STAT 525 FALL 2018

Chapter 18
ANOVA Diagnostics and Remedies

Professor Min Zhang

Overview

- General assumptions
  - Normally distributed error terms
  - Independent observations
  - Constant variance
- Will adapt diagnostics and remedial measures from regression
- Many are the same but others require slight modifications

Residuals

- Predicted values are the cell means

$$\hat{\mu}_i = \overline{Y}_{i.}$$

- Residuals are the difference between the observed and predicted

$$e_{ij} = Y_{ij} - \overline{Y}_{i.}$$

- Properties:
  - Same least squares properties
  - $\sum_j e_{ij} = 0$

Basic Plots

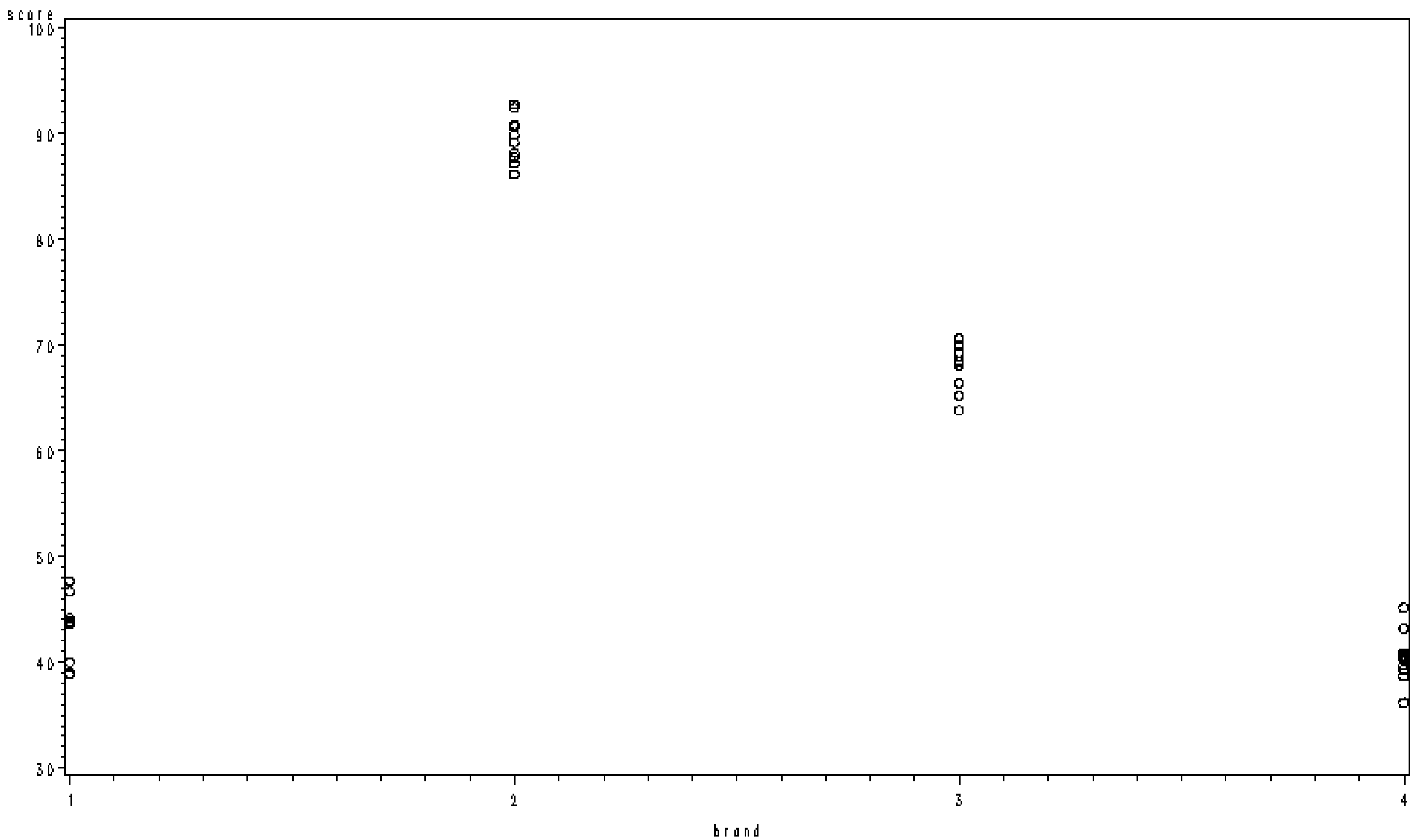
- Plot the data vs the factor levels
- Plot the residuals vs the factor levels
- Plot the residuals vs the fitted values
- Histogram of the residuals
- QQplot of the residuals

Example (Page 777)

- Experiment designed to study the effectiveness of four rust inhibitors
- Forty units were used in the experiment
- Units randomly and equally assigned to rust inhibitors ($n_i = 10$)
- Each unit exposed to severe weather conditions
- Y coded score (higher means less rust)
- X brand of rust inhibitor
  - $i = 1, 2, 3, 4$
  - $j = 1, 2, .., 10$

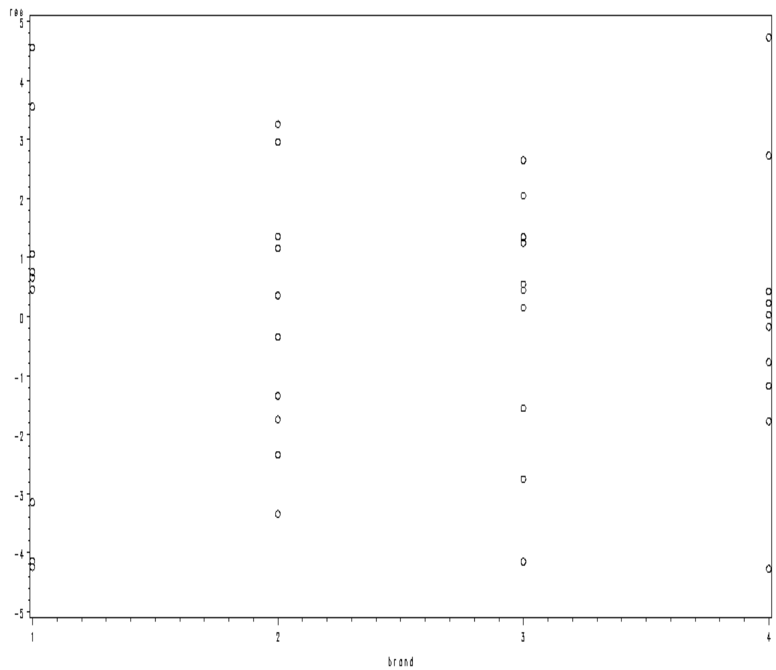
Scatterplot

```
options nocenter; goptions colors=('none');
data a1;
    infile 'u:\.www\datasets525\CH17TA02.txt';
    input score brand;

symbol1 v=circle i=none;
proc gplot data=a1;
    plot score*brand;
run; quit;
```

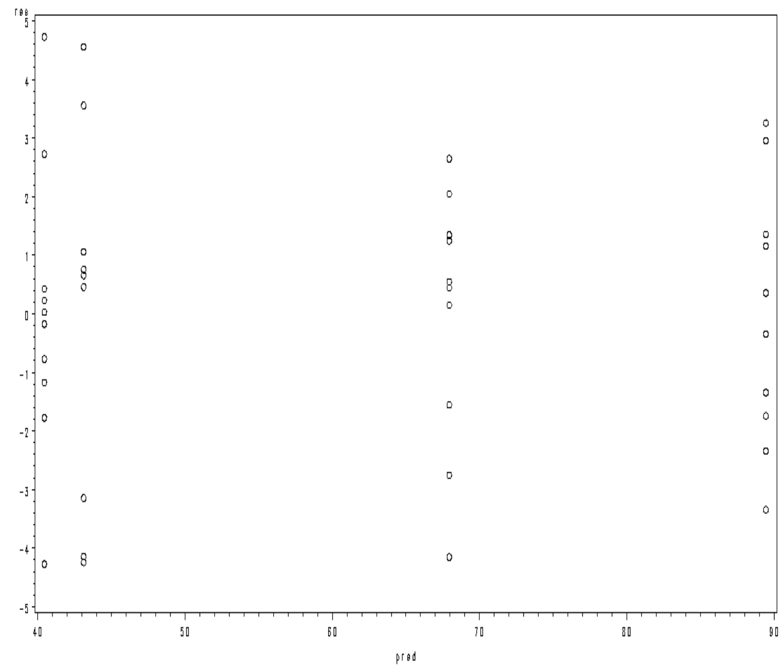


Residual Plots

```
proc glm data=a1;
    class brand;
    model score=brand;
    output out=a2 r=res p=pred;

proc gplot;
    plot res*(brand pred);
run; quit;
```



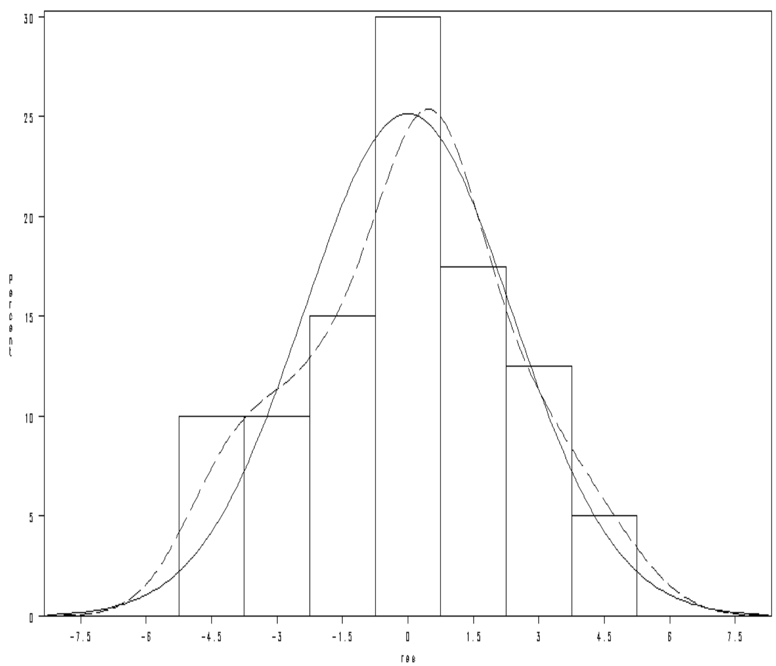
Residual vs. Brand



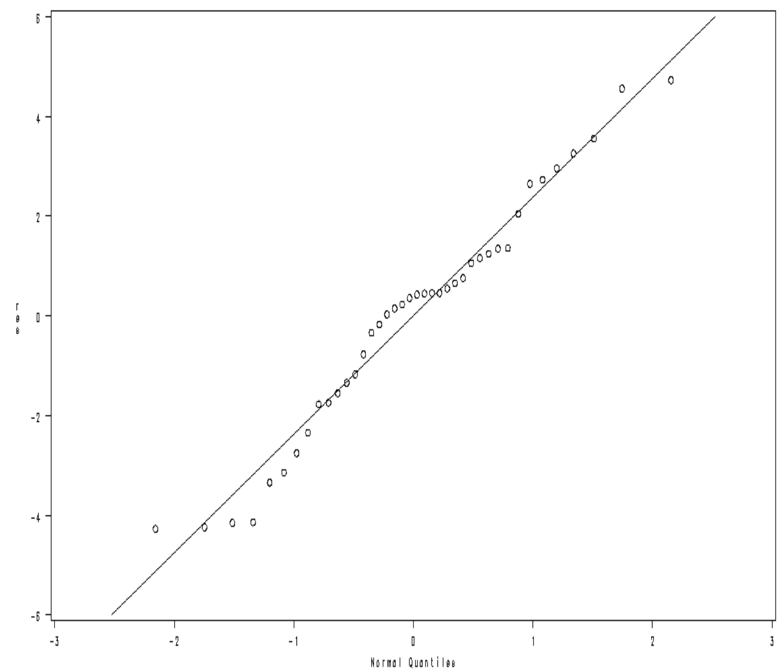
Residual vs. $\widehat{Y}_i$

Histogram & QQPlot

```
proc univariate noprint data=a2;
    histogram res / normal kernel(L=2);
    qqplot res / normal (L=1 mu=est sigma=est);
run; quit;
```



Histogram of Residuals



QQPlot of Residuals

Summary

- Look for
  - Outliers
  - Non-constant variance
  - Non-normal errors
- Can plot residuals vs time or other variables if available
  - Independent observations

Formal Tests

- Normality
  - Wilk-Shapiro
  - Anderson-Darling
  - Kolmogorov-Smirnov
- Homogeneity of Variance
  - Hartley test
  - Modified Levene test (aka Brown-Forsythe test in SAS)
  - Bartlett's

Homogeneity of Variance: Hartley Test

- It requires equal sample sizes across factor levels, i.e., $n_i = n$
- Hartley statistic,

$$H^* = \frac{\max(s_i^2)}{\min(s_i^2)} \sim H(r, n-1), \text{under } H_0$$

- Percentiles of H(r,df) are shown in Table B.10 (p. 1336).

Homogeneity of Variance: Modified Levene Test

- Called Brown-Forsythe test in SAS
- Test statistic,
  - Define $d_{ij} = |Y_{ij} - \tilde{Y}_i|$, with $\tilde{Y}_i$ the median at factor level i
  - Calculate $\bar{d}_{i\cdot} = \sum_j d_{ij}/n_i$, $\bar{d}_{\cdot\cdot} = \sum_i \sum_j d_{ij}/n_T$
  - Calculate

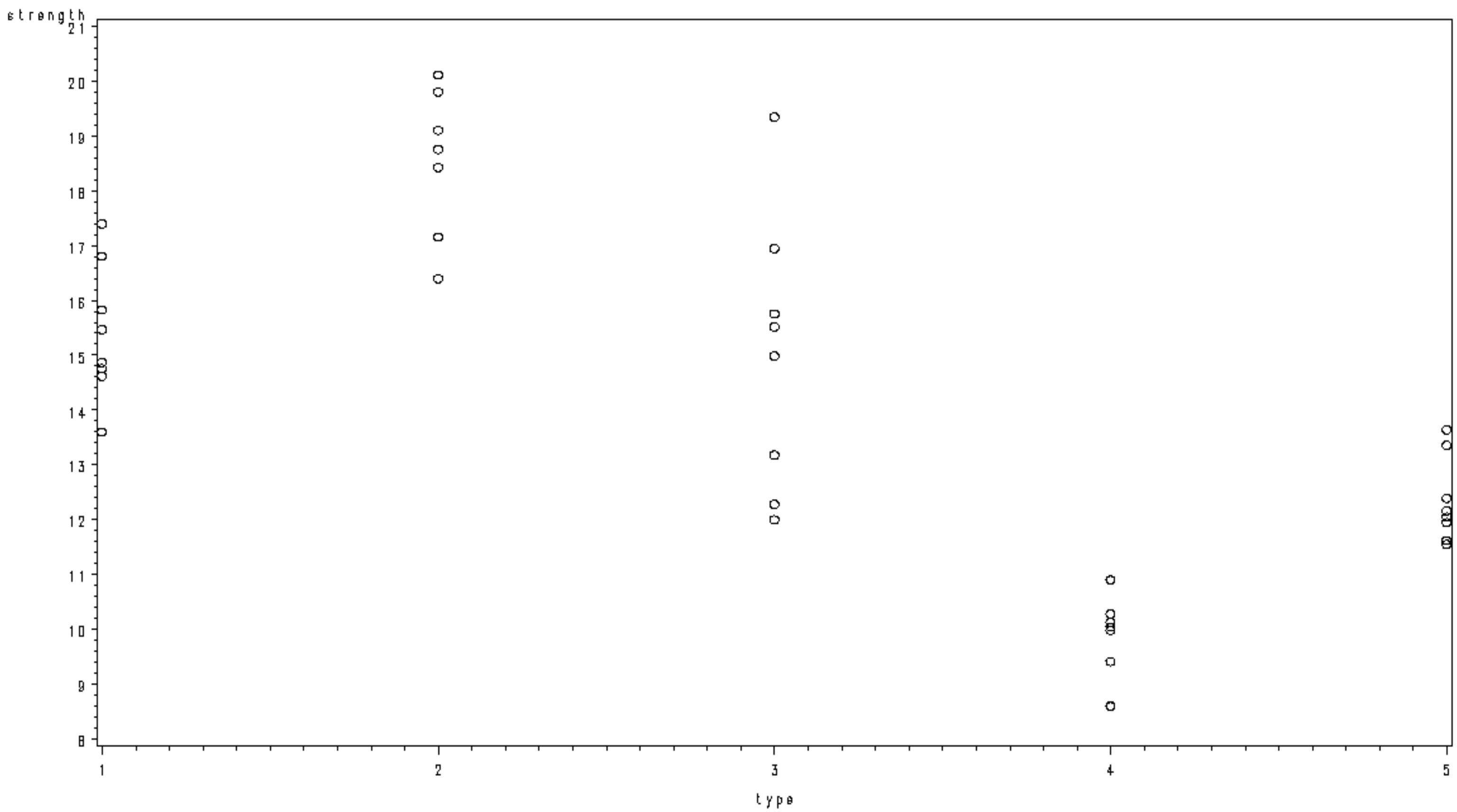
$$\begin{aligned} MSTR &= \sum_i n_i(\bar{d}_{i\cdot} - \bar{d}_{\cdot\cdot})^2/(r-1), \\ MSE &= \sum_i \sum_j (d_{ij} - \bar{d}_{i\cdot})^2/(n_T - r) \end{aligned}$$

  - $F^*_{BF} = MSTR/MSE \overset{approx}{\sim} F(r-1, n_T - r)$ under H_0
- Modified Levene test is often the best choice
  - Unlike the Hartley test, it is robust against departures from normality
  - It does not require equal sample sizes
- In `PROC GLM`, Use option `HOVTEST=BF` for `MEANS` statement

Example (Page 783)

- Experiment designed to assess the strength of five types of flux used in soldering wire boards
- Forty units were used in the experiment
- Units randomly and equally assigned to five types of flux ($n_i = 8$)
- Y – strength
- X – type of flux

```
data a1;
    infile 'u:\.www\datasets525\CH18TA02.DAT';
    input strength type;

/* Scatterplot */
proc gplot data=a1;
    plot strength*type;
run; quit;
```



```
/* Modified Levene Test */
proc glm data=a1;
    class type;
    model strength=type;
    means type / hovtest=bf clm;
run; quit;
```

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

                          Sum of         Mean
Source           DF      Squares       Square     F Value     Pr > F
type              4       9.3477       2.3369        2.94     0.0341
Error            35      27.8606       0.7960
```

```
Level of              -----------strength----------
type            N             Mean          Std Dev
1               8       15.4200000       1.23713956
2               8       18.5275000       1.25297076
3               8       15.0037500       2.48664397
4               8        9.7412500       0.81660337
5               8       12.3400000       0.76941536
```

Remedies

- Delete potential outliers
  - Is their removal important?
- Use weighted regression
- Box-Cox Transformation
- Non-parametric procedures

Variance Stabilization Transformations

- Consider response Y with E(Y)=μ_x and Var(Y)=$\sigma_x^2 = g(\mu_x)$
  - σ_x^2 depends on μ_x
- Want to find $\tilde{Y} = f(Y)$ such that Var($\tilde{Y}$)$\approx c$
  - What are the mean and var of $\tilde{Y}$?

Delta Method

Consider $f(Y)$ where $f'(\mu_x) \neq 0$

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

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

$$\text{Var}(\tilde{Y}) \approx [f'(\mu_x)]^2\text{Var}(Y) = [f'(\mu_x)]^2\sigma_x^2$$

- Want to choose f such that $[f'(\mu_x)]^2 g(\mu_x) \approx c$

Examples

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

Transformation Guides

- Regress log(s_i) vs log($\overline{Y_{i.}}$) $\rightarrow$ $\widehat{\lambda}$ =1-slope for $\tilde{Y} = Y^{\lambda}$
  - $f(\mu) = \mu^{\lambda} \Longrightarrow \log\sqrt{g(\mu)} = -\log\lambda + (1-\lambda)\log\mu$
  - When $\sigma_i^2 \propto \mu_i$ use $\sqrt{\ }$
  - When $\sigma_i \propto \mu_i$ use log
  - When $\sigma_i \propto \mu_i^2$ use $1/Y$
- For proportions, use arcsin($\sqrt{\ }$)
  - Use `arsin(sqrt(Y))` in SAS data step

Example (Page 783)

```
proc transreg data=a1;
    model boxcox(strength)=class(type);
run; quit;
```

```
  Lambda      R-Square    Log Like
   -1.50          0.86    -15.3143
   -1.25          0.86    -14.2378 *
   -1.00          0.86    -13.4223 *
   -0.75          0.86    -12.8608 *
   -0.50          0.85    -12.5428 *
   -0.25          0.85    -12.4549 <
    0.00 +        0.85    -12.5819 *
    0.25          0.84    -12.9078 *
    0.50          0.84    -13.4163 *
    0.75          0.83    -14.0919 *
    1.00          0.83    -14.9199
    1.25          0.82    -15.8868
    1.50          0.81    -16.9807

< - Best Lambda
* - Confidence Interval
+ - Convenient Lambda
```

- Log-transformation is suggested here.
- May also explore the relationship between s_i vs $\bar{Y}_{i\cdot}$ as shown on P.790-791.

Nonparametric Approach

- Based on ranking the observations and using the ranks
  - Rank Y_{ij} in ascending order from 1 to n_T, i.e., R_{ij}
  - Specify the score $d_{ij} = d(R_{ij})$
  - Apply One-Way ANOVA to d_{ij}, $1 \leq j \leq n_i$, $1 \leq i \leq r$
- Wilcoxon Scores, $d(R_{ij}) = R_{ij}$
  - Produces the Kruskal-Wallis test in the one-way ANOVA
  - Produces the Man-Whitney-Wilcoxon test for two-sample data ($r = 2$)
- Median scores, $d(R_{ij}) = 1[R_{ij} > (n_T + 1)/2]$
  - Produces the Brown-Mood test in the one-way ANOVA
  - Produces the median test for two-sample data ($r = 2$)
- SAS procedure `PROC NPAR1WAY`

Example (Page 783)

```
proc npar1way data=a1 median wilcoxon;
    class type;
    var strength;
run; quit;
```

```
         Wilcoxon Scores (Rank Sums) for Variable strength
                    Classified by Variable type

                       Sum of      Expected       Std Dev            Mean
type          N        Scores      Under H0      Under H0           Score
--------------------------------------------------------------------------
1             8         201.0         164.0     29.573377         25.1250
2             8         282.0         164.0     29.573377         35.2500
3             8         190.0         164.0     29.573377         23.7500
4             8          36.0         164.0     29.573377          4.5000
5             8         111.0         164.0     29.573377         13.8750

                 Average scores were used for ties.

   Kruskal-Wallis Test

Chi-Square         32.1634
DF                       4
Pr > Chi-Square     <.0001
```

```
Median Scores (Number of Points Above Median) for Variable strength
                    Classified by Variable type

                        Sum of      Expected       Std Dev          Mean
type          N         Scores      Under H0      Under H0         Score
--------------------------------------------------------------------------
1             8            7.0           4.0      1.281025        0.8750
2             8            8.0           4.0      1.281025        1.0000
3             8            5.0           4.0      1.281025        0.6250
4             8            0.0           4.0      1.281025        0.0000
5             8            0.0           4.0      1.281025        0.0000

                 Average scores were used for ties.


 Median One-Way Analysis

Chi-Square           28.2750
DF                         4
Pr > Chi-Square       <.0001
```

- χ^2-distributions, instead of F-distributions, are used due to the fact that the error variances are known in theory.
- Exact tests can be taken using the statement `EXACT MEDIAN WILCOXON;`
  - Recommended for small, sparse, skewed, or heavily tied dataset.

Chapter Review

- Diagnostics
  - Plots
  - Residual checks
  - Formal Tests
- Remedial Measures