Lecture7-ANOVA

Ruibin Xi

3/22/2020

First load the faraway library

```
library(faraway)
library(ggplot2)
library(tidyverse)
library(car)
```

算法包

Analysis of Variance (ANOVA)

We consider the coagulation data. The example dataset we will use is a set of 24 blood coagulation times. 24 animals were randomly assigned to four different diets and the samples were taken in a random order. This data comes from “Box, G. P., W. G. Hunter, and J. S. Hunter (1978). Statistics for Experimenters. New York: Wiley.”

```
data(coagulation, package="faraway")
head(coagulation)
```

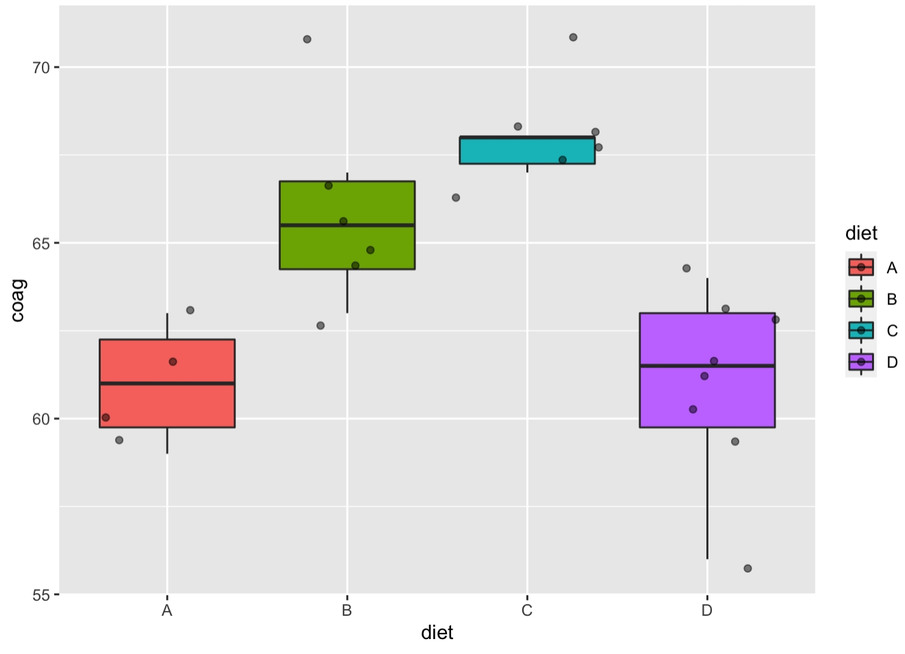
```
##   coag diet
## 1   62    A
## 2   60    A
## 3   63    A
## 4   59    A
## 5   63    B
## 6   67    B
```

diet是协变量(factor)，对应4个level

The first step is to plot the data - boxplots are useful. We are hoping not to see 1. Outliers — these will be apparent as separated points on the boxplots. 2. Skewness — this will be apparent from an asymmetrical form for the boxes. 3. Unequal variance — this will be apparent from clearly unequal box sizes.

```
coagulation %>% ggplot(aes(diet, coag,fill=diet)) + geom_boxplot(outlier.shape = NA) + geom_jitter(alpha=0.5)
```

按diet分组，画coag的boxplot,
box的宽度长度有区别，所以可以继续用线性模型



In this case, there are no obvious problems. For group C, there are only 4 distinct observations and one is somewhat separated which accounts for the slightly odd looking plot. Now let's fit the model.

```
lmod <- lm(coag ~ diet, coagulation)
summary(lmod)
```

linear model
lm(y~x)

```
##
## 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
```

We conclude from the small p-value for the F-statistic that there is some difference between the groups. Let's look at the design matrix

```
model.matrix(lmod)
```

查看回归里的x到底是什么，intercept默认对应第一个种类（dietA）

```
##    (Intercept) dietB dietC dietD
## 1            1     0     0     0
## 2            1     0     0     0
## 3            1     0     0     0
## 4            1     0     0     0
## 5            1     1     0     0
## 6            1     1     0     0
## 7            1     1     0     0
## 8            1     1     0     0
## 9            1     1     0     0
## 10           1     1     0     0
## 11           1     0     1     0
## 12           1     0     1     0
## 13           1     0     1     0
## 14           1     0     1     0
## 15           1     0     1     0
## 16           1     0     1     0
## 17           1     0     0     1
## 18           1     0     0     1
## 19           1     0     0     1
## 20           1     0     0     1
## 21           1     0     0     1
## 22           1     0     0     1
## 23           1     0     0     1
## 24           1     0     0     1
## attr(,"assign")
## [1] 0 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$diet
## [1] "contr.treatment"
```

n=24，24个样本

The anova function can give us the ANOVA comparisons

```
anova(lmod)
```

看不同的组中的均值是否有所区别

```
## Analysis of Variance Table
##
## Response: coag
##           Df Sum Sq Mean Sq F value    Pr(>F)
```

p很小，证明四组的均值不太一样

```
## 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
```

diet有4个level，k=4，自由度=4-1=3
residuals 自由度=n-k = 24-4=20

We can perform the analysis by excluding the intercept. In this case, it corresponds to another type of coding.

另一种方法：
不要intercept这一项

```
lmodi <- lm(coag ~ diet -1, coagulation)
summary(lmodi)
```

```
##
## Call:
## lm(formula = coag ~ diet - 1, 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|)
## dietA  61.0000     1.1832   51.55   <2e-16 ***
## dietB  66.0000     0.9661   68.32   <2e-16 ***
## dietC  68.0000     0.9661   70.39   <2e-16 ***
## dietD  61.0000     0.8367   72.91   <2e-16 ***
## ---
## 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:  0.9989, Adjusted R-squared:  0.9986
## F-statistic:  4399 on 4 and 20 DF,  p-value: < 2.2e-16
```

因为主要关心不同组间是否有区别，
所以这里estimate算出来结果和上面方法
不一样也没关系

Let’s look at the design matrix

```
model.matrix(lmodi)
```

```
##    dietA dietB dietC dietD
## 1      1     0     0     0
## 2      1     0     0     0
## 3      1     0     0     0
## 4      1     0     0     0
## 5      0     1     0     0
## 6      0     1     0     0
## 7      0     1     0     0
## 8      0     1     0     0
## 9      0     1     0     0
## 10     0     1     0     0
## 11     0     0     1     0
## 12     0     0     1     0
```

```
## 13     0     0     1     0
## 14     0     0     1     0
## 15     0     0     1     0
## 16     0     0     1     0
## 17     0     0     0     1
## 18     0     0     0     1
## 19     0     0     0     1
## 20     0     0     0     1
## 21     0     0     0     1
## 22     0     0     0     1
## 23     0     0     0     1
## 24     0     0     0     1
## attr(,"assign")
## [1] 1 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$diet
## [1] "contr.treatment"
```

We can also perform the analysis by just considering the intercept itself.

```
lmnull <- lm(coag ~ 1, coagulation)
summary(lmnull)
```

```
##
## Call:
## lm(formula = coag ~ 1, data = coagulation)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8.00  -2.25  -0.50   3.00   7.00
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  64.0000     0.7848   81.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.845 on 23 degrees of freedom
```

Compare the two models.

```
anova(lmnull,lmodi)
```

不加任何factor 的模型和加一个factor的模型来比较
得到与前一种方法一样的结果

```
## Analysis of Variance Table
##
## Model 1: coag ~ 1
## Model 2: coag ~ diet - 1
##   Res.Df RSS Df Sum of Sq      F    Pr(>F)
```

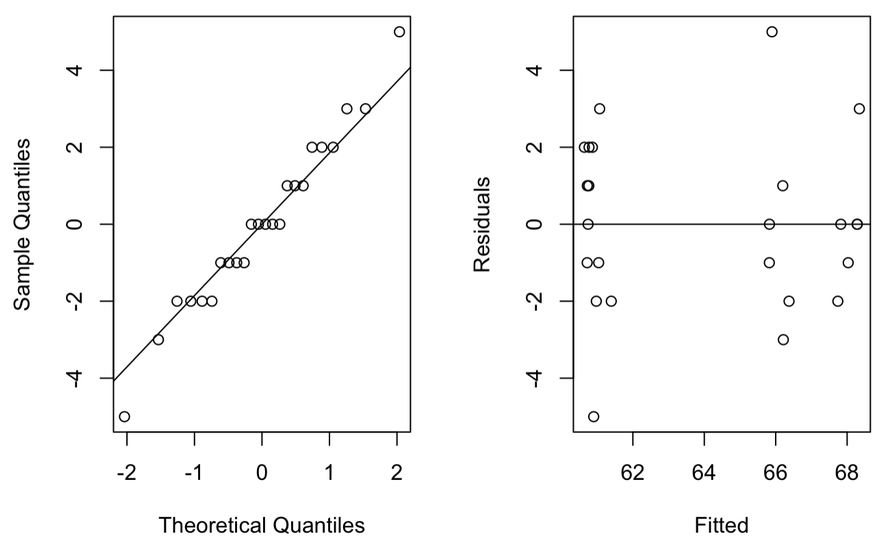
```
## 1     23 340
## 2     20 112  3       228 13.571 4.658e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now we perform some model diagnostic analysis.

```
#options(contrasts=c("contr.sum","contr.poly"))
#lmods <- lm(coag ~ diet , coagulation)
par(mfrow=c(1,2))
qqnorm(residuals(lmod))
qqline(residuals(lmod))
plot(jitter(fitted(lmod)),residuals(lmod),xlab="Fitted",ylab="Residuals")
abline(h=0)
```

QQplot看r分布

看y和r的相关性



We can use the barlett.test to test for homogeneity.

```
bartlett.test(coag ~ diet, coagulation)
```

homogeneity（假设2）的假设检验

```
##
##  Bartlett test of homogeneity of variances
##
```

p很大，不拒绝零假设，error的方差=0成立，符合假设2前提

```
## data:  coag by diet
## Bartlett's K-squared = 1.668, df = 3, p-value = 0.6441
```

We can construct the confidence interval using TukeyHSD.

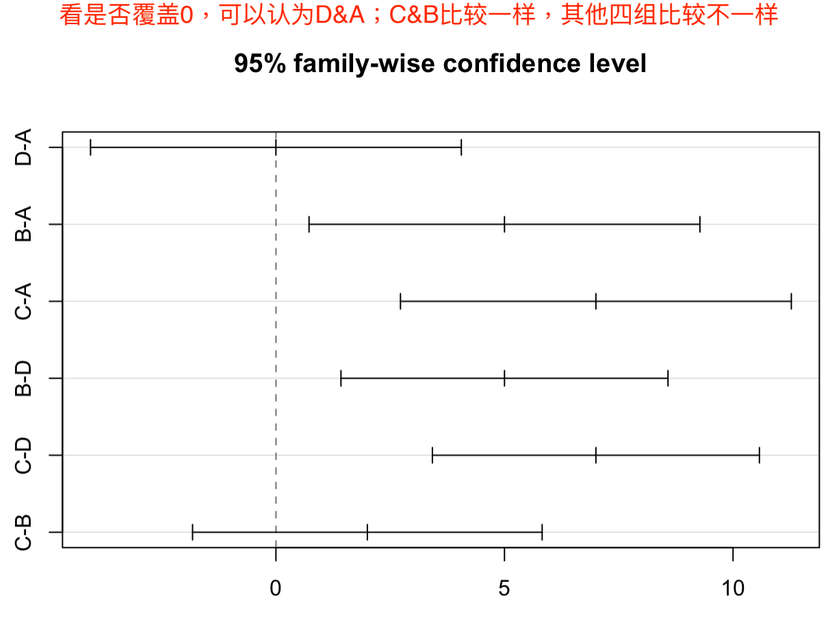
```
tci <- TukeyHSD(aov(lmod),conf.level=0.95,ordered=TRUE)
tci
```

两两之间的均值比较，算平行比较的置信度范围：A&B；A&C；A&D；B&C；B&D；C&D

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##     factor levels have been ordered
##
## Fit: aov(formula = lmod)
##
## $diet
##     diff        lwr       upr     p adj
## D-A    0 -4.0560438  4.056044 1.0000000
## B-A    5  0.7245544  9.275446 0.0183283
## C-A    7  2.7245544 11.275446 0.0009577
## B-D    5  1.4229056  8.577094 0.0044114
## C-D    7  3.4229056 10.577094 0.0001268
## C-B    2 -1.8240748  5.824075 0.4766005
```

```
plot(tci)
```

看是否覆盖0，可以认为D&A；C&B比较一样，其他四组比较不一样



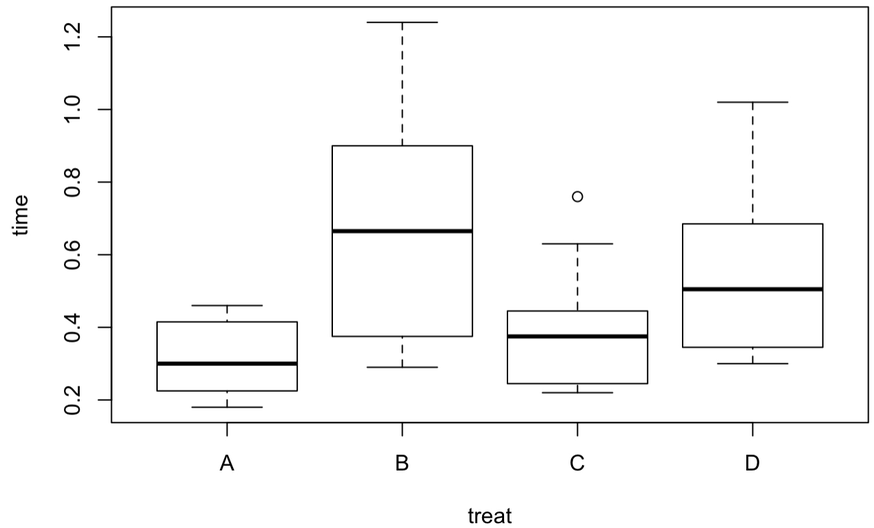
Now we consider the rats data. Here is a two-way anova design where there are 4 replicates. As part of an investigation of toxic agents, 48 rats were allocated to 3 poisons (I,II,III) and 4 treatments (A,B,C,D). The response was survival time in tens of hours.

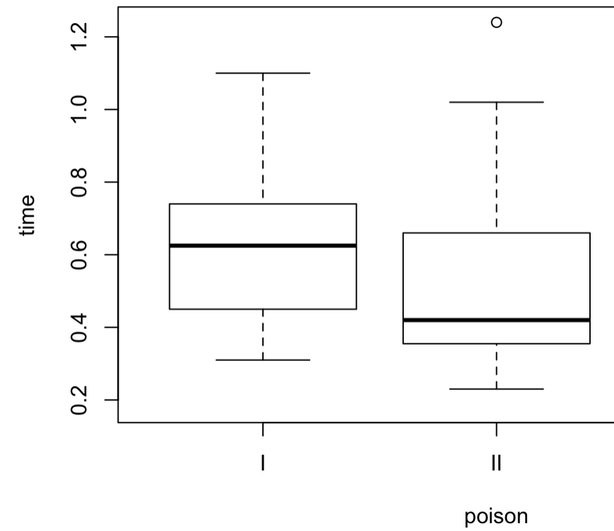
```
data(rats)
head(rats)
```

```
##   time poison treat
## 1 0.31      I     A
## 2 0.82      I     B
## 3 0.43      I     C
## 4 0.45      I     D
## 5 0.45      I     A
## 6 1.10      I     B
```

2个factor

```
plot(time ~ treat + poison, data=rats)
```



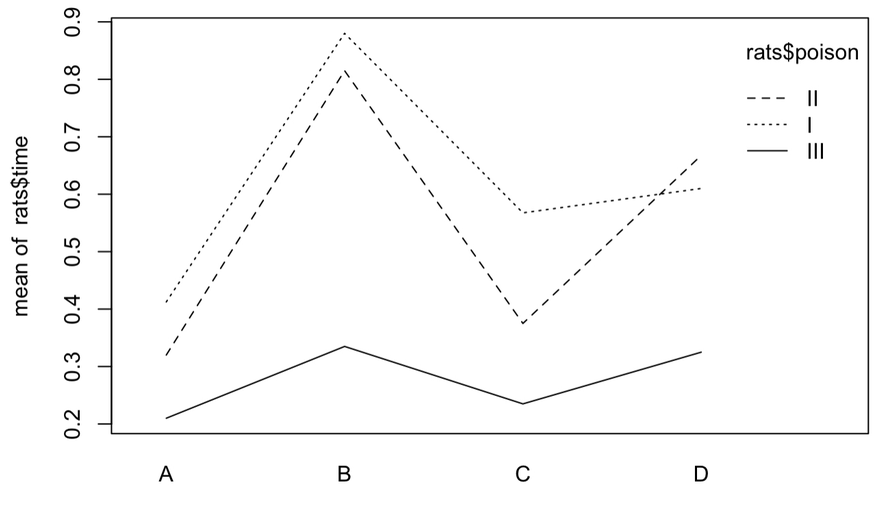


Some evidence of skewness can be seen, especially since it appears that variance is in some way related to the mean response. We now check for an interaction using graphical methods:

```
interaction.plot(rats$treat,rats$poison,rats$time)
```

x 分组依据 y

看两两factor之间是
negative interaction 还是
possitive interaction

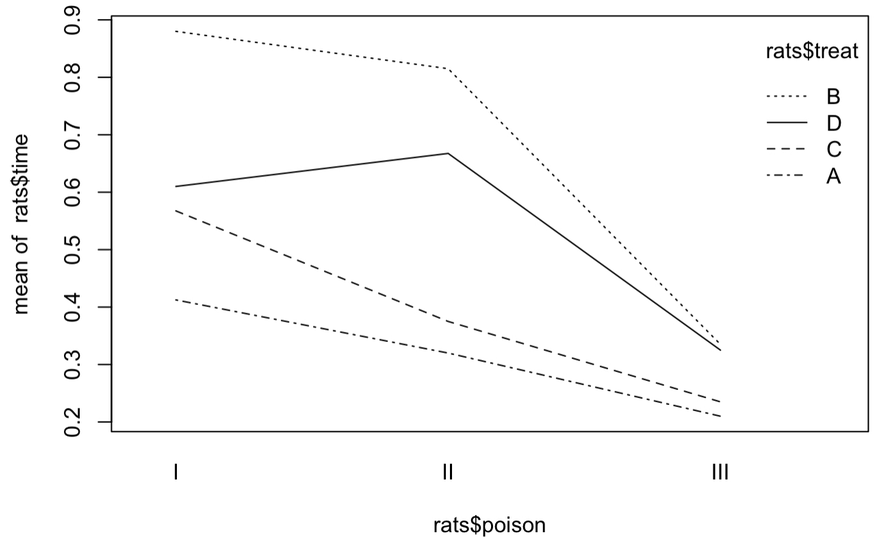


平行的线，推测poison和treat之间没有interaction

如果是相差比较大的线，则推测有比较明显的interaction

```
interaction.plot(rats$poison,rats$treat,rats$time)
```

x　　分组依据　　y



Do these look parallel? We have replication so we can directly test for an interaction effect.

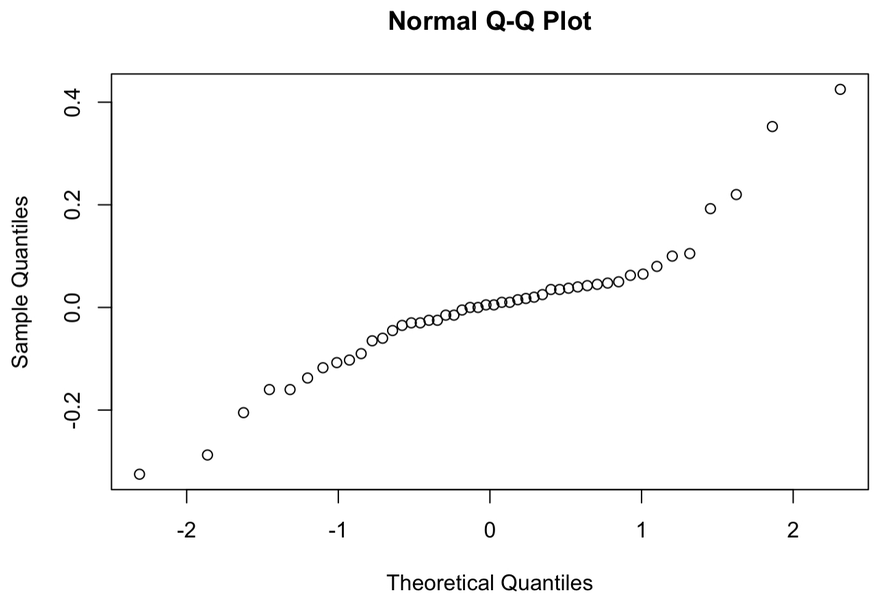
```
g <- lm(time~poison*treat, rats)
anova(g)
```

```
## 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
```

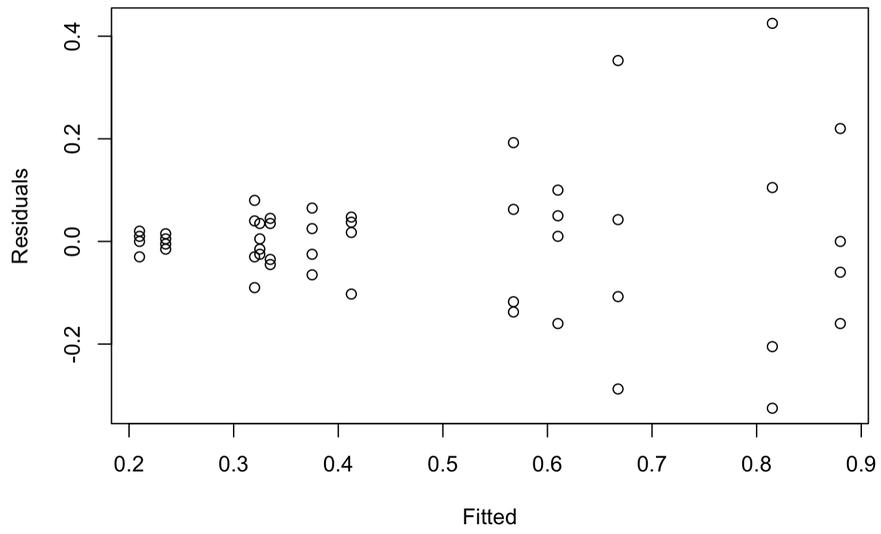
交互项

We see that the interaction effect is not significant but the main effects are. We check the diagnostics.

```
qqnorm(g$res)
```



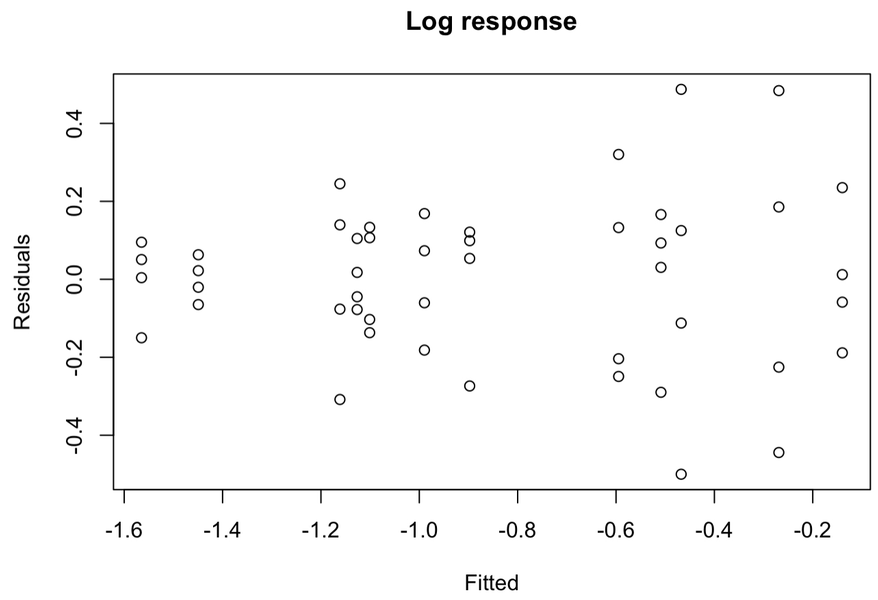

```
plot(g$fitted,g$res,xlab="Fitted",ylab="Residuals")
```



这个图数据之间的前窄后宽，出来结果可能不大好(x和y非线性关系），所以变换一下，用logx来替代x

Clearly there's a problem - perhaps transforming the data will help. Try logs first:

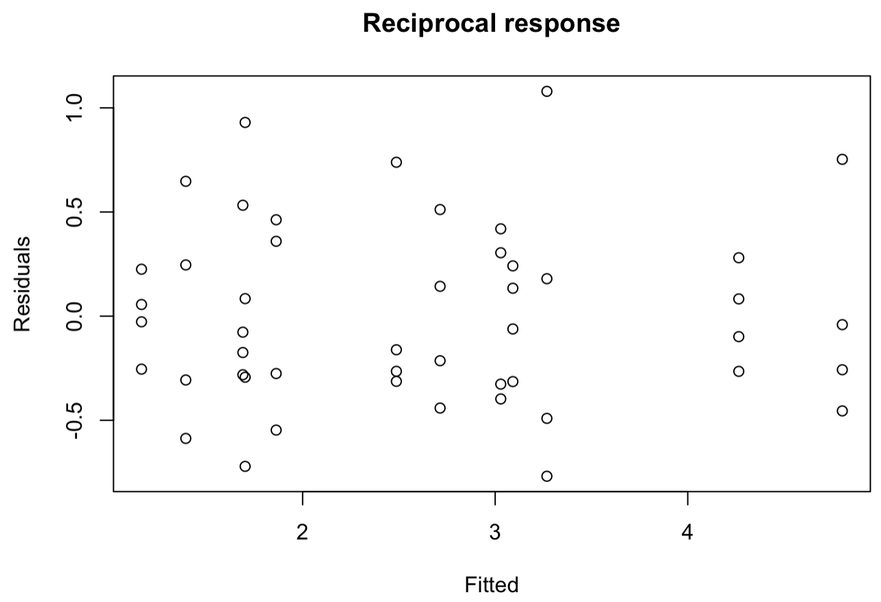
```
g <- lm(log(time) ~ poison*treat,rats)
plot(g$fitted,g$res,xlab="Fitted",ylab="Residuals",main="Log response")
```



这个图分布就平均一点了

Not enough so try the reciprocal:

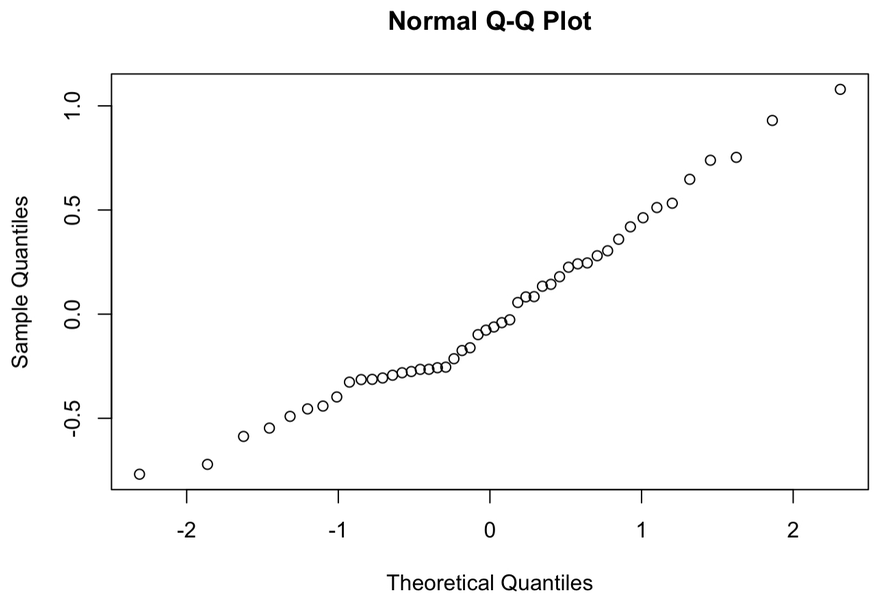
```
g <- lm(1/time ~ poison*treat,rats)
plot(g$fitted,g$res,xlab="Fitted",ylab="Residuals",main="Reciprocal response")
```



然后再尝试一下另一种x的转换，
再看qqplot，更明显地在一条直线上，这样数据更好，所以实际上是跟1/time有更明显地线性关系

Looks good - the reciprocal can be interpreted as the rate of dying. Better check the Q-Q plot again:

```
qqnorm(g$res)
```



This looks better than the first Q-Q plot. We now check the ANOVA table again, find the interaction is not significant, simplify the model and examine the fit:

```
anova(g)
```

```
## 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
```

```
g1 <- lm(1/time~poison+treat, rats)
summary(g1)
```

```
##
## 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
```

pvalue含义：poison3 - poison1(intercept) =0 的概率，p越小，决绝零假设，看estimate，差值越大，则y（time）相差越大

```
anova(g1)
```

```
## 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
```

Since the design is othorgonal, the order of the covariates will not influence the result of anova.

```
anova(lm(1/time~treat+poison, rats))
```

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

例子

However, this may not be the case for general data sets

```
mod <- lm(conformity ~ fcategory+partner.status, data=Moore)
anova(mod)
```

```
## Analysis of Variance Table
##
## Response: conformity
##                Df Sum Sq Mean Sq F value   Pr(>F)
## fcategory       2   3.73   1.867  0.0771 0.925974
## partner.status  1 212.21 212.214  8.7599 0.005098 **
## Residuals      41 993.25  24.226
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod1 <- lm(conformity ~ partner.status+fcategory, data=Moore)
anova(mod1)
```

```
## Analysis of Variance Table
##
## Response: conformity
##                Df Sum Sq Mean Sq F value   Pr(>F)
## partner.status  1 204.33 204.332  8.4345 0.005906 **
## fcategory       2  11.61   5.807  0.2397 0.787944
## Residuals      41 993.25  24.226
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

factor顺序不同，结果也有不同。
因为大多考虑最后一项放进来对之前结果的影响。

Clearly, we see that two anova tables are different. This is because anova used the so called type I SS (SAS convention).

Given two covariates, type I SS calculates SS(A), SS(B|A), SS(A*B|A,B) (if interaction is included) consecutively. The order will thus be important. Good thing is that all SSs sum up to SST. The Anova function in car package can calculate type II SSs.

In comparison, type II SS calulates SS(A|B), SS(B|A) and SS(A*B|A,B) and thus the order is not important.

```
Anova(mod)
```

大写的Anova对顺序没有要求

```
## Anova Table (Type II tests)
##
## Response: conformity
##                Sum Sq Df F value   Pr(>F)
## fcategory       11.61  2  0.2397 0.787944
## partner.status 212.21  1  8.7599 0.005098 **
## Residuals      993.25 41
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(mod1)
```

```
## Anova Table (Type II tests)
##
## Response: conformity
##                Sum Sq Df F value   Pr(>F)
## partner.status 212.21  1  8.7599 0.005098 **
## fcategory       11.61  2  0.2397 0.787944
## Residuals      993.25 41
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With the interaction term

```
mod <- lm(conformity ~ fcategory*partner.status, data=Moore)
Anova(mod)
```

```
## Anova Table (Type II tests)
##
## Response: conformity
##                          Sum Sq Df F value   Pr(>F)
## fcategory                 11.61  2  0.2770 0.759564
## partner.status           212.21  1 10.1207 0.002874 **
## fcategory:partner.status 175.49  2  4.1846 0.022572 *
## Residuals                817.76 39
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod1 <- lm(conformity ~ partner.status*fcategory, data=Moore)
Anova(mod1)
```

```
## Anova Table (Type II tests)
##
## Response: conformity
##                          Sum Sq Df F value   Pr(>F)
## partner.status           212.21  1 10.1207 0.002874 **
## fcategory                 11.61  2  0.2770 0.759564
## partner.status:fcategory 175.49  2  4.1846 0.022572 *
## Residuals                817.76 39
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

look at the detailed calculation of one way ANOVA using iris data

```
data(iris)
```

```
y = iris$Sepal.Width
group = iris$Species

y.mean.group = aggregate(y,by=list(group),mean)
n.ele.group = aggregate(y,by=list(group),length)
sst = sum((y-mean(y))^2)
sst
```

```
## [1] 28.30693
```

```
ssb = sum((y.mean.group[,2]-mean(y))^2*n.ele.group[,2])
ssb
```

```
## [1] 11.34493
```

```
sse = sst - ssb
sse
```

```
## [1] 16.962
```

```
msb = ssb/(length(unique(group))-1)
msb
```

```
## [1] 5.672467
```

```
mse = sse/(nrow(iris)-length(unique(group)))
mse
```

```
## [1] 0.1153878
```

```
F= msb/mse
F
```

```
## [1] 49.16004
```

```
qf(0.95,df1=2,df2=147)
```

```
## [1] 3.057621
```

```
pvalue=pf(F,lower.tail=FALSE,df1=2,df2=147)
```

```
pvalue
```

```
## [1] 4.492017e-17
```

Using anova function

```
lm.model = lm(Sepal.Width~Species-1,data=iris)
anova.model = anova(lm.model)
summary(anova.model)
```

```
##        Df          Sum Sq           Mean Sq           F value         Pr(>F)
##  Min.   :  3   Min.   :  16.96   Min.   :  0.1154   Min.   :4083   Min.   :0
##  1st Qu.: 39   1st Qu.: 366.08   1st Qu.:117.8730   1st Qu.:4083   1st Qu.:0
##  Median : 75   Median : 715.20   Median :235.6307   Median :4083   Median :0
##  Mean   : 75   Mean   : 715.20   Mean   :235.6307   Mean   :4083   Mean   :0
##  3rd Qu.:111   3rd Qu.:1064.32   3rd Qu.:353.3883   3rd Qu.:4083   3rd Qu.:0
##  Max.   :147   Max.   :1413.44   Max.   :471.1460   Max.   :4083   Max.   :0
##                                                     NA's   :1      NA's   :1
```

For more general hypothesis testing, we may use the glht (general linear hypothesis testing) function in the multcomp package.

```
library(multcomp)
contrast1 = c(-0.5,-0.5,1)
contrast2 = c(2,-1,-1)
contrast3 = c(1,-1,0)
constrats.tmp = rbind(contrast1,contrast2,contrast3)
iris.glht = glht(lm.model,linfct = constrats.tmp,alternative="two.sided")
summary(iris.glht)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = Sepal.Width ~ Species - 1, data = iris)
##
## Linear Hypotheses:
##                Estimate Std. Error t value Pr(>|t|)
## contrast1 == 0 -0.12500    0.05884  -2.125   0.0814 .
## contrast2 == 0  1.11200    0.11767   9.450   <0.001 ***
## contrast3 == 0  0.65800    0.06794   9.685   <0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Compare with the linear model estimates

```
lm.model
```

```
##
## Call:
## lm(formula = Sepal.Width ~ Species - 1, data = iris)
##
## Coefficients:
##     Speciessetosa  Speciesversicolor   Speciesvirginica
##             3.428              2.770              2.974
```

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```