Permutation tests for linear models in R

Robert E. Wheeler

2016-07-30

Abstract

An R package which uses permutation tests to obtain p-values for linear models. Standard R linear model functions have been modified to produce p-values obtained from permutation tests instead of from normal theory. These values are advantageous when the degrees of freedom for error is small or non-existent, as is the case with saturated experimental designs, or when the data is drawn from a non-normal population, or when there are apparent outliers. The package also supports ANOVA for polynomial models such as those used for response surfaces.

Contents

1 Introduction

Permutations are fundamental to statistical inference. Consider a simple experiment in which three levels of potash are applied to plots and the numbers of lettuce plants that emerge are tallied, as in Table 1.

Table 1: Lettuce growth experiment

Potash level	1			2			3		
No. Plants	449	413	326	409	358	291	341	278	312

There seems to be a downward trend in the data with increasing levels of potash: but is it real? The conventional way of deciding this nowadays would be to assume the observations are normally distributed and drawn from an infinite population of possible replications of this experiment. The first assumption cannot be checked, and the second requires a good deal of fancy, which doesn't seem to bother modern researchers very much; possibly because they have become used to it. Any assumption that involves infinity requires careful thought, since it is quite outside ordinary experience.

In any case, all that is available is this set of 9 observations. If there is a trend, then one measure of it is the slope of a line fitted through the data. This is given by subtracting the average of the level 3 observations from the average of the level 1 observations and dividing by 2: it is abut 43. The null hypothesis is that the level of potash does not effect plant growth, which means that the observed value of 43 is due to chance. If it is due to chance, then there is no connection between the values in Table 1 and the plots from which they come. In other words the first value, 449, could as easily have come from some other plot, and so too for the other values: none of them are tied to the plots shown. That being the case, it is reasonable to ask how frequently chance would create a slope of 43 or greater. If 43 is quite common, then it seems unlikely that the trend is real. If on the other hand, a slope of 43 or larger is rare, then the null hypothesis is suspect.

One can in fact estimate this chance, by permuting the observations in Table 1, and tallying the number of times a slope equal to or greater than 43 is obtained. This value is a well defined probability, in the same sense that the probability of snake eyes from a pair of fair dice is 1/36. It turns out that the probability that a slope will be equal to or greater than 43 is 0.039[12], making it a rare event in most peoples view; and leading to the conclusion that potash decreases fertility for this planting[3].

The rub, of course, is that the conclusion cannot be generalized to the effects of potash on lettuce plants without further assumptions. It is a perfectly correct conclusion for these particular plants, and it seems reasonable that it should apply to other plantings, which cries out for a replication of the experiment. If the same conclusion is reached from a number of replicated experiments under a variety of conditions, then one would have

[1]This is a one tailed test, and the corresponding F-test probability is 0.048.

[2]In addition, the randomization is over all $9! = 362880$ permutations, instead of over the 1680 combinations obtainable by switching observations only between different levels.

[3]This conclusion is only valid if the hypothesis is posed before the data is observed: the calculations are meaningless if a salient feature of the data is taken as a hypothesis after the fact.

reason to believe it in general. Replication requires care. For example, there may be other factors that influence fertility which should be taken into account, such as the soil gradient or the unequal exposure to sunlight, and repeat experiments might well show significant results due to inattention to these factors in the absence of a genuine trend.

Fechner [1860] ran up against such difficulties in establishing a just-noticeable difference for sensory measurement. He presented boxes with various weights to his subjects and recorded the point at which they were unable to make a judgment. He found it necessary to control for many extraneous factors such as the order of presentation and the hand that was used. It took many trials and considerable care to obtain his results.

Peirce and Jastrow [1884] repeated Fechner's work, but had a wonderful idea: instead of controlling the many factors, Peirce used a randomizing device which avoided many of the difficulties that Fechner had encountered. Any factor that might influence the results of the weighings could be expected to line up with the effect under investigation only by chance, making the results perhaps more variable, but more nearly correct. The just-noticeable difference of Fechner, thus became the point at which the probabilities of right and wrong judgments were equal.

Moreover, Peirce realized that this device enabled generalizations to be made:

> "The truth is that induction is reasoning from a sample taken at random to the whole lot sampled. A sample is a *random* one, provided it is drawn by such machinery, artificial or physiological, that in the long run any one individual of the whole lot would get taken as often as any other." [Peirce and Jastrow, 1884, p217].

This device of randomization was adopted by Fisher [1935] as a way to generalize the results from a particular experiment[4]. Fisher [1925-1952] then invented the idea of a permutation test, and provided justification for its use. Because of the computational difficulties, approximations such as the chi-squared distribution were used, and over time these approximations have replaced permutations as the preferred methodology: see [Fisher, 1935, p55]. What is now called the F distribution, in fact, was originally devised as an approximation for the permutation distribution of the variance ratio – see [Kempthorne, 1952, section 7.4]. Computers now make it possible to consider a direct use of permutation tests.

Thus permutation tests applied to suitably randomized experimental units offer a valid method of induction. The randomization is essential. The use of statistical tests, and in particular, the use of permutation tests for non-randomized units changes the inferences that are possible from the general to the particular. In other words the inferences are proper only for the units involved.

Scheffé [1959] has given a concise definition of permutation tests:

> "Permutation tests for a hypothesis exist whenever the joint distribution of the observations under the hypothesis has a certain kind of symmetry, namely,

[4]He must surely have been aware of Peirce's work, although he did not cite it.

> when there exists a set of permutations of the observations which leave the distribution the same (the distributions is invariant under a group of permutations)."

In other words, it must be possible under the hypothesis to "exchange" the observations, which occurs for linear models when the hypothesis is the usual null hypothesis and when the units have been selected at random from some specific population. If the null is true, then the observed sum of squares, SS, has the same distribution for all permutations of the observations, and a tally of the number of values of the SS which exceed that for the original ordering of the observations forms a critical region for the permutation test. The size of this region on a proportion scale is the p-value of the permutation test.

There is always the question of choosing the permutation group. For a single variable, one of course permutes all observations. For two variables in a table, one may permute each row independently of the row totals, but what about permuting all variables regardless of their row? An even more difficult decision is what to do about interactions. [Edgington, 1995, p133] argues for a very restricted definition which excludes may effects that are usually of interest to experimenters. The fact is, however, that all estimates are linear functions of the observations, as is discussed in Section 4.2, and the coefficients of these functionals depend only on the design. Each coefficient estimate is a function of all the observations, and thus permutation over observations is meaningful. An exception occurs when blocks need to be considered, since the linear functionals for such analyses are defined only over a subset of the observations: **R** deals with this by projecting the design and the observations into spaces orthogonal to the blocks, and permutation analyses seems to work well on these projections.

Of course, permutation tests do not assume a particular distribution and are more powerful in many cases such as for a mixture of distributions or distributions which depart substantially from the normal distribution, or when there are outliers. Simulations illustrating this are shown in Section 3.3

Permutation tests are clearly the method of choice for those vexing cases where there are no degrees of freedom for error such as for saturated experimental designs, as is illustrated in Section 2.6.

In those cases where the normal theory assumptions are adequately approximated, permutation tests are indistinguishable from the usual F-tests. Section 3.2 shows simulations illustrating this. In those cases where the p-values from permutation tests differ substantially from those for F-tests, a careful examination of the data is usually worthwhile.

2 Examples

This section illustrates the several functions in the lmPerm package with a dataset from [Cochran and Cox, 1957, p164].

The dataset is shown in Table(2) is a 3x3 factorial with 9 observations. The y values

are numbers of lettuce plants emerging, averaged over 12 plots. The factors are 3 levels of nitrogen, N, and 3 levels of potash, P. (The `Block` factor is not part of Cochran and Cox's data set: it will be used for a later illustration.) There are no degrees of freedom for error.

Cochran and Cox analyzed this data with an external estimate of the residual standard error. Their analysis indicated that both linear effects were significant.

Table 2: CC164, A 3x3 factorial

	y	P	N	Block
1	449	1	1	0
2	413	1	2	2
3	326	1	3	1
4	409	2	1	1
5	358	2	2	0
6	291	2	3	2
7	341	3	1	2
8	278	3	2	1
9	312	3	3	0

2.1 *lmp()* exact

The appropriate R function for an analysis of such a data set is *lm()*, but although it will estimate the coefficients of the linear model, it will not produce p-values because of the lack of an error estimate. The modified function is *lmp()*, and its output is shown in Table(3). As may be seen, the linear effects are not quite significant at the 5% level[5]. This suggests that Cochran and Cox's historical value may have been a tad small. Since there are no residuals the permutation criterion is the unscaled sum of squares, which does not provide as powerful an analysis as the scaled sum of squares as noted in Section 3.2.

The call for this analysis is

```
summary(lmp(y~P*N,data=CC164, perm="Exact"))
```

as indicated at the top of Table(3), which differs only in the `perm` parameter from the call that would be made to *lm()*. The `"Exact"`[6] argument to `perm` causes all `9!=362,880` permutations to be evaluated. The computer response time is not noticeably longer than for *lm()*, which is not the case for larger data sets. In fact 12 or so observations is near the limit of my patience at 3 minutes, while 14 is an overnighter. Is it any wonder, with computation times like these, that permutations are seldomly used? Before computers, only the simplest permutation calculations were possible, and although things are better nowadays, permutations are still only possible for small data sets. The statistical problems were resolved by the use of t and F distributions as approximations.

[5]This is a two tailed test, as are the tests in *lmp()* and `aovp()`.

[6]The `"Exact"` argument is redundant here, since it is the default.

Table 3: Exact permutation analysis of lettuce data

```
[1] "Settings:  unique SS "

Call:
lmp(formula = y ~ P * N, data = CC164, perm = "Exact")

Residuals:
ALL 9 residuals are 0: no residual degrees of freedom!

Coefficients:
        Estimate Pr(Exact)
P.L      -60.575    0.0786 .
P.Q        0.408    1.0000
N.L      -63.640    0.0643 .
N.Q        4.082    0.8929
P.L:N.L   47.000    0.4656
P.Q:N.L   24.249    0.7075
P.L:N.Q   42.724    0.5052
P.Q:N.Q   13.000    0.8524
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: NaN on 0 degrees of freedom
Multiple R-Squared:     1,        Adjusted R-squared:   NaN
F-statistic:   NaN on 8 and 0 DF,  p-value: NA
```

2.2 *lmp()* Prob

The alternative to evaluating all permutations is to sample from the possible permutations and to use estimates of p-values. Two methods for doing this are in the present package. The first uses a criterion suggested by Anscombe [1953] which stops the sampling when the estimated standard deviation of the p-value falls below some fraction of the estimated p-value. The second uses the sequential probability ratio of Wald [1947] to decide between two hypotheses about a p-value. There are of course other stopping rules.

Anscombe's method is controlled by setting `perm` to `"Prob"`. Thus one has the results shown in Table(4).

Table 4: Estimated permutation analysis of lettuce data

```
[1] "Settings:  unique SS "

Call:
lmp(formula = y ~ P * N, data = CC164, perm = "Prob")

Residuals:
ALL 9 residuals are 0: no residual degrees of freedom!

Coefficients:
        Estimate Iter Pr(Prob)
P.L     -60.5755  822   0.1095
P.Q       0.4082   51   1.0000
N.L     -63.6396 2265   0.0424 *
N.Q       4.0825   51   1.0000
P.L:N.L  47.0000  192   0.3438
P.Q:N.L  24.2487   51   0.7059
P.L:N.Q  42.7239  132   0.4318
P.Q:N.Q  13.0000   51   0.9020
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: NaN on 0 degrees of freedom
Multiple R-Squared:     1,        Adjusted R-squared:   NaN
F-statistic:   NaN on 8 and 0 DF,  p-value: NA
```

Note that `Pr(Exact)` has changed to `Pr(Prob)`. There is good agreement between the p-values in the two tables, but sampling being what it is, one must not expect certainty. The stopping rule is controlled by a parameter `Ca` which has a default value of 0.1. That is the sampling stops when the estimated standard deviation falls below 0.1 of the estimated p-value.

The `Iter` column reports the number of iterations (the sample size) required to meet the stopping rule – the minimum is set to 50. The number of iterations is a very small fraction of the `9!` possible permutations, so small in fact that had previous generations of statisticians explored this possibility they may well have used permutations more frequently. One only has to recall the still unequaled tables produced by Pearson [1933,1956] to realize that extensive computation was no hindrance to their efforts. Indeed groups of "computers" with mechanical calculators were employed in massive routine calculations, such as inverting matrices, up through the 1960's, when punched card calculators and the first computers became available.

2.3 *lmp()* SPR

Permutation calculations were first used to assess significance and their justification was as randomization tests Fisher [1935], that is tests which derive their validity from the randomization of the experimental data. They are not well suited to decision theory, and yet decisions theory can be used to assess p-values. If one chooses two hypotheses about the p-values, then one can use a sequential probability ratio test Wald [1947] to decide on when to stop the sampling. Table(5) illustrates this. It was obtained by setting the `perm` parameter to `"SPR"`. Acceptance of the null hypotheses is shown by 1's in the `Accept` column[7].

2.4 *lmp()* ANOVA

Analysis of variance tables may be produced. For "flat" data, one can use the call

```
anova(lmp(y~P*N,data=CC164))
```

with the result shown in Table(6).

2.5 *aovp()* Multistratum analyses

One may perform a multistratum analyses with a call to *aovp()* as shown in Table(7). The call was

```
summary(aovp(y~P*N+Error(Block),CC164)).
```

[7]For this illustration, the size of the acceptance region was set to 0.07 instead of the default 0.05, to insure that a 1 would appear in the `Accept` column.

Table 5: SPR permutation analysis of lettuce data

```
Call:
lmp(formula = y ~ P * N, data = CC164, perm = "SPR", p0 = 0.07,
    p1 = 0.08)

Residuals:
ALL 9 residuals are 0: no residual degrees of freedom!

Coefficients:
          Estimate       Iter    Pr(SPR) Accept
P.L      -60.57548 4206.00000    0.08250      0
P.Q        0.40825   35.00000    1.00000      0
N.L      -63.63961 1681.00000    0.05592      1
N.Q        4.08248   47.00000    0.76596      0
P.L:N.L   47.00000  126.00000    0.33333      0
P.Q:N.L   24.24871   40.00000    0.87500      0
P.L:N.Q   42.72392   78.00000    0.48718      0
P.Q:N.Q   13.00000   35.00000    1.00000      0

Residual standard error: NaN on 0 degrees of freedom
Multiple R-Squared:     1, Adjusted R-squared:   NaN
F-statistic:   NaN on 8 and 0 DF,  p-value: NA
```

Table 6: Anova permutation analysis of lettuce data

```
[1] "Settings:  unique SS "
Analysis of Variance Table

Response: y
          Df R Sum Sq R Mean Sq Pr(Exact)
P          2  11008.7    5504.3    0.2214
N          2  12200.0    6100.0    0.1893
P:N        4   4791.3    1197.8    0.8913
Residuals  0      0.0       NaN
```

The `Block` variable is fictitious, and is introduced just for this illustration. It has no meaning for the experiment.

The probability for the block stratum is unity, since the test is for a pooling of all components; and in any case there are only two degrees of freedom to permute.

An example with 7 blocks and 6 treatments taken from [Hald, 1952, Table 17.4], may be used to illustrate this point. An analysis similar to that of 7 is shown in Table 8.

The data seems to show a linear trend in the block means, so an additional variable, `L` was created as a linear contrast among blocks. The analysis is shown in Table 9, where it may be seen that there is a statistically significant linear trend.

2.6 Saturated designs

Saturated designs are experimental designs with no degrees of freedom for error, such as the main effect plans of Plackett and Burman [1946]. They are usually two level designs such as that shown in Table 10. This particular design and its analysis was discussed by Box [1988] in his critique of Taguchi metods.

Since there are no degrees of freedom for error in such designs, various techniques are used. The usual technique is to leave some of the factors unassigned, and use the pooled estimates from these to estimate error. Simulations show that a sharp bend in the power curve at 5 degrees of freedom which means that only only needs a few unassigned factors when using such designs.

In this case the original author assigned all factors, and chose to pool the smaller effect estimates for an estimate of error. This obviously biased the error estimate and produced several spurious effects as pointed out by Box [1988]. He analyzed the data using half-normal plots, and found only two significant effects, "E" and "G". Table 11 shows the *lmp()* analysis using permutations: it agrees with Box's analysis, as it should.

Table 7: Multistratum Anova permutation analysis of lettuce data

```
[1] "Settings:  unique SS "

Error: Block
Component 1 :
    Df R Sum Sq R Mean Sq Pr(Exact)
P:N  2   1970.7    985.33         1


Error: Within
Component 1 :
    Df R Sum Sq R Mean Sq Pr(Exact)
P    2  11008.7    5504.3   0.64167
N    2  12200.0    6100.0   0.08611 .
P:N  2   2820.7    1410.3   0.81944
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 8: Anova with a block stratum

```
> summary(aovp(Y~T+Error(block),Hald17.4))

[1] "Settings:  unique SS "

Error: block
Component 1 :
          Df R Sum Sq R Mean Sq
Residuals  6   1483.2     247.2


Error: Within
Component 1 :
          Df R Sum Sq R Mean Sq Iter Pr(Prob)
T          4   137.54    34.386 5000   0.0134 *
Residuals 24   236.60     9.858
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 9: Anova with a block stratum and a linear contrast among blocks

```
[1] "Settings:  unique SS "

Error: block
Component 1 :
          Df R Sum Sq R Mean Sq Pr(Exact)
L          1   1331.8   1331.79  0.001389 **
Residuals  5    151.4     30.28
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Error: Within
Component 1 :
          Df R Sum Sq R Mean Sq Iter Pr(Prob)
T          4   137.54    34.386 5000   0.0206 *
Residuals 24   236.60     9.858
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 10: Saturated experimental design with response SN

H	D	L	B	J	F	N	A	I	E	M	C	K	G	O	SN
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	6.2626
1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	4.8024
-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	21.0375
1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	15.1074
-1	-1	-1	1	1	1	1	-1	-1	-1	-1	1	1	1	1	14.0285
1	-1	1	1	-1	1	-1	-1	1	-1	1	1	-1	1	-1	16.6857
-1	1	1	1	1	-1	-1	-1	-1	1	1	1	1	-1	-1	12.9115
1	1	-1	1	-1	-1	1	-1	1	1	-1	1	-1	-1	1	15.0446
-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	1	1	17.6700
1	-1	1	-1	1	-1	1	1	-1	1	-1	1	-1	1	-1	17.2700
-1	1	1	-1	-1	1	1	1	1	-1	-1	1	1	-1	-1	6.8183
1	1	-1	-1	1	1	-1	1	-1	-1	1	1	-1	-1	1	5.4325
-1	-1	-1	1	1	1	1	1	1	1	1	-1	-1	-1	-1	15.2724
1	-1	1	1	-1	1	-1	1	-1	1	-1	-1	1	-1	1	11.1976
-1	1	1	1	1	-1	-1	1	1	-1	-1	-1	-1	1	1	9.2436
1	1	-1	1	-1	-1	1	1	-1	-1	1	-1	1	1	-1	4.6836

Table 11: *lmp()* analysis of a saturated design.

```
[1] "Settings:  unique SS "

Call:
lmp(formula = SN ~ ., data = Quinlan)

Residuals:
ALL 16 residuals are 0: no residual degrees of freedom!

Coefficients:
   Estimate Iter Pr(Prob)
H1   0.8138   95   0.5158
D1   0.8069   51   0.7255
L1  -0.4041   75   0.5733
B1  -0.2917   51   0.9216
J1   0.3332   51   0.9608
F1  -1.1057  205   0.3317
N1  -0.2779   51   0.8039
A1   1.1433  127   0.4409
I1  -0.4888   51   0.6667
E1  -3.5971 5000   0.0054 **
M1  -0.2202   51   0.7843
C1  -1.1409   99   0.5051
K1   1.1894  206   0.3301
G1  -2.3740 1236   0.0752 .
O1  -0.2153   51   0.8039
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: NaN on 0 degrees of freedom
Multiple R-Squared:     1,        Adjusted R-squared:   NaN
F-statistic:   NaN on 15 and 0 DF,  p-value: NA
```

2.7 Overfitting

The response in some "experiments" is almost error free with the fluctuations due only to the slight variations in the measuring process. An analysis assuming normal errors can be misleading in such cases. The problem is more common in industrial settings than it is for sociological or biological work. For example some data cited by [Faraway, 2005, page 190] involved measurements on thermoplastic composite strength subjected to several levels of laser power and tape speed. There were no replicate measurements and the design is saturated if an interaction is included. Table 12 shows a permutation analysis, and it may be seen from the mean squares that an analysis using F-tests, with the interaction as the error term, would show both factors to be significant. It is more likely that the data is almost without error, and the permutation results are the correct ones.

2.8 Polynomial Models

Models for response surfaces use polynomial models such as

$$E(Y) = \beta_0 + \sum_i \beta_i x_i + \sum_{ij} \beta_{ij} x_i x_j,$$

where the $\{x_i\}$ are numerical variables.

R has little support for such models. In particular, it is not possible to perform an ANOVA with them: each column of the incidence matrix is treated independently, and there is no way to pool them into sources, even though the several columns involving the same variable may represent vectors spanning a space.

`lmPerm` collects together the appropriate terms in response surface models and produces a correct ANOVA table. For efficient experimental designs, such an analysis will be meaningful, especially if the variables are centered. Table 13 illustrates this. The first two variables, A and B, are numeric, and the third, C, is a factor.

The function `poly.formula()` enables a few special functions, such as `quad()` to be included in the formula. The following special functions are available.

```
> poly.formula(Y~quad(A,B,C))

Y ~ (A + B + C)^2 + I(A^2) + I(B^2) + I(C^2)

> poly.formula(Y~cubic(A,B,C))

Y ~ (A + B + C)^3 + I(A^2) + I(B^2) + I(C^2) + I(A^3) + I(B^3) +
    I(C^3)

> poly.formula(Y~cubicS(A,B,C))

Y ~ (A + B + C)^3 + I(A * B * (A - B)) + I(A * C * (A - C)) +
    I(B * C * (B - C))
```

Table 12: Permutation analysis of tape composite data

```
> data(composite)

> anova(lmp(strength~laser*tape,composite))

[1] "Settings:  unique SS "
Analysis of Variance Table

Response: strength
           Df R Sum Sq R Mean Sq Pr(Exact)
laser       2  224.184   112.092  0.003571 **
tape        2   48.919    24.459  0.557143
laser:tape  4   10.503     2.626  0.990079
Residuals   0    0.000       NaN
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 13: Polynomial model in two variables and a factor

```
> anova(lmp(poly.formula(Y~quad(A,B)+C),simDesignPartNumeric))

[1] "Settings:  unique SS : numeric variables centered"
Analysis of Variance Table

Response: Y
          Df R Sum Sq R Mean Sq Iter Pr(Prob)
A          2  2391.02   1195.51 5000   0.0190 *
B          2   459.41    229.70  381   0.2598
A:B        1    77.80     77.80   99   0.5051
C          2   111.67     55.84   98   0.6224
Residuals  6   829.87    138.31
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.9 Multiple responses

The lhs of a formula may reference a matrix in the calling environment, and `lm()` will produce an analysis for each column as a response variable. `aov()` will produce a multivariate analysis. Both `lmp()` and `aovp()` produce analyses for each column as the response variable. Many datasets contain variables for both the lhs and rhs of the formula, and it is often convenient to specify the lhs using variables from the argument `data`. The function `multResp()` may be used to do this. The dataset `Plasma` contains three dependent variables, and Table 14 shows the analysis.

2.10 Mixture experiments

Continuous variables may represent mixtures, as for example, mixtures of components in a paint. In these, the variables are constrained to sum to a constant, usually unity. An example of a design for three mixture components is shown in Table (15).

Because of the constraint, ordinary polynomial models will have redundant terms. This may be dealt with by appending non-estimable constraints to the design, or by reformulating polynomial models to account for the constraint. The constraint method may be useful in the analysis in those cases where it is desired to interpret the coefficients. Cox [1971] treats this problem. Cox's results enable models to be build containing both mixture and non-mixture variables. At the present time neither `lmp()` nor `aovp()` support such constraints. For models containing only mixture variables, one may use models that have been given by Scheffé [1958], and elaborated upon by Gorman and Hinman [1962]. The Scheffé models for three variables are shown in Table (16). Note that the constant term is omitted from these models, which among other things, means that they are unaffected by block effects.

An analysis of the experiment in Table (15) is shown in Table (17). In this case, as is common for mixture experiments, there is no error in the data, and the analysis is performed only to create a prediction equation.

2.11 ANOVA types

An ANOVA table summarizes the information about sources in a linear model. Sources are of course groups of terms, such as the contrasts associated with a factor. For normal theory, the likelihood ratio test of the null hypothesis that the source is without effect is obtained as the difference of two residual sums of squares (SS). For example, the difference in the two SS from the following models.

$$\begin{aligned} y &= \mu + \alpha + \quad + \epsilon \\ y &= \mu + \alpha + \beta + \epsilon \end{aligned}$$

If the design is balanced the SS obtained from successively testing all sources add up to the total SS for a model with the constant as its only term. A balanced design is one in

Table 14: Multiple response analysis

```
> data(Plasma)

> anova(lmp(multResp(Amin,Pct,sinpoly)~.,Plasma))

[1] "Settings:  unique SS : numeric variables centered"
Analysis of Variance Table

Response: Amin
          Df R Sum Sq R Mean Sq Iter Pr(Prob)
W          1    17332     17332  849   0.1060
mTorr      1    47423     47423 5000   0.0144 *
cm         1   393343    393343 5000   <2e-16 ***
sccm       1     2260      2260  120   0.4583
Residuals  6    27379      4563
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Analysis of Variance Table

Response: Pct
          Df R Sum Sq R Mean Sq Iter Pr(Prob)
W          1   1.2202    1.2202  453  0.18102
mTorr      1   3.3014    3.3014 1431  0.06569 .
cm         1   3.9105    3.9105 2522  0.03846 *
sccm       1   0.5036    0.5036  122  0.45082
Residuals  6   3.6763    0.6127
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Analysis of Variance Table

Response: sinpoly
          Df R Sum Sq R Mean Sq Iter Pr(Prob)
W          1 0.027588  0.027588 5000  0.01920 *
mTorr      1 0.019333  0.019333 1888  0.05032 .
cm         1 0.297234  0.297234 5000  < 2e-16 ***
sccm       1 0.000283  0.000283   51  0.90196
Residuals  6 0.017287  0.002881
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 15: ghoctane: An octane-blending experiment

```
> data(ghoctane)
```

	X1	X2	X3	ON
1	1.000	0.000	0.000	100.8
2	0.000	1.000	0.000	85.2
3	0.000	0.000	1.000	86.0
4	0.500	0.500	0.000	88.8
5	0.500	0.000	0.500	90.3
6	0.000	0.500	0.500	85.5
7	0.333	0.333	0.333	88.3
8	0.150	0.595	0.255	86.6
9	0.300	0.490	0.210	87.6

Table 16: Scheffé models

linear	$X_1 + X_2 + X_3$
quadratic	$X_1 + X_2 + X_3 + X_1X_2 + X_1X_3 + X_2X_3$
special cubic	$X_1 + X_2 + X_3 + X_1X_2 + X_1X_3 + X_2X_3 + X_1X_2X_3$
cubic	$X_1 + X_2 + X_3 + X_1X_2 + X_1X_2 + X_2X_3 +$ $X_1X_2(X_1 - X_2) + X_1X_3(X_1 - X_3) + X_2X_3(X_2 - X_3)$
	R models
linear	`X1+X1+X3 -1`
quadratic	`(X1+X2+X3)^2 -1`
special cubic	`(X1+X2+X3)^3 -1`
cubic	`poly.formula(Y~cubicS(X1,X2,X3) -1)`

which the mean centered columns of the incidence matrix for the sources are orthogonal to each other. If the design is not balanced, the SS do not add up. One can of course apply ANOVA to any design, balanced or not, but it is seldom appropriate to do this for severely unbalanced data since the statistical tests will not be independent of the other sources in the model. A controversy exists about the ordering of sources in the model; in particular, there are different opinions about whether or not it is appropriate to test a main effect source when its interaction is included in the model. In other words, does the difference in SS for the following models provide an appropriate test? Here ϕ is the interaction between the other two sources. This problem is discussed in Section 4.2.

$$\begin{aligned} y &= \mu + \alpha + \quad + \phi + \epsilon \\ y &= \mu + \alpha + \beta + \phi + \epsilon \end{aligned}$$

In this section we will discuss an unbalanced dataset and analyze it in two ways.

Table 18 shows an unbalanced dataset reproduced in [Scheffé, 1959, p140]. The data shows the average weight of litters of rats of four genotypes reared by mothers of four genotypes. Is it the litter genotype or the mother genotype that matters most?

Table 17: Octane blending analysis

```
> anova(lmp(ON~.^3-1,ghoctane))

[1] "Settings:  unique SS "
Analysis of Variance Table

Response: ON
          Df R Sum Sq R Mean Sq Pr(Exact)
X1         1  10169.7   10169.7   0.05075 .
X2         1   7422.5    7422.5   0.05144 .
X1:X2      1     12.0      12.0   0.05568 .
X3         1   7410.1    7410.1   0.05145 .
X1:X3      1      6.5       6.5   0.06520 .
X2:X3      1      0.0       0.0   0.76062
X1:X2:X3   1      0.9       0.9   0.09946 .
Residuals  2      0.1       0.0
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 18: Rat genotype data

	Mother			
Litter	A	F	I	J
A	61.5	55.0	52.5	42.0
A	68.2	42.0	61.8	54.0
A	64.0	60.2	49.5	61.0
A	65.0		52.7	48.2
A	59.7			39.6
F	60.3	50.8	56.5	51.3
F	51.7	64.7	59.0	40.5
F	49.3	61.7	47.2	
F	48.0	64.0	53.0	
F		62.0		
I	37.0	56.3	39.7	50.0
I	36.3	69.8	46.0	43.8
I	68.0	67.0	61.3	54.5
I			55.3	
I			55.7	
J	59.0	59.5	45.2	44.8
J	57.4	52.8	57.0	51.5
J	54.0	56.0	61.4	53.0
J	47.0			42.0
J				54.0

The degree of unbalance in this data is not severe, as may be seen from the following. The columns of the incidence matrix are not orthogonal because of the unequal number of observations in the cells.

```
> data(ratGenotype)
> replications(~litter*mother,ratGenotype)

$litter
litter
 A  B  I  J
17 15 14 15

$mother
mother
 A  B  I  J
16 14 16 15

$`litter:mother`
      mother
litter A B I J
     A 5 3 4 5
     B 4 5 4 2
     I 3 3 5 3
     J 4 3 3 5
```

An analysis of this data is shown in Table 19. This analysis is sequential and tests each source conditional on the preceding sources in the table. It shows that the rearing is more important than the litter genotype.

A second analysis is shown in Table 20. This analysis makes unique tests on each source. That is the SS for each source is conditional on all other sources in the table. It reports essentially the same results, even though the SS are slightly different. (The parameter `seqs` is redundant since `aovp()` will perform a unique analysis by default.)

There may be a problem with unique tests for certain coding of the sources. Let C be a contrast matrix for a source, b, and suppose that there is a second factor with three levels. The columns of the incidence matrix for b will be:

$$\begin{bmatrix} C \\ C \\ C \end{bmatrix}$$

If the column sums of C are zero, then the cross product of the columns of b and the second factor will also be zero, and the two factors are said to be orthogonal; the SS for the two factors will be independent. This happens for a balanced design. If the design is unbalanced, some of the rows will be missing and the cross product will not be exactly

Table 19: Sequential Analysis of rat genotype data

```
> anova(lmp(wt~litter*mother,ratGenotype,seqs=TRUE))

[1] "Settings:  sequential SS "
Analysis of Variance Table

Response: wt
              Df R Sum Sq R Mean Sq Iter Pr(Prob)
litter         3    60.16    20.052  303   0.7393
mother         3   775.08   258.360 5000   0.0042 **
litter:mother  9   824.07    91.564 2923   0.1187
Residuals     45  2440.82    54.240
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 20: Unique Analysis of rat genotype data

```
> anova(lmp(wt~litter*mother,ratGenotype,seqs=FALSE))

[1] "Settings:  unique SS "
Analysis of Variance Table

Response: wt
              Df R Sum Sq R Mean Sq Iter Pr(Prob)
litter         3    27.66     9.219   66   1.0000
mother         3   671.74   223.913 5000   0.0038 **
litter:mother  9   824.07    91.564 2503   0.1674
Residuals     45  2440.82    54.240
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 21: Unique Analysis of rat genotype data using a contrast with non-zero sums.

```
> anova(lmp(wt~litter*mother,ratGenotype,contrasts=list(mother=contr.treatment(4))))

[1] "Settings:  unique SS "
Analysis of Variance Table

Response: wt
          Df R Sum Sq R Mean Sq Iter Pr(Prob)
mother2   12   1599.2    133.26 5000   0.0072 **
Residuals 45   2440.8     54.24
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

zero. Most of the time the columns sums are near enough to zero, so that this does not matter. On the other hand, if the column sums of C are quite different from zero, the departure from orthogonality can be considerable. Since an interaction is the arithmetic product of two factors, this will cause the interaction to be non-orthogonal to the factors, and the unique SS will reflect this because the factors are corrected for the interaction. Table 21 shows the consequence of using a `contr.treatment()` contrast, whose columns do not sum to zero. It should be compared with Table 20. The cross products in the interaction sum to zero for the `mother` columns because the `litter` contrasts sum to zero, but this does not happen for the `litter` columns. This changes the SS for `litter` since its columns are now dependent on the interaction.

3 Statistical Considerations

3.1 Randomized tests and Permutation tests

It is first of all best to be clear about the distinction between randomization tests and permutation tests. A randomization test uses permutations to obtain p-values and derives its legitimacy from the design of the experiment where the "information" is built in by randomizing the allocation of treatments to trials. A permutation test is simply the permutation part separated from the logical foundation of a randomization. It can be a valid method of inference but it is not based on randomization assignments of treatments to trials.

A leisurely exposition on randomization tests is given by Kempthorne and Doerfler [1969]. In simplest terms, one considers a set of experimental trials involving material (plots, machines, chemical vials, etc.) and treatments (hot-cold, fertilizer, additive, etc.) to be applied to the material. One usually attempts to make the material as homogeneous as possible, but there is a point of diminishing returns in this and so there will always be some variation in the material. Randomized assignment of treatments to material

prevents irregularities in the material from lining up with treatment applications, but this is not the logic for a randomized test. The logic is as follows. Imagine the experiment at its completion when all the measurements will have been made. Each item will have a response value. If the treatment had been the waving of a magic wand over some of the material, one would not expect this to have had any effect and the measurements would be exactly the same as if no wand had been waved over them. If, however, on inspection a "substantial portion" of those items that had been charmed turned out to have more extreme measurements than the uncharmed ones, then one would be compelled to concede that something very strange had happened. Either the wand had indeed had an effect, or the results might have occurred because of some unrecognized factor, such as trickery. There is no way to be sure. If, however, the charmed items had been selected at random, these difficulties would disappear, and like it or not, one would have to consider the wand as a possibility and rush out to repeat the experiment.

Judging the "substantial portion" is done by permuting the observations, which obtains its legitimacy from the randomization that was used to choose the items to be charmed. For if the magic wand had no effect, as we believed to start, then it can have nothing to do with the measurements that were made and they would be the same no matter which items were charmed. It is quite reasonable to ask, then, how the "substantial portion" observed fares when we examine the observations again and pretend that some other possible set of items were charmed. Is our "substantial portion" really large, or is it just largish. The only way to find out is to consider all the random allocations that might have been and to calculate the results using the observed data. Thus one permutes the data, calculates statistics, and tallies how many randomizations would have produced values as large or larger than that which was observed. The resulting tally divided by the total number of possible permutations is the p-value, a quite legitimate probability, and as good a measure of evidence as can be found. This probability assesses the information that was built into the experiment during its design by the use of a random mechanism.

On the other hand, a permutation test does not involve randomization of the material. If it is not simply a calculational exercise, it must be judged by other means; one of which is the selection of material and treatments from some population. If one has two populations, one of witches and one of muggles (about as realistic a pair of populations as many that are booted about), and if one selects a sample from each and sets them a task, then one may judge the efficacy of the these populations in performing this task with the aid of a permutation test. The null hypothesis is that witches don't exist, and those who claim to be are no more capable of producing charms than are muggles. Say the task is to curdle cream, One might pour bowls of cream for all participants, allow the witches to do their thing, and at some agreed upon time inspect the bowls for curdling. If upon examining the permutation p-value one found it exceeding the 5% level for the witches then the null would be rejected; but this p-value would depend for its validity on the random selection of individuals from the populations which stands in for the randomization of treatments used in a randomization test.

3.2 Comparing permutation and standard tests

There is no question about the use of permutation and randomization tests for paired samples or even for two samples, for surely these were the things used in the earliest discussions of permutation and randomization tests. Fisher's first example (Fisher [1935] used some data of Charles Darwin to illustrate his ideas and the data was in the form of matched pairs. This sort of experiment has been repeated many times and forms a staple in textbooks describing randomized tests. The fact that Darwin probably did not randomize his data is of no importance (an objection raised by some) because Fisher said "On the hypothesis that the two series of seeds are random samples from identical populations, and that their sites have been assigned to members of each pair independently at random ...": [Fisher, 1935, p44]. Fisher was illustrating a method and not performing an analysis with respect to the subject matter.

For more complicated designs, however, it is possible to randomize in different ways, and since the elements that one may exchange in a randomization (permutation) test are determined by the randomization it is possible to calculate slightly different p-values in different ways. [Edgington, 1995, p133] is adamant about the need to randomize among levels of one factor for each set of levels of the other factors: thus for a two-way table with factors A and B, he performs a randomization of A for each of the levels of B. This very conservative approach makes the treatment of multi-way tables or blocking factors difficult, and pretty nearly eliminates any methodologies for unbalanced designs.

[Manly, 1998, p130] has studied this problem and presents simulations which indicate that there is little to choose between several methods for main effects when power is considered. In particular, the simulation power values are about the same when Edgington's rules are used as when all observations are randomized. Randomization over all values is of course the easier method. In addition, Manly's simulations show that randomization over the residuals is a viable alternative: i.e., for a two-way table with data values x_{ijk}, where i and j index the two variables, and k is the replication index, the residuals are $r_{ijk} = x_{ijk} - x_{i..} - x_{.j.} + x_{...}$, where dots indicate averaging; thus randomizing the residuals makes the test conditional on the observed marginal means.

This can be carried further by considering the projection of the observations into subspaces, as is done with *aov()* for error subspaces. The *aov()* function fits the model described in the `Error` part of the formula and uses the Q matrix of the qr decomposition to project both the response and the model into orthogonal subspaces in which the other terms in the formula are fit. Thus if `Error` describes a block variable, then the formula terms are fit to data orthogonal to blocks, which amounts to correcting the data within each block for the block mean. The residuals described in the previous paragraph are dependent since they must sum to zero, and it might be argued that one might well consider permuting all but one of them. On the other hand, the orthogonal projection for `Error` describing a block reduces the number of elements by one in each block, and the permutation for all blocks involves only number of observations minus number of blocks.

In addition, the test statistic makes a difference. For analysis of variance the sum of squares, SS, for an effect is a natural choice, but analyses based on this statistic are not as powerful as those based on a sum of squares scaled by the residual sum of squares. The

reason is that when effects are present the overall level of the data changes, inflating the SS and making them dependent on this level. The effects of this are illustrated in Table 23. The functions *lmp()* and *aovp()* use the scaled SS by default.

The functions *lmp()* and *aovp()* are modifications of the R functions *lm()* and *aov()* that permute the projected response values and output the resulting p-values in place of normal theory values. The efficacy of this may be judged by contrasting the two. [Manly, 1998, p130] has done this for a number of procedures. A portion of his Table 7.6, filled out with simulations using *lmp()* and *aovp()*, is shown in Table 23.

Manly used a 24 observation data set involving two variables in a 4x2 design with 3 replicate observations per cell as shown in Table 22. He randomized the response 1000 times and applied each of the procedures to the resulting data. He tabulated the number of times each analysis produced a significant p-value at the 5% level – an estimate of the power of the test against the alternative hypothesis. Eight alternative hypotheses were generated by adding values to the main effects and interactions: these are indicated by the columns v1,v2,v3 in Table 22. The row labels of Table 23 indicate the v columns added to a randomization of the response values; the hi values are obtained by doubling the v values[8]. The "S" columns refer to tests based on SS, while the "F" columns refer to SS scaled by the residual SS.

The "F-distribution" columns represent the usual F-test values. The "lmp()" and "aovp()" columns were obtained by using these functions. The "aovp()" analysis had a block variable included to create two strata with perhaps different errors.

As may be seen from this table, there is little to choose between permutation tests and F-tests, even when the permutation tests are based on projections for blocking. The difficulties expressed by [Edgington, 1995, p133] do not seem to be of practical importance. The loss of power when unscaled SS are used should be noted.

3.3 Outliers and non-normal data

Linear models assuming normality are fairly insensitive to violation of the assumptions when the sample sizes are large. For small samples with few degrees of freedom for error, the size and power of the tests is degraded. Table 24 shows what happens when the data follows a gamma distribution. The values in this table were obtained from 1000 simulations for a 10 observation dataset comprising a single factor at two levels, drawing five observations from the null distribution with a shape parameter, α, of 0.5, and five observations from several alternatives with shape parameters varying from 0.01 to 1.00. The p-values and powers with respect to a 5% tests, are averaged over the 1000 simulations. The calculation assuming normal errors is clearly less satisfactory than the permutation calculation.

Unequal variances effect both normal-theory and permutation tests about the same, as may be seen in Table 25. This table was obtained from 1000 simulations for 10 observations with powers calculated for a 5% test.

[8]Manly's added values were designed to produce effects in each of the three spaces independently, but

Table 22: Data used by [Manly, 1998, p126] for power calculaitons

	month	sex	Y	v1	v2	v3
1	June	Male	13	100	0	0
2	June	Male	242	100	0	0
3	June	Male	105	100	0	0
4	June	Female	182	100	300	0
5	June	Female	21	100	300	0
6	June	Female	7	100	300	0
7	July	Male	8	200	0	0
8	July	Male	59	200	0	0
9	July	Male	20	200	0	0
10	July	Female	24	200	300	0
11	July	Female	312	200	300	0
12	July	Female	68	200	300	0
13	August	Male	515	300	0	0
14	August	Male	488	300	0	0
15	August	Male	88	300	0	0
16	August	Female	460	300	300	300
17	August	Female	1223	300	300	300
18	August	Female	990	300	300	300
19	September	Male	18	0	0	0
20	September	Male	44	0	0	0
21	September	Male	21	0	0	0
22	September	Female	140	0	300	300
23	September	Female	40	0	300	300
24	September	Female	27	0	300	300

unfortunately the additions to the interaction space are not orthogonal to the other spaces.

Table 23: Power simulations, after [Manly, 1998, p130]. p-values outside the range 3.6% to 6.4%, when true value is 5% are boldfaced.

	Unscaled SS						Scaled SS					
	lmp()			**F-distribution**			**lmp()**			**aovp()**		
Effects	S1	S2	S12	F1	F2	F12	F1	F2	F12	F1	F2	F12
1 None	5	7	5	4	5	5	5	5	6	5	5	5
2 Lo 1	23	4	**3**	20	5	5	24	6	7	30	5	5
3 Hi 1	75	**1**	**1**	71	4	4	71	5	4	90	6	5
4 Lo 2	**2**	57	**2**	4	56	4	4	58	5	5	74	5
5 Hi 2	**0**	100	**0**	5	100	4	4	100	4	4	100	5
6 Lo 1,2	16	55	**1**	21	53	4	20	56	5	33	63	5
7 Hi 1,2	30	99	**0**	73	100	3	70	100	4	90	100	6
8 Lo 1,2,12	10	86	2	51	90	11	29	89	10	42	97	16
9 Hi 1,2,12	12	100	0	100	100	32	91	100	34	99	100	53

Table 24: Power simulations when the data follows a gamma distribution

α_{null}	α_{alt}	F p-val	Perm p-val	F power	Perm power
0.5	0.01	0.11	0.03	0.32	0.91
0.5	0.05	0.17	0.12	0.24	0.63
0.5	0.10	0.24	0.21	0.18	0.42
0.5	0.50	0.47	0.51	0.02	0.05
0.5	1.00	0.37	0.38	0.11	0.16

Table 25: Power simulations when the variances are unequal

μ_{null}	σ_{null}	μ_{alt}	σ_{alt}	F p-val	Perm p-val	F power	Perm power
0	1	1.3	1	0.16	0.17	0.42	0.40
0	1	1.3	4	0.43	0.45	0.12	0.12
0	1	1.3	8	0.47	0.50	0.08	0.10
0	1	1.3	16	0.49	0.51	0.07	0.10

The exclusion of outliers is a vexed question for which there seems to be no general answer. In addition to the problems of choosing criteria for excluding obviously extreme values, there seem to be situations in which outliers become apparent only in multivariate situations, [Barnett and Lewis, 1978, p245]. Table 26 shows a power comparison for a mixture of normal distributions; the alternative distribution is a mixture of a $N(0, 1.3)$ distribution and with 10% outliers at the values given. The simulations are for 10 observations averaged over 1000 repetitions and the power is for a 5% test. As may be seen, modest outliers have little effect on either method, but larger outliers effect permutation calculations less than normal theory calculations.

Table 26: Power simulations when there are outliers

μ_{null}	μ_{alt}	μ_{out}	F p-val	Perm p-val	F power	Perm power
0	0	0	0.50	0.50	0.06	0.05
0	1.3	1	0.18	0.19	0.40	0.38
0	1.3	2	0.13	0.14	0.50	0.49
0	1.3	5	0.11	0.11	0.44	0.53
0	1.3	10	0.16	0.11	0.07	0.53

4 Technical Details

4.1 Permutations

Since permutations are computer intensive calculations, only the fastest algorithms should be used. To the best of my knowledge, the minimal effort general purpose permutation algorithm is pair exchange; that is successive item pairs are exchanged in a pattern that traverses all possible permutations. Of course special purpose routines can be written that take advantage of special characteristics of the data: see Baker [1995] for a survey of methodologies.

The idea employed here is to observe that if one has a set of permutations of n-1 elements which have been generated with minimal exchanges, then the set of n permutations obtained by adding a new element at each possible position is also a minimal exchange set, since each new permutation involves only a pairwise exchange of the new element with one of the old elements. The permutation {1,2,3} thus generates the permutations {1,2,3,4}, {1,2,4,3},{1,4,2,3},{4,1,2,3}, each requiring only one pair exchange. A scheme for doing the bookkeeping in this was given by Reingold et al. [1977]. This scheme is embodied in the c code `permute.c`, and made available in the R package through the R function `permute()`. A call to `permute()` produces a list containing pair indexes to be swapped. Using these indexes successively will traverse the possible permutations of a set of elements one pair-swap at a time.

The fact that all possible permutations are traversed by pairwise exchanges enables the fast generation of statistics calculated from permutations when the statistics are such that they may be updated by calculations involving only the exchanged elements. This is

the case when sums of squares are used:all that is involved is the backing out from the sum of squares the contributions from the two pairs, and the addition back in of the switched calculations. Since all permutations are generated, the distribution of the statistic is well calculated.

Pairwise exchanges are also used when the p-values are estimated by randomly sampling from all permutations. Of necessity, the first few permutations generated in this fashion are similar to the starting one, which raises the question of the representativeness of the sampling. Simulations indicate that starting pairwise permutations from the observed values is indeed a bad idea. The convergence to the correct value is very slow. One idea is to perform a complete randomization every so often, and indeed this works, but simply doing a complete randomization at the start seems to be just as good. Simulations show that the difference between randomizing every time, and randomizing once at the start and then using pair exchange, produce equally good estimates of the p-values. Besag and Clifford [1989] have studied this problem and provide justification for the procedure adopted here. The cautious user can control the randomization with the `nCycle` parameter.

4.2 Linear functionals

The usual statistics for linear models are derived from linear functionals of the observations; that is if y is the vector of observations, then all of the usual statistics are of the form $v'y$ for some vector v which does not depend on the observations. If $Ey = X\beta$, and if $QR = X$, where $Q'Q = I$, is the qr decomposition of X, then the least squares estimates of the coefficients are $\hat{\beta} = R^{-1}Q'y$ which are linear functionals of y. If the observations are uncorrelated with unit variances, then the covariance matrix of the least squares estimates is $(R'R)^{-1}$ The squares of the elements of the vector $Q'y$ are the sums of squares for the individual coefficients – the squares of linear functionals. If Q_b is a matrix of the columns of Q corresponding to a source, then the sum of squares for this source is given by summing the elements of $(Q_b'y)^2$: in other words by summing the squares of linear functionals of the y's. This is in fact the methodology used in the LINPACK and LAPACK routines called by *lm()* and *aov()* in *R*.

This methodology produces a sequential breakdown of the sums of squares, as pointed out in the *anova.lm()* writeup, in which the sum of squares for each source is dependent on sum of squares for the preceding sources. SAS calls this Type I ANOVA. There is no problem if the sources are orthogonal, since then the sums of squares are unaffected by the ordering; nor is there a problem when all sources are single degree of freedom sources, as are regression coefficients, since again the source ordering does not affect the sums of squares. The difficulty occurs for non-orthogonal (unbalanced) designs where the sources have more than one degree of freedom. In this case the order of the sources becomes important, since earlier sources constrain later sources.

There is no agreement on how to deal with this problem. One can formulate hypotheses about sources individually, which is usually acceptable if the sources are all main effects but difficulties clearly arise for interactions: should one test main effects when the interaction is significant? There is a considerable literature on the subject.

Part of the problem is due to a data mining mentality in which terms are added and subtracted to the model in the hope of achieving some goal. The fact that this invalidates the statistical tests seems to be overlooked even by those who should know better. For example it has been suggested that one ought to routinely omit high level terms unless they are significant[9].

ANOVA is most appropriately used for designed experiments, and for such experiments the model must be supplied in advance of the data collection – there is no room to fiddle with it after the fact. A computer program should provide significance tests for all sources. It is up to the experimenter to decide whether or not they are relevant.

A significant interaction means that there is a significant contrast in the interaction space, and almost always implies that the marginal sources are irrelevant. Sometimes this contrast is of little interest; for example, it might be due to a single anomalous cell which can be explained away. One should look at the interaction means to make a decision. On the other hand, a non-significant interaction means that the marginal sources are indeed of interest, and their statistical tests relevant.

The likelihood ratio test under normal theory is a test of two hypotheses, one involving a source and one not involving the source. The calculations that lead to the test statistic are made by placing the source of interest at the end of the model, which means that the source is conditional on the other sources. In common parlance it is "corrected" for all other sources, which seems wrong to many users, especially when main effects are "corrected" for interactions. The fact that single degree of freedom sources (regression coefficients) are tested by the same rubric is commonly ignored. This "correction" aspect is however an artifact of the calculation procedure and not the *raison d'être* of the test, which rather is a test of a hypothesis.

If one wants to test a main effect in a model with an interaction, and if one wants to use a likelihood ratio test, then the test is the test as stated, and it is equivalent to writing the model with the source as the last source. That this involves placing the interaction before the main effect, is quite irrelevant, since the test is the test is the test. `lmPerm` provides a "unique" test for all sources. How they are interpreted is a matter for the user, not the program.

For example Table 28 shows an analysis of Tab1 in Table 27. There is a non-significant interaction, but a significant main effect. Clearly the main effect is of interest and informative. Had the interaction been significant, however as is shown in Table 29 for Tab2, the main effect might be of less interest, although it clearly indicates that the first row in the table contains larger values on the average than the second. The statistical test might be ignored in this case. Part of the controversy has to do with the display of such information: the fear is that it will be misused by occasional statistical users. The fear is real, but twisting oneself into knots to avoid the problem doesn't seem to make sense. It might be better to abandon the traditional and very useful ANOVA table in favor of some display that shows the user Tab2, which makes the relationship very clear.

The technique used by `lmPerm` may be described by considering the model $y = X\beta + \epsilon$,

[9]Although this is impropper, one might implement some sort of regret criteria that would allow for data snooping.

Table 27: Two way tables

Tab1	B1	B2	B3	B.
A1	4.18	3.51	3.64	3.78
A2	0.74	0.59	0.58	0.64
A.	2.46	2.05	2.11	

Tab2	B1	B2	B3	B.
A1	4.18	3.51	0.62	2.77
A2	0.74	0.59	0.58	0.64
A.	2.46	2.05	0.6	

Table 28: Analysis of Tab1

Tab1	Df	R Sum Sq	R Mean Sq	Iter	Pr(Prob)	
A	1	10.6	10.6	5000.0	0.0188	*
B	1	0.5	0.5	51.0	0.8235	
A:B	1	0.1	0.1	51.0	0.6863	
Residuals	23	2.9	0.1			

where X is the incidence matrix of the design, β a vector of unknown parameters and ϵ a diagonal matrix of uncorrelated N(0,1) errors. The likelihood ratio test for source b (corresponding to a subset of the elements of β) contrasts the residuals from two models: H_0: model without source b; H_1: model with source b. The difference in the residual sums of squares $SS_b = SS_0 - SS_1$ is used as the test statistic: under normal theory it is distributed as a multiple of a chi-squared variable. This difference is a quadratic function involving the inverse, V_b^{-1}, of that part of the covariance matrix under H_1 corresponding to the source b: it is shown in the Appendix that $SS_b = \hat{\beta}_b^T V_b^{-1} \hat{\beta}_b$, where $\hat{\beta}_b$ is that portion of the coefficient vector under H_1 corresponding to b.

Permutation p-values are obtained from the sums of squares $SS_{bi} = \hat{\beta}_{bi}^T V \hat{\beta}_{bi} : i = 1 \ldots N$ where $\hat{\beta}_{bi}$ is obtained from the ith permutation of the y vector. If SS_{b0} is the sum of squares from the original ordering of the elements of y, then the p-values are given by $P = (\sum(SS_{bi}/SS_{ri} > SS_{b0}/SS_{r0}))/N$, where SS_{ri} and SS_{r0} are residual sums of squares. If N is the total number of permutations possible, then the p-values are exact, otherwise they are estimates: of course the scaling by the residual sums of squares is omitted for saturated models. Statistics other than sums of squares have merit, but are not implemented in these functions. It should be noted that scaling the sums of squares increases the power of the test: see Manly [1998].

Table 29: Analysis of Tab2

Tab2	Df	R Sum Sq	R Mean Sq	Iter	Pr(Prob)	
A	1	26.2	26.2	5000.0	0.0004	***
B	1	13.9	13.9	5000.0	0.0032	**
A:B	1	11.5	11.5	5000.0	0.0118	*
Residuals	23	5.2	0.2			

5 Derivation of $SS_b = \hat{\beta}_b^T V^{-1} \hat{\beta}_b$

5.1 Non-singular $X^T X$

Consider a partitioned incidence matrix $X = (X_1, X_2)$, and two hypotheses:

$$\begin{array}{rrcl} H_1: & y & \sim & N(X\beta_1, \sigma^2 I) \\ H_2: & y & \sim & N(X_1\beta_2, \sigma^2 I), \end{array}$$

with least squares estimates:

$$\begin{array}{rrcl} & y_1 & = & X_1\hat{\beta}_a + X_2\hat{\beta}_b \\ \text{and} & y_2 & = & X_1\hat{\beta}_2. \end{array}$$

The two least squares minimums are given by

$$\begin{array}{rcccl} L_b & = & (y-y_1)^T(y-y_1) & = & y^T y - y_1^T y_1 \\ L_a & = & (y-y_2)^T(y-y_2) & = & y^T y - y_2^T y_2, \end{array}$$

and the sum of squares of interest is

$$SS_b \;=\; L_a - L_b \;=\; y_1^T y_1 - y_2^T y_2.$$

We will transform X so that $y_1^T y_1$ is the sum of two parts, one of which is $y_2^T y_2$, and the other the desired quadratic function of $\hat{\beta}_b$.

Transform X into an orthogonal matrix Z by $Z = XT$, where

$$T \;=\; \begin{pmatrix} T_a & * \\ 0 & T_b \end{pmatrix} \qquad T^{-1} \;=\; \begin{pmatrix} T_a^{-1} & * \\ 0 & T_b^{-1} \end{pmatrix},$$

and the asterisks denote matrices of no interest.

Now

$$I = Z^T Z = T^T X^T X T, \tag{1}$$

so that

$$(X^T X)^{-1} \;=\; TT^T \;=\; \begin{pmatrix} * & * \\ * & T_b T_b^T \end{pmatrix} \;=\; \begin{pmatrix} * & * \\ * & V \end{pmatrix},$$

and $\sigma^2 V = Cov(\hat{\beta}_b)$.

Applying the transformation gives

$$\begin{array}{rcccccccl} y_1 & = & X\hat{\beta}_1 & = & XTT^{-1}\hat{\beta}_1 & = & Z\hat{\gamma}_1 & = & Z_1\hat{\gamma}_a + Z_2\hat{\gamma}_b \\ & & & & & & & = & y_a + y_b, \end{array}$$

and because $Z_1^T Z_2 = 0$, $\hat{\gamma}_1$ and $\hat{\gamma}_2$ are estimated independently, so

$$Z_1\hat{\gamma}_a = X_1 T_a \hat{\gamma}_a = X_1 T_a T_a^{-1}\hat{\beta}_2 = X_1\hat{\beta}_2 = y_2.$$

Since

$$y_a^T y_b = \gamma_a^T Z_1^T Z_2 \gamma_b = 0,$$

one has

$$y_1^T y_1 = y_a^T y_a + y_b^T y_b.$$

It follows that

$$\begin{array}{rcccccl} SS_b & = & y_1^T y_1 - y_2^T y_2 & = & y_a^T y_a + y_b^T y_b - y_2^T y_2 & = & y_b^T y_b \\ & = & \hat{\gamma}_b^T Z_2^T Z_2 \hat{\gamma}_b & = & \hat{\gamma}_b^T \hat{\gamma}_b & = & \hat{\beta}_b^T (T_b^{-1})^T T_b^{-1} \hat{\beta}_b \\ & & & & & = & \hat{\beta}_b^T (T_b T_b^T)^{-1} \hat{\beta}_b \\ & & & & & = & \hat{\beta}_b^T V^{-1} \hat{\beta}_b. \end{array}$$

5.2 Singular $X^T X$

The proof goes thru for singular $X^T X$ by choosing T_a, T_b such that instead of equation (1) one has $D^- = Z^T Z$, where D is diagonal with elements 0 and 1. Then TDT^T is a G inverse of $X^T X$ and $T_b D_2 T_b^T = V$ with $V^- = (T_b^T)^{-1} D^- T_b^{-1}$, hence $SS_b = \beta_b^T V^- \beta_b$ where β_b is the least squares estimate obatined by using TDT^T.

References

F.J. Anscombe. Sequential estimation. *J. R. Statist. Soc. B*, 15:1–29, 1953.

Rose D. Baker. Modern permutation test software. In E.G. Edgington, editor, *Randomization Tests*, chapter Appendix. Marcel Decker, 1995.

Vic Barnett and Toby Lewis. *Outliers in statistical data.* John Wiley and Sons, Inc., New York, 1978.

J. Besag and P. Clifford. Generalized monte carlo significance tests. *Biometrika*, 76(4): 633–642, 1989.

G. Box. Signal-to-noise ratios, performance criteria, and transformations. *Technometrics*, 30(1):1–17, 1988.

W.G. Cochran and G. M. Cox. *Experimental Designs.* John Wiley and Sons,Inc., New York, second edition, 1957.

D.R. Cox. A note on polynomial response functions for mixtures. *Biometrika*, 58(1): 155–159, 1971.

Eugene S. Edgington. *Randomization Tests.* Marcel Decker, New York, N.Y., 1995.

Julian Faraway. *Linear models with R.* Chapman and Hall, New York, 2005.

Gustav Theodor Fechner. *Elements of Psychophysics.* Holt, Rinehart and Winston, New York, N.Y., 1860. Translation of Elemente de Psychophysic, Vol1 by Helmut E. Adler.

R.A. Fisher. *Statistical Methods For Reserarch Workers.* Hafner, New York, N.Y, 1925-1952.

R.A. Fisher. *The Design Of Experiments.* Hafner, New York, N.Y, 1935.

J.W. Gorman and J.E. Hinman. Simplex lattice designs for multicomponent systems. *Technometrics*, 4(4):463–487, 1962.

A. Hald. *Statistical theory with engineering applications.* Wiley, New York, N.Y., 1952.

O. Kempthorne. *The design And Anaysis Of Experiments.* Wiley, New York, N.Y, 1952.

Oscar Kempthorne and T.E. Doerfler. The behaviour of some signifcance tests under experimental randomization. *Biometrika*, 56(2):231–248, 1969.

Bryan F.J. Manly. *Randomization, Bootstrap And MonteCarlo Methods In Biology, Second. Edition.* Chapman & Hall, New York, 1998.

Karl. Pearson. *Tables Of The Incomplete Beta-Function.* Cambridge Published for the Biometrika Trustees At the University Press, Cambridge, UK, 1933,1956.

Charles Peirce and Joseph Jastrow. On small differences of sensation. *Memoirs of the National Academy of Sciences for 1884*, 3:75–83, 1884.

R.L. Plackett and J.P. Burman. The design of optimal multifactorial experiments. *Biometrika*, 33:305–325, 1946.

E.M. Reingold, Jurg Nievergelt, and Narsingh Deo. *Combinatorial Algorithms Theory and Practice.* Prentice Hall, New Jersey, 1977.

H. Scheffé. Experiments with mixtures. *Jour. Roy. Statist. Soc (B)*, 20:344–360, 1958.

H. Scheffé. *The Analysis of Variance.* Wiley, New York, 1959.

A. Wald. *Sequential Analysis.* Wiley, New York,N.Y., 1947.