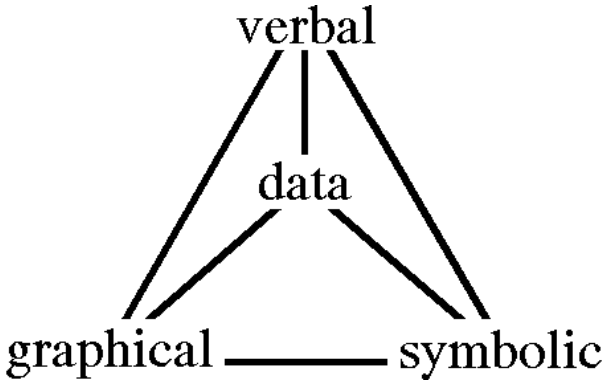
Handouts in Quantitative Biology

D. C. Schneider
Memorial University of Newfoundland
St. John's

September 2000



Part I Units and Dimensions

Part II. The General Linear Model.

PART III

Part I Units and Dimensions

Table 1. Base and supplementary units in the SI system.

Quantity	Unit	Abbreviation
Length	metre	m
Mass	kilogram	kg
Time	second	s
Thermodynamic temperature	kelvin	K
Amount of substance	mole	mol
Luminous intensity	candela	cd
Electrical current	ampere	A
Planar angle	radian	rad
Solid angle	steradian	sr

Table 2. Standard multiples of ratio scale units.

Name	Multiple	Abbreviation	Example
pico	10^{-12}	p	pW
nano	10^{-9}	n	nW
micro	10^{-6}	:	:W
milli	10^{-3}	m	mW
centi	10^{-2}	c	cW
deci	10^{-1}	d	dW
	10^{0}		W
deca	10^{1}	da	daW
hecto	10^{2}	h	hW
kilo	10^{3}	k	kW
mega	10^{6}	M	MW
giga	10^{9}	G	GW

Table 3. Units that commonly occur in biology.

Quantity		Unit Name	Unit Symbol	Equivalent Units
Acceleration	angular			$rad\,s^{-2}$
	linear			$m\,s^{-2}$
Area		square metre	m^2	
		hectare	ha	$10^4\,m^2$
Concentration				$mol\,m^{-3}$
Energy (work)		joule	J	$N\,m$
		kilocalorie	kcal	$4185\,J$
Energy flux				$J\,m^{-2}\,s^{-1}$
Force		newton	N	$kg\,m\,s^{-2}$
Frequency		hertz	Hz	s^{-1}
Light	Luminance			$cd\,m^{-2}$
	Luminous flux	lumen	lm	$cd\,sr$
	Illuminance	lux	lx	$lm\,m^{-2}$
		footcandle	fc	$10.764\,lx$
	Photon flux	einstein	E	$1\,mole$
Mass density				$kg\,m^{-1}$
Mass flow				$kg\,s^{-1}$
Mass flux				$kg\,m^{-2}\,s^{-1}$
Power		watt	W	$J\,s^{-1}$
Pressure (stress)		pascal	Pa	$N\,m^{-2}$
Surface tension				$N\,m^{-1}$
Velocity	angular			$rad\,s^{-1}$
	linear			$m\,s^{-1}$
Viscosity	dynamic			$Pa\,s$
	kinematic			$m^2\,s^{-1}$
Volume		cubic metre	m^3	
		litre	l	$10^{-3}\,m^3$
Volume flow rate				$m^3\,s^{-1}$
Wavelength				m
Wavenumber				m^{-1}

Table 4. Rules for working with dimensions.
From D.S. Riggs (1963) *The Mathematical Approach to Physiological Problems*. MIT Press.

1. All terms in equation must have the same dimensions.
   Terms separated by + ! or = .
2. Multiplication and division must be consistent with rule 1.
3. Dimensions are independent of magnitude.
   dx/dt is the ratio of infinitesimals,
   but still has dimensions of x/t = Length/Time.
4. Pure numbers (e, B) have no dimensions.
   Exponents and percentages have no dimensions.
5. Multiplication by a dimensionless number does not change dimensions.

Working with Dimensions--Examples.

1. According to Holligan et al 1984 (*Marine Ecology Progress Series* 17:201) the vertical flux of nutrients through the ocean's thermocline is:

$$F_N = K_V \;)N /)Z$$

were F_N is the vertical flux of nutrients (milligram-atoms $m^{!2}\ s^{!1}$)
K_V is the vertical eddy diffusivity ($10^{!4}\ m^2\ s^{!1}$)
)N is the nitrate difference across the thermocline (mg-atoms)
)Z is the thickness of the thermocline (metres)

Write out dimensions beneath each symbol in the equation.
Is this equation dimensionally homogeneous? _____

Work out the dimensions of)N required to make the equation homogeneous _____

Work out the units of)N required to make the equation homogeneous _____

M = Mass
$M\ L^{!1}$ = mass gradient
$M\ L^{!2}$ = mass density
$M\ L^{!3}$ = mass concentration

Based on this,)N must be the difference in nitrate _____ across the thermocline.

2. A series of experimental measurements by Holligan *et al* suggest that the vertical flux of nutrients through the thermocline follows an exponential relation:

$$F_N = "(K_V \partial N / \partial Z)^{3/4}$$

What units does " have? _____

What dimensions does " have? _____

3. Another series of experiments by Holligan *et al* suggest that nutrient flux depends upon the temperature gradient across the thermocline.

$$F_N = \$ (\partial T/\partial Z)^{!1/3}$$

$$\partial T/\partial Z = {}^{o}C/\text{metre}$$

What units does $ have? _____

What dimensions does $ have? _____

Elementary statistics courses for biologists tend to lead to the use of a stereotyped set of tests:
1 without critical attention to the underlying model involved;
2 without due regard to the precise distribution of sampling errors;
3 with little concern for the scale of measurement;
4 careless of dimensional homogeneity;
5 without considering the ideal transformation;
6 without any attempt at model simplification;
7 with too much emphasis on hypothesis testing and too little emphasis on parameter estimation.

M.J. Crawley. 1993. *GLIM for Ecologists*. (London, Blackwell)

Euclidean and Fractal Dimensions in Biology -- References

Gunther, B. 1975. Dimensional analysis and the theory of biological similarity. *Physiological Reviews* 55: 659-698.

Hastings, H. M. and G. Sugihara. 1993. *Fractals: a User's Guide for the Natural Sciences.* Cambridge University Press.

Mandelbrot, B.B. 1977. *Fractals: Form, Chance, and Dimension.* San Francisco: Freeman.

Pennycuick, C.J. *Newton Rules Biology: A Physical Approach to Biological Problems.* Oxford University Press.

Platt, T.R. and W. Silvert. 1981. Ecology, physiology, allometry, and dimensionality. *Journal of Theoretical Biology* 93: 855-860.

Schneider, D.C. 1994. *Quantitative Ecology: Spatial and Temporal Scaling.* San Diego: Academic Press.

Stahl, W.R. 1961, 1962. Dimensional analysis in mathematical biology. *Bulletin of Mathematical Biophysics* 23: 355-376, 24: 81-108.

Sugihara, G., B. Grenfell, and R.M. May. 1990. Applications of fractals in ecology. *Trends in Resereach in Ecology and Evolution*. 5: 79-87.

<short, highly readable account, including how to estimate km^d>

West, B.J. and A.L. Goldberger. 1987. Physiology in fractal dimensions. *American Scientist* 75: 351-365.

Part II. The General Linear Model.

Notation for Frequency Distributions and Probability Functions.

There is no standard notation for the 4 forms of a frequency distribution--the notation for probability functions will vary from text to text. The probability function gives the expected frequency (for a population), while the frequency distribution gives the observed frequency (for a sample). The probability function can thus be considered a model for the frequency distribution obtained from data.

Here is a notational convention that distinguishes the 4 ways that a single frequency distribution or probability function can be expressed. The observed or empirical form is tabulated from data or by repeated subsampling. The expected form is calculated from a mathematical expression.

	Observed (sample)	Expected (population)
Frequency	$F(Q = k)$	$E(F(Q = k)$ Expected value of frequency that $Q = k$
Relative Frequency	$P(Q = k)$	$E(P(Q = k)$ Expected value of frequency that $Q = k$, as a proportion
Cumulative Frequency	$F(Q \leq k)$	$E(F(Q \leq k)$ Expected value of cumulative frequency of Q less than or equal to k
Cumulative Relative Frequency	$P(Q \leq k)$	$E(P(Q \leq k)$ Expected value of the frequency of $Q \leq k$, as a proportion

Other names for $E(P(Q = k)$: probability mass function, probability density function, pdf for short.

Other notation for $E(P(Q = k)$: $Pr(Q = k)$ $Pr(X = x)$ $Pr(X = x)$

Other names for $E(P(Q \leq k)$: cumulative distribution function, cdf for short.

Other notation for $E(P(Q \leq k)$: $Pr(Q \leq k)$ $F(x)$ $P(X \leq x)$ $Pr(X \leq x)$

Table 5. Key for choosing the frequency distribution of a statistic.

Statistic is the population mean	
If data are normal or cluster around a central value	
If sample is large (n > 30)	Normal distribution
If sample is small (n < 30	t distribution
If data are Poisson	Poisson distribution
If data are Binomial	Binomial distribution
If data do not cluster around central value, examine residuals (deviations from the mean)	
If residuals are normal or cluster around a central value	
If sample is large (n > 30)	Normal distribution
If sample is small (n < 30)	t distribution
If residuals are not normal	Empirical (bootstrap)
Statistic is the population variance	
If data are normal or cluster around a central value	Chi-square
If data do not cluster around central value	
If sample is large (n > 30)	Chi-square
If sample is small (n < 30	Empirical (bootstrap)
Statistic is the ratio of two variances (ANOVA tables)	
If data are normal or cluster around a central value	F-distribution
If data do not cluster around a central value, calculate residuals	
If residuals are normal or cluster around a central value	F-distribution
If residuals do not cluster around central values	
If sample is large (n > 30)	F-distribution
If sample is small (n < 30)	Empirical
Statistic is none of the above	
Search statistical literature for appropriate distribution or confer with statistician	
If not in literature or cannot be found	Empirical

Empirical distributions are generated by taking all permutations, by sampling permutations, or by subsampling (bootstrap methods).

Table 6. Generic recipe for calculating a confidence limit.

1. State population; state the statistic of interest.
2. Calculate an estimate of the statistic from data
3. Determine the distribution of the estimate.
4. State tolerance for Type I error.
5. Write a probability statement about the estimate or statistic.
6. Plug values into the statement to obtain confidence limits.
7. Make a statement about the probability that the line (or limits) include the true value.
   This statement is not about the statistic or estimate.

Strangely, the motto chosen by the founders of the Statistical Society in 1834 was *Aliis exterendum*, which means "Let others thrash it out." William Cochran confessed that "it is a little embarrassing that statisticians started out by proclaiming what they will not do."
E. A. Gehan and N. A. Lemak. 1995. *Statistics in Medical Research: Developments in Clinical Trials* (Plenum Press).

Fisher's famous paper of 1922, which quantified information almost half a century ago, may be taken as the fountainhead from which developed a flow of statistical papers, soon to become a flood. This flood, as most floods, contains flotsam much of which, unfortunately, has come to rest in many text books. Everyone will have his own pet assortment of flotsam; mine include most of the theory of significance testing, including multiple comparison tests, and non parametric statistics.

John Nelder, Rothamsted Experimental Station. (Fisher's successor as Director of the Statistics Department, and pioneer of generalised linear models). From: *Mathematical Models in Ecology*, British Ecological Society Symposium 1971.

Table 7. Generic recipe for decision making with statistics.

1. State population, conditions for taking sample.
2. State the model or measure of pattern . . . . . ST
3. State Null Hypothesis about the population . . . . . H_o
4. State Alternative Hypothesis . . . . . H_a
5. State criterion (tolerance) for Type I error . . . . . α
6. State frequency distribution that gives probability of outcomes when the Null Hypothesis is true. Choices are:
   Permutations, i.e. distribution of all possible outcomes when H_o is true;
   Empirical distribution obtained by random sampling of all possible outcomes when H_o is true;
   Cumulative distribution function (cdf) that applies when H_o is true;
   State assumptions when using a cdf such as normal, F, t, or chisquare.
7. Calculate the statistic. This is the observed outcome.
8. Calculate the p-value for the observed outcome relative to distribution of outcomes when H_o is true.
9. If p less than α then reject H_o and accept H_a
   If p greater than α then accept H_o.
10. Report statistic, p-value, sample size.
    Declare decision.

Equivalent method (less informative) based on just a table, no computer

8. Calculate outcome corresponding to α
9. If observed outcome > outcome @ α then reject H_o, accept H_a.
   If observed outcome # outcome @ α then accept H_o.
10. Report statistic, p-value, and sample size. Declare decision.

This latter method is less informative, because the observed p-value does not get reported. This method was made necessary by the cumbersome tables for frequency distribution. With modern computers it is possible to calculate an exact p-value for any statistic (e.g. Minitab, or programs in hand-held calculators). The method of reporting an exact p-value is preferred to the method based on tables.

Table 8 Generic Recipe for decision making with the General Linear Model.

1. State population, and conditions for taking sample.
2. Construct the model: state the response variable;
   state the explanatory variable(s);
   state type of measurement scale for each of these;
   write model relating response to explanatory variables.
3. State H_A about parameters (means and slopes) of the model
4. State H_o about parameters.
5. State tolerance for type I error as criterion for significance, α.
6. Estimate parameters from data.
7. Use parameters to calculate residuals; plot against explanatory variable.
   If bowls or arches are evident, revise the form of the model (back to step 2)
8. Calculate the variance explained by the model.
   Partition both Var(Response variable) and df according to model
   Table SS, df, MS, F (by computer usually).
9. Calculate Type I error (the p-value) from density function (F or t distribution).
10. Check assumptions for use of p-value from density function.
    residuals homogeneous ? (residual versus fit plot)
    residuals normal ? (histogram of residuals, quantile or normal score plot)
    residuals independent ? (plot residuals versus residuals at lag 1)
11. If: assumptions are not met, sample small ($n < 30$), and p near α
    then compute acceptable p-value from empirical distribution of F-statistic,
    using randomization methods.
12. Declare decision: If $p < \alpha$ then reject H_o and accept H_A
    If $p \geq \alpha$ then accept H_o and reject H_A
13. Report F-ratio, p-value (not α), df.
    Declare decision.

This is a modification of the Generic Recipe for Hypothesis testing.
The pattern is stated as an equation; the summary statistic is the F-ratio.
The equation links one or more response variables to one or more explanatory variables, via parameters (means and slopes).
This equation is used to set up the ANOVA table, to partition the degrees of freedom, and to partition the total sum of squares: $SS_{total} = (n-1) * Var(Y) = (n-1) * s^2$

For reports, state in the methods section the critical value α;
state that the residuals were examined for normality, homogeneity, and independence;
state that randomization methods were used to compute Type I error, if assumptions were not met and sample size was small.

Table 9. Commonly used tests, based on the General Linear Model.

Analysis	Response Variable	Explanatory Variable	Interaction?	Comments
t-test	1 ratio	1 nominal	Absent	compares two means
1-way ANOVA	1 ratio	1 nominal	Absent	compares 3 or more means in 1 category
2-way ANOVA	1 ratio	2 nominal	Present	tests for interactive effects
				compares means in 2 categories, if no interaction
Paired Comparison	1 ratio	2 nominal	Assumed Absent	compares 2 means in 1 category, controlled for 2nd category (blocks or units)
Randomized Blocks	1 ratio	2 nominal	Assumed Absent	compares 3 or more means in 1 category, controlled for 2nd category (blocks or sampling units)
Hierarchical ANOVA	1 ratio	$2 nominal	Absent	nested comparisons of means
ANCOVA	1 ratio	$ 1 ratio $ 1 nominal	Present	compares two or more slopes
			Absent	compares means, controlled for slopes
Regression	1 ratio	1 ratio	Absent	tests linear relation of response to explanatory
Multiple Regression	1 ratio	$ ratio	Assumed Absent	tests linear relation to 2 explanatory variables relation expressed as a plane

GLM: One-Way ANOVA (srbx9_1.out)

Scutum width data from Box 9.1 in Sokal and Rohlf (1995), page 210.
Width of scutum (in ：m) of tick larvae *Haemaphysalis leporispalustris* in samples taken from each of 4 hosts (rabbits).

Begin by reading data and labelling variables.

```
MTB > read 'a:srbx9_1.dat' c1 c2;
SUBC> nobs = 37.
     37 ROWS READ

 ROW     C1   C2

   1    380    1
   2    376    1
   3    360    1
   4    368    1
  .   .   .

MTB > name c1 'width' c2 'host'
```

Write a general linear model relating the response variable to the explanatory variable.

Width = $\$_o$ + $\$_X X$ + ,
grand mean host deviations residuals

Width = Host means + residuals

There are 4 host means,
each equaling the grand mean plus 1 of 4 host deviations ($\$_o + \$_x$).

Next, estimate the parameters, $\$_o$ (1 value) and $\$_x$ (4 values).

```
MTB > describe 'width'

                N     MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
width          37   359.70   360.00   359.61    12.46     2.05

MTB > describe 'width' by 'host'

             host        N     MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
width           1        8   372.25   373.00   372.25     7.36     2.60
                2       10   354.40   353.00   353.75    11.92     3.77
                3       13   355.31   354.00   355.00     8.92     2.47
                4        6   361.33   366.00   361.33    15.27     6.23
```

Print out data equations:
Data(width) = fits(hosts) + residuals

```
MTB > name c3 'fits' c4 'res'
MTB > print 'width' 'fits' 'res'

 ROW  width      fits       res                fits  =   b.o    +  b.x  X

   1    380   372.250    7.7500                      380 = 372.250 + 7.7
                                                 = 359.703 + 12.5473 + 7.75
   2    376   372.250    3.7500
   3    360   372.250  -12.2500
   4    368   372.250   -4.2500
   5    372   372.250   -0.2500
   6    366   372.250   -6.2500
   7    374   372.250    1.7500
   8    382   372.250    9.7500
   9    350   354.400   -4.4000                       350 = 354.4 ! 4.4
                                                  = 359.703 -  5.3027 ! 4.4
  10    356   354.400    1.6000
  11    358   354.400    3.6000
  12    376   354.400   21.6000
  13    338   354.400  -16.4000
  14    342   354.400  -12.4000
  15    366   354.400   11.6000
  16    350   354.400   -4.4000
  17    344   354.400  -10.4000
  18    364   354.400    9.6000
  19    354   355.308   -1.3077          354 = 359.703 - 4.3950 - 1.3077
  20    360   355.308    4.6923
  21    362   355.308    6.6923
  22    352   355.308   -3.3077
  23    366   355.308   10.6923
  24    372   355.308   16.6923
  25    362   355.308    6.6923
  26    344   355.308  -11.3077
  27    342   355.308  -13.3077
  28    358   355.308    2.6923
  29    351   355.308   -4.3077
  30    348   355.308   -7.3077
  31    348   355.308   -7.3077
  32    376   361.333   14.6667             376 = 359.703 + 1.63 + 14.67
  33    344   361.333  -17.3333
  34    342   361.333  -19.3333
  35    372   361.333   10.6667
  36    374   361.333   12.6667
  37    360   361.333   -1.3333
```

Based on the model written for this data, execute an analysis of variance. In Minitab, use either the ANOVA or GLM command.

```
MTB > anova 'width' = 'host';
SUBC> fits c3;
SUBC> residuals c4.

Factor     Type Levels Values
host      fixed      4     1     2     3     4

Analysis of Variance for width

Source      DF         SS         MS       F      P
host         3       1808      602.6    5.26  0.004
Error       33       3778      114.5
Total       36       5586      155.2
```

Use residuals to check assumptions.

A. Structural model acceptable ? No need to check for bowls and arches, because model does not contain a slope (straight line).

B1. Sum(errors) = 0 ?

Sum should be zero because means were used with no transformations. Check this.

```
MTB > describe 'res'

                N     MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
res            37     0.00    -0.25    -0.05    10.24     1.68

              MIN      MAX       Q1       Q3
res        -19.33    21.60    -7.31     8.68
```

B2. Errors independent ?   Check for association between neighbouring values.
Plot each value against that of its neighbour.
If no association, then plot will resemble shotgun blast.
If errors are not independent, then points will tend downward from upper left,
or upward from lower left.

```
MTB > let c5 = lag('res')
MTB > name c5 'res_lag1'
MTB > plot c5 c4

res_lag1-
        -         *
        -                                            *
      15+       *
        -                        *  *                        *    *
        -                        *  *      *
        -              *          *            *
        -             *          *         *      *              *
       0+                      *            *
        -               *   *         * *          *
        -                   *           *                   *
        -                                         *
        -          *             *        *          *
     -15+             *
        -     *                                    *
        -
        -
          --------+---------+---------+---------+---------+--------res
              -16.0      -8.0       0.0       8.0      16.0

        N* = 1
```

No evidence of association between neighbouring value.

The assumption of independent errors can also be examined with a formal test.

```
MTB > runs 'res'
     res
    K =     0.0000

    THE OBSERVED NO. OF RUNS =  20
    THE EXPECTED NO. OF RUNS =  19.4865
    18 OBSERVATIONS ABOVE K   19 BELOW
              THE TEST IS SIGNIFICANT AT  0.8640
              CANNOT REJECT AT ALPHA = 0.05
```

Are there too many or too few runs of positive or negative values in the sequence ?
No, there are not an improbably large number of runs in the data.
This confirms graphical analysis, that residuals are independent. There is no evidence that residuals are associated, in order of presentation.

B3. Var(errors) = fixed value ? (i.e. homogeneous across groups?)
To test this assumption, plot residual versus fitted values.

```
MTB > plot 'res' 'fits'

res      -
         -    *
         -      *
       15+                      *
         -    * *                2
         -    *                                                  2
         -      3
         -    2 *                                                2
        0+      *                *                                 *
         -    2 2                                                 *
         -      2                                                 *
         -    *
         -    * 2                                                 *
      -15+    *
         -                       2
         -
         -
           +---------+---------+---------+---------+---------+------fits
        353.5     357.0     360.5     364.0     367.5     371.0
```

Plot shows similar vertical distribution in all 4 groups, i.e., no cones from left to right or right to left. var(errors) is similar across groups.

B4. Normally distributed errors ?

```
MTB > hist 'res'
  Histogram of res   N = 37

 Midpoint   Count
     -20       1  *
     -15       3  ***
     -10       4  ****
      -5       8  ********
       0       5  *****
       5       6  ******
      10       6  ******
      15       3  ***
      20       1  *
```

Plot shows residuals are normally distributed.

This assumption is tested last.
It is the least important assumption because it has the least effect on accurate calculation of the p-value, and hence least effect on accurate estimation of Type I error.

GLM: Regression (srbx1412.out)

Egg mass data from Box 14.12 in Sokal and Rohlf (1995), p 546.

Mass (to nearest 100 gram) of unspawned female cabezon fish *(Scorpaenichthys marmoratus*) and number of eggs (thousands) produced.

```
MTB > read 'a:srbx1412.dat' c1 c2;
SUBC> nobs = 11.
     11 ROWS READ

 ROW    C1     C2

   1    14     61
   2    17     37
   3    24     65
  .   .   .

MTB > name c1 'Wt' c2 'Eggs'
MTB > regress c2 on 1 c1

The regression equation is
Eggs = 19.8 + 1.87 Wt

Predictor       Coef       Stdev    t-ratio        p
Constant       19.77       10.55       1.87    0.094
Wt            1.8700      0.3325       5.62    0.000

s = 10.15       R-sq = 77.8%     R-sq(adj) = 75.4%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    1      3260.9      3260.9     31.63    0.000
Error         9       927.9       103.1
Total        10      4188.7

MTB > let c3 = 19.77 + 1.87*'Wt'
MTB > let c4 = 'Eggs' - c3
MTB > name c3 'fits' c4 'res'
MTB > print 'Eggs' 'fits' 'res'

 ROW   Eggs    fits       res                          Here are the data equations.

   1     61   45.95   15.0500
   2     37   51.56  -14.5600
   3     65   64.65    0.3500
   4     69   66.52    2.4800
   5     54   70.26  -16.2600
   6     93   81.48   11.5200
   7     87   83.35    3.6500
   8     89   88.96    0.0400
   9    100   94.57    5.4300
  10     90   96.44   -6.4400
  11     97   98.31   -1.3100
```

Compute sums of squares SS from data equations.

```
MTB > describe 'Eggs' 'fits' 'res'

                N      MEAN    MEDIAN    TRMEAN     STDEV    SEMEAN
Eggs           11     76.55     87.00     78.33     20.47      6.17
fits           11     76.55     81.48     77.53     18.06      5.44
res            11     -0.00      0.35      0.13      9.63      2.90

              MIN       MAX        Q1        Q3
Eggs        37.00    100.00     61.00     93.00
fits        45.95     98.31     64.65     94.57
res        -16.26     15.05     -6.44      5.43

MTB > let k1 = 10*stdev('Eggs')**2
MTB > let k2 = 10*stdev('fits')**2
MTB > let k3 = 10*stdev('res')**2
MTB > print k1 k2 k3
K1       4188.73
K2       3261.02
K3       927.866
```

Compare these to computations in ANOVA table.

Use residuals to check assumptions.
A. Linear relation of response to explanatory variable ?

```
MTB > plot c4 c3
        -
        -
res     -   *
      12+                                  *
        -
        -
        -                                    *           *
        -                    *
       0+                  *                       *
        -                                                  *
        -                                                *
        -
     -12+
        -     *
        -                  *
        -
          ------+---------+---------+---------+---------+---------+fits
               50        60        70        80        90       100
```

No bowls or arches, so linear model acceptable.

B1. Sum(errors) = 0 ? Yes, because method of least squares was used to estimate slope.

B2. Errors independent ?

```
MTB > let c6 = lag('res')
MTB > plot 'res' c6
          -    *
        10+
          -
res       -                                   *
          -                                                 *
          -                                    *
         0+         *                               *
          -                    *
          -
          -                                        *
          -
       -10+
          -
          -                                                      *
          -                                   *
          -
       -20+
            +---------+---------+---------+---------+---------+------C6
         -18.0      -12.0      -6.0        0.0        6.0       12.0

           N* = 1
```

Points scattered, no evidence of trends, neighbouring residuals appear independent.

B3. Var(errors) fixed ? Re-examine plot of residuals vs fits for change in scatter.
No evidence of increased scatter in plot.
(no cones, going left to right or right to left).

B4. Residuals normally distributed ?

```
MTB > root c4
 BIN    COUNT     RAWRS      DRRS        SUSPENDED ROOTGRAM

   1      0.0      -0.0     -0.01          .          -          .
   2      2.0       1.9      2.03          .           ++++++++++
   3      0.0      -0.6     -0.80          .      -----           .
   4      1.0      -1.1     -0.63          .       ----           .
   5      4.0       0.3      0.25          .           ++         .
   6      2.0      -1.1     -0.49          .        ---           .
   7      1.0      -0.2      0.04          .           +          .
   8      1.0       0.8      1.08          .           ++++++     .
   9      0.0      -0.0     -0.04          .          -           .

 IN DISPLAY, VALUE OF ONE CHARACTER IS .2       OO
```

```
MTB > hist c4                                          Are the residuals normally distributed?
 Histogram of res   N = 11
Midpoint   Count
     -15       2  **
     -10       0
      -5       1  *
       0       4  ****
       5       2  **
      10       1  *
      15       1  *

MTB > let c5 = nscores(c4)
MTB > plot c5 c4
          -                                                                  *
c5        -
          -
     1.0+                                                        *
          -                                              *
          -
          -                                         *
          -                                    *
     0.0+                                  *
          -                               *
          -                            *
          -
          -                    *
    -1.0+       *
          -
          -                                                  Not a straight line.
          -    *
            +---------+---------+---------+---------+---------+---------+------res
      -18.0     -12.0      -6.0       0.0       6.0      12.0
```

Errors not normal and sample size small (n = 11), so randomization should be used if p-value needed to be defended. However, the estimate of Type I error ($p < 0.001$) is so far from the tradition criterion " = 5% that better estimate from randomization will not change conclusion, that egg mass production is related to body mass.

GLM: Two-way ANOVA (srbx11_2.out).

Oxygen consumption data from Box 11.2 in Sokal and Rohlf (1995), page 332.

Oxygen consumption (∶l O_2 (mg body wt)$^{!1}$ min$^{!1}$) by two species of limpet at three salinities.

```
MTB > read 'a:srbx11_2.dat' c1-c3;
SUBC> nobs=48.
     48 ROWS READ

 ROW      C1   C2     C3

   1    7.16    1    100
   2    6.78    1    100
   3   13.60    1    100
   4    8.93    1    100
  .  .  .
MTB > name c1 'oxy'  c2 'sp'   c3 'sal'
MTB > anova 'oxy' = 'sp' 'sal' 'sp'*'sal';
SUBC> fits c4;
SUBC> residuals c5.

Factor     Type Levels Values
sp        fixed      2     1     2
sal       fixed      3    50    75   100

Analysis of Variance for oxy

Source      DF         SS         MS       F      P
sp           1      16.64     16.638    1.74  0.194
sal          2     181.32     90.661    9.48  0.000
sp*sal       2      23.93     11.963    1.25  0.297
Error       42     401.52      9.560
Total       47     623.41     13.264

MTB > name c4 'fits'  c5 'res'

MTB > print c1-c5

 ROW     oxy   sp    sal      fits       res

   1    7.16    1    100   10.5612  -3.40125
   2    6.78    1    100   10.5612  -3.78125
   3   13.60    1    100   10.5612   3.03875
   4    8.93    1    100   10.5612  -1.63125
   5    8.26    1    100   10.5612  -2.30125
   6   14.00    1    100   10.5612   3.43875
   7   16.10    1    100   10.5612   5.53875
   8    9.66    1    100   10.5612  -0.90125
   9    5.20    1     75    7.8900  -2.69000
  10    5.20    1     75    7.8900  -2.69000
```

Calculate the Sums of Squares for ANOVA table, using Minitab.

```
MTB > describe 'oxy';
SUBC> by 'sp'.
               sp        N     MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
oxy             1       24   10.208    9.740   10.045    3.493    0.713
                2       24    9.031    9.765    8.872    3.765    0.769

MTB > describe 'oxy';
SUBC> by 'sal'.
              sal        N     MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
oxy            50       16   12.250   11.455   12.201    3.200    0.800
               75       16    7.614    6.835    7.439    2.678    0.669
              100       16    8.995    8.595    8.854    3.473    0.868

MTB > set into c8
DATA>(8.995  7.6138  12.25)16                          The means for each salinity
MTB > end

MTB > set into c9
DATA>(10.2083  9.0308)24                               The means for each species
MTB > end

MTB > describe c1-c5 c8 c9
                N     MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
oxy            48    9.620    9.740    9.475    3.642    0.526
sp             48   1.5000   1.5000   1.5000   0.5053   0.0729
sal            48    75.00    75.00    75.00    20.63     2.98
fits           48    9.620    9.226    9.600    2.173    0.314
res            48    0.000   -0.983   -0.055    2.923    0.422
C8             48    9.620    8.995    9.591    1.964    0.283
C9             48   9.6195   9.6195   9.6195   0.5950   0.0859

MTB > let k1 = stdev('oxy')*stdev('oxy')*47                          Sums of Squares
MTB > let k2 = stdev('sp')*stdev('sp')*47
MTB > let k3 = stdev('sal')*stdev('sal')*47
MTB > let k4 = stdev('fits')*stdev('fits')*47
MTB > let k5 = stdev('res')*stdev('res')*47
MTB > let k8 = stdev(c8)*stdev(c8)*47
MTB > let k9 = stdev(c9)*stdev(c9)*47
MTB > print k1-k5 k8 k9
K1       623.407       SS total        df = 47
K2       12.0000
K3       20000.0
K4       221.885       SS model        df =  5
K5       401.521       SS res          df = 42
K8       181.318       SS salinity     df =  2
K9       16.6381       SS species      df =  1
```

A Linearity assumption. No need to check, no straight lines in model..
B1. Errors sum to zero ? Yes, because parameters were estimated by least squares.
B2. Errors independent ? Yes, based on graphical display (no pattern).

```
MTB > let c8 = lag('res')
 MTB > plot c8 'res'
    C8   -
         -                  *
         -                        *
      5.0+                           * *
         -                  *  *
         -                 *  ***  *                    *    *
         -                   *    *            *
         -                                     **
      0.0+                        *
         -              *  ** * * *  *           *
         -                 * * * * *             2   *   *  2     *
         -   *         *   *       *             *
         -                                         2
     -5.0+
         -                                  *
         -
         -
           ------+---------+---------+---------+---------+---------+res
              -5.0       -2.5        0.0        2.5        5.0        7.5
      N* = 1
 MTB > corr c7 c8
 Correlation of res and C8 = -0.148
```

B3 Residuals homogeneous ? Yes

```
 MTB > plot 'res' 'fits'
   res   -                                                      *
         -                                        *
      5.0+          *                                          *
         -    **
         -    22    *                             2
         -                                                    *2
         -          *                                          *
      0.0+                                                    **
         -    *2    2                             *           2
         -    32    *                             2           3*
         -    *     2                             *            *
         -     *                                  *
     -5.0+
         -                                                     *
         -
         -
           +---------+---------+---------+---------+---------+------fits
          7.0       8.0       9.0      10.0      11.0      12.0
```

Residuals from an ANOVA model can be plotted only at limited number of places along the X-axis. So pattern is judged by imagining that residuals have been erased between these locations along the X-axis.

B3. Errors are homogeneous. Residuals do not show cone-shaped pattern, with increasing spread going from left to right or from right to left.

B4. Errors normal ?

```
MTB > hist 'res'
Histogram of res   N = 48
Midpoint   Count
      -6       1  *
      -5       0
      -4       2  **
      -3       5  *****
      -2      12  ************
      -1       8  ********
       0       2  **
       1       2  **
       2       3  ***
       3       7  *******
       4       2  **
       5       2  **
       6       1  *
       7       1  *
```

Residuals look bimodal, rather than normal.
Compare observed distribution of residuals to normal distribution, using a rootogram with 95% confidence limits.

```
MTB > root 'res'
BIN   COUNT    RAWRS     DRRS       SUSPENDED ROOTGRAM
  1     0.0     -1.4    -1.58        .  --------          .
  2     1.0     -0.2     0.07        .          +         .
  3     0.0     -1.8    -1.89        .----------          .
  4     2.0     -0.7    -0.25        .         --         .
  5     5.0      1.4     0.79        .           ++++     .
  6    12.0      7.6     2.76        .           ++++++++++++++
  7     8.0      3.0     1.23        .           +++++++   .
  8     2.0     -3.3    -1.57        .  --------          .
  9     2.0     -3.2    -1.52        .  --------          .
 10     3.0     -1.7    -0.73        .      ----          .
 11     7.0      3.0     1.37        .          +++++++   .
 12     2.0     -1.1    -0.49        .       ---          .
 13     2.0     -0.2     0.03        .          +         .
 14     1.0     -0.5    -0.17        .          -         .
 15     1.0      0.1     0.31        .          ++        .
 16     0.0     -1.0    -1.24        .   -------          .

 IN DISPLAY, VALUE OF ONE CHARACTER IS .2       O
```

Bimodality of residuals again evident

Compute and plot normal equivalent deviates (nscores) to evaluate normality of residuals.

```
MTB > nscores 'res' c7
MTB > plot 'res' c7
          -
res       -                                                          *
          -                                                      *
       5.0+                                                  * *
          -                                              **
          -                                         *2 2 **     <--not normal
          -                                      2*
          -                                    **
          -                                   **
       0.0+                                 2*2*2
          -                          2****2*2*          <-----not normal
          -                  * ** 2
          -          *  *
      -5.0+
          - *
          -
             --------+---------+---------+---------+---------+--------C7
                 -1.60     -0.80      0.00      0.80      1.60
```

All three graphical analyses show bimodality in residuals.

The residuals are strongly bimodal. Is this because the data were bimodal?

```
MTB > hist 'oxy'
Histogram of oxy   N = 48
Midpoint   Count
       4       3  ***
       6      12  ************
       8       6  ******
      10      11  ***********
      12       6  ******
      14       7  *******
      16       1  *
      18       2  **

MTB > root 'oxy'
 BIN    COUNT     RAWRS      DRRS        SUSPENDED ROOTGRAM

   1      0.0      -3.3     -2.77      --------------           .
   2      3.0      -1.1     -0.41         .       ---           .
   3     12.0       5.5      1.89         .          ++++++++++.
   4      6.0      -2.3     -0.76         .      ----           .
   5     11.0       2.3      0.80         .          ++++       .
   6      6.0      -1.4     -0.42         .       ---           .
   7      7.0       1.9      0.87         .          +++++      .
   8      1.0      -1.8     -1.05         .    ------           .
   9      2.0       0.7      0.70         .          ++++       .
  10      0.0      -0.6     -0.89         .     -----           .

  IN DISPLAY, VALUE OF ONE CHARACTER IS .2       OO
```

The data are not as strongly bimodal as the residuals. Bimodality becomes more evident after the effects of species and salinity have been removed. There is evidence that some additional factor is operating. Such a source of heterogeneity should be included in the analysis. However, there is no way to revise the model to include an additional factor, because this is a text example without enough information to do this.

The residuals do not look normal:

- the histogram looks bimodal
- the rootogram shows deviation from normal distribution
- the normal scores do not fall on a straight line.

The p-value from a theoretical F-distribution cannot be trusted.

A randomization test is in order, to obtain better estimate of p-value.

Randomize data and execute ANOVA repeatedly to obtain better estimate of p-value.

```
MTB > sample 48 'oxy' c7;
MTB > anova c7 = 'sp' 'sal' 'sp'*'sal'

Factor     Type Levels Values
sp        fixed      2     1     2
sal       fixed      3    50    75   100

Analysis of Variance for C7

Source      DF         SS         MS       F      P
sp           1      6.586      6.586    0.45  0.504
sal          2      3.545      1.772    0.12  0.885
sp*sal       2      5.315      2.657    0.18  0.833
Error       42    607.961     14.475
Total       47    623.407     13.264

MTB > stack c10 .18 c10                    accumulate random F_sp x sal
MTB > stack c11 .45 c11                    accumulate random F_sp
MTB > stack c12 .12 c12                    accumulate random F_sal
```

Compare this random partitioning of the variance in O_2 consumption to the partitioning based on the observations before randomization. The "explained" variance after randomization has, as expected, dropped.
It has dropped from (16.64 + 181.32 + 23.93 = 221.89) to (6.586 + 3.545 + 5.315 = 15.4).
It has dropped from 36% to 2.5% (= 15.4/623.407).

Repeat the analysis on another randomization of the response variable Y.
Continue to accumulate the random F-ratios in columns c10 c11 and c12.

Now compute the p-values:

```
MTB > hist c10;
SUBC> start 1.251.                                  * this is the observed value of the F-ratio

Histogram of C10   N = 21
16 Obs. below the first class
Midpoint   Count
     1.3       0
     1.5       0
     1.7       3  ***
     1.9       1  *
     2.1       0
     2.3       0
     2.5       0
     2.7       0
     2.9       0
     3.1       1  *                    so   p =  5/21  =  0.24 no significant interaction
```

```
MTB > hist c11;                                     Compute p-value for species term.
SUBC> start 1.740.

Histogram of C11   N = 21
18 Obs. below the first class
Midpoint   Count
     1.7       2  **
     2.1       0
     2.5       0
     2.9       0
     3.3       0
     3.7       0
     4.1       0
     4.5       1  *                                so  p = 3/21
```

p = 3/21 = 0.14 Therefore conclude no significant species effect.
However, this p-value is based on few randomizations; several hundred would be needed to obtain good estimate of p-value (Type I error).

```
MTB > hist c12;                                     Compute p-value for salinity effect.
SUBC> start 9.483.
 Histogram of C12   N = 21
21 Obs. below the first class                       so   p < 1/21
```

p < 1/21 hence p < 0.0476

Small p-value suggests that salinity has significant effect on oxygen consumption, however, more randomizations would be needed to obtain good estimate of p-value (Type I error).

GLM: Randomized Blocks (srbx11_4.out)

Genotype data from Box 11.4 in Sokal and Rohlf (1995), page 351.

```
MTB > read 'srbx11_4.dat' c1 c2 c3;
SUBC> nobs = 12
MTB > name c1 'weight'  c2 'blocks' c3 'gtype'
MTB > print c1-c3

 ROW  weight  blocks  gtype

   1   0.958       1      1
   2   0.971       2      1
   3   0.927       3      1
   4   0.971       4      1
   5   0.986       1      2
   6   1.051       2      2
   7   0.891       3      2
   8   1.010       4      2
   9   0.925       1      3
  10   0.952       2      3
  11   0.829       3      3
  12   0.955       4      3
```

Test variation in weight among genotypes, ignoring blocks.

```
MTB > anova 'weights' = 'gtype';
SUBC> fits c4;
SUBC> residuals c5.

Factor     Type Levels Values
gtype     fixed      3     1     2     3

Analysis of Variance for weights

Source      DF         SS         MS       F      P
gtype        2   0.009717   0.004859    1.71  0.235
Error        9   0.025575   0.002842
Total       11   0.035292   0.003208
```

Error MSE is 0.002842

To obtain more sensitive test, include effects of blocks

Error MSE reduced to 0.000697

Error MSE reduced by factor of 0.002842/0.000697
= 4.1
= 410 %

```
MTB > anova 'weight' = 'blocks' 'gtype';
SUBC> fits c4;
SUBC> residuals c5.

Factor     Type Levels Values
blocks    fixed      4     1     2     3     4
gtype     fixed      3     1     2     3

Analysis of Variance for weight

Source      DF         SS         MS        F      P
blocks       3   0.021391   0.007130   10.23  0.009
gtype        2   0.009717   0.004859    6.97  0.027
Error        6   0.004183   0.000697
Total       11   0.035292   0.003208
```

Use residuals to check assumptions.
A. Structural model acceptable ? Regression variables not included in model, so no need to check this assumption (no need to check for bowls or arches).

B1. Sum(res) = 0 ? This assumption listed for completeness, but usually no need to check because sum will be zero when least squares methods is used, as in most statistical packages.

B2. Errors independent ? Check for association between neighbouring values.
Plot each residual against neighbouring value.

```
MTB > let c10 = lag('res')
MTB > plot 'res' c10
          -
res       -  *
          -
          -
     0.025+                      *
          -
          -          *
          -
          -                       *
     0.000+    *                          *
          -               *
          -                                                    *
          -                      *
          -
    -0.025+                      *                          *
          -
          -
            --+---------+---------+---------+---------+---------+---------+----C10
         -0.024     -0.012     0.000     0.012     0.024     0.036
 N* = 1
```

Some tendency to non-independence, points drift downward from left to right.

B3. Var(errors) = fixed value ? Plot residuals versus fits.

```
MTB > name c4 'fits'  c5 'res'
MTB > plot 'res' 'fits'
          -
res       -                  *
          -
          -
     0.025+                                                    *
          -
          -                                    *
          -
          -                               *
     0.000+                                        *                   *
          -                                           *       *
          -                                                   *
          -      *
          -
    -0.025+                           *                           *
          -
          -
            --+---------+---------+---------+---------+---------+---------+----fits
          0.840     0.875     0.910     0.945     0.980     1.015
```

No evidence in plot of increase or decrease in spread of residual values, going from left to right (low to high fitted values).

B4. Errors normal ?

```
MTB > hist 'res'

Histogram of res   N = 12
 Midpoint   Count
   -0.02       3  ***
   -0.01       1  *
    0.00       4  ****
    0.01       2  **
    0.02       0
    0.03       1  *
    0.04       1  *                    <-----Does not look normal
                                       looks skewed toward large values, to the right
MTB > root 'res'
 BIN   COUNT    RAWRS     DRRS       SUSPENDED ROOTGRAM

   1     0.0     -1.3    -1.53        .  --------          .
   2     3.0      1.3     0.98        .           +++++     .
   3     1.0     -1.4    -0.78        .      ----           .
   4     4.0      1.5     0.91        .           +++++     .
   5     2.0     -0.0     0.13        .           +         .
   6     0.0     -1.2    -1.44        .  --------          .
   7     1.0      0.4     0.64        .           ++++      .
   8     1.0      0.8     1.11        .           ++++++    .
   9     0.0     -0.1    -0.12        .          -          .

 IN DISPLAY, VALUE OF ONE CHARACTER IS .2       OO
```

Histogram shows some degree of skewness, but this deviation is not serious, judging from comparison of frequency distribution with normal distribution (rootogram).

Sample size is small, the residuals deviate somewhat from normal and from independence, and the calculated Type I error ($p = 0.027$) is not all that far from the traditional criterion of $\alpha = 5\%$, hence this p-value should be checked with a randomization test.

GLM: Paired comparisons. Randomized blocks, a = 2 (srbx11_5.out)

Facial width data from Box 11.5 in Sokal and Rohlf (1995), page 353
Facial width of 15 individuals at ages 5 and 6

```
MTB > read 'a:srbx11_5.dat' c1 c2
     15 ROWS READ

 ROW     C1     C2
   1   7.33   7.53
   2   7.49   7.70
   3   7.27   7.46
  .   .   .

MTB > stack c1 c2 c3
MTB > set into c4
DATA> (0 1)15
MTB > end
MTB > set into c5
DATA> 2(1 2 3 4 5 6 7 8 9 10 11 12 13 14 15)
MTB > end
MTB > name c3 'fw' c4 'age' c5 'ind'
MTB > print c3 c4 c5

 ROW     fw   age   ind
   1   7.33     0     1
   2   7.49     0     2
   3   7.27     0     3
   4   7.93     0     4
   5   7.56     0     5
   6   7.81     0     6
   7   7.46     0     7
   8   6.94     0     8
   9   7.49     0     9
  10   7.44     0    10
  11   7.95     0    11
  12   7.47     0    12
  13   7.04     0    13
  14   7.10     0    14
  15   7.64     0    15
  16   7.53     1     1
  17   7.70     1     2
```

Read data,

reorganize to model format by stacking response variable into one column,

then set up two explanatory variables.

Compute the variance of the response variable var(fw). This gives SS_{total}

```
MTB > let k1 = stdev(c3)*stdev(c3)
MTB > print k1
K1       0.101640                    = Var(fw)
MTB > let k2 = 29*k1
MTB > print k2
K2       2.94755                     = SS_total
```

Use ANOVA command to partition SS_{total} according to GLM model statement.
Interaction term assumed absent.
Computational formula for F-ratios shown in boldface type.

```
MTB > anova 'fw' = 'age' 'ind';
SUBC> fits c6;
SUBC> residuals c7.

Factor     Type Levels Values
age       fixed      2    0     1
ind       fixed     15    1     2     3     4     5     6     7     8     9
                         10    11    12    13    14    15

Analysis of Variance for fw

Source      DF         SS         MS       F      P
age          1    0.30000   0.300000  388.89  0.000  F = MS(age)/MS(e)
ind         14    2.63675   0.188339  244.14  0.000  F = MS(Ind)/MS(e)
Error       14    0.01080   0.000771                 MS(e) = SS(e)/14
Total       29    2.94755   0.101640
```

Slopes used in model ? . . . . . No
Therefore no need to check assumption A, linear relation of response to explanatory.

Sample size small ? . . . . . Yes (n = 30)
Therefore check assumptions concerning errors ? (B1 B2 B3 B4)

p-value close to traditional criterion of 5% ? . . . . . No
Conclude that decision won't change, even if assumptions are not met.

Check assumptions if p-value itself needs to be defended, rather than the decision.

B1 Sum of residuals equal zero ? . . . . . Yes
(Because parameters were estimated from data by least squares)

B2 Residuals independent ? . . . . . Yes
(Plot of residual versus neighbouring value shows no pattern)

B3 Residuals homogeneous ? . . . . . Yes

```
MTB > plot 'res' 'fits'
          -                                                  *
res       -
          -                                  *
          -
     0.025+           *                     *      *
          -         *                                 *
          -
          -                                  *
          -     *              *       2         *
     0.000+                      *          *          *        *
          -             *                **       2
          -                             *
          -
          -                   *                                      *
    -0.025+                     *            *          *
          -                                  *
          -
          -                                                  *
           --------+---------+---------+---------+---------+---------+fits
                 7.00      7.25      7.50      7.75      8.00      8.25
```

B4 Residuals normal ? . . . . . Yes

```
MTB > hist 'res'
Histogram of res   N = 30
Midpoint   Count
   -0.04       1  *
   -0.03       2  **
   -0.02       4  ****
   -0.01       2  **
    0.00       8  ********
    0.01       6  ******
    0.02       2  **
    0.03       4  ****
    0.04       1  *
```

Conclude that p-values are correctly estimated.

Is paired comparisons better than simple comparison of means ? (i.e. more sensitive, lower type II error, better able to detect a true difference ?)

Yes. When same data are analyzed with simple t-test (compares two means), the error MS is larger, the F-ratio is smaller, the p-value is larger, and it is not significant (p = 0.086). With this test, one concludes (erroneously) that facial width does not differ between 5 and 6 year olds.

```
MTB > aovoneway c1 c2
 ANALYSIS OF VARIANCE
SOURCE     DF         SS         MS         F         p
FACTOR      1     0.3000     0.3000      3.17     0.086
ERROR      28     2.6475     0.0946
TOTAL      29     2.9475
                                       INDIVIDUAL 95 PCT CI'S FOR MEAN
                                       BASED ON POOLED STDEV
LEVEL       N       MEAN      STDEV  ----+---------+---------+---------+--
C1         15     7.4613     0.2997   (---------*----------)
C2         15     7.6613     0.3151                (----------*----------)
                                       ----+---------+---------+---------+--
POOLED STDEV =    0.3075               7.35      7.50      7.65      7.80
```

A better analysis was attained by statistical control for variation among individuals.

GLM: Hierarchical ANOVA (srbx10_1.out)

Winglength data from Box 10.1 in Sokal and Rohlf (1995), page 276.
Winglengths from each of 4 mosquitoes reared in each of 3 cages.

```
MTB > read 'a:srbx10_1.dat' c1-c3;
SUBC> nobs =24.
     24 ROWS READ

 ROW     C1   C2   C3
   1   58.5    1    1
   2   59.5    1    1
   3   77.8    2    1
   4   80.9    2    1

MTB > name c1 'wlngth' c2 'female' c3 'cage'
MTB > anova 'wlngth' = 'cage' 'female'('cage');
SUBC> random 'cage' 'female'('cage');                Note use of random command
SUBC> fits c4;                                       to compute correct F-ratios.
SUBC> residuals c5.

Factor         Type Levels Values
cage         random      3     1   2   3
female(cage) random      4     1   2   3   4

Analysis of Variance for wlngth

Source          DF         SS         MS       F      P          This is correct.
cage             2     665.68    332.838    1.74  0.230
female(cage)     9    1720.68    191.186  146.88  0.000
Error           12      15.62      1.302
Total           23    2401.97    104.434

MTB > anova 'wlngth' = 'cage' 'female'('cage');

Factor         Type Levels Values
cage          fixed      3     1     2     3
female(cage)  fixed      4     1     2     3     4

Analysis of Variance for wlngth

Source          DF         SS         MS       F      P      This is NOT correct.
cage             2     665.68    332.838  255.70  0.000
female(cage)     9    1720.68    191.186  146.88  0.000
Error           12      15.62      1.302
Total           23    2401.97    104.434
```

```
MTB > name c4 'fits' c5 'res'                    Now calculate ANOVA SS from data equations.
MTB > print c1-c5                                          wlngth = fits + res

 ROW  wlngth  female   cage    fits       res

   1    58.5       1      1   59.00  -0.50000
   2    59.5       1      1   59.00   0.50000
   3    77.8       2      1   79.35  -1.55000
   4    80.9       2      1   79.35   1.55000
   5    84.0       3      1   83.80   0.20000
   6    83.6       3      1   83.80  -0.20000
   7    70.1       4      1   69.20   0.90000
   8    68.3       4      1   69.20  -0.89999
   9    69.8       1      2   69.80   0.00000
  10    69.8       1      2   69.80   0.00000
  11    56.0       2      2   55.25   0.75000
  12    54.5       2      2   55.25  -0.75000
  13    50.7       3      2   50.00   0.70000
  14    49.3       3      2   50.00  -0.70000
  15    63.8       4      2   64.80  -1.00000
  16    65.8       4      2   64.80   1.00001
  17    56.6       1      3   57.05  -0.45000
  18    57.5       1      3   57.05   0.45000
  19    77.8       2      3   78.50  -0.70000
  20    79.2       2      3   78.50   0.70000
  21    69.9       3      3   69.55   0.35001
  22    69.2       3      3   69.55  -0.35000
  23    62.1       4      3   63.30  -1.20000
  24    64.5       4      3   63.30   1.20000

MTB > set into c6
DATA> (72.84 59.96 67.10)8                                  These are the cage means.
DATA> end
MTB > describe c1 c4 c5 c6
                 N     MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
wlngth          24    66.63    67.05    66.63    10.22     2.09
fits            24    66.63    67.00    66.61    10.19     2.08
res             24    0.000    0.000    0.000    0.824    0.168
C6              24    66.63    67.10    66.65     5.38     1.10

MTB > let k1 = 10.22*10.22*23                               Compute Sums of Squares
MTB > let k2 = 10.19*10.19*23
MTB > let k3 = .824*.824*23
MTB > let k4 = 5.38*5.38*23
MTB > print k1 k2 k3 k4
K1       2402.31      SS total
K2       2388.23      SS model        (cages, female within cages)
K3       15.6164      SS within
K4       665.721      SS groups       (cages)
```

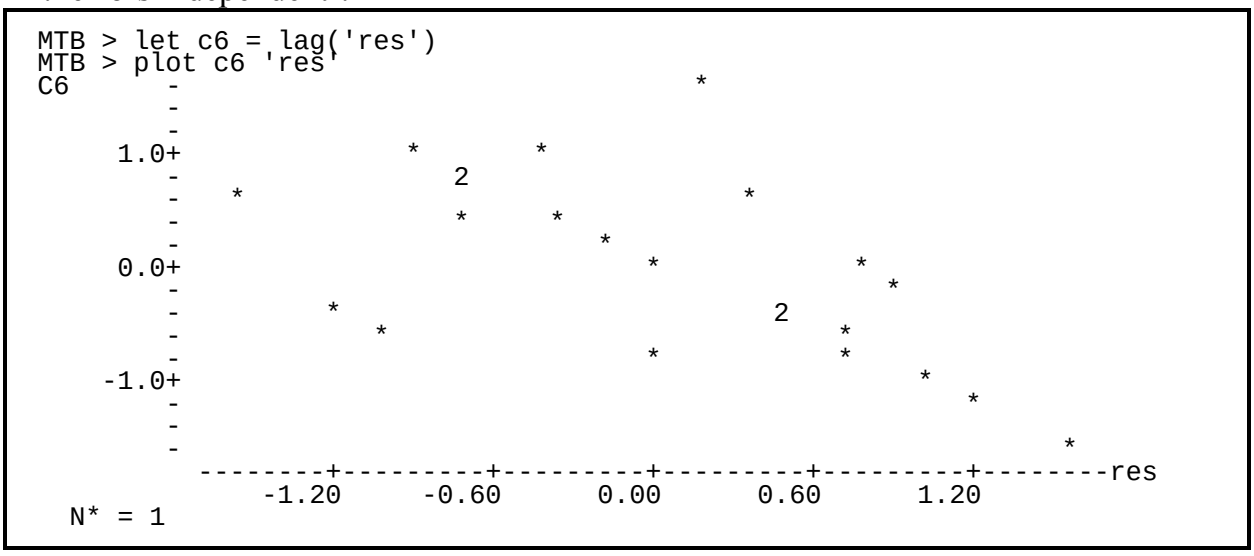
Use residuals to check assumptions.

Model contain estimates of slopes of straight lines ? . . . . . No
Therefore no need to check assumption A1, linearity.

Sample size small ? . . . . . Yes (n = 24)
Therefore p-value may be incorrect if assumptions not met.

B1 Residuals sum to zero ? . . . . . Yes
Because parameters estimated by least squares.

B2. errors independent ?



Residuals not independent of neighbour, in order of data listing.

Examination of data equations on previous page reveals that this arises from regular alternation of high and low values, when means are computed for only two data points (winglengths) in each unit (female). To remove this problem, offset residuals by 2 rather than 1, then examine graph for independence.

Lag command to offset residuals, then plot.

```
MTB > let c7 = lag(c6)
MTB > plot c7 c5
C7      -                          *
        -
        -
     1.0+                             *       *
        -            *                            *
        -                      *                              *
        -         *                               *
        -                                            *
     0.0+                 *                        *
        -              *
        -   *             *                               *
        -                                   *          *
        -                 *           *
    -1.0+                      *
        -
        -
        -                                *
          --------+---------+---------+---------+---------+--------res
              -1.20     -0.60      0.00      0.60      1.20
   N* = 2
```

Residuals independent.

B3. var(errors) = constant (errors homogeneous)?

```
MTB > plot 'res' 'fits'
res     -                                                *
        -
        -                         *
     1.0+                            *     *
        -      *       *
        -                   *                           *
        -                *                 *
        -                                                       *
     0.0+                                   2
        -                                                       *
        -                *  *              *
        -      *                                        *
        -              *                   *
    -1.0+                            *
        -                         *
        -
        -                                                *
          ----+---------+---------+---------+---------+---------+--fits
           49.0      56.0      63.0      70.0      77.0      84.0
```

Graph shows that errors constant. There is no evidence of increasing spread of residuals (cones) going from left to right or right to left.

B4. Errors normal ?

```
MTB > hist 'res'
Histogram of res   N = 24

Midpoint   Count
    -1.5       1  *
    -1.0       3  ***
    -0.5       6  ******
     0.0       4  ****
     0.5       5  *****
     1.0       4  ****
     1.5       1  *
```

Yes, histogram looks close to normal.

Assumption met. Therefore p-values are estimated correctly from F-distribution.

Source	F-ratio	p-value
cage	1.74	0.230
female(cage)	146.88	<0.001

GLM: Analysis of Covariance--Homogeneity of slopes. (brussard.out)

Heterzygosity data collected by Th. Dobzhansky (1948) *Genetics* 33: 158-176.
Data on Inversion heterozygosity (assuming Hardy Weinberg equilibrium) of 3rd chromosome inversions in *Drosophila pseudoobscura* (col 3: HDps = %) and *Drosophila persimilis* (col 2) (HDp = %) in relation to altitude in Yosemite Park. (c1: Elev = ft)

Data reported by P.F. Brussard 1984. Geographic patterns and environmental gradients: The central-marginal model in Drosophila revisited. *Annual Review of Ecology and Systematics* **15**: 25-64.

```
   850  0.59  0.70
  3000  0.37  0.69
  4600  0.41  0.71
  6200  0.40  0.70
  8000  0.31  0.70
  8600  0.18  0.62
 10000  0.20  0.68

Elev    HDp   HDps
```

```
MTB > read 'a:brussard.dat' c1-c3;
SUBC> nobs=7.
      7 ROWS READ

 ROW      C1     C2     C3

   1     850   0.59   0.70
   2    3000   0.37   0.69
   3    4600   0.41   0.71
  .  .  .
MTB > name c1 'alt' c2 'Dpers' c3 'Dpseu'
```

Begin with regression of heterozygosity on altitude in *D. persimilis*

```
MTB > regress 'Dpers' 1 'alt'

 The regression equation is
Dpers = 0.580 -0.000039 alt

Predictor       Coef       Stdev    t-ratio        p
Constant     0.58006     0.05287      10.97    0.000
alt      -0.00003880  0.00000798      -4.86    0.005

s = 0.06394     R-sq = 82.5%     R-sq(adj) = 79.0%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    1    0.096644    0.096644     23.64    0.005
Error         5    0.020442    0.004088
Total         6    0.117086
```

Detectable decrease in heterozygosity with altitude in *D. persimilis*.

Next, regression of *D. pseudoobscura* heterozygosity on altitude.

```
MTB > regress 'Dpseu' 1 'alt'

The regression equation is
Dpseu = 0.712 -0.000004 alt

Predictor       Coef       Stdev    t-ratio        p
Constant     0.71166     0.02432      29.26    0.000
alt      -0.00000440  0.00000367      -1.20    0.284

s = 0.02942     R-sq = 22.3%     R-sq(adj) = 6.8%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    1   0.0012449   0.0012449      1.44    0.284
Error         5   0.0043265   0.0008653
Total         6   0.0055714

Unusual Observations
Obs.     alt     Dpseu       Fit Stdev.Fit  Residual   St.Resid
  6     8600    0.6200    0.6738    0.0149   -0.0538     -2.12R

R denotes an obs. with a large st. resid.
```

Change in heterozygosity with altitude not detectable in *D. pseudoobscura*, but was detectable in *D. persimilis*. Is this apparent difference significant ?
I.e., do the regression slopes differ in the two species ?

```
MTB > stack c2 c3 c4
MTB > stack c1 c1 c5
MTB > set into c6
MTB > end
MTB > name c4 'Hyz' c5 'alti' c6 'sp'
MTB > print c4-c6
 ROW    Hyz    alti   sp
   1   0.59     850   -1
   2   0.37    3000   -1
   3   0.41    4600   -1
   4   0.40    6200   -1
   5   0.31    8000   -1
   6   0.18    8600   -1
   7   0.20   10000   -1
   8   0.70     850    1
   9   0.69    3000    1
  10   0.71    4600    1
  11   0.70    6200    1
  12   0.70    8000    1
  13   0.62    8600    1
  14   0.68   10000    1
```

To address this question, the data will have to be re-organized, with one response variable (heterozygosity), and two explanatory variables (species and altitude).

Slopes are compared in an Analysis of Covariance ANCOVA.

However, the ANCOVA command in Minitab does not compare slopes (!)

```
 MTB > ancova 'Hyz' = 'sp';
 SUBC> covariate 'alti';
 SUBC> fits c8;
 SUBC> residuals c9.

 Factor   Levels Values
 sp            2    -1     1

 Analysis of Covariance for Hyz

 Source      DF     ADJ SS         MS       F      P
 Covariates   1    0.05991   0.059913   10.50  0.008
 sp           1    0.39111   0.391114   68.57  0.000
 Error       11    0.06274   0.005704
 Total       13    0.51377   0.039521

 Covariate     Coeff     Stdev   t-value         P
 alti      -0.000022  0.000007    -3.241     0.008
```

The ANCOVA command does not include the interaction term (which is the term that compares slopes).

To compare slopes, write the complete model, then use the GLM command.

```
 MTB > glm 'hyz' = 'Hyx' = 'alti' 'sp' 'alti'*'sp';
 SUBC> covariate 'alti';
 SUBC> fits c8;
 SUBC> residuals c9.

Factor   Levels Values
sp            2     0     1

Analysis of Variance for Hyz

Source     DF     Seq SS     Adj SS     Adj MS       F      P
alti        1    0.05991    0.05991    0.05991   24.19  0.000
sp          1    0.39111    0.01267    0.01267    5.11  0.047
sp*alti     1    0.03798    0.03798    0.03798   15.33  0.003
Error      10    0.02477    0.02477    0.00248
Total      13    0.51377

Term         Coeff     Stdev  t-value      P
Constant   0.64586   0.02910    22.20  0.000
alti     -0.000022  0.000004    -4.92  0.000
alti*sp  -0.000017  0.000004    -3.92  0.003

 Unusual Observations for Hyz

Obs.      Hyz       Fit Stdev.Fit  Residual   St.Resid
  2  0.370000  0.463666  0.026013 -0.093666     -2.21R
 R denotes an obs. with a large st. resid.
```

Model assumes straight line relation of response variable to explanatory variable.
Use residuals to check assumption concerning linear model (Assumption A)

```
MTB > name c8 'fits' c9 'res'
MTB > plot 'res' 'fits'
          -                      *
     0.050+
          -             *                          *
res       -
          -                                                  2*
          -     *                     *                      *
     0.000+                                                     **
          -
          -
          -
    -0.050+                                                  *
          -
          -           *
          -
          -                                *
    -0.100+
           ----+---------+---------+---------+---------+---------+--fits
            0.20      0.30      0.40      0.50      0.60      0.70

MTB > plot 'res' 'alti'
          -                                *
     0.050+
          -     *                                    *
res       -
          -                          *       *       *
          -                          *                       2
     0.000+
          -     *          *
          -
          -
          -
    -0.050+                                               *
          -
          -                                               *
          -
          -                 *
    -0.100+
           +---------+---------+---------+---------+---------+------alti
           0       2000      4000      6000      8000     10000
```

Note that residuals can be plotted against an explanatory variable as well as against the fitted values, when these are not the same.

Model is acceptable. No bowls or arches.

Sample size small (n = 14) so p-value may not be calculated correctly if error assumptions not met. Examine assumptions before taking p-value as correct.

B1 sum(error) = 0 ? GLM uses least squares, so sum(error) will be zero.

B2 errors independent ?

```
MTB > let c9 = lag('res')
MTB > plot c9 'res'
          -                                                  *
     0.050+
          -    *         *
C9        -
          -                   *                     *   *
          -                                *                          *
     0.000+                                *         *
          -
          -
          -
    -0.050+                                          *
          -
          -                                         *
          -
          -                                         *
    -0.100+
           ----+---------+---------+---------+---------+---------+---------+--res
         -0.090    -0.060    -0.030     0.000     0.030     0.060
 N* = 1
```

Some evidence for non-independence: on left side a downward trend, with upward trend on right side.

B3 var(errors) = constant ?
Plot of residuals vs fits (see above) show no cones. However, residuals show greater scatter one species than in other.

```
MTB > plot 'res' 'sp'
          - *
    0.050+
          - 2
res       -
          -                                                    3
          - 2                                                  *
    0.000+
          -                                                    2
          -
          -
          -
   -0.050+                                                     *
          -
          - *
          -
          - *
   -0.100+
           +---------+---------+---------+---------+---------+------sp
        0.00      0.20      0.40      0.60      0.80      1.00
```

B4 errors normal ?

```
 MTB > hist 'res';
 SUBC> increment .04.

Histogram of res   N = 14

Midpoint   Count
 -0.0800       2  **
 -0.0400       1  *
  0.0000       7  *******
  0.0400       3  ***
  0.0800       1  *
```

Summary of analysis of errors. Some evidence for non-independence, some evidence for greater variance in error in one species than other, but residuals look acceptably normal.

Conclusion. p-value (p=0.003) may not be correct, but it is far enough from the traditional criterion of " = 5% that the decision is unlikely to change if the p-value were computed more accurately.

Reject null hypothesis (H_o: slopes equal), accept alternative (H_A: slopes not equal). However, if conclusion had to be defended (e.g. lives or livelihoods depended on it) then compute p-value more accurately. Better estimate of p-value can be obtained by randomization. Better estimate can also be obtained by using better model for errors. See: GLIM for Ecologists (1993) by M. Crawley. London: Blackwell

GLM: Analysis of Covariance--Statistical Control (CrwTb9_1.out)

Seed production data from Table 9.1 in Crawley (1993)

Data on seed production (fruit = mg dry wt) of a biennial plant with and without grazing by rabbits.

Inital plant size measured as diameter (mm) at top of rootstock.

Data from Table 9.1 in
GLIM for Ecologists (1993)
by M. Crawley.
London: Blackwell.

```
59.77  6.225  80.31  8.988
60.98  6.487  82.35  8.975
14.73  4.919 105.1   9.844
19.28  5.130  73.79  8.508
34.25  5.417  50.08  7.354
35.53  5.359  78.28  8.643
87.73  7.614  41.48  7.916
63.21  6.352  98.47  9.351
24.25  4.975  40.15  7.066
64.34  6.930 116.1  10.25
52.92  6.248  38.94  6.958
32.35  5.451  60.77  8.001
53.61  6.013  84.37  9.039
54.86  5.928  70.11  8.910
64.81  6.264  14.95  6.106
73.24  7.181  70.70  7.691
80.64  7.001  71.01  8.515
18.89  4.426  83.03  8.530
75.49  7.302  52.26  8.158
46.73  5.836  46.64  7.382
fruit  root   fruit  root
ungrazed      grazed
```

```
MTB > read 'crwtb9_1.dat' c1 c2 c3 c4
Entering data from file: crwtb9_1.dat
     20 rows read.

MTB > stack c1 c3 c5
MTB > stack c2 c4 c6
MTB > name c5 'fruit' c6 'root'
MTB > set into c7
DATA> (0 1)20                                          0 = Ungrazed
DATA> end                                              1 = grazed
MTB > name c7 'grazing'
MTB > describe 'fruit';
SUBC> by 'grazing'.                  Mean value of seed production apparently greater
                                                       in grazed areas (!)
          grazing        N     MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
fruit           0       20    50.88    54.24    50.84    21.76     4.87
                1       20    67.94    70.85    68.21    24.97     5.58

MTB > describe 'fruit'

                N     MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
fruit          40    59.41    60.88    59.04    24.68     3.90
```

This is contrary to what we expect, which is less seed production in grazed plants

Plants allocated to grazed areas were larger than those allocated to ungrazed. To remove effects of plant size, use root stock diameter as a covariate. Root size can be used as control variable if the slope relating fruit to root in ungrazed area is same as slope in grazed area, if slopes are same. Test the homogeneity of slopes by examining interaction term.

```
MTB > glm 'fruit' = 'root' 'grazing' 'root'*'grazing';
SUBC> covariate 'root';
SUBC> fits c8;
SUBC> residuals c9.

Factor   Levels Values
grazing       2     0     1

Analysis of Variance for fruit

Source         DF     Seq SS     Adj SS     Adj MS       F      P
root            1    16800.4    18791.6    18791.6  402.57  0.000
grazing         1     5266.7      157.1      157.1    3.37  0.075
grazing*root    1        4.6        4.6        4.6    0.10  0.754
Error          36     1680.5     1680.5       46.7
Total          39    23752.2

MTB > name c8 'fits' c9 'res'
MTB > plot 'res' 'fits'                    Model based on two slopes acceptable
                                                    No bowls or arches.
 res     -
         -                          *
         -
        12+
         -                              *         *
         -   *                  *               *         *
         -                    * *  **
         -               **                     * *               *    *
         0+           *       *  2       **           2 *
         -     *           *           *        *     *
         -                 *                       **
         -                                         *
         -          * *
       -12+                                  *         *
         -
         -                              *
         -
           ------+---------+---------+---------+---------+---------+fits
                 20        40        60        80       100       120

MTB > hist 'res'                           Variance of residuals constant (above)
Histogram of res   N = 40
 Midpoint   Count                          Variance of residuals normally distributed
     -15       1  *
     -10       5  *****
      -5       5  *****
       0      16  ****************
       5      10  **********
      10       2  **
      15       1  *
```

The slopes for grazed and ungrazed areas do not differ ($F_{1,36} = 0.10$ p = 0.754)
Therefore a model with a single slope term (and no interaction term) can be used to remove effects of plant size (= root size).

```
MTB > glm 'fruit' = 'root' 'grazing';
SUBC> covariate 'root';
SUBC> fits c8;
SUBC> residuals c9.

Factor   Levels Values
grazing       2     0     1

 Analysis of Variance for fruit

Source     DF     Seq SS     Adj SS     Adj MS       F      P
root        1      16800      19155      19155  420.60  0.000
grazing     1       5267       5267       5267  115.64  0.000
Error      37       1685       1685         46
Total      39      23752

MTB > plot 'res' 'fits'
 res      -
          -                          *
          -
         12+
          -                           *         *
          -   *                    *            *
          -                      * * * *                  *
          -                   **                  *                  *
         0+     *         *      2       **      *       **        *
          -           *     *           *        *       *
          -                 *                     ** *
          -                                       *
          -          *  *
        -12+                                *         *
          -
          -                              *
          -
            ------+---------+---------+---------+---------+---------+fits
                 20        40        60        80       100       120
MTB > hist 'res'
Histogram of res   N = 40
Midpoint   Count
     -15       1  *
     -10       4  ****
      -5       6  ******
       0      17  *****************
       5       9  *********
      10       2  **
      15       1  *
```

Calculate intercepts of two parallel lines to determine difference in seed production between grazed and ungrazed areas.

a1 = Mean(Y1) ! b*Mean(X1) = 50.88 ! 23.6*6.053 = !91.729 **ungrazed**

a2 = Mean(Y2) ! b*Mean(X2) = 67.94 ! 23.6*8.309 = !127.82.729 **grazed**

Grazed plants produce fewer seeds (127.82 ! 91.82) than ungrazed.

GLM: Multiple Regression. (sctb17_1.out)

Soil phosphorus data, Table 17.2.1 in Snedecor and Cochrane (1980)
Plant available soil phosphorus (ppm) in 17 Iowa soils at 20 deg C in relation to inorganic and organic phosphorus.

```
MTB > read 'a:sctb17_1.dat' c1 c2 c3 c4;
SUBC> nobs = 17.
MTB > name c1 'sample' c2 'inorgP' c3 'orgP' c4 'availP'
```

Begin with analysis of response variable availP relative to one explanatory variable, inorganic phosphorus.

```
MTB > plot c4 c2
        -                                                     *
availP  -                                          2
        -                    * *                *
      90+
        -
        -
        -                   *
        -      *                                   *
      75+                     *
        -        *
        -
        -
        -    *
      60+    2
        -
        -      *    *
        -                        *
          --+---------+---------+---------+---------+---------+----inorgP
          0.0       6.0      12.0      18.0      24.0      30.0
```

```
MTB > regress c4 1 c2

The regression equation is
availP = 62.6 + 1.23 inorgP

Predictor       Coef       Stdev    t-ratio        p
Constant      62.569       4.452      14.05    0.000
inorgP        1.2291      0.3058       4.02    0.001

s = 11.92       R-sq = 51.9%     R-sq(adj) = 48.6%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    1      2295.2      2295.2     16.15    0.001
Error        15      2131.2       142.1
Total        16      4426.5

Unusual Observations
Obs.  inorgP    availP       Fit Stdev.Fit  Residual   St.Resid
 10     12.6     51.00     78.06      2.93    -27.06     -2.34R
R denotes an obs. with a large st. resid.
```

Next, analysis of the response variable, availP, relative to the other explanatory variable organic phosphorus oP

```
MTB > plot c4 c3
          -                                        *
availP    -                                    *       *
          -                  * *               *
        90+
          -
          -
          -                                    *
          -                                          *              *
        75+                          *
          -       *
          -
          -
          -                                             *
        60+           *           *
          -
          -             *           *
          -                                                *
            ------+---------+---------+---------+---------+---------+orgP
                  20        30        40        50        60        70
```

```
MTB > regress c4 1 c3

The regression equation is
availP = 65.4 + 0.262 orgP

Predictor       Coef       Stdev    t-ratio        p
Constant       65.38       13.49       4.85    0.000
orgP          0.2622      0.3124       0.84    0.414

s = 16.79       R-sq = 4.5%      R-sq(adj) = 0.0%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    1       198.6       198.6      0.70    0.414
Error        15      4227.8       281.9
Total        16      4426.5
```

The variance explained by organic phosphorus in the soil is 198.6 out of SST = 4426.5, compare this to variance explained by inorganic phosphorus in soil, 2295.2 out of SST of 4426.5

Next, multiple regression. Response variable availP expressed as function of both explanatory variables, ioP and oP.

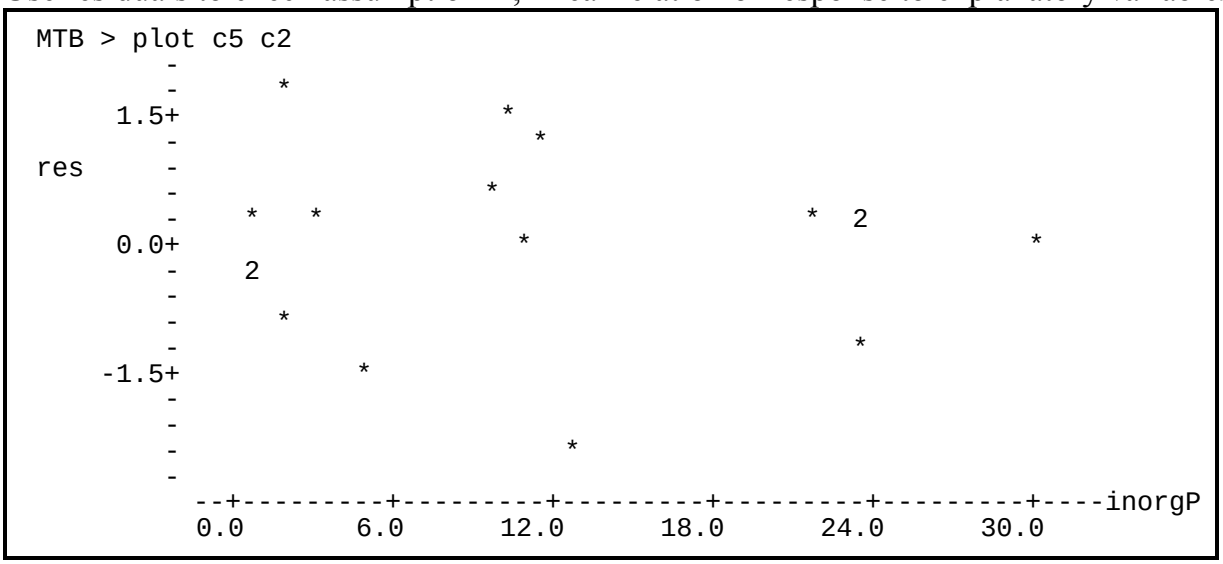
```
MTB > regress c4 2 c3 c2 [residuals c5]

The regression equation is
availP = 66.5 - 0.111 orgP + 1.29 inorgP

Predictor       Coef       Stdev    t-ratio        p
Constant      66.465       9.850       6.75    0.000
orgP         -0.1110      0.2486      -0.45    0.662
inorgP        1.2902      0.3428       3.76    0.002

s = 12.25       R-sq = 52.5%     R-sq(adj) = 45.7%

Analysis of Variance
SOURCE       DF          SS          MS         F        p
Regression    2      2325.2      1162.6      7.75    0.005
Error        14      2101.3       150.1
Total        16      4426.5

SOURCE       DF      SEQ SS
orgP          1       198.6  ..................... same as previous analysis of oP
inorgP        1      2126.5  ................... not the same as previous analysis of ioP

Unusual Observations
Obs.    orgP    availP       Fit Stdev.Fit  Residual   St.Resid
 10     58.0     51.00     76.28      4.98    -25.28     -2.26R
 R denotes an obs. with a large st. resid.
```

Is model acceptable? Plot residuals against fitted values calculated from equation. Or plot against the explanatory variables ioP and oP.

```
MTB > name c5 'res'
MTB > plot c5 c3
          -
          -                                                    *
       1.5+                   *
          -                  *
res       -
          -                                  *
          -       *                          * *       *   *
       0.0+                                  *               *
          -           *           *
          -
          -                          *
          -                                                  *
      -1.5+            *
          -
          -
          -                                                        *
          -
            ------+---------+---------+---------+---------+---------+orgP
                  20        30        40        50        60        70
```

Model acceptable, based on plot of residuals against first explanatory variable, ioP.

Use residuals to check assumption A, linear relation of response to explanatory variables.



Model acceptable based on plot of residuals against second explanatory variable, ioP.

If the two explanatory variables are correlated with each other, then the partitioning of the variance will depend on the order they are entered in the regression model.

```
MTB > correlate c2-c4
                inorgP      orgP
orgP             0.399
availP           0.720     0.212
```

```
MTB > regress c4 2 c2 c3 [residuals c5]
     The regression equation is
     availP = 66.5 + 1.29 inorgP - 0.111 orgP

Predictor       Coef       Stdev    t-ratio        p
Constant      66.465       9.850       6.75    0.000
inorgP        1.2902      0.3428       3.76    0.002
orgP         -0.1110      0.2486      -0.45    0.662

s = 12.25       R-sq = 52.5%     R-sq(adj) = 45.7%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    2      2325.2      1162.6      7.75    0.005
Error        14      2101.3       150.1
Total        16      4426.5

SOURCE       DF      SEQ SS
inorgP        1      2295.2   ..... same as analysis of ioP only
orgP          1        29.9   ..... not the same as analysis of oP only
```

Compare this analysis (ioP first in regression statement) with previous analysis (oP first).

Check assumptions concerning errors.
B1 sum(erorrs) = 0 ? Yes, because least squares was used by regression command.
B2 errors independent ?

```
MTB > let c10 = lag('res')
MTB > plot c10 'res'
          -                                          *
       1.5+                                                  *
          -      *
C10       -
          -                                                      *
          -                     *  *       * *          *
       0.0+                                             *
          -                 *                      *
          -
          -                                    *
          -                                        *
      -1.5+                                                          *
          -
          -
          -                                    *
          -
            --+---------+---------+---------+---------+---------+---------+----res
           -2.40      -1.60      -0.80       0.00       0.80       1.60
   N* = 1
```

Residuals scattered throughout plot, no evidence of non-independence.

B3 var(errors) = constant ? Yes, see above, residuals vs both explanatory variables

B4 errors normal ?

```
MTB > name c5 'res'
MTB > hist 'res'

Histogram of res   N = 17

Midpoint   Count
    -2.5       1  *
    -2.0       0
    -1.5       1  *
    -1.0       2  **
    -0.5       1  *
     0.0       4  ****
     0.5       5  *****
     1.0       0
     1.5       3  ***
```

Try constructing histogram with wider increment.

```
MTB > hist 'res';
SUBC> increment 1.

Histogram of res   N = 17

Midpoint   Count
   -2.00       1  *
   -1.00       3  ***
    0.00       9  *********
    1.00       3  ***
    2.00       1  *
```

This illustrates how informal (visual) interpretation of histogram depends on the interval used in constructing the histogram.

Conclusions. Linear model acceptable. Error assumptions met, so p-value calculated from cumulative distribution function (cdf F) can be trusted.

GLM: Revision of Model. (srbx14_9.out)

Membrane potential data from Box 14.9 of Sokal and Rohlf (1995), page 504.

Membrane potential (millivolts) for 4 different cation systems, as a function of the logarithm of the activity ratio of various concentrations. Here is a plot of the data.

```
MTB > read 'srbx14_9.dat' c1-c3;
SUBC> nobs=21.
Entering data from file: srbx14_9.dat
     21 rows read.
MTB > name c1 'Vmem' c2 'aR' c3 'Gr'
MTB > plot 'Vmem'  'aR'
         -
         -                                                   *
       40+                                                *   *
         -                                                 *
 Vmem    -                                        *           *
         -                                                 *
         -                                         *
       20+                               *                       *
         -                                *        *         *
         -
         -                                          *
         -              *           *
        0+
         -  *            *                 *
         -               *
         -   *
         -
           --------+---------+---------+---------+---------+--------aR
               -1.50     -1.00     -0.50      0.00      0.50
```

Here is analysis of covariance, using GLM command.

```
MTB > glm 'Vmem' = 'aR' 'Gr' 'aR'*'Gr';
SUBC> covariates 'aR';
SUBC> fits c4;
SUBC> residuals c5.
  Factor   Levels Values
  group         4     1     2     3     4
Analysis of Variance for mempt
  Source         DF     Seq SS     Adj SS     Adj MS        F       P
  aR              1    4197.01    3192.09    3192.09   876.71   0.000
  Gr              3    1768.58    1413.83     471.28   129.44   0.000
  Gr*aR           3       0.80       0.80       0.27     0.07   0.973
  Error          13      47.33      47.33       3.64
  Total          20    6013.72
Term            Coeff      Stdev   t-value       P
Constant      18.9633     0.4734     40.06   0.000
aR            20.9990     0.7092     29.61   0.000
aR*Gr
           1   -0.333      1.734     -0.19   0.851
           2    0.070      1.115      0.06   0.951
           3   0.3951     0.9434      0.42   0.682

Unusual Observations for mempt
Obs.    Vmem        Fit Stdev.Fit  Residual   St.Resid
 10  -10.8000  -13.5338    1.4835    2.7338      2.28R
 R denotes an obs. with a large st. resid.
```

Linear model acceptable ?

```
MTB > name c4 'fits' c5 'res'
MTB > plot 'res' 'fits'
         -      *
         -                                                 *
      2.0+           *                                         *
         -
 res     -          *                                *
         -
         -               *              *               *   *
      0.0+                                  *     *
         -
         -                       *               *
         -                    *            *         *
         -                *    *             *
     -2.0+
         -
         -
         -                               *
         -
           ------+---------+---------+---------+---------+---------+fits
               -12         0        12        24        36        48
```

Residuals indicate that linear model is not acceptable.

A series of models were examined:

$$\log(Vmem) = \log(aR) + Gr + Gr*\log(aR)$$
$$\log(Vmem) = aR + Gr + Gr*aR$$
$$Vmem = \log(aR) + Gr + Gr*\log(aR)$$
$$1/Vmem = 1/aR + Gr + Gr*(1/aR)$$
$$Vmem^2 = aR + Gr + Gr*aR$$

All resulted in bowls or arches, when plotted against fitted values. The high (or low) point was around aR = 0.7, when residuals were plotted against aR. This suggested a two level model: Level = above 0.7 or below 0.7. The model was written with three explanatory variables, and one interaction variable, which tests whether heterogeneity of slopes (aR*Gr) depends upon level.

```
MTB > let c4 = ('aR'+.7)/abs('aR'+.7)                    = above or below 0.7
MTB > name c4 'Lvl'
MTB > glm 'Vmem' = 'Lvl' 'aR'  'Gr'  'aR'*'Gr'*'Lvl'
                                          I
   * ERROR * Model is non-hierarchical at I.
```

Model poorly constructed. Model will need to be revised.

Revise model by including aR*Gr and aR*Lvl. This asks whether relation of Vmem to aR depends on group Gr (as before). It also consider whether relation of Vmem to aR depends on Lvl (aR*Lvl).

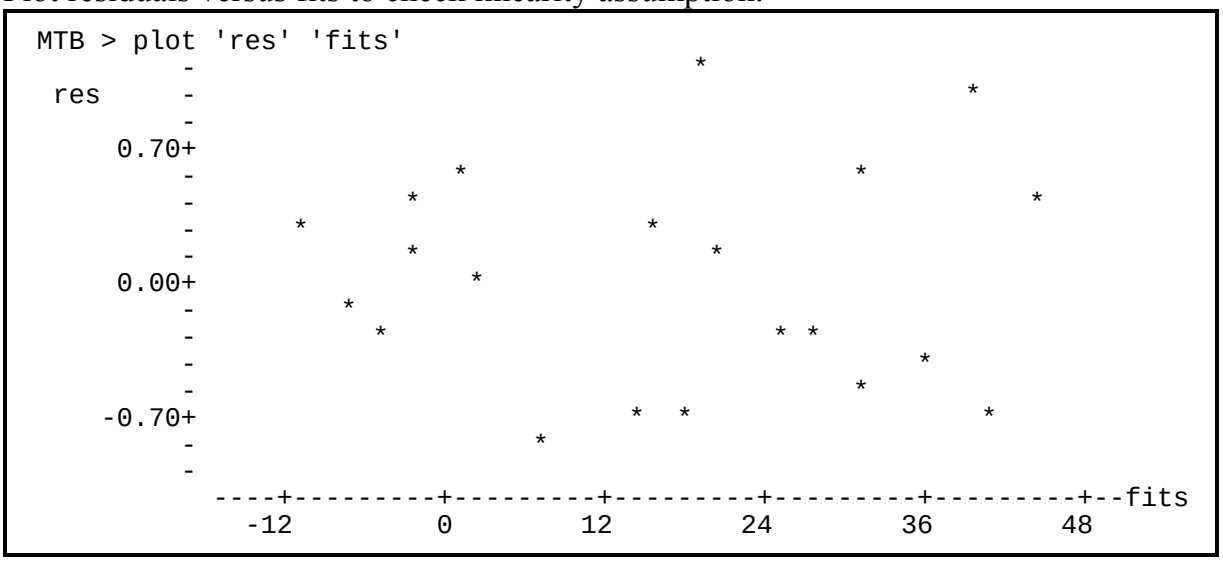
```
MTB > glm 'Vmem' = 'Lvl' 'aR'  'Gr'  'aR'*'Gr' 'aR'*'Lvl';
SUBC> covariate 'aR';
SUBC> fits c5;
SUBC> residuals c6.
  Factor   Levels Values
    Lvl           2    -1     1
    Gr            4     1     2     3     4
Analysis of Variance for Vmem
  Source     DF     Seq SS     Adj SS     Adj MS       F      P
  Lvl         1    2956.08      14.08      14.08   24.25  0.000
  aR          1    1255.08     360.51     360.51  620.81  0.000
  Gr          3    1760.80    1254.03     418.01  719.83  0.000
  Gr*aR       3       3.11      15.59       5.20    8.95  0.003
  Lvl*aR      1      32.27      32.27      32.27   55.56  0.000
  Error      11       6.39       6.39       0.58
  Total      20    6013.72
Term          Coeff     Stdev  t-value      P
Constant   13.6120    0.9720    14.00  0.000
aR         17.8394    0.7160    24.92  0.000
aR*Gr  1   -2.7116    0.7738    -3.50  0.005
       2   -0.7598    0.4743    -1.60  0.138
       3    1.8763    0.4179     4.49  0.000
aR*Lvl -1  -5.5381    0.7430    -7.45  0.000
```

The results from this model indicate that relation between Vmem and aR depends on Level ($t = \pm 7.45$ $p < 0.001$):

aR*Lvl -1 -5.5381 0.7430 -7.45 0.000

Before interpreting model, however, the linearity assumption is checked.

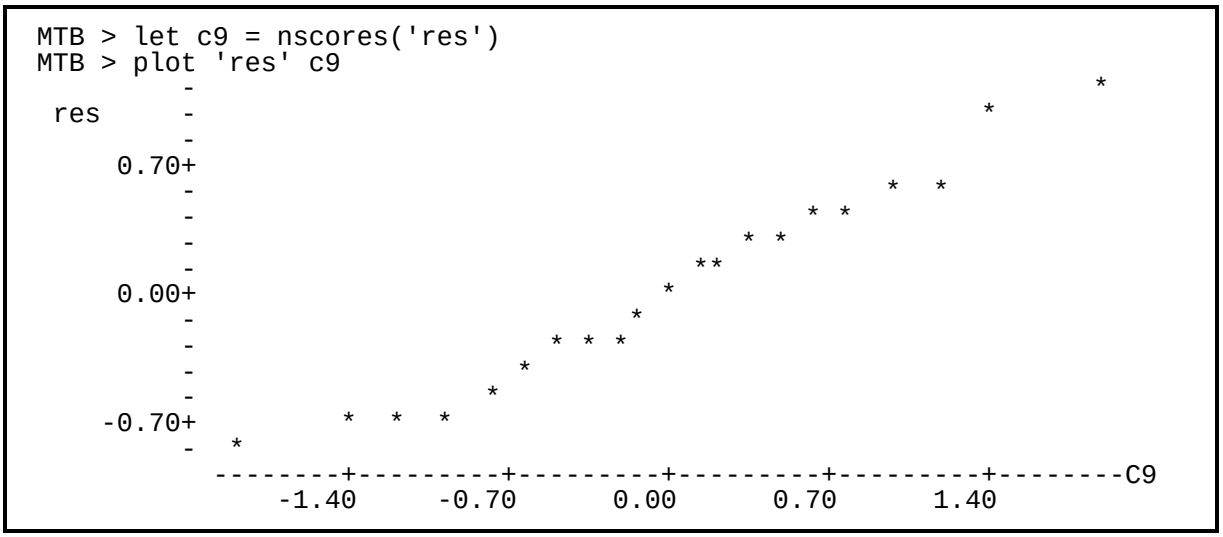
Plot residuals versus fits to check linearity assumption.



Model acceptable. No bowls or arches.

Can the computed p-value be trusted?
Graph of residuals vs fits shows no cones, so residuals are homogenous..

Next, errors normal ?



Conclusion: residuals are homogeneous and close to normal,
so p-values in ANOVA table printed by Minitab are acceptable.

```
MTB > hist 'res';
SUBC> increment .3.

Histogram of res   N = 21
Midpoint   Count
  -0.900       1  *
  -0.600       4  ****
  -0.300       4  ****
   0.000       3  ***
   0.300       5  *****
   0.600       2  **
   0.900       1  *
   1.200       1  *
```

Decisions: The relation of Vmem to aR depends on level; the slope that relates Vmem to aR is greater at high aR levels (aR > 0.7) than at lower levels (aR < 0.7). Also, the relation of Vmem to aR depends on group ($F = 8.95$, $p = 0.003$). This conclusion differs from that arising from the initial analysis, which would have been that relation of Vmem to aR is uniform across groups ($F = 0.07$, $p = 0.973$).

This example shows how a hidden source of heterogeneity can obscure a relation, resulting in Type II error (false acceptance of H_o). In this case the source of heterogeneity was a change in slope: greater slope at high aR than low aR values.

This example shows how model revision can improve the analysis of data.

PART III

Binomial Response Variable (srbx17_8.out)

Beetle colouration data from Box 17.8 in Sokal and Rohlf (1995), page
Two-way analysis of colour pattern frequency.

```
MTB > set into c1
DATA> 29 273 8 64
MTB > end
MTB > set into c2
DATA> 11 191 31 64
MTB > end
```

```
MTB > let c3 = c1 + c2
MTB > let k1 = sum(c1) + sum(c2)
MTB > let c3 = c3/k1
MTB > let c4 = c3*sum(c1)/k1
MTB > let c5 = c3*sum(c2)/k1
```

Compute expected proportions in each cell of the two-way table.

```
MTB > print c4 c5

 ROW         C4         C5

   1   0.033227   0.026386
   2   0.385429   0.306076
   3   0.032396   0.025726
   4   0.106325   0.084435
```

Print the expected proportions p, one for each cell of the two-way table.

```
MTB > stack c1 c2 c6
MTB > stack c4 c5 c7
MTB > let c7 = c7*k1
MTB > let c8 = c6 - c7
MTB > name c6 'f' c7 'pN' c8 'res'
```

Compute the fitted values fits = pN.
Then compute the residuals, based on the data equations:
Observed = Fits + Residual

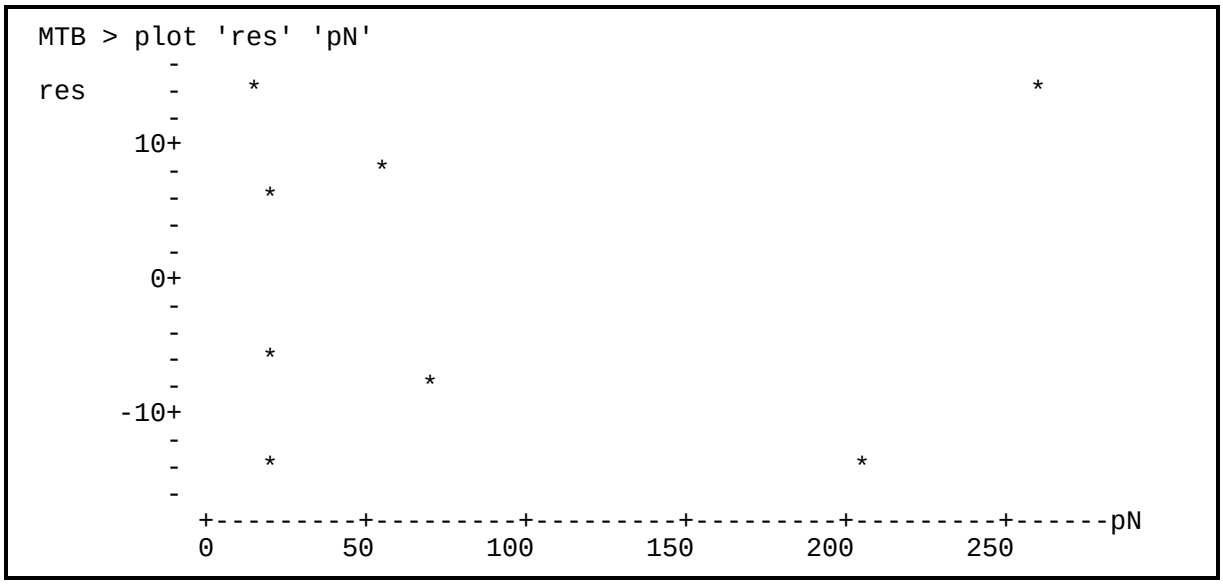
```
MTB > print c6 c7 c8

 ROW      f         pN        res

   1     29    22.295     6.7049
   2    273   258.623    14.3770
   3      8    21.738   -13.7377
   4     64    71.344    -7.3443
   5     11    17.705    -6.7049
   6    191   205.377   -14.3770
   7     31    17.262    13.7377
   8     64    56.656     7.3443
```

Print the 8 data equations.

Check assumptions
A Bowls or arches ? . . . . . Model does not contain lines, so no need to check.
B1 Residuals sum to zero ? . . . . . Yes
B2 Residuals independent ? . . . . . Yes

```
MTB > let c9= lag(c1)
MTB > corr c9 c8

Correlation of C9 and C8 = 0.045
```

B3 Dispersion of residuals around zero homogeneous across the plot ? . . . . . Yes



B4 Residuals normal ? . . . . . No

```
 MTB > hist 'res'

Histogram of res   N = 8

Midpoint   Count
     -15       2  **
     -10       0
      -5       2  **
       0       0
       5       2  **
      10       0
      15       2  **
```

We do not expect the residuals to be normal, because the data are counts.

A plot shows that the residuals, as expected, are not normal.

The residuals do not follow a normal distribution (B4). They are independent (B2) and identically distributed (B3), or "iid" for short.

The model is f = pN + error

This is a general**ized** linear model because it employs a **non-normal** error structure.
The error structure in this case arises from a binomial response variable (red or not red)
To test whether the data (f) fit the expected values (pN) we are going to compute a G-statistic, an overall measure of the goodness of fit of observed to expected values.
Because the errors are iid, we can use a Chisquare distribution to compute a p-value from the G-statistic.

```
MTB > let c9 = 'f'*log('f'/'pN')
MTB > let c9 = 2*c9
MTB > name c9 '2lnL'
MTB > print c6 c7 c8 c9

 ROW      f         pN        res       2lnL

   1     29    22.295     6.7049    15.2499
   2    273   258.623    14.3770    29.5389
   3      8    21.738   -13.7377   -15.9937
   4     64    71.344    -7.3443   -13.9051
   5     11    17.705    -6.7049   -10.4708
   6    191   205.377   -14.3770   -27.7233
   7     31    17.262    13.7377    36.2987
   8     64    56.656     7.3443    15.6019
```

Compute log likelihood ratios from residuals and fitted values.
The smaller the log likelihood ratio, the better the fit.

Compute G-statistic = 2 @ sum of the log likelihood ratios.

```
MTB > let k2 = sum('2lnL')
MTB > print k2
K2       28.5964
```

The larger the G-statistic, the poorer the fit of the data to the model (expected value).
Is the deviation of the data (f) from the expected value (pN) due to chance alone?

Compute p-value for the G-statistic.
The p-value is computed for three degrees of freedom. df_{row} @ df_{col} = 1@3.

```
MTB > cdf 28.5964;
SUBC> chisquare 3.

 28.5964     1.0000
```

cdf reports the proportion of outcomes smaller than the observed outcome (G-statistic = 28.6)
In this case the proportion is nearly 100% (to 4 decimal places).
The p-value is thus reported with 4 decimal places: $p < 0.0001$

Poisson Response Variable (Donax.out)

Shell colour data from *Bulletin of Marine Science* 32: 343.

```
MTB > print c1 c2

 ROW     C1    C2

   1     24     4
   2    118    35
   3     90    38
   4    139    40
```

Natural selection on a polymorphic bivalve *Donax variabilis*.
Predated in C2, unpredated in C1

Dark
Rays
Tinge
White

```
MTB > stack c1 c2 [into] c3
MTB > set into c4
DATA> (371 117)4
MTB > end
MTB > let c5 = c1 + c2
MTB > print c1-c5

 ROW     C1    C2     C3     C4     C5
   1     24     4     24    371     28
   2    118    35    118    371    153
   3     90    38     90    371    128
   4    139    40    139    371    179
   5                   4    117     28
   6                  35    117    153
   7                  38    117    128
   8                  40    117    179
```

For each cell (C3) compute expected value (C7) and log likelihood (C8), based on column totals (C4) and row totals (C5).

```
MTB > let c6 = (c4/488)*(c5/488)
MTB > let c7 = 488*c6
MTB > name c3 'f' c4 'coltot' c5 'rowtot'
MTB > name c6 'p_hat' c7 'f_hat'
MTB > let c8 = 2*('f'*loge('f'/'f_hat'))
MTB > name c8 '2lnL'
```

```
MTB > print c3-c8
 ROW      f  coltot  rowtot      p_hat     f_hat      2lnL
   1     24     371      28   0.043621    21.287    5.7582
   2    118     371     153   0.238356   116.318    3.3890
   3     90     371     128   0.199409    97.311  -14.0593
   4    139     371     179   0.278861   136.084    5.8940
   5      4     117      28   0.013756     6.713   -4.1421
   6     35     117     153   0.075169    36.682   -3.2864
   7     38     117     128   0.062886    30.689   16.2410
   8     40     117     179   0.087943    42.916   -5.6292
```

Compute goodness of fit statistic G = 2 Σ2 ln L, twice the sum of log likelihoods.

```
MTB > let k1 = sum(c8)
MTB > print k1
K1       4.16515
MTB > cdf 4.16515;
SUBC> chisquare 3.
    4.1652    0.7558
MTB > stop
```

p = 1 ! 0.77558 = 0.224 > " = 5%

No significant difference in proportions between predated and unpredated *D. variabilis*.

The null hypothesis H_o was accepted so turn to consideration of Type II error, that of missing a real difference.

The largest difference between observed and expected was 10 bivalves, in the category 'tinge.' This difference becomes significant at " = 5% if it rises to 15 rather than 10, or 15/117 = 13% of the collection.

Conclusion: Selection differential was less than 13%.
We can be sure that selection was less than 13%, given the sample size we were able to obtain.

Correlation (srbx15_7.out)

Thorax length data from Box 15.7 in Sokal and Rohlf (1995), p 594.
Total length of 15 aphid stem mothers and the mean thorax length of their parthenogenetic offspring.

```
MTB > read 'a:srbx15_7.dat' c1 c2;
SUBC> nobs = 15.
     15 ROWS READ

MTB > name c1 'ltot' c2 'thor'
MTB > plot c2 c1
          -
      6.40+                                                       *
          -                                                   *
thor      -         *                 *      *                    *
          -                           *                 *
          -               *                        * *
          -
      5.60+                     *      *
          -
          -
          -
          -
      4.80+
          -     *
          -
          -
          -       *
      4.00+
            --+---------+---------+---------+---------+---------+---------+----ltot
            6.0       7.2       8.4       9.6      10.8      12.0

MTB > plot c1 c2
     12.0+                                                                 *
          -                                                    *
ltot      -                                                          *
          -                                              *
          -
      10.0+                                        2
          -
          -                                          *
          -                                        *   *
          -                                    *
       8.0+                                 *
          -
          -                                     *
          -
          - *             *                                    *
       6.0+
          -
            ------+---------+---------+---------+---------+---------+thor
               4.40      4.80      5.20      5.60      6.00      6.40
```

Judging from these graphs, a linear model of association did not look acceptable. The following models were then investigated by transforming one or both variables, plotting, and examining the plot to see if it was linear (no bowls or arches).

ltot	log(lthor)
log(lot)	lthor
log(ltot)	log(lthor)
ltot	1/lthor
ltot	$lthor^3$

The last two were a slight improvement over the first three, but none of the plots could be viewed as linear.

Next, try a model based on monotonic relation: thorax length increases monotonically with total length. That is, variables are associated on a rank scale.

```
MTB > rank c1 c3
MTB > rank c2 c4
MTB > name c3 'Rltot'
MTB > name c4 'Rthor'

MTB > plot c3 c4
          -
     15.0+                                                          *
          -                                              *
Rltot     -                                                    *
          -                              *
          -                         *
     10.0+                          *
          -                                   *
          -                                *
          -                                         *
          -              *
      5.0+          *
          -                    *
          -                                              *
          -  *
          -      *
      0.0+
          ------+---------+---------+---------+---------+---------+Rthor
              2.5       5.0       7.5      10.0      12.5      15.0

 MTB > corr c3 c4

Correlation of Rltot and Rthor = 0.649
```

This is called the Spearman Rank correlation coefficient. It is a measure of monotonic relation. It measures the linear relation between the **ranks** of the variables.

How does this measure of monotonic association compare with a measure of linear association?

```
MTB > corr c1 c2 m1
Correlation of ltot and lthor = 0.650
```

This is the Pearson correlation, a measure of the linear association between the variables. In this example, the measure of linear association turns out to be the same as the measure of monotonic association.

So far 6 different models have been tried, none could be considered acceptable, based on lack of bowls or arches in the residuals (deviations from line), as judged by eye. Perhaps the problem is that the data are heterogeneous. There appears to be a positive relation, but some of the data points do not conform to this relation. In particular, it seems that any thorax length is possible at low total lengths (ltot < 7 micrometer units). Let's assume that something different is happening at low total lengths, and just examine the relation between variables when ltot > 7 micrometer units.

```
MTB > let c1(5) = 0/0
MTB > let c1(5) = 0/0
                  J
*** VALUES OUT OF BOUNDS DURING OPERATION AT J

MTB > let c1(8) = 0/0
MTB > let c1(9) = 0/0
MTB > plot c1 c2
ltot    -
        -                                                          *
        -                                  *
    11.2+                                                  *
        -
        -                                *
        -              *
        -              *
     9.6+
        -                                     *
        -
        -           *                  *           *
        -
     8.0+
        -   *
        -             *
        -
          ------+---------+---------+---------+---------+---------+lthor
             5.60      5.76      5.92      6.08      6.24      6.40

         N* = 3
```

This looks acceptably linear.

Now compute Pearson correlation, placing the coefficient into k1 for later use.

```
MTB > corr c1 c2 m1

Correlation of ltot and lthor = 0.664

MTB > copy m1 c3 c4
MTB > let k1 = c3(2)
MTB > print k1
K1       0.663741
```

Next compute t-statistic, with H_o that the true correlation is zero.

```
MTB > let k2 = k1*sqrt((12-2)/(1-k1**2))
MTB > print k2
K2       2.80620
```

Compute p-value from cumulative distribution function, for t distribution.

```
MTB > cdf k2;
SUBC> t 10.
    2.8062    0.9907
MTB > let k3 = (1-.9907)*2
MTB > print k3
K3       0.0186000
```

Note multiplication by 2, the cumulative distribution function yields proportion of outcomes smaller than t = 2.8062, which comes to 99.07% of the outcomes.
The right tail is thus approximately 1 - 0.9907 = 0.93% and both tails together comes to approximately 1.8% (p = 0.0186 exactly).

Summary.
For non-linear (monotonic) model, use ranks. Compute rank correlation.
For linear model (relation described by straight line) use Pearson correlation.

Multivariate Analysis -- References

Cooley, W. W. and P. R. Lohnes (1971). *Multivariate Data Analysis*. Wiley & Sons, New York.

Gittens, R. Canonical Analysis. *Biomathematics* **12**. Springer-Verlag, Berlin.

Ludwig, J. A. and J. F. Reynolds (1988). *Statistical Ecology*. Wiley & Sons, New York.

Kim, J. and C. W. Mueller (1978). *Introduction to Factor Analysis. What it is and How to do it*. Sage Publications, London.

Morrison, D. F. (1976). *Multivariate Statistical Methods*. McGraw-Hill, New York.

Pielou, E. C. (1984). *The Interpretation of Ecological Data*. Wiley & Sons, New York.

Seal, H. L. (1964). *Multivariate Statistical Analysis for Biologists*. Methuen, London.

Van de Geer, J. P. (1971). *Introduction to Multivariate Analysis for the Social Sciences*. W. H. Freeman, San Francisco.

Most statistical packages (such as SAS, BMDP, SYSTAT, SPSS) include references.

There are aspects of statistics other than its being intellectually difficult that are barriers to learning. For one thing, statistics does not benefit from a glamorous image that motivates students to persist through tedious and frustrating lessons….there are no TV dramas with a good-looking statistician playing the lead, and few mother's chests swell with pride as they introduce their son or daughter as "the statistician."
C.T. Le and J.R. Boen. 1995. *Health and Numbers: Basic Statistical Methods*. Wiley.

Autocorrelated Data -- References

Box, G. E. P. and G. H. Jenkins (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco.

<the basic text in time series analysis>

Cressie, N. A. C. (1991). *Statistics for Spatial Data*. John Wiley, New York

<extensive treatment of topic, fairly mathematical>

Diggle, P. J. (1983). *Statistical Analysis of Spatial Point Patterns*. Academic Press, London.

<somewhat mathematical, emphasizes use of randomization tests>

Griffith, D. A. (1987). *Spatial Autocorrelation*. Resource Publications in Geography, American Society of Geographers.

<accessible treatment with examples>

Platt, T. and K. L. Denman (1975). Spectral analysis in ecology. *Annual Review of Ecology and Systematics* **6**: 189-210.

<reviews one technique: analysis in the frequency domain>

Ripley, B. D. (1981). *Spatial Statistics*. Academic Press, London.

<comprehensive coverage of topics, fairly mathematical>

Upton, G. J. and B. Fingleton (1985). *Spatial Data Analysis by Example*. Vol. I. Point Pattern and Quantitative Data. John Wiley & Sons, Chichester.

<highly accessible because of examples; short on conceptual linkages>

Most statistical packages (such as SAS, BMDP, SYSTAT, SPSS) include references.

GLM: Autocorrelated Data (codacf.out)

Cod (*Gadus morhua*) catch data.

Catches from the northwest Atlantic, NAFO division 2J3KL are divided into Canadian offshore, other offshore, and inshore.

$Total_{offshore}$ = Other + $Can_{offshore}$. Catches in tonnes = 10^3 kg.

```
MTB > read 'a:cod.dat' c1-c4;
SUBC> nobs = 30.
MTB > let c5 = c3 - c2
MTB > name c1 'yr' c2 'other' c3 'totoff' c4 'inshore' c5 'canoff'
MTB > plot c4 c1
  160000+      * *
        -
inshore -           *
        -           *
        -              *
  120000+         *
        -                * *                         *
        -                   * *                        *       *
        -                       *                 *      *
        -                                       *
   80000+                        *             *    *     *   *
        -                                    *              *
        -                          * *
        -                                  *
        -
   40000+                             *   *
        -                               *
          +---------+---------+---------+---------+---------+------yr
      1956.0    1962.0    1968.0    1974.0    1980.0    1986.0
```

Are the inshore catches serially correlated?

```
MTB > acf c4
 ACF of inshore
          -1.0 -0.8 -0.6 -0.4 -0.2  0.0  0.2  0.4  0.6  0.8  1.0
            +----+----+----+----+----+----+----+----+----+----+
  1   0.816                          XXXXXXXXXXXXXXXXXXXXX
  2   0.636                          XXXXXXXXXXXXXXXXX
  3   0.537                          XXXXXXXXXXXXXX
  4   0.401                          XXXXXXXXXXX
  5   0.222                          XXXXXXX
  6   0.074                          XXX
  7  -0.069                        XXX
  8  -0.170                      XXXXX
  9  -0.245                    XXXXXXX
 10  -0.299                   XXXXXXXX
 11  -0.360                 XXXXXXXXXX
 12  -0.360                 XXXXXXXXXX
 13  -0.343                 XXXXXXXXXX
 14  -0.335                  XXXXXXXXX
 15  -0.293                   XXXXXXXX
```

Yes. Inshore catches are strongly correlated. r = +0.816 at lag of 1 year. This means that if catches are high in one year, they will be high the year before or the year after. Catches negatively correlated at lag of 11 years (r = ! 0.36).

What is best model to describe the relation? The two choices are moving average and autoregressive. Moving average means that catch in any one year depends on combined effects of several previous years. Autoregressive means that catch in any one year is related primarily to effects during a fixed time previously.

The shape of the autocorrelation function suggests that this catch is best described as moving average. Check this by computing the partial autocorrelation with PACF command

```
MTB > pacf c4
PACF of inshore
          -1.0 -0.8 -0.6 -0.4 -0.2  0.0  0.2  0.4  0.6  0.8  1.0
            +----+----+----+----+----+----+----+----+----+----+----+
  1   0.816                          XXXXXXXXXXXXXXXXXXXX
  2  -0.089                        XXX
  3   0.134                          XXXX
  4  -0.183                      XXXXXX
  5  -0.183                      XXXXXX
  6  -0.082                        XXX
  7  -0.160                       XXXXX
  8   0.028                           XX
  9  -0.052                         XX
 10  -0.010                          X
 11  -0.131                       XXXX
 12   0.057                          XX
 13  -0.063                        XXX
 14  -0.054                         XX
 15   0.047                           XX
```

The shape of the partial autocorrelation function also indicates that catch is related to several prior years (moving average) rather than to year at fixed time in past.

Conclusions:

Inshore catches strongly autocorrelated.

A moving average model is best guess for a statistical model.

Next Analysis: Can inshore catches be predicted from offshore catches?

```
MTB > regress c4 1 c5;
SUBC> residuals c8.

The regression equation is
inshore = 95000 - 0.028 canoff

Predictor       Coef       Stdev    t-ratio        p
Constant       95000        7851      12.10    0.000
canoff       -0.0285      0.1338      -0.21    0.833

s = 32914       R-sq = 0.2%      R-sq(adj) = 0.0%

Analysis of Variance

SOURCE       DF          SS           MS         F        p
Regression    1    49014084    49014084      0.05    0.833
Error        28 30333534208  1083340544
Total        29 30382548992

Unusual Observations
Obs.  canoff   inshore       Fit Stdev.Fit  Residual   St.Resid
  1     4515    159492     94871      7477     64621      2.02R
R denotes an obs. with a large st. resid.
```

Is this model acceptable? Check assumption A, linear relation.

```
MTB > plot c8 c5

C8      -
        -
        -   **
   50000+   *
        -   * *
        -
        -   *2                                  *
        -      *   *                                2          *
       0+         *             *
        -           *          *         *                   *  *
        -    *  *                                              *
        -  **  *
        -
  -50000+  2
        - *
        -
        -
          +---------+---------+---------+---------+---------+------canoff
          0     25000     50000     75000    100000    125000
```

No bowls or arches, so linear model acceptable.

Next, investigate the assumptions concerning errors.

B1 sum(errors) = 0 ? Yes, because least squares used in regression.

B2 errors independent ?
The catches are strongly autocorrelated, so residuals are also likely to be autocorrelated. If the residuals are autocorrelated, then p-values based on this model will be in error because the residuals won't be independent.

```
MTB > acf c8                                                  are residuals autocorrelated?
   ACF of C8
          -1.0 -0.8 -0.6 -0.4 -0.2  0.0  0.2  0.4  0.6  0.8  1.0
            +----+----+----+----+----+----+----+----+----+----+
  1   0.815                           XXXXXXXXXXXXXXXXXXXX
  2   0.636                           XXXXXXXXXXXXXXXX
  3   0.536                           XXXXXXXXXXXXX
  4   0.400                           XXXXXXXXXX
  5   0.218                           XXXXX
  6   0.067                           XX
  7  -0.082                         XXX
  8  -0.185                       XXXXX
  9  -0.262                     XXXXXXX
 10  -0.318                    XXXXXXXX
 11  -0.381                  XXXXXXXXXX
 12  -0.381                  XXXXXXXXXX
 13  -0.362                   XXXXXXXXX
 14  -0.351                   XXXXXXXXX
 15  -0.303                    XXXXXXXX
```

The residuals are not independent. p-value cannot be trusted.

```
MTB > differences 1 c4 c6
MTB > name c6 'inshd1'
MTB > print c4 c6

 ROW  inshore  inshd1

   1   159492       *          16    35181   -6467
   2   157286   -2206          17    41213    6032
   3   119363  -37923          18    59939   18726
   4   138511   19148          19    72623   12684
   5   144548    6037          20    81455    8832
   6   131328  -13220          21    85822    4367
   7   110527  -20801          22    96523   10701
   8   110843     316          23    80038  -16485
   9   101859   -8984          24   113049   33011
  10   101037    -822          25   106423   -6626
  11    97224   -3813          26    97721   -8702
  12    76588  -20636          27    79883  -17838
  13    62539  -14049          28    72369   -7514
  14    62052    -487          29    78747    6378
  15    41648  -20404          30   101925   23178
```

To solve the problem take the differences from one year to the next, in the response variable (inshore catch).

Taking the difference usually reduces the autocorrelation.

To check this, examine autocorrelation of the differenced variable.

```
MTB > acf c6
ACF of inshd1
          -1.0 -0.8 -0.6 -0.4 -0.2  0.0  0.2  0.4  0.6  0.8  1.0
            +----+----+----+----+----+----+----+----+----+----+----+
  1   0.006                          X
  2  -0.003                          X
  3  -0.048                         XX
  4   0.099                          XXX
  5  -0.034                         XX
  6   0.171                          XXXXX
  7  -0.164                       XXXXX
  8  -0.061                         XXX
  9  -0.081                         XXX
 10   0.064                           XXX
 11  -0.072                         XXX
 12   0.066                           XXX
 13   0.058                           XX
 14   0.037                           XX
 15  -0.152                       XXXXX
```

```
MTB > pacf c6
PACF of inshd1
          -1.0 -0.8 -0.6 -0.4 -0.2  0.0  0.2  0.4  0.6  0.8  1.0
            +----+----+----+----+----+----+----+----+----+----+----+
  1   0.006                          X
  2  -0.003                          X
  3  -0.048                         XX
  4   0.100                          XXX
  5  -0.036                         XX
  6   0.172                          XXXXX
  7  -0.168                       XXXXX
  8  -0.063                         XXX
  9  -0.065                         XXX
 10   0.021                           XX
 11  -0.039                         XX
 12   0.042                           XX
 13   0.129                           XXXX
 14   0.014                           X
 15  -0.144                       XXXXX
```

Autocorrelation in response variable is usually reduced by taking differences.

Now examine whether **change** in the inshore catch (inshore catch after differencing) is related to offshore catch.

```
MTB > regress c6 1 c5;
SUBC> residuals c9.

The regression equation is    inshd1 = - 4333 + 0.0603 canoff

29 cases used 1 cases contain missing values           (1956 lost from analysis)

Predictor       Coef       Stdev    t-ratio        p
Constant       -4333        3798      -1.14    0.264
canoff       0.06033     0.06364       0.95    0.352

s = 15509       R-sq = 3.2%      R-sq(adj) = 0.0%

Analysis of Variance
SOURCE       DF          SS          MS         F        p
Regression    1   216159680   216159680      0.90    0.352
Error        27  6493937152   240516192
Total        28  6710096896

Unusual Observations
Obs.  canoff    inshd1       Fit Stdev.Fit  Residual   St.Resid
  3     4676    -37923     -4051      3611    -33872     -2.25R
 24    94457     33011      1366      4559     31645      2.13R
```

Check the residuals for autocorrelation.

```
MTB > acf c9
ACF of C9
          -1.0 -0.8 -0.6 -0.4 -0.2  0.0  0.2  0.4  0.6  0.8  1.0
           +----+----+----+----+----+----+----+----+----+----+----+
  1  -0.002                          X
  2   0.001                          X
  3  -0.070                        XXX
  4   0.051                          XX
  5  -0.103                       XXXX
  6   0.095                          XXX
  7  -0.224                    XXXXXXX
  8  -0.130                       XXXX
  9  -0.132                       XXXX
 10   0.031                          XX
 11  -0.090                        XXX
 12   0.077                          XXX
 13   0.095                          XXX
 14   0.094                          XXX
 15  -0.094                        XXX
```

Residuals no longer autocorrelated for new model (based on differencing)

Conclusion: When we remove the autocorrelation present in the inshore catch series, we find that the inshore catches are not related to offshore catches.

Numerical Methods. Finding the sample size (srex9_6.out)

Exercise 9.6 from Sokal and Rohlf (1995), page 268

What sample size should be used to be 80% certain of observing a true difference between two means as small as a tenth of a millimeter, at the 5% level of significance?

First compute the error Mean square = 0.2496

This is better estimate than total variance = 25.6819/99 = 0.2594

```
MTB > read 'srex9_5.dat' c1-c5;
SUBC> nobs=20.
MTB > stack c1-c5 c6;
SUBC> subscripts c7.
MTB > name c6 'b_lngth' c7 'gr'
MTB > anova c6 = c7

Analysis of Variance for b_lngth

Source      DF         SS         MS       F      P
gr           4     1.9734     0.4933    1.98  0.104
Error       95    23.7085     0.2496
Total       99    25.6819
```

n = unknown

F^2 estimated as $s^2 = 0.2496$ (see above)

* = 0.10 and $*^2 = 0.01$

< = a (n ! 1)

" = 5%

P = 80%

match cdf computations in Minitab to t-values for example in Box 9.14 page 263

$t_{0.05[<]}$ = 2.642 in text, for < = 4(20 ! 1) = 76

$t_{2(1-0.80)[<]}$ = 0.847 in text, for < = 4(20 ! 1) = 76

```
MTB > invcdf .01;
SUBC> t 76.
    0.0100   -2.3764
MTB > invcdf .005;
SUBC> t 76.
    0.0050   -2.6421
MTB > invcdf .4;
SUBC> t 76.
    0.4000   -0.2542
MTB > invcdf .2;
SUBC> t 76.
    0.2000   -0.8464
```

use 0.005 and 0.20 for box 9.14

Use 0.005 and 0.20 for box 9.14 therefore use 0.025 and 0.20 for exercise 9.6

Compute k1 = $2(F/*)^2$

```
MTB > let k1 = 2*(0.2496)/(0.01)
```

Guess n = 20, hence < = 2*(20 ! 1) = 38

```
MTB > invcdf 0.025 k2;
SUBC> t 38.
MTB > invcdf 0.2 k3;
SUBC> t 38.
MTB > let k4 = k1*(k2 + k3)**2     ≤ n
MTB > print k1 k2 k3 k4
K1       49.9200
K2       -2.02439
K3       -0.851178
K4       412.782    ≤ n
```

t value stored into k2

t value stored into k3

≤ n in Box 9.14
Both t-values are negative, the sum becomes positive when squared.

```
MTB > invcdf 0.025 k2;
SUBC> t 822.
MTB > invcdf 0.2 k3;
SUBC> t 822.
MTB > let k4 = k1*(k2 + k3)**2
MTB > print k2 k3 k4
K2       -1.96285
K3       -0.842055
K4       392.745      ≤ n
```

Guess n = 412
hence < = 822

```
MTB > invcdf .025 k2;
SUBC> t 782.
MTB > invcdf .2 k3;
SUBC> t 782.
MTB > let k4 = k1*(k2 + k3)**2
MTB > print k4 k3 k2
K4       392.804       = n
K3       -0.842103
K2       -1.96301
MTB > stop
```

Guess n = 392
hence < = 782

No change from last iteration

Sample size is n = 392 for stated power and Type I error (= 5%).