STAT 461: Lab 7 - ANOVA Diagnostics

1 Assumptions in the ANOVA Model

The one-way ANOVA model with effects coding is written

$$Y_{it} = \mu + \tau_i + \epsilon_{it}, \quad i = 1, 2, \ldots, v \quad t = 1, 2, \ldots, r_i$$

$$\epsilon_{it} \stackrel{iid}{\sim} N(0, \sigma^2)$$

The v treatments are indexed by i and the number of replicates receiving the i-th treatment is r_i.

We have so far focused our analysis on the treatment means. In particular, we constructed the ANOVA table to find a test statistic to test the null hypothesis that all treatment means are equal. We also found ways to test for pairwise differences between treatment means, and tests for estimable contrasts (linear combinations) of treatment means.

All of these tests were constructed under the assumption that $\epsilon_{it} \stackrel{iid}{\sim} N(0, \sigma^2)$. We can decompose this assumption into three parts:

1.1 Independence of Errors

We assume that each trial is independent of all other trials, except for the effect τ_i of the treatment on the mean. Statistical independence of two trials means that knowing the result of one trial doesn't change the distribution of the other trial.

The most common causes of dependence in experimental data are confounding factors - something measured or unmeasured that affects the experiment. Randomization is a critical technique in experimental design because it can minimize the effect of any confounder.

1.2 Normality of Errors

We assume in the ANOVA model that the error terms are normally-distributed with zero mean. If the data are not normally-distributed, but instead come from some other distribution (exponential or binomial, for example), then we may not be able to trust our p-values, which were built by assuming normality.

1.3 Equal Error Variance Across Treatments

The next assumption is that all error terms have the same variance σ^2. It is common to see that different treatments seem to result in different means AND ALSO different variances.

2 Assessing ANOVA Assumptions

Before interpreting the results of any hypothesis tests, we need to check the three assumptions above.

2.1 Assessing Independence

The best way to assess whether or not one can assume independence of errors is to talk to the people who conducted the experiment. In particular, you should look for any potential confounding factor that was outside the randomization. For example, consider the greenhouse experiment in which 24 plants were potted and given one of four possible treatments (3 kinds of fertilizer and the control with no fertilizer). The 24 pots are arranged in two rows on a table in the greenhouse. The treatments were assigned randomly so that 6 plants received each treatment. Randomly assigning the treatments to the 24 pots randomizes over the spatial location of the pot on the greenhouse, and so will minimize any effect due to location (for example, if the table is near the door, one side of the table may be closer to the door and may experience colder air as the door is opened and closed).

Later in this course, we will consider approaches to modeling some dependence between trials through blocking factors or explicitly modeling confounding factors (ANCOVA). If such approaches are not appropriate, there may still be ways to account for dependence, but they are beyond the scope of this course.

2.2 Assessing Normality

Under the ANOVA model, $Y_{it} \sim N(\mu + \tau_i, \sigma^2)$. If the true means $\{\mu + \tau_i\}$ were known, then we could find the errors:
$$\epsilon_{it} = Y_{it} - (\mu + \tau_i) \sim N(0, \sigma^2).$$
As we don't know the true means, we could estimate them using the least-squares estimates. Our resulting estimates of the errors ϵ_{it} are called the **residuals** $\hat{\epsilon}_{it}$:

$$\hat{\epsilon}_{it} = Y_{it} - \overline{Y}_{i\cdot}, \quad i = 1, \ldots, v \quad t = 1, \ldots, r_i$$

If the residuals are approximately normal, then we can conclude that our ANOVA model assumption of normality is justified. We can assess normality of the residuals by looking at a histogram of the residuals, or by looking at a **normal probability plot**. A normal probability plot (also called a QQ-plot) plots the theoretical quantiles of a normal distribution against the estimated quantiles computed from the residuals. If the residuals are normally distributed, then the QQ-plot will be a straight line. Departures from normality, such as skew or heavy tails, will be seen as departures from the straight line. See the plots on the following page for examples.

Assessing normality can be hard at times, as our eyes are naturally drawn to any deviation from normality. Here are some suggestions:

- Small departures from normality are ok. The hypothesis tests we have discussed are only strongly affected if the data are VERY not normal, like some of the examples above.

- Even data that are normally distributed will show some departure from normality.

- It is often helpful to simulate data from an ANOVA model with the right number of observations and look at the QQ-plot from the residuals of the simulated data. Doing this multiple times will give you a sense of what sort of variation in histograms and QQ-plots is to be expected.

2.3 Assessing Constant Error Variance

The best ways to check for constant error variance among treatments are to plot the data (or a boxplot of the data) and to calculate the sample variances for each treatment.

The ANOVA hypothesis tests are not strongly affected by small differences in variance between treatments, especially if the sample sizes for each treatment are equal. As a **rule of thumb**, the difference in variance is
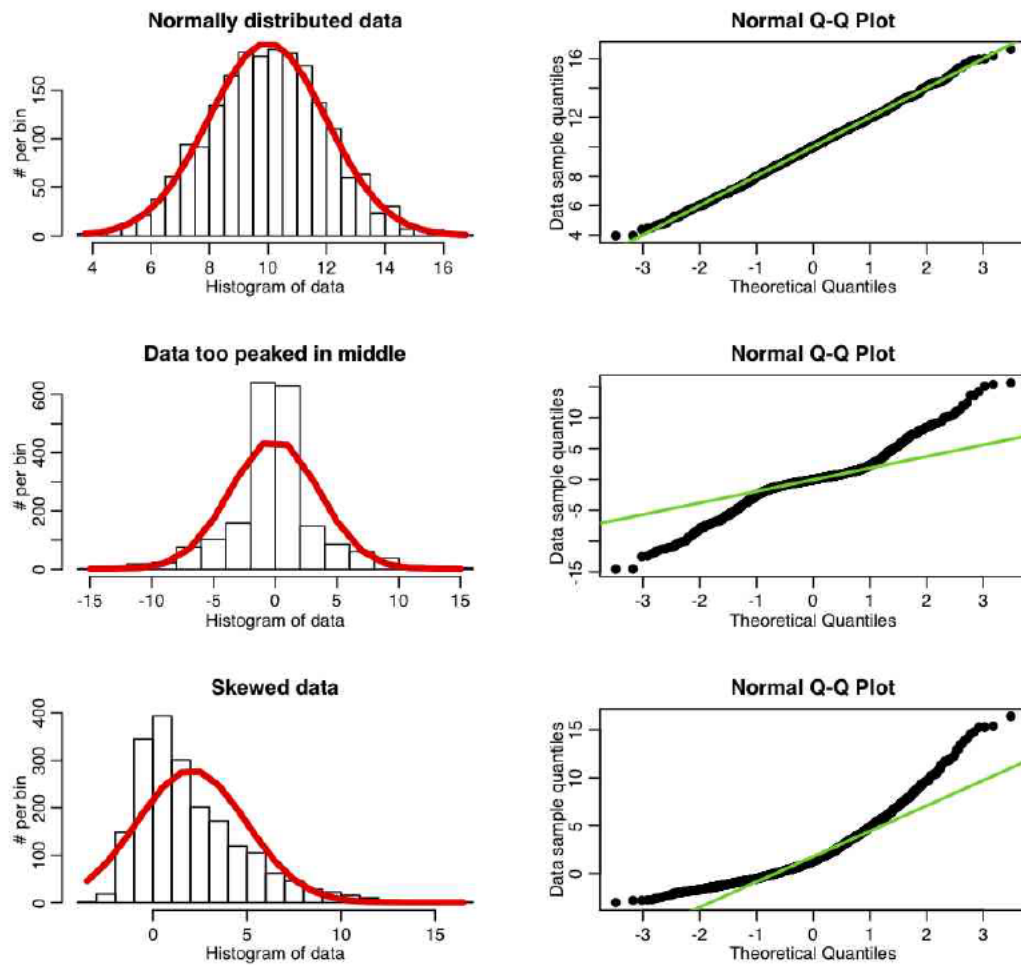
Figure 1:

too great to trust our hypothesis test results if the following is true:

$$\frac{\text{the largest treatment variance}}{\text{the smallest treatment variance}} > 3$$

This gives us a simple check for constant error variance. We can calculate the variance for each treatment:

$$s_i^2 = \frac{1}{r_i - 1} \sum_{t=1}^{r_i} (Y_{it} - \bar{Y}_{i.})^2$$

and if the ratio of the largest to the smallest variance is less than 3 we can trust the results of our hypothesis test. If the ratio is greater than 3, we will need to make some changes before trusting our hypothesis tests.

3 Addressing Normality and Constant Error Variance Through Transformation

We can often correct non-normality and unequal variances by **transforming the response variable**. That is, we find a transformation $f(\cdot)$ such that the transformed ANOVA model

$$f(Y_{it}) = \mu + \tau_i + \epsilon_{it}, \quad i = 1, 2, \ldots, v \quad t = 1, 2, \ldots, r_i$$

$$\epsilon_{it} \overset{iid}{\sim} N(0, \sigma^2)$$

results in residuals that are approximately normally-distributed with equal variances. When choosing a transformation, we often want to compress large values, and so one of the following transformations often works very well:

$$\sqrt{Y_{it}} \quad , \quad (Y_{it})^{1/3} \quad , \quad \log(Y_{it}).$$

For proportion data, the arcsin transformation is also very common.

$$\sin^{-1}(Y_{it}).$$

While there are some rules of thumb for choosing transformations (See Ch 18 of KNNL or Ch 5 of DV for details), it is simple to try multiple transformations to see which one results in residuals that are the closest to normal.

3.1 Interpreting ANOVA Results After Transformation

It is important to note that under the transformed model, we should state all hypothesis tests and results using the transformed function. For example, if we used a log transform, then the hypothesis of equal treatment means

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_v$$

is that the expected value of the log-response $\log(Y_{it})$ is the same for all treatments.

If the transformation $f(\cdot)$ is monotonic, it preserves the rank of the data and the treatment means. For example, if $W < Z$, then $\log(W) < \log(Z)$, so if we used a log transform of our response and found significant differences between the means of the first and second treatments on the log-transfored data, we could still conclude that there is a significant difference between the means of the first and second treatments on the original (not transformed) scale. This does not hold for transformations that are not monotonic, like for example squaring the response $f(Y_{it}) = Y_{it}^2$.

4 Example: Server Failure Times

A company has 3 servers in different cities (locations). They are interested in testing whether or not some locations experience more server outages than others. They record the waiting time between 5 server outages for each of the 3 locations. The data and code to analyze this data are given below.

```
loc<-c(1,1,1,1,1,
       2,2,2,2,2,
       3,3,3,3,3)
time<-c(4.41, 100.65, 14.45, 47.13, 85.21,
        8.24, 81.16, 7.35, 12.29, 1.61,
        106.19, 33.83, 78.88, 342.81, 44.33)
df<-data.frame(loc=as.factor(loc), time=time)
```

Run ANOVA on Untransformed Data
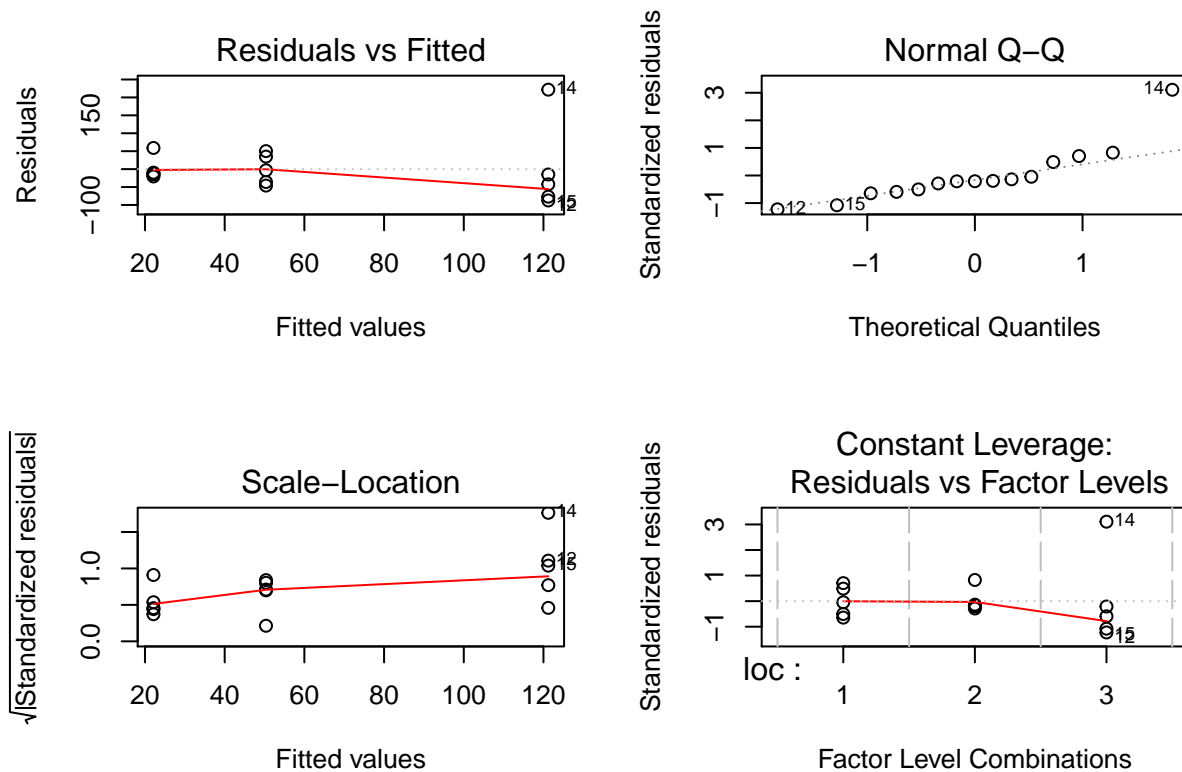
We first fit our One-Way ANOVA model like normal.

```
fit1<-aov(time~loc, data=df)
summary(fit1)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## loc          2  26053   13027    2.05  0.171
## Residuals   12  76239    6353
```

The results of the ANOVA F-test indicate that there aren't any differences between means, but the distribution and p-value of this test rely on the assumption that the errors are independent normal random variables with constant variance. We can check these visually. If you ask R to "plot" an aov object, it will give you multiple residual plots, which we will use to assess the assumptions of the model.

Diagnostics

```
par(mfrow=c(2,2))
plot(fit1)
```

Residuals vs Fitted

Normal Q–Q

Scale–Location

Constant Leverage: Residuals vs Factor Levels

loc :

The top-left plot ("Residuals vs Fitted") plots the residuals $(Y_{it} - \bar{Y}_{i\cdot})$ on the Y-axis, and the estimated means $(\bar{Y}_{i\cdot})$ on the x-axis. This is a good plot to use to check the constant variance assumption. Each verticle line of points is a different treatment. If each treatment has the same variance, then the "spread" on the points should be roughly equal. If the spread of points looks like a "trumpet" shape, or if there is one treatment that has vastly larger spread than others, then the assumption of constant variance is violated. In this case, the variance is clearly NOT constant, as the variance of treatments with small means (on the left of the plot) is much smaller than the variance of treatments with large means (on the right of the plot).

The top-right plot ("Normal Q-Q") shows a Normal quantile-quantile plot for the residuals. If they are normally-distributed, they will be mostly on the dotted line. Look back at Section 2.2 of this lab for examples of acceptable QQ plots. In this case, the residuals look pretty Normal, except for one point at the top right of the plot. Overall, I would say that the assumption of normality is not violated for this data.

The bottom two plots are less useful for us. These plots can help identify outliers, but with the small sample sizes we have for most experimental data it is very difficult to identify outliers, so we won't worry about that in this class (but see STAT 462 for a thorough treatment of this subject!).

Transform the response variable

From the above plots, we have determined that the assumption of constant variance is not met. We thus try to transform the response variable, fit the ANOVA model again on the transformed data, and see if the residuals from this transformed response variable are normally-distributed. We will try the square-root, log, and inverse transformations.
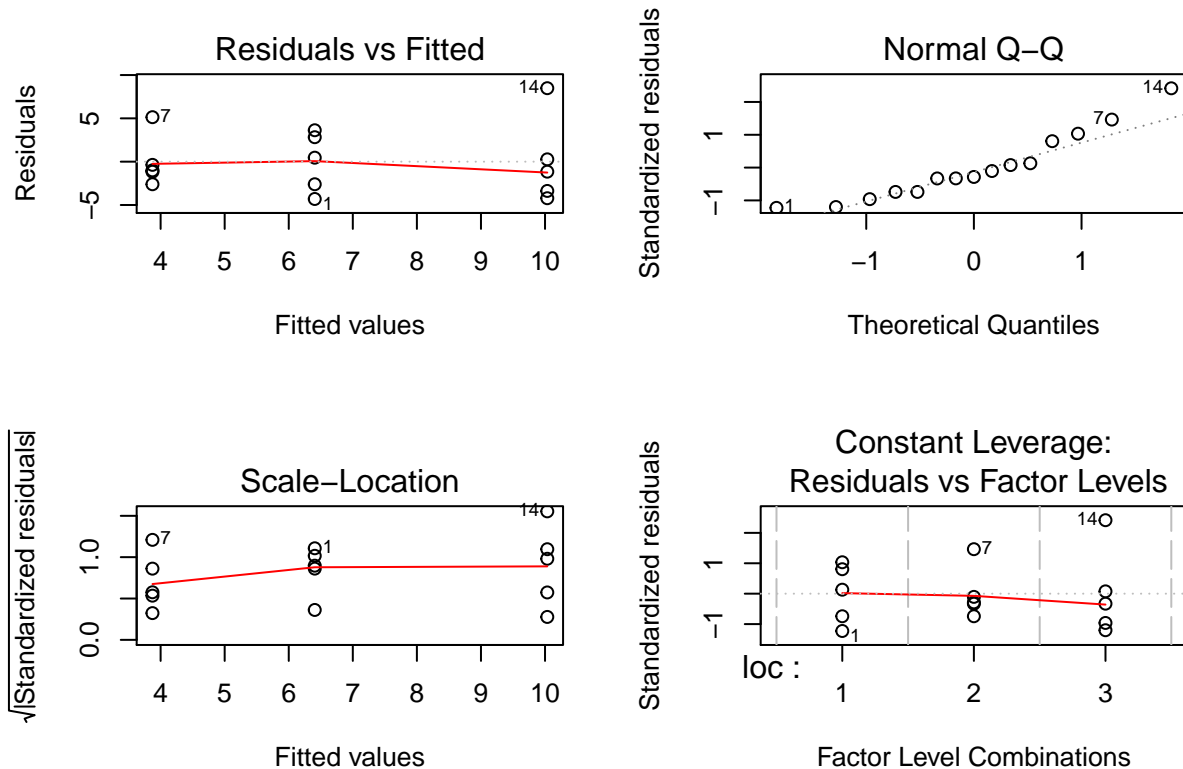
```r
df$SqrtTime<-sqrt(time)
df$LogTime<-log(time)
df$InvTime<-1/time
```

Run ANOVA on sqrt(response)

```
fit2<-aov(SqrtTime~loc, data=df)
summary(fit2)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## loc          2  95.93   47.97   3.114 0.0814 .
## Residuals   12 184.83   15.40
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow=c(2,2))
plot(fit2)
```


These residuals look pretty good. The constant variance assumption is now met (see the top-left plot). The normal QQ plot (top-left) shows some skew, with both high and low values above the dotted line, but overall this is much better.

Run ANOVA on 1/response

```
fit3<-aov(InvTime~loc, data=df)
summary(fit3)
```

```
##             Df  Sum Sq Mean Sq F value Pr(>F)
## loc          2 0.08472 0.04236   1.875  0.196
## Residuals   12 0.27114 0.02259
```

```
par(mfrow=c(2,2))
plot(fit3)
```
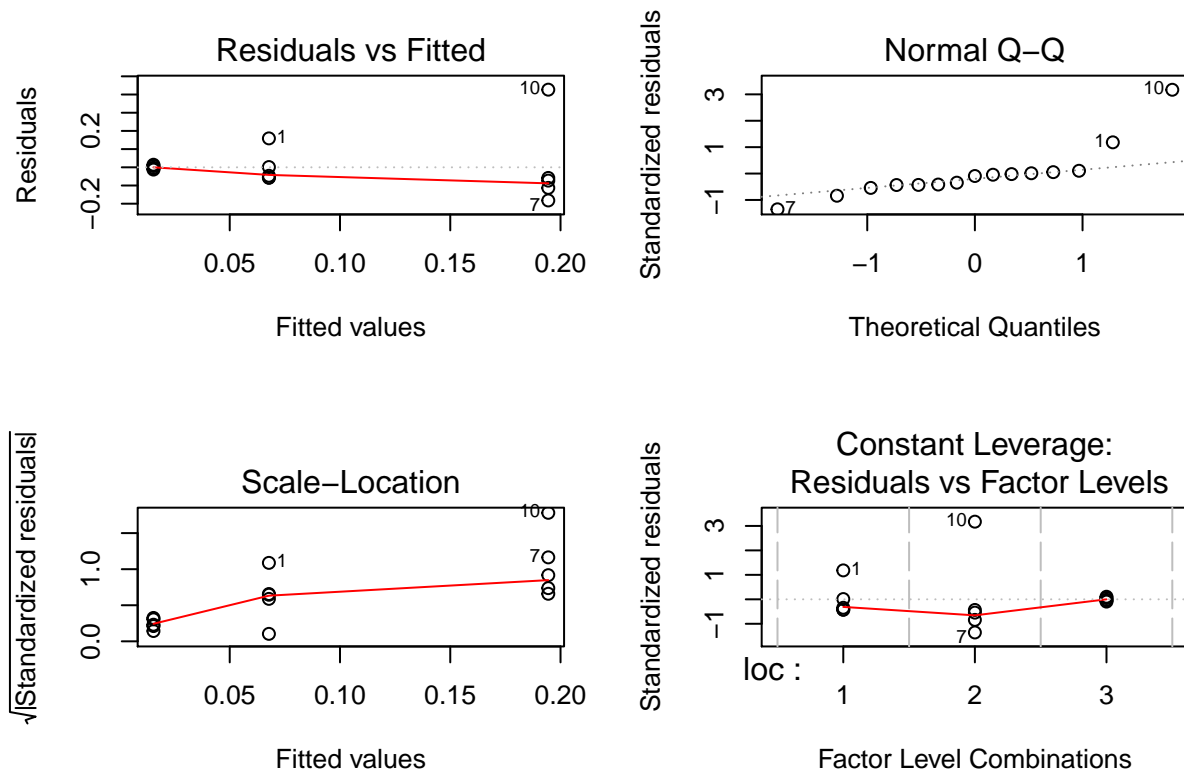
These residuals don't look good at all. The "trumpet" shape in the residuals vs fitted plot still show non-constant variance. So this transformation is not helpful.
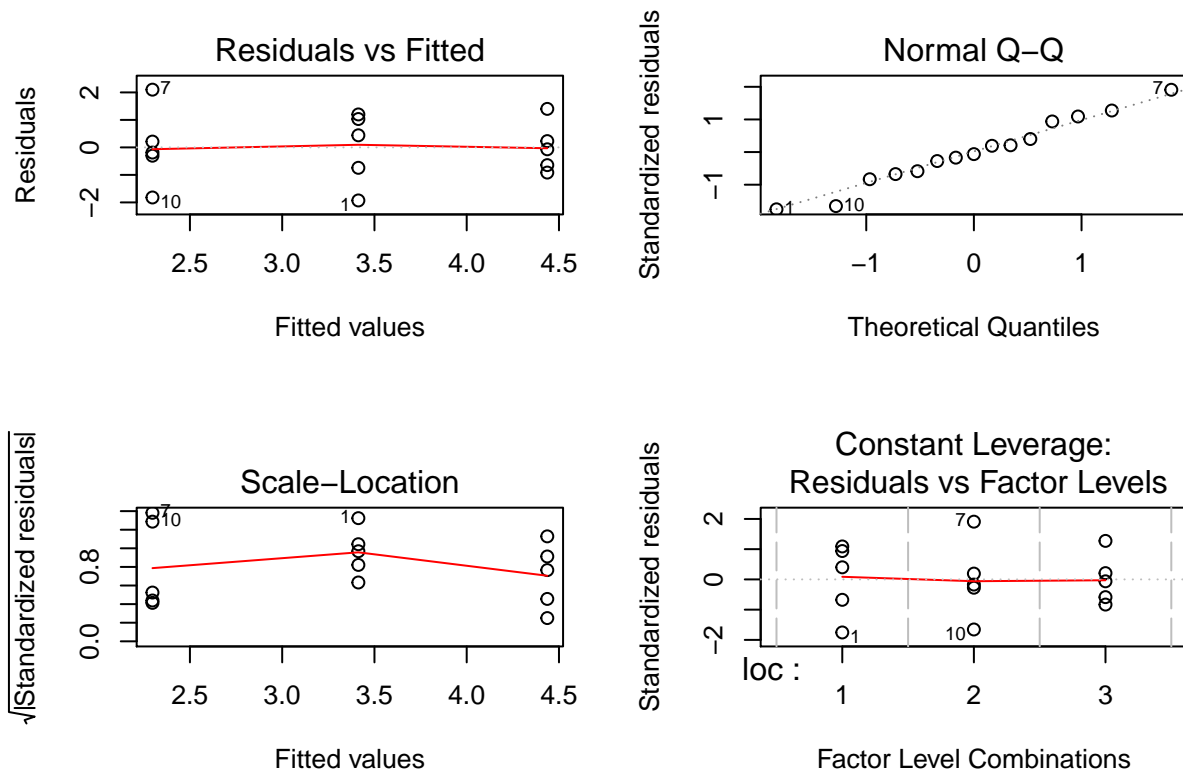
Run ANOVA on log(response)

```r
fit4<-aov(LogTime~loc, data=df)
summary(fit4)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## loc          2  11.45   5.726   3.789  0.053 .
## Residuals   12  18.14   1.511
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
par(mfrow=c(2,2))
plot(fit4)
```

Residuals vs Fitted

Residuals

Fitted values

Normal Q–Q

Standardized residuals

7

1 10

Theoretical Quantiles

Scale–Location

√|Standardized residuals|

7
8 10

1

Fitted values

**Constant Leverage:
Residuals vs Factor Levels**

Standardized residuals

7

1
10

loc :

1   2   3

Factor Level Combinations

These residuals look very good. The residuals vs fitted plot shows fairly constant variance, and the normal QQ plot shows very good normality. Thus, we will take this transformation and analyze it. If we look at the ANOVA table, we see that the p-value for the F-test of $H_0 : \tau_1 = \tau_2 = \tau_3$ is just above 0.05, so we do not reject the null hypothesis and conclude that there is no evidence to suggest that there are differences in means.

Note that this p-value is very different than the p-value under the untransformed data. P-values and hypothesis tests are ONLY valid when the residuals are approximately normal, and have approximately constant variance. We need to check assumptions BEFORE interpreting hypothesis tests of any kind!

Homework Assignment

Your homework will consider an experiment on battery life for different types and brands of battery. Two brands (a name brand and a generic brand) of two types (Alkaline and "Heavy Duty") of batteries were tested to see how long they could run continuously. This results in four categories, `AlkName` is for name-brand alkaline batteries, `AlkGen` is for generic alkaline batteries, `HDName` is for heavy duty name-brand batteries, and `HDGen` is for generic heavy duty batteries. Four batteries of each type were tested and the times to battery failure are recorded as below. Use the code below to read in the data:

```
type<-c("AlkName","AlkName","AlkName","AlkName","AlkGen","AlkGen","AlkGen","AlkGen",
        "HDName","HDName","HDName","HDName","HDGen","HDGen","HDGen","HDGen")
life<-c(100.668, 77.734,79.210,95.063,206.880,153.347,165.980,196.000,
        14.951,18.063,11.111,12.840,15.340,22.090,15.734, 14.440)
batt<-data.frame(type=type, life=life)
```

1. Plot the data.

2. For the battery data, do the following:

(a) Write out the one-way ANOVA model for this data.

(b) Show residual plots for this model. Are the residuals approximately normal? Justify your answer.

(c) Is the assumption of constant error variance among treatments justified? Explain your answer.

3. Now consider using the square-root of the battery life as a response variable. Repeat (a)-(c) above for this transformation.

4. Now consider using the log of the battery life as a response variable. Repeat (a)-(c) above for this transformation.

5. Now consider using the square of the battery life as a response variable. Repeat (a)-(c) above for this transformation.

6. Which of the four models you have fit has residuals that best satisfy the assumptions of the ANOVA model? Explain your choice.

7. For the model you chose in question 6 above, are there any significant pairwise differences in mean lifetime of different battery types? If so, state which are different, and provide p-values, test statistics, and null hypotheses for the hypothesis tests used.

8. **Hot Dogs** A study was conducted to compare the calories and sodium in hot dogs made with different types of meat.

Read the data into R using the following commands:

```
hotdog=read.table("hotdog.txt",header=TRUE)
hotdog
```

and plot calories as a response variable with the type of meat on the x-axis. Your plot could be either a boxplot or a plot with one dot for each hot dog.

Answer the following question: "Are there differences in the average calories of hot dogs made with different kinds of meat?". To answer this question, write down a statistical model (clearly state the response variable, treatment levels, number of replicates, ...), express the above question as a testable null hypothesis, and report the p-value of the test statistic under the null hypothesis. Conduct an analysis of pairwise differences if it helps you clarify where there are differences in mean calories. Your answer should include all R code used, and the important R output. If you need to transform the response variable, do so, but you do NOT need to provide details of the transformations that you tried, but did not ultimately select. Only report the best transformation, and only show residual plots, ANOVA tables, and any other results for this transformation.