Checking Model Assumptions

Hao Zhang

🔷 **Assumptions of ANOVA:**

1. **Normality**: Residuals ~ Normal
2. **Equal variance**: Across treatments
3. **Independence**: Errors uncorrelated
4. **Correct model form**: Linear in means

Introduction

Recall the one-way ANOVA model

$$Y_{it} = \mu + \tau_i + \epsilon_{it} \epsilon_{it} \sim N\left(0, \sigma^2\right) t = 1, \ldots, r_i, \quad i = 1, \ldots, v \epsilon_{it}\text{'s mutually independent}$$

It involves a few assumptions such as the error terms having a normal distribution with equal variance. The confidence intervals and the p-values are all based on the assumptions. A serious violation of the assumptions may render the inferences invalid. Good practice is to check these assumptions after ANOVA.

In this lecture, we will discuss strategies to check the assumptions and some remedies when some assumptions are violated. We will primarily focus on residual plots as a tool for checking model assumptions though some formal tests (such as for normality) do exist.

- ***Check the form of the model*** - are the mean responses for the treatments adequately described by $E\left(Y_{it}\right) = \mu + \tau_i, i = 1, \ldots, v?$ This is usually not an issue for one-way model, but needs to be examined for more complex models.
- ***Check for outliers*** - are there any unusual observations (outliers)?
- ***Check for independence*** - do the error variables ϵ_{it} appear to be independent?
- ***Check for constant variance*** - do the error variables ϵ_{it} have similar variances for each treatment?
- ***Check for normality*** - do the error variables ϵ_{it} appear to be a random sample from a normal distribution?

Residuals and Residual Plots

Residuals vs Errors

```
Error=Response–Mean
```

```
Residual=Response – Estimate of Mean
```

Errors are unknown, but the residuals are calculated from the sample data. Residuals are estimates of errors.

For the one-way ANOVA model,

Error: $e_{it} = Y_{it} - \mu_i$.

Residuals: $\hat{e}_{it} = y_{it} - \hat{\mu}_i$ where $\hat{\mu}_i$ is the least squares estimate of μ_i. Hence,

$$\hat{e}_{it} = y_{it} - \bar{y}_{i\cdot}$$

Scaling Residuals

It is desirable to scale the residuals to have a variance to be 1 or close to 1.

$$z_{it} = \frac{\hat{e}_{it}}{\sqrt{\text{ssE}/(n-1)}} \text{ MSE}$$

or

$$z_{it} = \frac{\hat{e}_{it}}{\sqrt{\text{ssE}/(n-v)}}. \text{ MSE}$$

Either way is fine.

Residual Plots

A residual plot is a plot of the standardized residuals z_{it} against another variable (e.g., treatment), the choice of which depends on the

assumption being checked.

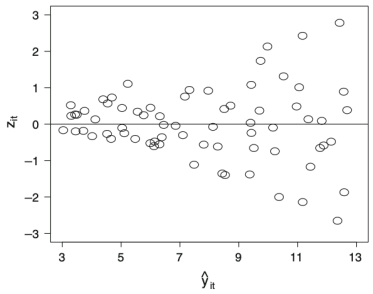
- ***The Fit of the Model:*** plot residuals versus treatment. Lack of fit is indicated if the residuals exhibit a nonrandom pattern about zero.
- ***Outliers:*** Plot residuals versus treatments. Any residual that is beyond -3 or 3 may be considered as an outlier.
- **Equal Variances:** Residuals versus fitted values.

Residuals > ±3
↳ Outlier!

- **Residual vs Treatment** → Checks model fit and outliers
- **Residual vs Fitted Values** → Checks constant variance
- **Normal Q-Q Plot** → Checks normality

Checking Equal Variances

Plot residuals versus fitted values. The most common pattern of non-constant variance is that the error variance increases as the mean response increases.



Variance-Stabilizing Transformation

When the variances depend on the means in the following way

$$\sigma_i^2 = k\mu_i^q$$

where k and q are constant, the following transformation $h\left(y_{it}\right)$ can be chosen to make the variances close to each other:

$$h\left(y_{it}\right) = \begin{cases} \left(y_{it}\right)^{1-(q/2)} & \text{if } q \neq 2 \\ \ln\left(y_{it}\right) & \text{if } q = 2 \text{ and all } y_{it} \text{ 's are nonzero} \\ \ln\left(y_{it}+1\right) & \text{if } q = 2 \text{ and some } y_{it} \text{ 's are zero.} \end{cases}$$

The choice of q can be determined by the slope of the regression

$$\ln\left(s_i^2\right) = \ln(k) + q\left(\ln\left(\bar{y}_{i\cdot}\right)\right) + \text{error.}$$

Read Example 5.6.2 in the text.

Satterthwaite's approximation

When no suitable transformation is available, one can consider Satterthwaite's approximation. This approximation uses a degree of freedom for inferences about contrasts that is calculated from sample data.

Then an approximate $100(1-\alpha)\%$ confidence interval for a single treatment contrast $\Sigma c_i \tau_i$ is

$$\sum c_i \bar{y}_{i\cdot} \pm w\sqrt{\sum \frac{c_i^2}{r_i} s_i^2}$$

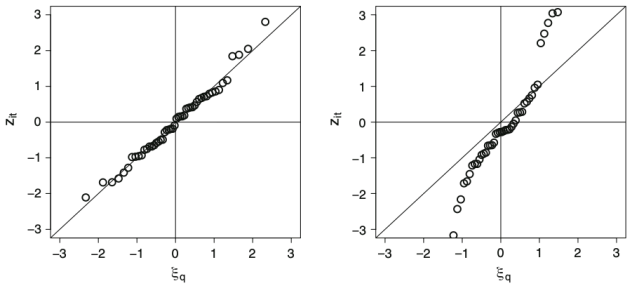
where $w = t_{df,\alpha/2}$ and

$$df = \frac{\left(\sum c_i^2 s_i^2 / r_i\right)^2}{\sum \frac{\left(c_i^2 s_i^2 / r_i\right)^2}{(r_i - 1)}}.$$

This idea of adjusting degrees of freedom can be extended to Tukey's method for all pairwise comparisons.

Checking the Normality Assumption

This assumption is checked using a normal probability plot, which is a plot of the standardized residuals against their normal scores. Normal scores are percentiles of the standard normal distribution. If the points lie on a line approximately, it indicates the data have a normal distribution. See the examples below.



(a) Normal distribution (b) Distribution with heavy tails

SAS Example

For the mung beans experiment, here is how to plot a normal probability plot to check normality.

```
* mung.bean.sas, mung bean experiment, Table 5.8, p121;
;
DATA MUNGBEAN;
  INPUT ORDER WATER MEDIUM LENGTH;
  TRTMT = 2*(WATER-1) + MEDIUM;
  LINES;
   1  2  2  13.2
   2  3  2  11.6
   3  1  2   0.0
   4  1  2   0.6
   5  3  1   5.1
   6  3  2   2.3
   7  1  2   9.5
   8  3  2   6.7
   9  3  2   2.5
  10  3  2  10.6
  11  2  2  14.8
  12  1  2  11.3
  13  2  2  10.7
  14  1  1   1.5
  15  1  1   1.1
  16  2  1   5.2
  17  1  2  12.6
  18  1  1   1.3
  19  3  2  10.8
  20  2  2  13.8
  21  3  1   3.3
  22  3  2  15.9
  23  2  1   0.4
  24  2  2   9.6
  25  3  2   9.0
  26  3  1   0.2
  27  1  2   8.1
  28  3  1   3.9
  29  1  2   7.8
  30  1  1   0.9
  31  2  1   3.6
  32  3  1   7.0
  33  3  1   9.5
  34  2  2   0.0
  35  1  1   8.5
  36  2  1   2.8
  37  1  2   7.3
  38  3  1  11.1
  39  1  1  10.6
  40  2  2   0.6
  41  3  1   6.2
  42  1  1   3.5
  43  1  1   7.4
  44  2  2   8.2
  45  2  1  12.3
  46  2  1  14.1
  47  2  1   0.3
  48  2  1   1.8
;
run;
```

```
PROC GLM data=mungbean;
  CLASS TRTMT;
  MODEL LENGTH = TRTMT;
  OUTPUT OUT=MUNGBN2 PREDICTED=YPRED RESIDUAL=E RESIDUAL=Z;
;
PROC STANDARD STD=1.0; VAR Z;
PROC RANK NORMAL=BLOM; VAR Z; RANKS NSCORE;
PROC PRINT;
;
run;
* Plotting standardized residuals versus run order;
PROC SGPLOT data=mungbn2;
  SCATTER X=ORDER Y=Z;
  XAXIS LABEL = 'Order';
  YAXIS LABEL = 'Standardized Residuals';
  REFLINE 0 / AXIS=Y;
;
run;
* Plotting standardized residuals versus normal scores;
PROC SGPLOT;
  SCATTER X=NSCORE Y=Z;
  XAXIS VALUES = (-4 to 4 by 2) LABEL = 'Normal Scores';
  YAXIS LABEL = 'Standardized Residuals';
  REFLINE 0 / AXIS=Y;
  REFLINE 0 / AXIS=X;
;
run;
```

🔷 What to Look For:

- **Non-random patterns** → model misfit
- **Fan shape** → unequal variances
- **Non-linearity on Q-Q** → non-normality
- **Outliers**: Residuals > ±3

🔷 Fixing Problems

- **Unequal variance**:
  - Use transformation: log, square root, etc.
  - Use **Satterthwaite's approximation** for contrasts
- **Non-normality**:
  - Transform data
  - Use non-parametric tests (if needed)