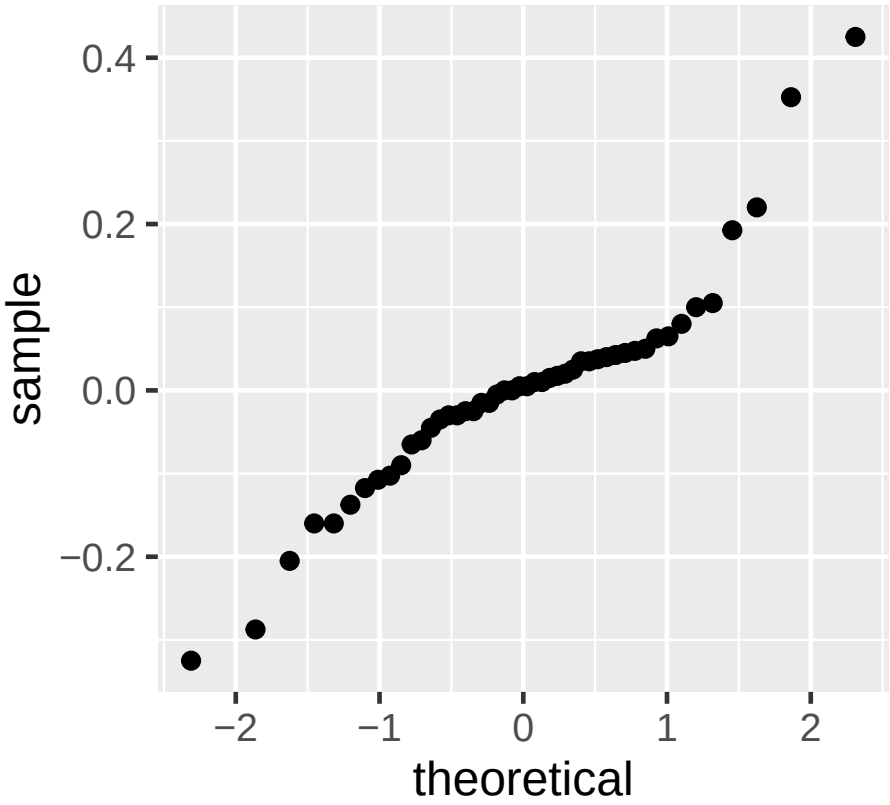
Two-way anova - an example

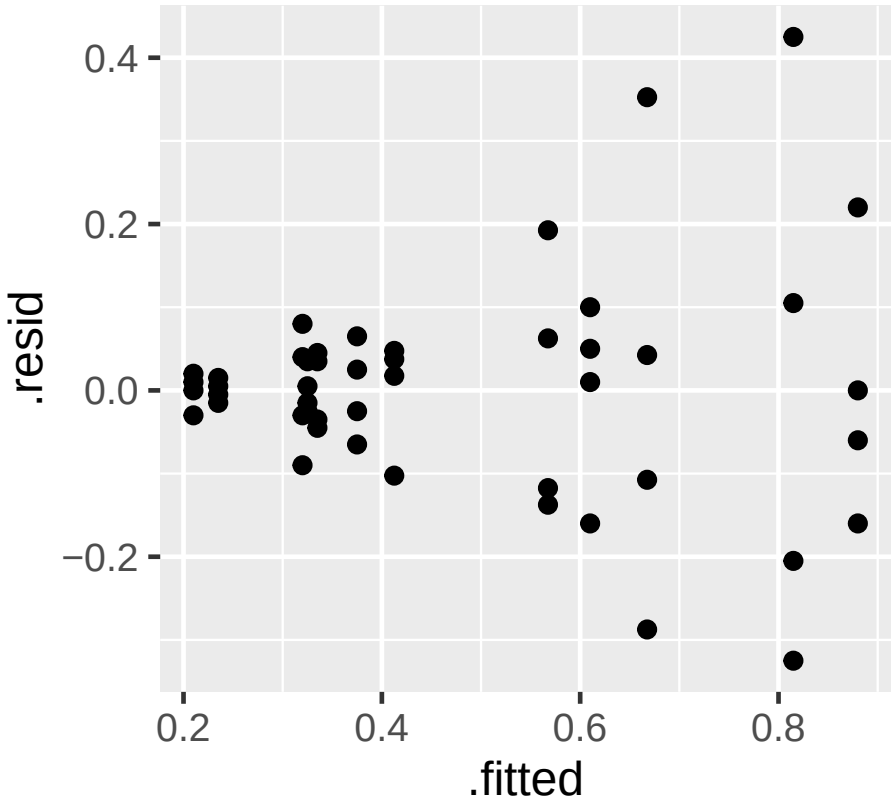
```r
# fit the full model
summary(fit.1)
```

```
##
## Call:
## lm(formula = time ~ poison * treat, data = rats)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32500 -0.04875  0.00500  0.04312  0.42500
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.41250    0.07457   5.532 2.94e-06 ***
## poisonII         -0.09250    0.10546  -0.877   0.3862
## poisonIII        -0.20250    0.10546  -1.920   0.0628 .
## treatB            0.46750    0.10546   4.433 8.37e-05 ***
## treatC            0.15500    0.10546   1.470   0.1503
## treatD            0.19750    0.10546   1.873   0.0692 .
## poisonII:treatB   0.02750    0.14914   0.184   0.8547
## poisonIII:treatB -0.34250    0.14914  -2.297   0.0276 *
## poisonII:treatC  -0.10000    0.14914  -0.671   0.5068
## poisonIII:treatC -0.13000    0.14914  -0.872   0.3892
## poisonII:treatD   0.15000    0.14914   1.006   0.3212
## poisonIII:treatD -0.08250    0.14914  -0.553   0.5836
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1491 on 36 degrees of freedom
## Multiple R-squared:  0.7335, Adjusted R-squared:  0.6521
```

Two-way anova - an example

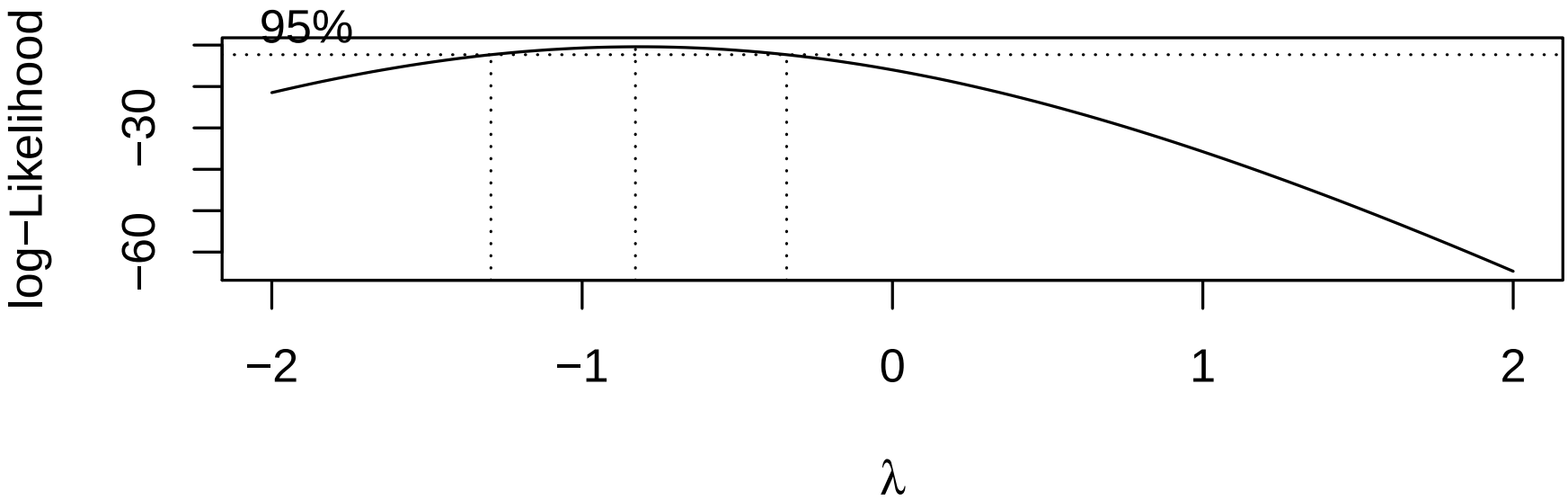
```
library(ggplot2)
p1.1<-ggplot(fit.1, aes(sample = .resid)) + stat_qq()
p1.2<-ggplot(fit.1, aes(x=.fitted, y=.resid)) + geom_point()
grid.arrange(p1.1, p1.2, ncol=2)
```





Two-way anova - an example

```
library(MASS)
boxcox(fit.1)
```

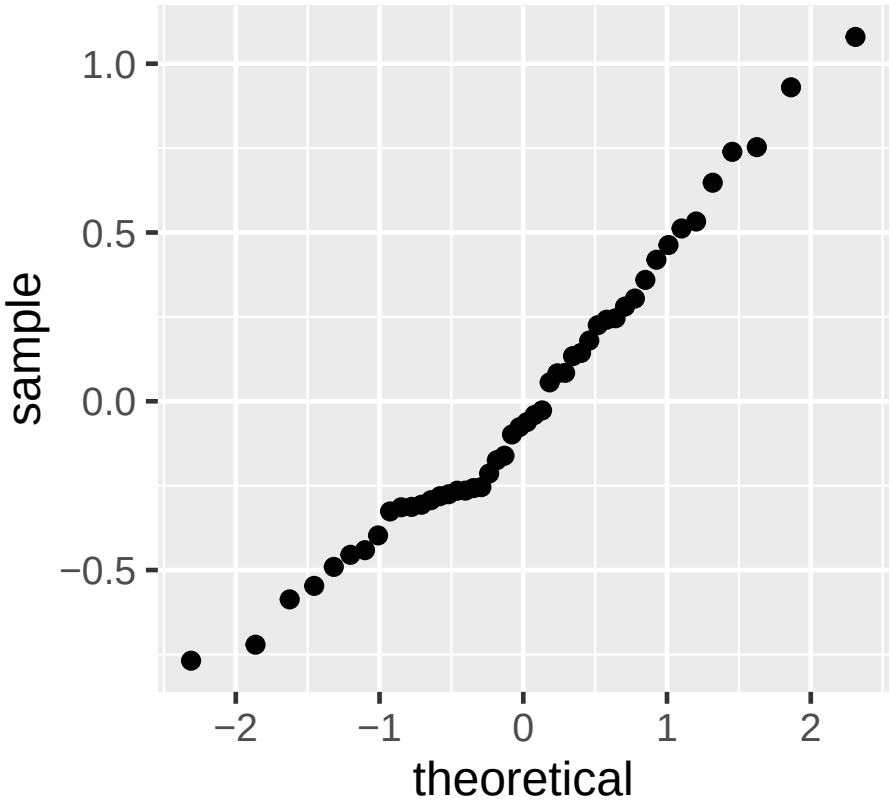


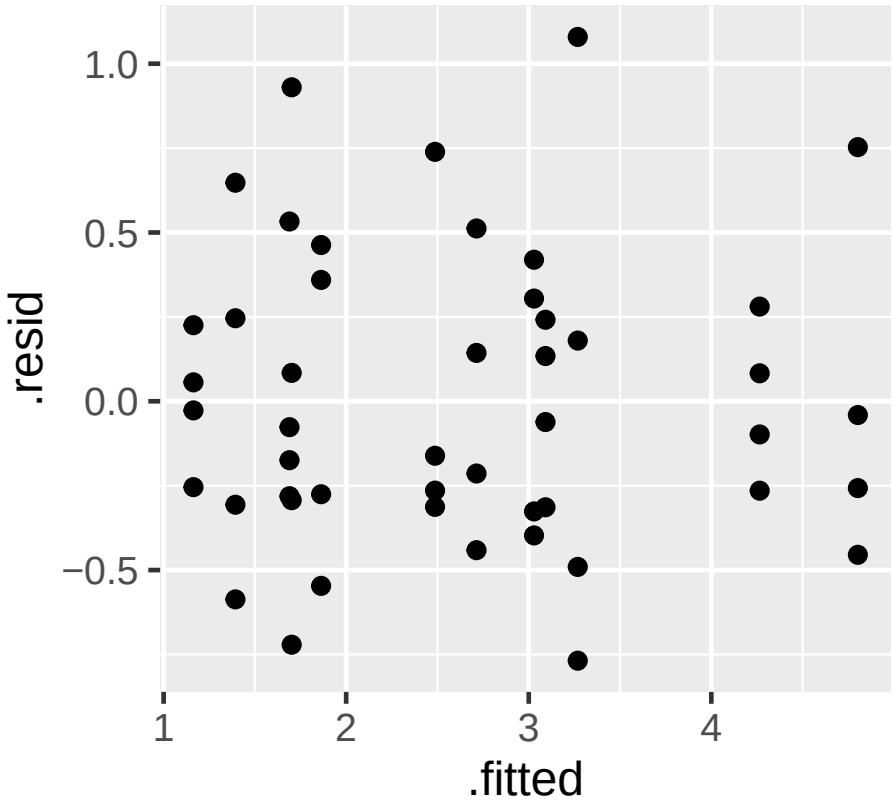
Two-way anova - an example

```
# fit the full model
fit.2<-lm(1/time~poison*treat ,data=rats)
anova(fit.2)

## Analysis of Variance Table
##
## Response: 1/time
##              Df Sum Sq Mean Sq F value    Pr(>F)
## poison        2 34.877 17.4386 72.6347 2.310e-13 ***
## treat         3 20.414  6.8048 28.3431 1.376e-09 ***
## poison:treat  6  1.571  0.2618  1.0904    0.3867
## Residuals    36  8.643  0.2401
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Two-way anova - an example

```
library(ggplot2)
p2.1<-ggplot(fit.2, aes(sample = .resid)) + stat_qq()
p2.2<-ggplot(fit.2, aes(x=.fitted, y=.resid)) + geom_point()
grid.arrange(p2.1, p2.2, ncol=2)
```





Two-way anova - an example

```
# fit a reduced model
fit.3<-lm(1/time~poison+treat ,data=rats)
anova(fit.3)

## Analysis of Variance Table
##
## Response: 1/time
##           Df Sum Sq Mean Sq F value    Pr(>F)
## poison     2 34.877 17.4386  71.708 2.865e-14 ***
## treat      3 20.414  6.8048  27.982 4.192e-10 ***
## Residuals 42 10.214  0.2432
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Two-way anova - an example

```
summary(fit.3)


##
## Call:
## lm(formula = 1/time ~ poison + treat, data = rats)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.82757 -0.37619  0.02116  0.27568  1.18153
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.6977     0.1744  15.473  < 2e-16 ***
## poisonII      0.4686     0.1744   2.688  0.01026 *
## poisonIII     1.9964     0.1744  11.451 1.69e-14 ***
## treatB       -1.6574     0.2013  -8.233 2.66e-10 ***
## treatC       -0.5721     0.2013  -2.842  0.00689 **
## treatD       -1.3583     0.2013  -6.747 3.35e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4931 on 42 degrees of freedom
## Multiple R-squared:  0.8441, Adjusted R-squared:  0.8255
## F-statistic: 45.47 on 5 and 42 DF,  p-value: 6.974e-16
```

Two-way anova - an example

```
fit.3.aov<-aov(1/time~poison+treat ,data=rats)
TukeyHSD(fit.3.aov)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = 1/time ~ poison + treat, data = rats)
##
## $poison
##             diff        lwr       upr     p adj
## II-I   0.4686413 0.04505584 0.8922267 0.0271587
## III-I  1.9964249 1.57283950 2.4200103 0.0000000
## III-II 1.5277837 1.10419824 1.9513691 0.0000000
##
## $treat
##           diff        lwr         upr     p adj
## B-A -1.6574024 -2.1959343 -1.11887050 0.0000000
## C-A -0.5721354 -1.1106673 -0.03360355 0.0335202
## D-A -1.3583383 -1.8968702 -0.81980640 0.0000002
## C-B  1.0852669  0.5467351  1.62379883 0.0000172
## D-B  0.2990641 -0.2394678  0.83759598 0.4550931
## D-C -0.7862029 -1.3247347 -0.24767096 0.0018399
```

Replication

- It's important that the observations observed in each cell are genuine replications.
- If this is not true, then the observations will be correlated and the analysis will need to be adjusted.
- Data where the replicates are correlated can be handled with repeated measures models.