Statistics 514: Design of Experiments

Topic 7a

Topic Overview

This topic will cover

- Assumptions
- Diagnostics
- Remedies

All models are wrong; some models are useful

Rubin's Commandments (Expurgated Version)

1. Thou shalt know that thou must make assumptions.
2. Thou shalt know thy assumptions.
3. Thou shalt not believe thy assumptions.
4. Thou shalt know the deviations and consequences of deviation from thy assumptions.

Can we check our assumptions?

Owen's Categorical Imperative: The answer to all worthwhile statistical questions is "It depends..."

Model Assumptions

$$y_{i,j} = \mu + \tau_i + e_{i,j}, \begin{cases} i = 1, \ldots, a \\ j = 1, \ldots, n_i \end{cases}$$

Assumptions

Model-based:

- Grand mean μ is constant for all (i, j).
- Effects add (instead of $y_{i,j} = \mu\tau_i e_{i,j}$).
- Error term has expected value 0. ($E(e_{i,j}) = 0$.)
    - Impossible to check for 1-factor model.
    - Even if wrong, get predictive and hypothesis-testing power.

Errors (second order):

- Independent observations
- Constant variance
- Normally distributed

  Consequences of deviation:

  - Predictions of treatment effects are fine (unbiased).
  - Can cause untrustworthy inference (false optimism).
  - Use residuals (biased downward for estimating true error):

$$e_{i,j} \geq r_{i,j} = \hat{e}_{i,j} = y_{i,j} - \hat{\tau}_i - \hat{\mu}$$

Model Adequacy and Diagnostics

Visual Inspections and Tests

- Normality
  - Histogram of residuals
  - Normal probability plot/QQ plot
  - Shapiro-Wilks/Kolmogorov-Smirnov Test
- Variance
  - Plot $\hat{\epsilon}_{i,j}$ vs $\hat{y}_{i,j}$ (residual plot)
  - Bartlett's or Levene's Test
- Independence
  - Plot $\hat{\epsilon}_{i,j}$ vs time/space.
  - Plot $\hat{\epsilon}_{i,j}$ vs variable of interest.
- Outliers
  - Is it influential? With and without analysis (depends on where point is in design)
  - Formal tests (e.g., standardized residuals)
  - Investigate why the result may occur; don't try to eliminate

Tests vs Diagnostics

- What does the p-value of a test tell us about the validity of ANOVA inference?
- Usual answer: if validity is in doubt, run another test and compare results.

Independence

- Plot of the residuals over time
  - Is there a drift or pattern as trials proceed?
  - Durbin-Watson statistic:

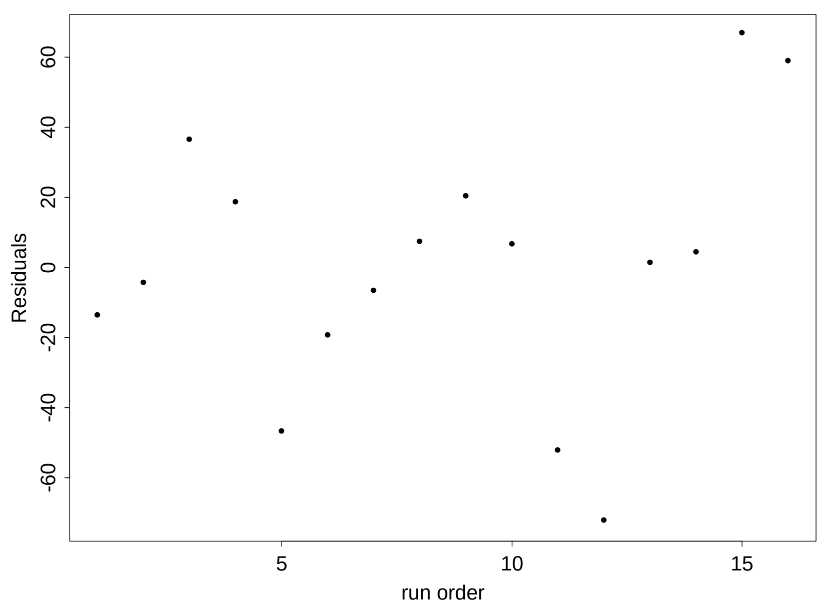
$$DW = \frac{\sum_{k=1}^{n-1}(r_k - r_{k+1})^2}{\sum_{k=1}^{n} r_k^2}.$$

    If $DW < 2$, then evidence for positive serial dependence.

- Spatial association
  - Variogram: plot $dist_{spatial}(x_i, x_j)$ vs $(r_i - r_j)^2$
  - If increasing (might require smoothing), then there's association.

Is There Serial Correlation?

$$DW = 1.42$$



Usual Fix

Plot residuals versus relevant variables

- Often variables omitted from analysis (lucky if not confounded)

- Experimental conditions (e.g., temp)
- May result in inclusion of factor in next experiment
- Easier to record and include factors in larger model and eliminate them if they don't seem relevant.

Constant Variance

- Often experiments with non-constant variance
- Why concern?
  - Comparison of treatments depends on MSE
  - Incorrect intervals and comparison results
- Does not affect F-test dramatically (balanced)
  - big n_i assigned to group with big variance $\Rightarrow$ p-value $>$ true p (pessimism)
  - big n_i $\rightarrow$ small variance $\Rightarrow$ p-value $<$ true p (optimism)
- **Symptom:** Size of residual associated with predicted value
- Residual plot
  - Plot $\hat{e}_{i,j}$ vs $\hat{y}_{i,j}$ (mean at factor level i).
  - Is the range constant for different levels of $\hat{y}_{i,j}$? Is there a pattern?

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

$$\begin{aligned} q &= (N-a)\log_{10} S_p^2 - \sum_{i=1}^{a}(n_i - 1)\log_{10} S_i^2 \\ c &= 1 + \frac{1}{3(a-1)}\left(\sum_{i=1}^{a}(n_i-1)^{-1} - (N-a)^{-1}\right). \end{aligned}$$

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 $\chi_0^2 > \chi^2_{\alpha,a-1}$.

Remark: very sensitive to normality assumption

Levene's Test

- Compute $d_{i,j} = |y_{i,j} - m_i|$, where m_i is median for $i = 1, 2, \ldots, a$, and $j = 1, 2, \ldots, n_i$.
- Compare samples using usual ANOVA F-test.

```
options ls=80 ps=65;

title1 'Diagnostics Example';

/* Read in the data set */
data one;
 infile 'h:\System\Desktop\tensile.dat';
 input percent strength time;

/* Perform ANOVA, test equal variances (hovtest=), store residuals and
   predicted values in a SAS data set called diag */
proc glm data=one;
 class percent;
 model strength=percent;
 means percent / hovtest=bartlett hovtest=levene;
 output out=diag p=pred r=res;

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

/* Generate a QQ-plot and histogram */
proc univariate data=diag noprint;
 var res;  qqplot res / normal (L=1 mu=est sigma=est);
 histogram res / normal; run;
run;

/* Plot residuals vs time to see if there is a pattern.  Will
   use different smoothing levels */
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;
```

The GLM Procedure

Dependent Variable: strength

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			

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

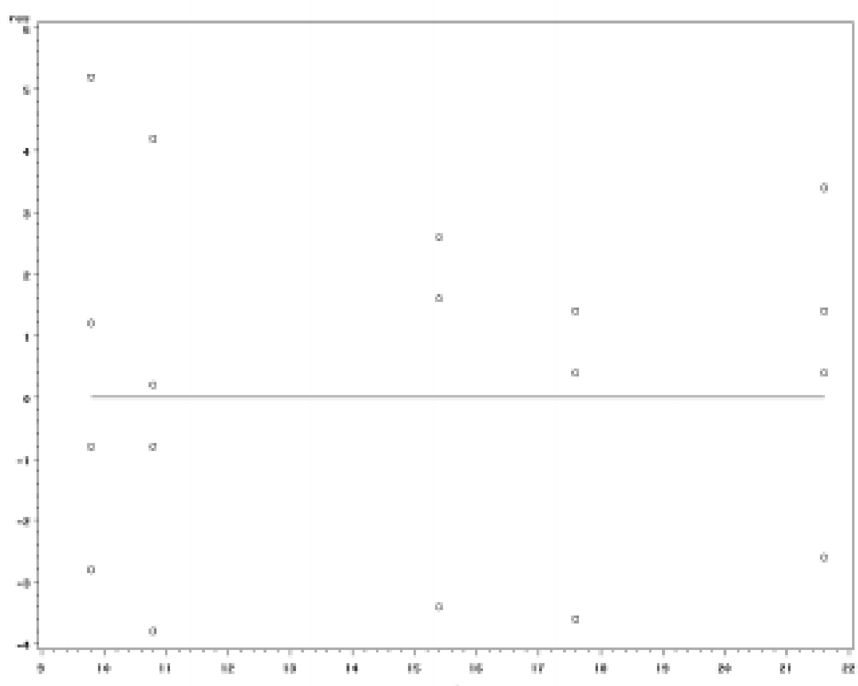
Levene's Test for Homogeneity of strength Variance
ANOVA of Squared Deviations from Group Means

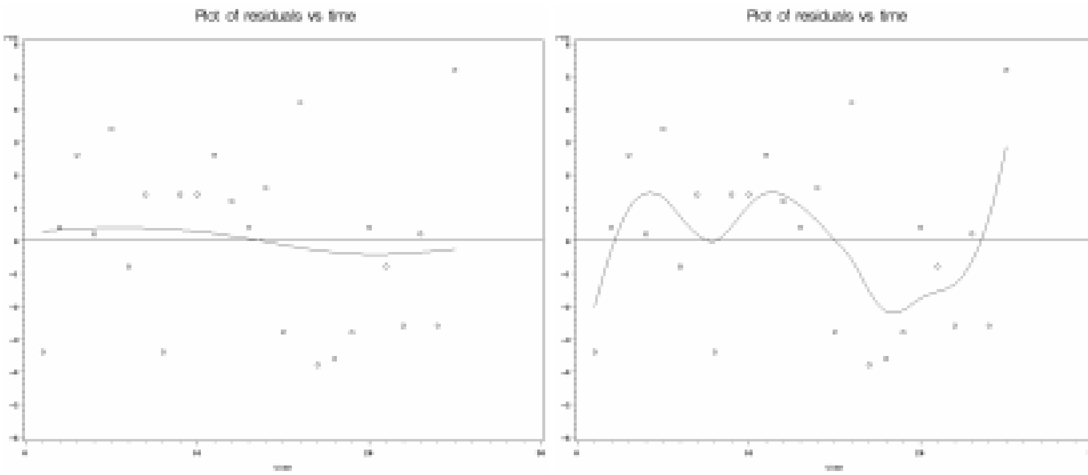
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
percent	4	91.6224	22.9056	0.45	0.7704
Error	20	1015.4	50.7720		

Bartlett's Test for Homogeneity of strength Variance

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

Residual Plot





Remedies

"Variance-Stabilizing" Transformations

- Consider response Y with $\mathrm{E}(Y) = \mu_x$ and $\mathrm{Var}(Y) = \sigma_x^2$.
- Define $Z = f(Y)$. What is the mean and variance of Z?

Delta Method

$$\text{Consider } f(Y), \text{ where } f'(\mu_x) \neq 0.$$
$$f(Y) \approx f(\mu_x) + (Y - \mu_x) f'(\mu_x)$$

$$\begin{aligned} \mathrm{E}(Z) = \mathrm{E}(f(Y)) &\approx \mathrm{E}(f(\mu_x)) + \mathrm{E}((Y - \mu_x) f'(\mu_x)) = f(\mu_x) \\ \mathrm{Var}(Z) &\approx [f'(\mu_x)]^2 \mathrm{Var}(Y) = [f'(\mu_x)]^2 \sigma_x^2 \end{aligned}$$

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

Examples

$g(\mu)$	Distribution
μ	Poisson(λ) ($\mathrm{Var}(Y) = \mathrm{E}(Y) = \lambda$)
$\mu(1-\mu)$	Binomial(n, p) ($\hat{p} = Y/n$; $\mathrm{Var}(\hat{p}) = p(1-p)/n$)
$\mu^{2\beta}$	(Box-Cox)
μ^2	(Box-Cox)

Transformation	New Variance
$f(y) = \int \frac{1}{\sqrt{\mu}} d\mu \rightarrow f(Y) = \sqrt{Y}$	1/(4n)
$f(y) = \int \frac{1}{\sqrt{\mu(1-\mu)}} d\mu \rightarrow f(Y) = \arcsin(\sqrt{Y})$	$\frac{1}{4}$
$f(y) = \int \mu^{-\beta} d\mu \rightarrow f(Y) = Y^{1-\beta}$	
$f(y) = \int \frac{1}{\mu} d\mu \rightarrow f(Y) = \log(Y)$	

Box-Cox Transformations

- Perform analysis of variance on

$$y^{\lambda} = \begin{cases} \frac{y^{\lambda}-1}{\lambda \dot{y}^{\lambda-1}} & \lambda \neq 0 \\ \dot{y}\log(y) & \lambda = 0 \end{cases},$$

  where $\dot{y}$ is the geometric mean of the observations

$$\dot{y} = \left(\prod_{i=1}^{a} \prod_{j=1}^{n_i} y_{i,j} \right)^{1/N}$$

- y to the power of λ rescaled for direct comparison
- Find λ which minimizes SS_E.

Transformations?

- On transformed data, not actually comparing means of data.
- However, instead of

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_a$$

  can think of randomization null hypothesis as

$$H_0 : F_1 = F_2 = \ldots = F_a,$$

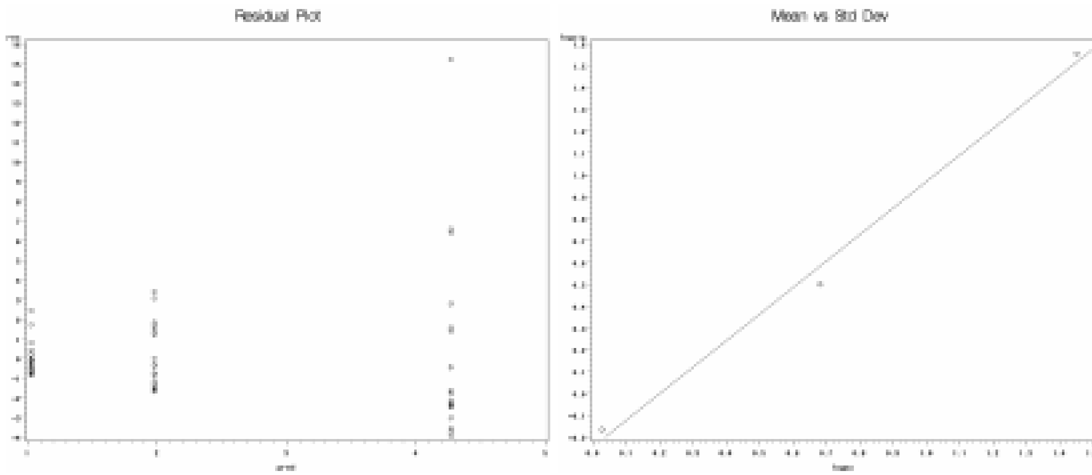
  where $F_1, \ldots, F_a$ are the distributions of the a factor levels.
- Comparing means is a first-order test of this equivalence. (Could compare medians, variances, skewnesses...)
- By comparing transformed means, maximize our ability to discriminate.
- Can fix outlier/normality problem

trans.sas

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

title1 'Increasing Variance Example';

data one;
 infile 'h:\System\Desktop\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;

/* This computes the sample mean (mu) and standard deviation (sigma) at
   each treatment level and outputs this information to SAS file "two" */
proc univariate data=one noprint;
 var resp; by trt;
 output out=two mean=mu std=sigma;

/* This creates a new SAS data set "three" which contains the natural
   logs of both the sample means and standard deviations */
data three;
 set two;
 logmu = log(mu);
 logsig = log(sigma);

/* Perform a regression analysis on these log transformed values */
proc reg;
 model logsig = logmu;

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

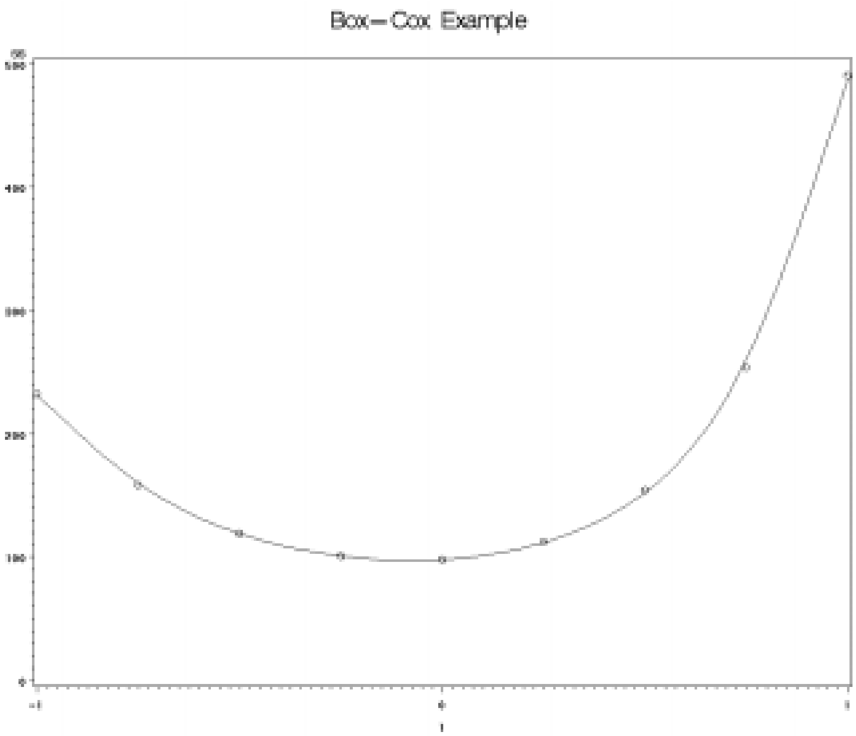
trans1.sas

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




```
/* Read in data set and take log of each response.  We will use the
   average of the log values to compute the geometric mean */
data one;
 infile 'h:\System\Desktop\boxcox.dat';
 input trt resp;
 logresp = log(resp);

/* This is computing the average of the log values.*/
proc univariate data=one noprint;
 var logresp; output out=two mean=mlogresp;

/* By taking the antilog of the mean log value, we get the geometric mean.
   This then runs through values of lambda ranging between -1 and 1 (with
   a special computation when l=0) computing the "transformed" value of
   each response. */
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 l eq 0 then den = 1;
    yl=(resp**l -1)/den;   if l=0 then yl=ydot*log(resp);
    output;
 end;
 keep trt yl l;

/* We now run GLM for each value of l outputing certain summary info
   into a data set called "four".  All we really need is the value of
   l and the residual SS */
proc sort data=three out=three;  by l;
proc glm data=three noprint outstat=four;
 class trt;  model yl=trt;  by l;
```

```
/* This throws everything else away except for l and the residual SS */
data five;  set four;
 if _SOURCE_ eq 'ERROR';  keep l SS;

symbol1 v=circle i=sm50;
proc gplot;
 plot SS*l;
run;
```



Unusual Observations

- Can use residual plot to detect unusual observations
- Unusual = large $|\hat{e}_{i,j}|$.
- Sometimes typographical error (double-entry data entry)
- Otherwise worth investigation
  - Helpful to save detailed lab notes
- Formal test to see if unusual
- But does not answer if in error

- **Influential:** Does observation make a difference if excluded?
- Perform analysis with and without observation(s).

Near-Zero/Truncated Values

- Can have heavily skewed observations (near zero)
  - Concentration of rare contaminant
  - Number of defects in assembly line
  - Number of birds at a given site
- Measurements may be truncated
  - Concentration of rare contaminant (not detectable)
  - Lifetime of component (does not fail)

Kruskal-Wallis Test: a nonparametric alternative

a treatments H_0: a treatments are identical.

- Rank the observations $y_{i,j}$ in ascending order
- Replace each observation by its rank $R_{i,j}$ (assign average for tied observations).
- Test statistics

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

  where

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

- Decision Rule: reject H_0 if $H > \chi^2_{\alpha,a-1}$.

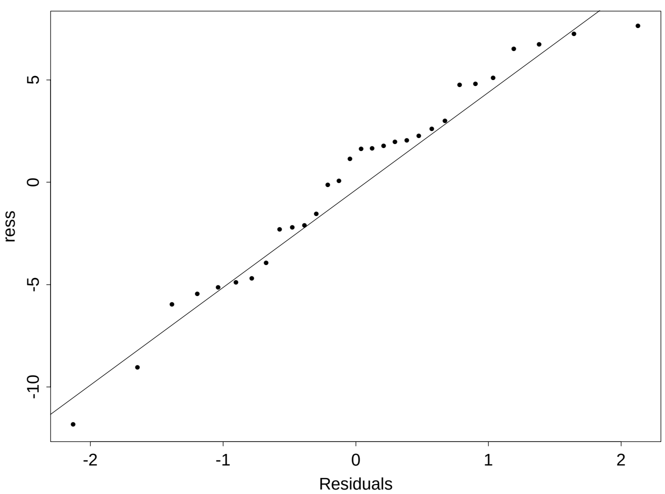
Relation to F_0

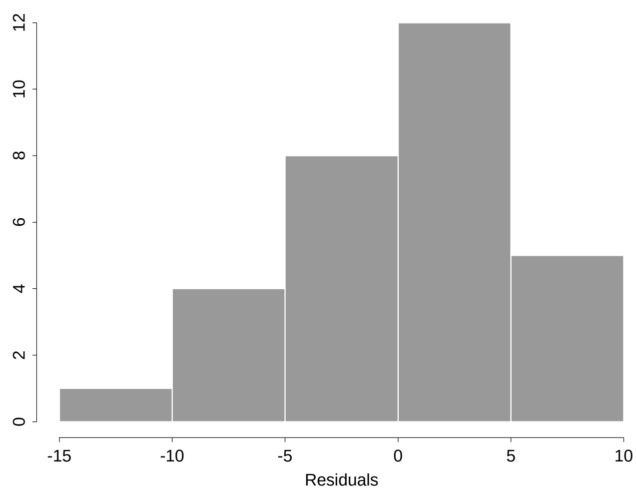
$$F_0 = \frac{H/(a-1)}{(N-1-H)/(N-a)}$$

Same as performing F-test on ranks.

nonpar.sas

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

data new;
 input strain nitrogen @@;
 cards;
  1 19.4 1 32.6 1 27.0 1 32.1 1 33.0
  2 17.7 2 24.8 2 27.9 2 25.2 2 24.3
  3 17.0 3 19.4 3  9.1 3 11.9 3 15.8
  4 20.7 4 21.0 4 20.5 4 18.8 4 18.6
  5 14.3 5 14.4 5 11.8 5 11.6 5 14.2
  6 17.3 6 19.4 6 19.1 6 16.9 6 20.8
 ;

proc npar1way;
 class strain;
 var nitrogen;

run;
```





Multiple R-Squared: 0.3795

F-statistic: 17.12 on 1 and 28 degrees of freedom, the p-value is 0.0002899

```
The NPAR1WAY Procedure

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

                          Sum of      Expected       Std Dev          Mean
strain        N           Scores      Under H0      Under H0         Score
----------------------------------------------------------------------------
1             5            130.0         77.50     17.961885         26.00
```

2	5	111.0	77.50	17.961885	22.20
3	5	40.0	77.50	17.961885	8.00
4	5	88.0	77.50	17.961885	17.60
5	5	23.0	77.50	17.961885	4.60
6	5	73.0	77.50	17.961885	14.60

```
                  Average scores were used for ties.
   Kruskal-Wallis Test

Chi-Square         21.6593
DF                       5
Pr > Chi-Square     0.0006

 Median Scores (Number of Points Above Median) for Variable nitrogen
                    Classified by Variable strain
```

strain	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
1	5	5.0	2.50	1.038068	1.00
2	5	4.0	2.50	1.038068	0.80
3	5	1.0	2.50	1.038068	0.20
4	5	3.0	2.50	1.038068	0.60
5	5	0.0	2.50	1.038068	0.00
6	5	2.0	2.50	1.038068	0.40

```
                  Average scores were used for ties.

 Median One-Way Analysis

Chi-Square         13.5333
DF                       5
Pr > Chi-Square     0.0189
```

Adjusted Degrees of Freedom

If null is true, but variances different in different groups, then

$$F_0/b \sim F(\nu_1, \nu_2),$$

where

$$\begin{aligned} b &= \frac{N-a}{N(a-1)} \frac{\sum_i (N-n_i)\sigma_i^2}{\sum_i (n_i-1)\sigma_i^2} \\ \nu_1 &= \frac{[\sum_i (N-n_i)\sigma_i^2]^2}{[\sum_i n_i\sigma_i^2]^2 + N\sum_i (N-2n_i)\sigma_i^4} \\ \nu_2 &= \frac{[\sum_i (n_i-1)\sigma_i^2]^2}{\sum_i (n_i-1)\sigma_i^4} \end{aligned}$$

Brown-Forsythe Method

Analogue of t-test with non-equal variances

$$BF = \frac{\sum_{i=1}^{a} n_i(\bar{y}_{i.} - \bar{y}_{..})^2}{\sum_{i=1}^{a} s_i^2(1 - \frac{n_i}{N})} \sim F_{a-1,\nu},$$

where

$$\begin{aligned} d_i &= s_i^2\left(1 - \frac{n_i}{N}\right) \\ \nu &= \frac{\sum d_i^2}{\sum d_i^2/(n_i - 1)} \\ s_i^2 &= \text{sample variance in treatment } i \end{aligned}$$

Normality

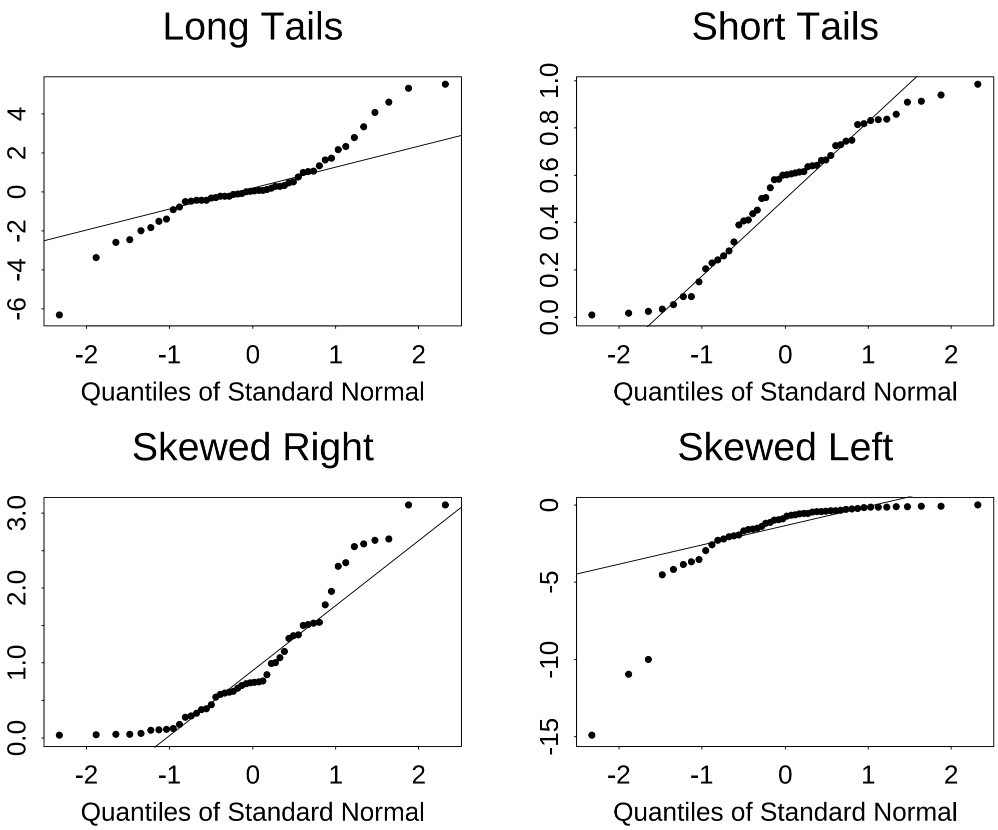
- If everything else is fine, least serious problem to have.
- Most common deviations:
  - Skewness
  - Kurtosis (heavy tails)
  - Truncated values: points to more serious missing data problem

Assessment of residuals: Histogram (bell-shaped?), QQplot, compare CDF's

- Plot sorted residuals r_i against $\frac{i-\frac{3}{8}}{n+\frac{1}{4}}$ quantiles of normal distribution (could use expectation of ith order statistic from sample of size n from normal distribution).

Testing: Shapiro-Wilk (preferred), Kolmogorov-Smirnov, etc.

Diagnosing Problems



tensile.sas

```
options ls=75 ps=60 nocenter;
goptions colors=(none) device=win target=winprtm rotate=landscape ftext=swiss
   hsize=8.0in vsize=6.0in htext=1.5 htitle=1.5 hpos=60 vpos=60
   horigin=0.5in vorigin=0.5in;

data one;
 infile 'h:\System\Desktop\tensile.dat';
 input percent strength time;

title1 ' example';
proc print data=one; run;

proc glm;
 class percent; model strength=percent;
 output out=oneres p=pred r=res; run;

proc univariate data=oneres pctldef=4;
 var res;  qqplot res / normal (L=1 mu=est sigma=est);
 histogram res / normal; run;
```

```
The UNIVARIATE Procedure
```

Fitted Distribution for res

Parameters for Normal Distribution

Parameter	Symbol	Estimate
Mean	Mu	0
Std Dev	Sigma	2.591653

Goodness-of-Fit Tests for Normal Distribution

Test	---Statistic----		-----p Value-----	
Kolmogorov-Smirnov	D	0.16212279	Pr > D	0.088
Cramer-von Mises	W-Sq	0.08045523	Pr > W-Sq	0.203
Anderson-Darling	A-Sq	0.51857191	Pr > A-Sq	0.177