Topic 5: Model adequacy checking

Montgomery: chapter 3

Prof. Lingling An
University of Arizona

Example: Tensile Strength

- Investigate the tensile strength of a new synthetic fiber. The factor is the weight percent of cotton used in the blend of the materials for the fiber and it has five levels.

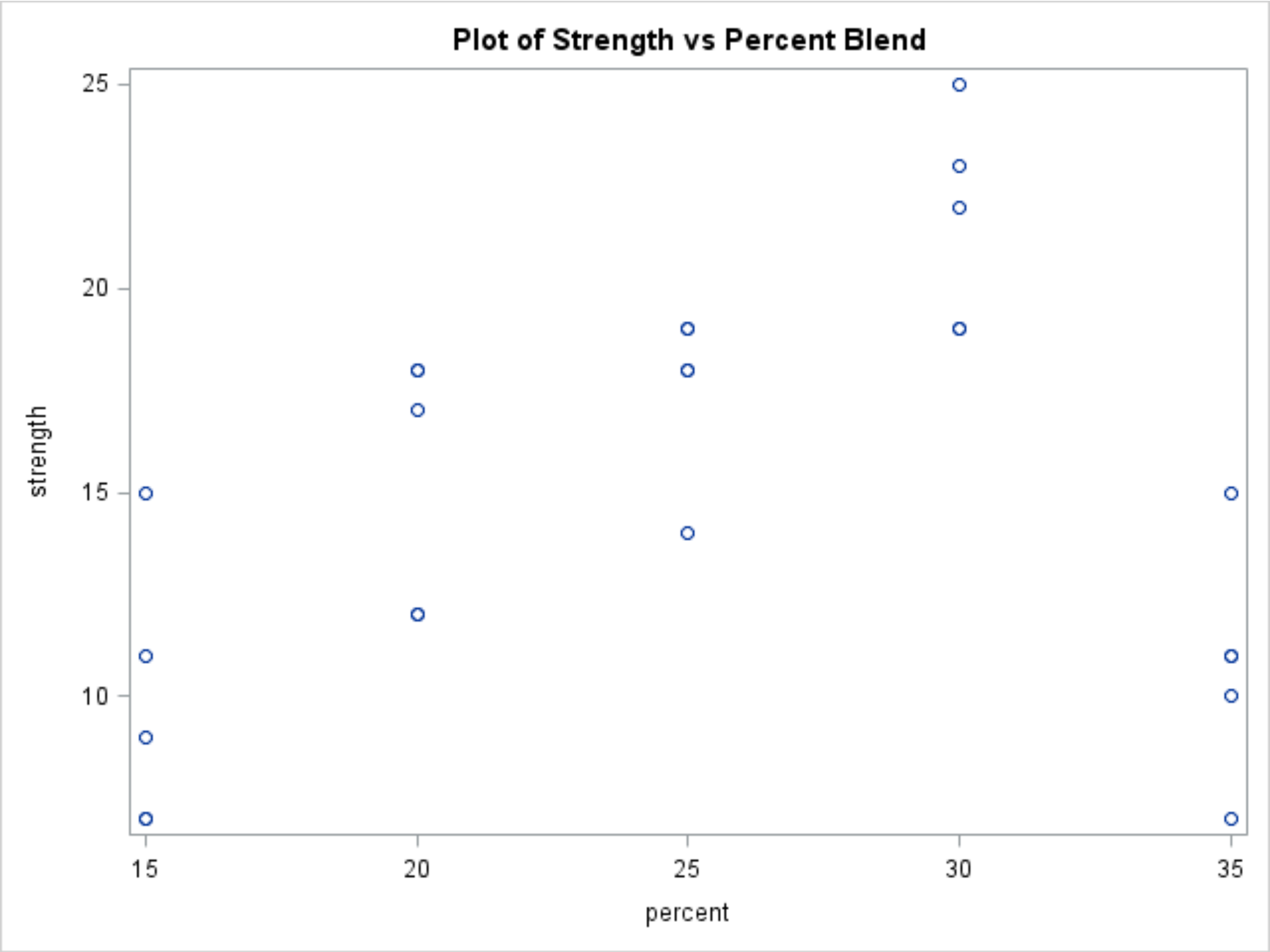
percent	tensile strength						
of cotton	1	2	3	4	5	total	average
15	7	7	11	15	9	49	9.8
20	12	17	12	18	18	77	15.4
25	14	18	18	19	19	88	17.6
30	19	25	22	19	23	108	21.6
35	7	10	11	15	11	54	10.8

SAS code

```
options ls=75 ps=60 nocenter;

data one;
 infile 'D:\TEACHING\T_stat571B\lab\sas_data
\tensile.dat';
 input percent strength time;
 run;

title1 'Tensile Strength  example';
proc print data=one;
run;
```

```
title1 'Plot of Strength vs Percent Blend';
proc sgplot data=one;
scatter x=percent y=strength;
run;

proc boxplot;
 plot strength*percent/boxstyle=skeletal
pctldef=4;
 run;

title1 ‘ANOVA analysis';
 proc glm data=one;
 class percent;
 model strength=percent;
 output out=diag p=pred r=res;
run;
```

Scatter plot



The GLM Procedure
Dependent Variable: strength

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	475.760000	118.9400000	14.76	<.0001
Error	20	161.200000	8.0600000		
Corrected Total	24	636.960000			

R-Square	Coeff Var	Root MSE	strength Mean
0.746923	18.87642	2.839014	15.04000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
percent	4	475.7600000	118.9400000	14.76	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
percent	4	475.7600000	118.9400000	14.76	<.0001

Model checking and diagnoistic

- **Checking assumptions** is important
  - Have we fit the right model?
  - Normality
  - Independence
  - Constant variance

$$
\begin{array}{ccccc}
y_{ij} & = & (\overline{y}_{..} + (\overline{y}_{i.} - \overline{y}_{..})) & + & (y_{ij} - \overline{y}_{i.}) \\
y_{ij} & = & \hat{y}_{ij} & + & \hat{\epsilon}_{ij} \\
\text{observed} & = & \text{predicted} & + & \text{residual}
\end{array}
$$

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

- Normality
  - Histogram of residuals
  - Normal probability plot / QQ plot
  - Shapiro-Wilk Test
- 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
  - Plot $\hat{\epsilon}_{ij}$ vs variable of interest
- Outliers

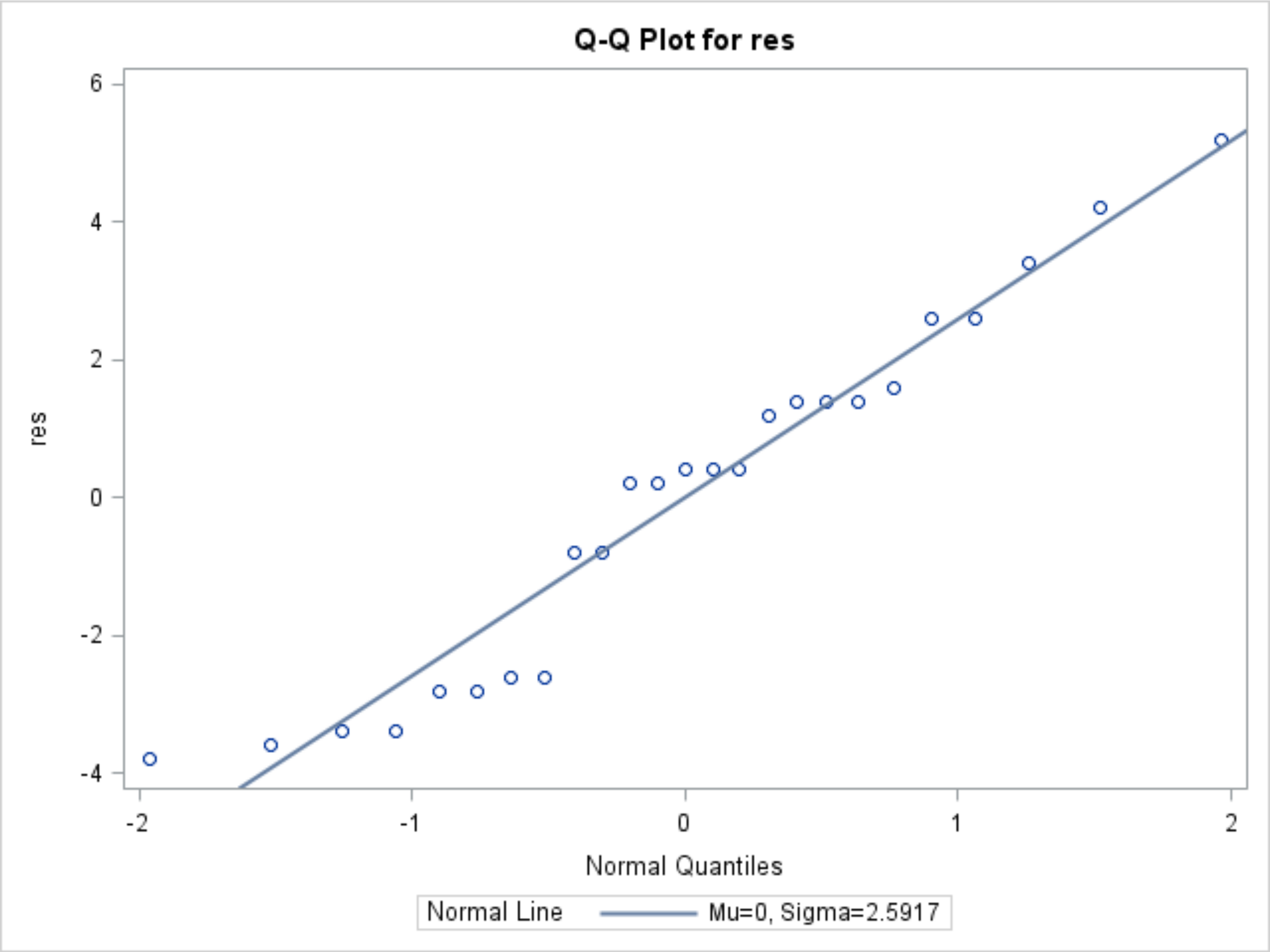
Normality Checking in the ANOVA

- Examination of **residuals** (see text, Sec. 3.4, p80)

$$e_{ij} = y_{ij} - \hat{y}_{ij}$$
$$= y_{ij} - \bar{y}_{i.}$$

- **Residual plots** are very useful – e.g., Q-Q plot

```
title "normality checking";
proc univariate data=diag normal;
var res;
qqplot res/normal(mu=est sigma=est
color=red L=1);
run;
```


Q-Q Plot for res
6
4
2
0
-2
-4
res
-2
-1
0
1
2
Normal Quantiles
Normal Line
Mu=0, Sigma=2.5917

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.943868	Pr < W	0.1818
Kolmogorov-Smirnov	D	0.162123	Pr > D	0.0885
Cramer-von Mises	W-Sq	0.080455	Pr > W-Sq	0.2026
Anderson-Darling	A-Sq	0.518572	Pr > A-Sq	0.1775

Outliers checking

- Use standardized residuals to check if there is outliers

$$d_{ij} = \frac{e_{ij}}{\sqrt{MSE}}$$

- > 3 or <(-3) is a potential outlier
- Be careful for removing outliers
- Read text on p82

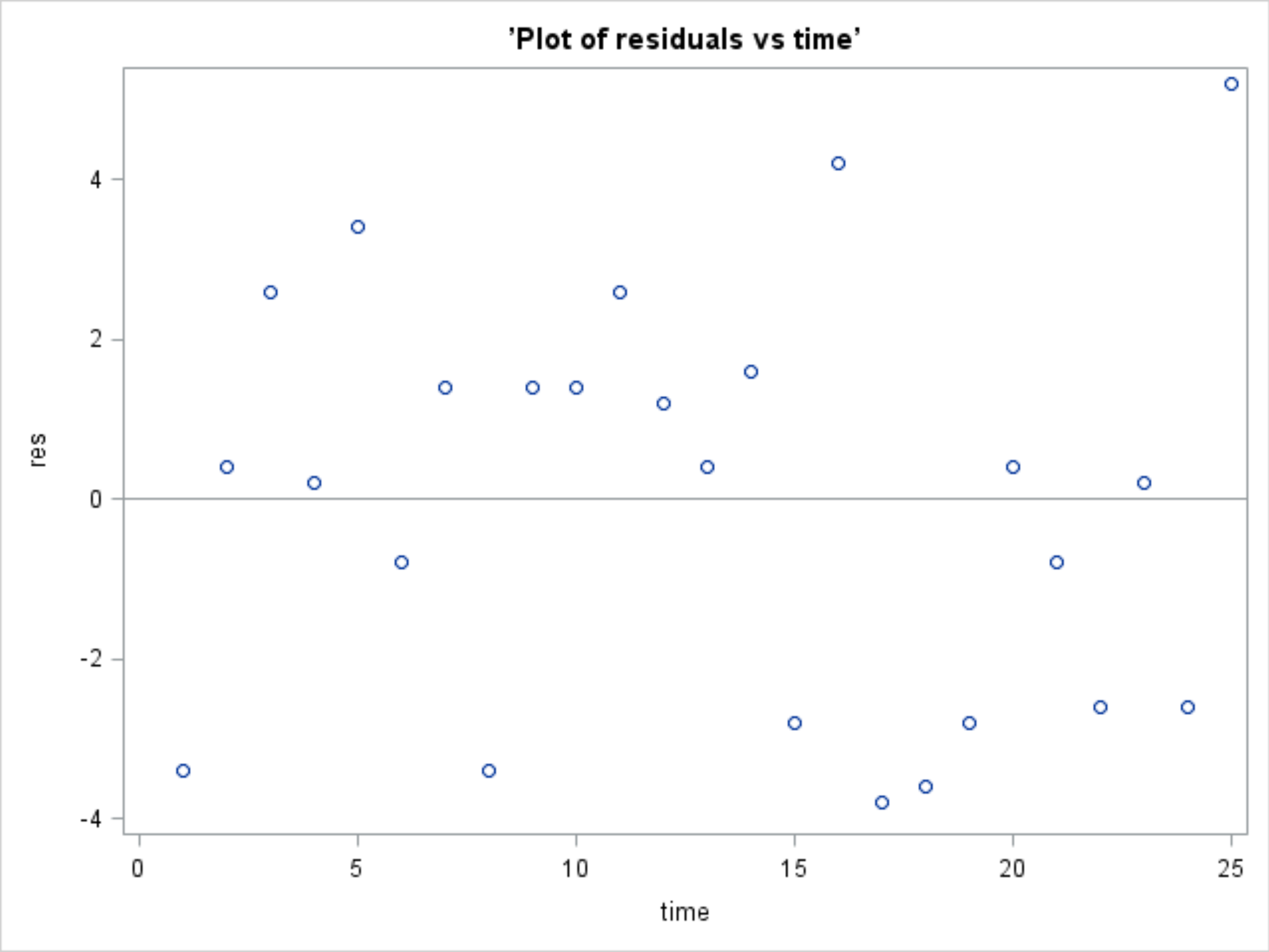
SAS code: checking outlier

```
data outlier;
set diag;
stdres=res/sqrt(8.06);
run;

proc print data=outlier;
run;
```

Obs	percent	strength	time	pred	res	stdres
1	15	7	15	9.8	-2.8	**-0.98626**
2	15	7	19	9.8	-2.8	**-0.98626**
3	15	15	25	9.8	5.2	**1.83162**
4	15	11	12	9.8	1.2	**0.42268**
5	15	9	6	9.8	-0.8	**-0.28179**
6	20	12	8	15.4	-3.4	**-1.19760**
7	20	17	14	15.4	1.6	**0.56358**
8	20	12	1	15.4	-3.4	**-1.19760**
9	20	18	11	15.4	2.6	**0.91581**
10	20	18	3	15.4	2.6	**0.91581**
11	25	14	18	17.6	-3.6	**-1.26805**
12	25	18	13	17.6	0.4	**0.14089**
13	25	18	20	17.6	0.4	**0.14089**
14	25	19	7	17.6	1.4	**0.49313**
15	25	19	9	17.6	1.4	**0.49313**
16	30	19	22	21.6	-2.6	**-0.91581**
17	30	25	5	21.6	3.4	**1.19760**
18	30	22	2	21.6	0.4	**0.14089**
19	30	19	24	21.6	-2.6	**-0.91581**
20	30	23	10	21.6	1.4	**0.49313**
21	35	7	17	10.8	-3.8	**-1.33849**
22	35	10	21	10.8	-0.8	**-0.28179**
23	35	11	4	10.8	0.2	**0.07045**
24	35	15	16	10.8	4.2	**1.47939**
25	35	11	23	10.8	0.2	**0.07045**

SAS code: independence checking

```
title1 'Plot of residuals vs time';
proc sgplot data=diag;
scatter y=res x=time;
refline 0;
run;
```


'Plot of residuals vs time'
res
time
4
2
0
-2
-4
0
5
10
15
20
25

Constant variance checking

- 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.

Constant variance - 2

Bartlett's Test

- $H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_a^2$
- Test statistic: $\chi_0^2 = 2.3026\frac{q}{c}$

  where

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

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

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

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

Remark: sensitive to normality assumption.

Constant variance - 3

Modified Levene's Test

- For each fixed i, calculate the median m_i of $y_{i1}, y_{i2}, \dots, y_{in_i}$.
- Compute the absolute deviation of observation from sample median:

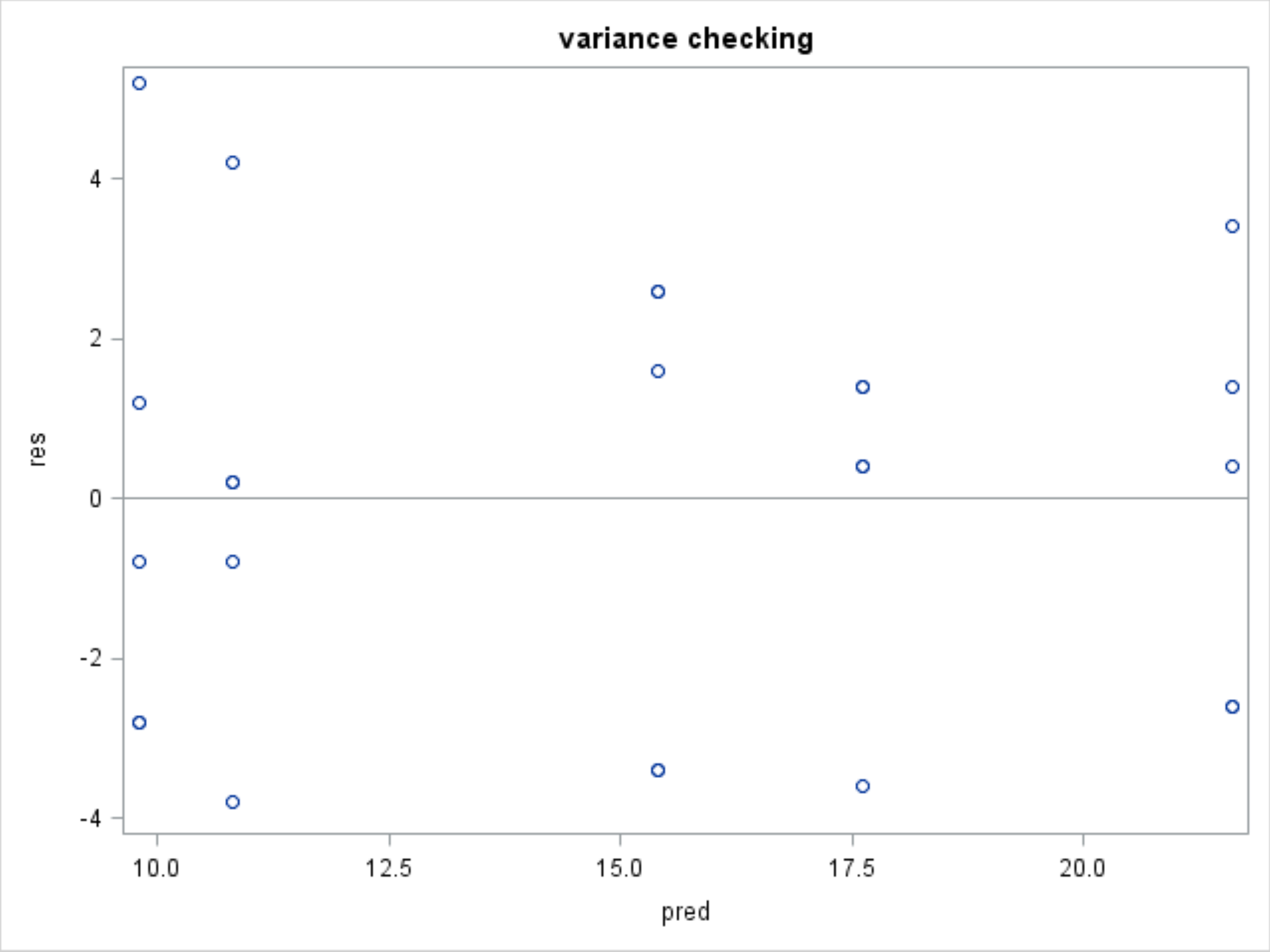
$$d_{ij} = |y_{ij} - m_i|$$

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 $H_0 : \sigma_1^2 = \dots = \sigma_a^2$.

SAS code

```
title 'variance checking';
proc glm data=one;
class percent;
model strength=percent;
means percent / hovtest=bartlett
hovtest=levene;
output out=diag2 p=pred r=res;
run;


proc sgplot data=diag2;
scatter x=pred y=res;
refline 0;
run;
```


variance checking
res
4
2
0
-2
-4
10.0
12.5
15.0
17.5
20.0
pred

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		

Bartlett's Test for Homogeneity of strength Variance			
Source	DF	Chi-Square	Pr > ChiSq
percent	4	0.9331	0.9198

Non-constant Variance: Impact and Remedy

- Why concern?
  - Comparison of treatments depends on MSE
  - Incorrect intervals and comparison results
- Variance-Stabilizing Transformations
  - Common transformations

    $\sqrt{x}$, $\log(x)$, $1/x$, $\arcsin(\sqrt{x})$, and $1/\sqrt{x}$
  - Box-Cox transformations
    1. approximate the relationship $\sigma_i = \theta\mu_i^\beta$, then the transformation is $X^{1-\beta}$
    2. use maximum likelihood principle

    * Distribution often more "normal" after transformation

Ideas for Finding Proper Transformations

- Consider response Y with mean E(Y)=μ and variance Var(Y)=σ^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):

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

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

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

- 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: Var($\tilde{Y}$) does not depend on μ.
- Have shown Var($\tilde{Y}$)$\approx [f'(\mu)]^2\sigma^2 \approx [f'(\mu)]^2 g(\mu)$
- Want to choose f such that $[f'(\mu)]^2 g(\mu) \approx c$

Examples

$g(\mu) = \mu$	(Poisson)	$f(X) = \int \frac{1}{\sqrt{\mu}} d\mu \to f(X) = \sqrt{X}$
$g(\mu) = \mu(1-\mu)$	(Binomial)	$f(X) = \int \frac{1}{\sqrt{\mu(1-\mu)}} d\mu \to f(X) = \arcsin(\sqrt{X})$
$g(\mu) = \mu^{2\beta}$	(Box-Cox)	$f(X) = \int \mu^{-\beta} d\mu \to f(X) = X^{1-\beta}$
$g(\mu) = \mu^2$	(Box-Cox)	$f(X) = \int \frac{1}{\mu} d\mu \to f(X) = \log X$

Identify Box-Cox Transformation Using Data: Approximate Method

- From the previous slide, if $\sigma = \theta\mu^{\beta}$, the transformation is

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

  So it is crucial to estimate β based on data y_{ij}, $i = 1, \ldots, a$.

- We have $\log\sigma_i = \log\theta + \beta\log\mu_i$

- 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 $i = 1, \ldots, a$.

- 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: Formal Method

1 . For a fixed λ, perform analysis of variance on

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

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

- Repeat 1 and 2 for various λ in an interval, e.g., [-2,2], and record SS$_E(\lambda)$

3 Find λ_0 which minimizes SS$_E(\lambda)$ and pick up a meaningful λ in the neighborhood of λ_0. Denote it again by λ.

4 The transformation is:

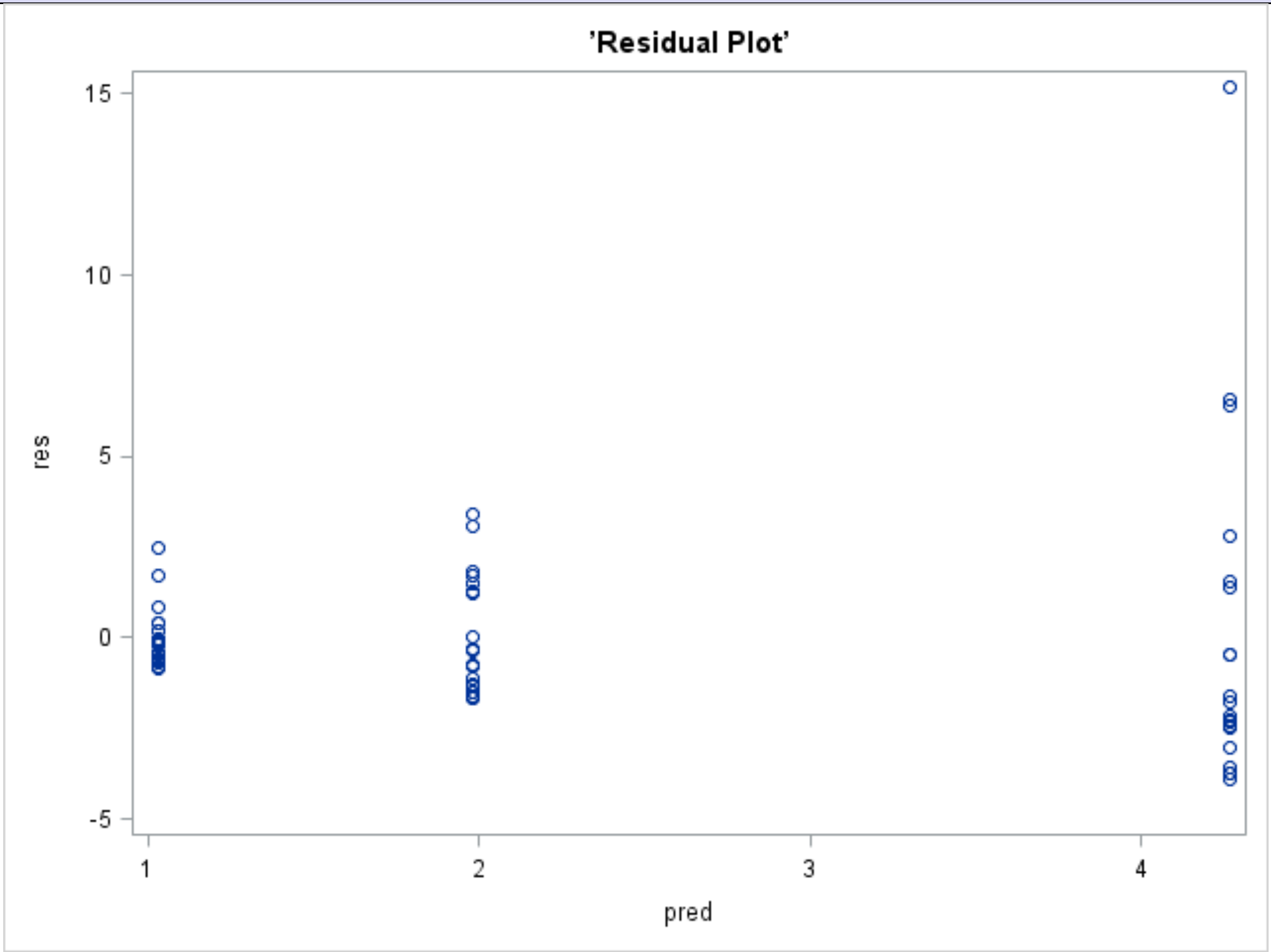
$$\tilde{y}_{ij} = y_{ij}^{\lambda_0} \text{ if } \lambda_0 \neq 0;$$
$$\tilde{y}_{ij} = \log y_{ij} \text{ if } \lambda_0 = 0.$$

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
```

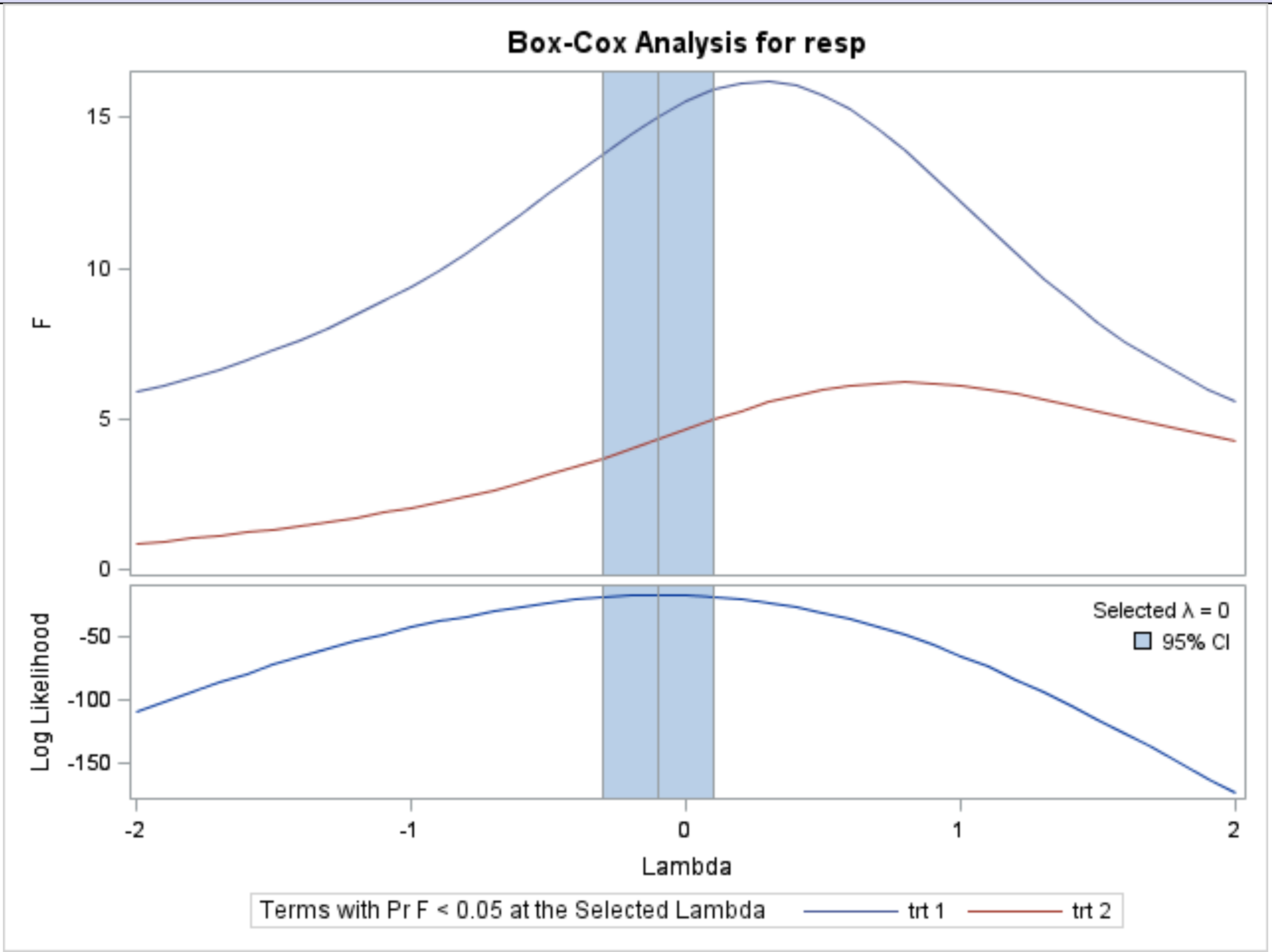
SAS code

```
data two;
 infile 'D:\TEACHING\T_stat571B\lab
\sas_data\boxcox.dat';
 input trt resp;
 run;

 proc glm data=two;
class trt;
model resp=trt;
output out=three p=pred r=res;
run;

title1 'Residual Plot';
proc sgplot data=three;
scatter y=res x=pred;
run;
```


'Residual Plot'
15
10
5
0
-5
res
1
2
3
4
pred

Use boxcox transformation on the response variable

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



Box-Cox Analysis for resp
F
15
10
5
0
Log Likelihood
-50
-100
-150
Selected λ = 0
95% CI
-2
-1
0
1
2
Lambda
Terms with Pr F < 0.05 at the Selected Lambda
trt 1
trt 2

Nonparametric methods for ANOVA

H_0: *a* treatments are equal. H_a: at least one not equal.
(But normality assumption is unsatisfied)

- Kruskal-Wallis Test
  - Rank the observations y_{ij} in ascending order
  - Replace each observation by its rank R_{ij} (assign average for tied observations)
  - Test statistic
    - $H = \frac{1}{S^2}\left[\sum_{i=1}^{a} \frac{R_{i.}^2}{n_i} - \frac{N(N+1)^2}{4}\right] \approx \chi^2_{a-1}$
    - 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 $H > \chi^2_{\alpha,a-1}$.

SAS code

```
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;
```

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		
Average scores were used for ties.					

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

Last slide

- Read Sections: sections 3.4 and 3.11