ANOVA – Chapter 3

Outline

- ANOVA – what is it…
- How to compute ANOVA
  - Examples
- Assumption checking
- Extensions (next class):
  - Unbalanced samples
  - Differing variances
  - Regression / fitted models

ANOVA

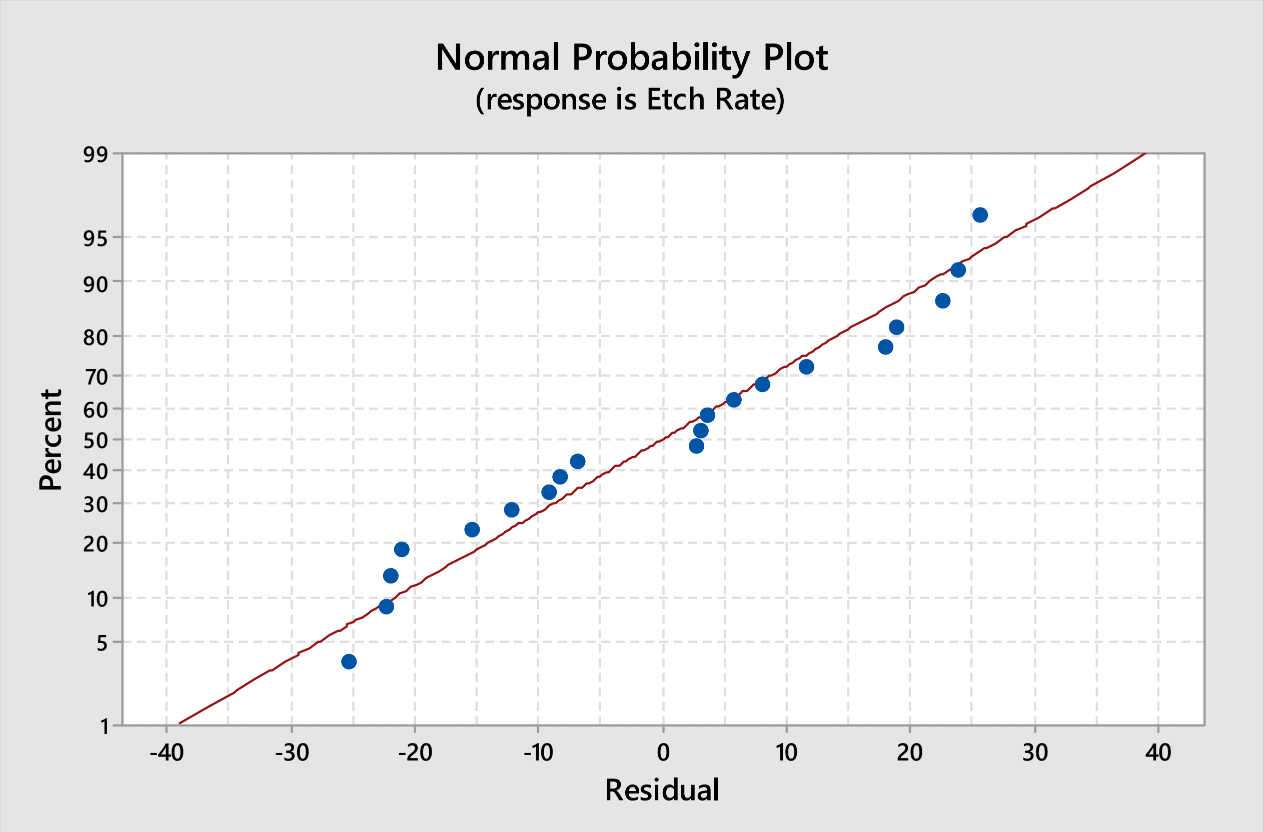
- Assume *a* set of samples
  - Each experiment has *n* replicants
  - Want to see if any of the sets are “significantly” different from the others
- Compute:
  - The overall mean
  - The mean for each set
  - The total sum of squares (SS_T)
  - The sum of squares of just the means (times *n*) ($SS_{Treatments}$)
  - The sum of squares of the errors = SS_T - $SS_{Treatments}$

Examples:

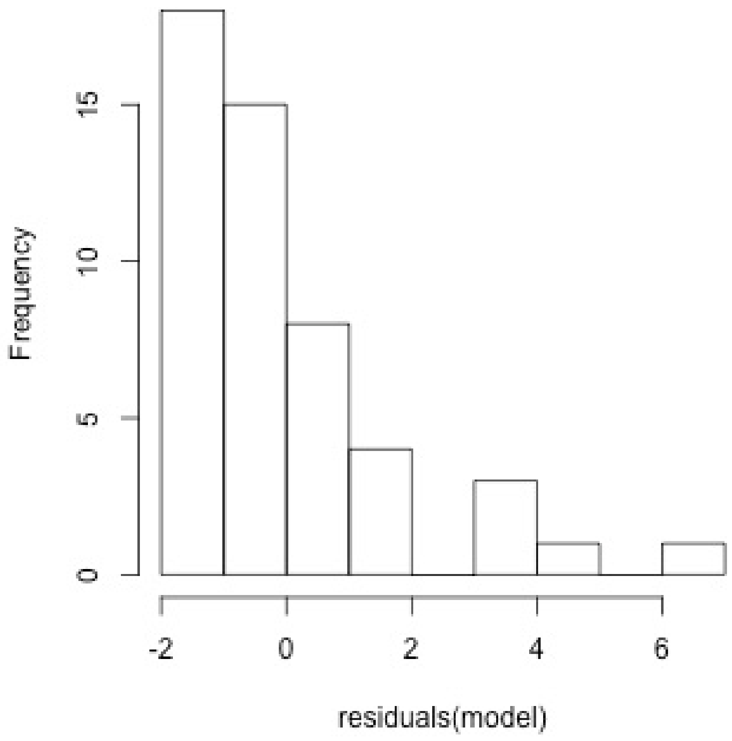
- 3.17
- 3.29

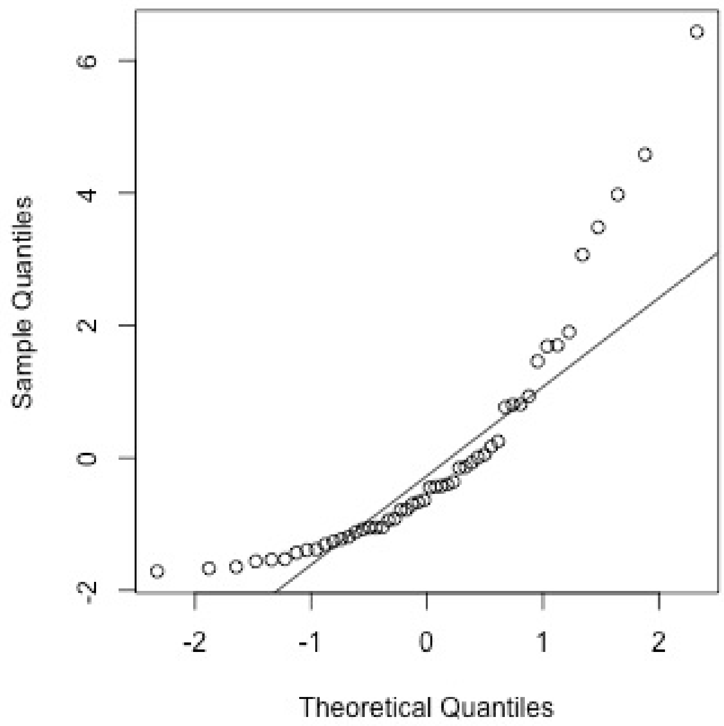
3.4.1. Normal probability plot of residuals

- NPP of residuals combine observations in *a* treatments in one plot
  - Chapter 2 for t-test: NPP of raw data at 2 conditions.
  - In ANOVA, it is usually more effective to work with residuals
- No evidence of a violation of the normality assumption

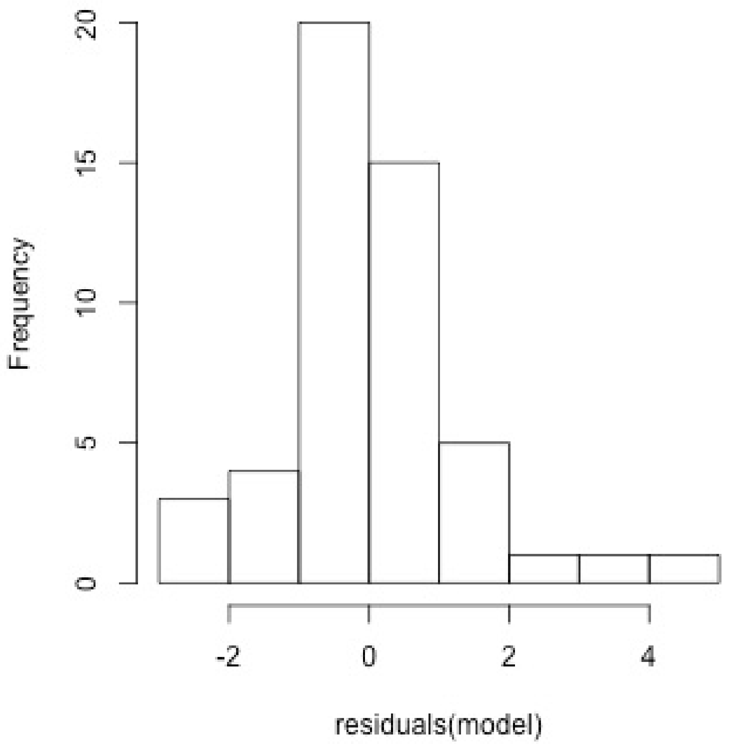


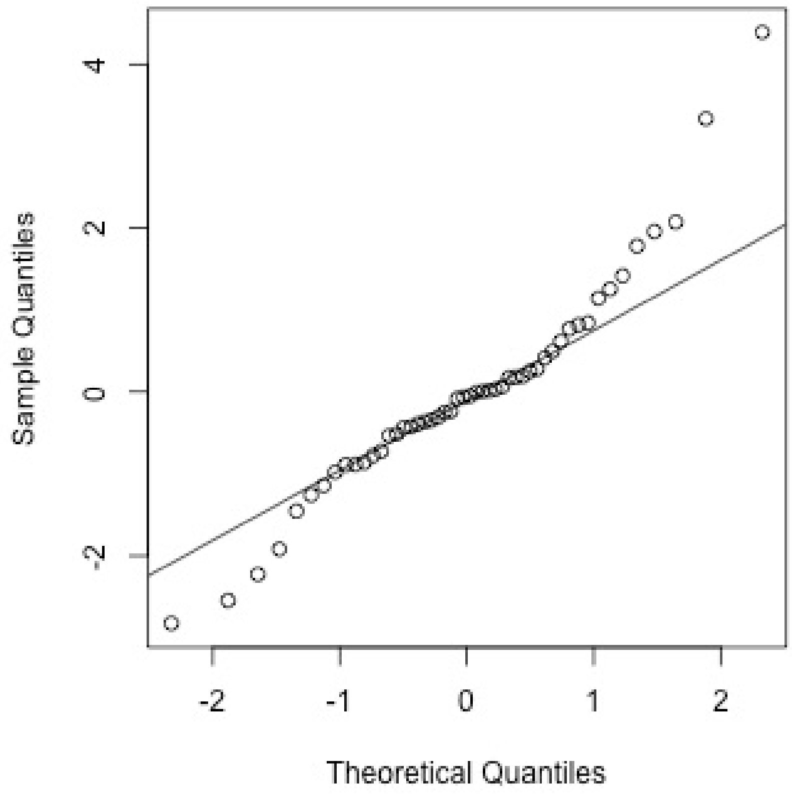
Departures from normality - Skewed residuals





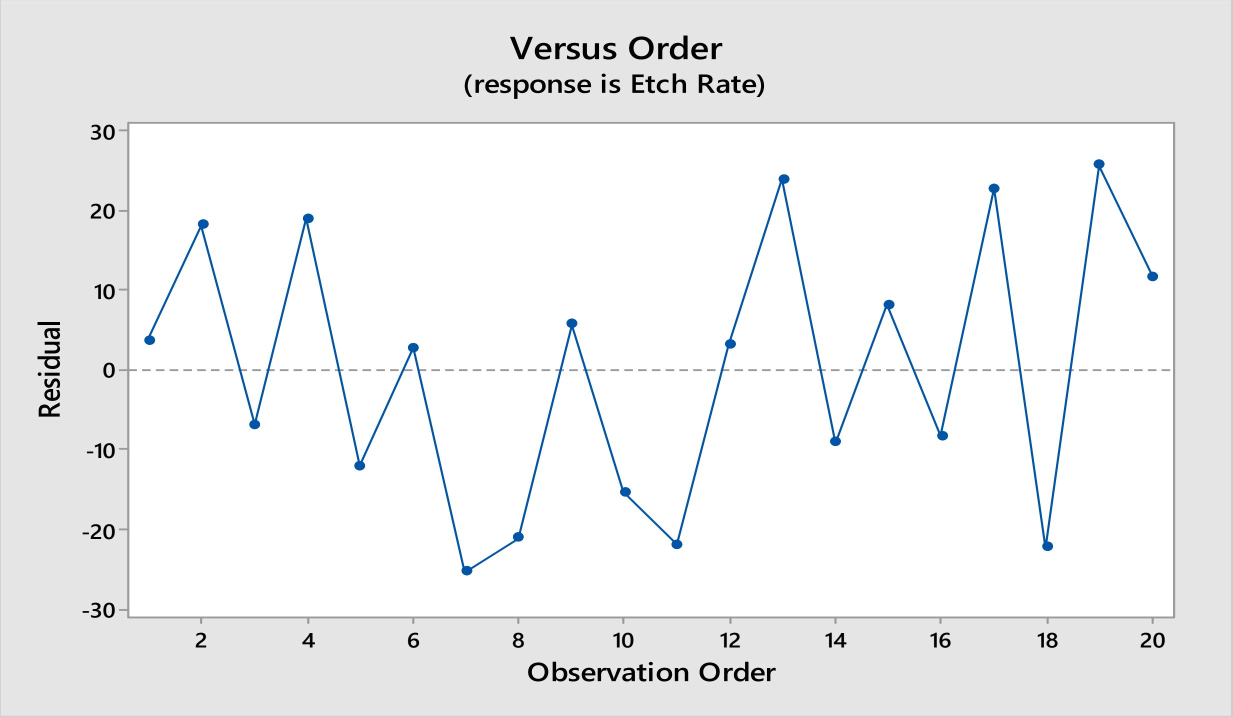
Departures from normality - Heavy tailed residuals





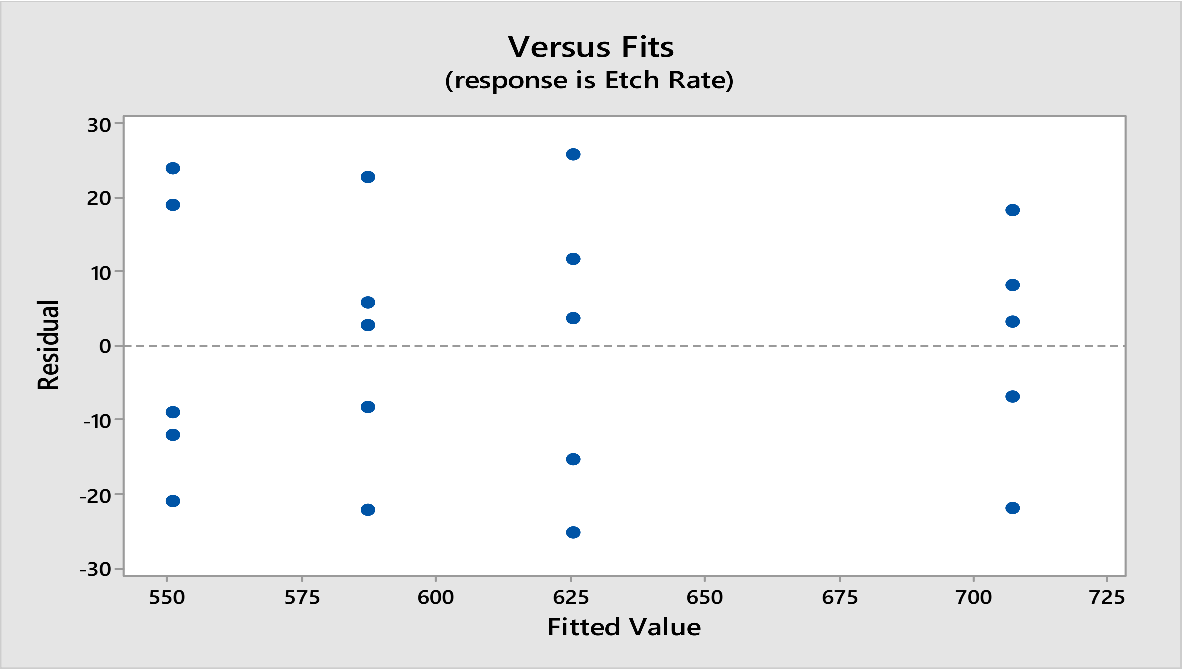
3.4.2. Check for independence

- Residuals versus run order: To satisfy independence assumption, there should not be strong correlation between residuals
  - Tendency to have "runs" of positive or negative residuals indicates strong correlation and indicates that the observations are not independent
  - In this experiment there is no indication of correlation between residuals



3.4.3. Check for constant variance

- Residuals versus fitted values $\hat{y}_{ij}$
  - To satisfy the constant variance assumption at all treatments plot should look like a “parallel band” centered about zero
  - In this experiment there is no indication of non-constant variation at different treatments



Moving beyond the basic assumptions

■ **TABLE 3.3**

The Analysis of Variance Table for the Single-Factor, Fixed Effects Model

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Between treatments	$SS_{\text{Treatments}} = n \sum_{i=1}^{a} (\bar{y}_{i.} - \bar{y}_{..})^2$	$a - 1$	$MS_{\text{Treatments}}$	$F_0 = \frac{MS_{\text{Treatments}}}{MS_E}$
Error (within treatments)	$SS_E = SS_T - SS_{\text{Treatments}}$	$N - a$	MS_E	
Total	$SS_T = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{..})^2$	$N-1$		

Varying Replicate count (Unbalanced Data)

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{..}^2}{N}$$

$$SS_{Treatments} = \sum_{i=1}^{a} \frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{N}$$

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2$$

$$SS_{Treatments} = \sum_{i=1}^{a} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$$

Examples…

- Divide into 2 groups
  - Group A – From the same model
  - Group B – From different models
  - Do 5 batches, each batch having random (4,8) elements
  - Same model

$$\mathcal{N}(10, 1)$$

  - Different model
    - $\mathcal{N}(10, 1)$ for (3 or 4) batches
    - $\mathcal{N}(8, 1)$ for (2 or 1) batches

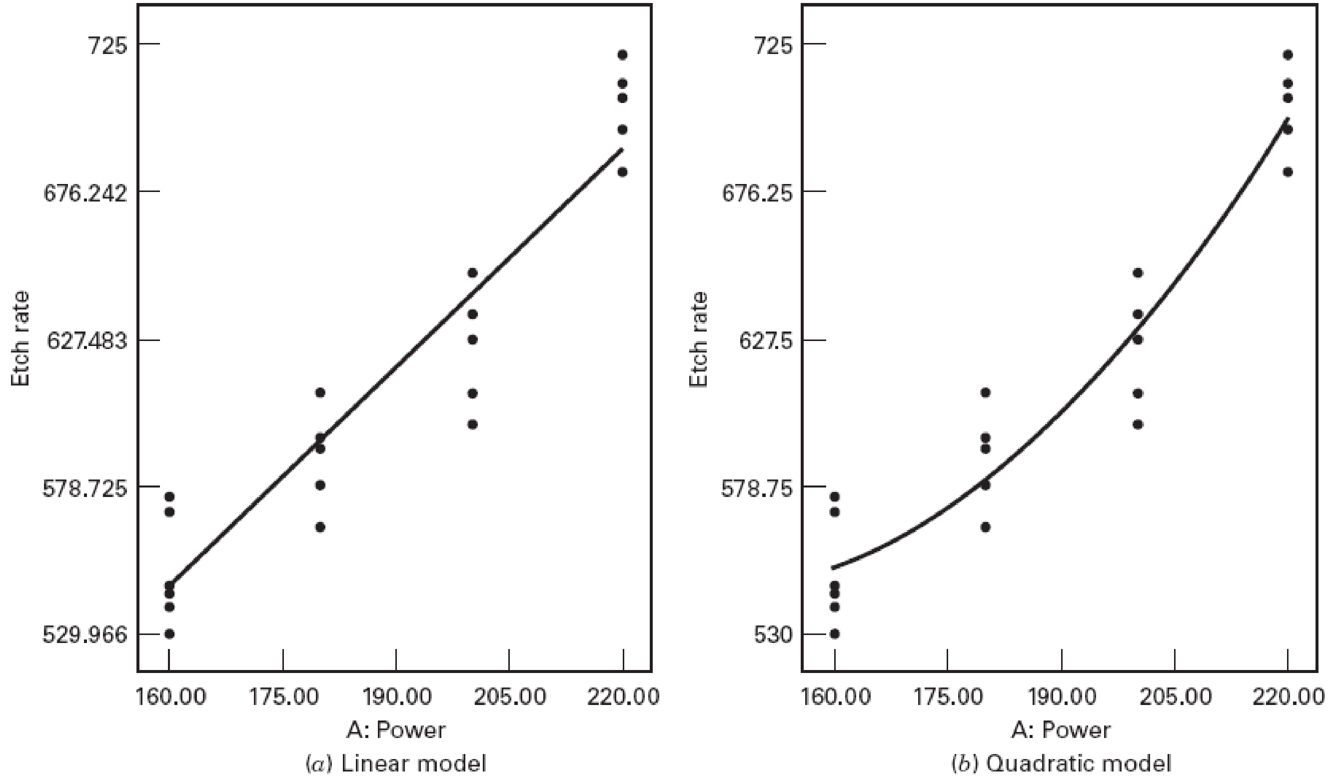
Linear Regression

- $\bar{y}_{i\cdot}$ modeled as linear model $y = \beta_1 + \beta_2 x$
- 2 degrees of freedom for overall model (a-2!)

- Quadratic?

The Regression Model

$$\hat{y} = 137.62 + 2.527x \qquad \hat{y} = 1147.77 - 8.2555\,x + 0.028375\,x^2$$



■ **FIGURE 3.10** Scatter diagrams and regression models for the etch rate data of Example 3.1

Examples…

- Divide into 2 groups
  - Group A – Follows linear regression model
  - Group B – Does not follow linear regression model
  - Do 5 batches, each batch having random (4,8) elements
  - Same model

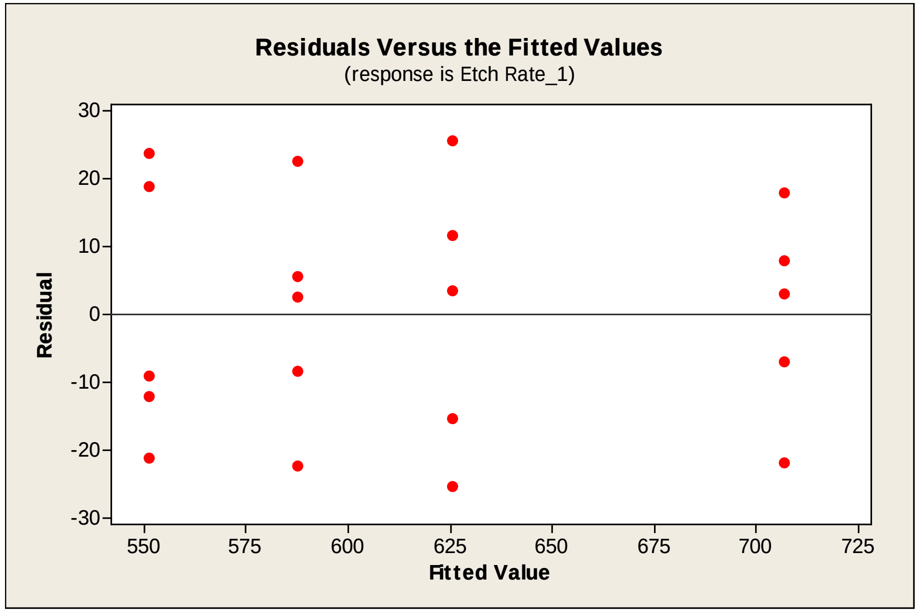
$$\mathcal{N}(10, 1)$$

  - Different model
    - $\mathcal{N}(10, 1)$ for (3 or 4) batches
    - $\mathcal{N}(8, 1)$ for (2 or 1) batches

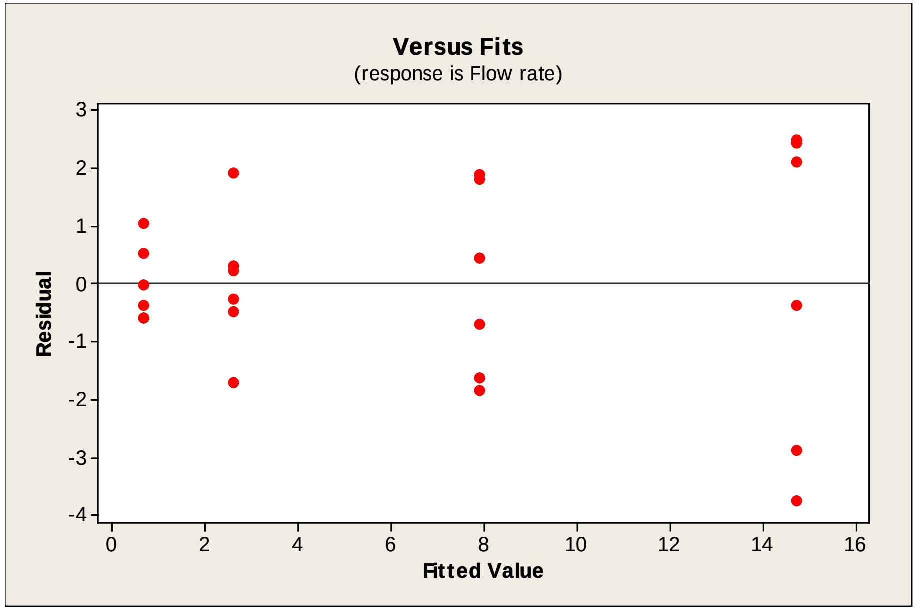
Unbalanced with regression?

Variance stabilizing transformation

- When the error variance is not constant
  - response must be transformed to have constant variance→ variance-stabilizing transformation
  - Use a power transformation of original data as $y^* = y^\lambda$
  - Run the ANOVA on the transformed data y^*



Constant variance



nonconstant variance

Choice of transformation when distribution of y is approximately known

- In general, the standard deviation could be proportional to a power of the mean

$$\sigma_y \propto \mu^\alpha$$

- We hope that transformed response ($y^* = y^\lambda$) standard deviation will be stabilized

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

  - Choose $\lambda = 1 - \alpha$ so that

$$\sigma_{y^*} \propto \text{constant}$$

Choice of transformation when distribution of y is approximately known

- If the theoretical distribution of y is known then transformation is chosen accordingly

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 3.5 (p. 85) – Flood flow estimation

- Determine whether four different methods that measure flood flow frequency produce equivalent results
  - Measure peak discharge(ft^3/sec) with 4 methods
  - Each method is tested 6 times

Method			Peak discharge data			
1	0.34	0.12	1.23	0.7	1.75	0.12
2	0.91	2.94	2.14	2.36	2.86	4.55
3	6.31	8.37	9.75	6.09	9.82	7.24
4	17.15	11.82	10.95	17.2	14.35	16.82