Lecture 3: ANOVA Diagnostics and Remedial Measures

STA 106: Analysis of Variance

Suggested reading: ALSM Chapter 18

Xiner Zhou
Department of Statistics, University of California, Davis

ANOVA Diagnostics and Remedial Measures

- ANOVA Model Assumptions
- Check assumptions
- Remedies and concerns about assumption violations

ANOVA Model Assumptions

$Y_{ij} = \mu_i + \epsilon_{ij}$

- ϵ_{ij} error term

We assume $\epsilon_{ij}'s$ are independently and identically distributed as $N(0,\sigma^2)$, for all i, j

$$\epsilon_{ij} \sim N(0,\sigma^2)$$

ANOVA Model Assumptions

- Independence of error terms
  - Independent random samples are obtained
  - would be violated, e.g. clustered within family or geographic region or time or other factors that give rise to similarity among units
- constancy or homogeneity of error variance
- Normality of error terms

often come together, detect violations using residuals

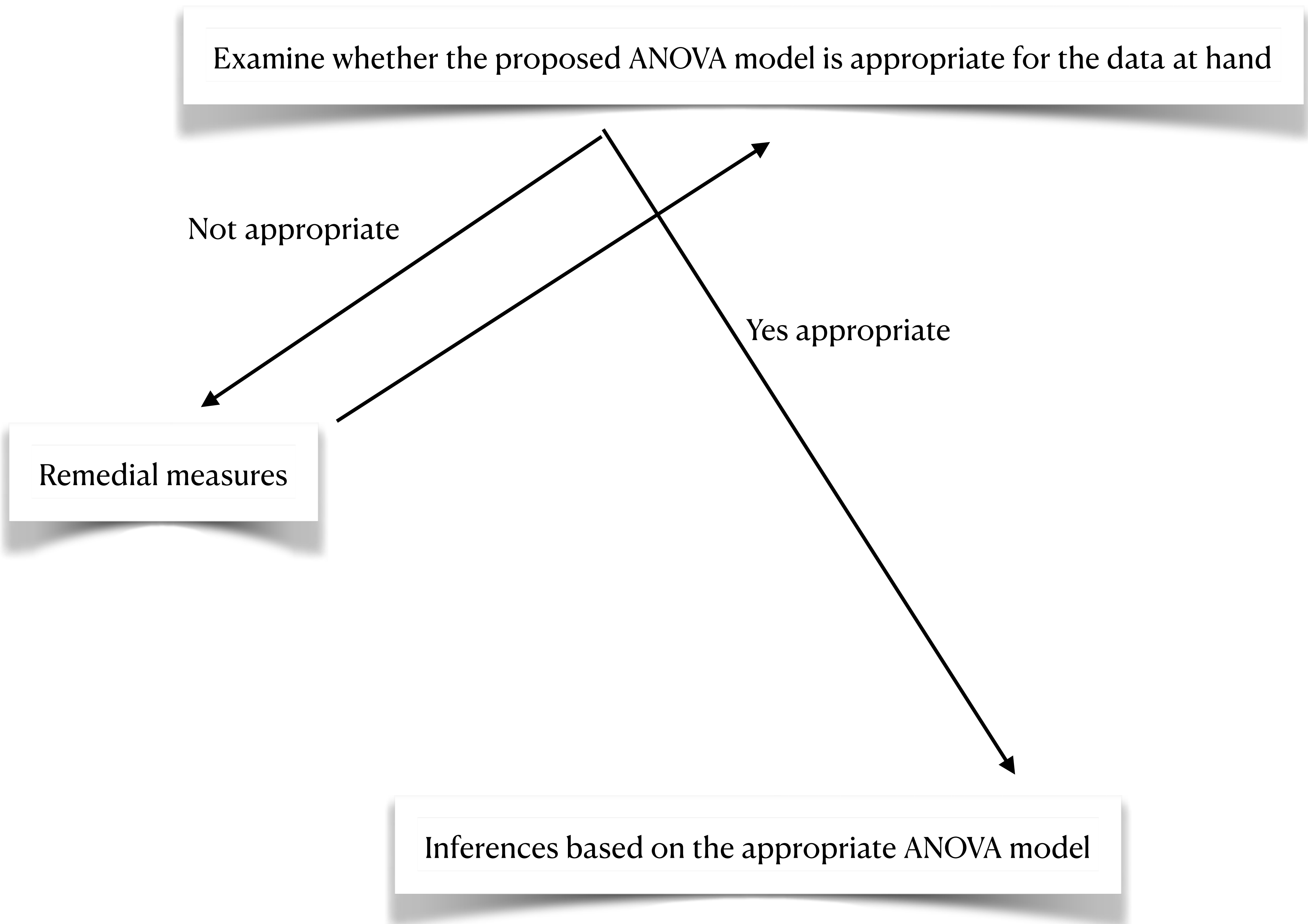
Residual diagnostics

$e_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \bar{Y}_{ij}$

What does “ANOVA model is appropriate” mean?

It is not necessary, nor possible, that an ANOVA model fit the data perfectly.

The major purpose of examination of the appropriateness of the model is to detect serious departure from the assumptions made by ANOVA Model



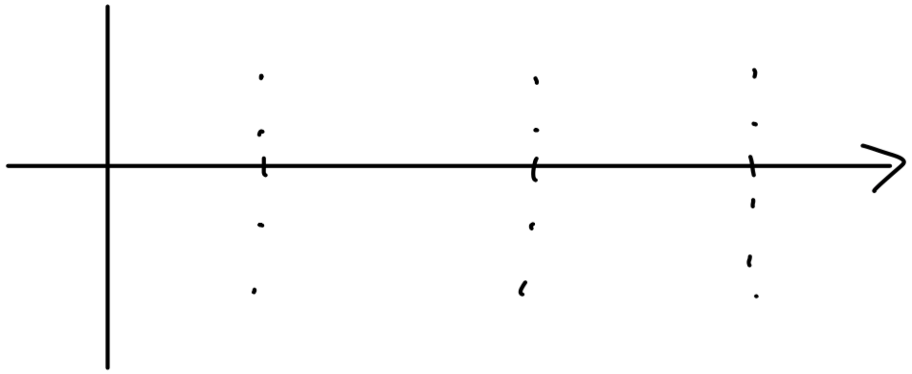
ANOVA Diagnostics and Remedial Measures

- ANOVA Model Assumptions
- Check assumptions
- Remedies and concerns about assumption violations

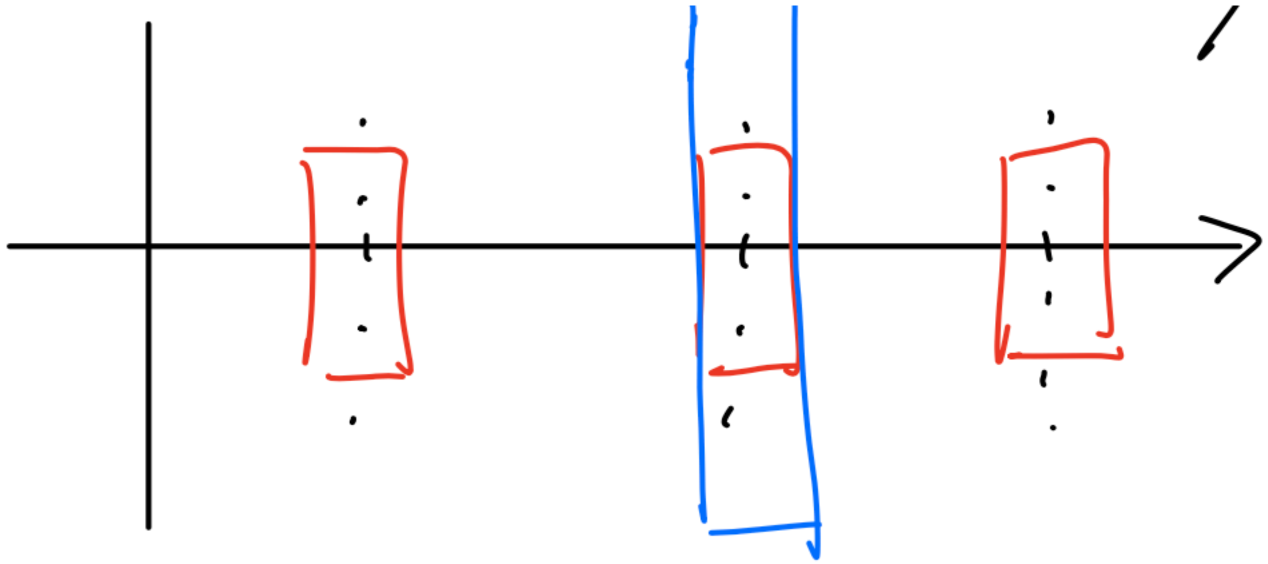
Constancy or Homogeneity of Error Variance

1. Graphical Way

Plot residuals against treatment or factor levels
If Constancy of error variance holds, then we would expect to see the same extent of scatter of residuals around zero, across all factor levels



When sample size is large, box plots of residuals arranged vertically can give a concise summary of the extent of scatters



Constancy or Homogeneity of Error Variance

2. Formal Statistical Tests

Hartley test

To test equal variance using independent samples of equal size selected from r normal populations:

$$H_0: \sigma_1^2 = \cdots = \sigma_r^2 \quad \text{vs} \quad H_a: \text{not all } \sigma_i^2 \text{ are equal}$$

$$H^* = \frac{\max(s_i^2)}{\min(s_i^2)} \overset{H_0}{\sim} H(r, \underset{=n-1}{df})$$

The Hartley test stat is based on the largest sample variance & smallest sample...

H^* near 1 supports H_0

large H^* supports H_a.

where s_i^2 is sample variance of n_i obs Y_{ij} for ith factor level.

$$s_i^2 = \frac{\sum_{j=1}^{n} (Y_{ij} - \bar{Y}_{i.})^2}{n-1} = \frac{\sum_{j=1}^{n} e_{ij}^2}{n-1}$$

Hartley test

- simple to compute
- Only applicable if
  - sample size are equal (or almost equal) across factor levels
  - error terms are normally distributed, quite sensitive to departure from normality

Constancy or Homogeneity of Error Variance

2. Formal Statistical Tests

Brown-Forsythe test (a nonparametric test)

- more difficult to compute
- No sample size requirement
- Robust to departure from normality

To test equal variance using independent samples from r populations:

$H_0: \sigma_1^2 = \dots = \sigma_r^2$ vs H_a: not all σ_i^2 are equal

idea: compare absolute deviations of Y_{ij} obs

$d_{ij} = |Y_{ij} - \tilde{Y}_{i}| = |e_{ij}|$ $\quad \tilde{Y}_{i}$: median

translate original response $Y_{ij} \rightarrow d_{ij}$ which measures the variability

ANOVA pattern.

variance = expected distance square $\sigma_i^2 \approx E(d_{ij}^2)$

r variances σ_i^2's are equal $\Rightarrow$ $E(d_{1j}) = \dots = E(d_{rj})$

same exp. value of absolute deviation

not equal $\Rightarrow$ $E(d_{ij}) \neq E(d_{rj})$

So. testing equal variance is the same as F test for equal factor level means with resp. $\{d_{ij}\}$

$\Rightarrow F_{BF}^* = \dfrac{MSTR}{MSE}$

where $MSTR = \dfrac{\sum n_i (\bar{d}_{i.} - \bar{d}_{..})^2}{r-1}$

$MSE = \dfrac{\sum_i \sum_j (d_{ij} - \bar{d}_{i.})^2}{n_T - r}$

$\bar{d}_{i.} = \dfrac{\sum_j d_{ij}}{n_i}$

$\bar{d}_{..} = \dfrac{\sum \sum d_{ij}}{n_T}$

$F_{BF}^* \overset{H_0}{\sim} F(r-1, n_T - r)$

Large F_{BF}^* value indicates: heterogeneity of error term variances across factor levels.

Example

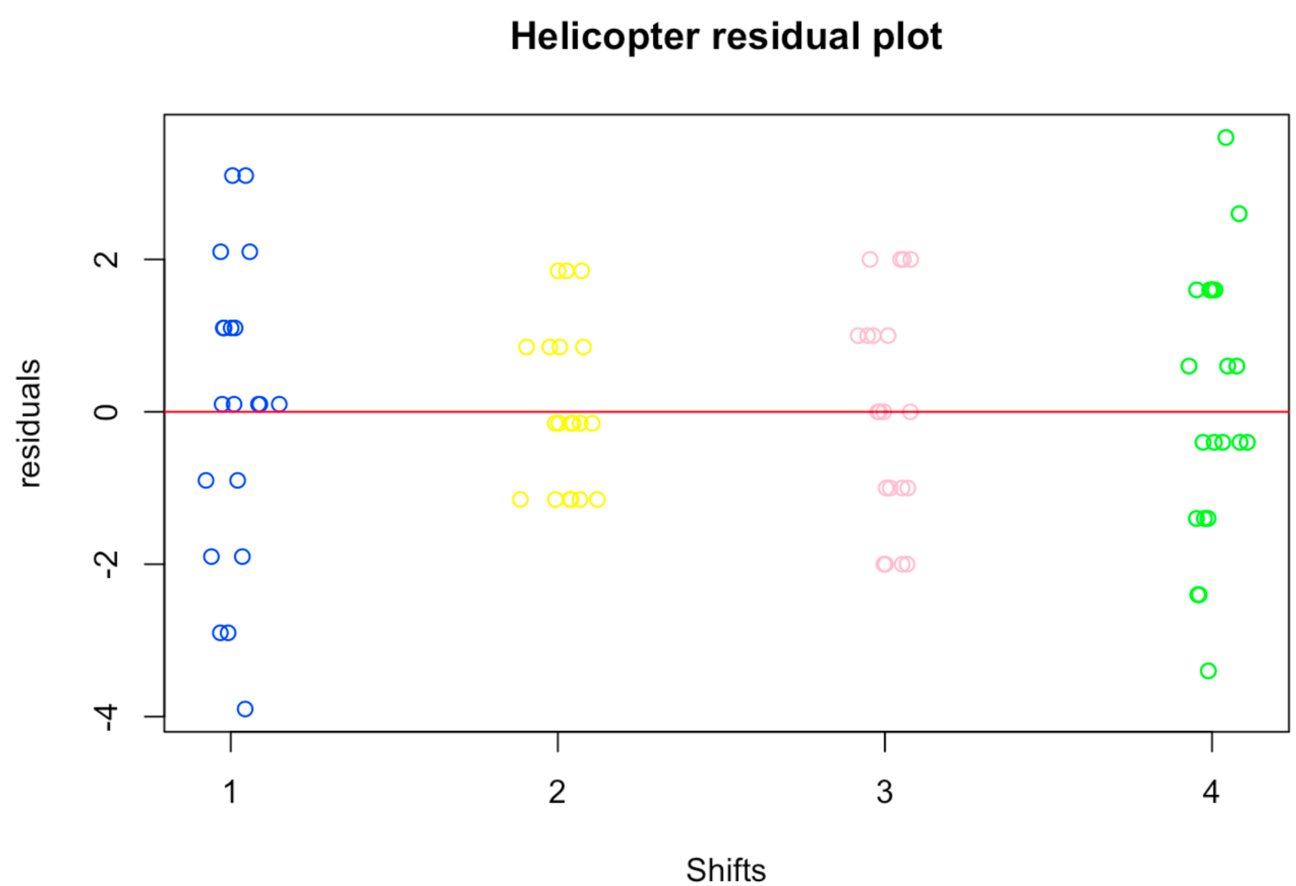
Helicopter service.

An operations analyst in a sheriff"s department studied how frequently their emergency helicopter was used during the past year, by time of day:

shift 1: 2 A.M.-8 A.M.
shift 2: 8 A.M.-2 P.M.
shift 3: 2 P.M.-8 P.M.
shifi 4: 8 P.M.-2 A.M..

Random Samples of size 20 for each shift were obtained. Since the data are counts. the analyst was concerned about the normality and equal variances assumptions of ANOVA model.

To study whether or not the error variances are equal in a graphical way. What are your findings?



The residual plot suggests that the error variance for the shift 2 and 3 tend to be smaller than that for shift 1 and 4 , though due to small sample sizes for each treatment, the differences might just due to randomness.

Example

To study whether or not the error variances are equal using formal tests. State the alternatives. decision rule, and conclusion. What are your findings? Are your results consistent?

Code

```
##
##  Hartley's maximum F-ratio test of homogeneity of variances
##
## data:  y by i
## F Max = 3.2727, df = 19, k = 4, p-value = 0.05991
```

Hide

```
# or implememnt by yourself
library(SuppDists)

# p-value appraoch
si2=rep(NA,4)
for(i in 1:4){
  si2[i]=var(helicopter$y[helicopter$i==i])
}

## test statistics
H=max(si2)/min(si2)
## p value
pmaxFratio(H, df=20-1, k=4, lower.tail=FALSE)
```

```
## [1] 0.05990844
```

Hide

```
# Brown-Forsythe test

# compute absolte deviations from median
Ymedian = rep(0,4)
for(i in 1:4)
{
  Ymedian[i] = median(helicopter$y[helicopter$i==i])
}

d = abs(helicopter$y-Ymedian[helicopter$i])

# ANOVA for d
mu_hat = rep(0,4)
for(i in 1:4)
{
  mu_hat[i] = mean(d[helicopter$i==i])
}

# fitted value for d
d_hat=mu_hat[helicopter$i]

MSTR = sum((d_hat-mean(d))^2)/(r-1)
MSE = sum((d-d_hat)^2)/(n_T-r)
F_BF = MSTR/MSE
## test statistic value
F_BF
```

```
## [1] 1.700657
```

Hide

```
## p value
pf(F_BF, r-1,n_T-r, lower.tail=FALSE)
```

```
## [1] 0.1739958
```

Both the Hartley and Brown-Forsythe test do not reject the null, therefore we can conclude that the data does not provide enough evidence that equal variance assumption is violated.

Normality of Error Variance

Small departures from normality do not create any serious problems
Major departure should be of concern.

Factor level sample size

- **small**
  - combine residuals for all factor levels into one. Check normality for the pooled sample
    - if no major difference in error variance is detected
      - residuals
    - if major difference in error variance is detected
      - studentized residuals

$$r_{ij} = \frac{e_{ij}}{s\{e_{ij}\}}$$

- **large**
  - check normality separately for each factor level

Because non-normal may be detected solely due to we combine error terms having unequal variance

Normality of Error Variance

Graphical Way

Distribution Plots: Histogram

- helpful for detecting serious departure
- sample size must be reasonably large for histogram to convey reliable information about the shape of the distribution

Normality of Error Variance

Graphical Way: Normal Probability Plot or Q-Q Plot

Each residual is plotted against its expected value under normality.

Suppose $e_{ij} \sim N(0, \sigma^2)$

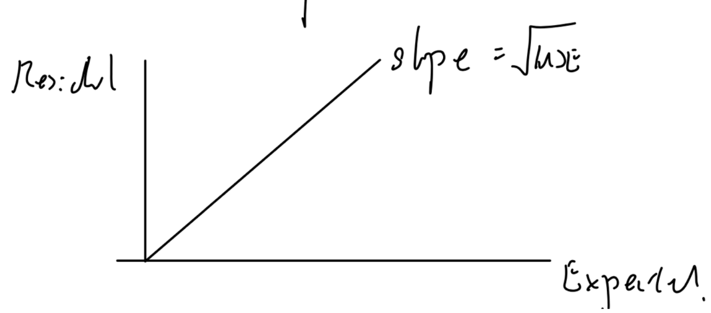
where $\sigma^2 \approx MSE$

then: order residuals $e_{(1)}, e_{(2)}, \dots e_{(k)}$

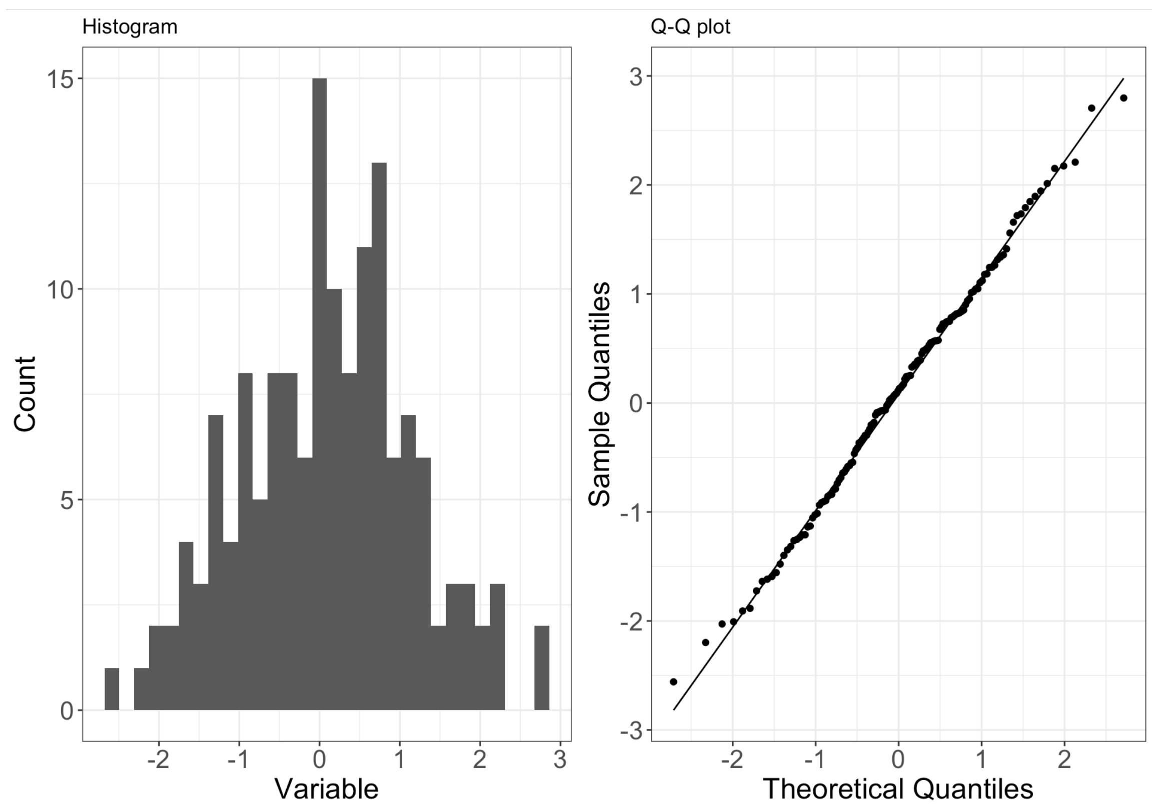
$e_{(k)}$: k-th smallest obs in a random sample of n

$$E[e_{(k)}] = \sqrt{MSE}\ \underbrace{z\left(\frac{k-0.375}{n+0.25}\right)}_{z(p):\ p\text{-th Quantile of } N(0,1)}$$

$e_{(k)} \approx E[e_{(k)}]$ for most of obs.



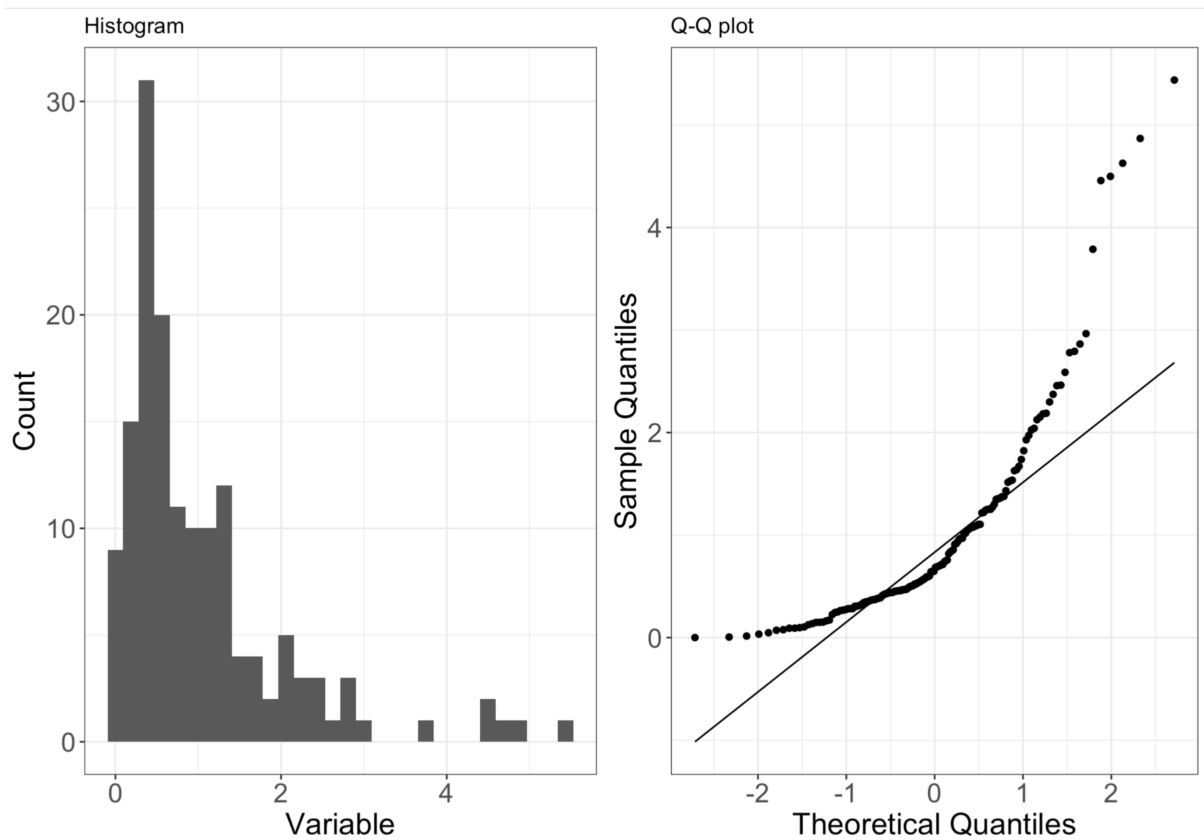
Normality of Error Variance



Normally distributed data

On a Q-Q plot, normally distributed data appears as roughly a straight line, although the ends of the Q-Q plot often start to deviate from the straight line.

Normality of Error Variance

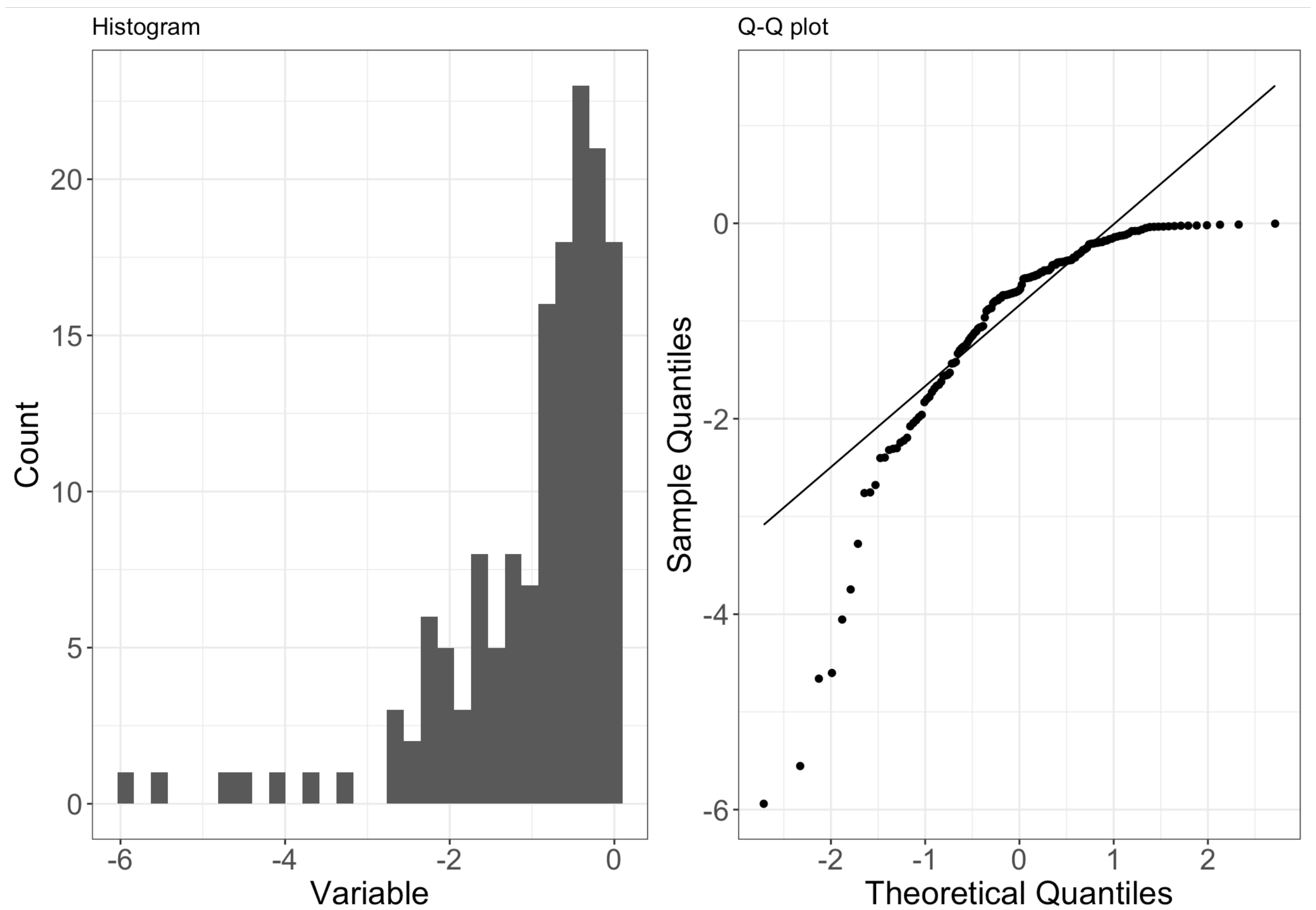


Right-skewed data

There are more data on the right tail of the distribution, than a normal distribution would.

Mostly, residual quantiles are larger than what is expected to be if it's truly normal, so the residuals are all shifted or skewed towards the right

Normality of Error Variance

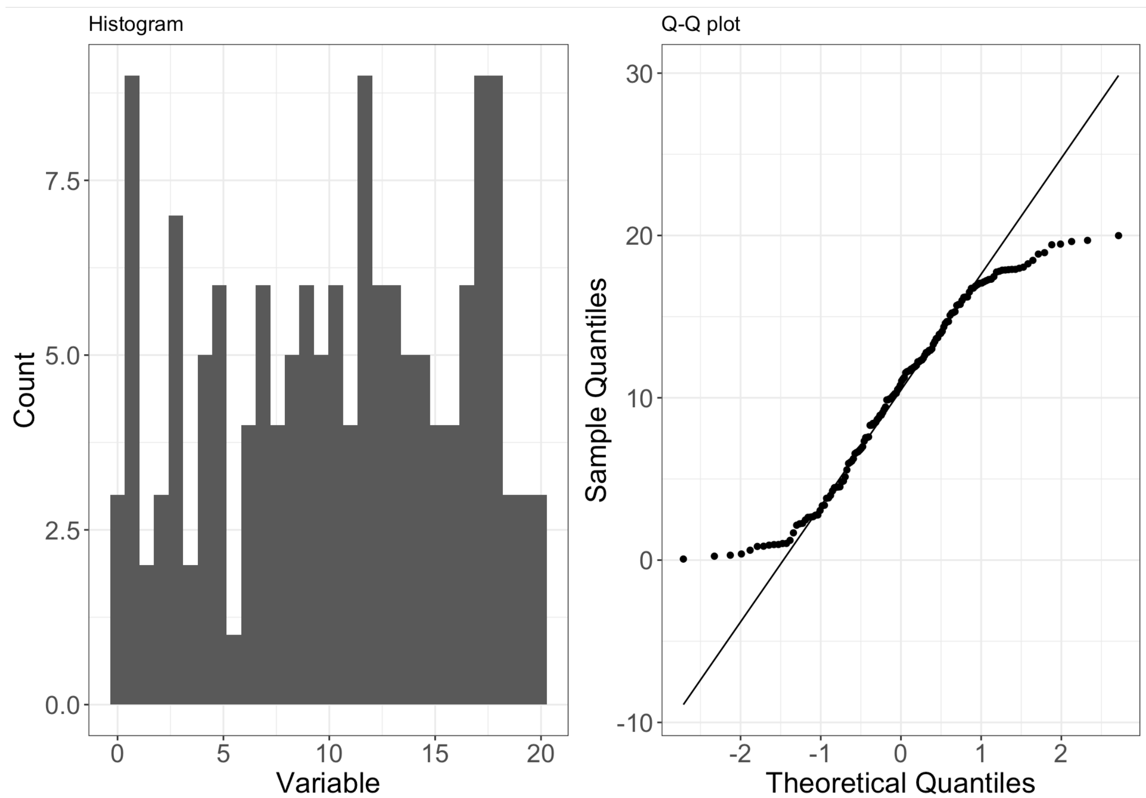


Left-skewed data

There are more data on the left tail of the distribution, than a normal distribution would.

Mostly, residual quantiles are smaller than what is expected to be if it's truly normal, so the residuals are all shifted or skewed towards the left

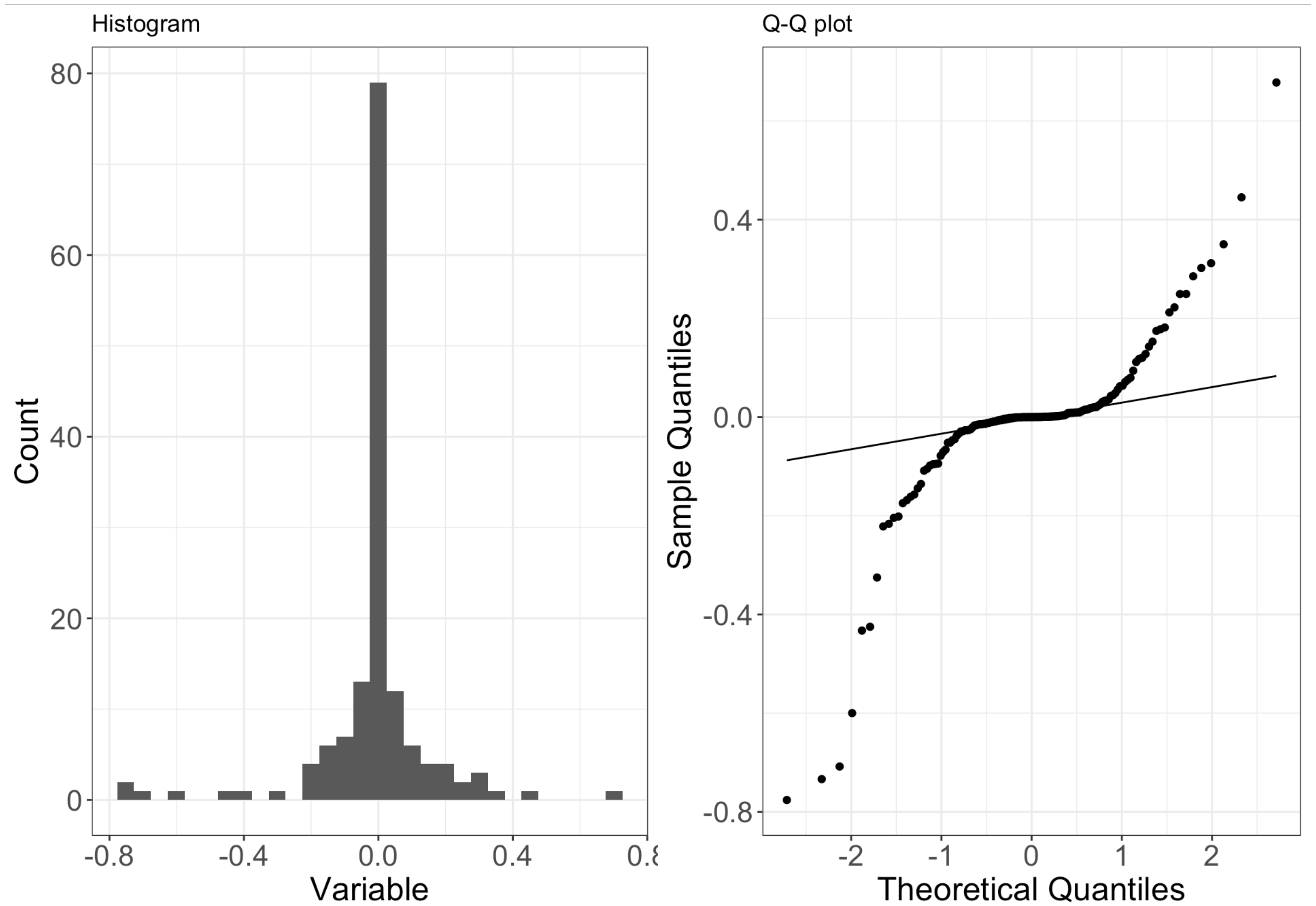
Normality of Error Variance



Under-dispersed data

It exhibits patterns resembles right-skewed data on the left tail, and left-skewed data on the right tail, so the residuals is more concentrated on the center, thus less dispersed than normally distributed data.

Normality of Error Variance



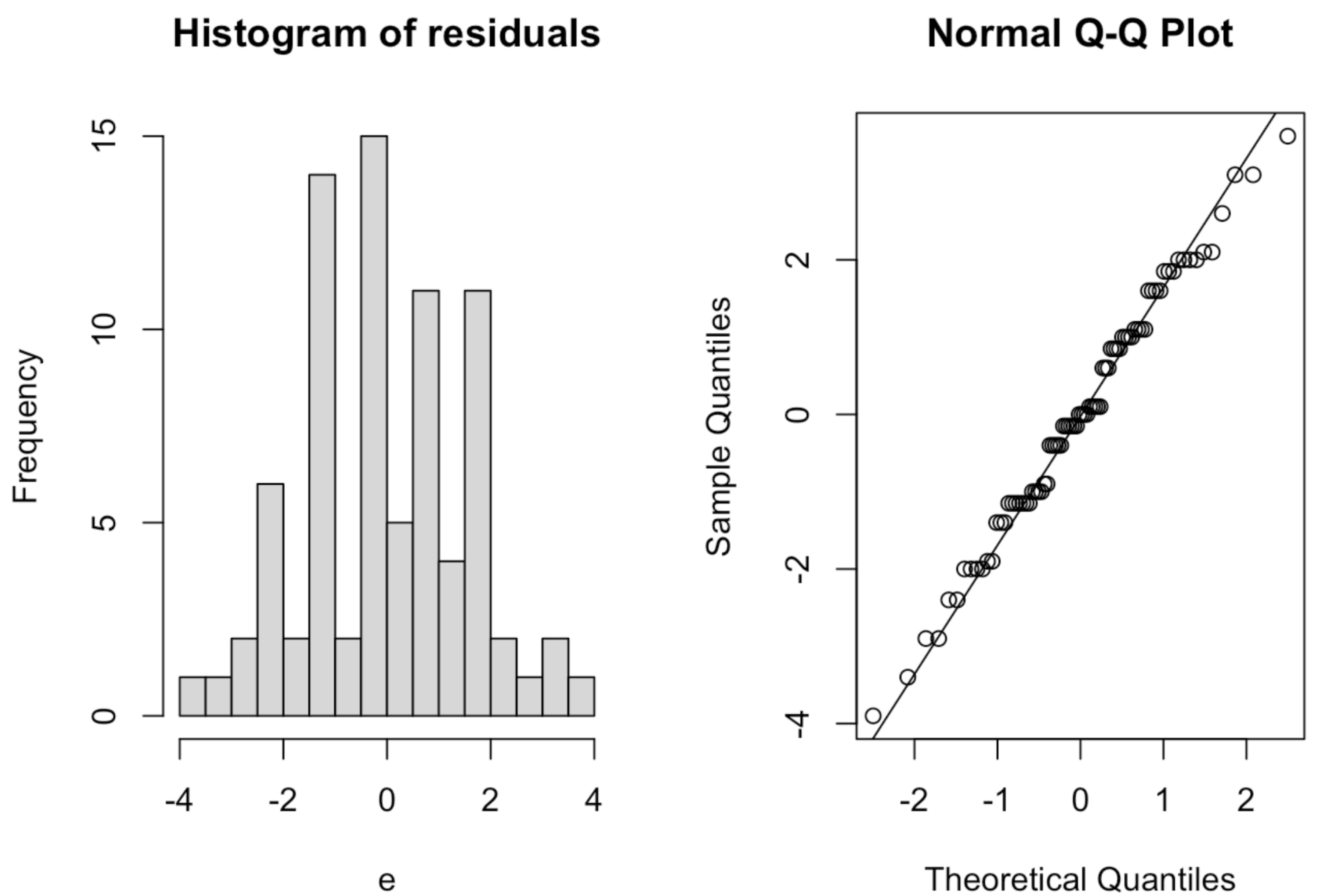
Over-dispersed data

It exhibits patterns resembles left-skewed data on the left tail, and right-skewed data on the right tail, so the residuals is fatter tails than a normal distribution, thus more dispersed than normally distributed data.

Example

To study whether or not the normality assumption is met.

Since each factor level only has 20 observations, it is considered not large, we check normality by combining the residuals.



The histogram shows the data is roughly symmetrically and bell-shaped. On the Q-Q plot, data appears as roughly a straightline. The normality assumption is met.

Example

Summarize your findings about whether the ANOVA assumptions are met.

The independence assumption is automatically satisfied by the set up of the experiemnt. No violation of he equal variance of error term assumption is detected by Hartley and Brown-Forsythe tests.

No violation of normality is detected by checking histogram and Q-Q plot.

Therefore, there is no evidence that the ANOVA assumptions are violated.

However, be aware that those tests only suggest that no violation is detected, they do not confirm these assumptions actually hold for the data at hand definitively.

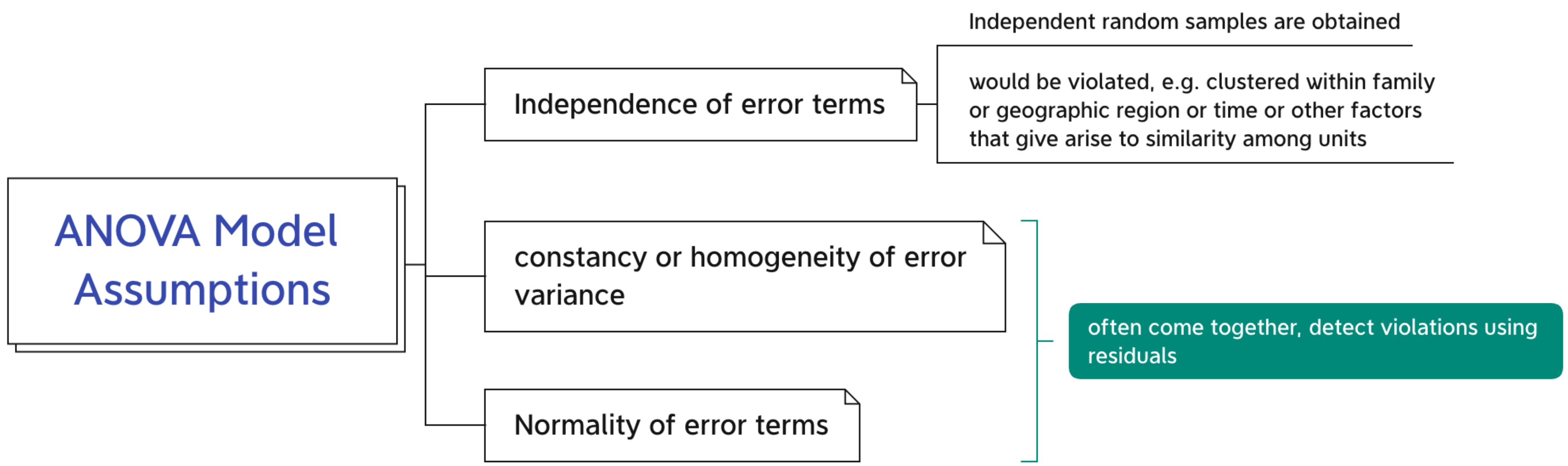
ANOVA Diagnostics and Remedial Measures

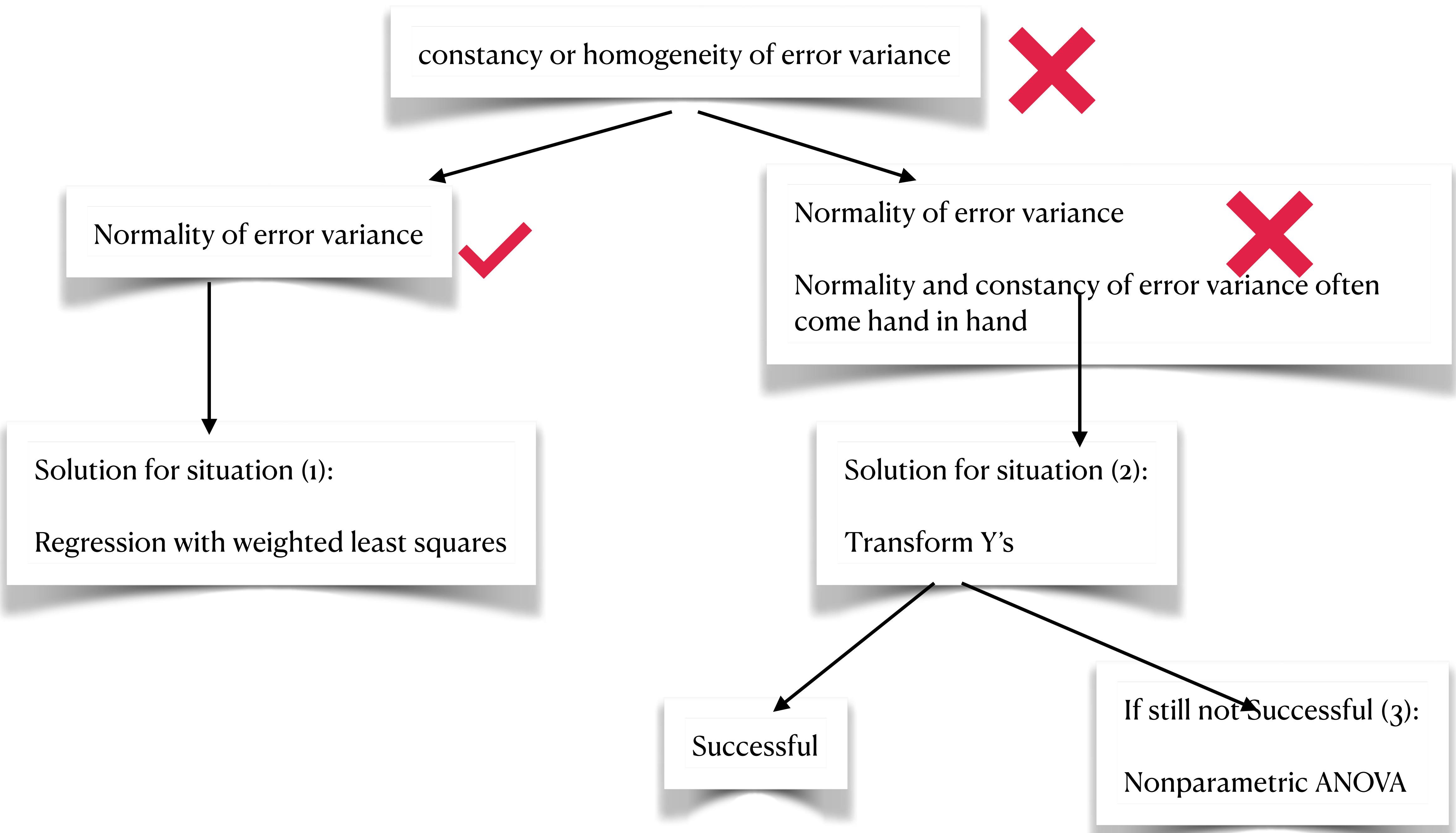
ANOVA Model Assumptions

Check assumptions

Remedies and concerns about assumption violations

Remedies


ANOVA Model Assumptions
Independence of error terms
Independent random samples are obtained
would be violated, e.g. clustered within family or geographic region or time or other factors that give rise to similarity among units
constancy or homogeneity of error variance
often come together, detect violations using residuals
Normality of error terms

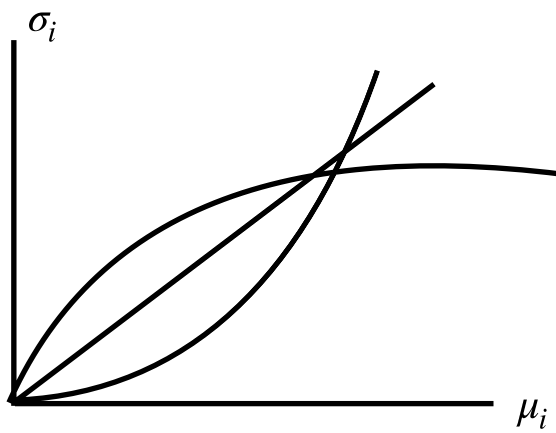

constancy or homogeneity of error variance
Normality of error variance
Normality of error variance
Normality and constancy of error variance often come hand in hand
Solution for situation (1):
Regression with weighted least squares
Solution for situation (2):
Transform Y's
Successful
If still not Successful (3):
Nonparametric ANOVA

Remedies: Transform Y

Solution for (2): Transform Y's

- brings the distribution of error terms more close to normal
- Equalize the error variances across all factor levels

Solutions
- Practical Guide
- Box-Cox Transformation



Variance σ^2 increase when the mean μ increase

$\sigma_i \approx c\sqrt{\mu}_i$ $\quad Y' = \sqrt{Y}$ square root transformation

$\sigma_i \approx c\mu_i$ $\quad Y' = log_e Y$ log- transformation

$\sigma_i \approx c\mu_i^2$ $\quad Y' = \frac{1}{Y}$ reciprocal transformation

Plot these against factor levels:

$$\frac{s_i^2}{\bar{Y}_{i.}} = c$$

$$\frac{s_i}{\bar{Y}_{i.}} = c$$

$$\frac{s_i}{\bar{Y}_{i.}^{\,2}} = c$$

Approximate constancy of one of the three would suggest useful transformation, it will stabilize error variance and make the error distribution more close to normal.

Remedies: Transform Y

Solution for (2): Transform Y's

- brings the distribution of error terms more close to normal
- Equalize the error variances across all factor levels

Solutions
- Practical Guide
- Box-Cox Transformation

Box-Cox Transformation :

Uses power transformation of the data Y

Is a general data transformation technique (even outside ANOVA) used to stabilize variance and make the data more normally distributed

$$Y' = Y^{\lambda}$$

λ : a parameter to be determined by data using MLE for $Y_i^{\lambda} = \beta_1 + \beta_2 X_i + \epsilon_{ij}$

$$\lambda = 2 \quad Y' = Y^2$$

$$\lambda = .5 \quad Y' = \sqrt{Y}$$

$$\lambda = 0 \quad Y' = \log_e Y$$

$$\lambda = -.5 \quad Y' = \frac{1}{\sqrt{Y}}$$

$$\lambda = -1.0 \quad Y' = \frac{1}{Y}$$

Box-Cox is used only to provide a guide for selection a transformation, so overly precise results are not needed.

Choose a meaningful value of λ around the exact result provided by MLE.

Remedies: Transform Y

Solution for (2): Transform Y's

- brings the distribution of error terms more close to normal
- Equalize the error variances across all factor levels

Solutions
- Practical Guide
- Box-Cox Transformation

When such a transformation appropriately solve the non-constancy of error variance and normality, once can work with the transformed data for testing the equality of factor level means.

However, it is desirable present the results for estimating factor level effects in the original variable Y, for easier understanding of the meaning and significance of the results.

Effects of Departures form ANOVA Model Assumptions

What are the effects of any remaining departures from model assumptions?

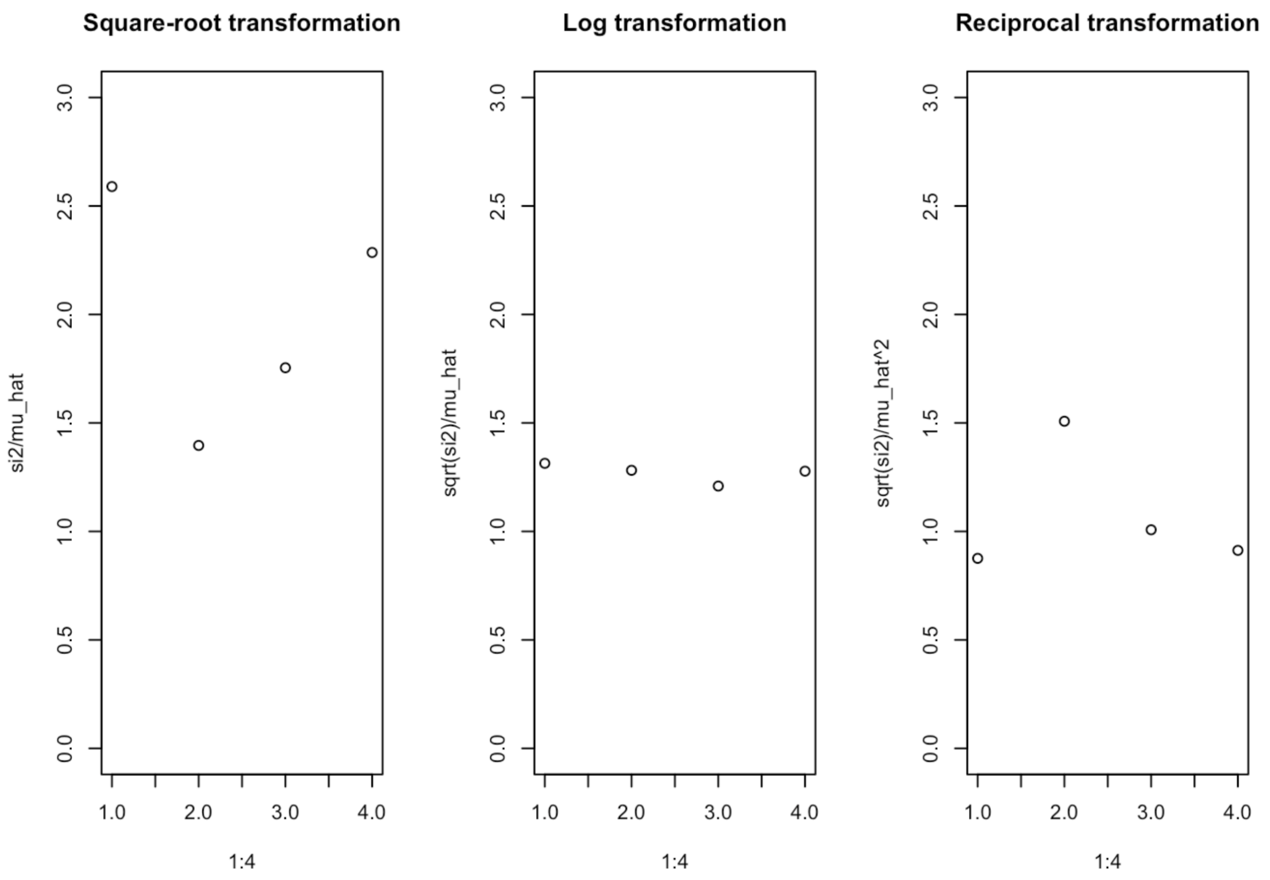
- Nonnormality
  - lack of normality is not a serious problem, as long as it's not extreme
- Unequal Error Variances
  - Inference based on F test (F test for equality of factor level means, Sheffe multiple comparision), is not affected substantially, but other inference procedures do
- Non-independence of Error Terms
  - Serious effects on inference
  - important to prevent in the design stage
  - more advanced modelling technique can rescue

Example

Regardless of your conclusions, practice with transformations.

For each shift. calculate $\bar{Y}_i$, and s_i.

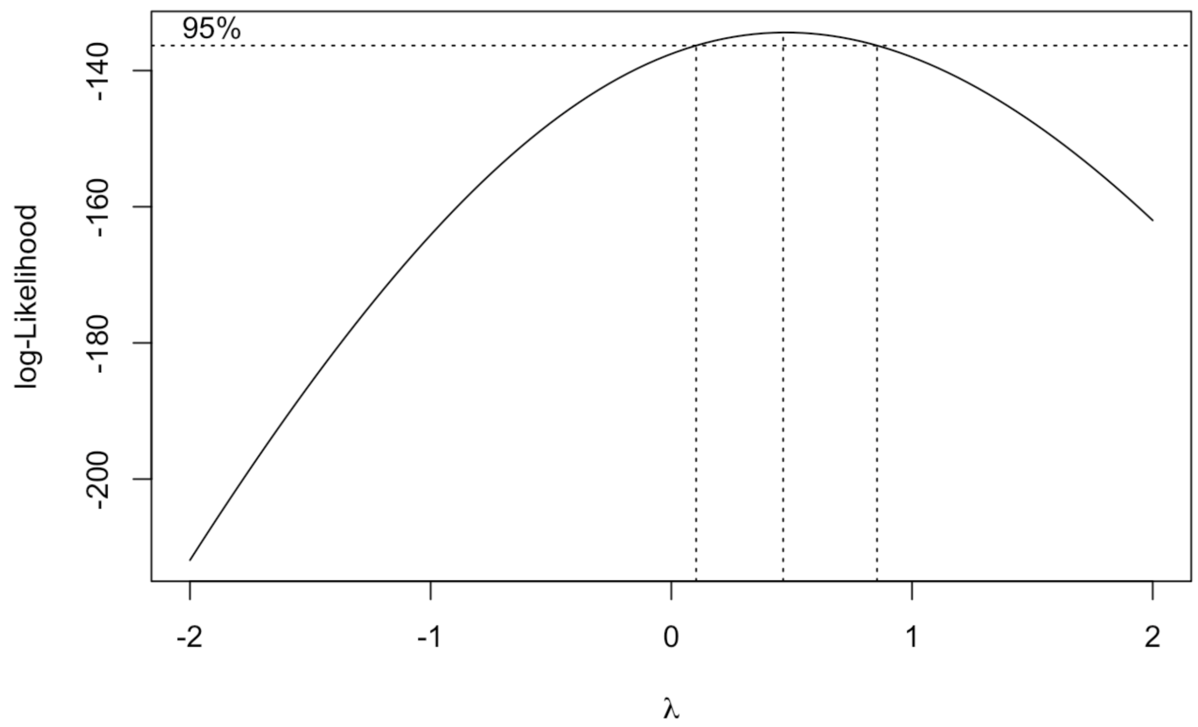
Examine the relation and determine the transformation that is most appropriate here. What do you conclude?



The plot suggests that log transformation of Y is appropriate.

Example

Use the Box-Cox procedure to find an appropriate power transformation of Y.
first adding the constant 1 to each Y observation. Does $\lambda = .5$a square-rocst transformation appear to be reasonable based on the Box-Cox procedure?

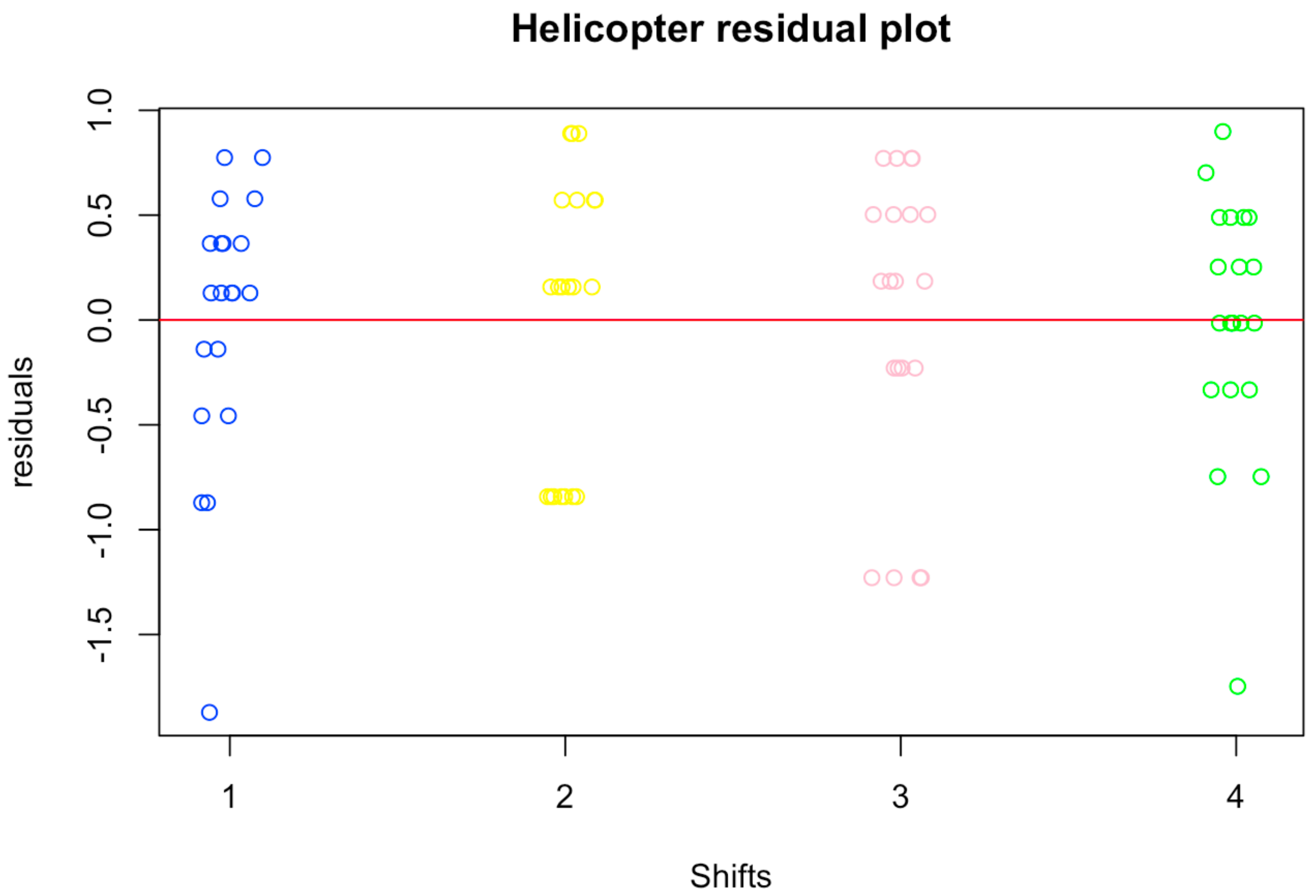


Box-Cox is used only to provide a guide for selection of transformation, so overly precise result is not needed.
Choose a meaningful value around the MLE.

In this case, the likelihood function is maximized around 0.5, it corresponds to a square-root transformation, although log transformation ($\lambda = 0$) is not bad choice either.

Example

The analyst decided to apply the square root transformation $Y' = \sqrt{Y}$ and examine its effectiveness. Obtain the transformed response data. fit ANOVA model. and obtain the residuals.



Example

Test whether the ANOVA model assumptions are met in the transformed data. What are your findings about the effectiveness of the transformation?

```
##
##  Hartley's maximum F-ratio test of homogeneity of variances
##
## data:  y by i
## F Max = 1.4067, df = 19, k = 4, p-value = 0.8832
```

Code

```
## [1] 0.8832147
```

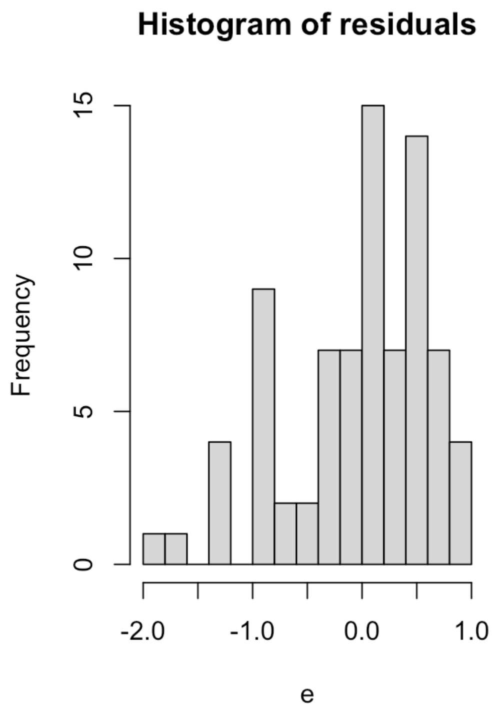
Code

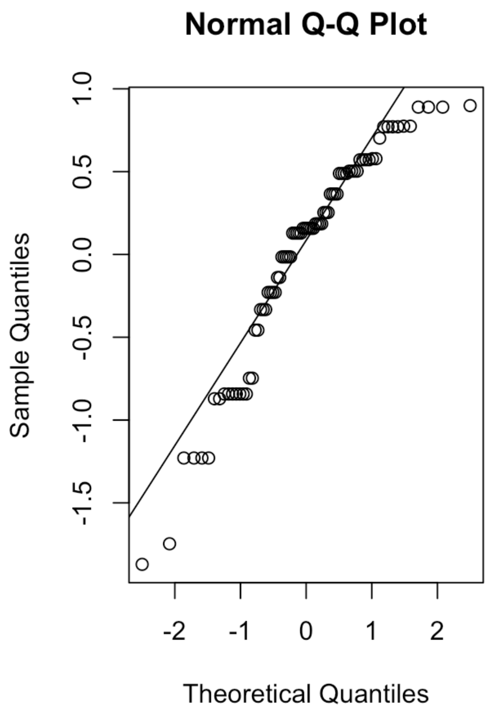
```
## [1] 0.3862421
```

Code

```
## [1] 0.7632153
```

Both the Hartley and Brown-Forsythe test do not reject the null with large p-values, therefore we can conclude that the equal variance assumption is not violated.

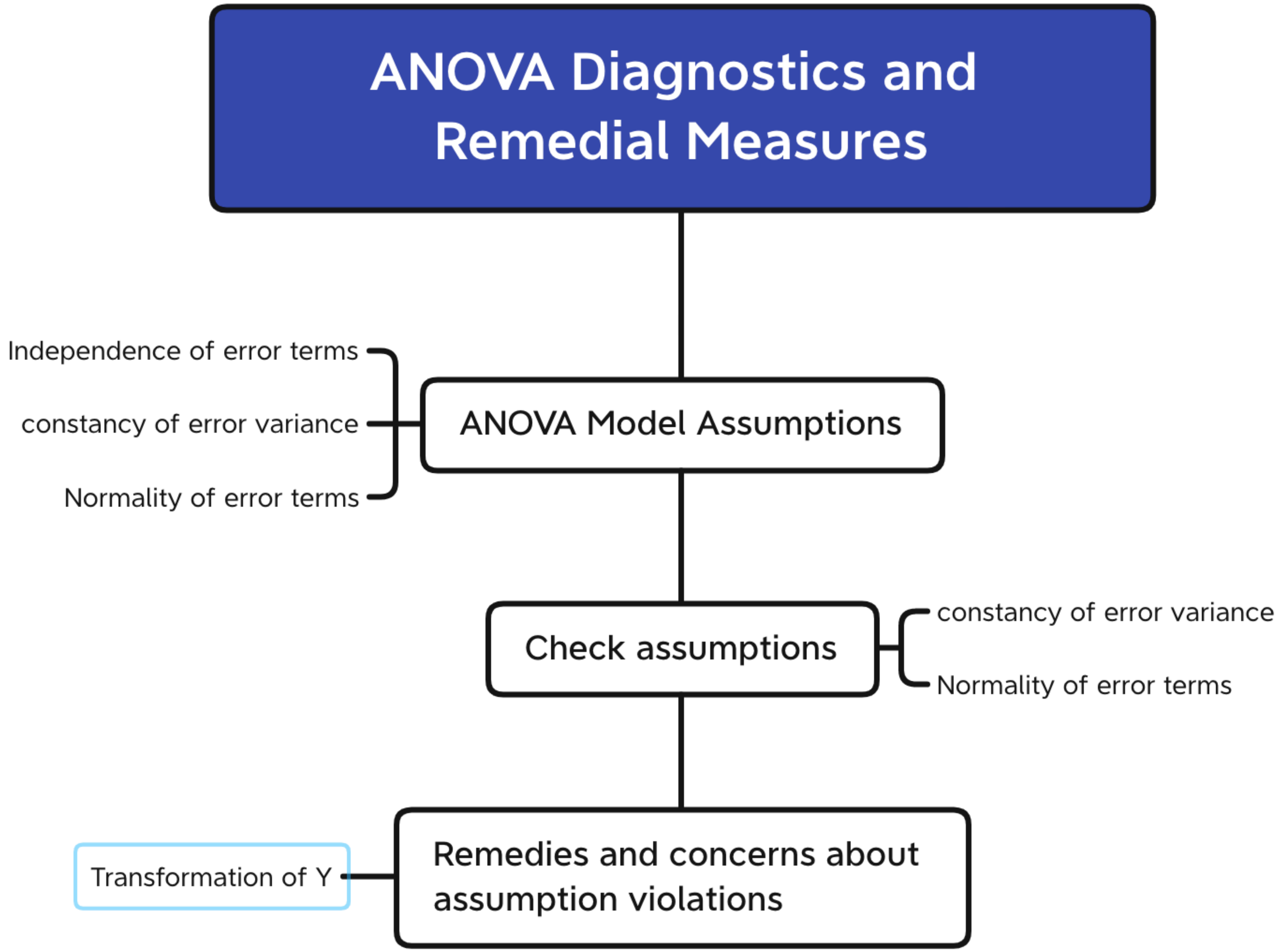




The QQ-plot

suggests that the transformed data might be a little left-skewed. But since normality is not a serious concern for ANOVA model, we can proceed analysis with the square-root transformed data.

Summary



ANOVA Diagnostics and Remedial Measures
Independence of error terms
constancy of error variance
Normality of error terms
ANOVA Model Assumptions
Check assumptions
constancy of error variance
Normality of error terms
Transformation of Y
Remedies and concerns about assumption violations