Lec 3: Model Adequacy Checking

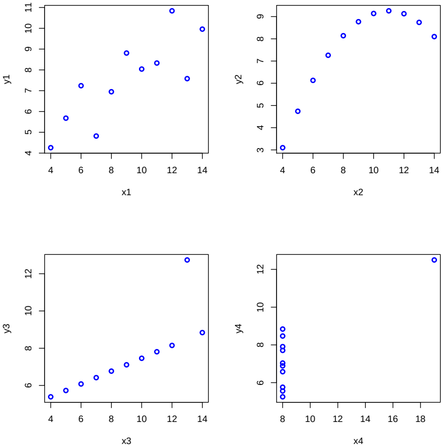
Ying Li

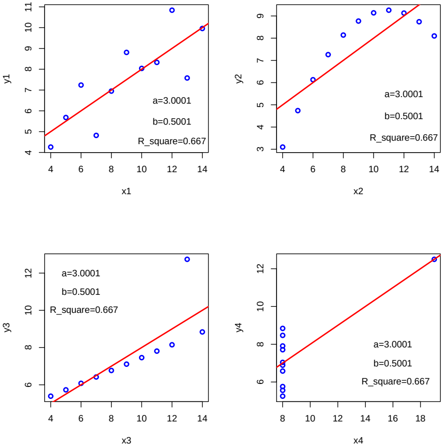
November 16, 2011

Model validation

- Model validation is a very important step in the model building procedure. (one of the most overlooked)
- A high R^2 value does not guarantee that the model fits the data well.
- Use of a model that does not fit the data well can not provide good answers to the underlying scientific questions.

An interesting example: Ascombe dataset





a=3.0001
b=0.5001
R_square=0.667
x1
y1
a=3.0001
b=0.5001
R_square=0.667
x2
y2
a=3.0001
b=0.5001
R_square=0.667
x3
y3
a=3.0001
b=0.5001
R_square=0.667
x4
y4

Main Tool: Graphical residual Analysis

- Different types of plots of residuals (histgram, Plot of residuals in time sequence, plot of residuals versus fitted values) provide information on the adequacy of different aspects of the model.
- Graphical methods have an advantage over numerical methods in model validation
  - graphical methods: a broad range of complex aspects
  - numerical methods: narrowly focused on a particular aspect(a number)

Why use residuals?

If the model fit to the data were correct, the residuals would approximate the random errors.

- If the residuals appear to behave as the assumptions of the error it suggests the model fit the data well.
- Otherwise the model fits the data poorly.(non-normality, dependency, heteroscedasticity).

Two concepts

- **Homogeneity (Homoscedasticity)** : In statistics, a sequence or a vector of random variables is homoscedastic if all random variables in the sequence or vector have the same finite variance. This is also known as homogeneity of variance.
- **Heteroscedasticity**: a collection of random variables is heteroscedastic, or heteroscedastic, if there are sub-populations that have different variabilities than others

The assumptions for ANOVA

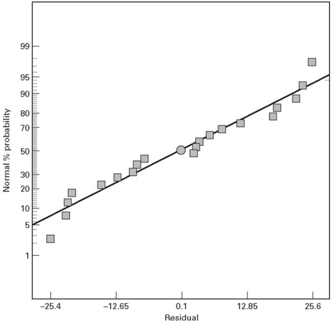
$$y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

Assumptions of the errors:

- ϵ_{ij} are normally distributed
- ϵ_{ij} are independent
- ϵ_{ij} has mean zero and constant variance σ^2.

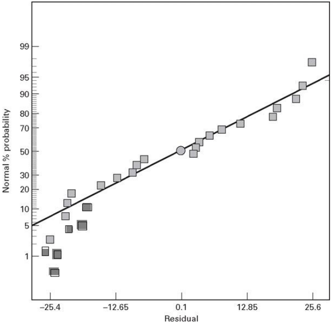
Check the normality

Normal probability plot.



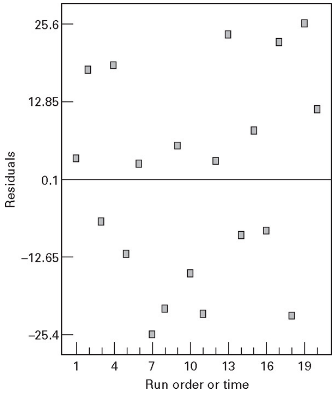
Check the normality

Normal probability plot.

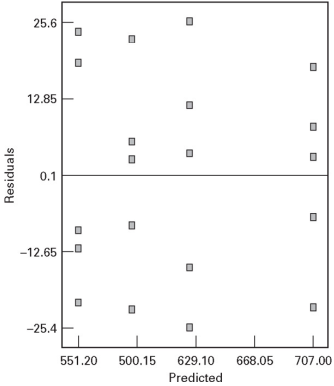


Plot of residuals in Time sequence

Detect the correlation between the residuals.

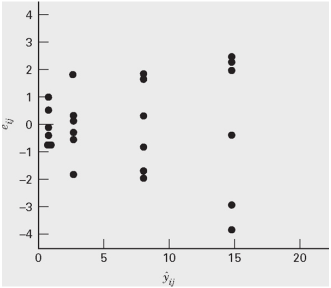


Plot of residuals versus fitted values



Plot of residuals versus fitted values

Detect the nonconstant variance.



Statistical Tests for Equality of Variance

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_a^2$$

$$H_1 : \text{above not true for at least one} \sigma_i^2$$

- Bartlett's test
- Modified Levene test

Bartlett's test

The test statistic is

$$\chi_0^2 = 2.3026\frac{q}{c}$$

where

$$q = (N-a)log_{10}S_p^2 - \sum_{i=1}^{a}(n_i-1)log_{10}S_i^2$$

$$c = 1 + \frac{1}{3(a-1)}(\sum_{i=1}^{a}(n_i-1)^{-1} - (N-a)^{-1})$$

$$S_p^2 = \frac{\sum_{i=1}^{a}(n_i-1)S_i^2}{N-a}$$

and S_i^2 is the sample variance of the ith population.

modified Levene test

The modified Levene test uses the absolute deviation of the observations y_{ij} in each treatment from the treatment **median** $\widetilde{y}_i$. Denote these deviations by

$$d_{ij} = |y_{ij} - \widetilde{y}_i|.$$

The modified Levene test then evaluates whether or not the means of these deviations are equal for all treatment.

modified Levene test

The modified Levene test uses the absolute deviation of the observations y_{ij} in each treatment from the treatment **median** $\widetilde{y}_i$. Denote these deviations by

$$d_{ij} = |y_{ij} - \widetilde{y}_i|.$$

The modified Levene test then evaluates whether or not the means of these deviations are equal for all treatment.
Apply the ANOVA F test.

- Bartlett's test, good accuracy, but very sensitive to normality assumption
- Modified Levene test, robust to departures from normality.

A dilemma

- Assume that we test for homogeneity and whether the residuals are normally distributed or not.
- The more observations, the easier it is to show that the requirement are not fulfilled.
- Conclusion: The validity of the ANOVA is reduced with the number of observation.

Recommendation

- ANOVA is robust against minor heteroscedasticity and minor deviations form the normal distribution.
- Do not use tests, but study the residual plot.

Data transformation

Suppose that the standard deviation of y is proportional to a power of the mean of y such that

$$\sigma_y \propto \mu^{\alpha}$$

We want to transform the data to yield a constant variance. Usually we use

$$y^{\star} = y^{\lambda}$$

Then it can be shown that

$$\sigma_{y^{\star}} \propto \mu^{\lambda+\alpha-1}.$$

If we set $\lambda = 1 - \alpha$, the variance of the transformed data $y^{\star}$ is constant.

How to find α

Estimate α empirically from the data. Consider $\sigma_{y_i} = \theta\mu_i^{\alpha}$. We make the logs

$$\log \sigma_{y_i} = \log \theta + \alpha \log \mu_i.$$

TABLE 3.9

Variance-Stabilizing Transformations

Relationship Between σ_y and μ	α	$\lambda = 1 - \alpha$	Transformation	Comment
$\sigma_y \propto$ constant	0	1	No transformation	
$\sigma_y \propto \mu^{1/2}$	1/2	1/2	Square root	Poisson (count) data
$\sigma_y \propto \mu$	1	0	Log	
$\sigma_y \propto \mu^{3/2}$	3/2	$-1/2$	Reciprocal square root	
$\sigma_y \propto \mu^2$	2	-1	Reciprocal	

Example

A civil engineer is interested in determining whether four different methods of estimating flood flow frequency produce equivalent estimates of peak discharge when applied to the same watershed.

TABLE 3.7
Peak Discharge Data

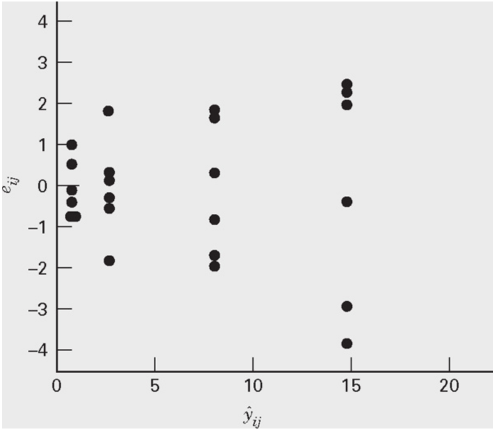
Estimation Method	Observations						$\bar{y}_{i.}$	$\tilde{y}_i$	S_i
1	0.34	0.12	1.23	0.70	1.75	0.12	0.71	0.520	0.66
2	0.91	2.94	2.14	2.36	2.86	4.55	2.63	2.610	1.09
3	6.31	8.37	9.75	6.09	9.82	7.24	7.93	7.805	1.66
4	17.15	11.82	10.95	17.20	14.35	16.82	14.72	15.59	2.77

Example

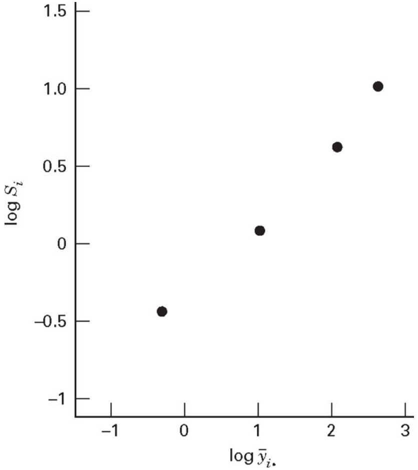
TABLE 3.8
Analysis of Variance for Peak Discharge Data

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-Value
Methods	708.3471	3	236.1157	76.07	<0.001
Error	62.0811	20	3.1041		
Total	770.4282	23			

Example

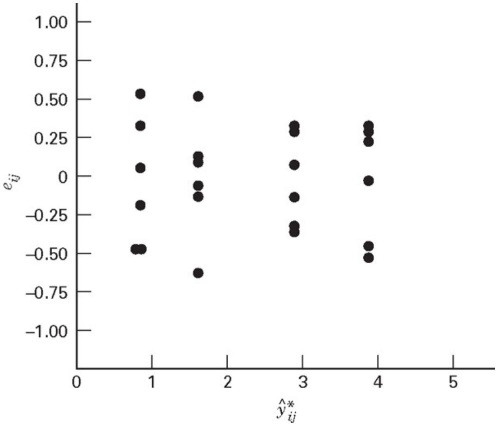


Example



Example

After square-root transformation



Box-Cox transformation family

In the Box-Cox transformation model it is assumed that there is such that a transformation of the observed data y according to:

$$y_\lambda = \left\{ \begin{array}{c} \frac{y^\lambda - 1}{\lambda}, \lambda \neq 0 \\ logy, \lambda = 0 \end{array} \right.$$

The Kruskal-Wallis Test

1. Rank the observation y_{ij} in ascending order.
2. Replace each observation it its rank R_{ij}.
3. The test statistic is

$$H = \frac{1}{S^2}[\sum_{i=1}^{a} \frac{R_{i.}^2}{n_i} - \frac{N(N+1)^2}{4}],$$

where

$$S^2 = \frac{1}{N-1}[\sum_{i=1}^{a} \sum_{j=1}^{n_i} R_{ij}^2 - \frac{N(N+1)^2}{4}].$$

4. If $n_i \geq 5$, and H_0 is ture, H approximately$\sim \chi_{a-1}$. If $H > \chi_{\alpha,a-1}$, then the null hypothesis is rejected.