STAT 210
Applied Statistics and Data Analysis
Experimental Design II

Joaquin Ortega

Hypothesis Tests

Hypothesis Test of No Treatment Effect

In an experiment involving k treatment levels, a hypothesis of interest is whether the treatments are different in terms of their effect on the response variable. Thus, the null hypothesis would be

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_k \quad \textit{vs.} \quad H_1 : \text{ at least two of the } \tau_i \text{ differ.}$$

Even though it looks like the null hypothesis involves nonestimable parameters, it can easily be recasted in terms of estimable contrasts:

$$H_0 : \tau_2 - \tau_1 = 0, \text{ and } \tau_3 - \tau_1 = 0 \text{ and } \ldots \text{ and } \tau_k - \tau_1 = 0.$$

There are other ways of rewriting H_0 in terms of estimable contrasts, but they will always depend on $k - 1$ distinct contrasts, which is the number of degrees of freedom for treatment.

Hypothesis Test of No Treatment Effect

As we saw in the previous lecture, the basic idea of the analysis of variance is that the error sum of squares measures how well the model fits the data.

One way of testing whether the treatments have an effect is to compare the sums of squares for the complete model with that obtained with a model that assumes that the null hypothesis is true.

This last model is known as the *reduced model* and is

$$y_{ij} = \mu + \epsilon_{ij}$$

with the same hypothesis as before for the noise.

Hypothesis Test of No Treatment Effect

The least squares estimator for μ in this model is the overall mean:

$$\hat{\mu} = \bar{y}_{\bullet\bullet}$$

and the error sum of squares for the reduced model is

$$\begin{aligned} SST &= \sum_i \sum_j (y_{ij} - \bar{y}_{\bullet\bullet})^2 = \sum_i \sum_j (y_{ij}^2 - 2y_{ij}\bar{y}_{\bullet\bullet} + \bar{y}_{\bullet\bullet}^2) \\ &= \sum_i \sum_j y_{ij}^2 - n\bar{y}_{\bullet\bullet}^2. \end{aligned} \tag{1}$$

If the null hypothesis is false and treatment effects differ, SSE should be smaller than SST.

Hypothesis Test of No Treatment Effect

The test is based on the difference $SST - SSE$, relative to the size of the SSE, i.e. $(SST - SSE)/SSE$ and we reject H_0 if this quantity is large.

We used before the notation SSA for $SST - SSE$ and called it the treatment sum of squares. Using (**??**) and (1), it is given by

$$SSA = \sum_i \sum_j y_{ij}^2 - n\bar{y}_{\bullet\bullet}^2 - \left(\sum_i \sum_j y_{ij}^2 - \sum_i r_i \bar{y}_{i\bullet}^2 \right) = \sum_i r_i \bar{y}_{i\bullet}^2 - n\bar{y}_{\bullet\bullet}^2.$$

Equivalently,

$$SSA = \sum_i r_i (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2. \tag{2}$$

Hypothesis Test of No Treatment Effect

As we mentioned before, this sum represents the difference between treatment means and corresponds to the comparison between groups.

It can be shown that

- SSE/σ^2 has a χ^2_{n-k} distribution
- SSA/σ^2 has a χ^2_{k-1} distribution under H_0,
- these variables are independent.

In consequence, under H_0 the quotient

$$\frac{SSA/\sigma^2(k-1)}{SSE/\sigma^2(n-k)} \sim F_{k-1,n-k} \tag{3}$$

and we can use this relation to test H_0.

Hypothesis Test of No Treatment Effect

Recall that we defined $MSE = SSE/(n-k)$ and define $MSA = SSA/(k-1)$, then (3) becomes

$$\frac{MSA}{MSE} \sim F_{k-1,n-k}$$

and if msE and msA represent the observed values of these variables, the decision rule for testing H_0 at level of significance α is

$$\text{reject } H_0 \text{ if } \quad \frac{msA}{msE} > F_{1-\alpha,k-1,n-k} \tag{4}$$

Hypothesis Test of No Treatment Effect

As we have seen, the values for the sums of squares, degrees of freedom, mean squares and F test are usually written in an Analysis of Variance table, such as Table 3 below.

Table 3: Anova table for the one-way analysis of variance

Source	SS	d.f.	MS	F_{obs}	Critical F
Treatment	SSA	$k-1$	$MSA = \frac{SSA}{k-1}$	$F = \frac{MSA}{MSE}$	`qf(1-`α`, k-1, n-k)`
Error	SSE	$n-k$	$MSE = \frac{SSE}{n-k}$		
Total	SST	$n-1$			

Computational Formulae

$SSA = \sum_i r_i \bar{y}_{i\bullet}^2 - n\bar{y}_{\bullet\bullet}^2$ $\qquad$ $SSE = \sum_i \sum_j y_{ij}^2 - \sum_i r_i \bar{y}_{i\bullet}^2$

$SST = \sum_i \sum_j y_{ij}^2 - n\bar{y}_{\bullet\bullet}^2$

Hypothesis Test of No Treatment Effect

An Anova table can be obtained in R using the aov function. The inputs for this function are the same as those for the lm function we used previously, but the summary of an object created with the aov function is the Anova table. For the tire tread example the code is

```
library(PASWR)
```

```
## Loading required package: e1071
```

```
## Loading required package: MASS
```

```
## Loading required package: lattice
```

```
mod0 <- lm(StopDist ~ tire, data = Tire)
mod1 <- aov(StopDist ~ tire, data = Tire)
summary(mod1)
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## tire          3   5673  1891.0   5.328 0.00732 **
## Residuals    20   7099   354.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

Hypothesis Test of No Treatment Effect

The results for treatment are in the line labeled `tire` while the results corresponding to the errors are in the line labeled `Residuals`.

Results for *total* can be obtained adding up the corresponding terms in the table.

The F value is the ratio MSA/MSE and the last column labeled `Pr(>F)` is the probability of exceeding the calculated F-value when the null hypothesis is true, i.e. it is the p-value for the F test.

In this example we would conclude that there are significant differences among the average braking distances for different treads at the 0.05 level and also at the 0.01 level.

Hypothesis Test of No Treatment Effect

The same table is produced using the function `anova()` on the model we obtained with the `lm` function (`mod0`).

```
anova(mod0)
```

```
## Analysis of Variance Table
##
## Response: StopDist
##           Df Sum Sq Mean Sq F value   Pr(>F)
## tire       3 5673.1 1891.04  5.3278 0.007316 **
## Residuals 20 7098.8  354.94
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Hypothesis Test of No Treatment Effect

To see the complete list of quantities calculated by the `aov` function type `names(mod1)`:

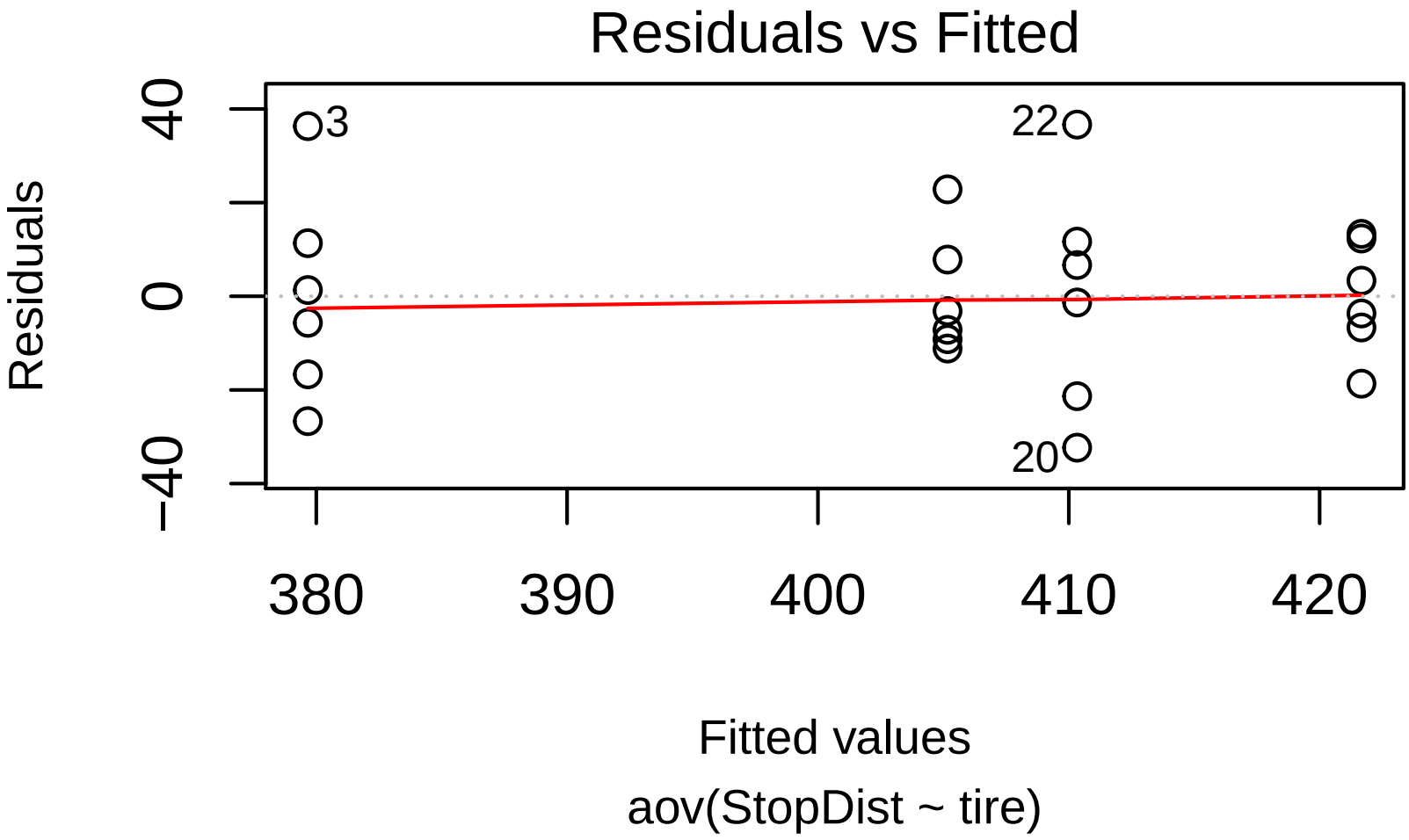
```
##  [1] "coefficients"  "residuals"     "effects"
##  [4] "rank"          "fitted.values" "assign"
##  [7] "qr"            "df.residual"   "contrasts"
## [10] "xlevels"       "call"          "terms"
## [13] "model"
```

```
## [1] "Df"      "Sum Sq"  "Mean Sq" "F value" "Pr(>F)"
```

Hypothesis Test of No Treatment Effect

Look at diagnostic plots to check the assumptions of the model.

```
plot(mod1, which=1, cex.lab=0.8, cex.sub=0.8)
```



Hypothesis Test of No Treatment Effect

Since we have six replications for each treatment level and there are only four x-values, the points appear vertically aligned at these values.

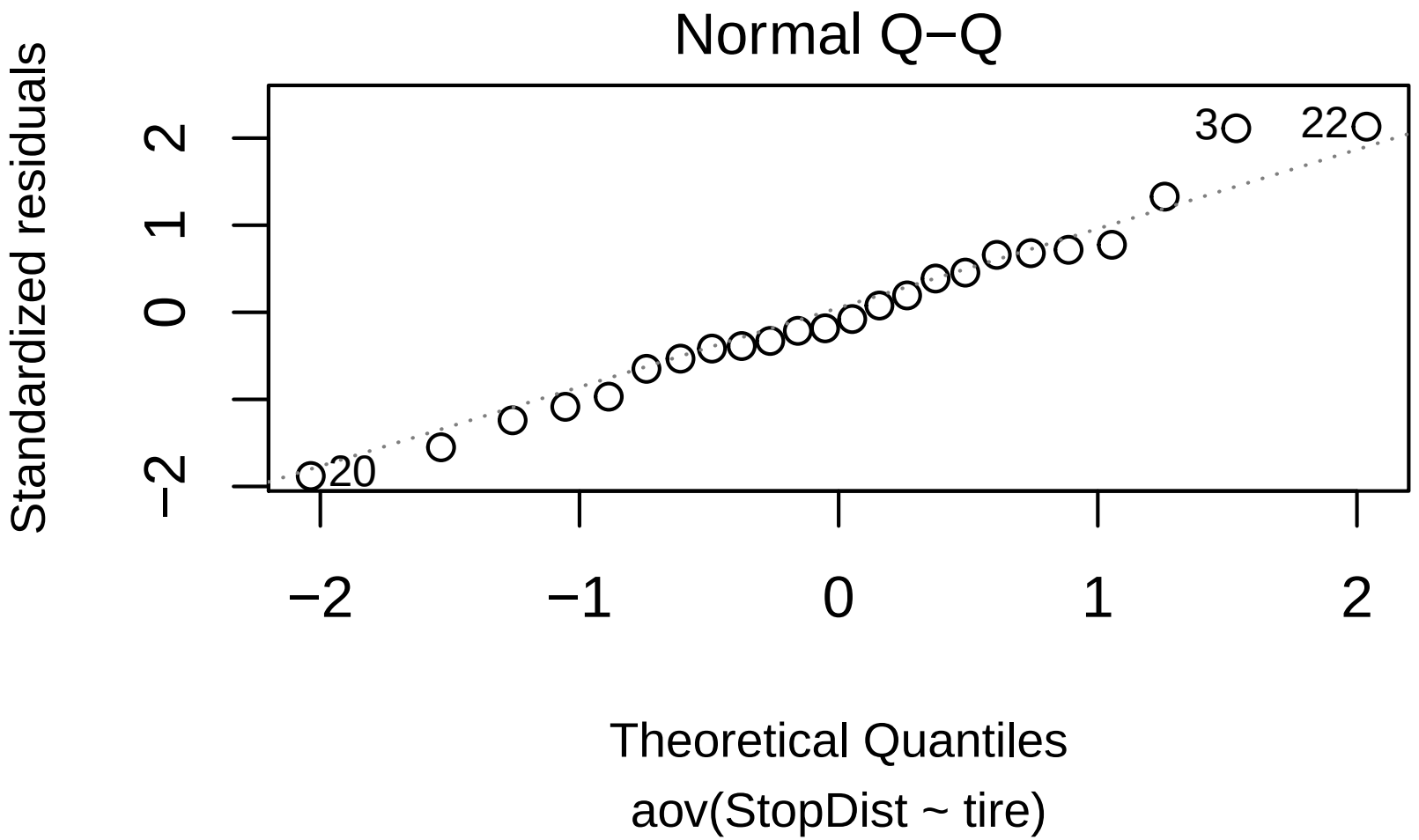
We look in this graphs for constant variance. We see that values in some cases appear to be more spread than in others, and this may be a sign of non-constant variance.

However, we only have a few points and this is difficult to determine.

Hypothesis Test of No Treatment Effect

We look for departures from normality in the standardized residuals. Considering the fit we observe, the normality assumption seems justified.

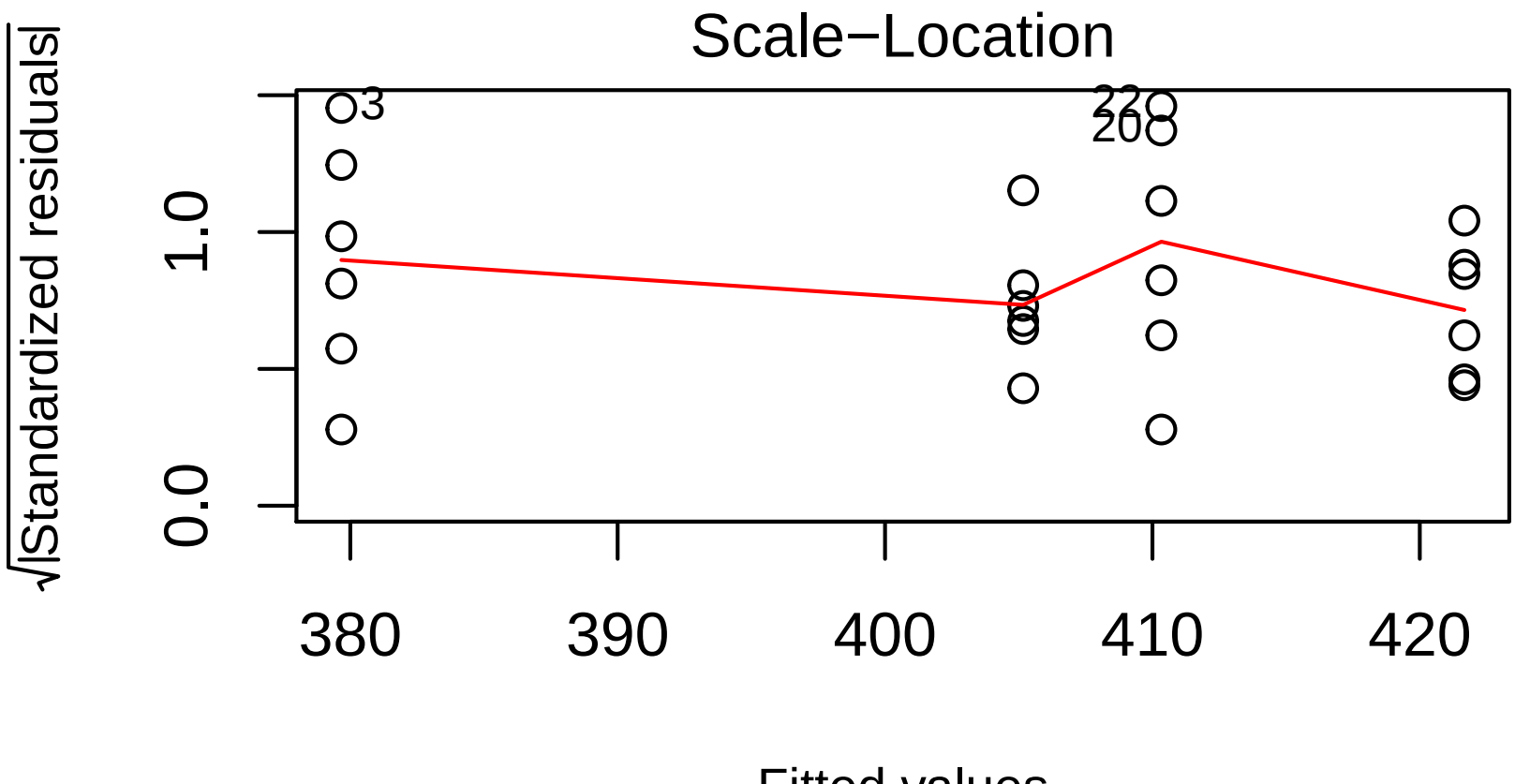
```
plot(mod1, which=2, cex.lab=0.8, cex.sub=0.8)
```



Hypothesis Test of No Treatment Effect

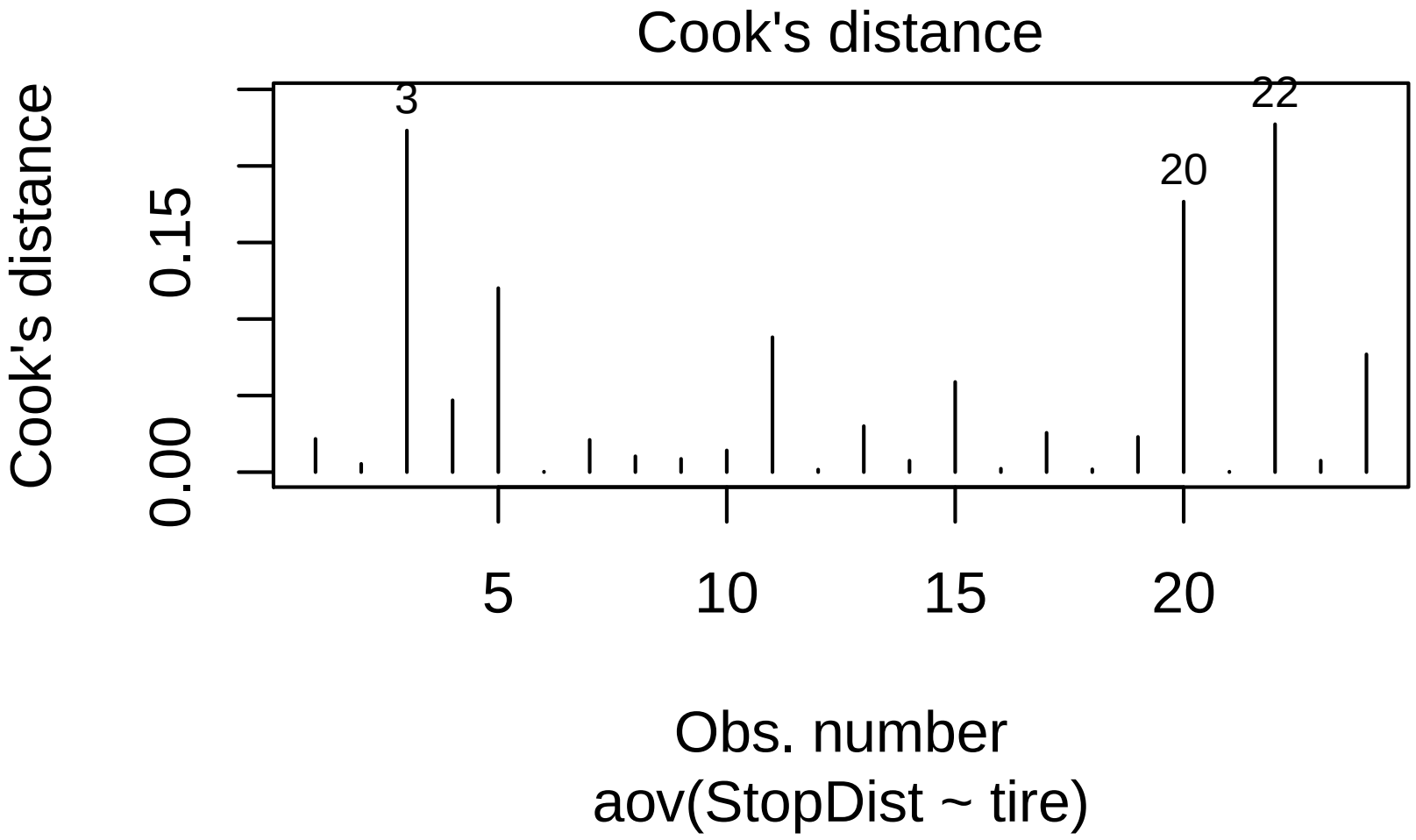
The third graph plots the square root of standardized residuals versus fitted values, and again we look for changes in the variance. As in the first graph, we see that some points seem to have more spread than others, perhaps pointing to heterocedasticity. However, there are very few points to be certain.

```
plot(mod1, which=3, cex.lab=0.8, cex.sub=0.8)
```



Hypothesis Test of No Treatment Effect

```
plot(mod1, which=4)
```



Hypothesis Test of No Treatment Effect

Another plot that is frequently useful is the plot of residuals versus the order in which the experiments have been performed.

This may help detecting if there is tendency in the results that should be taken into account.

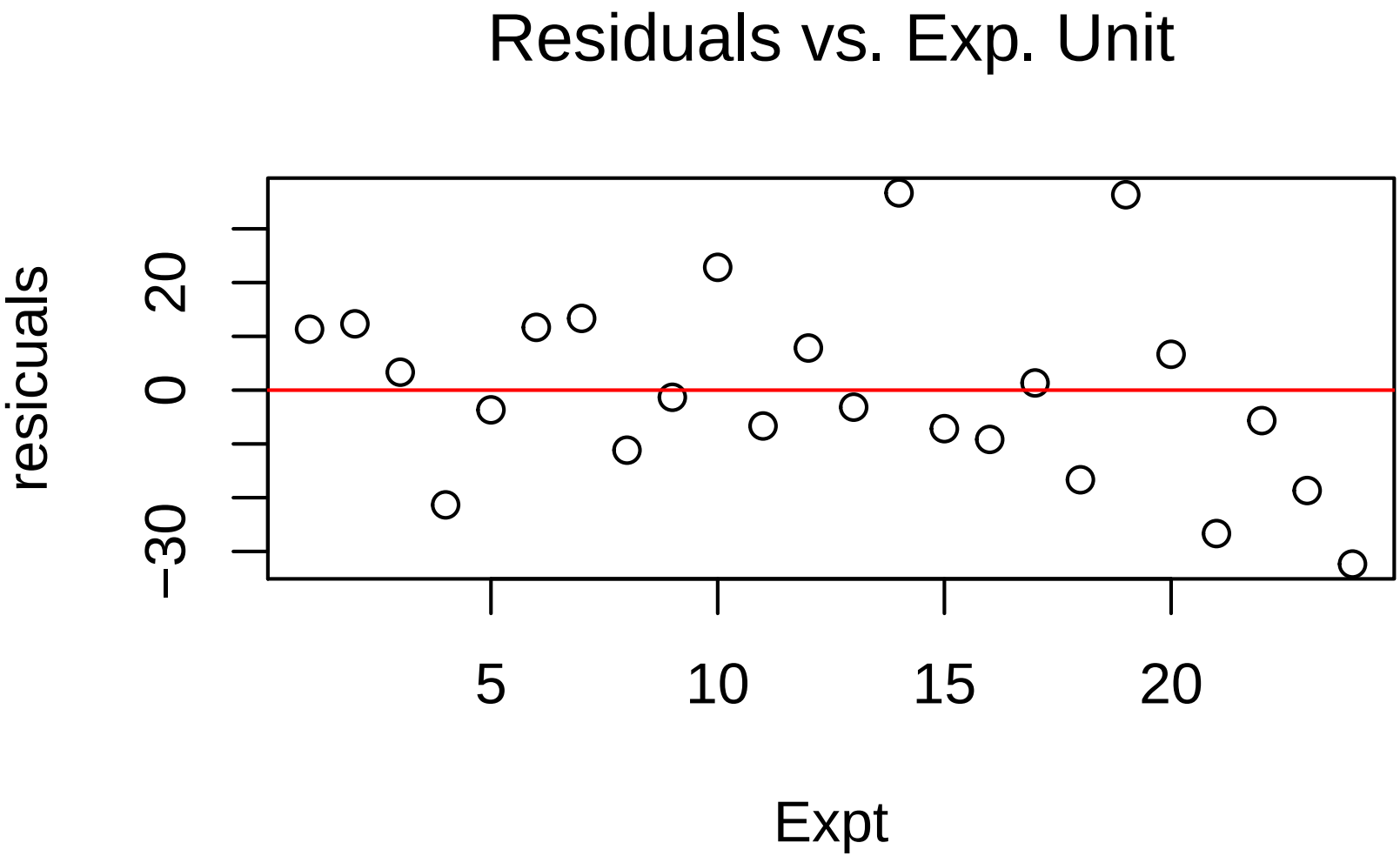
In this case that information is not available in the data set.

To illustrate the procedure, we will do the graph using a random ordering of the experiments.

To get the residuals for the model use the command `residuals(mode1)`.

Hypothesis Test of No Treatment Effect

```
plot(residuals(mod1) ~ sample(1:24,24), xlab='Expt',
     ylab='resicuals', main='Residuals vs. Exp. Unit',
     font.main=1)
abline(h=0, col='red')
```



Effect Sizes

To see the effect sizes in tabular form use `model.tables`

```
model.tables(mod1, se=T)
```

```
## Tables of effects
##
##  tire
## tire
##       A       B       C       D
## -24.542   0.958  17.458   6.125
##
## Standard errors of effects
##          tire
##         7.691
## replic.     6
```

Effect Sizes

Specifying means you get

```
model.tables(mod1,'means',  se=T)
```

```
## Tables of means
## Grand mean
##
## 404.2083
##
##  tire
## tire
##     A     B     C     D
## 379.7 405.2 421.7 410.3
##
## Standard errors for differences of means
##          tire
##         10.88
## replic.     6
```

Pairwise comparisons

Pairwise comparisons

If the result of the F test is to reject the null hypothesis of no treatment effects one is naturally interested in determining where the difference lies. For this, it becomes necessary to compare the individual groups.

This can be (partly) done looking at the regression coefficients using `summary` and `lm`.

Pairwise comparisons

```
summary(mod0)
```

```
##
## Call:
## lm(formula = StopDist ~ tire, data = Tire)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.333  -9.667  -2.250  11.417  36.667
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  379.667      7.691  49.363  < 2e-16 ***
## tireB         25.500     10.877   2.344 0.029497 *
## tireC         42.000     10.877   3.861 0.000973 ***
## tireD         30.667     10.877   2.819 0.010594 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.84 on 20 degrees of freedom
## Multiple R-squared:  0.4442, Adjusted R-squared:  0.3608
## F-statistic: 5.328 on 3 and 20 DF,  p-value: 0.007316
```

Pairwise comparisons

Recall that default coding for treatment in `R` means that the first value (`Intercept`) corresponds to the average value for the first treatment level $\hat{\mu} + \hat{\tau}_1$ while $\texttt{tireB} = \hat{\tau}_2 - \hat{\tau}_1$, $\texttt{tireC} = \hat{\tau}_3 - \hat{\tau}_1$, and $\texttt{tireD} = \hat{\tau}_4 - \hat{\tau}_1$

The t value in the table for `timeB` gives a t-test for comparing the first two groups (tread A vs. B, p-value $= 0.0294$), the next row in the table has a t-test for comparing groups 1 and 3 (tread A vs C, p-value$=0.0009$) and finally `tireD` compares tread A vs D, with p-value 0.011.

Pairwise comparisons

However, there is not a test for comparing groups 2 and 3, or 2 and 4 in the table.

One could redifine the factor `time` to include some of the missing comparison but then some others will dissapear.

This is not a convenient way to get all the tests, particularly if there are many levels for the treatment.

There exists a function for comparing all groups called `pairwise.t.test`.

However, one must correct for multiple testing.

Pairwise comparisons

The problem is that if we perform many tests, the probability of finding one of them to be significant by chance alone increases.

Consider one hundred statistical tests at the 5% level and assume all null hypothesis are true. We expect to reject 5 of them by chance alone. This is the expected value of Type I errors.

If the tests are independent, the probability of rejecting at least one hull hypothesis can be calculated using the binomial distribution:

```
1 - dbinom(0,100,0.05); 1 - dbinom(0,100,0.01)
```

```
## [1] 0.9940795
```

```
## [1] 0.6339677
```

Pairwise comparisons

One simple and frequently used method for correction is based on the Bonferroni inequalities and is known as the Bonferroni correction:

$$P(\cup_1^n B_i) \leq \sum_1^n P(B_i)$$

Thus, by dividing the significance level by the number of tests we get a test with true significance level smaller or equal to the nominal significance level.

This is equivalent to multiplying the p-values by the number of tests.

This test is conservative (tends to produce p values that are larger than they really are).

The Bonferroni correction should be applied at the planning stage of the experiment.

Pairwise comparisons

```
with(Tire, pairwise.t.test(StopDist, tire,
                           p.adjust.method = 'bonferroni'))
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  StopDist and tire
##
##   A      B      C
## B 0.1770 -      -
## C 0.0058 0.8696 -
## D 0.0636 1.0000 1.0000
##
## P value adjustment method: bonferroni
```

Pairwise comparisons

We have seen that the Bonferroni inequality can be used in the case of preplanned comparisons.

We will consider here briefly another method proposed by Tukey for comparisons of the form $H_0 : \tau_s = \tau_u$ for $s \neq u$ in favor of the alternative $H_1 : \tau_s \neq \tau_u$.

The procedure considers simultaneously all pairs of effects and adjusts the critical region by using the studentized range statistic instead of student's t-distribution.

Pairwise comparisons

The test is

$$\text{reject } H_0 \text{ if } \quad |\hat{\tau}_u - \hat{\tau}_s| > \sqrt{2} q_{I,n-k,1-\alpha/2} \hat{\sigma}(\bar{y}_{s\bullet} - \bar{y}_{u\bullet})$$

where $q_{I,n-k,1-\alpha}$ is the $1-\alpha$ percentile of the studentized range and $\hat{\sigma}(\bar{y}_{s\bullet} - \bar{y}_{u\bullet})$ is the estimated standard error for the difference between the averages.

If $X_1, \ldots X_I$ are independent random variables with a $N(\mu, \sigma^2)$ distribution and $R = \max_i X_i - \min_i X_i$ then $R/\hat{\sigma}$ follows the studentized range distribution.

This is only approximate when the sample sizes are unequal.

Pairwise comparisons

```
(mod1.tky <-TukeyHSD(mod1))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = StopDist ~ tire, data = Tire)
##
## $tire
##            diff          lwr      upr     p adj
## B-A  25.500000  -4.9446409 55.94464 0.1213153
## C-A  42.000000  11.5553591 72.44464 0.0049515
## D-A  30.666667   0.2220258 61.11131 0.0479540
## C-B  16.500000 -13.9446409 46.94464 0.4464584
## D-B   5.166667 -25.2779742 35.61131 0.9637307
## D-C -11.333333 -41.7779742 19.11131 0.7273681
```

Pairwise comparisons