

Lecture 4. Checking Model Assumptions: Diagnostics and Remedies

Montgomery: 3-4, 15-1.1

Model Assumptions

- Model Assumptions
 - 1 Model is correct
 - 2 Independent observations
 - 3 Errors normally distributed
 - 4 Constant variance

$$y_{ij} = (\bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..})) + (y_{ij} - \bar{y}_{i.})$$

$$y_{ij} = \hat{y}_{ij} + \hat{\epsilon}_{ij}$$

$$\text{observed} = \text{predicted} + \text{residual}$$

- Note that the predicted response at treatment i is $\hat{y}_{ij} = \bar{y}_{i.}$
- Diagnostics use predicted responses and residuals.

Diagnostics

- Normality
 - Histogram of residuals
 - Normal probability plot / QQ plot (refer to Lecture 3)
 - Shapiro-Wilk Test (refer to Lecture 3)
- Constant Variance
 - Plot $\hat{\epsilon}_{ij}$ vs \hat{y}_{ij} (residual plot)
 - Bartlett's or Levene's Test
- Independence
 - Plot $\hat{\epsilon}_{ij}$ vs time/space (refer to Lecture 3)
 - Plot $\hat{\epsilon}_{ij}$ vs variable of interest
- Outliers

Constant Variance

- In some experiments, error variance (σ_i^2) depends on the mean response

$$E(y_{ij}) = \mu_i = \mu + \tau_i.$$

So the constant variance assumption is violated.

- Size of error (residual) depends on mean response (predicted value)
- Residual plot
 - Plot $\hat{\epsilon}_{ij}$ vs \hat{y}_{ij}
 - Is the range constant for different levels of \hat{y}_{ij}
- More formal tests:
 - Bartlett's Test
 - Modified Levene's Test.

Bartlett's Test

- Uses sample variances as estimates of population variances
- $H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_a^2$
- Test statistic: $\chi_0^2 = 2.3026q/c$, where

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

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

$$S_i^2 = \frac{\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2}{n_i - 1} \quad (\text{sample variance at treatment } i)$$

$$S_p^2 = \frac{\sum_{i=1}^a (n_i - 1)S_i^2}{N - a} = \text{MS}_E \quad (\text{pooled variance})$$

- Decision Rule: reject H_0 when $\chi_0^2 > \chi_{\alpha, a-1}^2$.
Remark: sensitive to normality assumption.

Modified Levene's Test

- Use **mean absolute deviations** as estimates of population variances
- For each fixed i , calculate the median (Modified Levene) m_i of $y_{i1}, y_{i2}, \dots, y_{in_i}$.

- Compute the absolute deviation of observation from sample median:

$$d_{ij} = |y_{ij} - m_i|$$

for $i = 1, 2, \dots, a$ and $j = 1, 2, \dots, n_i$,

- Apply ANOVA to the deviations: d_{ij}
- Use the usual ANOVA F -statistic for testing $H_0 : \sigma_1^2 = \dots = \sigma_a^2$.

```
options ls=80 ps=65;

title1 'Diagnostics Example';

data one;
  infile 'c:\saswork\data\tensile.dat';
  input percent strength time;

proc glm data=one;
  class percent;
  model strength=percent;
  means percent / hovtest=bartlett hovtest=levene hovtest=bf;
  output out=diag p=pred r=res;

proc sort; by pred;
symbol1 v=circle i=sm50; title1 'Residual Plot';
proc gplot; plot res*pred/frame; run;

proc univariate data=diag normal noprint;
  var res; qqplot res / normal (L=1 mu=est sigma=est);
  histogram res / normal; run;
```

```
run;
```

```
proc sort; by time;  
symbol1 v=circle i=sm75;  
title1 'Plot of residuals vs time';  
proc gplot; plot res*time / vref=0 vaxis=-6 to 6 by 1;  
run;
```

```
symbol1 v=circle i=sm50;  
title1 'Plot of residuals vs time';  
proc gplot; plot res*time / vref=0 vaxis=-6 to 6 by 1;  
run;
```

Diagnostics Example

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	475.7600000	118.9400000	14.76	<.0001
Error	20	161.2000000	8.0600000		
Corrected Total	24	636.9600000			

Bartlett's Test for Homogeneity of strength Variance

Source	DF	Chi-Square	Pr > ChiSq
percent	4	0.9331	0.9198

Levene's Test for Homogeneity of strength Variance

ANOVA of Squared Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
percent	4	91.6224	22.9056	0.45	0.7704
Error	20	1015.4	50.7720		

Brown and Forsythe's Test for Homogeneity of strength Variance
ANOVA of Absolute Deviations from Group Medians

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
percent	4	4.9600	1.2400	0.32	0.8626
Error	20	78.0000	3.9000		

Non-constant Variance: Impact and Remedy

At different treatments ($i = 1, 2, \dots, a$), variances (σ_i^2) 's are different; in particular, the variance (σ_i^2) 's depend on treatment means (μ_i) 's, i.e.

$$\sigma_i^2 = g(\mu_i).$$

- Does not affect F-test dramatically when experiment is balanced
- Why concern?
 - Further comparison of treatments depends on MS_E
 - Lead to comparison results and confidence intervals.
- Variance-Stabilizing Transformations
 - Transform data y_{ij} to $f(y_{ij})$, e.g. y_{ij} to $\sqrt{y_{ij}}$, with the hope that the transformed data $f(y_{ij})$ do not violate the constant variance assumption.
 - f is called a variance-stabilizing transformation; \sqrt{y} , $\log(y)$, $1/y$, $\arcsin(\sqrt{y})$, and $1/\sqrt{y}$ are some commonly used transformations.
 - Transformations are also used as remedies for nonnormality

Ideas for Finding Proper Transformations

- Consider response Y with mean $E(Y)=\mu$ and variance $\text{Var}(Y)=\sigma^2$.
- That σ^2 depends on μ leads to nonconstant variances for different μ .
- Let f be a transformation and $\tilde{Y} = f(Y)$; What is the mean and variance of \tilde{Y} ?
- Approximate $f(Y)$ by a linear function (Delta Method):

$$f(Y) \approx f(\mu) + (Y - \mu)f'(\mu)$$

$$\text{Mean } \tilde{\mu} = E(\tilde{Y}) = E(f(Y)) \approx E(f(\mu)) + E((Y - \mu)f'(\mu)) = f(\mu)$$

$$\text{Variance } \tilde{\sigma}^2 = \text{Var}(\tilde{Y}) \approx [f'(\mu)]^2 \text{Var}(Y) = [f'(\mu)]^2 \sigma^2$$

- f is a good transformation if $\tilde{\sigma}^2$ does not depend on $\tilde{\mu}$ anymore. So, \tilde{Y} has constant variance for different $f(\mu)$.

Transformations

- Suppose σ^2 is a function of μ , that is $\sigma^2 = g(\mu)$
- Want to find transformation f such that $\tilde{Y} = f(Y)$ has constant variance: $\text{Var}(\tilde{Y})$ does not depend on μ .
- Have shown $\text{Var}(\tilde{Y}) \approx [f'(\mu)]^2 \sigma^2 = [f'(\mu)]^2 g(\mu)$
- Need to choose f such that $[f'(\mu)]^2 g(\mu) = \text{constant}$
- When $g(\mu)$ is known, f can be derived explicitly.

Examples (c is some unknown constant)

$$g(\mu) = c\mu \quad (\text{Poisson}) \quad f(Y) = \int \frac{1}{\sqrt{\mu}} d\mu \rightarrow f(Y) = \sqrt{Y}$$

$$g(\mu) = c\mu(1 - \mu) \quad (\text{Binomial}) \quad f(Y) = \int \frac{1}{\sqrt{\mu(1-\mu)}} d\mu \rightarrow f(Y) = \arcsin(\sqrt{Y})$$

$$g(\mu) = c\mu^{2\beta} (\beta \neq 1) \quad (\text{Box-Cox}) \quad f(Y) = \int \mu^{-\beta} d\mu \rightarrow f(Y) = Y^{1-\beta}$$

$$g(\mu) = c\mu^2 \quad (\text{Box-Cox}) \quad f(Y) = \int \frac{1}{\mu} d\mu \rightarrow f(Y) = \log Y$$

Box-Cox Transformations

- Assume $\sigma^2 = c\mu^{2\beta}$, then the variance-stabilizing transform should be

$$f(Y) = \begin{cases} Y^{1-\beta} & \beta \neq 1; \\ \log Y & \beta = 1 \end{cases}$$

These transformations are referred to as Box-Cox transformations.

Clearly it is crucial to know what β is.

As a matter of fact, β can be regarded as a parameter, and it can be estimated (identified) from data.

Identify Box-Cox Transformations: An Approximate Method

- From the assumption $\sigma^2 = c\mu^{2\beta}$, we have

$$\sigma_i^2 = c\mu_i^{2\beta} \text{ for treatments } i = 1, 2, \dots, a.$$

Take logarithm of both sides,

$$\log\sigma_i = \frac{1}{2}\log c + \beta\log\mu_i$$

- Let s_i and $\bar{y}_{i.}$ be the sample standard deviations and means. Because $\hat{\sigma}_i = s_i$ and $\hat{\mu}_i = \bar{y}_{i.}$, **approximately**,

$$\log s_i = \text{constant} + \beta\log\bar{y}_{i.},$$

where $i = 1, \dots, a$.

- We can plot $\log s_i$ against $\log\bar{y}_{i.}$, fit a straight line and use the slope to estimate β .

Identify Box-Cox Transformation: A Formal Method

Basic idea: try all possible transformations and choose the best one. For example, consider λ in an interval, e.g. $[-2, 2]$.

- 1 . Fix λ , transform data y_{ij} as follows,

$$y_{ij,\lambda} = \begin{cases} \frac{y_{ij}^\lambda - 1}{\lambda \dot{y}^{\lambda-1}} & \lambda \neq 0 \\ \dot{y} \log y_{ij} & \lambda = 0 \end{cases} \quad \text{where } \dot{y} = \left(\prod_{i=1}^a \prod_{j=1}^{n_i} y_{ij} \right)^{1/N} .$$

- 2 . Step 1 generates a transformed data $y_{ij,\lambda}$. Apply ANOVA to the new data and obtain its SS_E . Because SS_E depends on λ , it is denoted by $SS_E(\lambda)$.

- Repeat 1 and 2 for various λ in $[-2,2]$, and record $SS_E(\lambda)$

- 3 Find λ_0 that minimizes $SS_E(\lambda)$ and pick up a meaningful λ around λ_0 . Then the transformation is:

$$\tilde{y}_{ij} = y_{ij}^{\lambda_0} \text{ if } \lambda_0 \neq 0; \quad \tilde{y}_{ij} = \log y_{ij} \text{ if } \lambda_0 = 0.$$

An Example: `boxcox.dat`

```
trt response
1  0.948916
1  0.431494
1  3.486359
.  . . . .
.  . . . .

2  3.469623
2  0.840701
2  3.816014
2  1.234756
.  . . .
.  . . .

3  10.680733
3  19.453816
3  3.810572
3  10.832754
3  3.814586
```

Approximate Method: trans.sas

```
options nocenter ps=65 ls=80;
title1 'Increasing Variance Example';
data one;
  infile 'c:\saswork\data\boxcox.dat'; input trt resp;
proc glm data=one; class trt;
  model resp=trt; output out=diag p=pred r=res;

title1 'Residual Plot'; symbol1 v=circle i=none;
proc gplot data=diag; plot res*pred /frame;

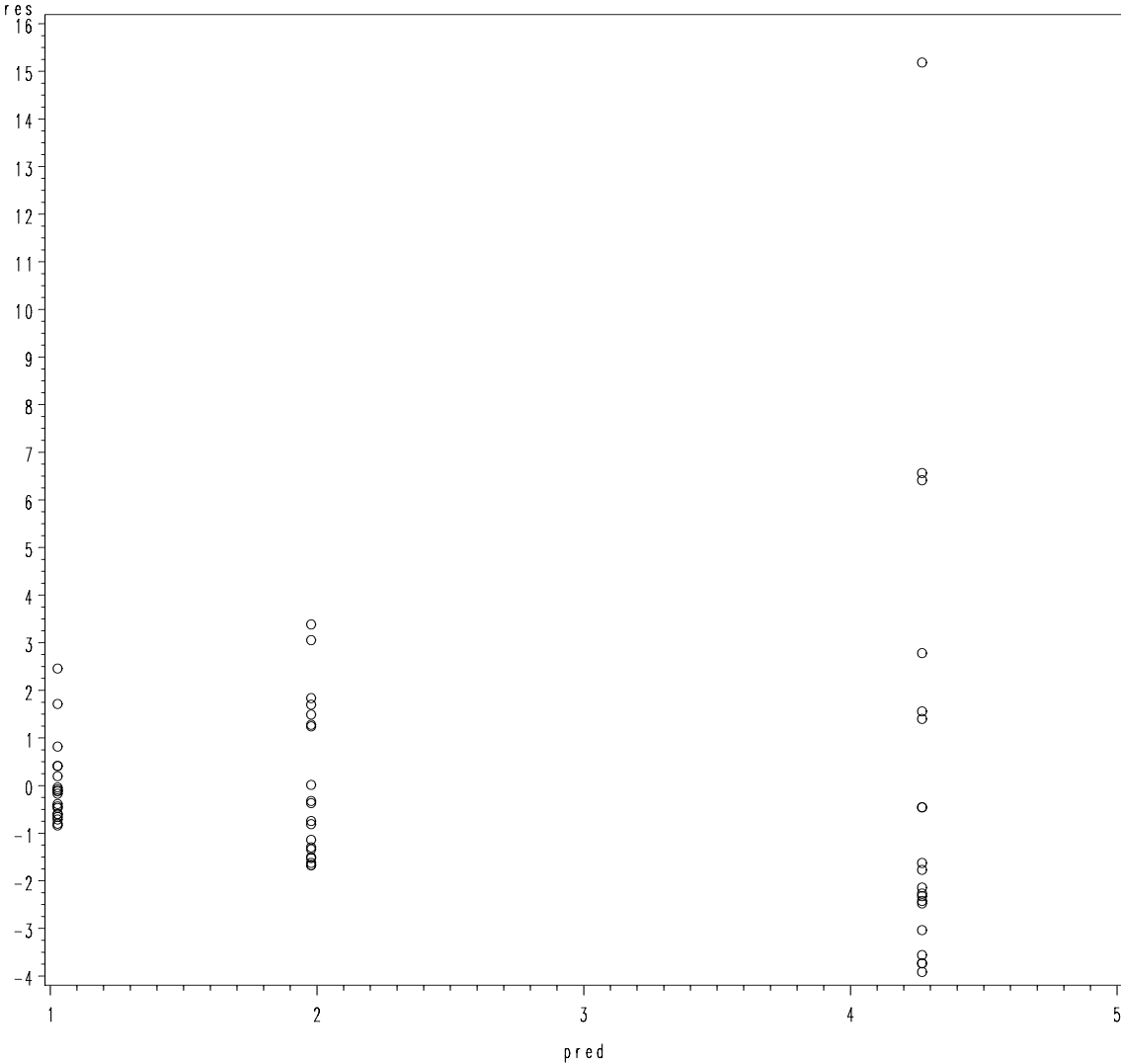
proc univariate data=one noprint;
  var resp; by trt; output out=two mean=mu std=sigma;
data three;
  set two; logmu = log(mu); logsig = log(sigma);

proc reg; model logsig = logmu;

title1 'Mean vs Std Dev'; symbol1 v=circle i=rl;
proc gplot; plot logsig*logmu / regeqn; run;
```

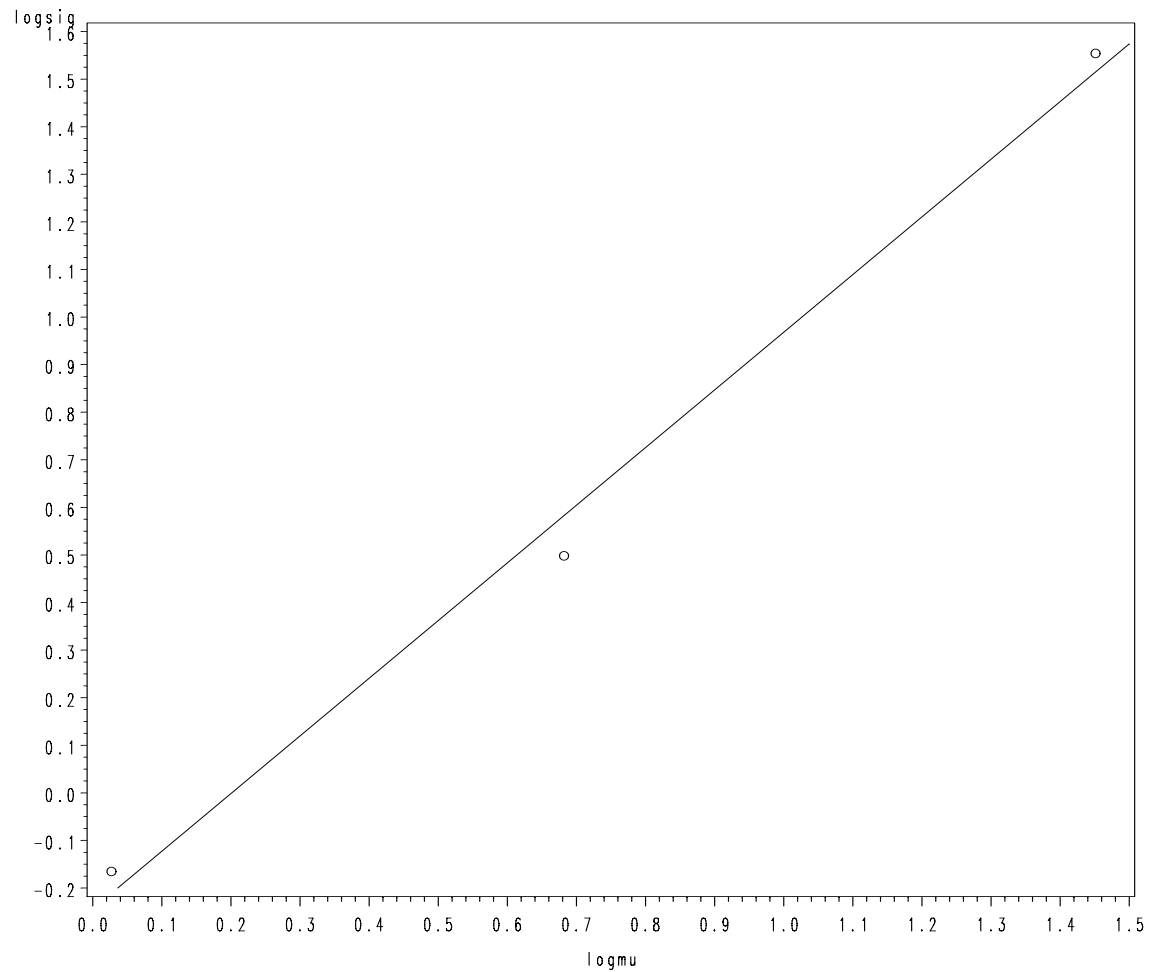
Residual Plot

Residual Plot



Plot of $\log s_i$ vs $\log \mu_i$

Mean vs Std Dev



Regression Equation:
 $\log s_i = -0.243928 + 1.212067 \cdot \log \mu_i$

Formal Method: trans1.sas

```
options ls=80 ps=65 nocenter;
title1 'Box-Cox Example';

data one;
  infile 'c:\saswork\data\boxcox.dat';
  input trt resp;
  logresp = log(resp);

proc univariate data=one noprint;
  var logresp; output out=two mean=mlogresp;

data three;
  set one; if _n_ eq 1 then set two;
  ydot = exp(mlogresp);
  do l=-1.0 to 1.0 by .25;
    den = l*ydot**(l-1);    if abs(l) eq 0 then den = 1;
    yl=(resp**l -1)/den;    if abs(l) < 0.0001 then yl=ydot*log(resp);
    output;
  end;
```

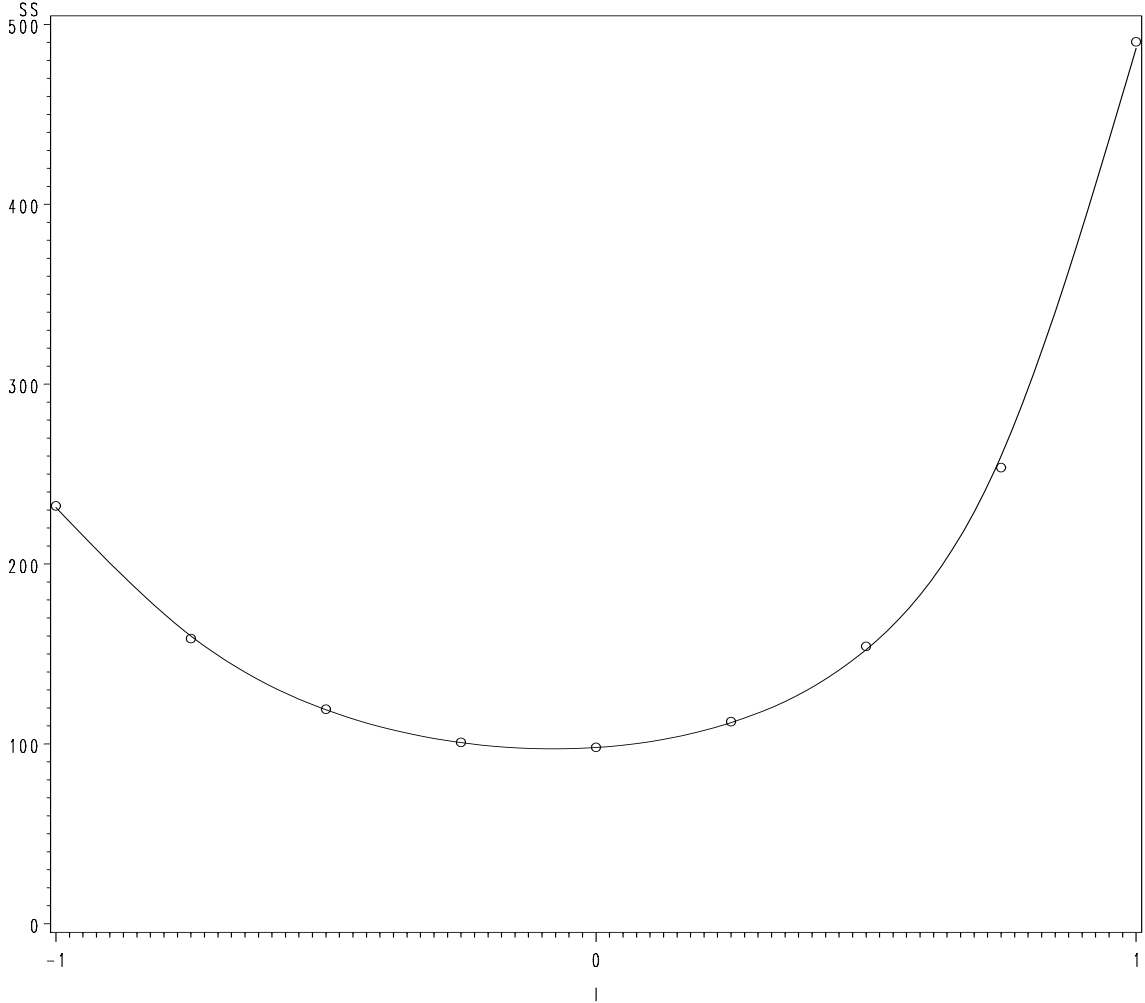
```
keep trt y1 l;  
  
proc sort data=three out=three; by l;  
proc glm data=three noprint outstat=four;  
  class trt; model y1=trt; by l;  
  
data five; set four;  
  if _SOURCE_ eq 'ERROR'; keep l SS;  
  
proc print data=five;  
run;  
  
symbol1 v=circle i=sm50;  
proc gplot;  
  plot SS*l;  
run;
```

$SS_E(\lambda)$ and λ

OBS	L	SS	OBS	L	SS
1	-2.00	2150.06	10	0.25	112.37
2	-1.75	1134.83	11	0.50	154.23
3	-1.50	628.94	12	0.75	253.63
4	-1.25	369.35	13	1.00	490.36
5	-1.00	232.32	14	1.25	1081.29
6	-0.75	158.56	15	1.50	2636.06
7	-0.50	119.28	16	1.75	6924.95
8	-0.25	100.86	17	2.00	19233.39
9	0.00	98.09			

Plot of $SS_E(\lambda)$ vs λ

Increasing Variance Example



Using Proc Transreg

```
proc transreg data=one;
model boxcox(y/lambda=-2.0 to 2.0 by 0.1)=class(trt); run;
```

The TRANSREG Procedure

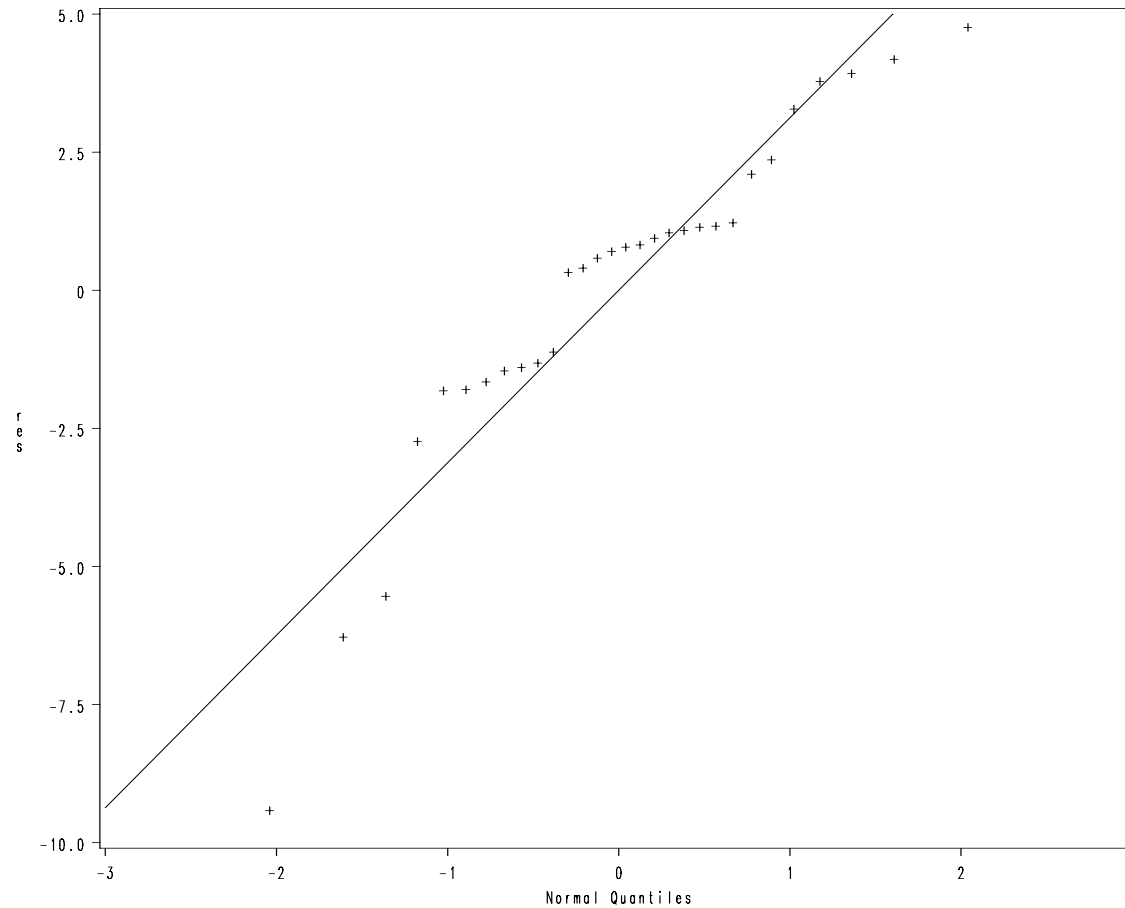
Transformation Information

for BoxCox(y)

Lambda	R-Square	Log Like	
-2.0	0.10	-108.906	
:	:	:	
-0.5	0.18	-22.154	
-0.4	0.19	-19.683	
-0.3	0.20	-17.814	*
-0.2	0.20	-16.593	*
-0.1	0.21	-16.067	<
0.0 +	0.21	-16.284	*
0.1	0.22	-17.289	*
0.2	0.22	-19.124	
0.3	0.22	-21.820	< Best Lambda
:	:	:	* Confidence Interval
2.0	0.10	-174.641	+ Convenient Lambda

Nonnormality

trt	nitrogen				
1	2.80	7.04	0.41	1.73	0.18
2	0.60	1.14	0.14	0.16	1.40
3	0.05	1.07	1.68	0.46	4.87
4	1.20	0.89	3.22	0.77	1.24
5	0.74	0.20	1.62	0.09	2.27
6	1.26	0.26	0.47	0.46	3.26



Test	---	Statistic	---	-----	p Value	-----
Shapiro-Wilk	W	0.910027	Pr < W	0.0149		

Kruskal-Wallis Test: a Nonparametric alternative

a treatments, H_0 : a treatments are not different.

- Rank the observations y_{ij} in ascending order
- Replace each observation by its rank R_{ij} (assign average for tied observations)

- Test statistic

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

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

- Decision Rule: reject H_0 if $H > \chi_{\alpha, a-1}^2$.
- Let F_0 be the F -test statistic in ANOVA based on R_{ij} . Then

$$F_0 = \frac{H/(a-1)}{(N-1-H)/(N-a)}$$

```
options nocenter ps=65 ls=80;

data new;
  input strain nitrogen @@;
  cards;
1  2.80  1  7.04  1  0.41  1  1.73  1  0.18
2  0.60  2  1.14  2  0.14  2  0.16  2  1.40
3  0.05  3  1.07  3  1.68  3  0.46  3  4.87
4  1.20  4  0.89  4  3.22  4  0.77  4  1.24
5  0.74  5  0.20  5  1.62  5  0.09  5  2.27
6  1.26  6  0.26  6  0.47  6  0.46  6  3.26
;
proc npar1way;
  class strain;
  var nitrogen;
run;
```

The NPAR1WAY Procedure

Analysis of Variance for Variable nitrogen

Classified by Variable strain

strain	N	Mean
--------	---	------

1	5	2.4320
2	5	0.6880
3	5	1.6260
4	5	1.4640
5	5	0.9840
6	5	1.1420

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
--------	----	----------------	-------------	---------	--------

Among	5	9.330387	1.866077	0.7373	0.6028
Within	24	60.739600	2.530817		

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable nitrogen
Classified by Variable strain

strain	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
1	5	93.00	77.50	17.967883	18.60
2	5	57.00	77.50	17.967883	11.40
3	5	78.50	77.50	17.967883	15.70
4	5	93.00	77.50	17.967883	18.60
5	5	68.00	77.50	17.967883	13.60
6	5	75.50	77.50	17.967883	15.10

Average scores were used for ties.

Kruskal-Wallis Test

Chi-Square 2.5709
DF 5
Pr > Chi-Square 0.7658