Lecture 21: Analysis of Variance

Pratheepa Jeganathan

11/08/2019

Recap

- What is a regression model?
- Descriptive statistics – graphical
- Descriptive statistics – numerical
- Inference about a population mean
- Difference between two population means
- Some tips on R
- Simple linear regression (covariance, correlation, estimation, geometry of least squares)
  - Inference on simple linear regression model
  - Goodness of fit of regression: analysis of variance.
  - F-statistics.
  - Residuals.
  - Diagnostic plots for simple linear regression (graphical methods).

Recap

- Multiple linear regression
  - Specifying the model.
  - Fitting the model: least squares.
  - Interpretation of the coefficients.
  - Matrix formulation of multiple linear regression
  - Inference for multiple linear regression
    - T-statistics revisited.
    - More F statistics.
    - Tests involving more than one β.
- Diagnostics – more on graphical methods and numerical methods
  - Different types of residuals
  - Influence
  - Outlier detection
  - Multiple comparison (Bonferroni correction)
  - Residual plots:
    - partial regression (added variable) plot,
    - partial residual (residual plus component) plot.

Recap

- Adding qualitative predictors
  - Qualitative variables as predictors to the regression model.
  - Adding interactions to the linear regression model.
  - Testing for equality of regression relationship in various subsets of a population

ANOVA

Outline

- One-way layout
- Two-way layout

ANOVA models

- Often, especially in experimental settings, we record *only* categorical variables.
- Such models are often referred to *ANOVA (Analysis of Variance)* models.
- These are generalizations of the two sample t-test.

Example: recovery time

- Suppose we want to understand the relationship between recovery time after surgery based on an patient's prior fitness.
- We group patients into three fitness levels: below average, average, above average.
- If a patient is in better shape before surgery, does it take less time to recover?

Example: recovery time

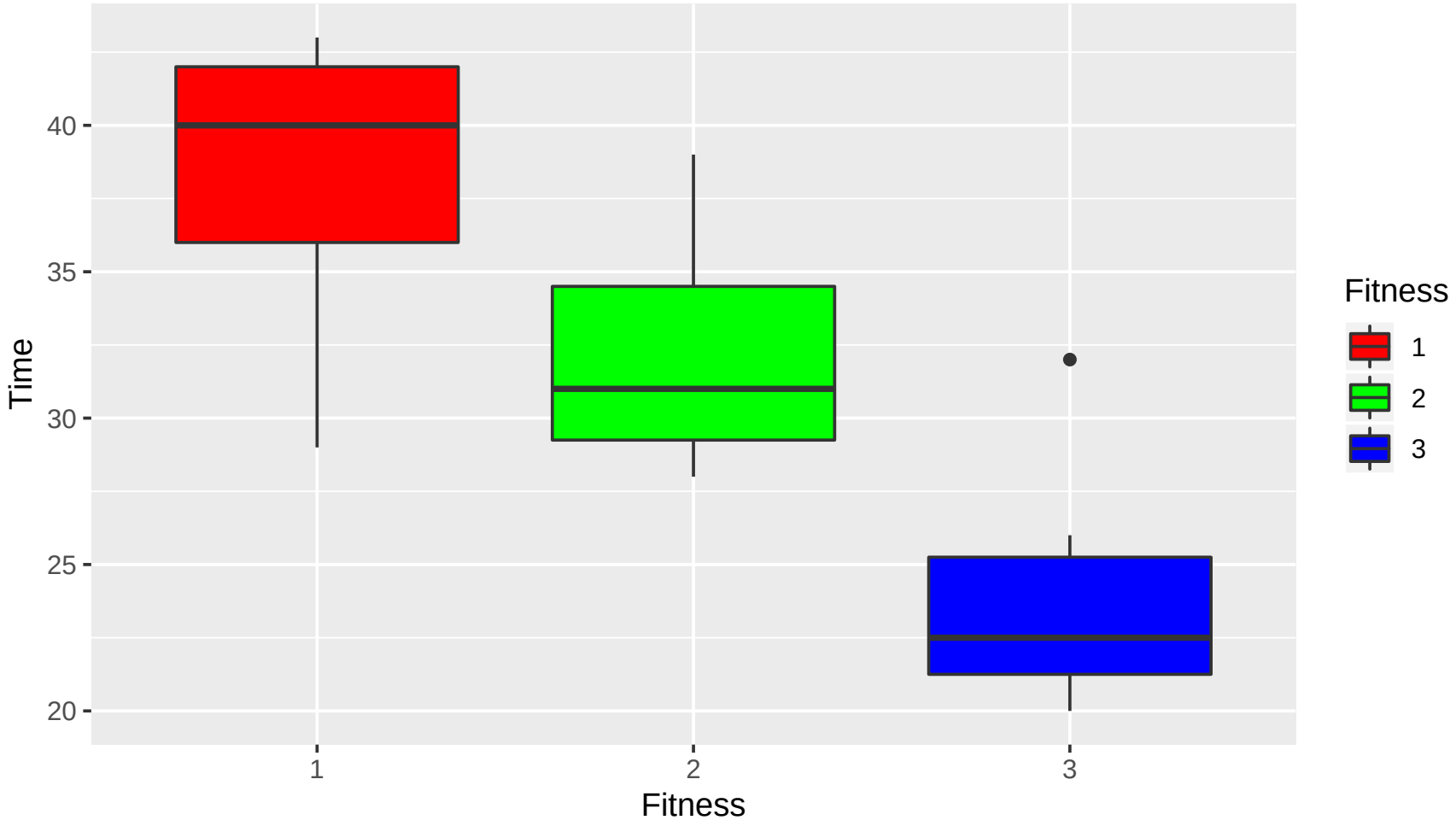
```
url = 'http://stats191.stanford.edu/data/rehab.csv'
rehab.table = read.table(url, header=T, sep=',')
rehab.table$Fitness <- factor(rehab.table$Fitness)
head(rehab.table)
```

```
##   Fitness Time
## 1       1   29
## 2       1   42
## 3       1   38
## 4       1   40
## 5       1   43
## 6       1   40
```

Example: recovery time

```
p = ggplot(data = rehab.table) +
  geom_boxplot(aes(x = Fitness,
    y = Time, fill = Fitness)) +
  scale_fill_manual(values =
      c('red','green','blue'))
```

Example: recovery time



- Boxplot shows that the reovery time for the `above average` fitness group is less than the `average` and `below average` group.

One-way ANOVA

- First generalization of two sample t-test: more than two groups.
- Observations are broken up into r groups with $n_i, 1 \leq i \leq r$ observations per group.
- Model:
$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \qquad \varepsilon_{ij} \overset{IID}{\sim} N(0, \sigma^2),$$
- Y_{ij} is the j-th measurement in i-th group, μ is the overall mean $\mu = \frac{1}{r}\sum_{i=1}^{r} \mu_i$, α_i is the main effect of group i on Y (That is, $\alpha_i = \mu_i - \mu$).
- Constraint: $\sum_{i=1}^{r} \alpha_i = 0$.
  - This constraint is needed for "identifiability".
  - This is "equivalent" to only adding $r - 1$ columns to the design matrix for this qualitative variable.

Fitting the model (One-way ANOVA)

- Model is easy to fit:

$$\widehat{Y}_{ij} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} = \overline{Y}_{i\cdot}$$

  - If observation is in i-th group: predicted mean is just the sample mean of observations in i-th group.

Testing (One-way ANOVA)

- Simplest question: is there any group (main) effect?

$$H_0 : \alpha_1 = \cdots = \alpha_r = 0$$

or

$$H_0 : \mu_1 = \cdots = \mu_r.$$

- Test is based on F-test with full model vs. reduced model.
  - Reduced model just has an intercept $Y_{ij} = \mu + \varepsilon_{ij}$.
- Other questions: is the effect the same in groups 1 and 2?

$$H_0 : \alpha_1 = \alpha_2?$$

Fitting the model (One-way ANOVA)

- `lm()` uses indicator variables.

```
rehab.lm = lm(Time ~ Fitness, data = rehab.table)
##summary(rehab.lm)
```

- `lm()` considers the `Fitness == 1` as the base level.

```
Call:
lm(formula = Time ~ Fitness, data = rehab.table)

Residuals:
   Min     1Q Median     3Q    Max
  -9.0   -3.0   -0.5    3.0    8.0

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   38.000      1.574  24.149  < 2e-16 ***
Fitness2      -6.000      2.111  -2.842  0.00976 **
Fitness3     -14.000      2.404  -5.824 8.81e-06 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.451 on 21 degrees of freedom
Multiple R-squared:  0.6176,    Adjusted R-squared:  0.5812
F-statistic: 16.96 on 2 and 21 DF,  p-value: 4.129e-05
```

H0: Time ~1 versus Ha: Time ~ Fitness

Fitting the model (One-way ANOVA)

- Upon inspection of the design matrix above, we see that the `(Intercept)` coefficient corresponds to the mean in `Fitness==1`, while `Fitness==2` coefficient corresponds to the difference between the groups `Fitness==2` and `Fitness==1`.

Fitting the model (One-way ANOVA)

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}.$$

- This is not the same *parameterization* we get when only adding $r - 1$ of 0-1 columns, but it gives the same *model*.
  - $\hat{Y}_{ij} = \bar{Y}_{i.}$
- The estimates of α's can be obtained from the estimates of β using R's default parameters.
- For a more detailed exploration into R's creation of design matrices, try reading the following tutorial on design matrices.

Fitting the model (One-way ANOVA)

- The design matrix is the indicator coding

```
head(model.matrix(rehab.lm))
```

```
##   (Intercept) Fitness2 Fitness3
## 1           1        0        0
## 2           1        0        0
## 3           1        0        0
## 4           1        0        0
## 5           1        0        0
## 6           1        0        0
```

- Recall that the rows of the `Coefficients` table above do not correspond to the α parameter.
  - `R` does use the $r - 1$ indicator variables.
  - `R` does not use the condition that α's and their sum would have to be equal to 0.

Fitting the model (One-way ANOVA)

- $\bar{Y}_{1\cdot}, \bar{Y}_{2\cdot}, \bar{Y}_{3\cdot}$ are as follows:

```
print(predict(rehab.lm,
  list(Fitness=factor(c(1,2,3)))))
```

```
##  1  2  3
## 38 32 24
```

```
c(mean(rehab.table$Time[rehab.table$Fitness == 1]),
  mean(rehab.table$Time[rehab.table$Fitness == 2]),
  mean(rehab.table$Time[rehab.table$Fitness == 3]))
```

```
## [1] 38 32 24
```

Fitting the model (One-way ANOVA)

```
overall_mean = mean(rehab.table$Time);overall_mean
```

```
## [1] 32
```

```
group_by(rehab.table, Fitness) %>%
  summarise(
    n_i = n(),
    hat_mean_i = mean(Time, na.rm = TRUE),
    hat_alpha_i = hat_mean_i - overall_mean,
    hat_sd_i = sd(Time, na.rm = TRUE)
  )
```

```
## # A tibble: 3 x 5
##   Fitness   n_i hat_mean_i hat_alpha_i hat_sd_i
##   <fct>   <int>      <dbl>       <dbl>    <dbl>
## 1 1           8         38           6     5.48
## 2 2          10         32           0     3.46
## 3 3           6         24          -8     4.43
```

ANOVA table

Source	SS	df	MS	$\mathbb{E}$ (MS)
Treatment	$\text{SSTR} = \sum_{i=1}^{r} n_i \left(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\right)^2$	$r-1$	$\text{MSTR} = \frac{\text{SSTR}}{r-1}$	$\sigma^2 + \frac{\sum_{i=1}^{r} n_i \alpha_i^2}{r-1}$
Error	$\text{SSE} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\cdot})^2$	$\sum_{i=1}^{r} (n_i - 1)$	$\text{MSE} = \frac{\text{SSE}}{\sum_{i=1}^{r} (n_i - 1)}$	σ^2

- Much of the information in an ANOVA model is contained in the ANOVA table.
- SSTR: sum of squares of treatment and SSE: Sum of squares of error.
- Note that *MSTR* measures "variability" of the "cell" means.
  - If there is a group effect we expect this to be large relative to *MSE*.
- We see that under $H_0 : \alpha_1 = \cdots = \alpha_r = 0$, the expected value of *MSTR* and *MSE* is σ^2.
  - This tells us how to test H_0 using ratio of mean squares, i.e. an F test.

Testing for any main effect

- Rows in the ANOVA table are, in general, independent.
- Therefore, under H_0

$$F = \frac{MSTR}{MSE} = \frac{\frac{SSTR}{df_{TR}}}{\frac{SSE}{df_E}} \sim F_{df_{TR},df_E}$$

  the degrees of freedom come from the *df* column in previous table.
- Reject H_0 at level α if $F \geq F_{1-\alpha,df_{TR},df_E}$.

ANOVA table

```
anova(rehab.lm)
```

```
## Analysis of Variance Table
##
## Response: Time
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Fitness    2    672  336.00  16.962 4.129e-05 ***
## Residuals 21    416   19.81
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Testing for any main effect

- Relationship between the columns in the above ANOVA table.

```
F = 336.00 / 19.81
pval = 1 - pf(F, 2, 21)
print(data.frame(F,pval))
```

```
##           F          pval
## 1 16.96113 4.129945e-05
```

ANOVA in R (using aov())

```
rehab.aov = aov(Time ~ Fitness,
  data = rehab.table)
summary(rehab.aov)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Fitness      2    672   336.0   16.96 4.13e-05 ***
## Residuals   21    416    19.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

- We can conclude at 5% significance level that at least one of the effects is non zero (or at least one of means μ_i is different from other)

Inference for linear combinations

- Suppose we want to "infer" something about

$$\sum_{i=1}^{r} a_i \mu_i,$$

where $\mu_i = \mu + \alpha_i$ is the mean in the i-th group.

  - For example:

$$H_0 : \mu_1 - \mu_2 = 0 \qquad \text{(same as } H_0 : \alpha_1 - \alpha_2 = 0\text{)?}$$

  - For example: Is there a difference between below average and average groups in terms of rehab time?

Inference for linear combinations

- We need to know

$$\text{Var}\left(\sum_{i=1}^{r} a_i \overline{Y}_{i\cdot}\right) = \sigma^2 \sum_{i=1}^{r} \frac{a_i^2}{n_i}.$$

- After this, the usual confidence intervals and t-tests apply.

Example

- Pairwise t-test
- $H_0 : \mu_i = \mu_k$ versus $H_a : \mu_i \neq \mu_k$ where $i \neq k = 1, 2, 3$.

```
pairwise.t.test(rehab.table$Time,
  rehab.table$Fitness,
  p.adjust.method = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  rehab.table$Time and rehab.table$Fitness
##
##   1       2
## 2 0.0293  -
## 3 2.6e-05 0.0067
##
## P value adjustment method: bonferroni
```

Example

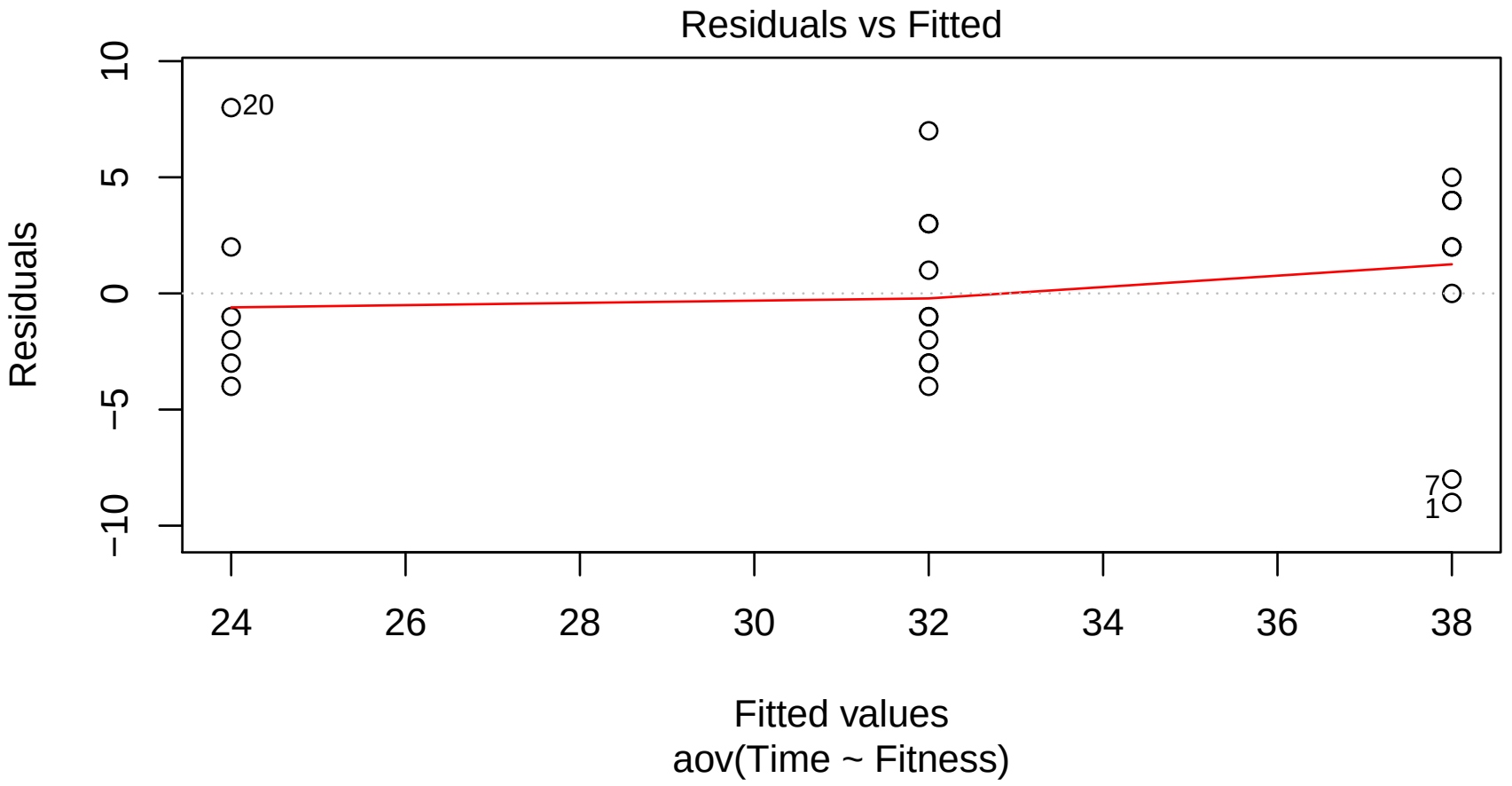
- Tukey's multiple pairwise-comparisons
- More about Tukey's multiple pairwise-comparisons

```
TukeyHSD(rehab.aov)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Time ~ Fitness, data = rehab.table)
##
## $Fitness
##     diff       lwr        upr     p adj
## 2-1   -6 -11.32141 -0.6785856 0.0253639
## 3-1  -14 -20.05870 -7.9413032 0.0000254
## 3-2   -8 -13.79322 -2.2067778 0.0060547
```

Diagnostics

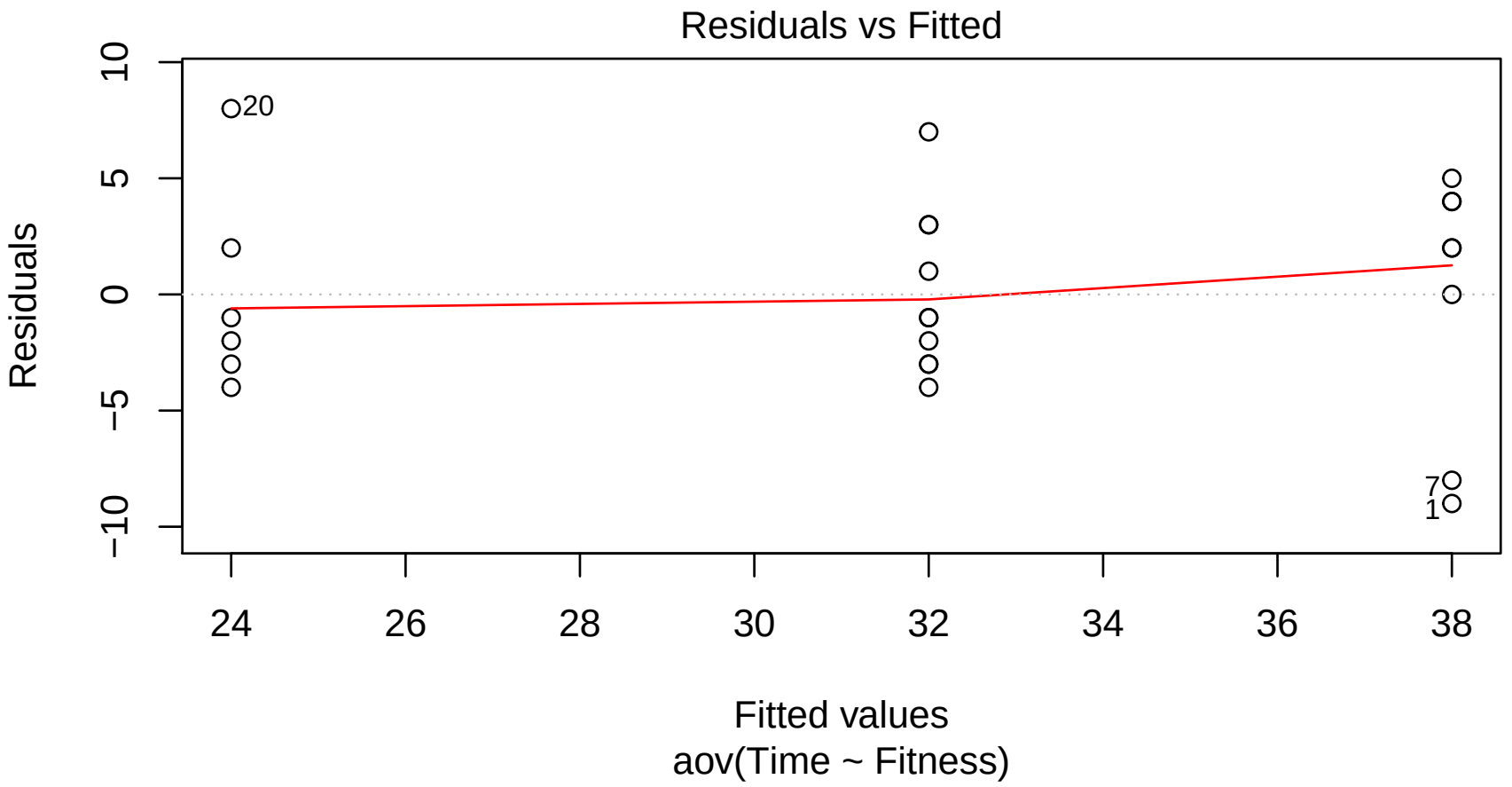
```
plot(rehab.aov, 1)
```



Diagnostics

- Variance is same in each group
- OK!

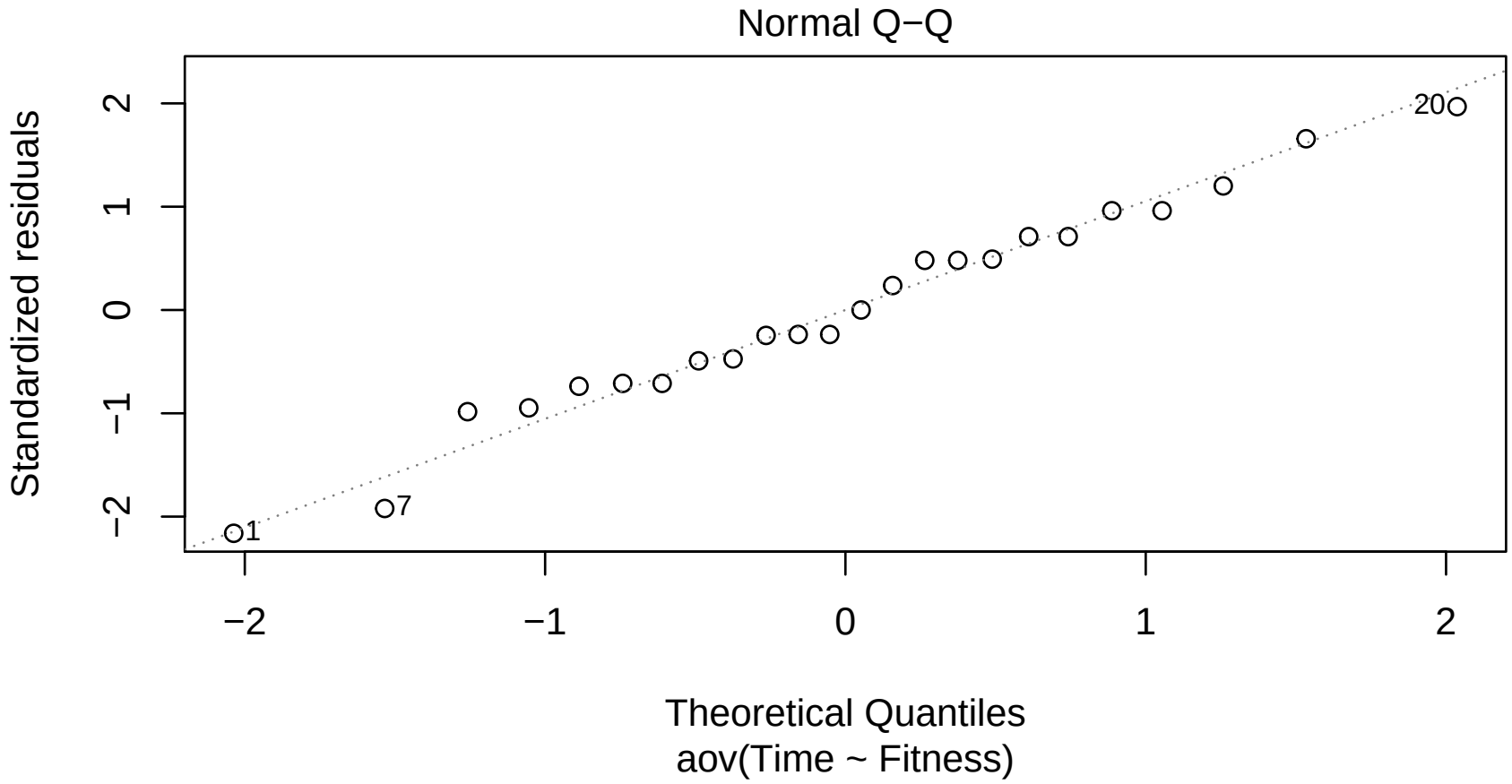
```
plot(rehab.aov, 1)
```



Diagnostics

- Normality assumption

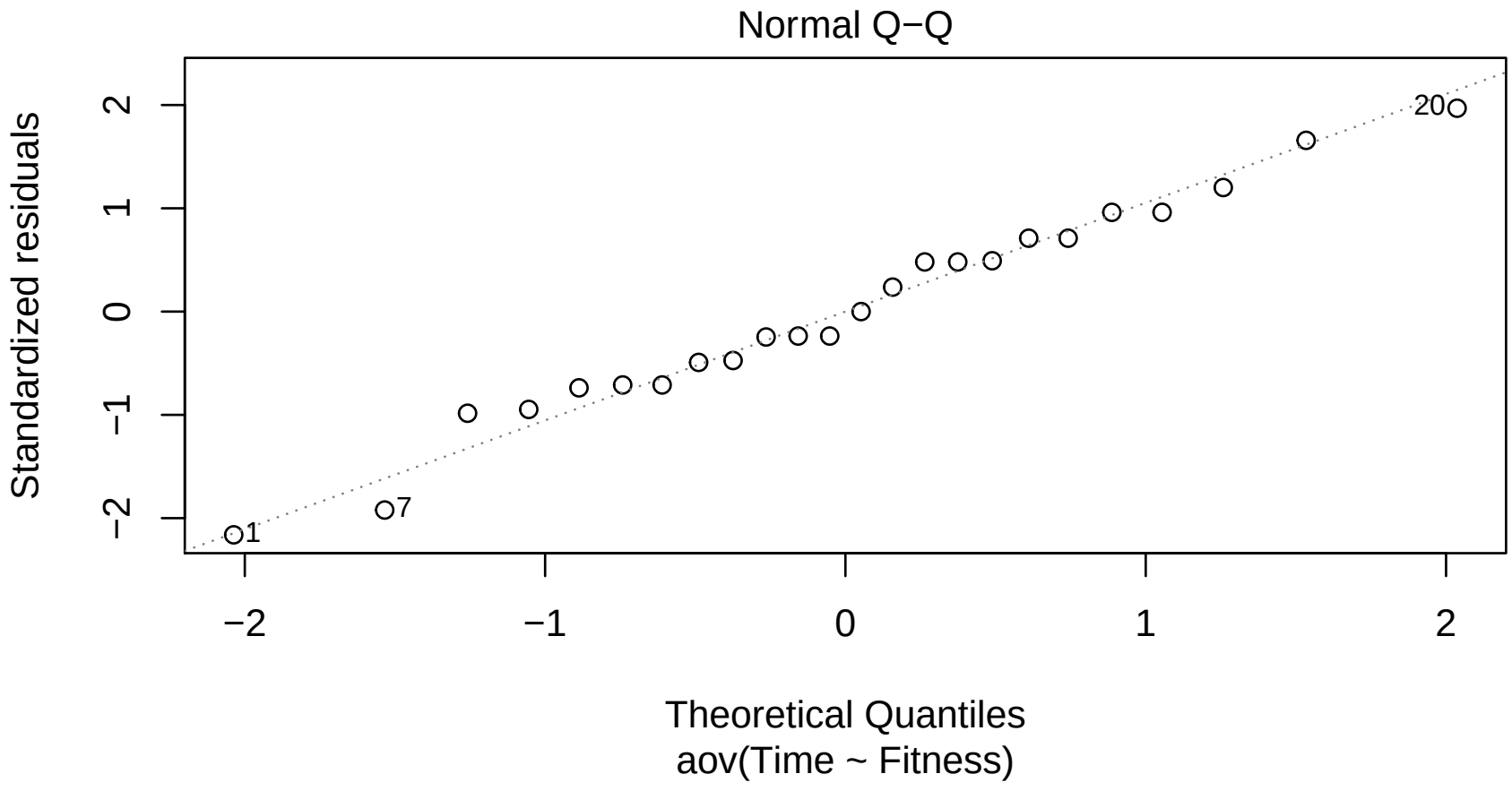
```
plot(rehab.aov, 2)
```



Diagnostics

- Normality assumption
- Looks OK!

```
plot(rehab.aov, 2)
```



Two-way ANOVA

- Often, we will have more than one variable we are changing.

Example

- After kidney failure, we suppose that the time of stay in hospital depends on weight gain between treatments and duration of treatment.
- We will model the `log` number of days as a function of the other two factors.

Variable	Description
Days	Duration of hospital stay (resposne)
Weight	How much weight is gained? (three levels)
Duration	How long under treatment for kidney problems? (two levels)

Example (Two-way ANOVA model)

```
url = 'http://statweb.stanford.edu/~jtaylo/stats191/data/ki
kidney.table = read.table(url, header=T)
kidney.table$D = factor(kidney.table$Duration)
kidney.table$W = factor(kidney.table$Weight)
kidney.table$logDays = log(kidney.table$Days + 1)
head(kidney.table)
```

```
##   Days Duration Weight ID D W   logDays
## 1    0        1      1  1 1 1 0.0000000
## 2    2        1      1  2 1 1 1.0986123
## 3    1        1      1  3 1 1 0.6931472
## 4    3        1      1  4 1 1 1.3862944
## 5    0        1      1  5 1 1 0.0000000
## 6    2        1      1  6 1 1 1.0986123
```

Two-way ANOVA model

- Second generalization of t-test: more than one grouping variable.
- Two-way ANOVA model:
  - r groups in first factor
  - m groups in second factor
  - n_{ij} in each combination of factor variables.
- Model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \qquad \varepsilon_{ijk} \sim N(0, \sigma^2).$$

- In kidney example, $r = 3$ (weight gain), $m = 2$ (duration of treatment), $n_{ij} = 10$ for all (i, j).

Questions of interest

Two-way ANOVA: main questions of interest

- Are there main effects for the grouping variables?

$$H_0 : \alpha_1 = \cdots = \alpha_r = 0, \qquad H_0 : \beta_1 = \cdots = \beta_m = 0.$$

- Are there interaction effects:

$$H_0 : (\alpha\beta)_{ij} = 0, 1 \leq i \leq r, 1 \leq j \leq m.$$

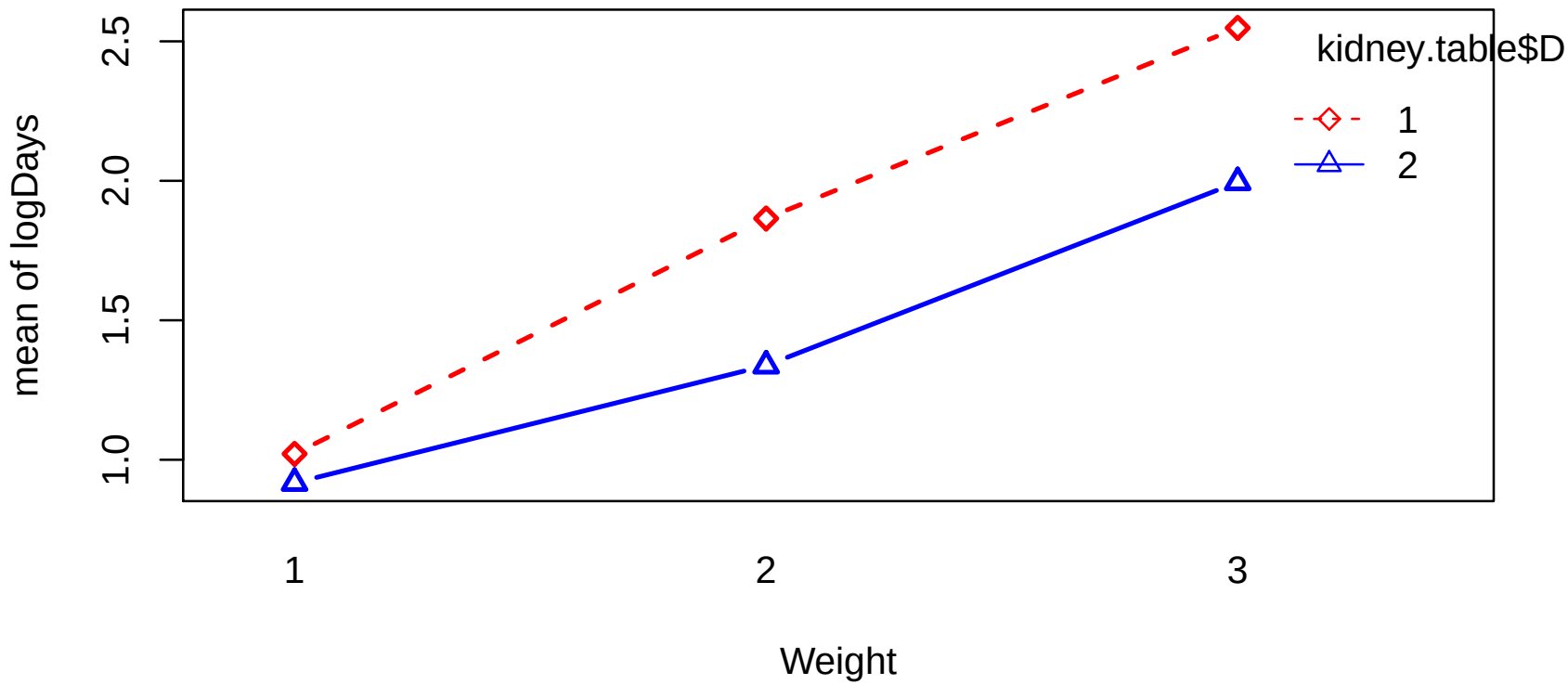
Interactions between factors

We've already seen these interactions in the IT salary example.

- An *additive model* says that the effects of the two factors occur additively – such a model has no interactions.
- An *interaction* is present whenever the additive model does not hold.

Interaction plot

```
interaction.plot(kidney.table$W, kidney.table$D,
  kidney.table$logDays, type='b',
  col=c('red', 'blue'), lwd=2,
  pch=c(23,24),
  xlab = "Weight",
  ylab = "mean of logDays")
```



Interaction plot

- When these broken lines are not parallel, there is evidence of an interaction.
- The one thing missing from this plot are errorbars.
  - The above broken lines are clearly not parallel but there is measurement error.
  - If the error bars were large then we might consider there to be no interaction, otherwise we might.

Parameterization

- Many constraints are needed, again for identifiability. Let's not worry too much about the details.
- Constraints:
  - $\sum_{i=1}^{r} \alpha_i = 0$
  - $\sum_{j=1}^{m} \beta_j = 0$
  - $\sum_{j=1}^{m} (\alpha\beta)_{ij} = 0, 1 \leq i \leq r$
  - $\sum_{i=1}^{r} (\alpha\beta)_{ij} = 0, 1 \leq j \leq m.$
- We should convince ourselves that we know have exactly $r * m$ free parameters.

Fitting the model

- Easy to fit when $n_{ij} = n$ (balanced)

$$\widehat{Y}_{ijk} = \overline{Y}_{ij\cdot} = \frac{1}{n}\sum_{k=1}^{n} Y_{ijk}.$$

- Inference for linear combinations of μ_{ij}'s

$$\text{Var}\left(\sum_{i=1}^{r}\sum_{j=1}^{m} a_{ij}\overline{Y}_{ij\cdot}\right) = \frac{\sigma^2}{n}\cdot\sum_{i=1}^{r}\sum_{j=1}^{m} a_{ij}^2.$$

- Usual t-tests, confidence intervals.

Fitting the model

```
group_by(kidney.table, W, D) %>%
  summarise(
    count = n(),
    hat_mean_ij = mean(logDays, na.rm = TRUE),
    hat_sd_ij = sd(logDays, na.rm = TRUE)
  )
```

```
## # A tibble: 6 x 5
## # Groups:   W [3]
##   W     D     count hat_mean_ij hat_sd_ij
##   <fct> <fct> <int>       <dbl>     <dbl>
## 1 1     1        10        1.02     0.831
## 2 1     2        10        0.917    0.759
## 3 2     1        10        1.87     0.751
## 4 2     2        10        1.34     0.726
## 5 3     1        10        2.55     0.693
## 6 3     2        10        1.99     0.619
```

ANOVA table

- In the balanced case, everything can again be summarized from the ANOVA table

Source	SS	DF	MS
A	$SSA = nm \sum_{i=1}^{r} \left(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots} \right)^2$	r-1	SSA/(r-1)
B	$SSB = nr \sum_{j=1}^{m} \left(\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots} \right)^2$	m-1	SSB/(m-1)
A:B	$SSAB = n \sum_{i=1}^{r} \sum_{j=1}^{m} \left(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots} \right)^2$	(m-1)(r-1)	SSAB/(m-1)(r-1)
ERROR	$SSE = \sum_{i=1}^{r} \sum_{j=1}^{m} \sum_{k=1}^{n} (Y_{ijk} - \overline{Y}_{ij\cdot})^2$	(n-1)mr	SSE/(n-1)mr

ANOVA table

Source	$\mathbb{E}(MS)$
A	$\sigma^2 + nm\frac{\sum_{i=1}^{r}\alpha_i^2}{r-1}$
B	$\sigma^2 + nr\frac{\sum_{j=1}^{m}\beta_j^2}{m-1}$
A:B	$\sigma^2 + n\frac{\sum_{i=1}^{r}\sum_{j=1}^{m}(\alpha\beta)_{ij}^2}{(r-1)(m-1)}$
ERROR	σ^2

Tests using the ANOVA table

- Rows of the ANOVA table can be used to test various of the hypotheses we started out with.
- For instance, we see that under $H_0 : (\alpha\beta)_{ij} = 0, \forall i, j$ the expected value of $SSAB$ and SSE is σ^2: use these for an F-test testing for an interaction.
- Under H_0,

$$F = \frac{MSAB}{MSE} = \frac{\frac{SSAB}{(m-1)(r-1)}}{\frac{SSE}{(n-1)mr}} \sim F_{(m-1)(r-1),(n-1)mr}.$$

Tests using the ANOVA table

```
kidney.aov = aov(logDays ~ D * W,
  data = kidney.table)
summary(kidney.aov)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## D             1  2.340   2.340   4.358   0.0416 *
## W             2 16.971   8.486  15.807 3.94e-06 ***
## D:W           2  0.636   0.318   0.592   0.5567
## Residuals    54 28.989   0.537
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

- $H_0 : (\alpha\beta)_{ij} = 0, \forall i, j$, we do not reject H_0 at 5% significance level.
- The main effects are significant at 5% significance level.

Tests using the ANOVA table

- Multiple pairwise comparison of effect of `Weight` groups

```
pairwise.t.test(kidney.table$logDays,
  kidney.table$W,
  p.adjust.method = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  kidney.table$logDays and kidney.table$W
##
##   1       2
## 2 0.030   -
## 3 2.8e-06 0.019
##
## P value adjustment method: bonferroni
```

Tests using the ANOVA table

- Multiple pairwise comparison of effect of `Duration` groups

```
pairwise.t.test(kidney.table$logDays,
  kidney.table$D,
  p.adjust.method = "bonferroni")
```

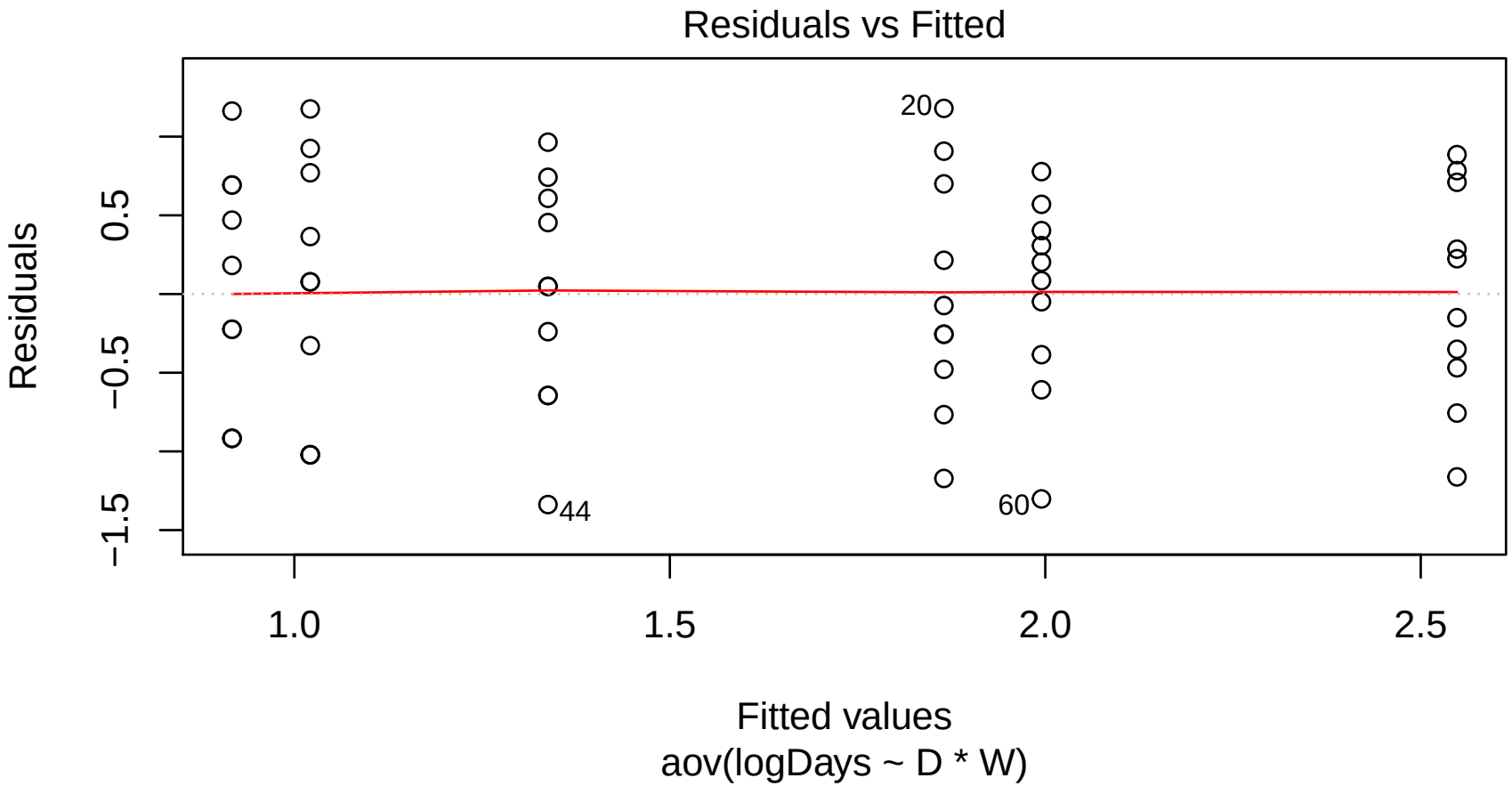
```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  kidney.table$logDays and kidney.table$D
##
##   1
## 2 0.093
##
## P value adjustment method: bonferroni
```

- We do not reject the H_0 (at 5% significance level) that the level of duration (of treatment) has no different effect on the number of stays in the hospital.

Diagnostics

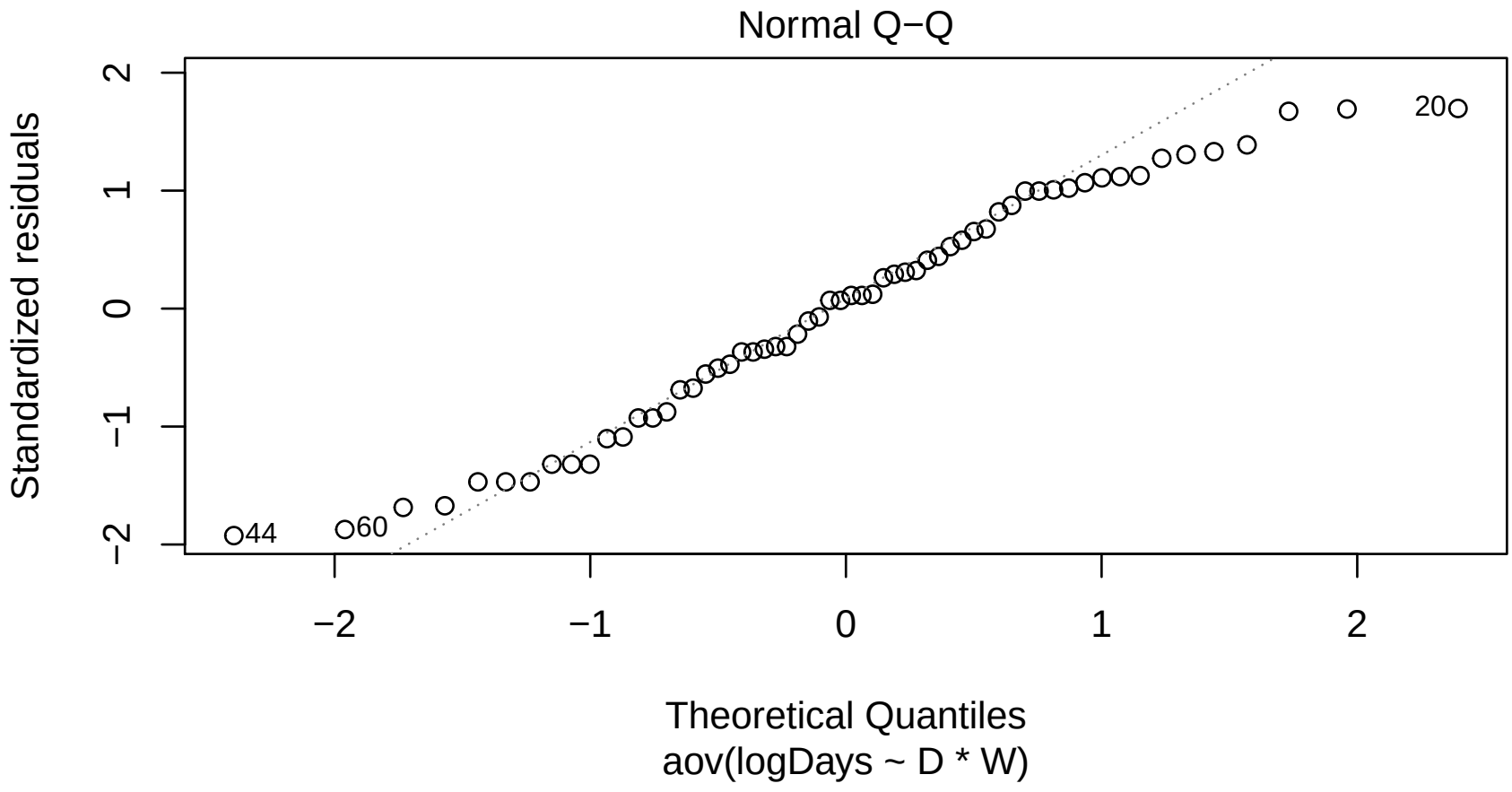
- homogeneity of variance - OK!

```
plot(kidney.aov, 1)
```



- Normality assumption - OK!

```
plot(kidney.aov, 2)
```



Fit using lm

$$Y_{ijk} = \beta_0 + \beta_1 D_1 + \beta_2 W_1 + \beta_3 W_2 + \beta_4 D_1 W_1 + \beta_5 D_1 W_2 + \epsilon$$

```
kidney.lm = lm(logDays ~ D*W,
   contrasts=list(D='contr.sum',
     W='contr.sum'), data = kidney.table)
#summary(kidney.lm)
```

```
Call:
lm(formula = logDays ~ D * W, data = kidney.table, contrasts = list(D = "contr.sum",
    W = "contr.sum"))

Residuals:
     Min       1Q   Median       3Q      Max
-1.33772 -0.51121  0.06302  0.62926  1.17950

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.61401    0.09459  17.063  < 2e-16 ***
D1           0.19747    0.09459   2.088   0.0416 *
W1          -0.64496    0.13377  -4.821  1.2e-05 ***
W2          -0.01264    0.13377  -0.095   0.9251
D1:W1       -0.14537    0.13377  -1.087   0.2820
D1:W2        0.06618    0.13377   0.495   0.6228
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7327 on 54 degrees of freedom
Multiple R-squared:  0.4076,    Adjusted R-squared:  0.3528
F-statistic: 7.431 on 5 and 54 DF,  p-value: 2.301e-05
```

Overall mean estimate

Ho: logDays ~ 1 versus Ha: logDays ~ D*W

Contrasts in R

- One level is the base level
- Compare other levels with the base level

```
contr.sum(3)
```

```
##   [,1] [,2]
## 1    1    0
## 2    0    1
## 3   -1   -1
```

```
contr.sum(2)
```

```
##   [,1]
## 1    1
## 2   -1
```

Model matrix in lm

- R uses indicator variables

```
head(model.matrix(kidney.lm))
```

```
##   (Intercept) D1 W1 W2 D1:W1 D1:W2
## 1           1  1  1  0     1     0
## 2           1  1  1  0     1     0
## 3           1  1  1  0     1     0
## 4           1  1  1  0     1     0
## 5           1  1  1  0     1     0
## 6           1  1  1  0     1     0
```

Finding predicted values using lm

- The most direct way to compute predicted values is using the `predict` function.
  - For example, $\bar{Y}_{11.}$ and the confidence interval for $\mu_{11.}$ are

```
predict(kidney.lm, list(D=factor(1),
  W=factor(1)), interval='confidence')
```

```
##        fit       lwr      upr
## 1 1.021156 0.5566306 1.485681
```

ANOVA using lm

```
anova(kidney.lm)
```

```
## Analysis of Variance Table
##
## Response: logDays
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## D          1  2.3397  2.3397  4.3583   0.04156 *
## W          2 16.9713  8.4856 15.8067 3.945e-06 ***
## D:W        2  0.6357  0.3178  0.5920   0.55675
## Residuals 54 28.9892  0.5368
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

- We can tests the interaction and the overall main effects (same as using aov)

Some caveats

- We can test the interaction using our usual approach.

```
anova(lm(logDays ~ D+W,
  data = kidney.table),
  lm(logDays ~ D*W,
    data = kidney.table))
```

```
## Analysis of Variance Table
##
## Model 1: logDays ~ D + W
## Model 2: logDays ~ D * W
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1     56 29.625
## 2     54 28.989  2   0.63566 0.592 0.5567
```

- But we cannot test the main effects using this approach

Some caveats

- Test the main effect of `Weight` factor variable using our ususal approach.

```
anova(lm(logDays ~ D,
  data = kidney.table),
  lm(logDays ~ D+W,
    data = kidney.table))
```

```
## Analysis of Variance Table
##
## Model 1: logDays ~ D
## Model 2: logDays ~ D + W
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     58 46.596
## 2     56 29.625  2    16.971 16.041 3.109e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Some caveats

- This F statistics value is not same as when we use anova(kidney.lm)

```
Analysis of Variance Table

Model 1: logDays ~ D
Model 2: logDays ~ D + W
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     58 46.596
2     56 29.625  2    16.971 16.041 3.109e-06 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Analysis of Variance Table

Response: logDays
          Df  Sum Sq Mean Sq F value    Pr(>F)
D          1  2.3397  2.3397  4.3583   0.04156 *
W          2 16.9713  8.4856 15.8067 3.945e-06 ***
D:W        2  0.6357  0.3178  0.5920   0.55675
Residuals 54 28.9892  0.5368
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Sum of squares

- Let's take Y response and 3 predictors X_1, X_2, X_3.
- $\text{SSR}(X_1, X_2, X_3)$: total variation explained by X_1, X_2, and X_3.
- $\text{SSR}(X_1|X_2)$: additional variation explained by X_1 when added to a model already containing X_2.

Extra sum of squares

- ESS measures the part of the SSE (sum of squares of error) that is explained by an added subset of predictors.
  - $\text{SSR}(X_1|X_2) = \text{SSE}(X_2) - \text{SSE}(X_1, X_2)$
  - Sequential sum of squares can be used to compute the total variation explained by X_1, X_2, and X_3. This computation of sum of squares is called the Type I sum of squares.

    $$\text{SSR}(X_1, X_2, X_3) = \text{SSR}(X_1) + \text{SSR}(X_2|X_1) + \text{SSR}(X_3|X_1, X_2)$$

  - Type I sum of squares can be used to test a term in the order they are listed in the model.

ANOVA table

- Need partial sum of squares $SSA = SS(A|B, AB)$, $SS(B|A, AB)$, $SSAB = SS(AB|A, B)$.
- We can compute the above sum of squares using Type III sum of squares.
- In the **balanced design** the ANOVA table is from the ANOVA table

Source	SS
A	$SSA = SS(A\|B, AB) = nm \sum_{i=1}^{r} \left(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots}\right)^2$
B	$SSB = SS(B\|A, AB) = nr \sum_{j=1}^{m} \left(\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots}\right)^2$
A:B	$SSAB = SS(AB\|A, B) = n \sum_{i=1}^{r} \sum_{j=1}^{m} \left(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots}\right)^2$
ERROR	$SSE = \sum_{i=1}^{r} \sum_{j=1}^{m} \sum_{k=1}^{n} (Y_{ijk} - \overline{Y}_{ij\cdot})^2$

Type I, II, III sum of squares

- Type I, II, III sum of squares will give different results (ANOVA table) for **unbalanced design**.
- `anova()` and `aov()` computes the sequential sum of squares
  - factors are tested in the order they are in the model.
- For the two-way layout, we want to test each term in the model in light of the every other term in the model.
- If the design is unbalanced, we can use `Anova()` (with contrasts sum) in `car` package to get the ANOVA table.

For the example

```
library(car)
Anova(lm(logDays ~ D * W, data = kidney.table,
         contrasts=list(D='contr.sum', W='contr.sum')),
      type = "III")
```

```
## Anova Table (Type III tests)
##
## Response: logDays
##              Sum Sq Df  F value    Pr(>F)
## (Intercept) 156.302  1 291.1532 < 2.2e-16 ***
## D             2.340  1   4.3583   0.04156 *
## W            16.971  2  15.8067 3.945e-06 ***
## D:W           0.636  2   0.5920   0.55675
## Residuals    28.989 54
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Recommended reading

Design and Analysis of Experiments, 10th Edition, Douglas C. Montgomery: PDF is available for an older edition.

Reference

- Lecture notes of Jonathan Taylor .