Topic 20 - Diagnostics and Remedies

STAT 525 - Fall 2013

Outline

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

Overview

- General assumptions
  - Normally distributed error terms
  - Independent observations
  - Constant variance
- Will adopt or adapt diagnostics and remedial measures from linear 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
  - $\sum_j e_{ij} = 0 \, \forall i$

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

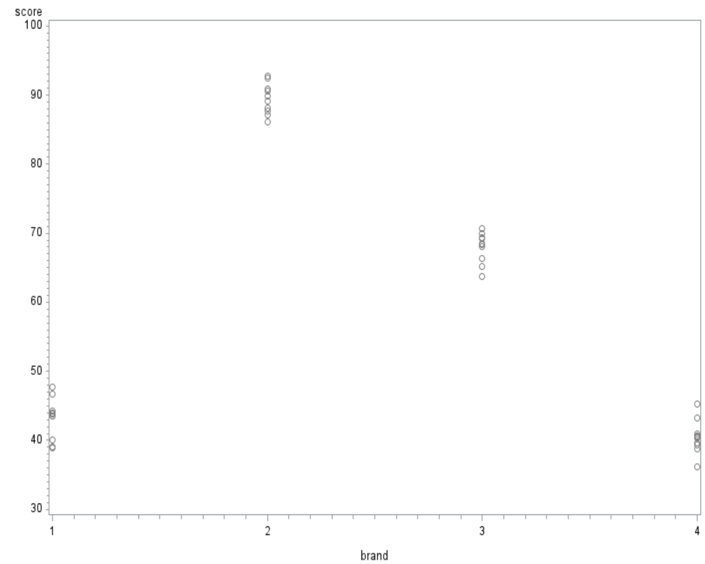
Example Page 777

- 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 (accelerated life study)
- Y coded score (higher means less rust)
- X brand of rust inhibitor
  - $i = 1, 2, 3, 4$
  - $j = 1, 2, .., 10$

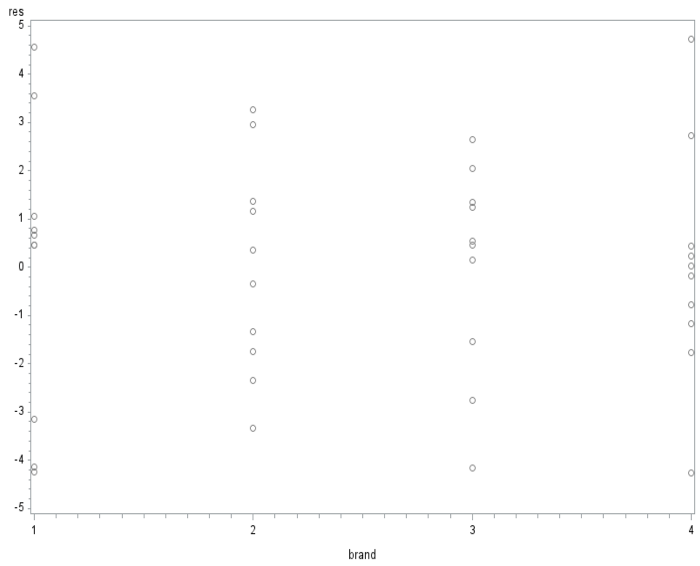
SAS Commands

```
data a1; infile 'u:\.www\datasets525\CH17TA02.txt';
   input score brand;

symbol1 v=circle i=none;
proc gplot;                               ***Scatterplot;
   plot score*brand;

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

proc gplot;                               ***Residual plots;
   plot res*(brand pred);

proc univariate noprint;                  ***Histogram and QQplot;
   histogram res / normal kernel(L=2);
   qqplot res / normal (L=1 mu=est sigma=est);
run;
```

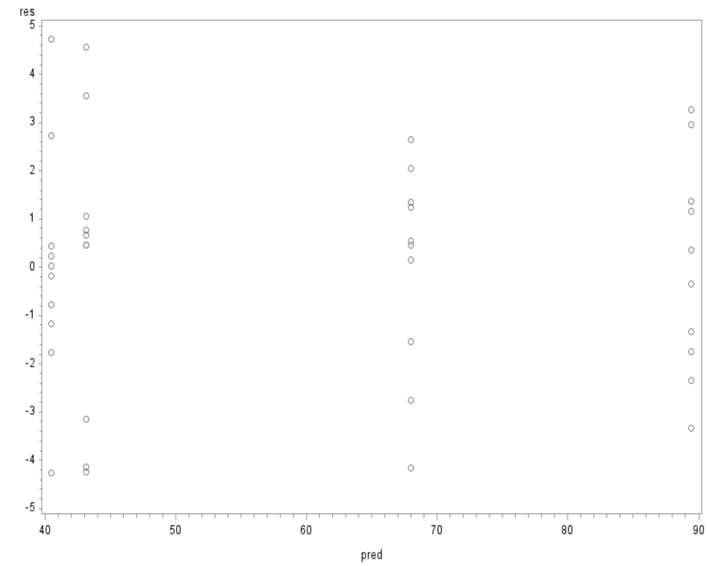

Scatterplot



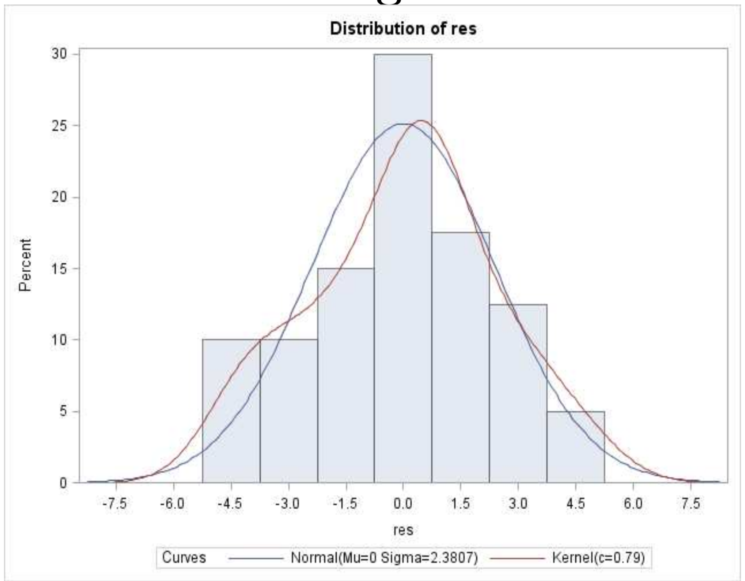
Residual Plot I



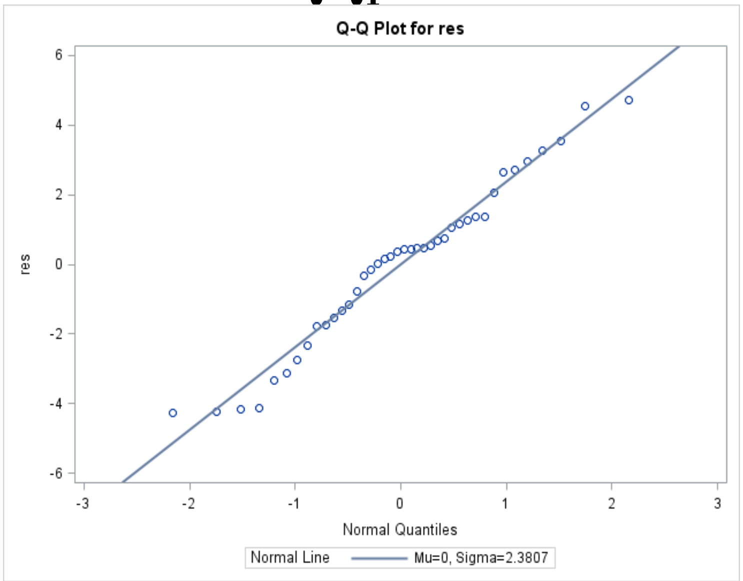
Residual Plot II



Histogram



QQplot



Summary

- Look for
  - Outliers
  - Non-constant variance
  - Non-normal errors
- Nothing too obvious in this example
- Can plot residuals vs time or other explanatory variable if they are available

Formal Tests

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

Homogeneity Tests

- Hartley statistic (Table B.10)

$$H = \frac{\max(\mathrm{s}_\mathrm{i}^2)}{\min(\mathrm{s}_\mathrm{i}^2)}$$

- Modified Levene - Brown-Forsythe
  - Same as regression approach
  - Groups are the factor levels
- Bartlett's test
  - Basically a likelihood ratio test

Homogeneity Tests

- Often caught in problem with assumptions
  - ANOVA is robust with respect to moderate deviations from normality
  - ANOVA results can be sensitive to homogeneity of variance, especially when there is unequal cell size
- Homogeneity tests are often sensitive to normality assumption
- Modified Levene's often best choice

Modified Levene's Test

- More robust against normality
- Considers the absolute deviation of each observation Y_{ij} about its factor level median $\tilde{Y}_i$

$$d_{ij} = \left|Y_{ij} - \tilde{Y}_i\right|$$

- Tests whether the expected value of these absolute deviations is equal across factor levels
- Simply performs ANOVA on these absolute deviations

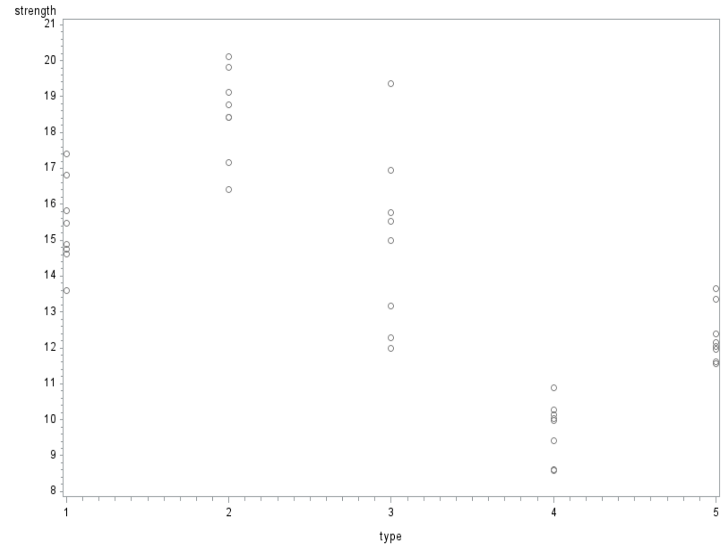
Example Page 783

- 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 flux type ($n_i = 8$)
- Y strength
- X type of flux

SAS Commands

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

proc gplot;                                ***scatterplot;
 plot strength*type;

proc glm;
 class type;
 model strength=type;
 means type / hovtest=bf;                  ***Modified Levene test;
 means type / hovtest=levene(type=abs);    ***Uses mean rather than median;
 lsmeans type / stderr cl;
```


Scatterplot



Output

```
Source     DF Sum of Squares   Mean Square  F Value  Pr > F
Model       4   353.6120850    88.4030213    41.93  <.0001
Error      35    73.7988250     2.1085379
Cor Total  39   427.4109100

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

            strength        Standard
type          LSMEAN           Error    Pr > |t|
1         15.4200000       0.5133880      <.0001
2         18.5275000       0.5133880      <.0001
3         15.0037500       0.5133880      <.0001
4          9.7412500       0.5133880      <.0001
5         12.3400000       0.5133880      <.0001
```


Remedies

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

SAS Commands

```
proc means data=a1;                            ***Obtain sample variances
   var strength;                                  for weights;
   by type;
   output out=a2 var=s2;
data a2; set a2; wt=1/s2;
data a3; merge a1 a2; by type;

proc glm data=a3;                              ***Weighted ANOVA;
   class type;
   model strength=type;
   weight wt;
   lsmeans type / stderr cl;

proc mixed data=a1;                            ***Mixed model with diff
   class type;                                    variances for all types;
   model strength=type / ddfm=kr;
   repeated / group=type;
```


GLM Output

```
                            Sum of
Source           DF        Squares   Mean Square  F Value  Pr > F
Model             4    324.2130988    81.0532747    81.05  <.0001
Error            35     35.0000000     1.0000000
Corrected Total  39    359.2130988

Least Squares Means
            strength         Standard
type          LSMEAN            Error     Pr > |t|      95% Confidence Limits
1         15.4200000        0.4373949       <.0001     14.532041     16.307959
2         18.5275000        0.4429921       <.0001     17.628178     19.426822
3         15.0037500        0.8791614       <.0001     13.218957     16.788543
4          9.7412500        0.2887129       <.0001      9.155132     10.327368
5         12.3400000        0.2720294       <.0001     11.787751     12.892249
```

MIXED Output

Cov Parm	Group	Estimate
Residual	type 1	1.5305
Residual	type 2	1.5699
Residual	type 3	6.1834
Residual	type 4	0.6668
Residual	type 5	0.5920

Fit Statistics

-2 Res Log Likelihood	122.1
AIC (smaller is better)	132.1
AICC (smaller is better)	134.2
BIC (smaller is better)	140.6

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
4	13.73	0.0082

MIXED Output

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
type	4	14.8	71.78	<.0001

Least Squares Means

Effect	type	Estimate	Standard Error	DF	t Value	Pr > |t|
type	1	15.4200	0.4374	7	35.25	<.0001
type	2	18.5275	0.4430	7	41.82	<.0001
type	3	15.0038	0.8792	7	17.07	<.0001
type	4	9.7413	0.2887	7	33.74	<.0001
type	5	12.3400	0.2720	7	45.36	<.0001

Summary

- GLM (weighted ANOVA) and MIXED analysis provide the same factor level estimates and standard errors.
- Without `ddfm=kr`, F test and df are also the same.
- With `ddfm=kr`, F test similar to Welch F test and factor level df more reasonable.
- Can consider groups of factor levels with similar variances
  - Group1=1 : Type 1 and 2
  - Group1=2 : Type 3
  - Group1=3 :Type 4 and 5

MIXED Output

Cov Parm	Group	Estimate
Residual	Group 1	1.5502
Residual	Group 2	6.1834
Residual	Group 3	0.6294

Fit Statistics

-2 Res Log Likelihood	122.1	
AIC (smaller is better)	128.1	
AICC (smaller is better)	128.9	***Much better fit
BIC (smaller is better)	133.2	

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
2	13.70	0.0011

MIXED Output

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
type	4	19.8	77.68	<.0001

Least Squares Means

Effect	type	Estimate	Standard Error	DF	t Value	Pr > |t|
type	1	15.4200	0.4402	14	35.03	<.0001
type	2	18.5275	0.4402	14	42.09	<.0001
type	3	15.0038	0.8792	7	17.07	<.0001
type	4	9.7413	0.2887	14	34.73	<.0001
type	5	12.3400	0.2720	14	43.99	<.0001

Delta Method

- Consider response X with $\mathrm{E}(X)=\mu_x$ and $\mathrm{Var}(X)=\sigma_x^2$
- Define $Y = f(X)$; What is the mean and var of Y?
- Consider the following Taylor series expansion

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

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

$$\mathrm{E}(Y)=\mathrm{E}(f(X))\approx \mathrm{E}(f(\mu_x)) + \mathrm{E}((X-\mu_x)f'(\mu_x))= f(\mu_x)$$

$$\mathrm{Var}(Y) \approx [f'(\mu_x)]^2\mathrm{Var(X)} = [\mathrm{f}'(\mu_\mathrm{x})]^2\sigma_\mathrm{x}^2$$

Transformations

- Suppose σ_x^2 depends on $\mu_x \rightarrow \sigma_x^2 = g(\mu_x)$
- Want to find $Y = f(X)$ such that $\mathrm{Var}(Y)\approx c$
- Have shown $\mathrm{Var}(f(X))\approx [f'(\mu_x)]^2\sigma_x^2$
- Want to choose f such that $[f'(\mu_x)]^2 g(\mu_x) \approx c$

Examples

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

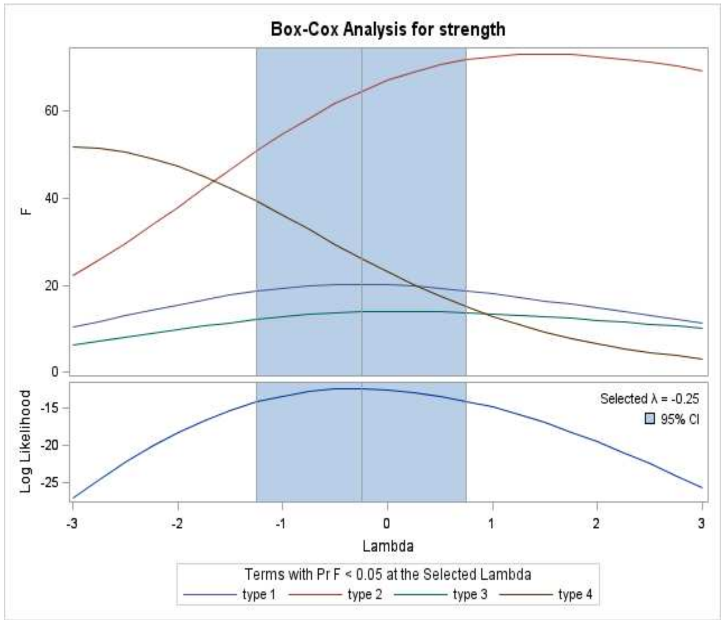
Transformation Guides

- Regress $\log(s_i)$ vs $\log(\overline{Y_{i.}}) \rightarrow \widehat{\lambda} = 1 - b_1$
- Use Proc TRANSREG

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

- When $\sigma_i^2 \propto \mu_i$ use $\sqrt{\quad}$
- When $\sigma_i \propto \mu_i$ use log
- When $\sigma_i \propto \mu_i^2$ use $1/Y$
- For proportions, use $\arcsin(\sqrt{\quad})$

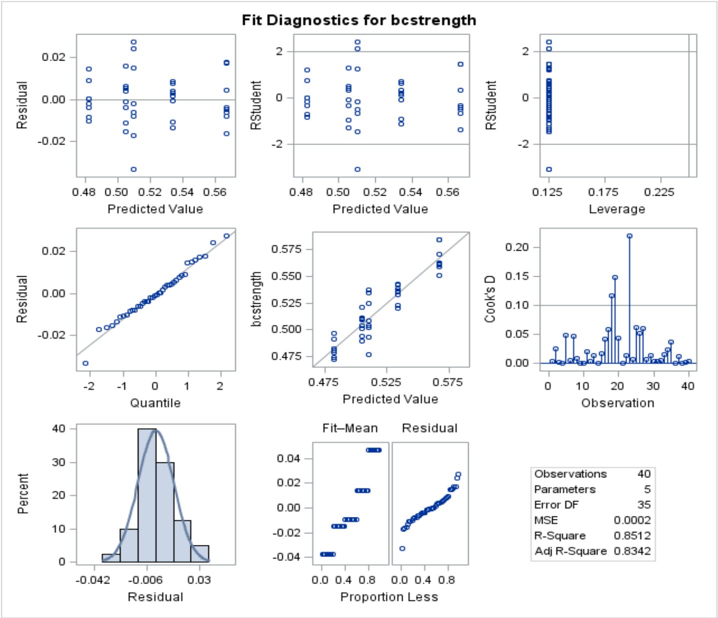
`arsin(sqrt(Y)) is SAS data step`



GLM Output - Box-Cox Transformation

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	0.03284060	0.00821015	50.05	<.0001
Error	35	0.00574192	0.00016405		
Corrected Total	39	0.03858252			

type	bcstrength LSMEAN	Standard Error	Pr > |t|	95% Confidence Limits	
1	0.50507700	0.00452845	<.0001	0.495884	0.514270
2	0.48230819	0.00452845	<.0001	0.473115	0.491501
3	0.50998508	0.00452845	<.0001	0.500792	0.519178
4	0.56659809	0.00452845	<.0001	0.557405	0.575791
5	0.53382040	0.00452845	<.0001	0.524627	0.543014



Nonparametric Approach

- Based on ranking the observations and using the ranks for inference
- SAS procedure NPAR1WAY
- Could also perform permutation test
  - Under H_0 all observations from population with same mean
  - Any observation could be assigned to any group
  - Permute observations numerous times
  - For each permutation compute test statistic
  - Compare observed statistic with this distribution

SAS Commands

```
proc npar1way data=a1;
 var strength;
 class type;

*Approximate approach

proc rank;
 var strength;
 ranks strength1;

proc glm;
 class type;
 model strength1=type;
run;
```


Output

Wilcoxon Scores (Rank Sums)

type	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean 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

Kruskal-Wallis Test

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

Output

Median Scores (Number of Points Above Median)

type	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean 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

Median One-Way Analysis

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

Background Reading

- KNNL Chapter 18
- knnl777.sas, knnl783.sas
- KNNL Chapter 19