Stat GR4315 Lecture 5

Jingchen Liu

Department of Statistics
Columbia University

Analysis of variance

- The decomposition

$$\sum(y_i - \bar{y})^2 = \sum(y_i - \hat{y}_i)^2 + \sum(\hat{y}_i - \bar{y})^2$$

$$SS_{total} = SS_{error} + SS_{regression}$$

- The analysis of variance table

	d.f.	Sum Sq	Mean Sq	F-value	p-value
x	$p-1$	$\sum(\hat{y}_i - \bar{y})^2$	$\frac{\sum(\hat{y}_i-\bar{y})^2}{p-1}$	$\frac{(n-p)\sum(\hat{y}_i-\bar{y})^2}{(p-1)\sum(y_i-\hat{y}_i)^2}$	$*$
Residuals	$n-p$	$\sum(y_i - \hat{y}_i)^2$	$\frac{\sum(y_i-\hat{y}_i)^2}{n-p}$		
Total	$n-1$	$\sum(y_i - \bar{y})^2$			

Analysis of variance

- The analysis of variance table of the Iris setosa data

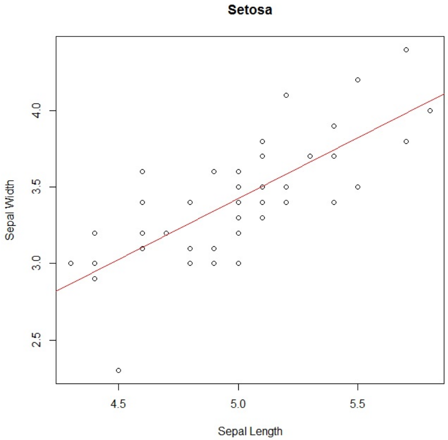
	d.f.	Sum Sq	Mean Sq	F-value	p-value
x	1	3.9	3.2	59.0	7×10^{-10}
Residuals	48	3.2	0.066		
Total	49	7.1			

- F-test and t-test

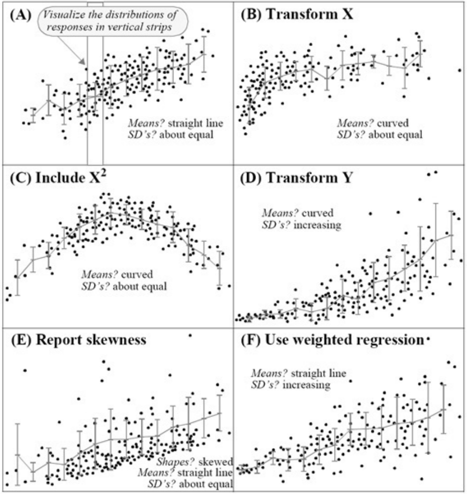
Inferential tools of simple linear model

- The four assumptions
- Point estimate
  - Understanding the least squares estimate
  - variance estimate
  - Frequentist's distribution
- Interval estimate
  - Regression coefficients
  - Prediction: conditional mean, future observation, simultaneous confidence band
- Hypothesis testing
  - Z-test
  - t-test: special case – two sample test
  - F-test: analysis of variance, R^2

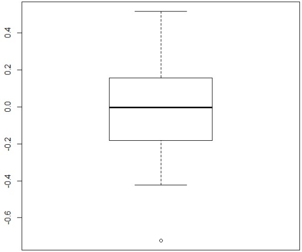
Diagnosis – linearity

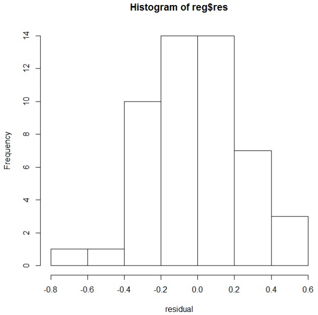


Some possible deviation away from the assumption

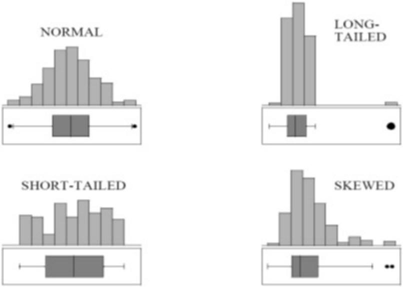


Diagnosis – graphical analysis

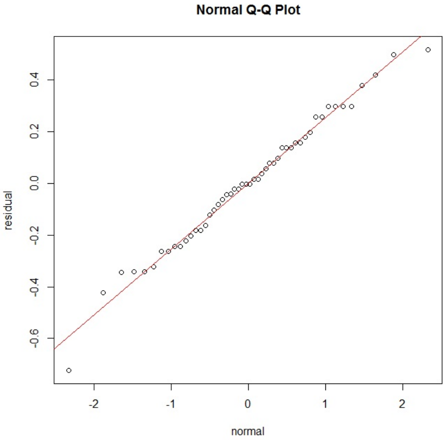




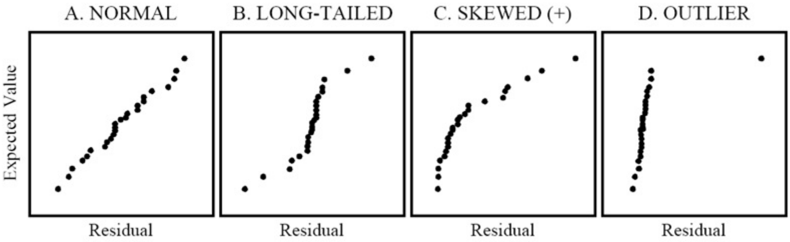
Diagnosis – box plot



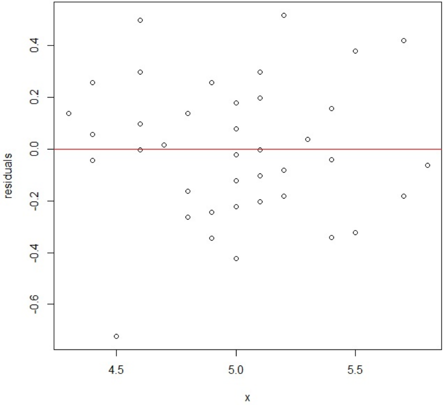
Diagnosis – quantile-quantile plot



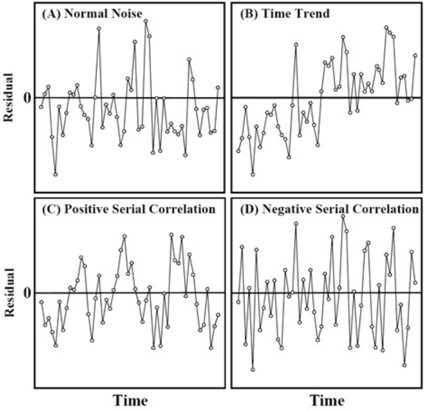
Diagnosis – quantile-quantile plot



Diagnosis – residuals



Some possible deviation away from the assumption



Diagnostic test

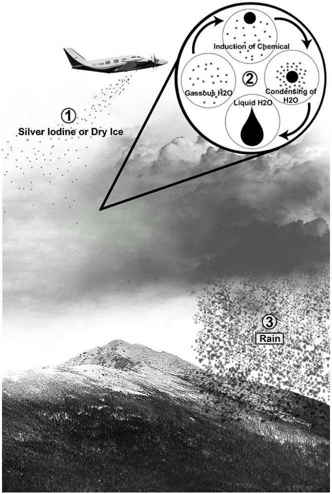
- Pure significant test
- Equal variance test: Brown-Forsythe test and Levene's test

Howard Levene

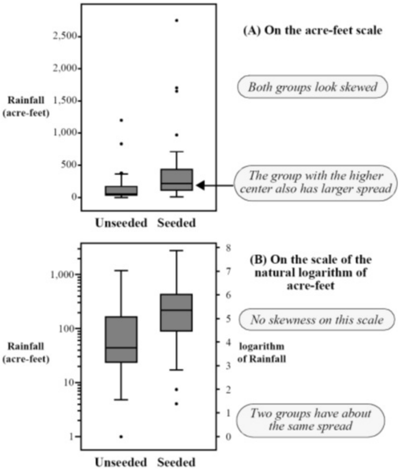
Transformation 101

- Logarithm!

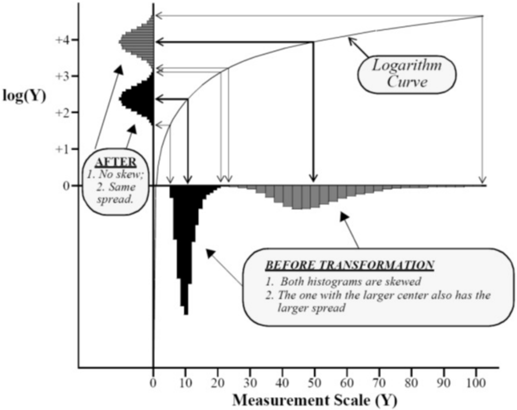
Logarithm


Induction of Chemical
Gassous H2O
Condensing of H2O
Liquid H2O
Silver Iodine or Dry Ice
Rain

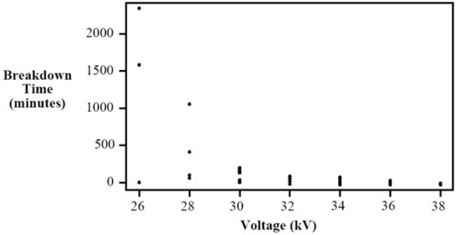
Logarithm


(A) On the acre-feet scale
Both groups look skewed
The group with the higher center also has larger spread
Rainfall (acre-feet)
2,500
2,000
1,500
1,000
500
0
Unseeded
Seeded
(B) On the scale of the natural logarithm of acre-feet
No skewness on this scale
logarithm of Rainfall
Two groups have about the same spread
1,000
100
10
1
8
7
6
5
4
3
2
1
0

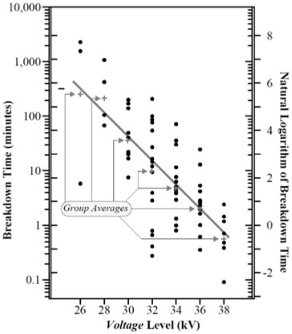
Logarithm


log(Y)
+4
+3
+2
+1
0
Logarithm Curve
AFTER
1. No skew;
2. Same spread.
BEFORE TRANSFORMATION
1. Both histograms are skewed
2. The one with the larger center also has the larger spread
0
10
20
30
40
50
60
70
80
90
100
Measurement Scale (Y)

Logarithm


2000
1500
1000
500
0
Breakdown
Time
(minutes)
26
28
30
32
34
36
38
Voltage (kV)

Logarithm


10,000
1,000
100
10
1
0.1
Breakdown Time (minutes)
8
6
4
2
0
-2
Natural Logarithm of Breakdown Time
Group Averages
26
28
30
32
34
36
38
Voltage Level (kV)

Box-Cox tranformation

$$f_\lambda(y)$$

- $f_\lambda(y) = (y^\lambda - 1)/\lambda$ if $\lambda \neq 0$; $f_\lambda(y) = \log y$, if $\lambda = 0$.
- Choice of λ: maximum likelihood estimate
- library{car}: box.cox.powers

Some other transformations

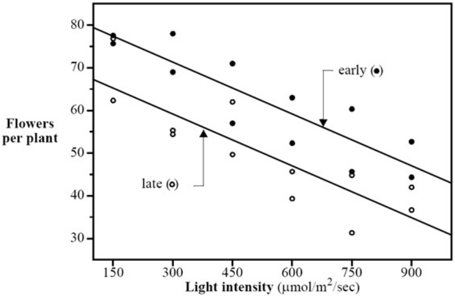
$$y \in [0, 1]$$

- Logit transform: $\log \frac{y}{1-y}$
- Probit transform: $F^{-1}(y)$ where $F(x) = P(Z \leq x)$
- etc.

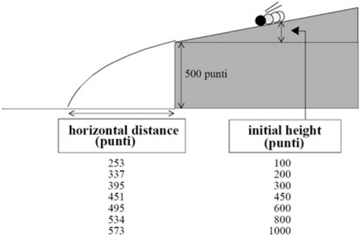
Multiple linear regression – motivation

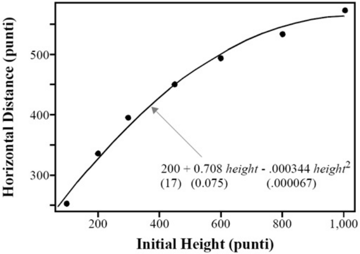
- Simple linear model in subgroups
- Nonlinear relationship
- Multiple predictors
- Variable selection

Multiple group

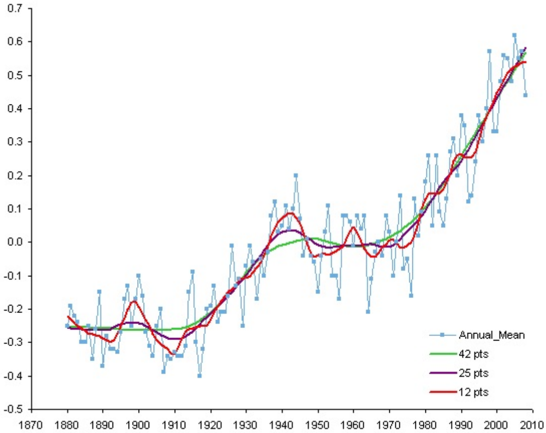


Nonlinearity





Local regression



Multiple linear regression

- Response variable y and p covariates x_1, ..., x_p
- Regression model

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$$

where $\varepsilon \sim N(0, \sigma^2)$

Matrix notation

- $$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

  ε_1, ..., ε_n are i.i.d. $N(0, \sigma^2)$.
- Write in short as
  $$Y = X\beta + \varepsilon$$
  where ε is multivariate normal with mean zero and covariance matrix $\sigma^2 I$.

Least squares estimate

- Least squares estimator

$$\hat{\beta} = \arg\min_{\beta} (Y - X\beta)^\top (Y - X\beta) = (X^\top X)^{-1} X^\top Y$$

- Derive it.

Frequentist distribution

- Unbiased distribution

$$E(\hat{\beta}) = \beta$$

- Variance and covariance

$$Var(\hat{\beta}) = \sigma^2(X^\top X)^{-1}.$$

- Computation of covariance matrix
- Multivariate normal distribution

About multivariate normal distribution

- Multivariate normal distribution with mean μ and covariance matrix Σ

$$f(x) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} e^{-\frac{(x-\mu)^\top \Sigma^{-1} (x-\mu)}{2}}$$

- Generating multivariate normal random variables

Variance estimation

- An unbiased estimator

$$\hat{\sigma}^2 = \frac{(Y - \hat{Y})^\top (Y - \hat{Y})}{n - p - 1}$$

- The distribution of $\hat{\sigma}^2$
- Prediction

$$\hat{Y} = X(X^\top X)^{-1} X^\top Y$$

- Hat matrix

$$H = X(X^\top X)^{-1} X^\top$$

Projection