Stat 428/528: Advanced Data Analysis 2

Chapters 5: Paired Experiments and Randomized Block Experiments

March 5, 2019

ANOVA: randomized block design

A randomized block design is often used instead of a completely randomized design in studies where there is extraneous variation among the experimental units that may influence the response.

- A significant amount of the extraneous variation may be removed from the comparison of treatments by **partitioning the experimental units into fairly homogeneous subgroups or blocks.**
- If the subjects come from different groups but are fairly homogeneous within these groups, then it might make sense to use a randomized block design, where you estimate the effect of being in different groups.

  —— Blocks for medical patients could be based on say, sex, age category, or whether or not the person smokes. Subjects within each block are then randomly assigned to the possible treatments.

ANOVA: randomized block design

In a randomized block design, you estimate effects contributed by each block as well as the treatment effects.

Example (Boys Shoes): want to measure the durability of the soles A and B.
10 boys were selected at random.
—— Each boy was given a pair of shoes.
——Each pair had 1 shoe with the old sole (Sole A) and 1 shoe with the new sole (sole B).
—— For each pair of shoes, the sole type was randomly assigned to the right or left foot.

- Block: boy
- Treatment: Sole A and Sole B

Example:
There might be differences in soil fertility in different plots of land.
—— Researchers wanted the effect of fertilizer (for example) to be estimated
——Needed to account for the fact that some plots of land might have had different types of soil
——Differences in crop yield depended on both the treatment (type of fertilizer) and block (type of soil).
—— The desire is to account for the effect of the soil when estimating the effect of the fertilizer.

ANOVA: randomized block design

Beecher (1959): treatments to relieve itching.

- 10 patient volunteers, all male and between 20 and 30 years old.
- 7 treatments : 5 drugs, a placebo, and no drug to relieve itching.
- Each subject was given a different treatment on seven study days.
  —— The time ordering of the treatments was randomized across days.
  ——Time ordering is not part of the statistical analysis but is scientifically a good idea
  ——This helps reduce any accident effect due to time ordering.
- Except on the no-drug day, the subjects were given the treatment intravenously, and then itching was induced on their forearms using an effective itch stimulus called cowage.
  —— The subjects recorded the duration of itching, in seconds.
- The data are given in the table below. From left to right the drugs are: papaverine, morphine, aminophylline, pentobarbitol, tripelenamine.

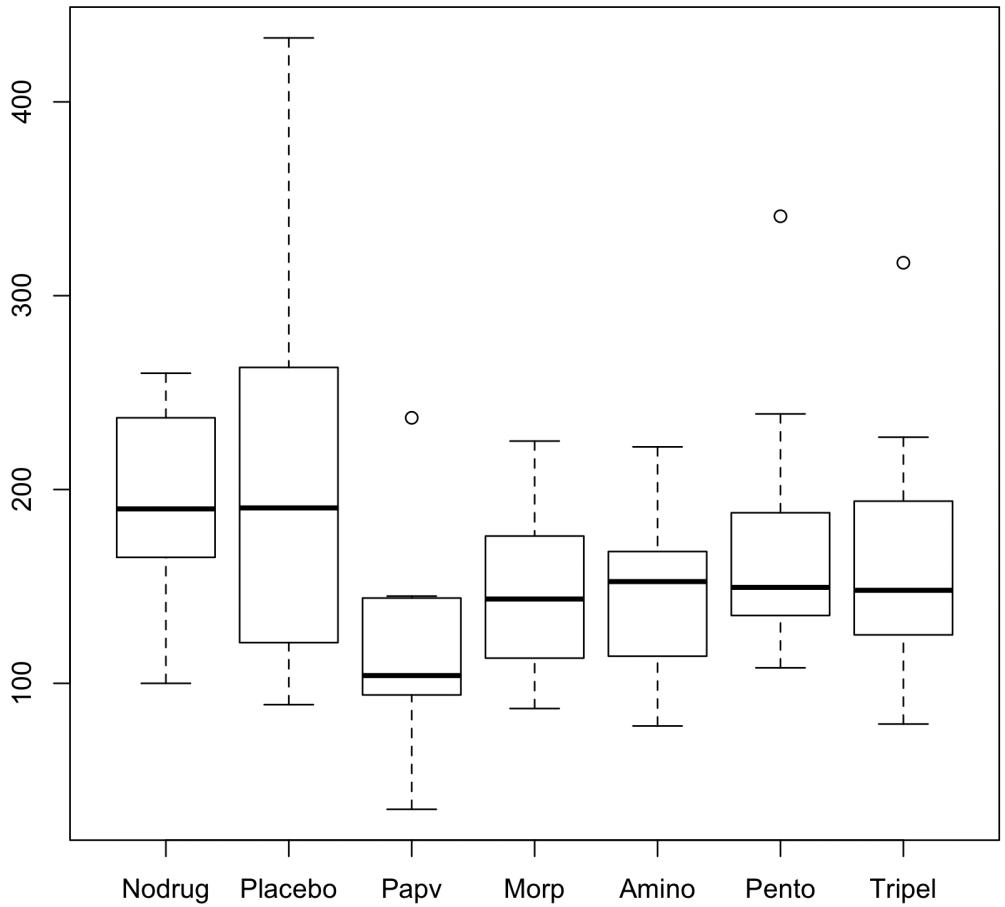
ANOVA: randomized block design

Here the subjejcts are treated as blocks because some subjects might have different mean levels of itchiness than others, and the effect of the treatments should have these differences accounted for.

Patient	Nodrug	Placebo	Papv	Morp	Amino	Pento	Tripel
1	174	263	105	199	141	108	141
2	224	213	103	143	168	341	184
3	260	231	145	113	78	159	125
4	255	291	103	225	164	135	227
5	165	168	144	176	127	239	194
6	237	121	94	144	114	136	155
7	191	137	35	87	96	140	121
8	100	102	133	120	222	134	129
9	115	89	83	100	165	185	79
10	189	433	237	173	168	188	317

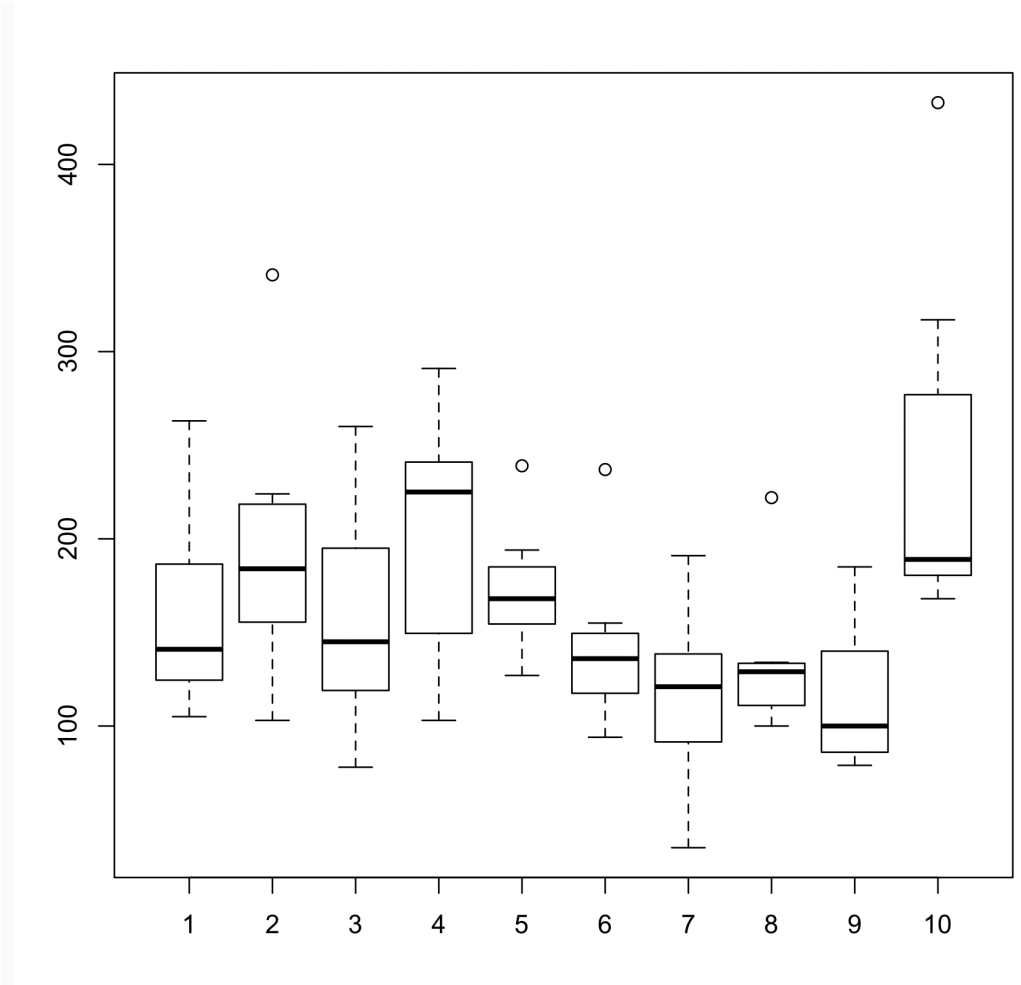
ANOVA: itching data example

Figure: boxplot of mean itching time for each treatment group



ANOVA: itching data example

Figure: boxplot of mean itching time for each individual



ANOVA: randomized block design

Now to write the model, y_{ij} again represents the jth treatment for the ith block. The model is

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

- each individual has their own mean.

We can also think of the model this way where $\mu_{ij} = \mu + \alpha_i + \beta_j$:

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

where $i = 1, 2, \cdots, I$ and $j = 1, 2, \cdots, J$

- μ is the grand mean
- α_i is the effect of block i (i.e., subject i), and β_j is the effect of treatment j.
- Less formally

$$\text{Response} = \text{Grand mean} + \text{Block effect} + \text{Treatment effect}$$

ANOVA: randomized block design

The ANOVA table can be written as follows,

- $\overline{y}_{..}$ is the mean of all IJ observations
- $\overline{y}_{i.}$ is the ith block sample mean (the average of the responses in the i th block)
- $\overline{y}_{.j}$ is the jth treatment sample mean (the average of the responses on the j th treatment)

Source	df	SS	$MS = SS/df$
Blocks	$I-1$	$J\sum_i(\overline{y}_{i.}-\overline{y}_{..})^2$	
Treatments	$J-1$	$I\sum_i(\overline{y}_{.j}-\overline{y}_{..})^2$	
Error	$(I-1)(J-1)$	$\sum_{ij}(y_{ij}-\overline{y}_{i.}-\overline{y}_{.j}+\overline{y}_{..})^2$	
Total	$IJ-1$	$\sum_{ij}(y_{ij}-\overline{y}_{..})^2$	

The MS (Mean square) column is filled in using SS/df for the same row.

ANOVA: randomized block design

Usually you are more interested in testing whether the treatment effects are 0 rather than whether the blocking effects are 0. In other words the hypothesis test of greatest interest is

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_J = 0$$

- Test statistic
$$F_{obs} = \frac{MS\ \text{Treat}}{MS\ \text{Error}}$$
- Under H_0, F statistic is distributed as F with $J-1$ numerator degrees of freedom and $(I-1)(J-1)$ denominator degrees of freedom.
- Reject H_0 if $F_{obs} > F_{crit}$

ANOVA: randomized block design

The randomized block design is used when blocks are very different but observations within blocks would be very similar if the null hypothesis of no treatment effect is true. However, you could test

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$$

using an F test based on

$$F_{obs} = \frac{MS \text{ Blocks}}{MS \text{ Error}}$$

- Under H_0, F statistic is distributed as F with $I - 1$ numerator degrees of freedom and $(I - 1)(J - 1)$ denominator degrees of freedom.
- Reject H_0 if $F_{obs} > F_{crit}$

The Block SS plus the Error SS is the Error SS from a one-way ANOVA comparing the J treatments.

- If the Block SS is large relative to the Error SS from the two-factor model, then the experimenter has eliminated a substantial portion of the variation that is used to assess the differences among the treatments.
- This leads to a more sensitive comparison of treatments than would have been obtained using a one-way ANOVA.

ANOVA: randomized block design

Under the sum constraint on the parameters

$$\sum_{i=1}^{I} \alpha_i = 0, \sum_{j=1}^{J} \beta_j = 0$$

The estimates for μ, α_i and β_j are

$$\widehat{\mu} = \overline{y}_{..}$$
$$\widehat{\alpha}_i = \overline{y}_{i.} - \overline{y}_{..}$$
$$\widehat{\beta}_j = \overline{y}_{.j} - \overline{y}_{..}$$
$$\widehat{\mu}_{ij} = \widehat{\mu} + \widehat{\alpha}_i + \widehat{\beta}_j$$

- the estimated treatment effect (for a particular treatment) is the average response for that treatment minus the overall mean
- the estimated block effect (for a particular block) is the mean response in that block (i.e., for that patient) minus the overall mean.

ANOVA: randomized block design

The model can be fitted in R.

```
> itch <- read.csv("http://statacumen.com/teach/ADA2/
ADA2_notes_Ch05_itch.csv")
> head(itch)
  Patient Nodrug Placebo Papv Morp Amino Pento Tripel
1       1    174     263  105  199   141   108    141
2       2    224     213  103  143   168   341    184
3       3    260     231  145  113    78   159    125
4       4    255     291  103  225   164   135    227
5       5    165     168  144  176   127   239    194
6       6    237     121   94  144   114   136    155
```

ANOVA: randomized block design

To analyze in R, the data should be in narrow format, with one column for the patient, one for the treatment, and one for the response.

```
> install.packages("reshape2")
> library(reshape2)
Warning message:
package ‘reshape2’ was built under R version 3.4.3
> R.Version()$version.string
[1] "R version 3.4.2 (2017-09-28)"
```

ANOVA: randomized block design

```
> itch.long  <- melt(itch
+                    , id.vars       = "Patient"
+                    , variable.name = "Treatment"
+                    , value.name    = "Seconds"
+ )
> head(itch.long)
  Patient Treatment Seconds
1       1    Nodrug     174
2       2    Nodrug     224
3       3    Nodrug     260
4       4    Nodrug     255
5       5    Nodrug     165
6       6    Nodrug     237
```

ANOVA: randomized block design

It is important to make the Patient ID a factor variable. Otherwise the patient ID is treated as quantitative!!

```
> itch.long$Patient <- factor(itch.long$Patient)
> attach(itch.long)
> model1 <- lm(Seconds ~ Patient + Treatment)
> library(car)
> Anova(model1,type=3)
Anova Table (Type III tests)

Response: Seconds
             Sum Sq Df F value    Pr(>F)
(Intercept) 155100  1 50.1133 3.065e-09 ***
Treatment    53013  6  2.8548  0.017303 *
Patient     103280  9  3.7078  0.001124 **
Residuals   167130 54
```

ANOVA: randomized block design

- Based on the output, there are significant differences among the treatments (p-value=0.017) and among patients (p-value=0.001).
  ———This suggests that it was important to take into account differences among patients.
- Now fit the model as a one-factor ANOVA (ignoring the effect of the individual patients), the evidence appears not as strong against the null hypothesis.

```
> model2 <- lm(Seconds ~ Treatment)
> Anova(model2,type=3)
Anova Table (Type III tests)

Response: Seconds
             Sum Sq Df F value    Pr(>F)
(Intercept) 364810  1 84.9935 2.709e-13 ***
Treatment    53013  6  2.0585   0.07082 .
Residuals   270409 63
```

```
> summary(model1)
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       188.286     26.598   7.079 3.07e-09 ***
TreatmentPlacebo   13.800     24.880   0.555  0.58141
TreatmentPapv     -72.800     24.880  -2.926  0.00501 **
TreatmentMorp     -43.000     24.880  -1.728  0.08965 .
TreatmentAmino    -46.700     24.880  -1.877  0.06592 .
TreatmentPento    -14.500     24.880  -0.583  0.56245
TreatmentTripel   -23.800     24.880  -0.957  0.34303
Patient2           35.000     29.737   1.177  0.24436
Patient3           -2.857     29.737  -0.096  0.92381
Patient4           38.429     29.737   1.292  0.20176
Patient5           11.714     29.737   0.394  0.69518
Patient6          -18.571     29.737  -0.625  0.53491
Patient7          -46.286     29.737  -1.557  0.12543
Patient8          -27.286     29.737  -0.918  0.36292
Patient9          -45.000     29.737  -1.513  0.13604
Patient10          82.000     29.737   2.758  0.00793 **
Multiple R-squared:  0.4832,Adjusted R-squared:  0.3397
F-statistic: 3.367 on 15 and 54 DF,  p-value: 0.00052
```

In this sort of example, predicting the reduction in seconds is probably not as interesting as learning whether the treatments were different from each other, and which treatments were most effective.

From the linear model output, we also get an F test with a p-value, which is a p-value for testing whether both variables together (blocks and treatments) are significantly different from 0. This is usually not as interesting as testing whether just treatments are different from each other taking blocks into account.

In order to test which treatments are significantly different from each other, we should take into account that we are doing multiple comparisons. The package `multcomp` can be used to help do multiple comparisons.

```
> install.packages("multcomp")
> library(multcomp)
> comp.itch <- glht(aov(model1),linfct =
  mcp(Treatment = "Tukey"))
> summary(comp.itch)
```

	Estimate	Std. Error	t value	Pr(>|t|)	
Placebo - Nodrug == 0	13.80	24.88	0.555	0.9978	
Papv - Nodrug == 0	-72.80	24.88	-2.926	0.0697	.
Morp - Nodrug == 0	-43.00	24.88	-1.728	0.6005	
Amino - Nodrug == 0	-46.70	24.88	-1.877	0.5039	
Pento - Nodrug == 0	-14.50	24.88	-0.583	0.9971	
Tripel - Nodrug == 0	-23.80	24.88	-0.957	0.9610	
Papv - Placebo == 0	-86.60	24.88	-3.481	0.0165	*
Morp - Placebo == 0	-56.80	24.88	-2.283	0.2712	
Amino - Placebo == 0	-60.50	24.88	-2.432	0.2052	
Pento - Placebo == 0	-28.30	24.88	-1.137	0.9135	
Tripel - Placebo == 0	-37.60	24.88	-1.511	0.7370	
Morp - Papv == 0	29.80	24.88	1.198	0.8920	
Amino - Papv == 0	26.10	24.88	1.049	0.9398	
Pento - Papv == 0	58.30	24.88	2.343	0.2434	
Tripel - Papv == 0	49.00	24.88	1.969	0.4454	
Amino - Morp == 0	-3.70	24.88	-0.149	1.0000	
Pento - Morp == 0	28.50	24.88	1.146	0.9107	
Tripel - Morp == 0	19.20	24.88	0.772	0.9867	
Pento - Amino == 0	32.20	24.88	1.294	0.8516	
Tripel - Amino == 0	22.90	24.88	0.920	0.9676	

Based on the output, the only comparison that is statistically significant at the .05 level is papaverine versus placebo (with p-value of 0.0165),
———-suggests that papaverine induces a lower mean itching time than placebo.

The second lowest adjusted p-value is for papaverine versus no drug (with p-value of 0.0697).
———-This suggests that there is some (but not overwhelming) evidence that this drug reduced itchiness.

The p-value for the Tukey multiple comparisons is based on the Tukey range distribution, which is similar to a t-test but results in different p-values.

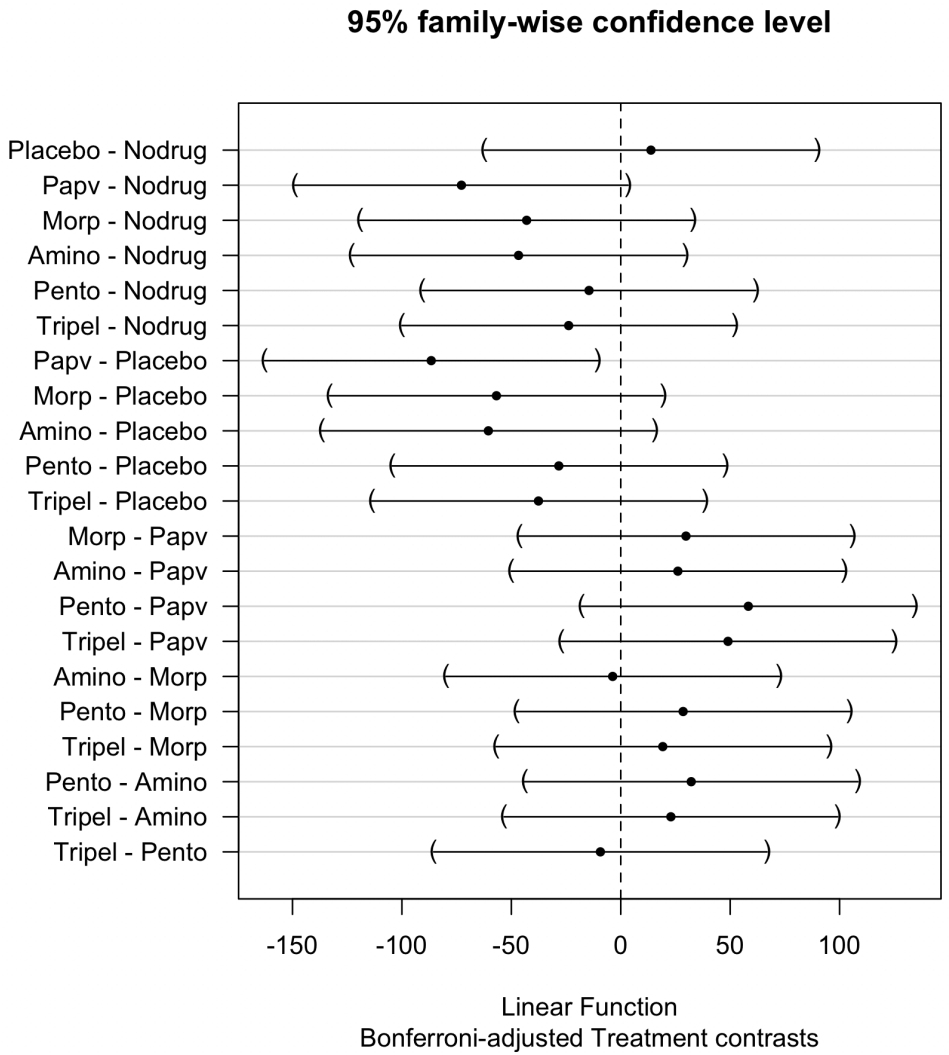
You could also do a Bonferroni correction instead.

```
summary(comp.itch, test = adjusted("bonferroni"))
```

	Estimate	Std. Error	t value	Pr(>|t|)	
Placebo - Nodrug == 0	13.80	24.88	0.555	1.000	
Papv - Nodrug == 0	-72.80	24.88	-2.926	0.105	
Morp - Nodrug == 0	-43.00	24.88	-1.728	1.000	
Amino - Nodrug == 0	-46.70	24.88	-1.877	1.000	
Pento - Nodrug == 0	-14.50	24.88	-0.583	1.000	
Tripel - Nodrug == 0	-23.80	24.88	-0.957	1.000	
Papv - Placebo == 0	-86.60	24.88	-3.481	0.021	*
Morp - Placebo == 0	-56.80	24.88	-2.283	0.554	
Amino - Placebo == 0	-60.50	24.88	-2.432	0.386	
Pento - Placebo == 0	-28.30	24.88	-1.137	1.000	
Tripel - Placebo == 0	-37.60	24.88	-1.511	1.000	
Morp - Papv == 0	29.80	24.88	1.198	1.000	
Amino - Papv == 0	26.10	24.88	1.049	1.000	
Pento - Papv == 0	58.30	24.88	2.343	0.479	
Tripel - Papv == 0	49.00	24.88	1.969	1.000	
Amino - Morp == 0	-3.70	24.88	-0.149	1.000	
Pento - Morp == 0	28.50	24.88	1.146	1.000	
Tripel - Morp == 0	19.20	24.88	0.772	1.000	
Pento - Amino == 0	32.20	24.88	1.294	1.000	

You can also plot confidence intervals for the differences between treatments as follows.

```
# plot the summary
op <- par(no.readonly = TRUE) # the whole list of settable par's.
# make wider left margin to fit contrast labels
par(mar = c(5, 10, 4, 2) + 0.1) # order is
c(bottom, left, top, right)
# plot bonferroni-corrected difference intervals
plot(summary(comp.itch, test = adjusted("bonferroni"))
    , sub="Bonferroni-adjusted Treatment contrasts")
par(op) # reset plotting options
```


95% family-wise confidence level
Placebo - Nodrug
Papv - Nodrug
Morp - Nodrug
Amino - Nodrug
Pento - Nodrug
Tripel - Nodrug
Papv - Placebo
Morp - Placebo
Amino - Placebo
Pento - Placebo
Tripel - Placebo
Morp - Papv
Amino - Papv
Pento - Papv
Tripel - Papv
Amino - Morp
Pento - Morp
Tripel - Morp
Pento - Amino
Tripel - Amino
Tripel - Pento
-150
-100
-50
0
50
100
Linear Function
Bonferroni-adjusted Treatment contrasts

The Bonferroni comparisons for Treatment suggest that

- Papaverine induces a lower mean itching time than placebo.
- All the other comparisons of treatments are insignificant.
- The comparison of Patient blocks is of less interest.

ANOVA: diagnostics

Part of an ANOVA or regression should ideally be diagnostic tests (although these are often not mentioned in scientific studies).

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

where $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$

ANOVA: diagnostics

- Typically, diagnostics are done visually and nor very formally, especially by examining residuals.
- In the itchiness study, there is only one observation for each combination of predictors/factors, so the normality would be impossible to assess looking at each combination of predictors separately.
- The residuals should all come from the same distribution, so there is still information in the residuals regarding the normality and constant variance assumptions.

If the `plot()` function is given saved model output, it will automatically generate diagnostic plots. For example

```
par(mfrow=c(1,3))
plot(lm.s.t.p, which = c(1,4,6))

par(mfrow=c(1,3))
plot(itch.long$Treatment, lm.s.t.p$residuals,
main="Residuals vs Treatment")
  # horizontal line at zero
  abline(h = 0, col = "gray75")

plot(itch.long$Patient, lm.s.t.p$residuals,
main="Residuals vs Patient")
  # horizontal line at zero
  abline(h = 0, col = "gray75")

# Normality of Residuals
library(car)
qqPlot(lm.s.t.p$residuals, las = 1, main="QQ Plot")
```

Figure: Diagnostic Plots 2

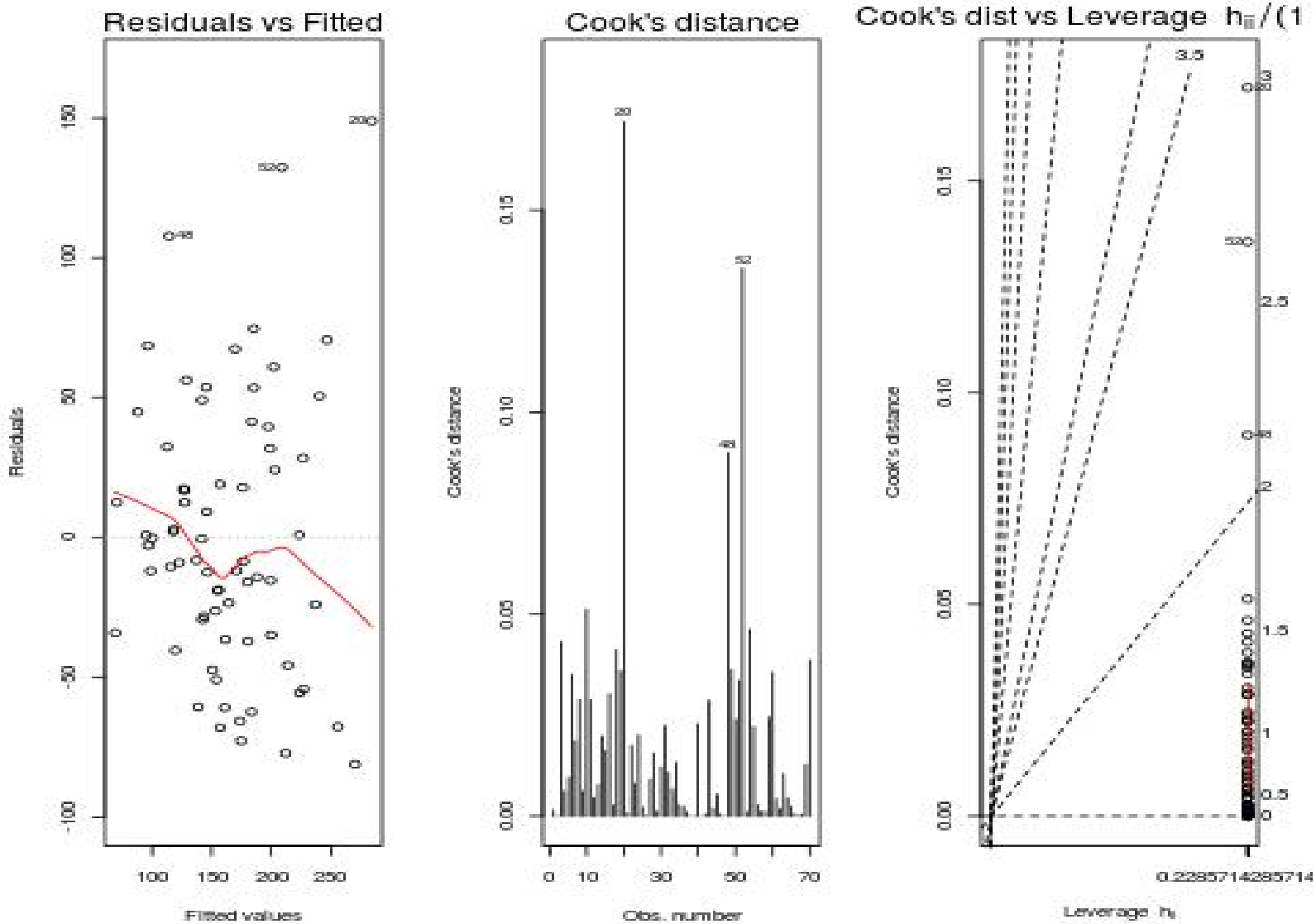
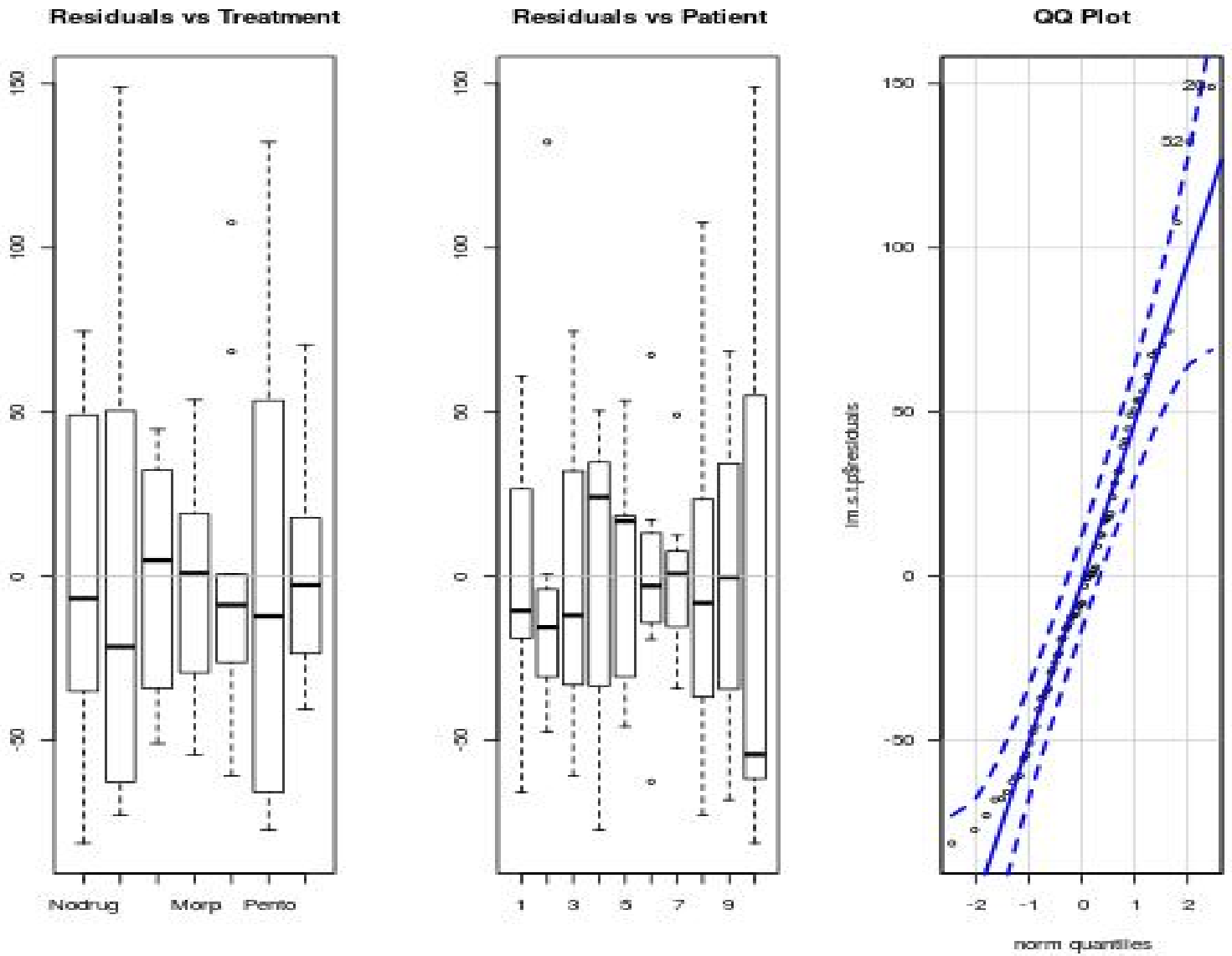


Figure: Diagnostic Plots 1



- The normal quantile (or QQ-plot) shows the residual distribution is slightly skewed to the right, in part, due to three cases that are not fitted well by the model (the outliers in the boxplots).
- These three cases are also the most influential cases (from Cooks distance)
- The plot of the studentized residuals against fitted values shows no obvious pattern.

ANOVA

Friedman Test

A non-parametric alternative to ANOVA in this situation (ANOVA with one treatment, and one blocking variable, and no replication within treatment-block combinations), you can use the Friedman Test (named after economist Milton Friedman),

—— similar to the Kruskal-Wallis test for one-way ANOVA.

More general tests for dealing with ANOVA alternatives based on rank are called Durbin tests.

```
# Friedman test for differences between groups
conditional on blocks.
#   The formula is of the form y ~ t | b,
#     where y, t and b give the data values (y)
#and corresponding groups/treatment (t) and blocks (b),
 respectively.
friedman.test(Seconds ~ Treatment | Patient,
data = itch.long)
Friedman rank sum test

data:  Seconds and Treatment and Patient
Friedman chi-squared = 14.887, df = 6, p-value = 0.02115
```

- Note that the syntax (and test) distinguishes blocks from treatments. Here you condition on the blocks (patients) $t|b$.
  The null hypothesis is that apart from an effect of blocks, the location parameter of y is the same in each of the groups.
- If you swap Treatment and Patient variables, then you are testing whether patients differ from each other, controlling for type of medication. This would also result in a statistically significant test (p-value $= .01$).
- The output suggests significant differences among treatments with a p-value of 0.02115, which supports the earlier conclusion.

ANOVA with two factors and replication

Generally, ANOVA can be run with more than two factors,

- some of which might be considered blocking variables (meaning we want to control for them),
- or we might be interested in all factors.
- often experiments are done with replication for different combinations of treatments.
  —— this is usually preferable to just having one observation for each combination (it is more data and allows better estimates of variability).

ANOVA with two factors and replication

Consider an experiment on beetles' survival time under different insecticides and doses

- Four different insecticides and three different doses (low, medium, high) are interested
  ——There are twelve combinations
  ——Suppose each combination is replicated four times, which results in 48 observations.
- Response: the survival time of the beetles.
  —— time is measured in fractions of a 10 minute interval. (So 0.4 means 4 minutes.)
- The doses of high, medium, and low, are really ordinal (we don't know if they are equally spaced, for example, but they can be ranked)
  —— ANOVA will treat them as qualitative, like having three different brands without knowing the rankings.

ANOVA with two factors and replication

```
beetles <- read.table("http://statacumen.com/teach/ADA2
/ADA2_notes_Ch05_beetles.dat"
                      , header = TRUE)
> beetles
     dose insecticide   t1   t2   t3   t4
1     low           A 0.31 0.45 0.46 0.43
2     low           B 0.82 1.10 0.88 0.72
3     low           C 0.43 0.45 0.63 0.76
4     low           D 0.45 0.71 0.66 0.62
5  medium           A 0.36 0.29 0.40 0.23
6  medium           B 0.92 0.61 0.49 1.24
7  medium           C 0.44 0.35 0.31 0.40
8  medium           D 0.56 1.02 0.71 0.38
9    high           A 0.22 0.21 0.18 0.23
10   high           B 0.30 0.37 0.38 0.29
11   high           C 0.23 0.25 0.24 0.22
12   high           D 0.30 0.36 0.31 0.33
```

ANOVA with two factors and replication

```
# make dose a factor variable and label the levels
beetles$dose <- factor(beetles$dose,
 labels = c("low","medium","high"))
> beetles$dose
 [1] low    low    low    low    medium medium medium
 medium high   high   high   high
Levels: low medium high
```

ANOVA with two factors and replication

As usual, we need to reshape the data into the long format. Here the columns should be dose, insecticide, and replicate.

```
library(reshape2)
beetles.long <- melt(beetles
                     , id.vars       = c("dose", "insecticide")
                     , variable.name = "number"
                     , value.name    = "hours10"
                )
str(beetles.long)
> str(beetles.long)
'data.frame': 48 obs. of  4 variables:
 $ dose       : Factor w/ 3 levels "low","medium",..: 1 1 1 1 2 2
 $ insecticide: Factor w/ 4 levels "A","B","C","D": 1 2 3 4 1 2 3
 $ number     : Factor w/ 4 levels "t1","t2","t3",..: 1 1 1 1 1 1
 $ hours10    : num  0.31 0.82 0.43 0.45 0.36 0.92 0.44 0.56 0.22
```

ANOVA with two factors and replication

```
> beetles.long
> head(beetles.long)
    dose insecticide number hours10
1    low           A     t1    0.31
2    low           B     t1    0.82
3    low           C     t1    0.43
4    low           D     t1    0.45
5 medium           A     t1    0.36
6 medium           B     t1    0.92
```

ANOVA with two factors and replication

```
> beetles.mean.di
     dose insecticide      m
1     low           A 0.4125
2     low           B 0.8800
3     low           C 0.5675
4     low           D 0.6100
5  medium           A 0.3200
6  medium           B 0.8150
7  medium           C 0.3750
8  medium           D 0.6675
9    high           A 0.2100
10   high           B 0.3350
11   high           C 0.2350
12   high           D 0.3250
```

ANOVA with two factors and replication

Balanced ANOVA examples have an advantage in interpretation

- marginal is calculated by average of the averages. For example, the average of low doses 0.618 is the average of the averages for each combination of low dose and insecticide $(0.413 + 0.880 + 0.568 + 0.610)/4 = 0.61775$.

Cell Means	Dose			
Insecticide	1	2	3	Insect marg
A	0.413	0.320	0.210	0.314
B	0.880	0.815	0.335	0.677
C	0.568	0.375	0.235	0.393
D	0.610	0.668	0.325	0.534
Dose marg	0.618	0.544	0.277	0.480

ANOVA with two factors and replication

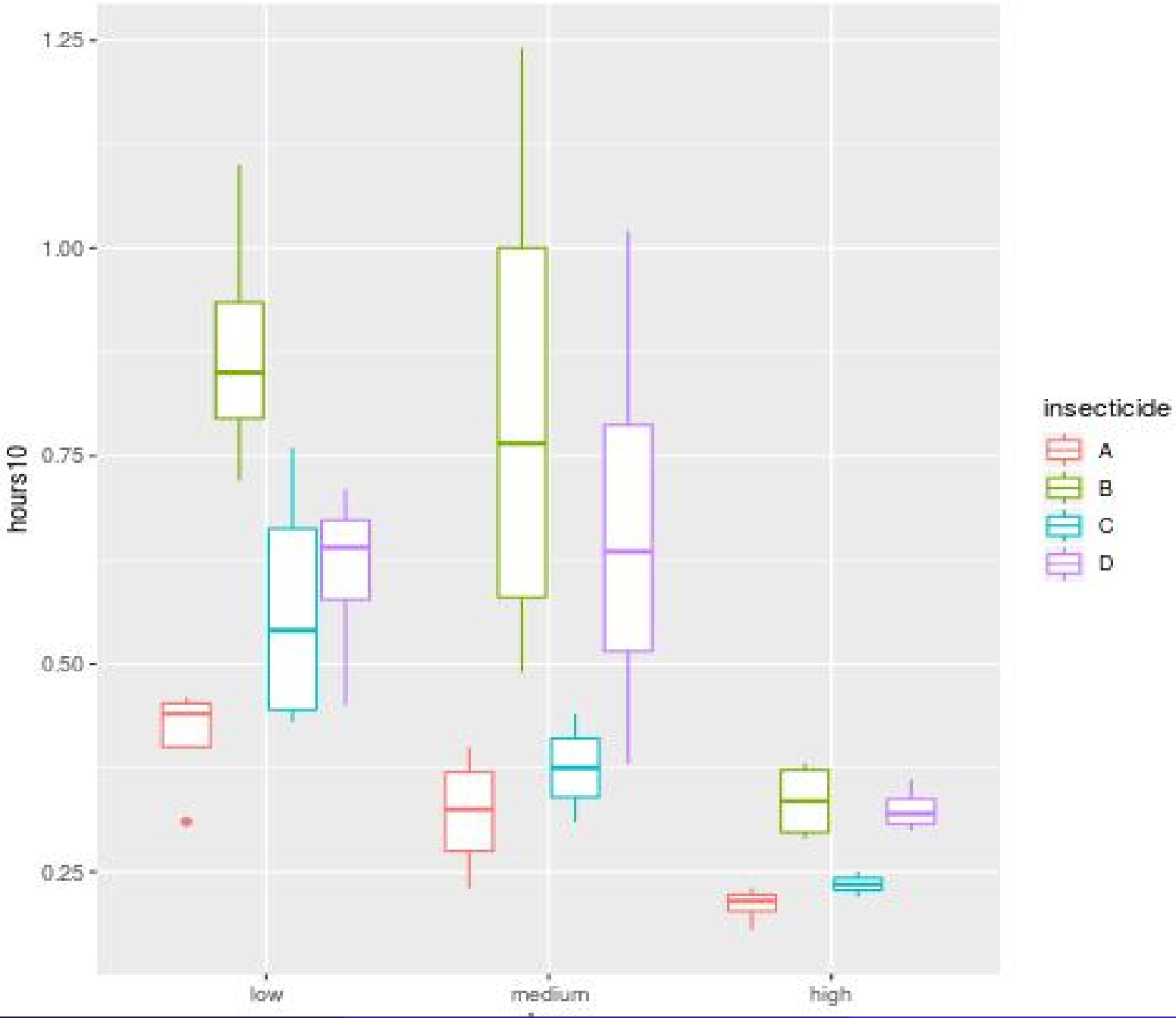
- looking at the margins, the survival time was lowest for insecticides A and C.
- Higher doses also lead to lower survival times on average
- the survival times are not equally spaced—the difference in average survival times between doses 3 versus 2 is larger than for doses 2 versus 1

Cell Means	Dose			
Insecticide	1	2	3	Insect marg
A	0.413	0.320	0.210	0.314
B	0.880	0.815	0.335	0.677
C	0.568	0.375	0.235	0.393
D	0.610	0.668	0.325	0.534
Dose marg	0.618	0.544	0.277	0.480

You can do boxplots for looking at the responses for combinations of predictors.

```
library(ggplot2)
  p <- ggplot(beetles.long, aes(x = dose, y = hours10,
  colour = insecticide))
  p <- p + geom_boxplot()
  print(p)
```

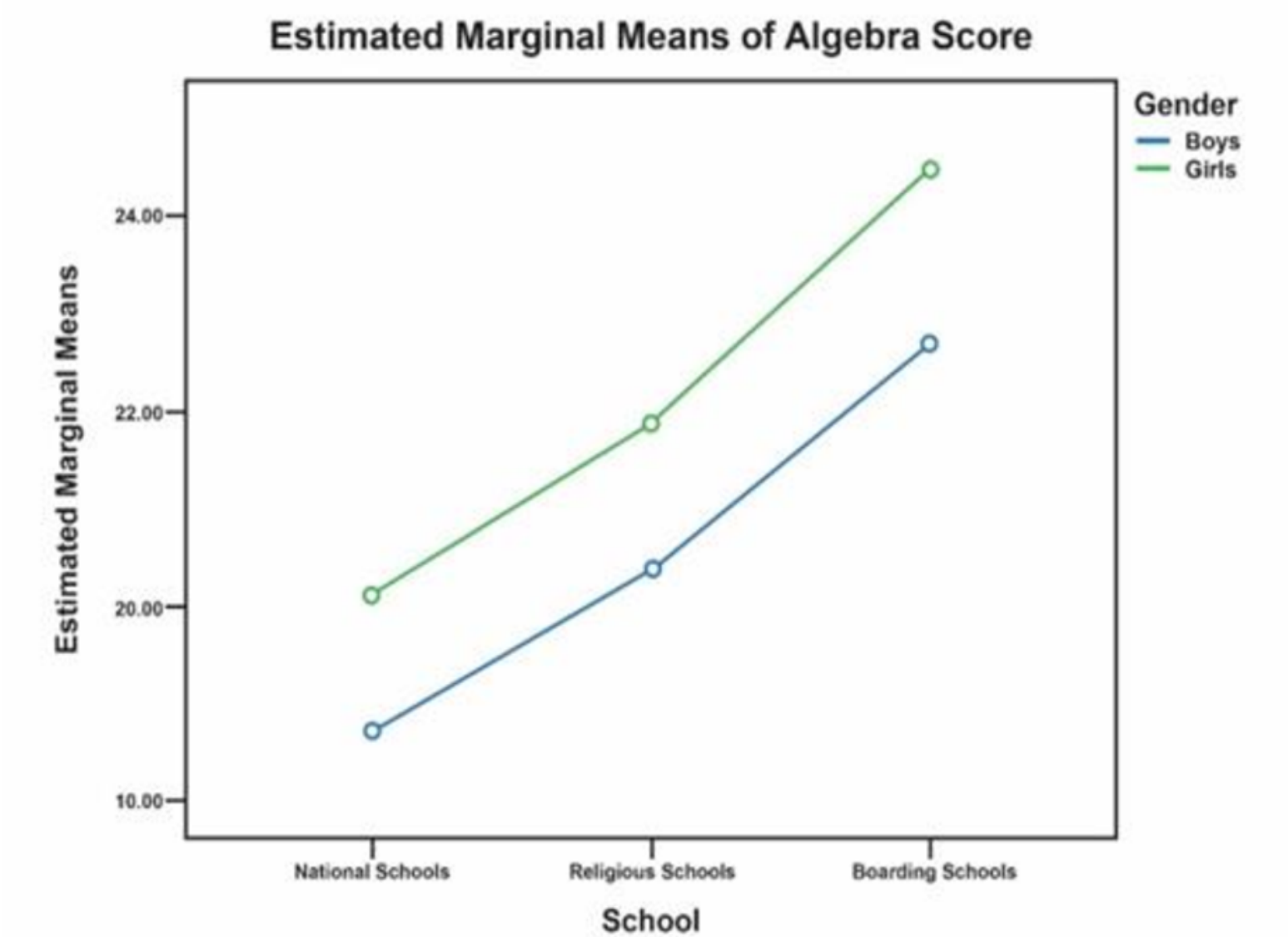
It looks like there are problems with the equal variances assumption! To make the assumptions not so badly violated, one possibility is to transform the data, such as using log of the survival times.


1.25
1.00
0.75
0.50
0.25
hours10
low
medium
high
insecticide
A
B
C
D

Interactions

To understand interaction, suppose you (conceptually) plot the means in each row of the population table, giving what is known as the population mean profile plot. In practice, we plot the sample mean profile plot.

No interaction is present: if the plot has perfectly parallel F1 profiles, as in the plot below for a 2×3 experiment. The levels of F1 and F2 do not interact.

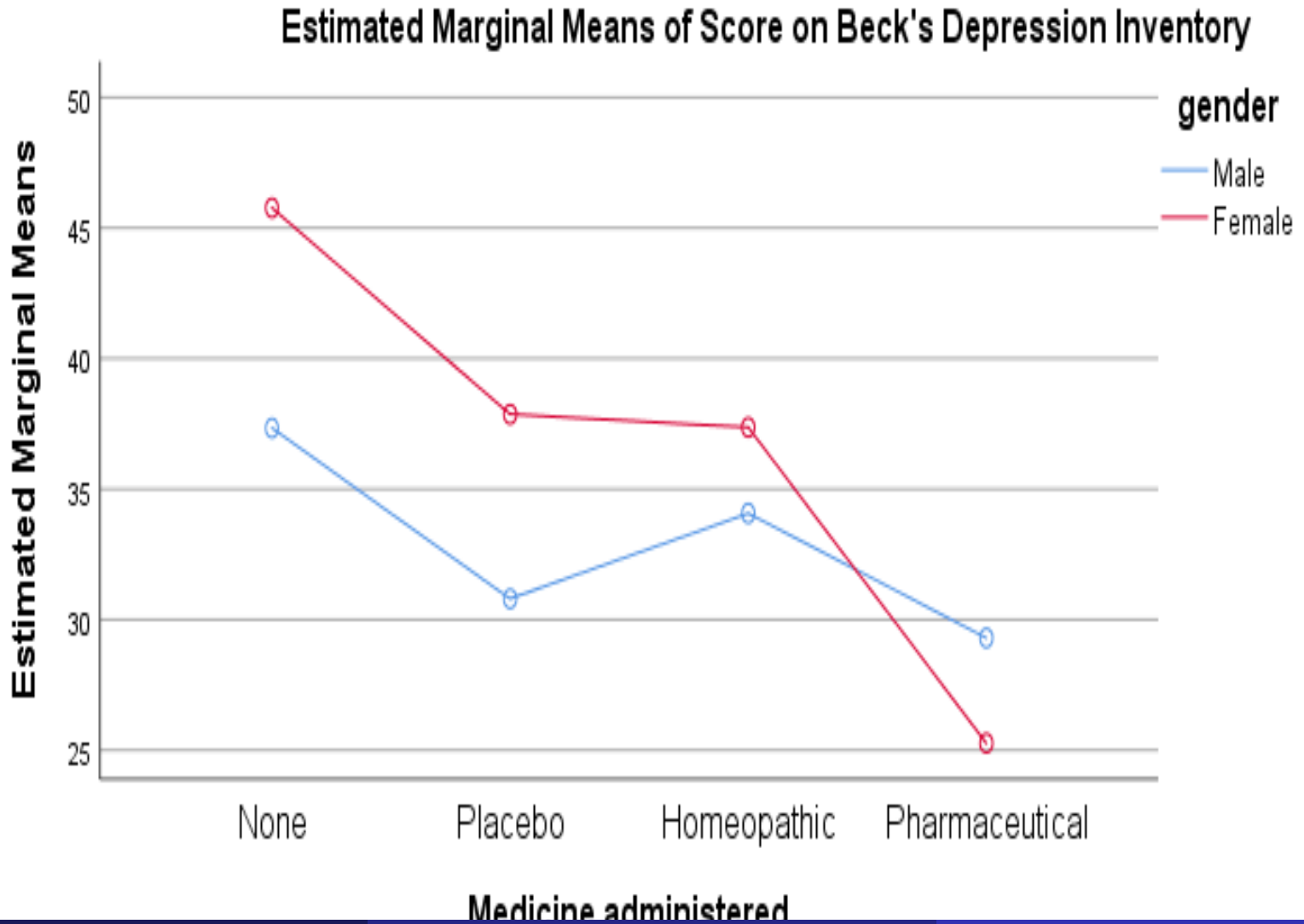


Parallel profiles

- $\mu_{ij} - \mu_{hj}$ is independent of j for each i and h
  ——-difference between levels of F1 does not depend on level of F2
- $$\mu_{ij} - \bar{\mu}_{i\cdot} = \mu_{hj} - \bar{\mu}_{h\cdot} \text{ for all } i, j, h$$
- $$\mu_{ij} - \bar{\mu}_{i\cdot} = \bar{\mu}_{\cdot j} - \bar{\mu}_{\cdot\cdot} \text{ for all } i, j$$
- $$\mu_{ij} - \bar{\mu}_{i\cdot} - \bar{\mu}_{\cdot j} + \bar{\mu}_{\cdot\cdot} \text{ for all } i, j$$

  ——interaction effect $(\alpha\beta)_{ij} = 0$ for all i, j

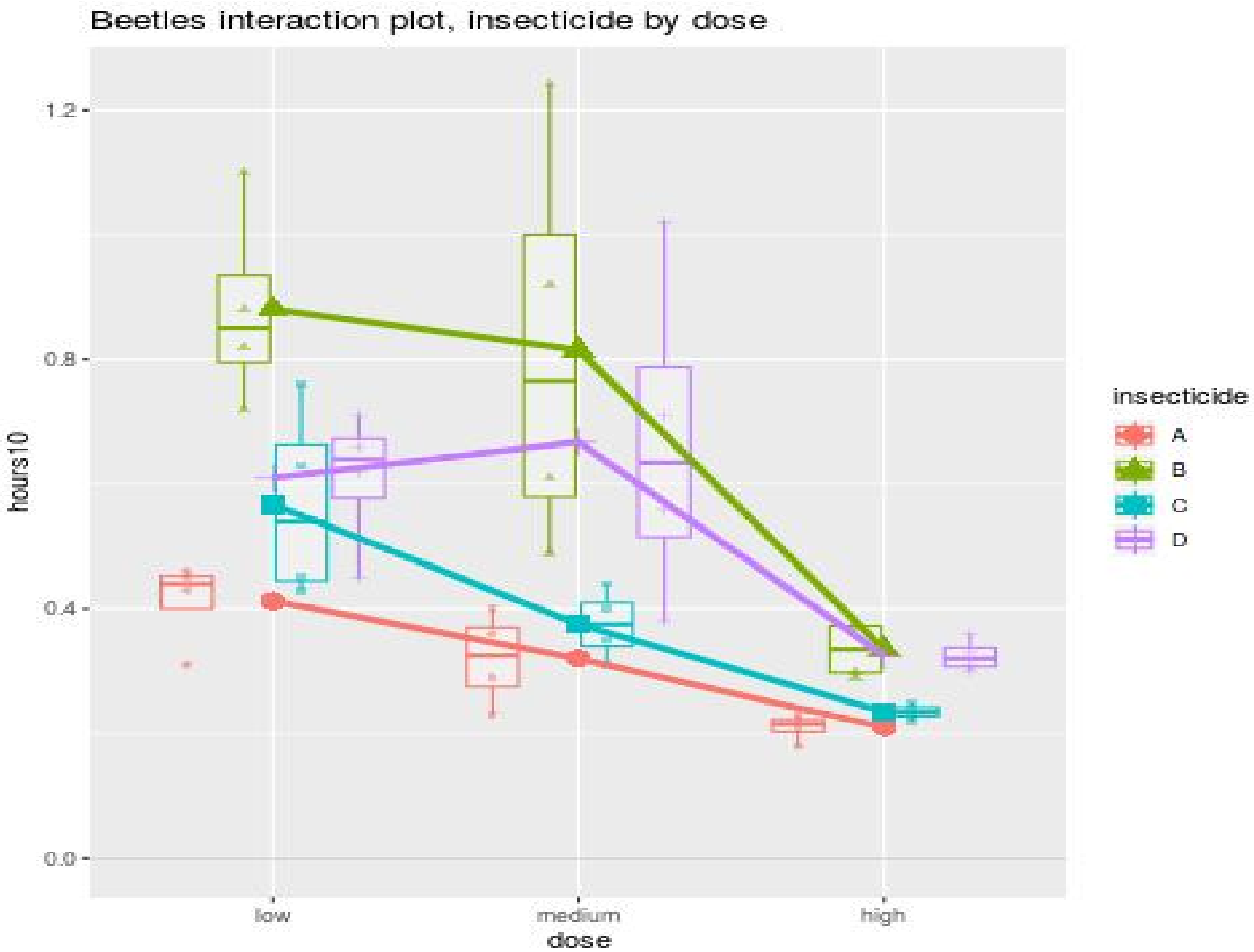
Interaction is present if the profiles are not perfectly parallel. An example of a profile plot for two-factor experiment (2×4) with interaction is given below.

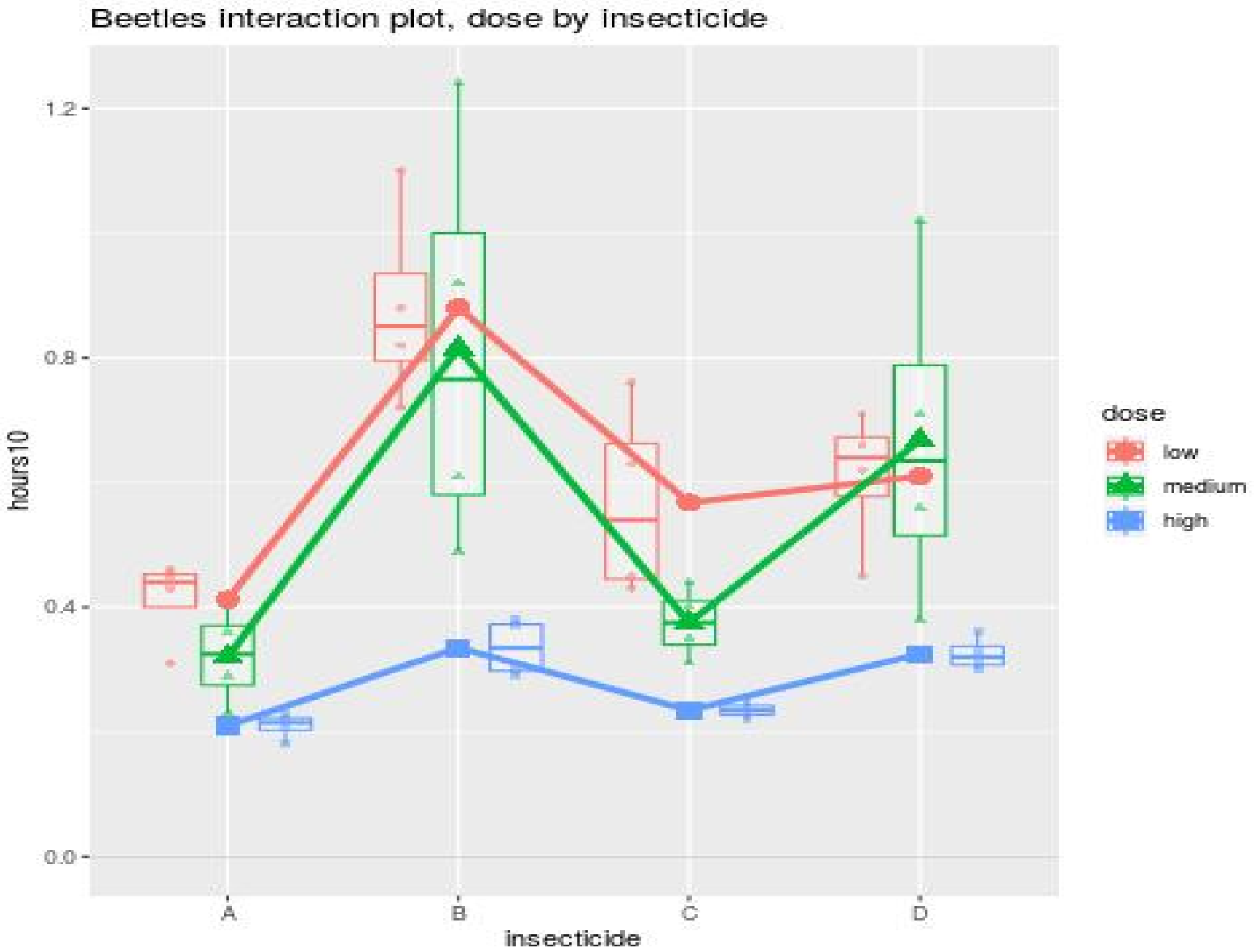


Comments on interactions:

- The roles of F1 and F2 can be reversed in the profile plots without changing the assessment of a presence or absence of interaction.
  ——It is often helpful to view the interaction plot from both perspectives.
- A qualitative check for interaction can be based on the sample means profile plot,
  ——but keep in mind that profiles of sample means are never perfectly parallel even when the factors do not interact in the population.
  ——The Interaction SS measures the extent of non-parallelism in the sample mean profiles.

Profile plots




Beetles interaction plot, dose by insecticide
hours10
1.2
0.8
0.4
0.0
A
B
C
D
insecticide
dose
low
medium
high

- The profile plots indicate that the main effects are significant
  ——the insecticides have noticeably different mean survival times averaged over doses, with insecticide A having the lowest mean survival time averaged over doses.
  ——higher doses tend to produce lower survival times.
- Interaction seems not significant.

Looking back at the table of cell means (slide 45), the idea is the differences between columns are similar, and the differences between rows are similar.

—— For example, going from dose 1 to dose 2 (low to medium), the change in average survival for insecticide A is (0.413-0.320) = 0.093 (i.e., .93 minutes or 55 seconds), and the difference for insecticide B is (0.880 - 0.815) = 0.712 (i.e., 39 seconds). Given the variability in the data, the change going from low to medium doses is similar for insecticides A and B.

Cell Means	Dose			
Insecticide	1	2	3	Insect marg
A	0.413	0.320	0.210	0.314
B	0.880	0.815	0.335	0.677
C	0.568	0.375	0.235	0.393
D	0.610	0.668	0.325	0.534
Dose marg	0.618	0.544	0.277	0.480

ANOVA with interaction

Consider a balanced two-factor experiment with K responses at each combination of the I levels of factor 1 (F1) with the J levels of factor 2 (F2).

- The total number of responses is KIJ, or K times the IJ treatment combinations.
- Let Y_{ijk} be the k th response at the i th level of F1 and the j th level of F2.
- A generic model for the experiment expresses Y_{ijk} as a mean response plus an error term:

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk}$$

  where μ_{ij} is the population mean response for the treatment defined by the i th level of F1 combined with the j th level of F2.
- The responses within and across treatment groups are assumed to be independent, normally distributed, and have constant variance.

The interaction model expresses the population means as

$$\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

——μ is a grand mean
——- α_i is the effect for the i th level of F1
——β_j is the effect for the jth level of F2
——-$(\alpha\beta)_{ij}$ is the interaction between the i th level of F1 and the j th level of F2.

- The interaction model is often written

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

where $i = 1, 2, \cdots, I, j = 1, 2, \cdots, J$ and $k = 1, 2, \cdots, K$.

- Informally,

Response = Grand mean+F1 effect+F2 effect+F1-by-F2 interaction+residual.

- The model with no interaction is called an additive model or main effects model, and is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

Population means

	Level of F2				
Level of F1	1	2	$\cdots$	J	F1 marg
1	μ_{11}	μ_{12}		μ_{1J}	$\bar{\mu}_{1\cdot}$
2	μ_{21}	μ_{22}		μ_{2J}	$\bar{\mu}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I	μ_{I1}	μ_{I2}		μ_{IJ}	$\bar{\mu}_{I\cdot}$
F2 marg	$\bar{\mu}_{\cdot 1}$	$\bar{\mu}_{\cdot 2}$	$\cdots$	$\bar{\mu}_{\cdot J}$	$\bar{\mu}_{\cdot\cdot}$

ANOVA with interaction

Under restrictions

$$\sum_{i=1}^{I} \alpha_i = 0, \sum_{j=1}^{J} \beta_j = 0, \sum_{i=1}^{I} \sum_{j=1}^{J} (\alpha\beta)_{ij} = 0$$

The main effects and interaction effects can be estimated as follows:

$$\begin{aligned}
\hat{\mu} &= \bar{y}_{..} \quad \text{the estimated grand mean} \\
\hat{\alpha}_i &= \bar{y}_{i.} - \bar{y}_{..} \quad \text{the estimated F1 effect } i = 1, 2, \ldots, I \\
\hat{\beta}_j &= \bar{y}_{.j} - \bar{y}_{..} \quad \text{the estimated F2 effect } j = 1, 2, \ldots, J \\
\widehat{(\alpha\beta)}_{ij} &= \bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..} \quad \text{the estimated cell interaction}
\end{aligned}$$

$$\bar{y}_{ij} = \frac{1}{K}\sum_{k} y_{ijk}$$

$$\bar{y}_{i\cdot} = \frac{1}{J}\sum_{c} \bar{y}_{ic}$$

$$\bar{y}_{\cdot j} = \frac{1}{I}\sum_{r} \bar{y}_{rj}$$

$$\bar{y}_{\cdot\cdot} = \frac{1}{IJ}\sum_{i,j} \bar{y}_{ij} = \frac{1}{I}\sum_{i} \bar{y}_{i\cdot} = \frac{1}{J}\sum_{j} \bar{y}_{\cdot j}$$

ANOVA with interaction

The ANOVA table (in the balanced case) is as follows:

Source	df	SS	MS
F1	$I-1$	$KJ\sum_i(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2$	MS F1=SS/df
F2	$J-1$	$KI\sum_j(\bar{y}_{\cdot j}-\bar{y}_{\cdot\cdot})^2$	MS F2=SS/df
Interaction	$(I-1)(J-1)$	$K\sum_{ij}(y_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}+\bar{y}_{\cdot\cdot})^2$	MS Inter=SS/df
Error	$IJ(K-1)$	$(K-1)\sum_{ij}s_{ij}^2$	MSE=SS/df
Total	$IJK-1$	$\sum_{ijk}(y_{ijk}-\bar{y}_{\cdot\cdot})^2$.	

Test of no F1 effect

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$$

using an F test based on

$$F_{obs} = \frac{MS \text{ F1}}{MS \text{ Error}}$$

- Under H_0, F statistic is distributed as F with $I - 1$ numerator degrees of freedom and $IJ(K - 1)$ denominator degrees of freedom.
- Reject H_0 if $F_{obs} > F_{crit}$
  ——H_0 is rejected when the F1 marginal means $\bar{y}_{i\cdot}$ vary significantly relative to the within sample variation. Equivalently, H_0 is rejected when the sum of squared F1 effects (between sample variation) is large relative to the within sample variation.

Test of no F2 effect

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_J = 0$$

- Test statistic
$$F_{obs} = \frac{MS\text{F2}}{MS\ \text{Error}}$$
- Under H_0, F statistic is distributed as F with $J-1$ numerator degrees of freedom and $IJ(K-1)$ denominator degrees of freedom.
- Reject H_0 if $F_{obs} > F_{crit}$

Test of no interaction

$$H_0 : (\alpha\beta)_{11} = \cdots = (\alpha\beta)_{IJ} = 0$$

- Test statistic

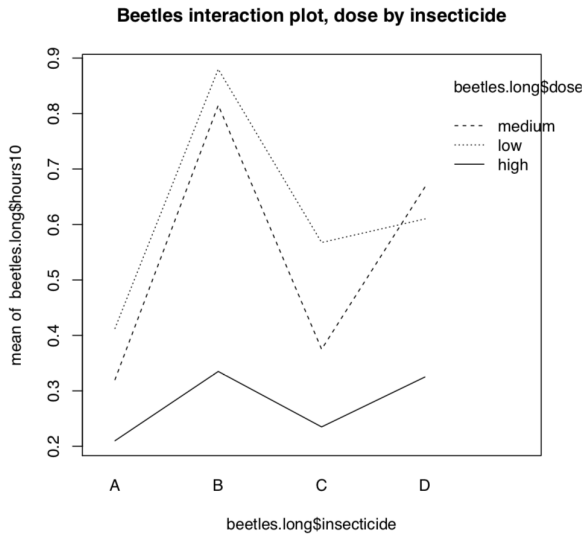
$$F_{obs} = \frac{MS\text{Interact}}{MS\ \text{Error}}$$

- Under H_0, F statistic is distributed as F with $(I-1)(J-1)$ numerator degrees of freedom and $IJ(K-1)$ denominator degrees of freedom.
- Reject H_0 if $F_{obs} > F_{crit}$

ANOVA with interaction

When there are two factors, it is possible that the effect of one factor depends on the value of the other factor. For this example, this could mean that the effect of the dose depends on the insecticide.

```
> lm.h.d.i.di <- lm(hours10 ~ dose + insecticide +
dose*insecticde)
> Anova(lm.h.d.i.di,type=3)
Anova Table (Type III tests)

Response: hours10
                  Sum Sq Df F value    Pr(>F)
(Intercept)      0.68063  1 30.6004 2.937e-06 ***
dose             0.08222  2  1.8482 0.1721570
insecticide      0.45395  3  6.8031 0.0009469 ***
dose:insecticide 0.25014  6  1.8743 0.1122506
Residuals        0.80072 36
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```




```
interaction.plot(beetles.long$dose, beetles.long$insecticide,
beetles.long$hours10 , main = "insecticide by dose")
interaction.plot(beetles.long$insecticide, beetles.long$dose,
beetles.long$hours10, main = "dose by insecticide")
```

ANOVA with interaction

The idea behind the plots is that we can see whether the effect of the insecticide depends on the dose, or similarly, whether the effect of the dose depends on the insecticide.

- In the left plot on the previous slide, there is a rank ordering of insecticides based on survival times.
  ——–Here lower survival times means a more effective insecticide, and for each dose, we appear to have that insecticide A has the lowest survival time, followed by C, then followed by D, and finally B.
- If there were a strong interaction between dose and insecticide, you might find that one insecticide is the most effective at low doses, while another is the the most effective at higher doses. In this case, the rank ordering of insecticides doesn't change much.

ANOVA with interaction

A statistical test for interaction is testing whether the lines in the interaction plot are parallel, taking into account variability in the data.
——This does not necessarily mean that the lines are straight, but that the spacing in between the lines doesn't change significantly from level to level of the factor on the horizontal axis.
—— An interaction can show up in the interaction plots either by curves crossing or by being significantly non-parallel.

Since the interaction is not significant, we'll drop the interaction term and fit the additive model with main effects only.

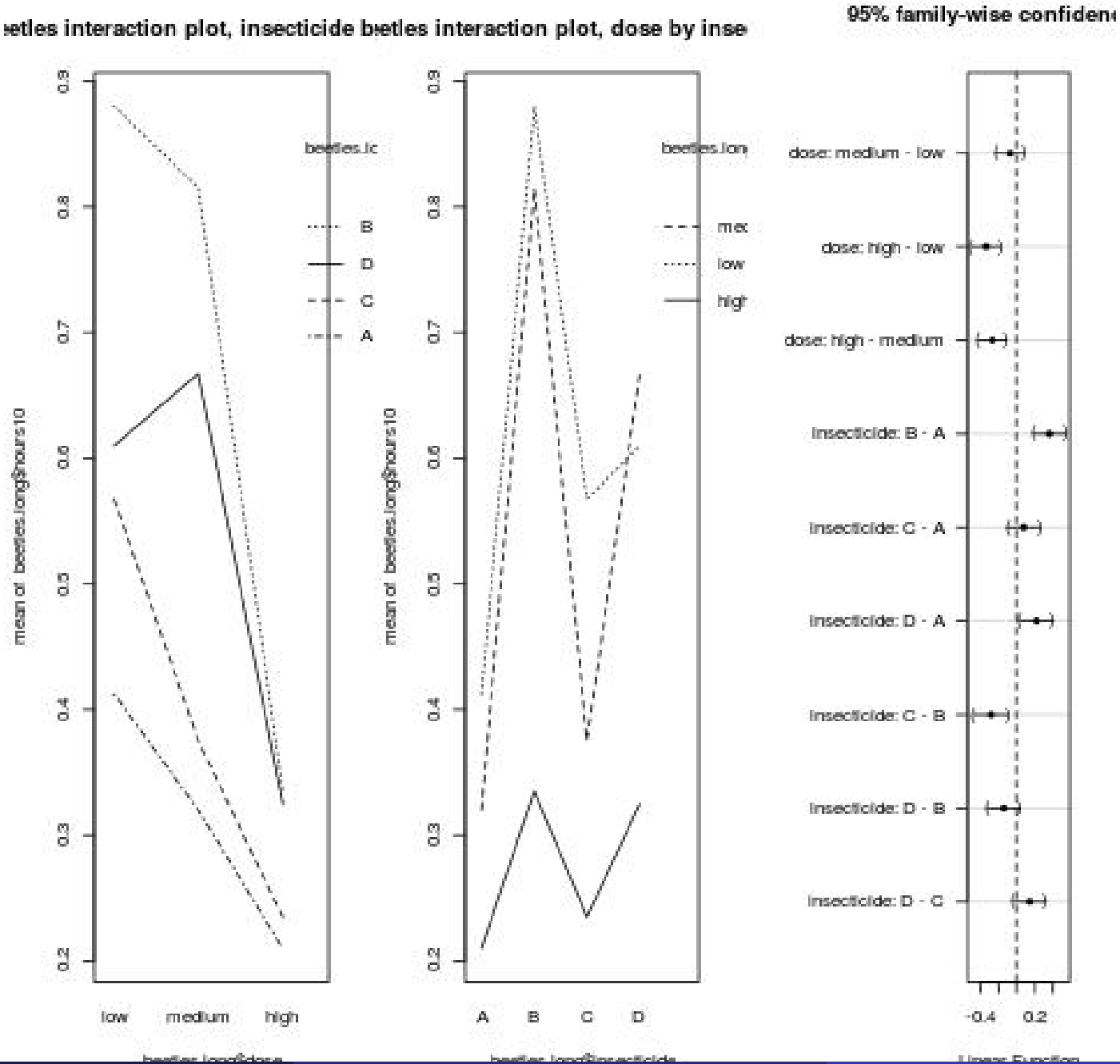
```
lm.h.d.i <- update(lm.h.d.i.di, ~ . - dose:insecticide )
library(car)
Anova(lm.h.d.i, type=3)
Response: hours10
             Sum Sq Df F value    Pr(>F)
(Intercept) 1.63654  1  65.408 4.224e-10 ***
dose        1.03301  2  20.643 5.704e-07 ***
insecticide 0.92121  3  12.273 6.697e-06 ***
Residuals   1.05086 42
```

Testing multiple factors

```
# Testing multiple factors is of interest here.
# Note that the code below corrects the p-values
#   for all the tests done for both factors together,
#   that is, the Bonferroni-corrected significance level
 is (alpha / (d + i))
#   where d = number of dose        comparisons
#   and   i = number of insecticide comparisons.

# correcting over dose and insecticide
library(multcomp)
glht.beetle.di <- glht(aov(lm.h.d.i),
 linfct = mcp(dose = "Tukey"
        , insecticide = "Tukey"))
summary(glht.beetle.di, test = adjusted("bonferroni"))
```

```
                             Estimate Std. Error t value Pr(>|t|)
dose: medium - low == 0   -0.07313    0.05592  -1.308 1.000000
dose: high - low == 0     -0.34125    0.05592  -6.102 2.55e-06 ***
dose: high - medium == 0  -0.26812    0.05592  -4.794 0.000186 ***
insecticide: B - A == 0    0.36250    0.06458   5.614 1.28e-05 ***
insecticide: C - A == 0    0.07833    0.06458   1.213 1.000000
insecticide: D - A == 0    0.22000    0.06458   3.407 0.013134 *
insecticide: C - B == 0   -0.28417    0.06458  -4.400 0.000653 ***
insecticide: D - B == 0   -0.14250    0.06458  -2.207 0.295702
insecticide: D - C == 0    0.14167    0.06458   2.194 0.304527
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
(Adjusted p values reported -- bonferroni method)
```

```
# plot the summary
  op <- par(no.readonly = TRUE) # the whole list of settable par's.
  # make wider left margin to fit contrast labels
  par(mar = c(5, 10, 4, 2) + 0.1)
  # order is c(bottom, left, top, right)
# plot bonferroni-corrected difference intervals
plot(summary(glht.beetle.di, test = adjusted("bonferroni"))
    , sub="Bonferroni-adjusted Treatment contrasts")
  par(op) # reset plotting options
```


setles interaction plot, insecticide
beetles interaction plot, dose by inse
95% family-wise confiden
mean of beetles.long$hours10
beetles.lo
B
D
C
A
low
medium
high
mean of beetles.long$hours10
beetles.lon
mec
low
high
A
B
C
D
dose: medium - low
dose: high - low
dose: high - medium
insecticide: B - A
insecticide: C - A
insecticide: D - A
insecticide: C - B
insecticide: D - B
insecticide: D - C
-0.4
0.2

Given the test for interaction, I would likely summarize the main effects assuming no interaction. For factor dose

- The average survival time decreases as the dose increases, with estimated mean survival times of 0.618, 0.544, and 0.276, respectively.
- A Bonferroni comparison shows that the population mean survival time for the high dose (averaged over insecticides) is significantly less than the population mean survival times for the low and medium doses (averaged over insecticides).
- The two lower doses are not significantly different from each other. This leads to two dose groups:

```
Doses:
1=Low   2=Med    3=High
0.618   0.544    0.276
------------     -----
```

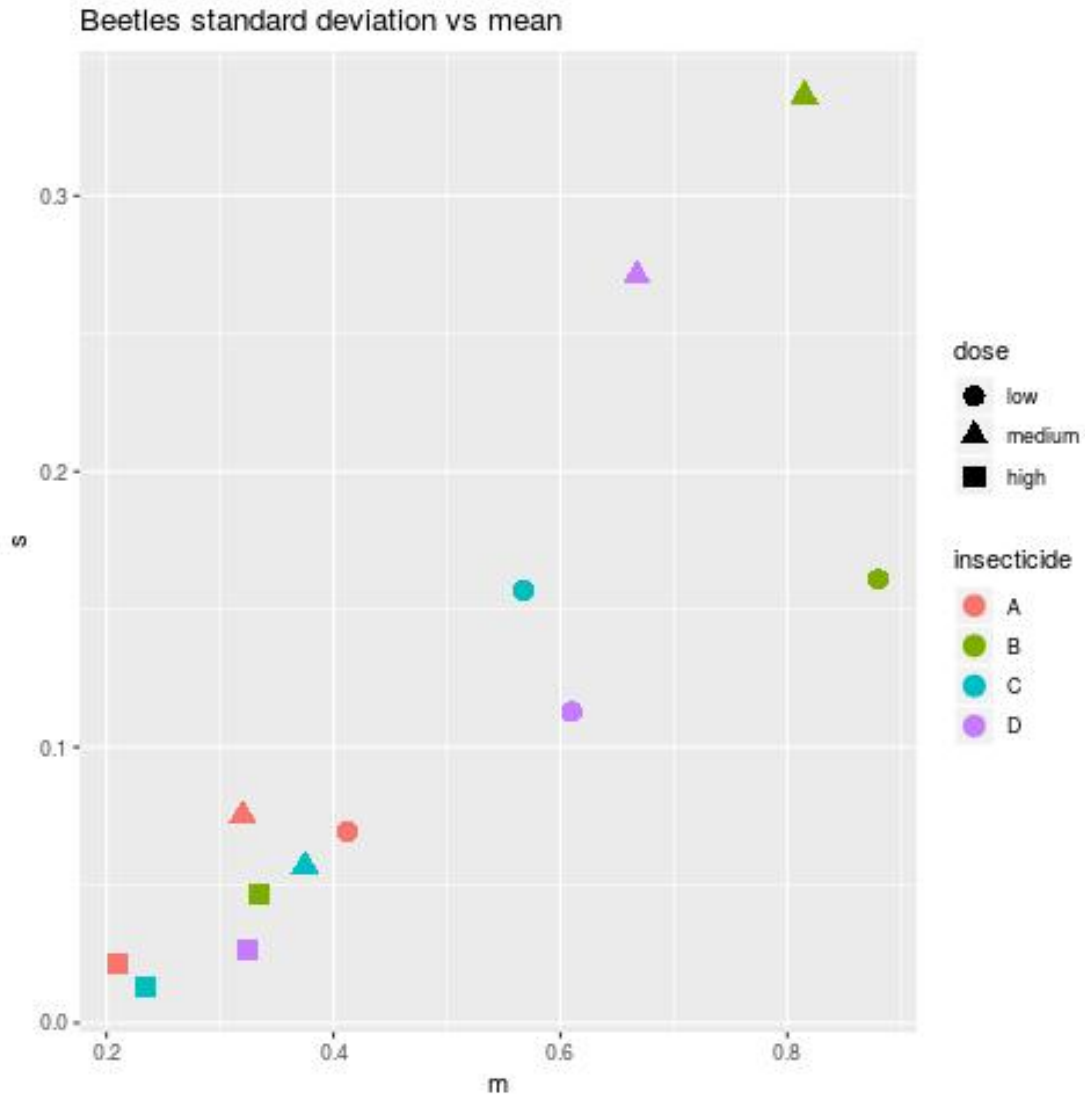
Given the test for interaction, for insecticides

- A is not significantly better than C, but is significantly better than B or D, regardless of the dose.
- The difference in marginal means for insecticides B and A of 0.677 - 0.314 = 0.363 is the expected decrease in survival time from using A instead of B, regardless of dose. This is also the expected decrease in survival times when averaged over doses.

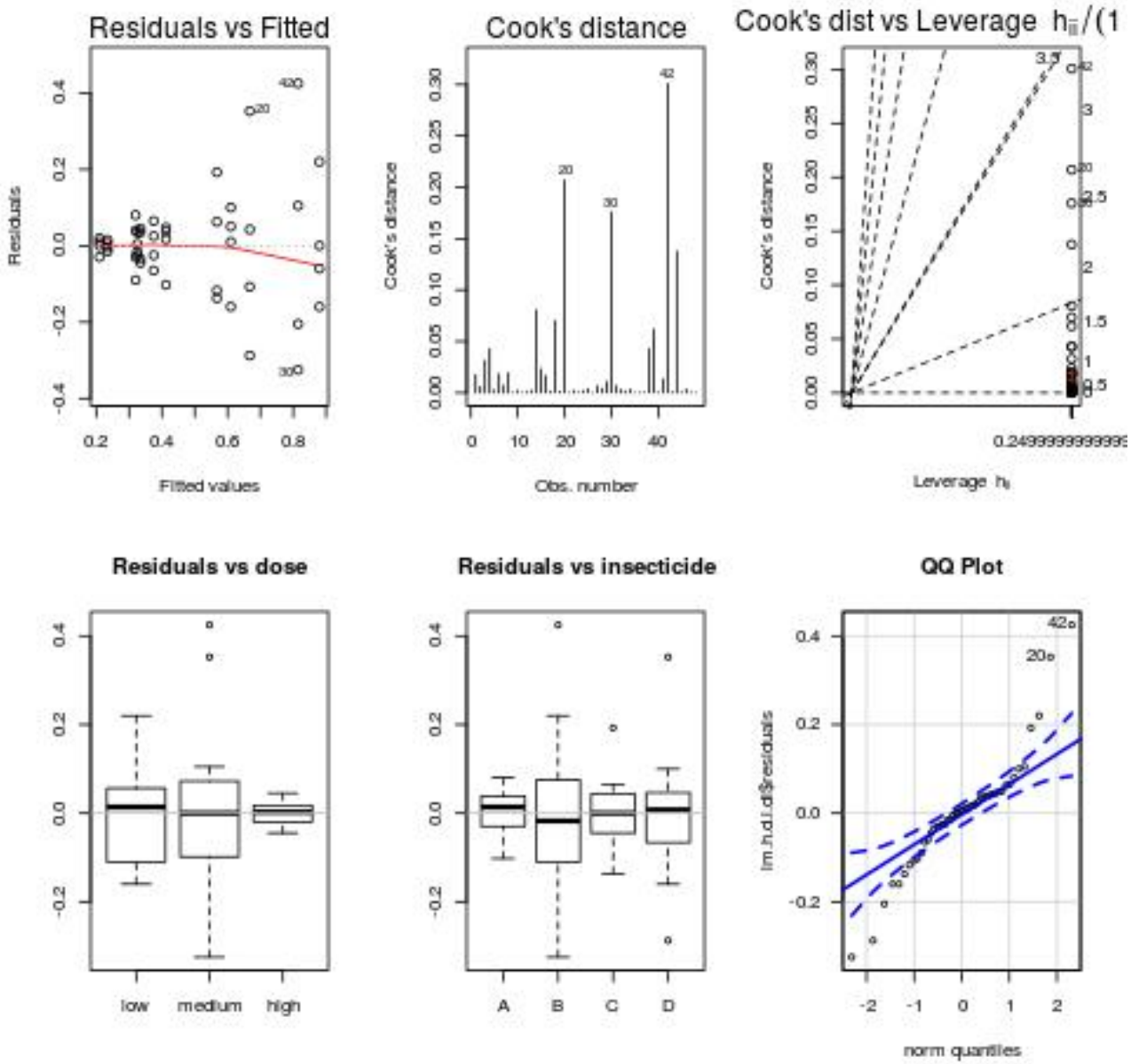
```
Insecticides:
B     D     C      A
0.677 0.534 0.393  0.314
----------
      ----------
            ----------
```

Diagnostics

Plot of the standard deviation vs mean shows an increasing trend.



Diagnostic plots of the model with interactions



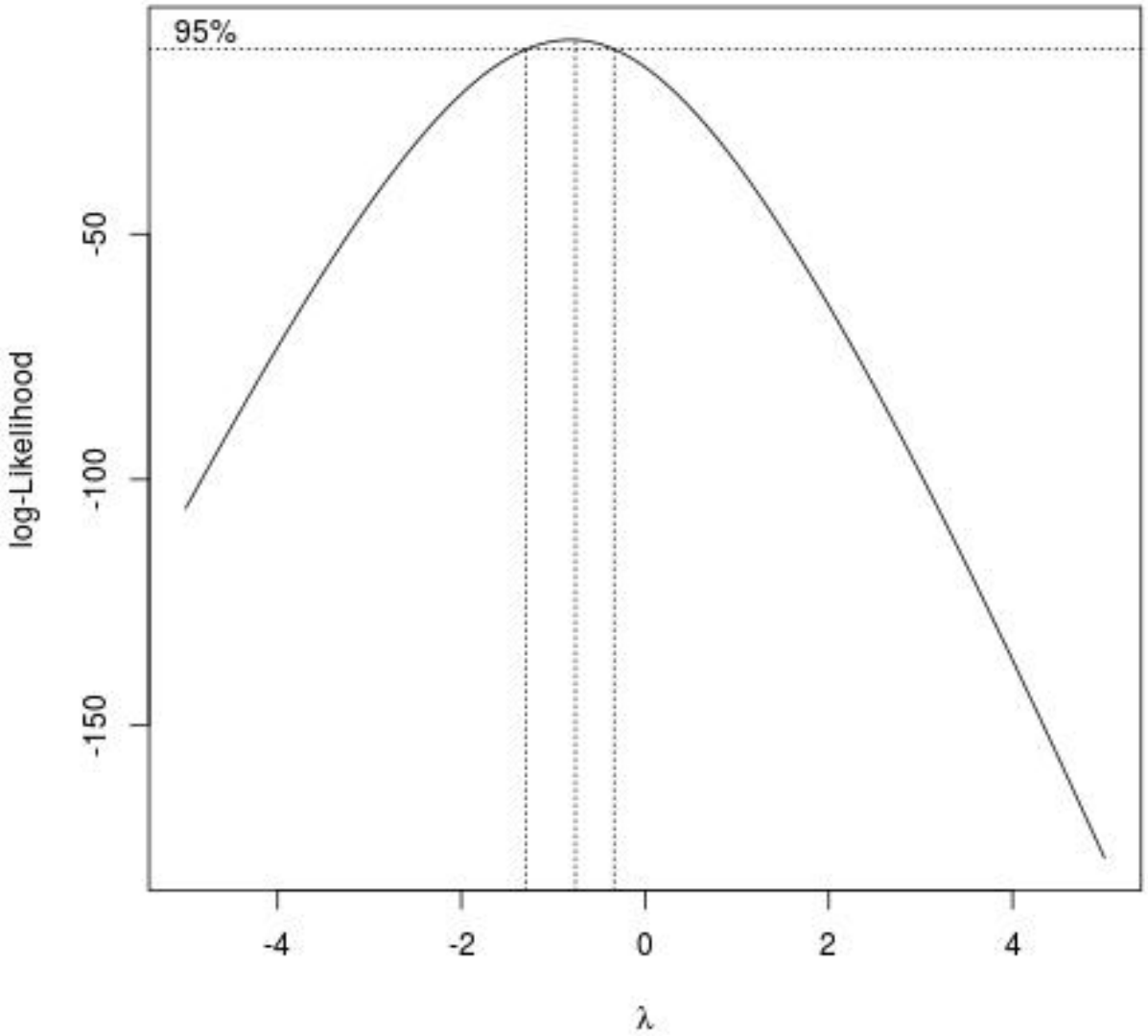
Diagnostic plots show the following features

- The normal quantile plot shows an "S" shape rather than a straight line, suggesting the residuals are not normal.
- The residuals vs the fitted (predicted) values show that the higher the predicted value the more variability (horn shaped). This could be seen from the plot of standard deviation v.s. mean.
- The plot of the Cooks distances indicate a few influential observations.

Transformations

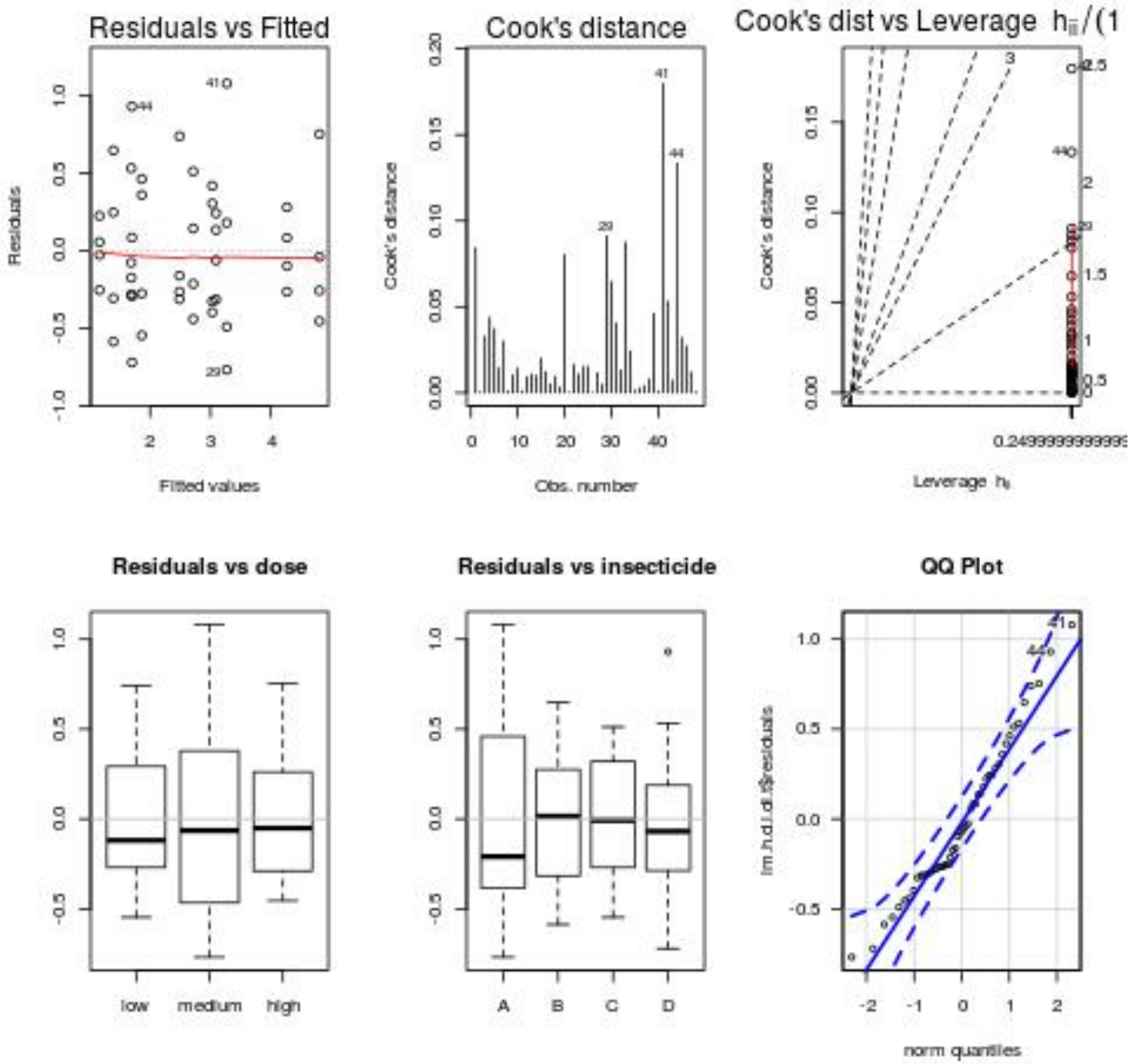
```
library(MASS)
  boxcox(lm.h.d.i.di, lambda = seq(-5, 5, length = 10),
  plotit = TRUE)
```

- $\lambda = -1$ is within the 95% confidence interval, we will try a transformation of $y^* = 1/y$
- the inverse survival time has a natural interpretation as the dying rate. For example, if you survive 2 hours, then 1/2 is the proportion of your remaining lifetime expired in the next hour.


95%
log-Likelihood
-50
-100
-150
-4
-2
0
2
4
λ

```
> lm.h.d.i.di.t <- lm(1/hours10 ~ dose*insecticide,
data = beetles.long)
>   Anova(lm.h.d.i.di.t, type=3)
Anova Table (Type III tests)

Response: 1/hours10
                  Sum Sq Df  F value    Pr(>F)
(Intercept)      24.7383  1 103.0395 4.158e-12 ***
dose             11.1035  2  23.1241 3.477e-07 ***
insecticide       3.5723  3   4.9598  0.005535 **
dose:insecticide  1.5708  6   1.0904  0.386733
Residuals         8.6431 36
```

Diagnostic plots of the transformed model



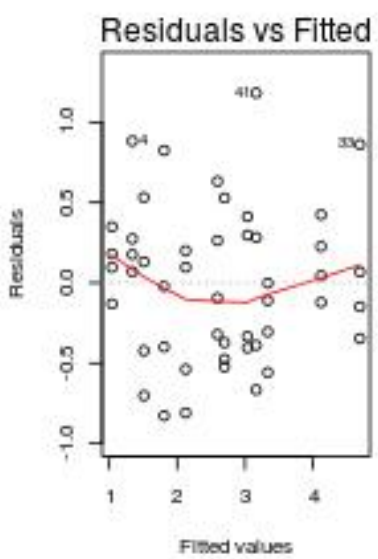
Diagnostic plots of the transformed model show the following features

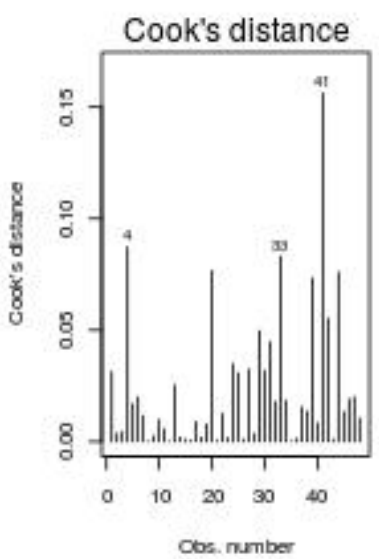
- The normal quantile plot shows a rough straight line, suggesting the residuals are normal.
- The residuals vs the fitted (predicted) values show a random pattern.
- The plot of the Cooks distances indicate a few influential observations, but none of them are greater than 1.
- Normality assumption and constant variance assumption seem not violated.

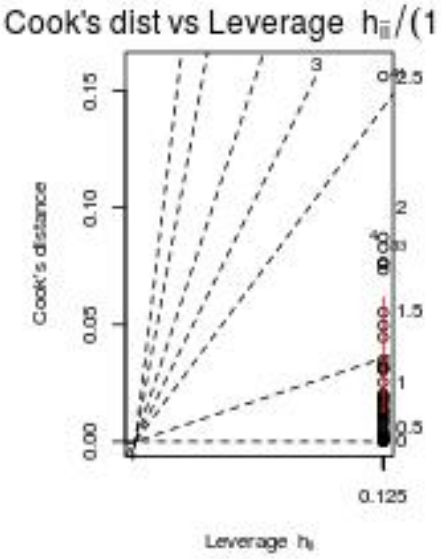
Drop the nonsignificant interaction term, refit model

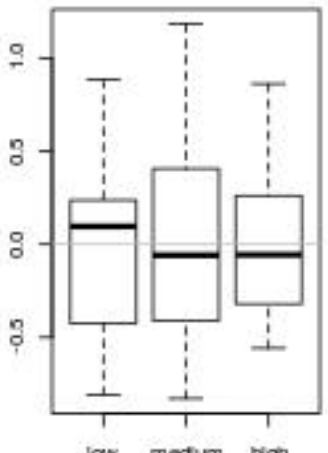
```
> lm.h.d.i.di.t2 <- lm(1/hours10 ~ dose+insecticide,
data = beetles.long)
>   Anova(lm.h.d.i.di.t2, type=3)
Anova Table (Type III tests)

Response: 1/hours10
             Sum Sq Df F value    Pr(>F)
(Intercept) 58.219  1 239.399 < 2.2e-16 ***
dose        34.877  2  71.708 2.865e-14 ***
insecticide 20.414  3  27.982 4.192e-10 ***
Residuals   10.214 42
```

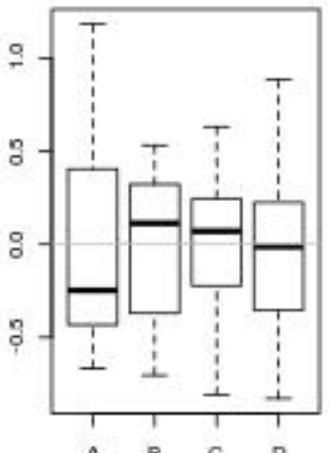
Diagnostic plots of the reduced transformed model

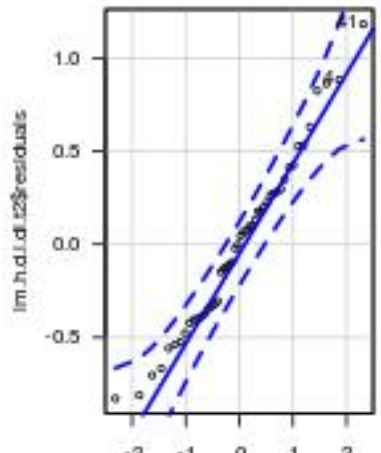










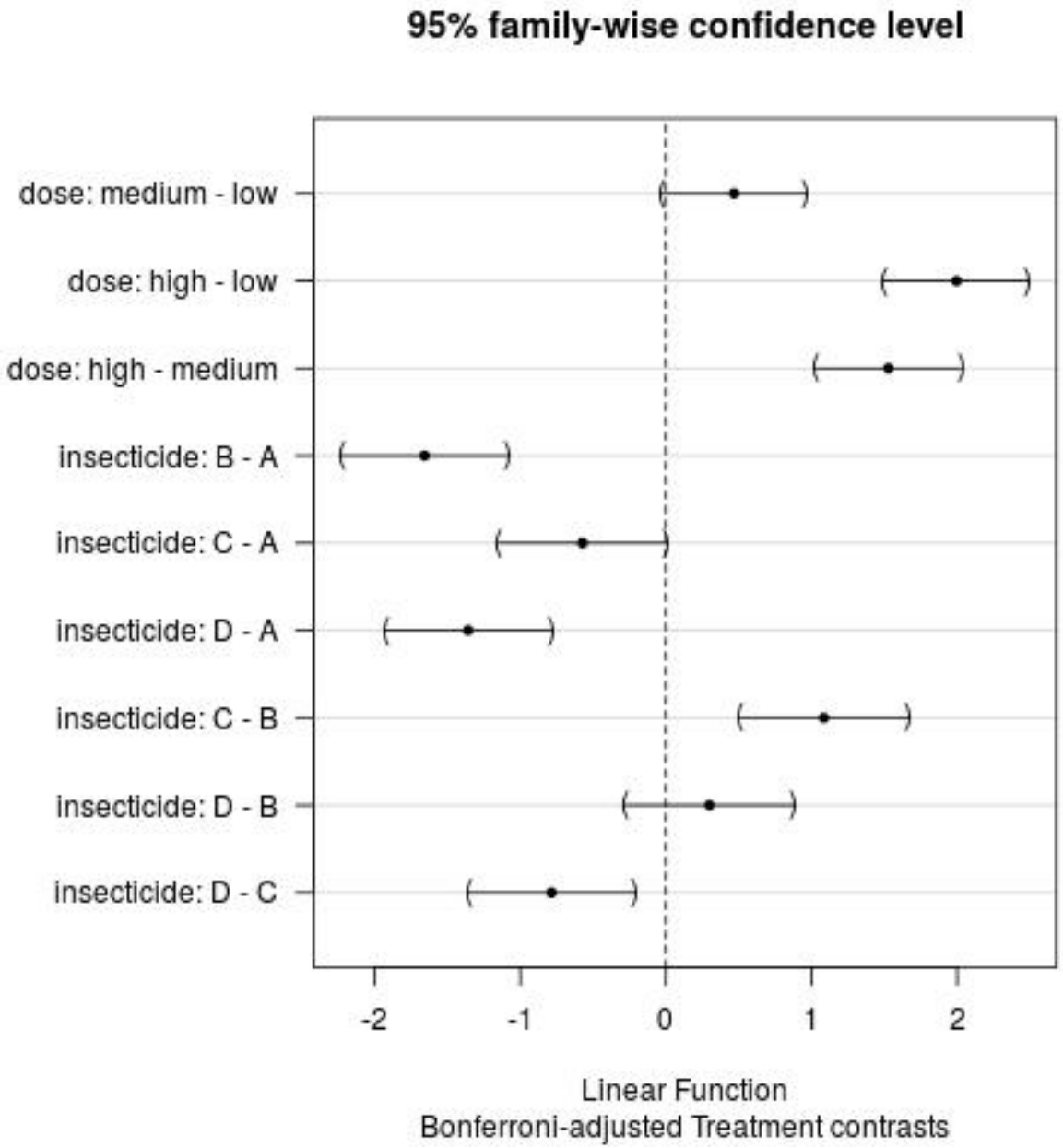


Diagnostic plots of the reduced transformed model show the following features

- The normal quantile plot shows a rough straight line, suggesting the residuals are normal.
- The residuals vs the fitted (predicted) values show a random pattern.
- The plot of the Cooks distances indicate a few influential observations, but none of them are greater than 1.
- Normality assumption and constant variance assumption seem not violated.

Multiple comparisons

```
library(multcomp)
  glht.beetle.di2 <- glht(aov(lm.h.d.i.di.t2),
  linfct = mcp(dose = "Tukey"
    , insecticide = "Tukey"))
  summary(glht.beetle.di2, test = adjusted("bonferroni"))

  # plot the summary
  op <- par(no.readonly = TRUE) # the whole list of
  settable par's.
  # make wider left margin to fit contrast labels
  par(mar = c(5, 10, 4, 2) + 0.1)
  # order is c(bottom, left, top, right)
  # plot bonferroni-corrected difference intervals
  plot(summary(glht.beetle.di2, test = adjusted("bonferroni"))
       , sub="Bonferroni-adjusted Treatment contrasts")
  par(op) # reset plotting options
```


95% family-wise confidence level
dose: medium - low
dose: high - low
dose: high - medium
insecticide: B - A
insecticide: C - A
insecticide: D - A
insecticide: C - B
insecticide: D - B
insecticide: D - C
-2
-1
0
1
2
Linear Function
Bonferroni-adjusted Treatment contrasts

For dose

- A Bonferroni comparison shows that the population mean dying rate for the high dose (averaged over insecticides) is significantly higher than the population mean dying rate for the low and medium doses (averaged over insecticides).
- The two lower doses are not significantly different from each other.

For insecticides

- A is not significantly better than C, but is significantly better than B or D, regardless of the dose.
- C is significantly better than B or D, B is not significantly better than D, regardless of the dose.

Multiple comparison, when interaction is important

Example: The maximum output voltage for storage batteries is thought to be influenced by

- the temperature in the location at which the battery is operated
- and the material used in the plates.

A scientist designed a two-factor study (a balanced 3-by-3 factorial experiment with four observations per treatment) to examine this hypothesis,

- Temperatures (50, 65, 80)
- Materials for the plates (1, 2, 3).
- Four batteries were tested at each of the 9 combinations of temperature and material type.
- The maximum output voltage was recorded for each battery.

ANOVA with interaction

```
> battery
  material temp  v1  v2  v3  v4
1        1   50 130 155  74 180
2        1   65  34  40  80  75
3        1   80  20  70  82  58
4        2   50 150 188 159 126
5        2   65 136 122 106 115
6        2   80  25  70  58  45
7        3   50 138 110 168 160
8        3   65 174 120 150 139
9        3   80  96 104  82  60
```

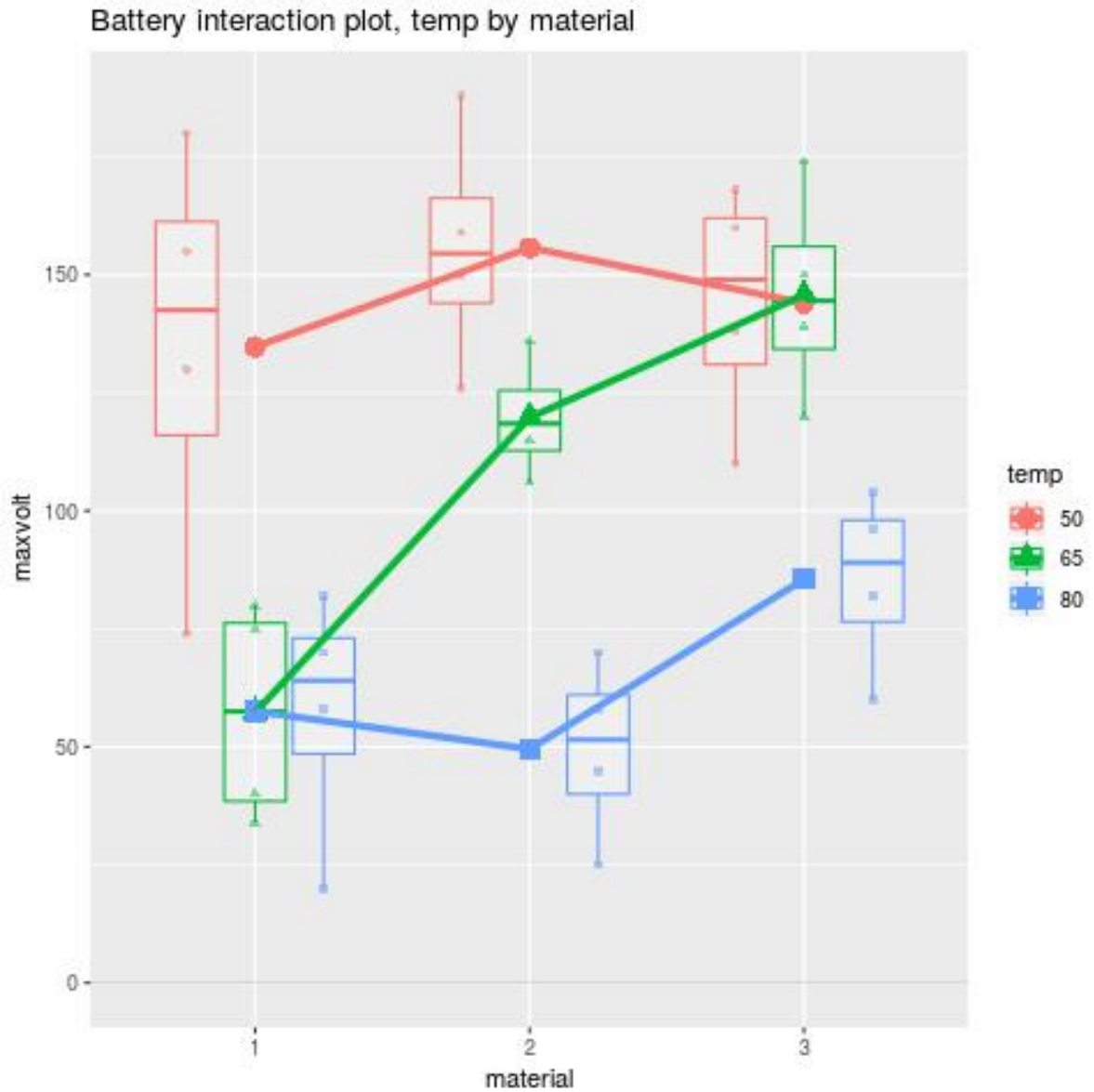
```
> battery.long
   material temp battery maxvolt
1         1   50      v1     130
2         1   65      v1      34
3         1   80      v1      20
4         2   50      v1     150
5         2   65      v1     136
6         2   80      v1      25
7         3   50      v1     138
8         3   65      v1     174
9         3   80      v1      96
10        1   50      v2     155
11        1   65      v2      40
12        1   80      v2      70
13        2   50      v2     188
14        2   65      v2     122
15        2   80      v2      70
16        3   50      v2     110
```

```
> table(battery.long$material,battery.long$temp)

   50 65 80
1  4  4  4
2  4  4  4
3  4  4  4
```

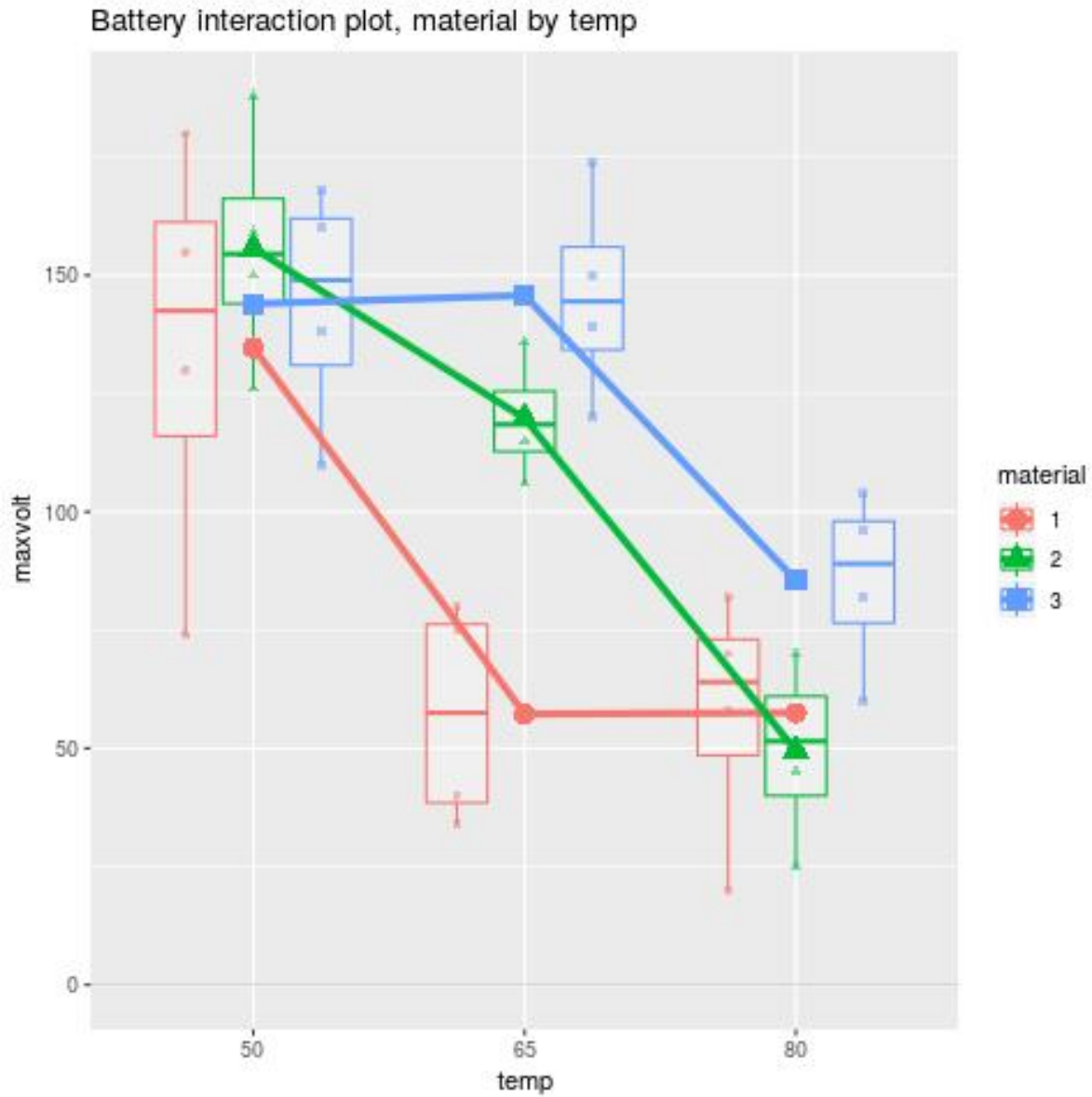
```
> lm.m.m.t.mt <- lm(maxvolt ~ material*temp,
data = battery.long)
>   library(car)
>   Anova(lm.m.m.t.mt, type=3)
Anova Table (Type III tests)

Response: maxvolt
              Sum Sq Df  F value    Pr(>F)
(Intercept)    72630  1 107.5664 6.456e-11 ***
material         886  2   0.6562 0.5268904
temp           15965  2  11.8223 0.0002052 ***
material:temp   9614  4   3.5595 0.0186112 *
Residuals      18231 27
```

- The two-way ANOVA table indicates that the main effect of temperature and the interaction are significant at the 0.05 level, the main effect of material is not.
  ——-note that the test for the main effect for material doesn't appear significant, but because the interaction is significant, you can't conclude that the materials are not significantly affecting the voltage.
  ——-if the model is made with the interaction term removed, then both material and temperature are significant. The p-value for material isn't significant only when the interaction is in the model.
- The profile plots of the material profiles have different slopes, which is consistent with the presence of a temperature-by-material interaction.

Profile plots



Profile plots



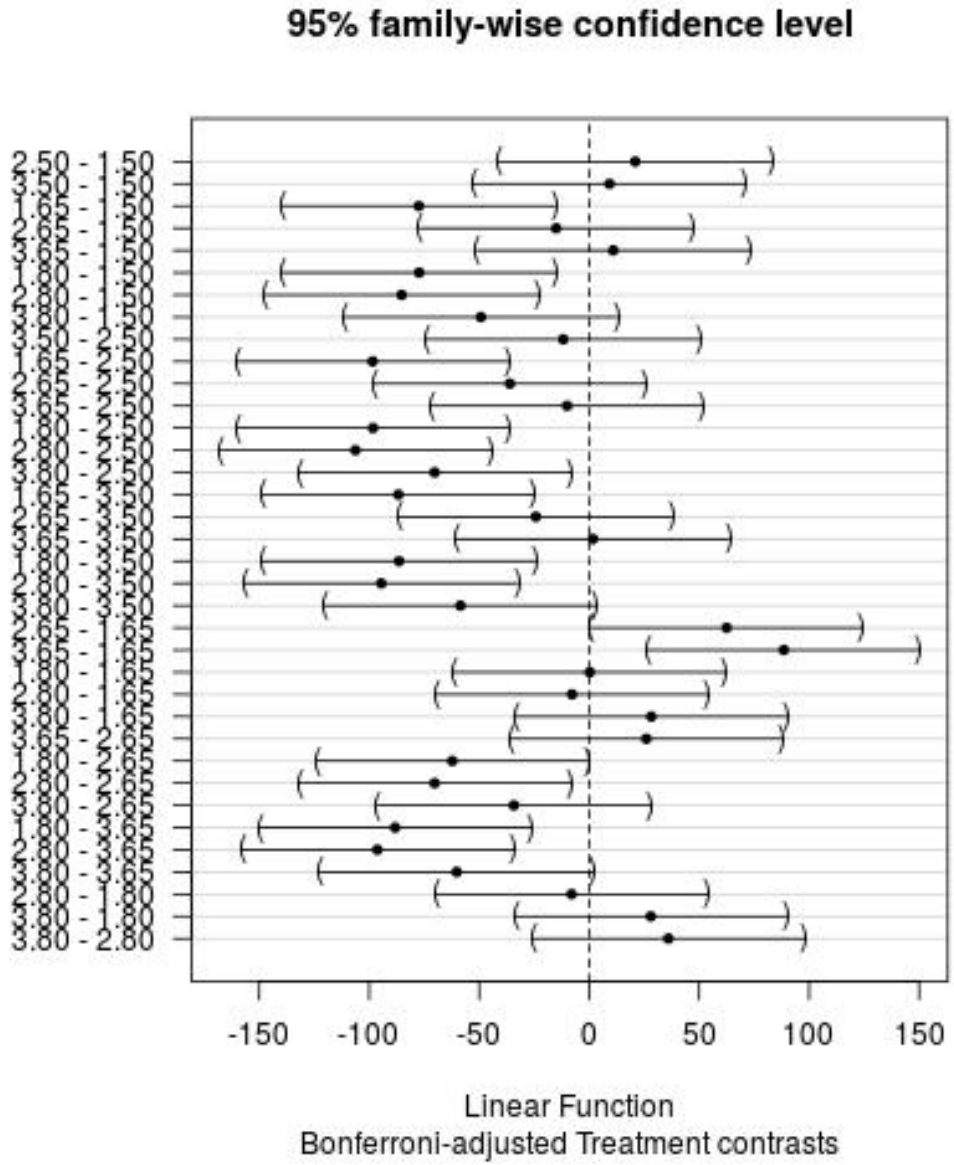
From the interaction plots, we see that as the temperature increases, the voltage tends to decrease for all three materials. However, for material 3, there is very little change from 50 to 65 degrees, and a big decrease from 65 to 80. For material 1, there is a large change in voltage from 50 to 65 degrees, and very little change from 65 to 80.

This suggests that the effect of temperature depends on the material, and similarly, the effect of the material depends on the temperature.

When there are model interactions, in general, we shall compare the different combinations of the two factors.

Comparing means (interaction is significant, compare the different combinations of temp*materials)

```
library(multcomp)
battery.long$mt
<- with(battery, interaction(material, temp))
lm.mt <- lm(maxvolt ~ mt, data = battery.long)
glht.battery <- glht(aov(lm.mt),
linfct = mcp(mt = "Tukey"))
summary(glht.battery, test = adjusted("bonferroni"))
# plot the summary
op <- par(no.readonly = TRUE) # the whole list of settable p
# make wider left margin to fit contrast labels
par(mar = c(5, 10, 4, 2) + 0.1)
 # order is c(bottom,
 left, top, right)
# plot bonferroni-corrected difference intervals
plot(summary(glht.battery, test = adjusted("bonferroni"))
      , sub="Bonferroni-adjusted Treatment contrasts")
```


95% family-wise confidence level
2.50 - 1.50
3.50 - 1.50
1.65 - 1.50
2.65 - 1.50
3.65 - 1.50
1.80 - 1.50
2.80 - 1.50
3.80 - 1.50
3.50 - 2
1.65 - 2.50
2.65 - 2.50
3.65 - 2.50
1.80 - 2.50
2.80 - 2.50
3.80 - 2.50
1.65 - 3.50
2.65 - 3.50
3.65 - 3.50
1.80 - 3.50
2.80 - 3.50
3.80 - 3.50
2.65 - 1.65
3.65 - 1.65
1.80 - 1.65
2.80 - 1.65
3.80 - 1.65
3.65 - 2.65
1.80 - 2.65
2.80 - 2.65
3.80 - 2.65
1.80 - 3.65
2.80 - 3.65
3.80 - 3.65
2.80 - 1.80
3.80 - 1.80
3.80 - 2.80
-150
-100
-50
0
50
100
150
Linear Function
Bonferroni-adjusted Treatment contrasts

The mean maximum output voltage for storage batteries for different combinations from material and temperature is as follows

	material	temp	m
1	1	50	134.75
2	1	65	57.25
3	1	80	57.50
4	2	50	155.75
5	2	65	119.75
6	2	80	49.50
7	3	50	144.00
8	3	65	145.75
9	3	80	85.50

Material 2 with temperature 50 produce highest mean output voltage of 155.75, and Material 2 with temperature 80 produce lowest mean output voltage of 49.50.

From multiple comparison of interactions,

- The mean maximum output voltage for storage batteries by material 1 (m1) and temperature 50 (t50) is significantly different from those by m2*t80, m1*t80, m1*t65
- The mean maximum output voltage for storage batteries by material 2 (m2) and temperature 50 (t50) is significantly different from those by m2*t80, m1*t80, m3*80, m1*t65
- The mean maximum output voltage for storage batteries by material 3 (m3) and temperature 50 (t50) is significantly different from those by m2*t80, m1*t80, m1*t65
- The mean maximum output voltage for storage batteries by material 1 (m1) and temperature 65 (t65) is significantly different from those by m3*t65
- The mean maximum output voltage for storage batteries by material 2 (m2) and temperature 65 (t65) is significantly different from those by m2*t80
- The mean maximum output voltage for storage batteries by material 3 (m3) and temperature 65 (t65) is significantly different from those by m2*t80, m1*t80.

- Temperature is important in producing high output voltage.
- We can do multiple comparison of temperature. But when interaction is significant, comparing mean of different combinations seem more appropriate and interesting.
- When interaction is significant, reduced model should include both main effects involved in interaction, even some of them are not significant.

Unbalanced Two-Factor Designs and Analysis

Sample sizes are usually unequal, or unbalanced, for the different treatment groups in an experiment.

- this has no consequence on the specification of a model
- with unbalanced designs, the Type I and Type III SS differ, as do the main effect averages given by means and lsmeans.
- Inferences might critically depend on which summaries are used.

Example: Rat insulin

The experiment consists of measuring insulin levels in rats a certain length of time after a fixed dose of insulin was injected into their jugular or portal veins.

- This is a two-factor study with two vein types (jugular, portal) and three time levels (0, 30, and 60).
- An unusual feature of this experiment is that the rats used in the six vein and time combinations are distinct.
  ——— The design is unbalanced, with sample sizes varying from 3 to 12.

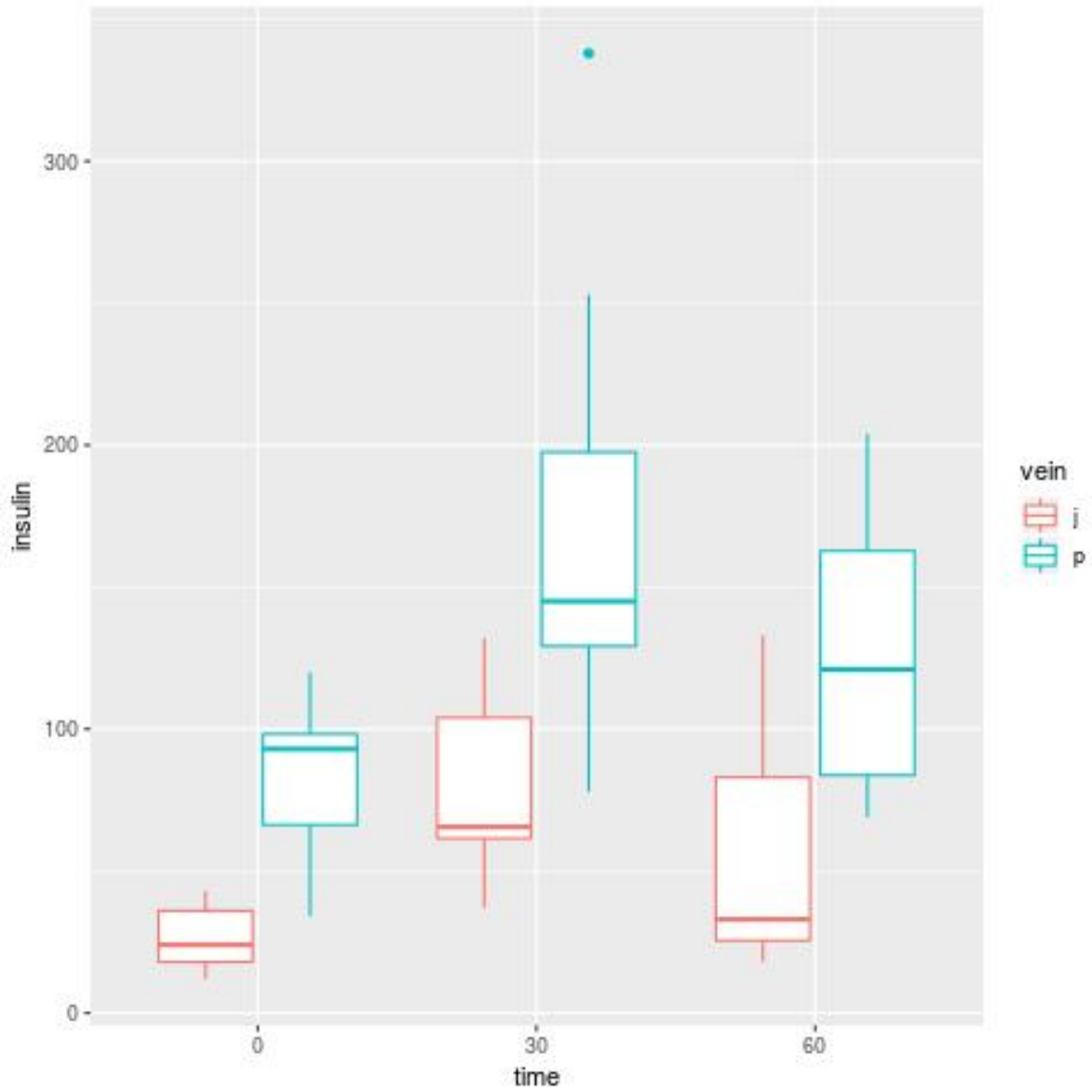
Unbalanced ANOVA

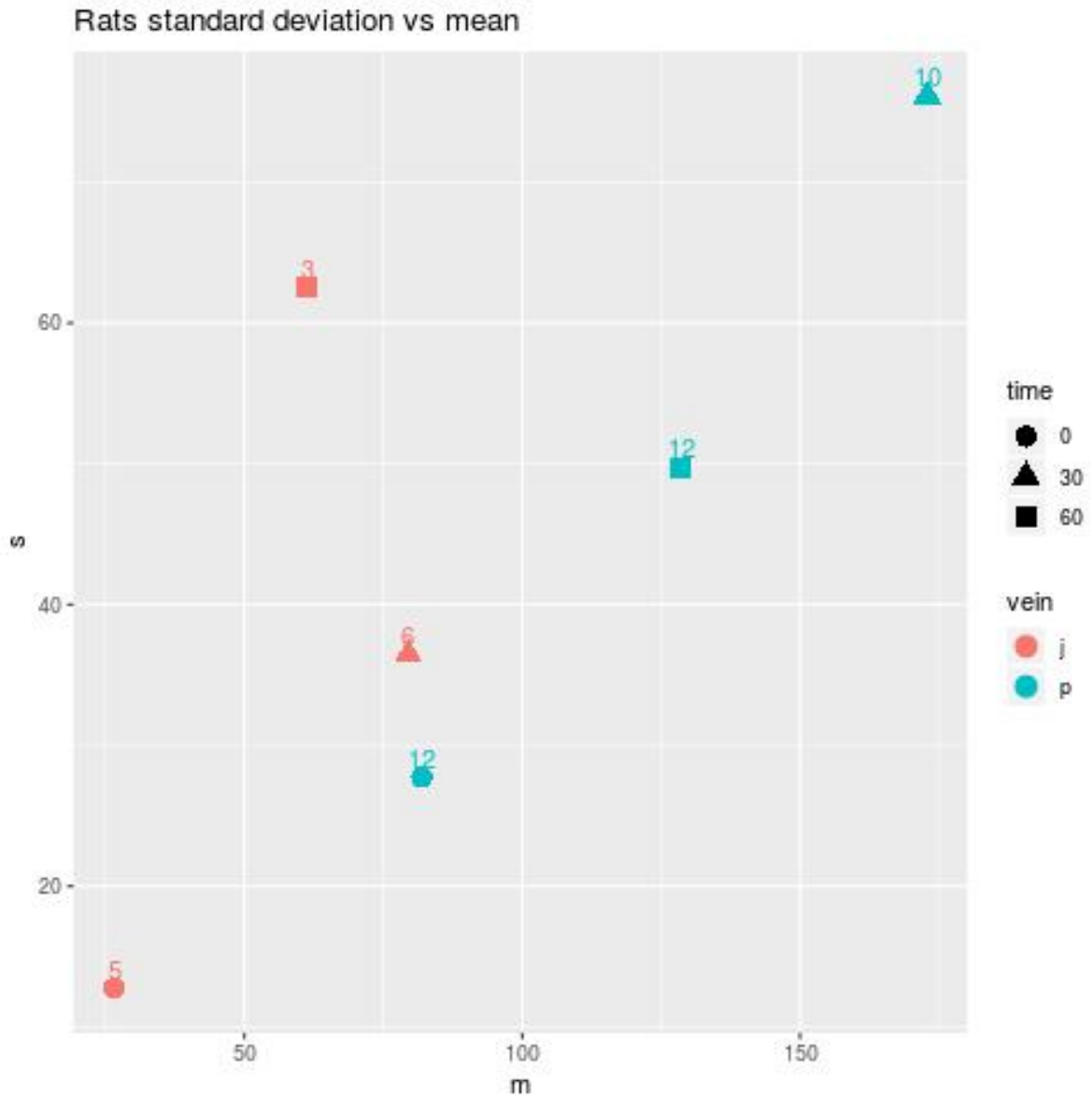
```
> rat
   vein time insulin
1     j    0      18
2     j    0      36
3     j    0      12
4     j    0      24
5     j    0      43
6     j   30      61
7     j   30     116
8     j   30      63
```

```
> table(rat$vein,rat$time)

   0 30 60
j  5  6  3
p 12 10 12
```

```
> rat.meansd.tv
  time vein          m         s  n
1    0    j   26.60000 12.75931  5
2    0    p   81.91667 27.74710 12
3   30    j   79.50000 36.44585  6
4   30    p  172.90000 76.11753 10
5   60    j   61.33333 62.51666  3
6   60    p  128.50000 49.71830 12
```

It appears the standard deviation increases with the mean.

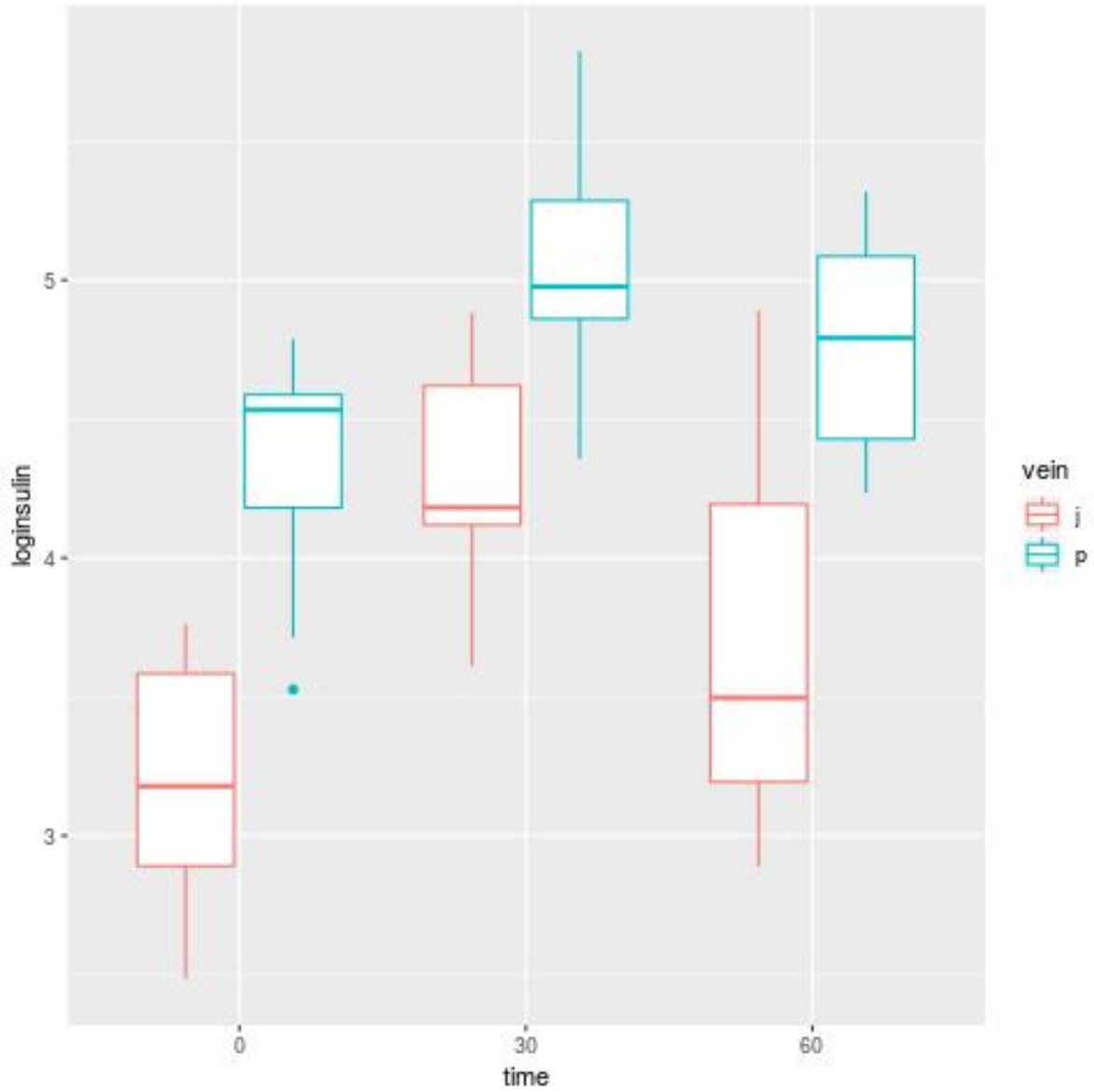


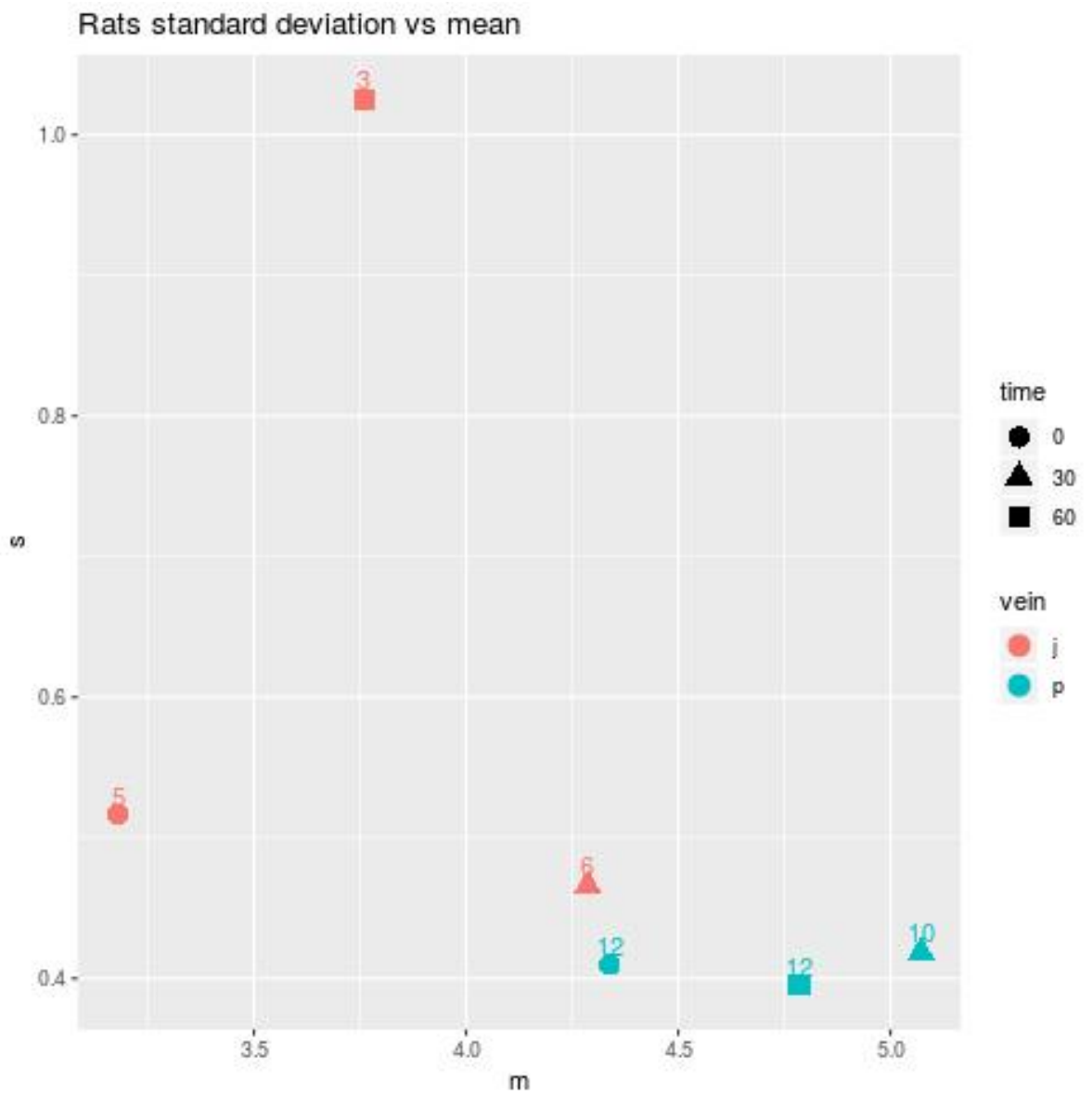

Rats standard deviation vs mean
s
m
60
40
20
50
100
150
time
0
30
60
vein
j
p

Transformation

We take the log of insulin to correct the problem. The variances are more constant now, except for one sample with only 3 observations which has a larger standard deviation than the others, but because this is based on such a small sample size, its not of much concern.

```
rat$loginsulin <- log(rat$insulin)
> rat.meansd.tv
  time vein        m         s  n
1    0    j 3.179610 0.5166390  5
2    0    p 4.338230 0.4096427 12
3   30    j 4.286804 0.4660571  6
4   30    p 5.072433 0.4185221 10
5   60    j 3.759076 1.0255165  3
6   60    p 4.785463 0.3953252 12
```


loginsulin
5
4
3
0
30
60
time
vein
i
p


Rats standard deviation vs mean
1.0
0.8
0.6
0.4
s
3.5
4.0
4.5
5.0
m
time
0
30
60
vein
j
p

Unbalanced ANOVA

For unbalanced ANOVA, type I versus type III sums of squares have a different meaning.

- Type I sums of squares are sequential, meaning that measure variation contributed by the variable given previous variables already in the model.
  ——This means that type I sums of squares are sensitive to the input order of the variables.
- Type III SS correspond to the reduction in Error SS achieved when an effect is added last to the model
- Typically we use type III sums of squares instead.

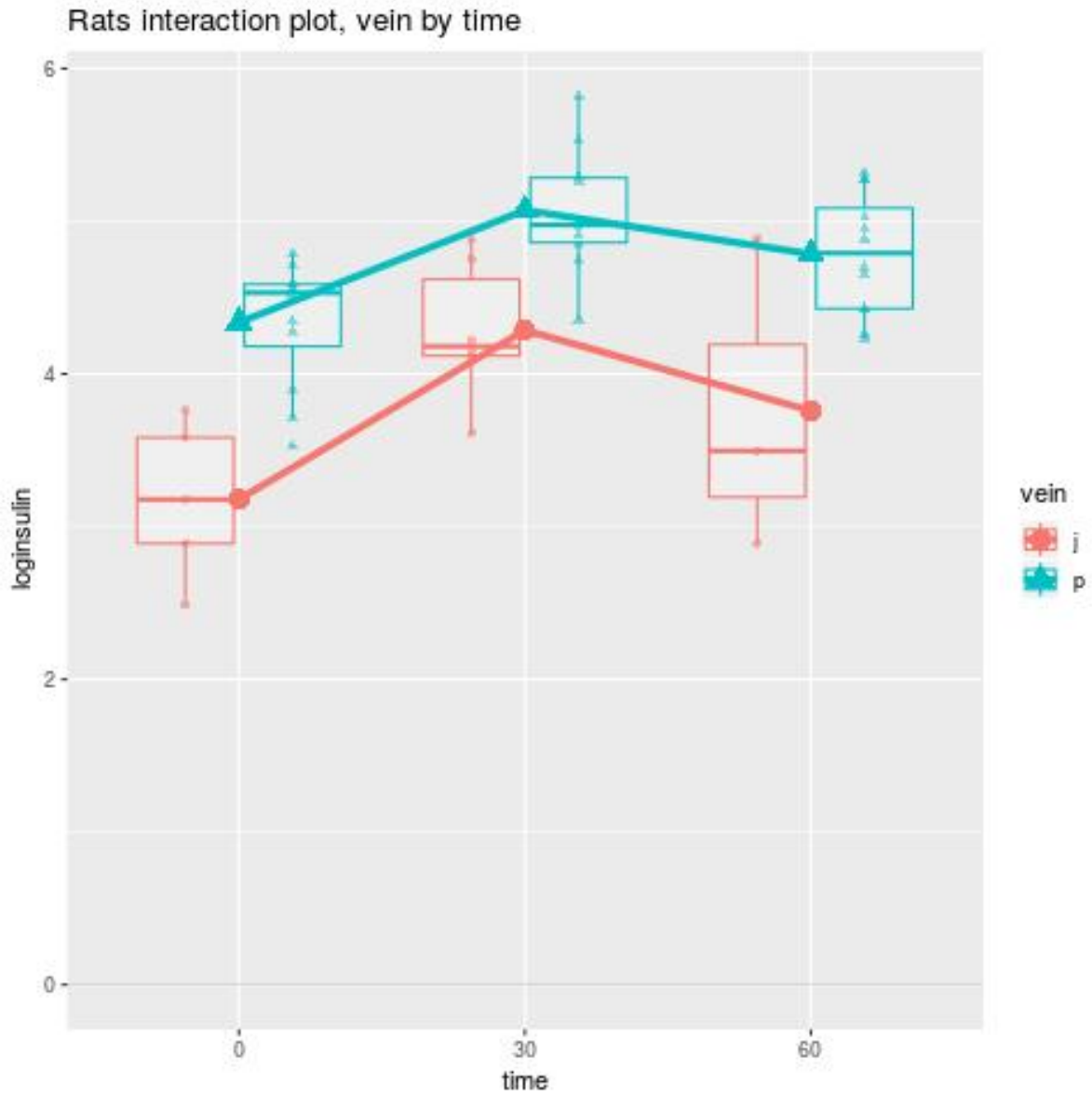
Usually the hypothesis of interest is about the significance of one factor while controlling for the level of the other factors. This equates to using type II or III SS.

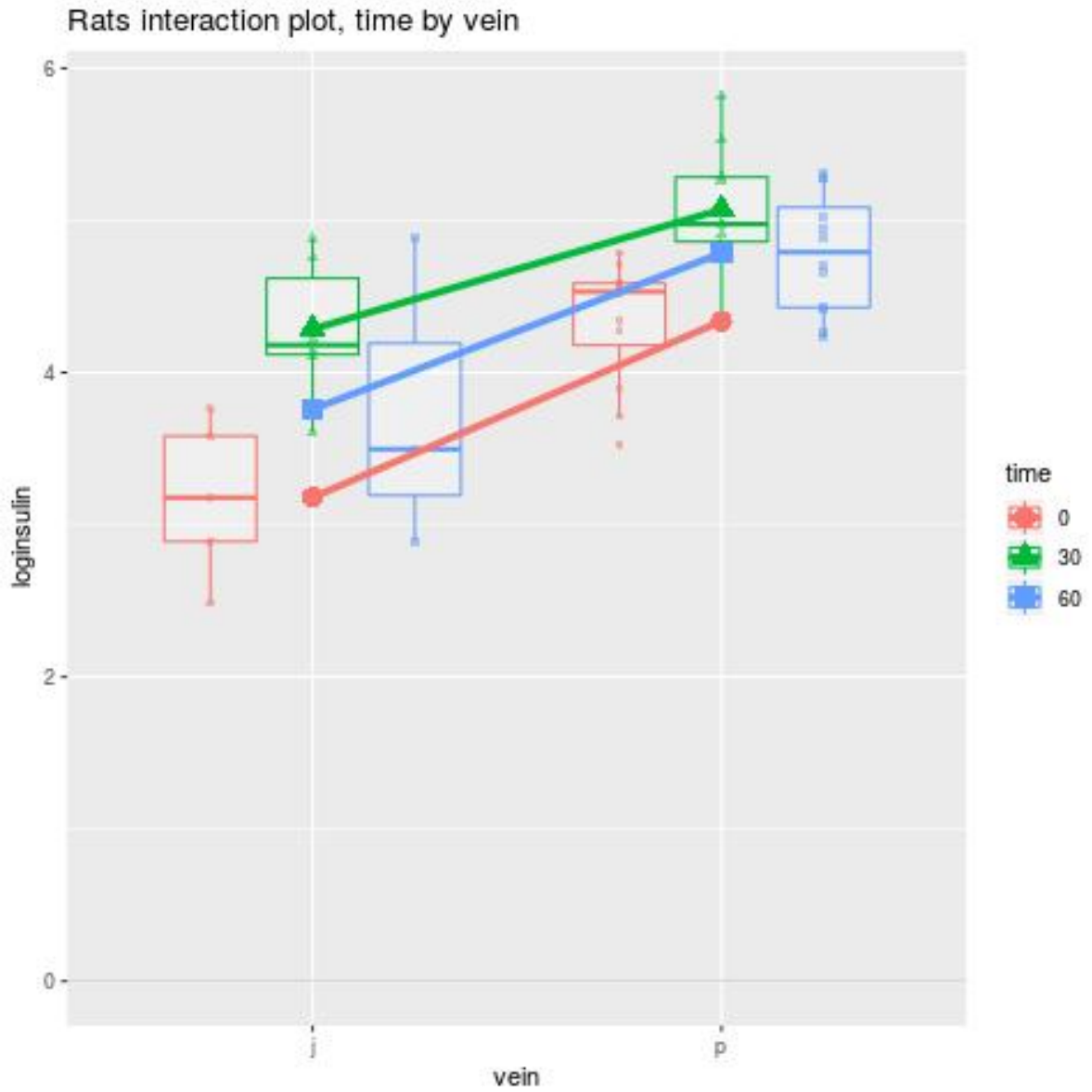
- When data is balanced, the factors are orthogonal, and types I, II and III all give the same results.
- In general, if there is no significant interaction effect, then type II is more powerful, and follows the principle of marginality. If interaction is present, then type II is inappropriate while type III can still be used, but results need to be interpreted with caution (in the presence of interactions, main effects are rarely interpretable).
- Therefore, typically, we use type III SS for unbalanced design.

```
lm.i.t.v.tv <- lm(loginsulin ~ time*vein, data = rat
                  , contrasts = list(time = contr.sum,
                  vein = contr.sum))
## CRITICAL!!!  Unbalanced design warning.
##   The contrast statement above must be
included identifying
##   each main effect with "contr.sum" in order for
 the correct
##   Type III SS to be computed.
##   See http://goanna.cs.rmit.edu.au/~fscholer/anova.php
library(car)
# type I   SS  (intercept SS not shown)
summary(aov(lm.i.t.v.tv))
# type III SS
Anova(lm.i.t.v.tv, type=3)
```

```
> # type I   SS  (intercept SS not shown)
> summary(aov(lm.i.t.v.tv))
            Df Sum Sq Mean Sq F value   Pr(>F)
time         2  5.450   2.725   12.18 6.74e-05 ***
vein         1  9.321   9.321   41.66 8.82e-08 ***
time:vein    2  0.259   0.130    0.58    0.565
Residuals   42  9.399   0.224
> # type III SS
> Anova(lm.i.t.v.tv, type=3)
Anova Table (Type III tests)

Response: loginsulin
             Sum Sq Df   F value    Pr(>F)
(Intercept) 668.54  1 2987.5842 < 2.2e-16 ***
time          6.18  2   13.7996 2.475e-05 ***
vein          9.13  1   40.7955 1.101e-07 ***
time:vein     0.26  2    0.5797    0.5645
Residuals     9.40 42
```

- Looking at the output, we see the Type I and Type III SS are different, except for the interaction term.
- The following roughly parallel profile plots indicate that interaction is not important


Rats interaction plot, vein by time
loginsulin
6
4
2
0
0
30
60
time
vein
i
p


Rats interaction plot, time by vein
loginsulin
6
4
2
0
j
p
vein
time
0
30
60

Means versus lsmeans

- The lsmeans (sometimes called adjusted means) for a single factor is an arithmetic average of cell means.
  —— For example, the mean responses in the jugular vein at times 0, 30, and 60 are 3.18, 4.29, and 3.76, respectively.
  ——The lsmeans for the jugular vein is thus

$$3.74 = (3.18 + 4.29 + 3.76)/3.$$

  —— This average gives equal weight to the 3 times even though the sample sizes at these times differ (5, 6, and 3).
- The means of 3.78 for the jugular is the average of the 14 jugular responses, ignoring time.
- If the cell sample sizes were equal, the lsmeans and means averages would agree.

Unbalanced, means and lsmeans don't match

```
> rat.mean.t  <- ddply(rat, .(time), summarise, m = mean(logi
> rat.mean.t
  time        m
1    0 3.997460
2   30 4.777822
3   60 4.580186

> lsmeans(lm.i.t.v.tv, list(pairwise ~ time),
adjust = "bonferroni")
NOTE: Results may be misleading due to involvement in interact
$`lsmeans of time`
 time   lsmean        SE df lower.CL upper.CL
 0    3.758920 0.1258994 42 3.504845 4.012996
 30   4.679619 0.1221403 42 4.433130 4.926108
 60   4.272270 0.1526754 42 3.964158 4.580381

Results are averaged over the levels of: vein
```

```
> rat.mean.v
  vein        m
1    j 3.778293
2    p 4.712019
>   # compare jugular mean above (3.778) with the lsmeans aver
> (3.179610 + 4.286804 + 3.759076)/3
[1] 3.74183
> lsmeans(lm.i.t.v.tv, list(pairwise ~ vein), adjust = "bonfer
NOTE: Results may be misleading due to involvement in interact
$`lsmeans of vein`
 vein   lsmean         SE df lower.CL upper.CL
 j    3.741830 0.13192664 42 3.475592 4.008069
 p    4.732042 0.08142689 42 4.567716 4.896369
```

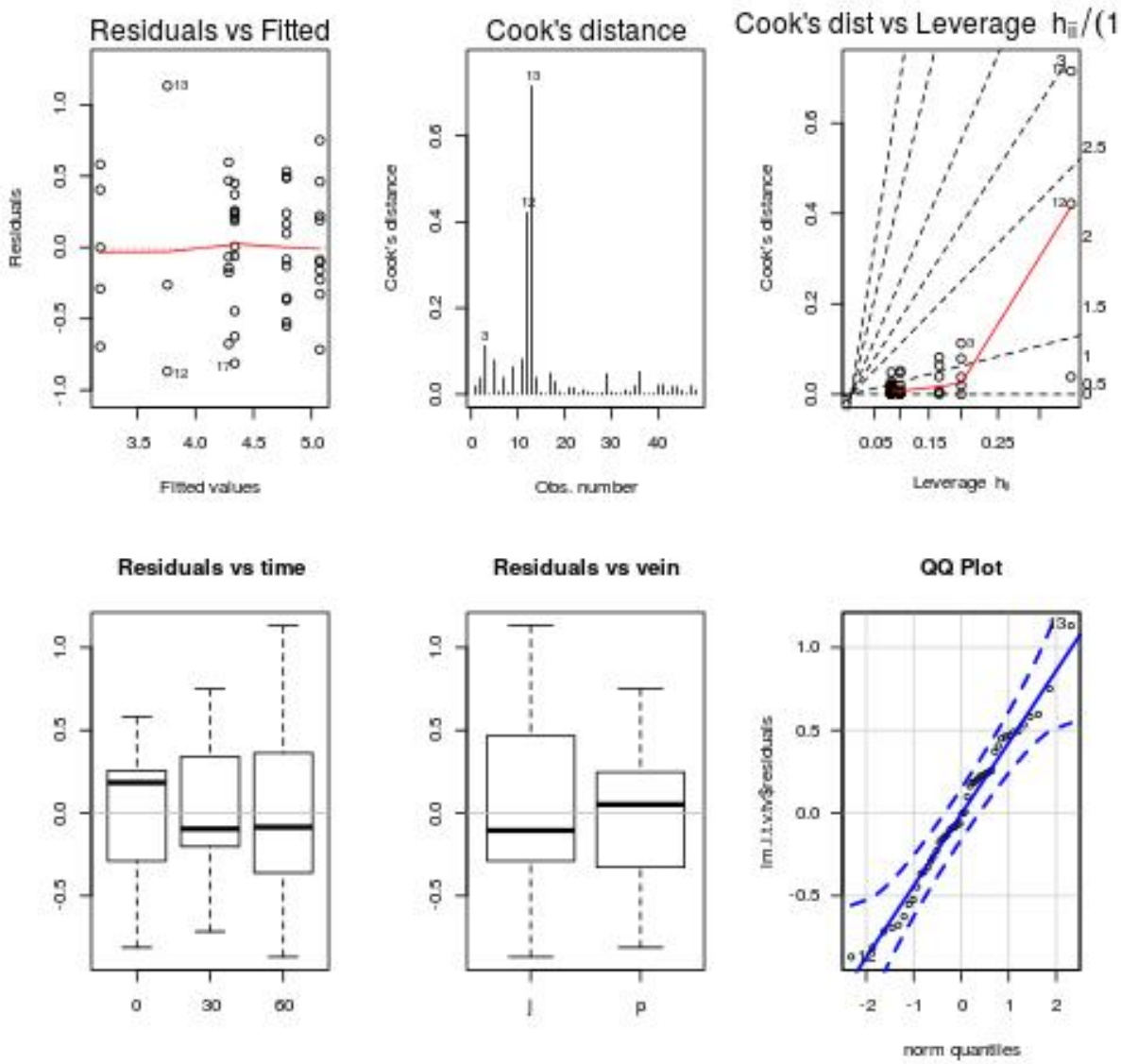
```
> # unbalanced, but highest-order interaction cell means will
> rat.mean.tv <- ddply(rat, .(time,vein), summarise, m = mean(
> rat.mean.tv
  time vein        m
1    0    j 3.179610
2    0    p 4.338230
3   30    j 4.286804
4   30    p 5.072433
5   60    j 3.759076
6   60    p 4.785463
```

```
> lsmeans(lm.i.t.v.tv, list(pairwise ~ time | vein), adjust =
$`lsmeans of time | vein`
vein = j:
 time   lsmean         SE df lower.CL upper.CL
 0    3.179610 0.2115533 42 2.752678 3.606542
 30   4.286804 0.1931208 42 3.897071 4.676538
 60   3.759076 0.2731141 42 3.207910 4.310243

vein = p:
 time   lsmean         SE df lower.CL upper.CL
 0    4.338230 0.1365570 42 4.062647 4.613814
 30   5.072433 0.1495908 42 4.770547 5.374320
 60   4.785463 0.1365570 42 4.509880 5.061047
```

Recall interaction model

```
> # interaction model
> lm.i.t.v.tv <- lm(loginsulin ~ time*vein, data = rat
+                   , contrasts = list(time = contr.sum, vein =
> Anova(lm.i.t.v.tv, type=3)
Anova Table (Type III tests)

Response: loginsulin
             Sum Sq Df   F value    Pr(>F)
(Intercept) 668.54  1 2987.5842 < 2.2e-16 ***
time          6.18  2   13.7996 2.475e-05 ***
vein          9.13  1   40.7955 1.101e-07 ***
time:vein     0.26  2    0.5797    0.5645
Residuals     9.40 42
```

Diagnostic plots



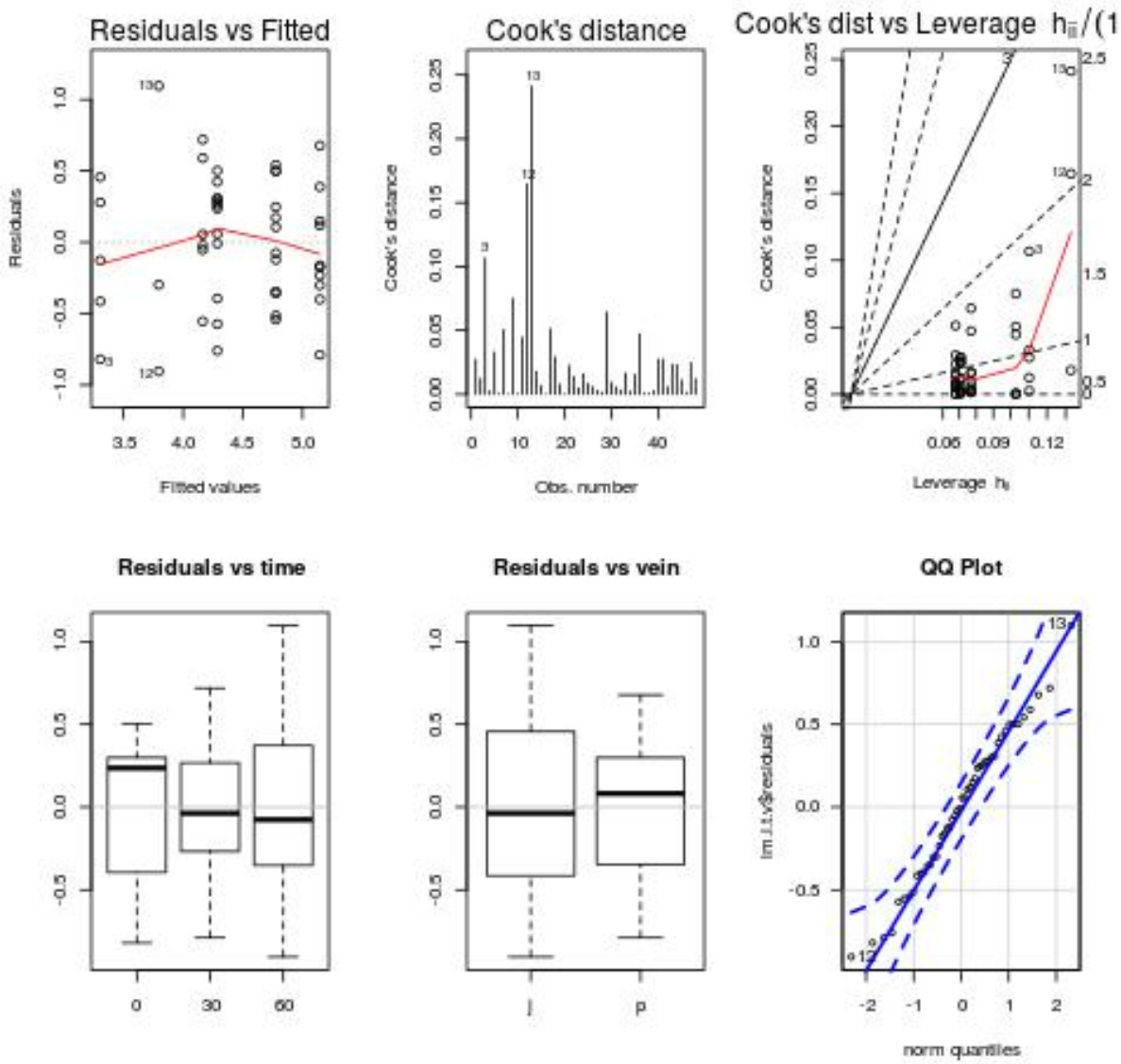
Diagnostic plots of the interaction model show the following features

- The normal quantile plot shows a rough straight line, suggesting the residuals are normal.
- The residuals vs the fitted (predicted) values show a random pattern.
- The plot of the Cooks distances indicate a few influential observations, but none of them are greater than 1.
- Normality assumption and constant variance assumption seem not violated.

Main effects model, removing non-significant interaction term

```
> lm.i.t.v <- lm(loginsulin ~ time+vein, data = rat
+                , contrasts = list(time = contr.sum,
vein = contr.sum))
> library(car)
Loading required package: carData
> Anova(lm.i.t.v, type=3)
Anova Table (Type III tests)

Response: loginsulin
             Sum Sq Df  F value    Pr(>F)
(Intercept) 708.02  1 3225.641 < 2.2e-16 ***
time          6.13  2   13.954 2.027e-05 ***
vein          9.32  1   42.467 5.863e-08 ***
Residuals     9.66 44
```

Diagnostic plots, main effect model



Normality test and constant variance test

```
> shapiro.test(lm.i.t.v$residuals)

Shapiro-Wilk normality test

data:  lm.i.t.v$residuals
W = 0.98609, p-value = 0.8343

> library(lmtest)
> bptest(loginsulin ~ time+vein, data = rat,studentize=FALSE)

Breusch-Pagan test

data:  loginsulin ~ time + vein
BP = 4.5673, df = 3, p-value = 0.2064
```

Diagnostic plots of the main effect model show the following features

- The normal quantile plot shows a rough straight line, suggesting the residuals are normal.
- The residuals vs the fitted (predicted) values show a random pattern.
- The plot of the Cooks distances indicate a few influential observations, but none of them are greater than 1.
- Normality assumption and constant variance assumption seem not violated.

Multiple comparison: unbalanced case, use lsmeans

```
> ### comparing lsmeans (unbalanced)
> library(lsmeans)
> ## compare levels of one factor at each level of
another factor separately
> # comapre different time levels
> lsmeans(lm.i.t.v, list(pairwise ~ time), adjust = "bonferror
$`lsmeans of time`
 time   lsmean         SE df lower.CL upper.CL
 0    3.795422 0.1177832 44 3.558045 4.032798
 30   4.655156 0.1186296 44 4.416074 4.894239
 60   4.285787 0.1291284 44 4.025546 4.546029

Results are averaged over the levels of: vein
Confidence level used: 0.95
```

```
$`pairwise differences of contrast`
 contrast   estimate        SE df t.ratio p.value
 0 - 30   -0.8597349 0.1636419 44  -5.254  <.0001
 0 - 60   -0.4903658 0.1665707 44  -2.944  0.0155
 30 - 60   0.3693690 0.1704300 44   2.167  0.1070

Results are averaged over the levels of: vein
P value adjustment: bonferroni method for 3 tests
```

```
> # comapre different vein levels
> lsmeans(lm.i.t.v, list(pairwise ~ vein),
adjust = "bonferroni")
$`lsmeans of vein`
 vein   lsmean         SE df lower.CL upper.CL
 j    3.754791 0.12659269 44 3.499660 4.009921
 p    4.736119 0.08054872 44 4.573784 4.898455

Results are averaged over the levels of: time
Confidence level used: 0.95

$`pairwise differences of contrast`
 contrast   estimate         SE df t.ratio p.value
 j - p    -0.9813288 0.1505883 44  -6.517  <.0001

Results are averaged over the levels of: time
```

```
# compare different combinations if interaction is
 significant
lsmeans(lm.i.t.v.tv, list(pairwise ~ vein*time),
adjust = "bonferroni")
```

The multiple comparisons of the significant main factors show that

- The mean log insulin level in rats after a fixed dose of insulin was injected are not significantly different for time levels of 30 and 60, while mean log insulin level for time level 0 is significantly lower than those for time levels 30 and 60.
- The mean log insulin level in rats after a fixed dose of insulin was injected through jugular vein is significantly lower than that of the portal vein.

use means or lsmeans, Type I or Type III SS?

Use lsmeans and Type III SS regardless of whether the design is balanced.

- The F -statistics based on Type III SSs are appropriate for unbalanced two- factor designs because they test the same hypotheses that were considered in balanced designs.
  ——That is, the Type III F -tests on the main effects check for equality in population means averaged over levels of the other factor.
  ——The Type III F -test for no interaction checks for parallel profiles.
- Given that the Type III F -tests for the main effects check for equal population cell means averaged over the levels of the other factor, multiple comparisons for main effects should be based on lsmeans.
- The Type I SS and F -tests and the multiple comparisons based on means should be ignored because they do not, in general, test meaningful hypotheses.