Introduction to Statistical Learning

Notes by Tim Brown, Yao-ban Chan and Owen Jones

Module 6: The Less than Full Rank Model and Experimental Design

Contents

1 Introduction, classification and factors

1.1 Introduction and one-way classification: factors - Faraway

The less than full rank model

In previous sections we used the linear model

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

in the knowledge (or assumption) that X, of dimension $n \times p$, is of full rank, i.e. $r(X) = p$.

This assumption is important because a full rank X implies that $X^T X$ is invertible, and therefore the normal equations

$$X^T X\mathbf{b} = X^T\mathbf{y}$$

have a unique solution.

The less than full rank model

Unfortunately, not all linear models fall into this category.

For example, consider the *the classification model with one factor having k levels*.

In this model, samples come from k distinct populations, with different characteristics. The *factor* is the variable that gives the population.

The term *factor* is the one used in R.

We wish to investigate the differences between these populations.

1.2 Example - one-way classification - MM5.1

One-way classification model

For example:

- A data scientist compares the performance of three different types of algorithms using different data sets;
- An economist studies the costs of four different approaches to developing new systems; or
- An engineer investigates the sulfur content in the five major coal seams in a particular geographic region.

One-way classification model

Let y_{ij} be the jth sample taken from the ith population. Then a natural model is

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij},$$

for $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, n_i$, where

- k is the number of populations/treatments;
- n_i is the number of samples from the ith population;
- μ is the overall mean;
- τ_i is a mean characteristic for population i.

One-way classification model

$$\begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{21} \\ y_{22} \\ \vdots \\ y_{k,n_k} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & \ldots & 0 \\ 1 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 1 & \ldots & 0 \\ 1 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & 0 & 0 & \ldots & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \vdots \\ \tau_k \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \vdots \\ \varepsilon_{k,n_k} \end{bmatrix}$$

$$\mathbf{y} \quad = \quad X \quad \boldsymbol{\beta} \quad + \quad \boldsymbol{\varepsilon}$$

The first column of X is the sum of the remaining columns, and therefore X is not of full rank.

One-way classification model

Example. Three different treatment methods for removing organic carbon from tar sand wastewater are compared: airflotation, foam separation, and ferric-chloride coagulation. A study is conducted and the amounts of carbon removed are:

AF	FS	FCC
34.6	38.8	26.7
35.1	39.0	26.7
35.3	40.1	27.0

One-way classification model

The linear model is

$$\begin{bmatrix} 34.6 \\ 35.1 \\ 35.3 \\ 38.8 \\ 39.0 \\ 40.1 \\ 26.7 \\ 26.7 \\ 27.0 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{31} \\ \varepsilon_{32} \\ \varepsilon_{33} \end{bmatrix}$$

$$\mathbf{y} \quad = \quad X \quad \boldsymbol{\beta} \quad + \quad \boldsymbol{\varepsilon}$$

The less than full rank model

The difficulty with a less than full rank model is that $X^T X$ is singular. This means that the normal equations do not have a unique solution.

However, the problem goes deeper than that: not only can we not estimate the parameters, but the parameters themselves are not well defined.

The less than full rank model

In a one-way classification model, the response variable from population i has a mean of $\mu + \tau_i$. Thus, for our carbon removal example we might have

$$\begin{aligned} \mu + \tau_1 &= 36 \\ \mu + \tau_2 &= 39 \\ \mu + \tau_3 &= 27. \end{aligned}$$

So our parameters might be $\mu = 34, \tau_1 = 2, \tau_2 = 5, \tau_3 = -7$.

However, we can also have $\mu = 30, \tau_1 = 6, \tau_2 = 9, \tau_3 = -3$.

In fact we can choose μ to be any real number, and still describe the system.

2 Reparameterisation

2.1 Reparameterisation - MM5.2

Reparametrization

Computer packages tackle the less than full rank model by converting it to a full rank model. We can then use all the machinery we have developed, in the knowledge that the least squares estimates and hypothesis tests have many desirable properties.

Example. Consider the one-way classification model with $k = 3$. The less than full rank model for this is

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij},$$

for $i = 1, 2, 3$, $j = 1, 2, \ldots, n_i$.

However, we can write the mean of each population as

$$\mu_i = \mu + \tau_i.$$

The variable i is the *factor*. In the carbon removal example, the factor is the different methods of removing the organic carbon.

Reparametrization

Then we can recast the model as

$$y_{ij} = \mu_i + \varepsilon_{ij},$$

with corresponding matrices

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}.$$

Reparametrization

The columns of X are now linearly independent, and so this is a full rank model that we can analyse. Simple matrix calculations give us

$$X^T X = \begin{bmatrix} n_1 & 0 & 0 \\ 0 & n_2 & 0 \\ 0 & 0 & n_3 \end{bmatrix}, \quad (X^T X)^{-1} = \begin{bmatrix} \frac{1}{n_1} & 0 & 0 \\ 0 & \frac{1}{n_2} & 0 \\ 0 & 0 & \frac{1}{n_3} \end{bmatrix}$$

$$X^T \mathbf{y} = \begin{bmatrix} \sum_{i=1}^{n_1} y_{1i} \\ \sum_{i=1}^{n_2} y_{2i} \\ \sum_{i=1}^{n_3} y_{3i} \end{bmatrix}, \quad \mathbf{b} = (X^T X)^{-1} X^T \mathbf{y} = \begin{bmatrix} \sum_{i=1}^{n_1} y_{1i}/n_1 \\ \sum_{i=1}^{n_2} y_{2i}/n_2 \\ \sum_{i=1}^{n_3} y_{3i}/n_3 \end{bmatrix}.$$

Reparametrization

Therefore, the least squares estimates for each of the population means are the means of the samples drawn from that population:

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}.$$

Linear functions of the parameters, of the form $\mathbf{t}^T\boldsymbol{\beta}$, are estimated using $\mathbf{t}^T\mathbf{b}$. For example, the function $\mu_1 - \mu_2$ is estimated by

$$\frac{1}{n_1}\sum_{i=1}^{n_1} y_{1i} - \frac{1}{n_2}\sum_{i=2}^{n_2} y_{2i}.$$

Reparametrization

The standard assumption that the errors are normally distributed with mean $\mathbf{0}$ and variance $\sigma^2 I$ is interpreted in this context to mean that *all* populations have a common variance σ^2 (but different means). The standard estimator for this variance is the residual sum of squares divided by the degrees of freedom:

$$s^2 = \frac{\mathbf{y}^T(I-H)\mathbf{y}}{n-p} = \frac{\mathbf{y}^T\mathbf{y} - \mathbf{y}^T X\mathbf{b}}{n-3},$$

where $H = X(X^TX)^{-1}X^T\mathbf{y}$ is the hat matrix for which $H\mathbf{y} = X\mathbf{b}$.

Reparametrization

That is

$$\begin{aligned}
& s^2 \\
&= \frac{1}{n-3}\left[\sum_{i=1}^{3}\sum_{j=1}^{n_i} y_{ij}^2 - \begin{bmatrix} \sum_{i=1}^{n_1} y_{1i} & \sum_{i=1}^{n_2} y_{2i} & \sum_{i=1}^{n_3} y_{3i} \end{bmatrix} \begin{bmatrix} \sum_{i=1}^{n_1} y_{1i}/n_1 \\ \sum_{i=1}^{n_2} y_{2i}/n_2 \\ \sum_{i=1}^{n_3} y_{3i}/n_3 \end{bmatrix}\right] \\
&= \frac{1}{n-3}\left[\sum_{i=1}^{3}\sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{3}\frac{1}{n_i}\left(\sum_{j=1}^{n_i} y_{ij}\right)^2\right] \\
&= \frac{1}{n-3}\sum_{i=1}^{3}\left[\sum_{j=1}^{n_i} y_{ij}^2 - \frac{1}{n_i}\left(\sum_{j=1}^{n_i} y_{ij}\right)^2\right].
\end{aligned}$$

Reparametrization

This can be written as a 'pooled' variance

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + (n_3-1)s_3^2}{(n_1-1)+(n_2-1)+(n_3-1)}$$

where s_i^2 are the individual population variance estimators

$$s_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}\left(y_{ij} - \frac{1}{n_i}\sum_{k=1}^{n_i} y_{ik}\right)^2.$$

The case of 2 populations was discussed in MAST90105 and our full rank linear model theory permits immediate generalisation to k populations.

Reparametrization

In general, it is always possible to re-parameterise a less than full rank model into a full rank model.

There are a number of different ways to do this and each way will generate different parameters with different estimates.

So it is important to have be precise about the way that this is done - we'll explore this through the lens of the practice in R.

Reparametrization

Example. Consider the *two-way* classification model (without interaction), with two levels of each of the two factors:

$$y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}, i, j = 1, 2.$$

The design matrix for this model is

$$X = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}.$$

Reparametrization

It is obvious that the first column is the sum of the next two columns, and also the sum of the 4th and 5th columns. Thus $r(X) = 3$.

This means that we have to reduce the rank of the matrix by 2.

This can be done by removing columns or by forming new columns which are linear combinations of the old ones.

But removing a parameter is the same as giving that column 0 weight and leaving the others at weight 1, so it is only necessary to consider reparameterising using linear combinations of the columns.

There analogue of the population means approach becomes more complicated so we'll approach this through a general framework.

2.2 Choosing parameters, contrast matrices and contrasts - Faraway 13.2

Framework for reparameterizing

Suppose there are n observations, p parameters, that the rank of the X matrix is $r < p$ and that $n \geq r$.

A linear combination of the columns is $X\mathbf{c}$, for some $p-$dimensional column vector $\mathbf{c}$.

Hence r linear combinations of the columns can be expressed as XC where C is a $p \times r$ matrix.

To assure that XC has the same column space as X, and therefore is full rank, it is assumed that C is of full rank, that is rank of $C = r$.

XC is of full rank?

Faraway (p.173) claims C, termed the *coding* or *contrast* matrix, can be chosen to be "anything that spans" r-dimensional space. This requires C just to be of full rank.

Unfortunately this can be misleading as the following example shows.

Suppose

$$X = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 1 & -1 \end{bmatrix} \quad C = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}.$$

Then

$$XC = \begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix},$$

which has rank 1, not $2 = r(X) = r(C)$.

XC is of full rank

Theorem 6.1. *Suppose X has r linearly independent columns and that the corresponding rows of C are also linearly independent. Then XC is full rank if, and only if, $I_r + DE$ is rank r where, if necessary by reordering the rows and columns of X and the rows of C, $X \,\&\, C$ have been partitioned as*

$$X = \left[\begin{array}{c|c} X_r & X_r D \\ \hline X^{n-r} & X^{n-r} D \end{array}\right] \quad C = \left[\begin{array}{c} C_r \\ \hline EC_r \end{array}\right],$$

X_r, X^{n-r}, D, C_r, E are respectively $r \times r, n-r \times r, r \times p-r, r \times r \,\&\, p-r \times r$ and X_r, C_r are both rank r.

You'll be led through the proof of this theorem as one of the exercises in Lab 8.

Reparameterisation through C

If $X \,\&\, C$ satisfy the conditions of the theorem, then to change parameters from $\boldsymbol{\beta}$ to $\boldsymbol{\gamma}$ and predictors from X to $Y = XC$ with $Y\boldsymbol{\gamma} = X\boldsymbol{\beta}$, the first r equations in the ordering specified in the theorem are

$$X_r(I_r + DE)C_r\boldsymbol{\gamma} = \left[\begin{array}{c|c} X_r & X_r D \end{array}\right]\boldsymbol{\beta}.$$

The matrix on the left is non-singular since it is the product of the three non-singular matrices $X_r, I_r + DE \,\&\, C_r$. Hence there is a unique solution for $\boldsymbol{\gamma}$:

$$\boldsymbol{\gamma} = (X_r(I_r + DE)C_r)^{-1}\left[\begin{array}{c|c} X_r & X_r D \end{array}\right]\boldsymbol{\beta}. \tag{1}$$

in terms of $\boldsymbol{\beta}$ once $X \,\&\, C$ have been partitioned.

Using the formula for the inverse of a product and factoring X_r gives:

$$\begin{aligned} \boldsymbol{\gamma} &= ((I_r + DE)C_r)^{-1}X_r^{-1}X_r\left[\begin{array}{c|c} I_r & D \end{array}\right]\boldsymbol{\beta}. \\ &= (I_r + DE)C_r)^{-1}(\boldsymbol{\beta}_r + D\boldsymbol{\beta}_{p-r}) \end{aligned} \tag{2}$$

where

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_r \\ \boldsymbol{\beta}_{p-r} \end{bmatrix}$$

and the dimensions of $\boldsymbol{\beta}_r, \boldsymbol{\beta}_{p-r}$ are $r \times 1, p-r \times 1$ respectively.

Different choices of C

For any less than full rank model, there will be a number of different ways to choose the matrix C and each choice will correspond to a different reparameterization.

Clearly, output from one of these choices will give different parameter estimates with different interpretations but:

Theorem 6.2. *Whatever the choice of the full rank matrix C in Theorem 6.1, the fitted values are the same, as are the residuals, residual sum of squares, regression sum of squares, and estimate of s^2.*

You'll be led through the proof of this theorem as one of the workshop in Lab 8.

Choices of C in R

The theorem allows us to use different choices of C, and the choice is one of convenience and interpretation in the context of the data.

The default option in R uses the command `contr.treatment`.

Consider a factor with 3 levels, so in this case $p = 4$ and $r = 3$.

By default, R includes an intercept term so the matrix C is the following prefaced by a column which is the first unit vector in 4 dimensions and topped by a row which is the transpose of the first unit vector in 3 dimensions, so that the XC will have the first column of X and otherwise not involve the intercept.

```
contr.treatment(3)
```

```
##   2 3
## 1 0 0
## 2 1 0
## 3 0 1
```

To simplify the presentation of our matrices, let $\mathbf{0}_m, \mathbf{1}_m$ denote column vectors of length m with entries $0, 1$ respectively. Let $\mathbf{e}_i$ be the unit column vector with 1 in the ith position and 0 elsewhere.

With parameter vector $\boldsymbol{\beta} = [\mu, \tau_1, \tau_2, \tau_3]^T$with $n_i, i = 1, \cdots, 3$ observations at level i, the less than full rank model has matrix:

$$X = \begin{bmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \mathbf{0}_{n_2} \\ \mathbf{1}_{n_3} & \mathbf{0}_{n_3} & \mathbf{0}_{n_3} & \mathbf{1}_{n_3} \end{bmatrix}.$$

The matrix C is:

$$C = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The effect of postmultiplying X by C is to form a new matrix the ith column of which is a linear combination of the columns of X whose weights are the ith column of C.

If a column of C is $\mathbf{e}_i$ for some i, then the corresponding linear combination of columns of X selects the ith column of X.

In this case, the choice in R given above selects the first, third and fourth columns of X.

Hence

$$Y = XC = \begin{bmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{1}_{n_2} & \mathbf{0}_{n_2} \\ \mathbf{1}_{n_3} & \mathbf{0}_{n_3} & \mathbf{1}_{n_3} \end{bmatrix}.$$

Treatment contrast parameters?

There are three parameters after reparameterization $\gamma_1, \gamma_2, \gamma_3$.

Since $X\boldsymbol{\beta} = Y\boldsymbol{\gamma}$, picking rows $1, n_1 + 1, n_1 + n_2 + 1$ gives:

$$\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} \gamma_1 \\ \gamma_1 + \gamma_2 \\ \gamma_1 + \gamma_3 \end{bmatrix},$$

recalling that $\mu_i = \mu + \tau_i$ is the population mean for the ith population.

Rearanging gives $\gamma_1 = \mu_1, \gamma_2 = \mu_2 - \mu_1, \gamma_3 = \mu_3 - \mu_1$.

That is, the intercept parameter is the population mean for the first level of the factor.

The other parameters are the *differences* between the population mean for the other levels with the population mean for the first level. (See Lab 8 Question 5 for use of equation 1.)

2.3 Example - 3 different maths classes

Exam marks example

We compare the marks of students in 3 different mathematics classes. There is another factor (IQ), but we ignore this for the time being.

```
maths <- read.csv("../data/maths.csv")
str(maths)

## 'data.frame': 30 obs. of  5 variables:
##  $ X      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ maths.y: int  81 84 81 79 78 79 81 85 72 79 ...
##  $ iq     : int  99 103 108 109 96 104 96 105 94 91 ...
##  $ class  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ class.f: int  1 1 1 1 1 1 1 1 1 1 ...

maths$class.f <- factor(maths$class.f)
```

Exam marks example

```
plot(maths$class, maths$maths.y)
```

Figure 1 shows the resulting boxplots.

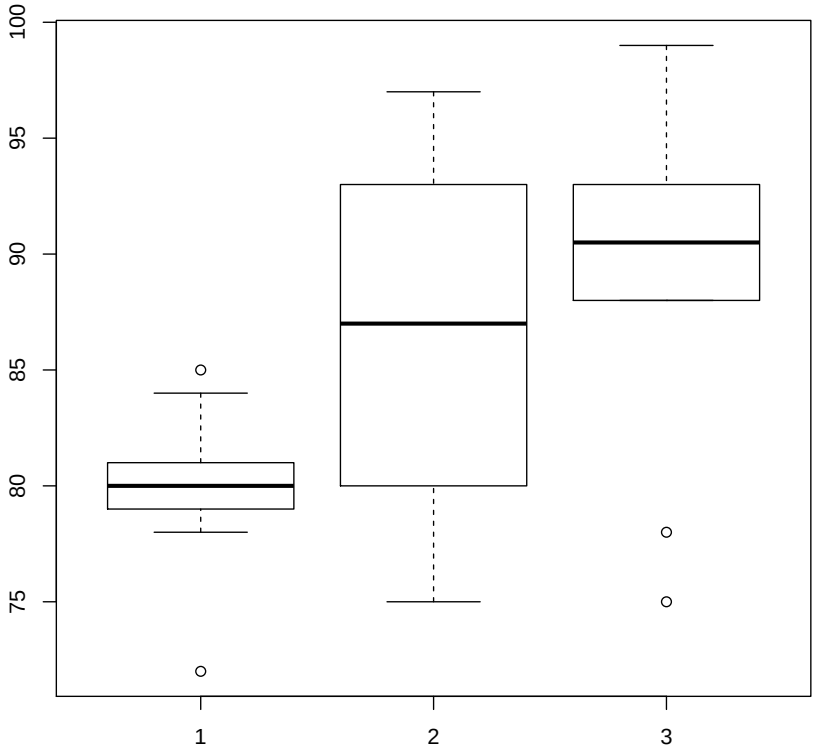


Figure 1: Boxplots of the marks in the three maths classes

Exam marks example

```
(y <- maths$maths.y)

##  [1] 81 84 81 79 78 79 81 85 72 79 85 78 93 80 83 95 90 89 97 75 90 75 99
## [24] 97 93 91 88 93 90 78

n <- dim(maths)[1]
k <- length(levels(maths$class.f))
X <- matrix(0,n,k+1)
X[,1] <- 1
X[maths$class.f==1,2] <- 1
X[maths$class.f==2,3] <- 1
X[maths$class.f==3,4] <- 1
```

Exam marks example - extract of X matrix

```
X

##       [,1] [,2] [,3] [,4]
##  [1,]    1    1    0    0
##  [2,]    1    1    0    0
##  [3,]    1    1    0    0
##  [4,]    1    1    0    0
##  [5,]    1    1    0    0
##  [6,]    1    1    0    0
##  [7,]    1    1    0    0
##  [8,]    1    1    0    0
##  [9,]    1    1    0    0
## [10,]    1    1    0    0
## [11,]    1    0    1    0
## [12,]    1    0    1    0
## [13,]    1    0    1    0
## [14,]    1    0    1    0
## [15,]    1    0    1    0
## [16,]    1    0    1    0
## [17,]    1    0    1    0
## [18,]    1    0    1    0
## [19,]    1    0    1    0
## [20,]    1    0    1    0
## [21,]    1    0    0    1
## [22,]    1    0    0    1
## [23,]    1    0    0    1
## [24,]    1    0    0    1
## [25,]    1    0    0    1
## [26,]    1    0    0    1
## [27,]    1    0    0    1
## [28,]    1    0    0    1
## [29,]    1    0    0    1
## [30,]    1    0    0    1
```

Exam marks example: reparametrisation by omitting 1's

```
Xre <- X[,-1]
(b <- solve(t(Xre) %*% Xre, t(Xre) %*% y))
```

```
##       [,1]
## [1,] 79.9
## [2,] 86.5
## [3,] 89.4
```

Exam marks example: reparametrisation using lm

```
modelre <- lm(y ~ 0 + X[,2] + X[,3] + X[,4])
summary(modelre)

##
## Call:
## lm(formula = y ~ 0 + X[, 2] + X[, 3] + X[, 4])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -14.40  -1.80   0.85   3.60  10.50
##
## Coefficients:
##        Estimate Std. Error t value Pr(>|t|)
## X[, 2]   79.900      2.053   38.92   <2e-16 ***
## X[, 3]   86.500      2.053   42.14   <2e-16 ***
## X[, 4]   89.400      2.053   43.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.492 on 27 degrees of freedom
## Multiple R-squared:  0.9948,Adjusted R-squared:  0.9942
## F-statistic:  1729 on 3 and 27 DF,  p-value: < 2.2e-16
```

The interpretation of the parameters is interesting here. They are the population means for students who would be taught in the three classes - identical conditions, same teachers etc but different years, so that they are not real populations but conceptual ones.

The next slide shows what happens if the `lm` command is used in the usual way.

To make the message clear, the option for `contr.treatment` has been entered, but this is the default.

The resulting C matrix has already been shown and the parameters described as the population mean for the first class, and the differences between the population means for the second class with the first, and the third class with the first.

```
modelstan <- lm(y ~ class.f, contrasts=contr.treatment,data = maths)
summary(modelstan)

##
## Call:
## lm(formula = y ~ class.f, data = maths, contrasts = contr.treatment)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -14.40  -1.80   0.85   3.60  10.50
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   79.900      2.053  38.922  < 2e-16 ***
## class.f2       6.600      2.903   2.273  0.03117 *
## class.f3       9.500      2.903   3.272  0.00292 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.492 on 27 degrees of freedom
## Multiple R-squared:  0.2941,Adjusted R-squared:  0.2418
## F-statistic: 5.625 on 2 and 27 DF,  p-value: 0.009077
```

3 Estimability

3.1 Definition - MM5.4

Estimability

Now we know how to reparameterize and interpret the new parameters, we might want to know which linear combinations of the parameters in the less than full rank model have the same estimates whichever is the choice of the reparameterization (ie of C).

Definition 6.3. In the general linear model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, a linear combination of the parameters $\mathbf{t}^T\boldsymbol{\beta}$ is said to be *estimable* if there exists a vector $\mathbf{c}$ such that $E[\mathbf{c}^T\mathbf{y}] = \mathbf{t}^T\boldsymbol{\beta}$, that is there is an *unbiased* estimator of $\mathbf{t}^T\boldsymbol{\beta}$ based on a linear combination of the observations $\mathbf{y}$.

3.2 Estimability theorems

Estimability theorems

Theorem 6.4. *In the linear model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, elements of $X\boldsymbol{\beta}$ are estimable.*

Proof. We know that $E[\mathbf{y}] = X\boldsymbol{\beta}$. Now take $\mathbf{e}_i$ to be the ith standard basis vector.

We have

$$\begin{aligned}(X\boldsymbol{\beta})_i &= \mathbf{e}_i^T X\boldsymbol{\beta} \\ &= \mathbf{e}_i^T E[\mathbf{y}] \\ &= E[\mathbf{e}_i^T\mathbf{y}]\end{aligned}$$

and so the ith element of $X\boldsymbol{\beta}$ is estimable.

3.3 Example - elements of $X\boldsymbol{\beta}$

Carbon example

Example. Consider the carbon removal example. We have

$$X = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{bmatrix}.$$

We know that we cannot estimate the parameter vector $\boldsymbol{\beta}$, because it is not uniquely determined.

Carbon example

However, the real quantities of interest are the mean responses from the three treatments. These are:

$$\begin{aligned} \mu + \tau_1 &= \begin{bmatrix} 1 & 1 & 0 & 0 \end{bmatrix} \boldsymbol{\beta} \\ \mu + \tau_2 &= \begin{bmatrix} 1 & 0 & 1 & 0 \end{bmatrix} \boldsymbol{\beta} \\ \mu + \tau_3 &= \begin{bmatrix} 1 & 0 & 0 & 1 \end{bmatrix} \boldsymbol{\beta} \end{aligned}$$

and each of these are elements of $X\boldsymbol{\beta}$. Therefore, they are estimable.

In a one-way classification model with any number of levels, $\mu + \tau_i$ is always estimable.

Independent of parameterization

Theorem 6.5. *Whatever the full rank parameterization, the BLUE estimator of an estimable combination of parameters is the same and is a linear combination of the least squares estimates for that parameterization.*

Proof. Suppose $\mathbf{t}^T\boldsymbol{\beta}$ is estimable. Then there is a column vector $\mathbf{c}$ so that $\mathbf{c}^T y$ is an unbiased estimator of $\mathbf{t}^T\boldsymbol{\beta}$.

But $E(\mathbf{c}^T y) = \mathbf{c}^T X\boldsymbol{\beta}$. Hence $\mathbf{t}^T\boldsymbol{\beta} = \mathbf{c}^T X\boldsymbol{\beta}$ for all values of $\boldsymbol{\beta}$.

Thus for a particular parameterisation, $y = Y\boldsymbol{\gamma} + \mathbf{e}$, we have $\mathbf{t}^T\boldsymbol{\beta} = \mathbf{c}^T X\boldsymbol{\beta} = \mathbf{c}^T Y\boldsymbol{\gamma}$.

The BLUE of this is $\mathbf{c}^T Y\mathbf{g}$ and this is a linear combination of the fitted values. Hence, it is the same whatever the parameterization.

3.4 Linear combinations

Linear combinations of estimable combinations

Theorem 6.6. *Let $\mathbf{t}_1^T\boldsymbol{\beta}, \mathbf{t}_2^T\boldsymbol{\beta}, \ldots, \mathbf{t}_k^T\boldsymbol{\beta}$ be estimable combinations of parameters, and let*

$$z = a_1\mathbf{t}_1^T\boldsymbol{\beta} + a_2\mathbf{t}_2^T\boldsymbol{\beta} + \ldots + a_k\mathbf{t}_k^T\boldsymbol{\beta}.$$

Then z is estimable, and the best linear unbiased estimator for z is a linear combination of the least squares estimators in any parameterization.

Linear combinations proof

Proof. By definition,

$$z = (a_1\mathbf{t}_1 + a_2\mathbf{t}_2 + \ldots + a_k\mathbf{t}_k)^T\boldsymbol{\beta}.$$

Therefore z is estimable, with BLUE a corresponding linear combination of least squares estimates in any parameterization.

Treatment contrasts

Of particular interest in many studies is the way different populations compare against each other. In a one-way classification model suppose the mean response for the ith population is$\mu + \tau_i$. To attach a numerical value to these comparisons, we form linear combinations

$$a_1\tau_1 + a_2\tau_2 + \ldots + a_k\tau_k,$$

where $\sum_{i=1}^{k} a_i = 0$.

These are called *treatment contrasts*. They wipe out the effect of the overall mean response in describing the mean differences between populations as we shall see soon.

Treatment contrasts

In a one-way classification model (with the mean of the ith population being $\mu + \tau_i$), any treatment contrast is estimable.

If

$$z = a_1\tau_1 + a_2\tau_2 + \ldots + a_k\tau_k$$

is a treatment contrast, then

$$\begin{aligned} z &= \sum_{i=1}^{k} a_k\mu + a_1\tau_1 + a_2\tau_2 + \ldots + a_k\tau_k \\ &= a_1(\mu + \tau_1) + a_2(\mu + \tau_2) + \ldots + a_k(\mu + \tau_k) \end{aligned}$$

is a linear combination of the estimable functions $\mu + \tau_i$, and is therefore estimable.

Treatment contrasts

Of particular interest among treatment contrasts is the contrast of the form $\tau_i - \tau_j$, for some $i \neq j$. This is because

$$\tau_i - \tau_j = (\mu + \tau_i) - (\mu + \tau_j)$$

is the difference between the mean responses in populations i and j.

We would expect to estimate this contrast by the corresponding difference in sample means, $\bar{y}_i - \bar{y}_j$.

3.5 Contrast sets - not in MM

`contr.treatment` and `contr.sum` in R

For the less than full rank model, R uses *contrasts* to record parameter estimates for a factor with k levels. The two main contrast sets are `contr.treatment` and `contr.sum`.

Each contrast set has the maximum number, k, of parameter combinations to estimate.

Label	`contr.treatment`	`contr.sum`
Intercept	$\mu + \tau_1$	$\mu + \frac{1}{k}\sum \tau_i$
level 1		$\tau_1 - \frac{1}{k}\sum \tau_i$
level 2	$\tau_2 - \tau_1$	$\tau_2 - \frac{1}{k}\sum \tau_i$
level 3	$\tau_3 - \tau_1$	$\tau_3 - \frac{1}{k}\sum \tau_i$
$\vdots$	$\vdots$	$\vdots$
level k-1	$\tau_{k-1} - \tau_1$	$\tau_{k-1} - \frac{1}{k}\sum \tau_i$
level k	$\tau_k - \tau_1$	

`contr.treatment` and `contr.sum` in R

The intercept term for `contr.treatment` (the default) is estimable because it is an element of $X\beta$. The intercept term for `contr.sum` is estimable because it is $E(\bar{y})$ (where $\bar{y}$ is defined as the simple average over the factor levels of the sample means (so it is the usual average if, and only if, there are the same number of observations at each factor level) - check this). The other terms are contrasts between estimable combinations (check this).

Exam marks example - illustration of `contr.treatment`

```
contrasts(maths$class.f) <- contr.treatment(3)
model <- lm(maths.y ~ class.f, data = maths)
summary(model)

##
## Call:
## lm(formula = maths.y ~ class.f, data = maths)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -14.40  -1.80   0.85   3.60  10.50
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   79.900      2.053  38.922  < 2e-16 ***
## class.f2       6.600      2.903   2.273  0.03117 *
## class.f3       9.500      2.903   3.272  0.00292 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.492 on 27 degrees of freedom
## Multiple R-squared:  0.2941,Adjusted R-squared:  0.2418
## F-statistic: 5.625 on 2 and 27 DF,  p-value: 0.009077
```

Exam marks example - illustration of `contr.sum`

```r
contrasts(maths$class.f) <- contr.sum(3)
model2 <- lm(maths.y ~ class.f, data = maths)
summary(model2)

##
## Call:
## lm(formula = maths.y ~ class.f, data = maths)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -14.40  -1.80   0.85   3.60  10.50
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   85.267      1.185  71.943  < 2e-16 ***
## class.f1      -5.367      1.676  -3.202  0.00348 **
## class.f2       1.233      1.676   0.736  0.46818
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.492 on 27 degrees of freedom
## Multiple R-squared:  0.2941,Adjusted R-squared:  0.2418
## F-statistic: 5.625 on 2 and 27 DF,  p-value: 0.009077
```

3.6 Diagnostic plots for a factor - illustration on exam marks - not in MM

R commands to produce diagnostic plots

```r
plot(model, which=1)
plot(model, which=2)
plot(model, which=3)
plot(model, which=5)
```

Figures 2, 3, 4 and 5 show the diagnostic plots for exam marks in the three classes.

4 Testability

4.1 Definition

Testability

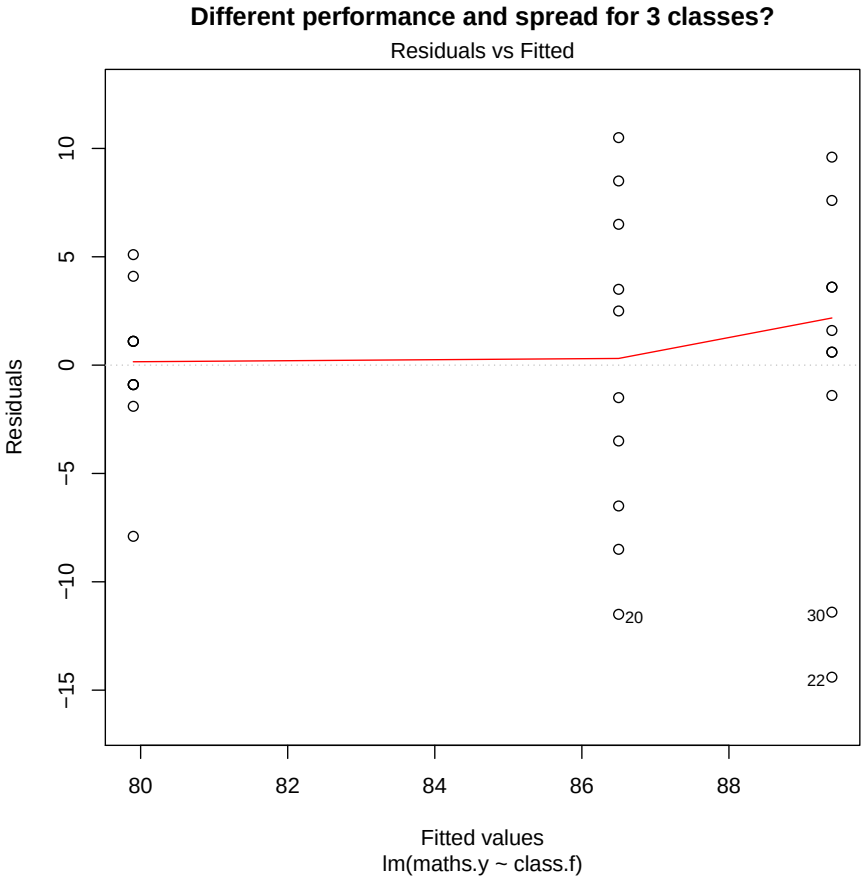

Different performance and spread for 3 classes?
Residuals vs Fitted
Residuals
10
5
0
−5
−10
−15
80
82
84
86
88
Fitted values
lm(maths.y ~ class.f)
20
30
22

Figure 2:

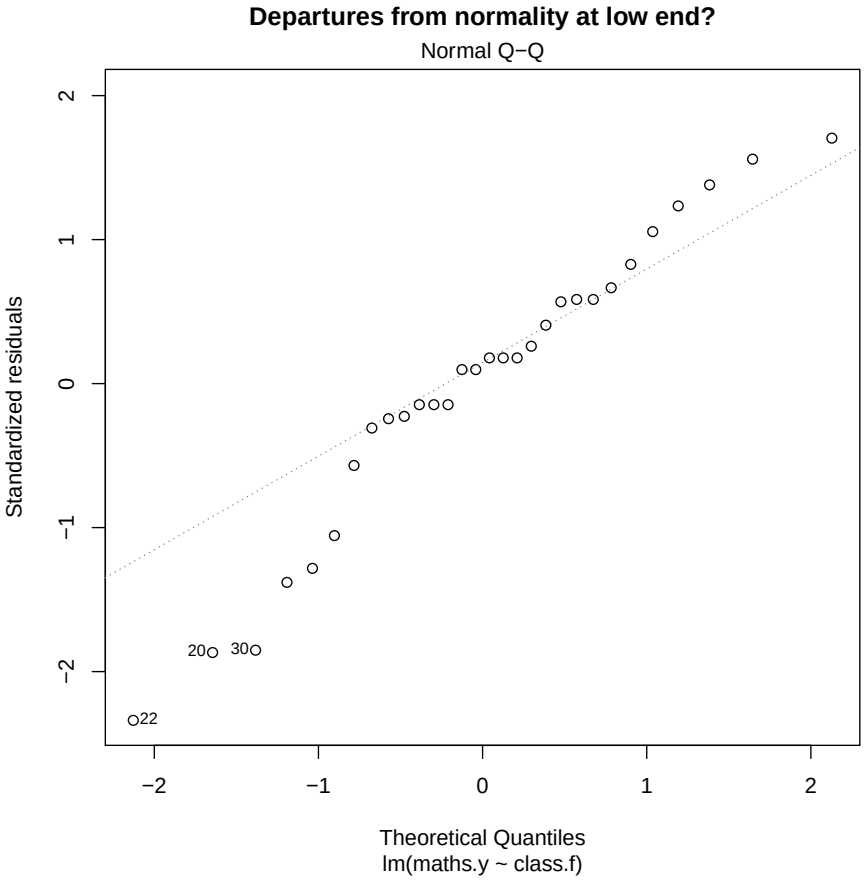

Departures from normality at low end?
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(maths.y ~ class.f)
20
30
22
−2
−1
0
1
2

Figure 3:

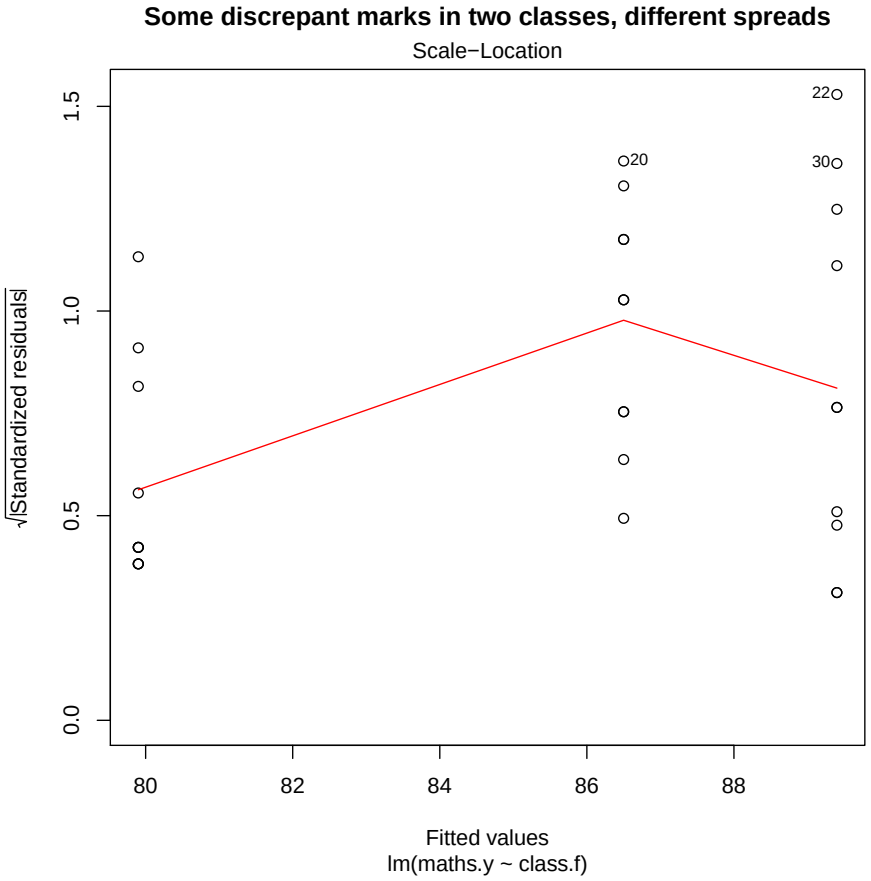

Some discrepant marks in two classes, different spreads
Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
20
22
30
80
82
84
86
88
Fitted values
lm(maths.y ~ class.f)

Figure 4:

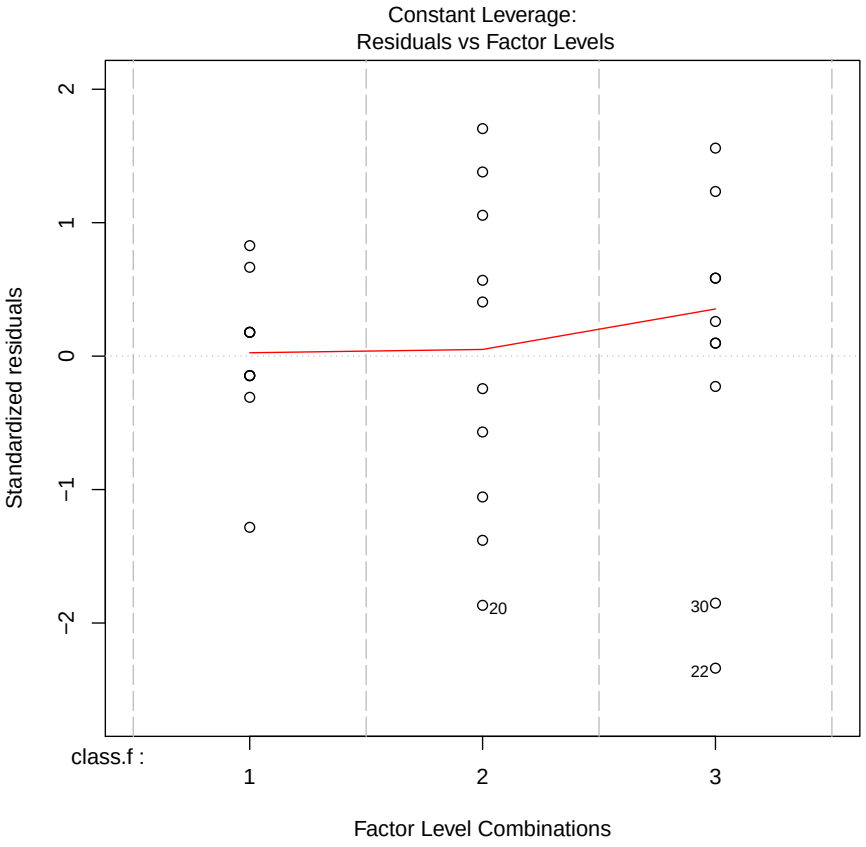

Constant Leverage:
Residuals vs Factor Levels
Standardized residuals
2
1
0
−1
−2
20
30
22
class.f :
1
2
3
Factor Level Combinations

Figure 5:

Definition 6.7. A hypothesis H_0 is testable if there exists a set of estimable functions $\mathbf{c}_1^T\boldsymbol{\beta}, \mathbf{c}_2^T\boldsymbol{\beta}, \ldots, \mathbf{c}_m^T\boldsymbol{\beta}$ such that H_0 is true if and only if

$$\mathbf{c}_1^T\boldsymbol{\beta} = \mathbf{c}_2^T\boldsymbol{\beta} = \ldots = \mathbf{c}_m^T\boldsymbol{\beta} = \mathbf{0},$$

and $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_m$ are linearly independent.

Testability

That is, a testable hypothesis is of the form $H_0 : C\boldsymbol{\beta} = \mathbf{0}$, where

$$C = \begin{bmatrix} \mathbf{c}_1^T \\ \vdots \\ \mathbf{c}_m^T \end{bmatrix}$$

is $m \times p$ of rank m, and each $\mathbf{c}_i^T\boldsymbol{\beta}$ is estimable.

4.2 One factor example

Testability

Example. Consider the one-way classification model with fixed effects and $k = 3$. The linear model that we use is

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}.$$

Consider the hypothesis that the means of all three populations are equal. This is equivalent to $H_0 : \tau_1 = \tau_2 = \tau_3$.

This hypothesis is true if and only if

$$\tau_1 - \tau_2 = 0$$

and

$$\tau_2 - \tau_3 = 0.$$

Testability

So we can express this hypothesis as $H_0 : C\boldsymbol{\beta} = \mathbf{0}$, where

$$C = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{bmatrix}.$$

$\tau_1 - \tau_2$ is a contrast, so it is estimable. Similarly, $\tau_2 - \tau_3$ is estimable. The rows of C are obviously linearly independent, so H_0 is testable.

4.3 How to test

Testability

Once we have determined that a hypothesis is testable, how can we test it?

Answer: Take any full rank parameterisation, and the estimates of the parameters will be the same, and the tests of hypotheses will be the same.

The F statistic for this hypothesis (for the reparameterized full rank case) is

$$\frac{(C\mathbf{b})^T[C(X^TX)^{-1}C^T]^{-1}C\mathbf{b}/m}{SS_{Res}/(n-p)},$$

which under the null hypothesis has an F distribution with m and $n-p$ degrees of freedom, where $m = r(C)$.

4.4 Carbon removal example

Carbon removal example

Example. Let us look at the carbon removal example from the previous section. We compare three methods of removing carbon from wastewater. The data is:

AF	FS	FCC
34.6	38.8	26.7
35.1	39.0	26.7
35.3	40.1	27.0

Carbon removal example

We test whether the populations have the same mean, i.e. $H_0 : \tau_1 = \tau_2 = \tau_3$. This can be written in matrix form as $H_0 : C\boldsymbol{\beta} = \mathbf{0}$, where

$$C = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{bmatrix}.$$

Carbon removal example

```
n <- 9
r <- 3
y <- c(34.6,35.1,35.3,38.8,39.0,40.1,26.7,26.7,27.0)
x <- factor(c(rep(1,3),rep(2,3),rep(3,3)))
anova(lm(y~x))
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x          2 241.98 120.990  558.42 1.526e-07 ***
## Residuals  6   1.30   0.217
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Carbon removal example

```
anova(lm(y~x,contrast=contr.sum(x)))
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x          2 241.98 120.990  558.42 1.526e-07 ***
## Residuals  6   1.30   0.217
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can reject H_0 firmly, so the populations are not all the same. It is still possible that *some* of the populations are the same, but not all of them.

4.5 Orthogonal contrasts

Orthogonal contrasts

A concept which is similar in nature to orthogonality is that of *orthogonal contrasts*.

Definition 6.8. In a one-factor model, two treatment contrasts $\sum_{i=1}^{k} a_i\mu_i$ and $\sum_{i=1} b_i\mu_i$ are orthogonal if (and only if)

$$\sum_{i=1}^{k} \frac{a_i b_i}{n_i} = 0.$$

If each sub-population is equally sampled, this reduces to

$$\sum_{i=1}^{k} a_i b_i = 0.$$

Orthogonal contrasts behave like orthogonal variables, in the sense that they can be tested independently of each other.

The sum of squares attributed to a hypothesis made up of multiple orthogonal contrasts can be broken down into sums of squares attributed to each of the individual contrasts.

5 Two-factor model

5.1 Notation

Two-factor models

In this section, we look at two-factor models (two-way classification), but the ideas extend easily any number of factors.

In a basic two-factor model, we assume that each level of each factor affects the overall mean μ by a specific amount. We name these effects τ_i for the ith level of factor 1 and β_j for the jth level of factor 2. Then our model is

$$y_{ijk} = \mu + \tau_i + \beta_j + \varepsilon_{ijk}, \qquad i = 1, \ldots, a, j = 1, \ldots, b, k = 1, \ldots, n_{ij}.$$

In matrix form (with 1 sample from each combination of factor levels):

Two-factor models

$$X = \left[\begin{array}{c|cccc|cccc} 1 & 1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 & 0 & 1 & & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & & \ddots & \vdots \\ 1 & 1 & 0 & \dots & 0 & 0 & 0 & \dots & 1 \\ \hline 1 & 0 & 1 & \dots & 0 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 & 0 & 1 & & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & & \ddots & \vdots \\ 1 & 0 & 1 & \dots & 0 & 0 & 0 & \dots & 1 \\ \hline \vdots & & & & & & & & \\ \hline 1 & 0 & 0 & \dots & 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 1 & 0 & 1 & & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & & \ddots & \vdots \\ 1 & 0 & 0 & \dots & 1 & 0 & 0 & \dots & 1 \end{array}\right], \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \vdots \\ \tau_a \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_b \end{bmatrix}$$

5.2 Estimable contrasts

Two-factor models

This model is known as the *additive* model. It assumes that the effects from each factor can be added to produce the overall effect.

Any hypothesis that we can test in a one-factor model can be tested for each factor in an additive two-factor model.

Theorem 6.9. *In an additive two-factor model, every contrast in the τ's is estimable. Similarly, every contrast in the β's is estimable.*

Two-factor models

The most common hypotheses that we will want to test are

$$\tau_1 = \tau_2 = \ldots = \tau_a$$

and

$$\beta_1 = \beta_2 = \ldots = \beta_b.$$

Because they are all composed of treatment contrasts for one factor, they are testable. We can use the theory already developed to test them.

5.3 Example: capsule dissolve time

Two-factor models

Example. We model the time taken to dissolve a capsule in a biological fluid. A study is conducted with 1 sample from each combination of factor levels and the following data found:

Time		Fluid type	
		Gastric	Duodenal
Capsule	A	39.5	31.2
Capsule	B	47.4	44

Two-factor models

The linear model is

$$\mathbf{y} = \begin{bmatrix} 39.5 \\ 47.4 \\ 31.2 \\ 44 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \beta_1 \\ \beta_2 \end{bmatrix}.$$

We test the hypotheses that there is no difference in the response for the levels of each of the two factors.

Two-factor models

The first factor (fluid type) gives the null hypothesis against the alternative of different values of τ.

$$H_0 : \tau_1 = \tau_2 \text{ or } \begin{bmatrix} 0 & 1 & -1 & 0 & 0 \end{bmatrix} \boldsymbol{\beta} = \mathbf{0}.$$

The second factor (capsule type) gives the hypothesis of different values of β.

$$H_0 : \beta_1 = \beta_2 \text{ or } \begin{bmatrix} 0 & 0 & 0 & 1 & -1 \end{bmatrix} \boldsymbol{\beta} = \mathbf{0}.$$

The next slide gives the default option of `contr.treatment` for both τ and β, and the following one gives it for `contr.treatment`. This is followed by the the model matrices and the analysis of variance.

```
##
## Call:
## lm(formula = y ~ tau + beta)
##
## Residuals:
##      1      2      3      4
##  1.225 -1.225 -1.225  1.225
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   38.275      2.122  18.039   0.0353 *
## tau2          -5.850      2.450  -2.388   0.2525
## beta2         10.350      2.450   4.224   0.1480
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.45 on 1 degrees of freedom
## Multiple R-squared:  0.9593,Adjusted R-squared:  0.8778
## F-statistic: 11.77 on 2 and 1 DF,  p-value: 0.2018
```

5.4 Sum contrasts

Sum contrasts

```
summary(models <- lm(y ~ tau + beta,
contrasts=
list(tau="contr.sum",beta="contr.sum")))
```

```
##
## Call:
## lm(formula = y ~ tau + beta, contrasts = list(tau = "contr.sum",
##     beta = "contr.sum"))
##
## Residuals:
##      1      2      3      4
##  1.225 -1.225 -1.225  1.225
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   40.525      1.225  33.082   0.0192 *
## tau1           2.925      1.225   2.388   0.2525
## beta1         -5.175      1.225  -4.224   0.1480
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.45 on 1 degrees of freedom
## Multiple R-squared:  0.9593,Adjusted R-squared:  0.8778
## F-statistic: 11.77 on 2 and 1 DF,  p-value: 0.2018
```

```
anova(modelt)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value Pr(>F)
## tau        1  34.222  34.222  5.7014 0.2525
## beta       1 107.123 107.123 17.8463 0.1480
## Residuals  1   6.002   6.002
```

```
anova(models)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value Pr(>F)
## tau        1  34.222  34.222  5.7014 0.2525
## beta       1 107.123 107.123 17.8463 0.1480
## Residuals  1   6.002   6.002
```

We cannot reject either null hypothesis, which is not surprising since there is only one data point for each combination of parameters.

In a two- or more-factor model, it is still possible to have orthogonal contrasts. This happens in particular if the design is *balanced*, which means that we have the same number of samples from each combination of factors.

In this case it is simple to show that any treatment contrast in the τs is orthogonal to any treatment contrast in the βs.

6 Interaction

6.1 Definition

Interaction

In some cases, it is possible that *interaction* between factors may occur.

Interaction happens when one factor affects the effect of another factor.

For example, if the effect of factor 1 when factor 2 is at level 1 is different from the effect of factor 1 when factor 2 is at level 2, then there is interaction.

Interaction

Example. Suppose that we are studying the effect of pressure and temperature on viscosity, and the *actual* means of the response variable for each of the combinations are given by:

		Pressure			
		1	2	3	4
Temperature	1	4	6	4	3
	2	8	2	7	5

Interaction

When the pressure is at level 1, changing the temperature from level 1 to level 2 results in an increase of viscosity of 4.

However, when the pressure is at level 2, changing the temperature from level 1 to level 2 results in a *decrease* of viscosity of 4!

In this case, the factors interact.

Interaction

If, on the other hand, the actual means were:

		Pressure			
		1	2	3	4
Temperature	1	4	6	4	3
	2	8	10	8	7

then there would be no interaction between the factors. Even though the factors themselves are significant, the *combination* of factor levels has no effect apart from the individual factor effects.

Interaction

An additive model assumes that there is no interaction between the factors, so the effects of the factor levels can be measured in isolation from the other factor(s).

If we have interaction, or want to test whether there is interaction, we must use a different model:

$$y_{ijk} = \mu + \tau_i + \beta_j + \xi_{ij} + \varepsilon_{ijk},$$

where ξ_{ij} is an interaction term which quantifies the effect of factor 1 being at level i at the same time that factor 2 is at level j.

Interaction

Example. Consider the previous example (dissolving a capsule in fluid). If we allow an interaction term, $\mathbf{y}$ stays the same, but the linear model becomes

$$X = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \beta_1 \\ \beta_2 \\ \xi_{11} \\ \xi_{12} \\ \xi_{21} \\ \xi_{22} \end{bmatrix}.$$

Interaction

In a two-factor model with interaction, we are often interested in testing whether there is interaction or not.

However, testing the presence of interaction is not quite as straightforward as it may seem. It seems like we would want to test the hypothesis

$$H_0 : \xi_{11} = \xi_{12} = \ldots = \xi_{1b} = \xi_{21} = \ldots = \xi_{ab},$$

but it turns out that this is not correct. We illustrate with an example.

Interaction

Example. Suppose we have a two-factor model with the following actual means:

		Factor I	
		1	2
Factor II	1	6	5
	2	6	5

There is clearly no interaction between the factors.

Interaction

One possible parameter set is

$$\mu = 0, \tau_1 = 5, \tau_2 = 4, \beta_1 = \beta_2 = 1,$$
$$\xi_{ij} = 0 \qquad \forall i, j.$$

However, an equally valid parameter set is

$$\mu = 0, \tau_1 = 2, \tau_2 = 1, \beta_1 = 3, \beta_2 = 2,$$
$$\xi_{11} = 1, \xi_{12} = 2, \xi_{21} = 1, \xi_{22} = 2.$$

Interaction

Thus, while $\xi_{ij} = 0$ for all i, j implies no interaction, it is not actually necessary.

Moreover, the hypothesis $H_0 : \xi_{ij} = 0 \, \forall i, j$ is not even testable.

Nor is $H_0 : \xi_{ij}$ the same $\forall i, j$.

Example

Consider a two-way classification where each factor has two levels, with one observation from each combination of levels. We have

$$X = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Example

We can write the hypothesis $H_0 : \xi_{ij}$ the same $\forall i, j$ as $C\boldsymbol{\beta} = \mathbf{0}$ where

$$C = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \end{bmatrix}.$$

This hypothesis is not testable since $\xi_{11} - \xi_{12}$ is not estimable. Why?

6.2 Testing for interaction

Interaction

Theorem 6.10. *For the linear model*

$$y_{ijk} = \mu + \tau_i + \beta_j + \xi_{ij} + \varepsilon_{ijk},$$

there is no interaction if and only if

$$(\xi_{ij} - \xi_{ij'}) - (\xi_{i'j} - \xi_{i'j'}) = 0,$$

for all $i \neq i', j \neq j'$.

Moreover these quantities are all estimable.

Interaction

Proof. A sample which has level i of factor 1 and level j of factor 2 has mean

$$\mu_{ij} = \mu + \tau_i + \beta_j + \xi_{ij}.$$

Now take two levels of factor 1 (i and i') and two levels of factor 2 (j and j').

No interaction between these levels means the difference in means that results from switching factor 2 from j to j' is the same whether factor 1 is at level i or i'.

Interaction

In other words,

$$\mu_{ij} - \mu_{ij'} = \mu_{i'j} - \mu_{i'j'},$$

and expanding gives

$$\mu + \tau_i + \beta_j + \xi_{ij} - \mu - \tau_i - \beta_{j'} - \xi_{ij'} = \mu + \tau_{i'} + \beta_j + \xi_{i'j} - \mu - \tau_{i'} - \beta_{j'} - \xi_{i'j'}$$

which reduces to

$$(\xi_{ij} - \xi_{ij'}) - (\xi_{i'j} - \xi_{i'j'}) = 0.$$

In order for there to be no interaction at all, we need this condition to hold for all i, i', j, j'.

The group means μ_{ij} are all elements of $X\boldsymbol{\beta}$, and thus linear combinations of them are estimable.

Interaction

Theorem 6.5 generates $ab(a-1)(b-1)$ equations. However, it can be shown that all but $(a-1)(b-1)$ of them are redundant.

Example. In a two-factor design with two levels in each factor, Theorem 6.5 shows that there is no interaction if and only if

$$\begin{aligned}
(\xi_{11} - \xi_{12}) - (\xi_{21} - \xi_{22}) &= 0 \\
(\xi_{21} - \xi_{22}) - (\xi_{11} - \xi_{12}) &= 0 \\
(\xi_{12} - \xi_{11}) - (\xi_{22} - \xi_{21}) &= 0 \\
(\xi_{22} - \xi_{21}) - (\xi_{12} - \xi_{11}) &= 0.
\end{aligned}$$

Interaction

It is easy to see that all of these equations are equivalent, so we need only test one.

This gives the hypothesis $H_0 : C\boldsymbol{\beta} = \mathbf{0}$, where

$$\begin{aligned} C &= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 \end{bmatrix}, \\ \boldsymbol{\beta} &= \begin{bmatrix} \mu & \tau_1 & \tau_2 & \beta_1 & \beta_2 & \xi_{11} & \xi_{12} & \xi_{21} & \xi_{22} \end{bmatrix}^T. \end{aligned}$$

Interaction considerations

Some things to consider when testing for interaction:

1) If we have one sample per combination of factors, it is impossible to account for or test for interaction.

This is because $r(X) = n$ and therefore $n-r$, the residual degrees of freedom, are 0.

Essentially we treat each combination of factors as a separate population. If we have one sample from each population, then we have no way to estimate the variance!

Interaction considerations

2) If we test for interaction and find that there is none, we theoretically should still use the residual sum of squares from the full model with interaction, unless there is a convincing data-related reason to think that there is no interaction.

For example, there might be a danger of over-fitting and prediction on another data set is required.

This follows from the same reasoning as using SS_{Res} for the full model in sequential tests: we cannot be sure that there is no interaction, we just haven't found any!

However, for practical purposes, this may take away too many degrees of freedom from SS_{Res}. So if you find no interaction, it's OK to use the SS_{Res} from an additive model.

Interaction considerations

3) It is possible to have interaction between three or more factors.

However, this is hard to test for and hard to interpret. In practice most people only look at two-factor interactions.

6.3 Example: engine

Engine example

We look at the effect of pre-chamber volume ratio and injection timing on the emission of noxious gas from an engine. The factors have 3 levels each.

```
str(engine)

## 'data.frame': 18 obs. of  3 variables:
##  $ gas   : num  6.27 8.08 7.34 5.43 8.04 7.87 6.94 7.48 8.61 6.51 ...
##  $ volume: Factor w/ 3 levels "low","medium",..: 1 2 3 1 2 3 1 2 3 1 ...
##  $ time  : Factor w/ 3 levels "short","medium",..: 1 1 1 1 1 1 2 2 2 2 ...
```

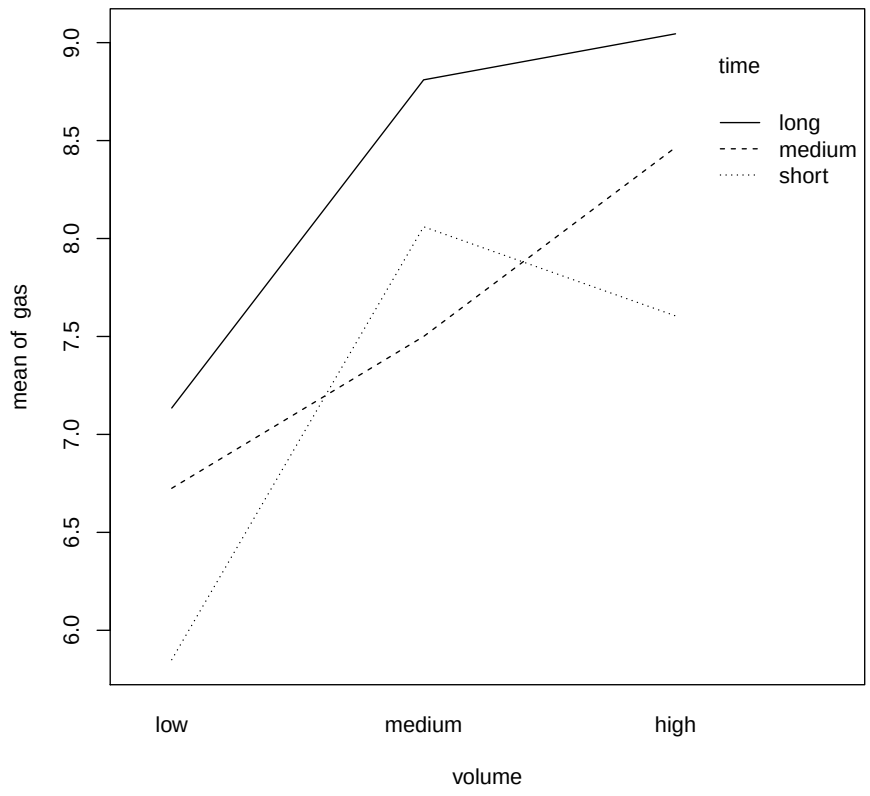
```
means
```

```
##       [,1]  [,2]  [,3]
## [1,] 5.850 6.725 7.135
## [2,] 8.060 7.500 8.810
## [3,] 7.605 8.465 9.045
```

Engine example

```
with(engine, interaction.plot(time, volume, gas))
```

Engine example



7 ANCOVA

7.1 Factor and quantitative variables

ANCOVA

We can also do analysis of covariance (ANCOVA) using the linear model framework.

In this case we have one (or more) categorical predictors and one (or more) continuous predictors. For example:

$$y_{ij} = \mu + \tau_i + \beta x_{ij} + \xi_i x_{ij} + \varepsilon_{ij}.$$

We can think of this simple model as fitting several regression lines, one to each population (assuming equal variances across populations).

ANCOVA

Interaction in this case means that the slopes of the regression lines (effect of continuous predictor) are different for each population.

A model without interaction assumes that the slopes are the same (but the intercepts may be different):

$$y_{ij} = \mu + \tau_i + \beta x_{ij} + \varepsilon_{ij}.$$

We fit these models using the less than full rank model.

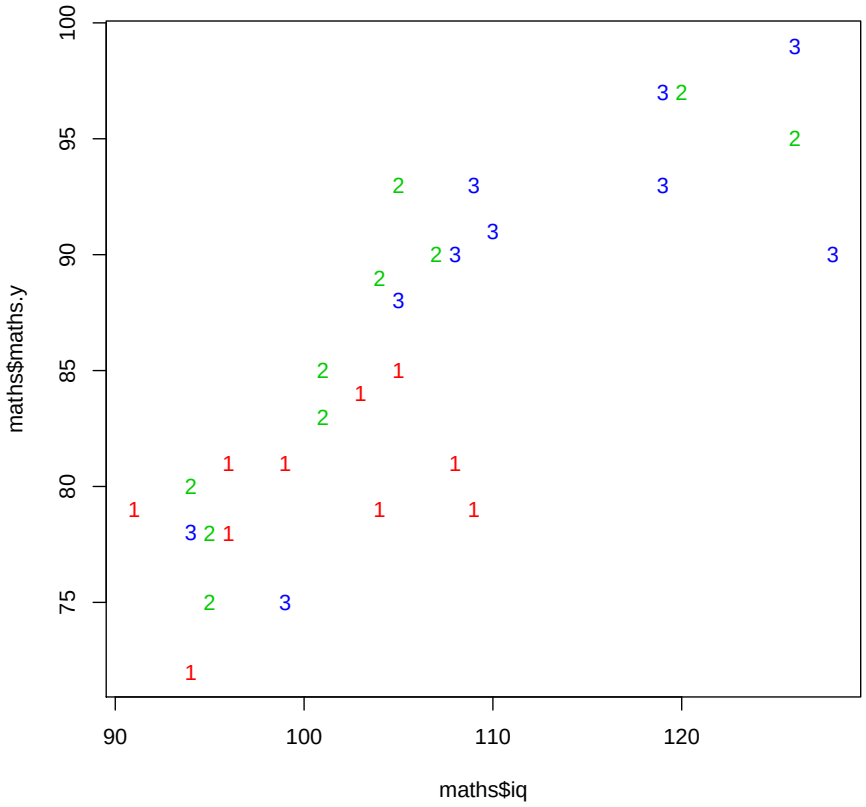
7.2 Example: exam marks

Exam marks example

The `maths` dataset also has another component: the IQ of the student and we can plot the maths score versus IQ with class colour coded.

```
plot(maths$iq, maths$maths.y, pch=array(maths$class.f),
col=maths$class+1)
```

Plot of maths versus iq colour coded for class



The full model includes `class` as a factor, `iq` as a variable, and allows for different slopes and intercepts for each class - `*` indicates to allow for this.

```
model <- lm(maths.y ~ class.f * iq, data=maths)
summary(model)


##
## Call:
## lm(formula = maths.y ~ class.f * iq, data = maths)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2507 -1.8312  0.9807  2.4711  6.3765
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  34.43315    9.68693   3.555  0.00161 **
## class.f1     18.32451   15.97135   1.147  0.26255
## class.f2    -12.63968   12.36337  -1.022  0.31681
## iq            0.47683    0.09327   5.112 3.13e-05 ***
## class.f1:iq  -0.20676    0.15688  -1.318  0.19996
## class.f2:iq   0.14060    0.11842   1.187  0.24674
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.043 on 24 degrees of freedom
## Multiple R-squared:  0.7566,Adjusted R-squared:  0.7059
## F-statistic: 14.92 on 5 and 24 DF,  p-value: 1.072e-06
```

Testing whether interaction is needed

```r
amodel <- lm(maths.y ~ class.f + iq, data = maths)
anova(amodel, model)
```

```
## Analysis of Variance Table
##
## Model 1: maths.y ~ class.f + iq
## Model 2: maths.y ~ class.f * iq
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1     26 423.42
## 2     24 392.36  2    31.062 0.95 0.4008
```

Interaction is not significant, so we remove the interaction term and fit an additive model - this amounts to the same slope for each class but different intercepts.

Summary of chosen model

```r
summary(amodel)
```

```
##
## Call:
## lm(formula = maths.y ~ class.f + iq, data = maths)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.137 -2.842  1.220  2.662  6.393
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.62522    8.58371   3.335  0.00257 **
## class.f1    -2.59713    1.12274  -2.313  0.02888 *
## class.f2     1.69790    1.04432   1.626  0.11605
## iq           0.53604    0.08093   6.623 5.03e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.036 on 26 degrees of freedom
## Multiple R-squared:  0.7373,Adjusted R-squared:  0.707
## F-statistic: 24.33 on 3 and 26 DF,  p-value: 1.032e-07
```

Is IQ necessary?

```r
basemodel <- lm(maths.y ~ class.f, data = maths)
anova(basemodel,amodel)
```

```
## Analysis of Variance Table
##
## Model 1: maths.y ~ class.f
## Model 2: maths.y ~ class.f + iq
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     27 1137.80
## 2     26  423.42  1    714.38 43.866 5.032e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Clearly IQ is significant.

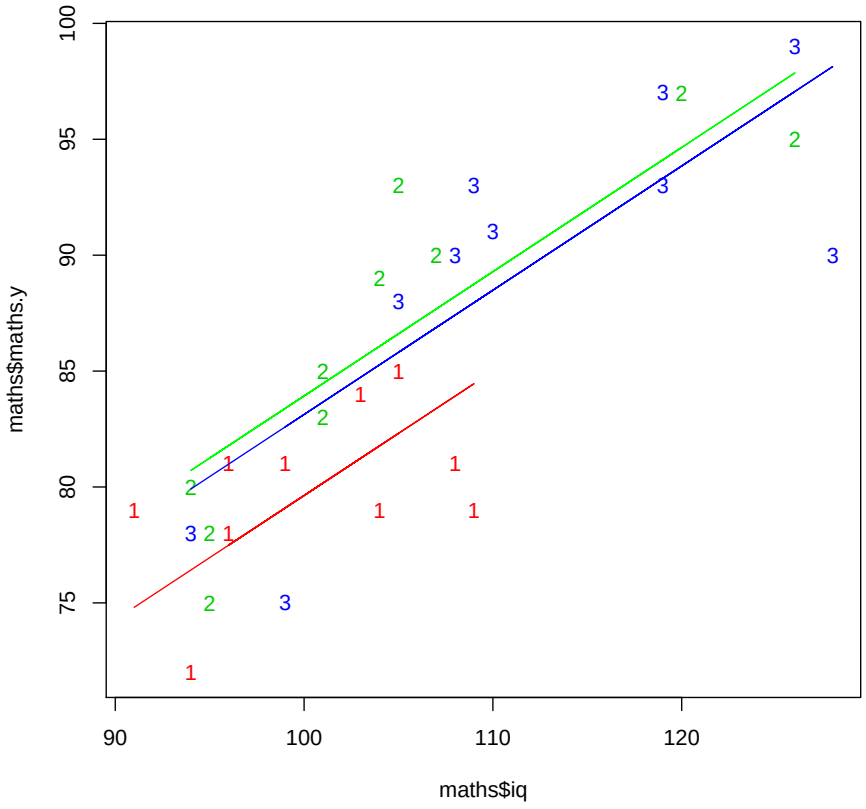
Exam marks example

```
basemodel <- lm(maths.y ~ iq, data = maths)
anova(basemodel, amodel)

## Analysis of Variance Table
##
## Model 1: maths.y ~ iq
## Model 2: maths.y ~ class.f + iq
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     28 518.13
## 2     26 423.42  2    94.707 2.9077 0.0725 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The class is not very significant. However, since it is so close, we will retain it. Remember that it was significant in the one-factor model!

The fitted ANCOVA model



Exam marks example

To spell out the full model including IQ as a less than full rank model:

```
maths <- read.csv("../data/maths.csv")
maths$class.f <- factor(maths$class)
y <- maths$maths.y
n <- 30
X <- matrix(0, n, 8)
X[,1] <- 1
X[cbind(1:n,maths$class+1)] <- 1
X[,5] <- maths$iq
X[cbind(1:n,maths$class+5)] <- maths$iq
r <- rankMatrix(X)[1]
```

Model matrix - less than full rank

```
X
```

```
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
##  [1,]    1    1    0    0   99   99    0    0
##  [2,]    1    1    0    0  103  103    0    0
##  [3,]    1    1    0    0  108  108    0    0
##  [4,]    1    1    0    0  109  109    0    0
##  [5,]    1    1    0    0   96   96    0    0
##  [6,]    1    1    0    0  104  104    0    0
##  [7,]    1    1    0    0   96   96    0    0
##  [8,]    1    1    0    0  105  105    0    0
##  [9,]    1    1    0    0   94   94    0    0
## [10,]    1    1    0    0   91   91    0    0
## [11,]    1    0    1    0  101    0  101    0
## [12,]    1    0    1    0   95    0   95    0
## [13,]    1    0    1    0  105    0  105    0
## [14,]    1    0    1    0   94    0   94    0
## [15,]    1    0    1    0  101    0  101    0
## [16,]    1    0    1    0  126    0  126    0
## [17,]    1    0    1    0  107    0  107    0
## [18,]    1    0    1    0  104    0  104    0
## [19,]    1    0    1    0  120    0  120    0
## [20,]    1    0    1    0   95    0   95    0
## [21,]    1    0    0    1  108    0    0  108
## [22,]    1    0    0    1   99    0    0   99
## [23,]    1    0    0    1  126    0    0  126
## [24,]    1    0    0    1  119    0    0  119
## [25,]    1    0    0    1  109    0    0  109
## [26,]    1    0    0    1  110    0    0  110
## [27,]    1    0    0    1  105    0    0  105
## [28,]    1    0    0    1  119    0    0  119
## [29,]    1    0    0    1  128    0    0  128
## [30,]    1    0    0    1   94    0    0   94
```

Model matrix - `contr.treatment`

```
model.matrix(~class.f*iq,data=maths)
```

```
##    (Intercept) class.f2 class.f3  iq class.f2:iq class.f3:iq
## 1            1        0        0  99          0          0
## 2            1        0        0 103          0          0
## 3            1        0        0 108          0          0
## 4            1        0        0 109          0          0
## 5            1        0        0  96          0          0
## 6            1        0        0 104          0          0
## 7            1        0        0  96          0          0
## 8            1        0        0 105          0          0
## 9            1        0        0  94          0          0
## 10           1        0        0  91          0          0
## 11           1        1        0 101        101          0
## 12           1        1        0  95         95          0
## 13           1        1        0 105        105          0
## 14           1        1        0  94         94          0
## 15           1        1        0 101        101          0
## 16           1        1        0 126        126          0
## 17           1        1        0 107        107          0
## 18           1        1        0 104        104          0
## 19           1        1        0 120        120          0
## 20           1        1        0  95         95          0
## 21           1        0        1 108          0        108
```

```
## 22           1        0        1  99           0          99
## 23           1        0        1 126           0         126
## 24           1        0        1 119           0         119
## 25           1        0        1 109           0         109
## 26           1        0        1 110           0         110
## 27           1        0        1 105           0         105
## 28           1        0        1 119           0         119
## 29           1        0        1 128           0         128
## 30           1        0        1  94           0          94
## attr(,"assign")
## [1] 0 1 1 2 3 3
## attr(,"contrasts")
## attr(,"contrasts")$class.f
## [1] "contr.treatment"
```

Model matrix - `contr.sum`

```
model.matrix(~class.f*iq,data=maths,
contrasts.arg = list(class.f="contr.sum"))
```

```
##    (Intercept) class.f1 class.f2  iq class.f1:iq class.f2:iq
## 1            1        1        0  99          99           0
## 2            1        1        0 103         103           0
## 3            1        1        0 108         108           0
## 4            1        1        0 109         109           0
## 5            1        1        0  96          96           0
## 6            1        1        0 104         104           0
## 7            1        1        0  96          96           0
## 8            1        1        0 105         105           0
## 9            1        1        0  94          94           0
## 10           1        1        0  91          91           0
## 11           1        0        1 101           0         101
## 12           1        0        1  95           0          95
## 13           1        0        1 105           0         105
## 14           1        0        1  94           0          94
## 15           1        0        1 101           0         101
## 16           1        0        1 126           0         126
## 17           1        0        1 107           0         107
## 18           1        0        1 104           0         104
## 19           1        0        1 120           0         120
## 20           1        0        1  95           0          95
## 21           1       -1       -1 108        -108        -108
## 22           1       -1       -1  99         -99         -99
## 23           1       -1       -1 126        -126        -126
## 24           1       -1       -1 119        -119        -119
## 25           1       -1       -1 109        -109        -109
## 26           1       -1       -1 110        -110        -110
## 27           1       -1       -1 105        -105        -105
## 28           1       -1       -1 119        -119        -119
## 29           1       -1       -1 128        -128        -128
## 30           1       -1       -1  94         -94         -94
## attr(,"assign")
## [1] 0 1 1 2 3 3
## attr(,"contrasts")
## attr(,"contrasts")$class.f
## [1] "contr.sum"
```

Interaction in R

With factors and quantitative variables, R uses the `contr.arg` to determine the coding for the factors, and the resulting parameters for the factor.

With additional quantitative variables, the product of the quantitative variables with the factor is included when the * is used in the model formula.

8 Causality

8.1 Causal relationships vs. correlation

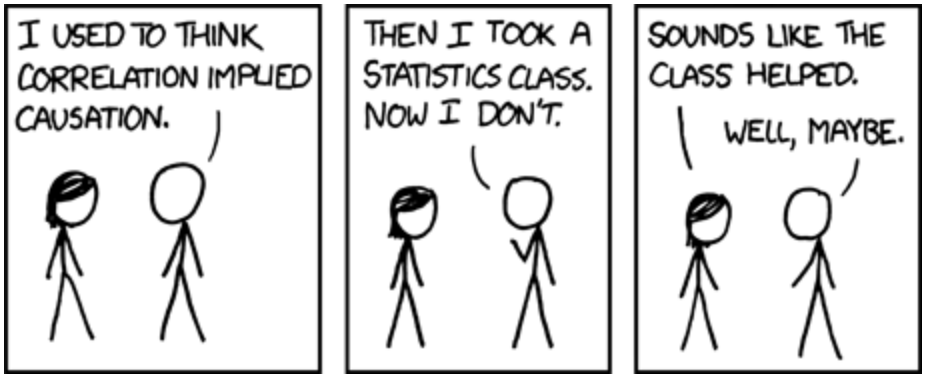
Experimental design

When we conduct an experiment involving two variables A and B, we seek to understand how A and B are related.

Does A cause B (causality)? Or vice versa? Or does something else affect both of them?

An **observational study** can not determine causality, only correlation. Instead you need a **designed experiment/trial**. Even then, we can never be sure.

Correlation does not imply causation!



(xkcd.com/552/, Creative Commons license)

8.2 Hill's criteria

Hill's criteria for determining causal links

- Strength of association
- Consistency of association (from one study to another)
- Consistent with existing knowledge
- Monotonic response - increasing A makes B more likely
- Temporal relationship - A must come before B
- Plausibility of alternatives - is there another explanation?
- Predictive value of link - can you use it?

[A.B. Hill 1897–1991]

A designed experiment can help determin whether there is a monotonic response, a temporal relationship and the plausibility of alternatives.

8.3 Example: false causality

Examples: false causality

Temporal:

- The more fireman the bigger the fire, therefore firemen cause fires.
- The more tourists waiting, the sooner Old Faithful will erupt, therefore tourists cause eruptions.

Inconsistency of association:

- Since the 1600's there has been global warming and a decrease in pirates, therefore pirates help stop global warming.

Examples: false causality

Confounding factor:

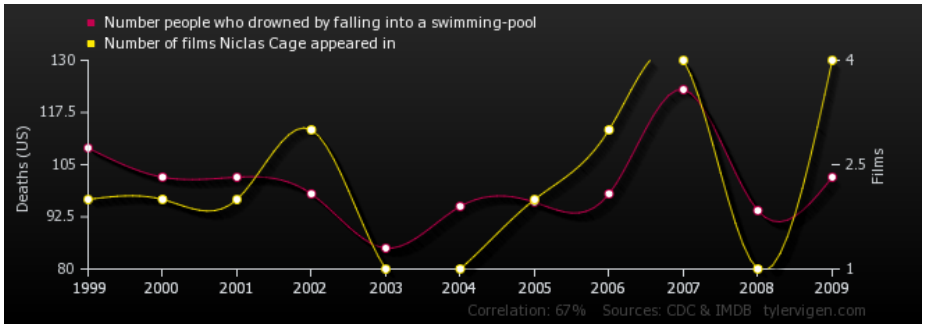
- Sleeping with shoes on causes headaches?
- Heavy drinking causes lung cancer?
- Sleeping with the light on as a child increases the risk of myopia?

In each case there is a better explanation, based on a confounding factor:

- Drinking the night before;
- Passive smoking;
- Myopic parents.

8.4 Example: Nicolas Cage causes drownings

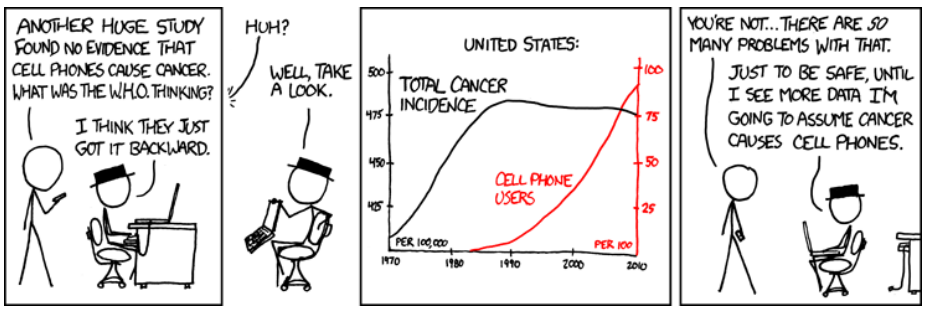
Nicolas Cage causes drownings



(`tylervigen.com/view_correlation?id=359`, Creative Commons license)

8.5 Example: cell phones cause cancer?

Cell phones cause cancer?



(xkcd.com/925/, Creative Commons license)

8.6 Example: life expectancy in Sweden and Panama

Example: life expectancy in Sweden and Panama

"Common sense" tells us that the residents of Sweden should have lower death rates than the residents of Panama.

But each year, a greater proportion of Swedish residents die. Why?

Example: life expectancy in Sweden and Panama

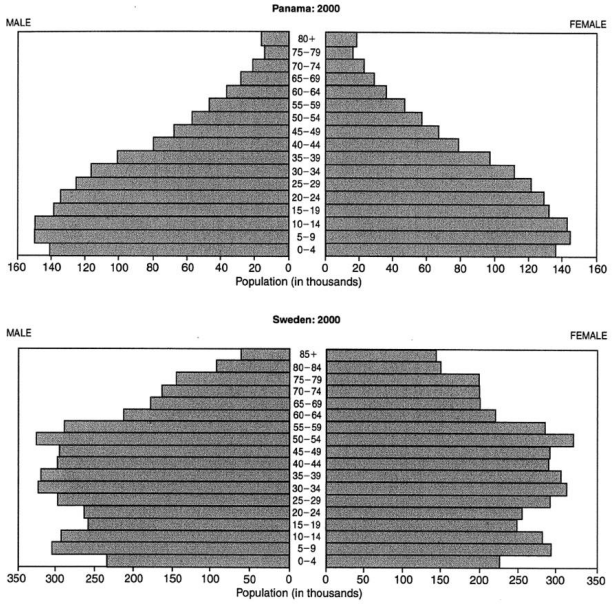
Older people die at a greater rate than younger people.

For individuals of the same age, the death rate among Swedes is less than the death rate among Panamanians.

BUT Sweden has a population that is older than that in Panama, so a greater proportion of Swedes die in any one year, despite the lower death rates within the age categories.

Age is a confounding factor when measuring death rates.

Example: life expectancy in Sweden and Panama



8.7 Example: smoking and xancer

Example: smoking and cancer

Consider the following mortality data, summarised from a study that looked at the smoking habits of a group of female residents of Whickham, UK, in the period 1972-1974, and then tracked their survival over the next twenty years.

smoker	non-smoker	total
139/582	230/732	369/1314
(24%)	(31%)	(28%)

Only 24% of the women who were smokers at the time of the intial survey died during the 20-year follow-up period, whereas 31% of the non-smokers died in the same period.

Does this difference indicate that women who were smokers fared better than women who were not smokers?

Example: smoking and cancer

Here is the same data *stratified* by age (at interview).

age	smoker	non-smoker	total
18–24	2/55 (4%)	1/62 (2%)	3/117 (3%)
25–34	3/124 (2%)	5/157 (3%)	8/281 (3%)
35–44	14/109 (13%)	7/121 (6%)	21/230 (9%)
45–54	27/130 (21%)	12/78 (15%)	39/208 (19%)
55–64	51/115 (44%)	40/121 (33%)	91/236 (39%)
65–74	29/36 (81%)	101/129 (78%)	130/165 (79%)
75+	13/13 (100%)	64/64 (100%)	77/77 (100%)
total	139/582 (24%)	230/732 (31%)	369/1314 (28%)

The death rate for smokers is higher for all age groups except 25-34 (where it is very close), but the age distribution is different for smokers and non-smokers.

Age is a confounding variable in this case.

8.8 Example: vitamin A during pregnancy

Example: vitamin A during pregnancy

To study the relation between the diet of pregnant women and the development of birth defects in their offspring, a US study interviewed more than 22,000 pregnant women early in their pregnancies.

The women were divided into cohorts according to the amount of vitamin A in their diet, from food or from supplements:

daily vitamin A	birth defects
0–5000 IU	51/11083 (0.0046%)
5001–8000 IU	54/10585 (0.0051%)
8001–10000 IU	9/763 (0.0118%)
10001+ IU	7/317 (0.0221%)

Example: vitamin A during pregnancy

These data indicate that the prevalence of these defects increased with increasing intake of vitamin A.

But there is a confounding factor: vitamin A increases the probability of a foetus with defects surviving to full term.

9 Design principles

9.1 4 Principles

Principles of experimental design

Experimental design aims to avoid the effects of confounding factors, known and unknown, by:

- **Control and comparison**
- **Blocking**
- **Randomisation**
- **Blind and double blind testing**

9.2 Control

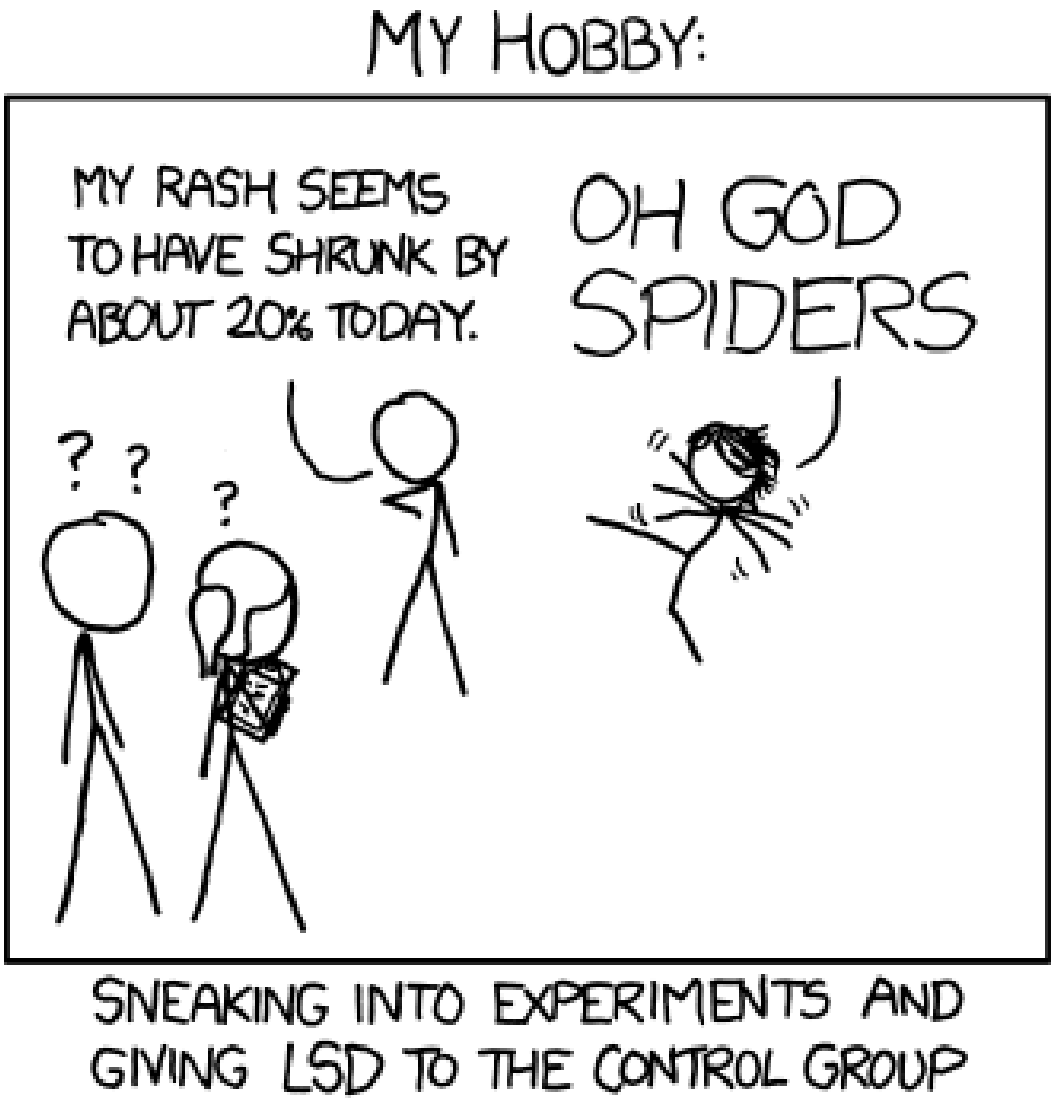
Control

Keep everything the same except the variables you know about. That is, compare groups where you vary just known variables.

For example, all test subjects are non-smoking males aged 18–25, and we compare the effect of two types of diet supplement. The consequence is that the test results only apply to this subpopulation.

Often a group with no treatment is used as a basis for comparison. This is called a **control group**.

Don't do this at home (or anywhere else)!



(xkcd.com/790/, Creative Commons license)

9.3 Blocking

Blocking

Given a known confounding factor, e.g. gender or age, partition the population into blocks which are homogeneous in the confounding variables.

Then, within each block assign treatment and control units, so the effect of the treatment variable can be judged within each block.

9.4 Randomisation

Randomisation

There may be *lurking* factors: confounding factors we are unaware of.

The solution is to randomly assign units to different treatments (possibly within blocks).

Examples: randomisation

A self-selected or voluntary sample is invariably biased.

First Salk polio vaccine trial: In this trial, parents could choose if their children were to participate. Poor families were less likely to participate, but poor children were naturally less likely to get polio.

Literary Digest poll: Readers were asked if they would vote for Roosevelt. Responders were self-selected, and were biased.

Survey	Sample size	Roosevelt's %
Literary Digest	2,400,000	43%
Gallup	50,000	56%
Election		62%

Examples: randomisation

Choosing from a list: Twenty people are selected to trial a new drug. We need to assign ten to the control group and ten to the treatment group.

Their last names are: Chang, Chao, Cheng, Chou, Chu, Gordon, Hsu, Huang, Hu, Li, Liu, MacGregor, MacIntosh, MacKenzie, MacMillan, Munro, Murray, Shannon, Stewart, Urquhart.

Choosing names alphabetically would introduce a cultural bias (for example, diet, which depends on your cultural background, could affect the drug's performance). Randomly assigning people to each group avoids this.

9.5 Blind and double blind testing

Blind and double blind testing

In a blind experiment, the patients do not know if they are being treated or not.

In a double blind experiment, neither the patients nor those administering treatments know which treatment is being given to whom.

We do this to avoid response bias and the placebo effect.

Example: mild polio can be confused with flu, and in the first Salk polio vaccine trial, doctors were more likely to diagnose children who had not been vaccinated.

Blind and double blind testing

Example: gastric freezing was a treatment proposed for ulcer patients. The idea was to reduce acid secretion by cooling the stomach and so relieve ulcers. This was achieved by pumping a freezing liquid into a balloon in the stomach.

An experiment reported in the Journal of the American Medical Association showed that gastric freezing did reduce acid secretion and relieve ulcer pain.

The treatment was safe and easy and was widely used for several years.

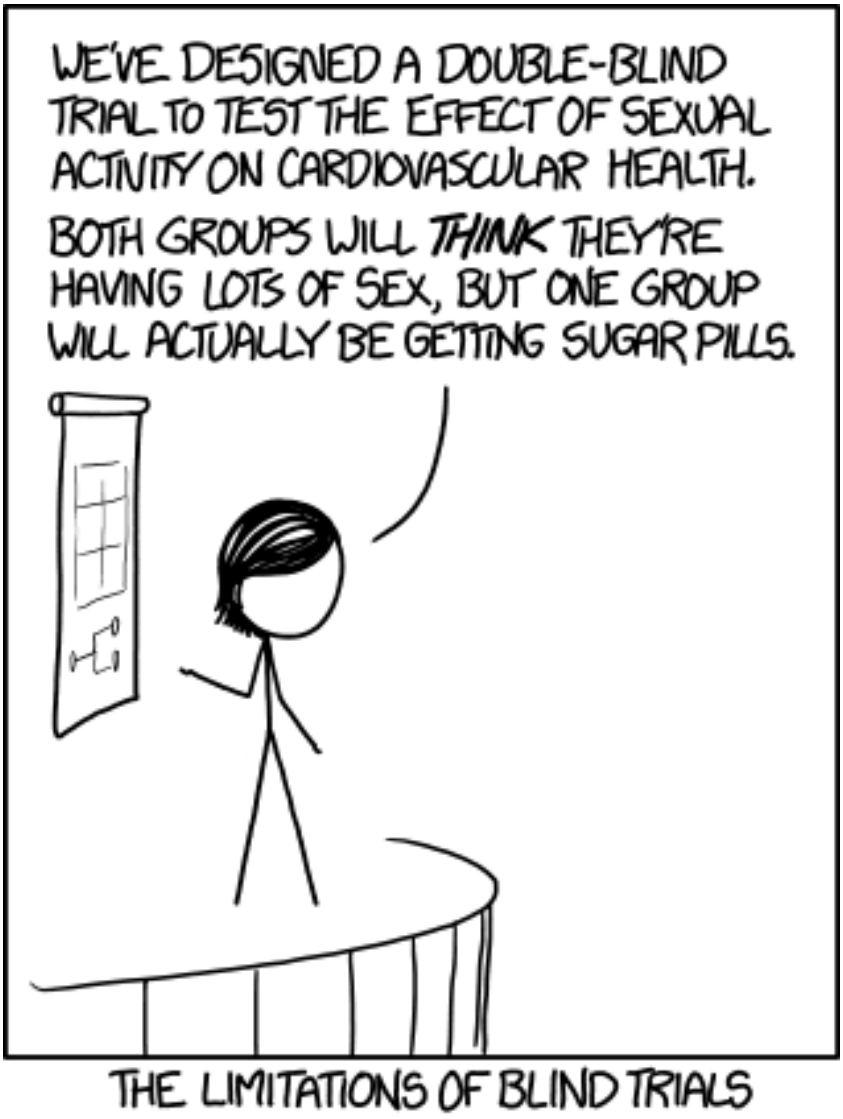
Blind and double blind testing

Unfortunately, the reported effect was just a placebo effect.

A better-designed experiment, done several years later, divided ulcer patients into two groups. One group was treated by gastric freezing as before (the treatment group), while the other group received a placebo treatment in which the solution in the balloon was at body temperature rather than freezing (a control group).

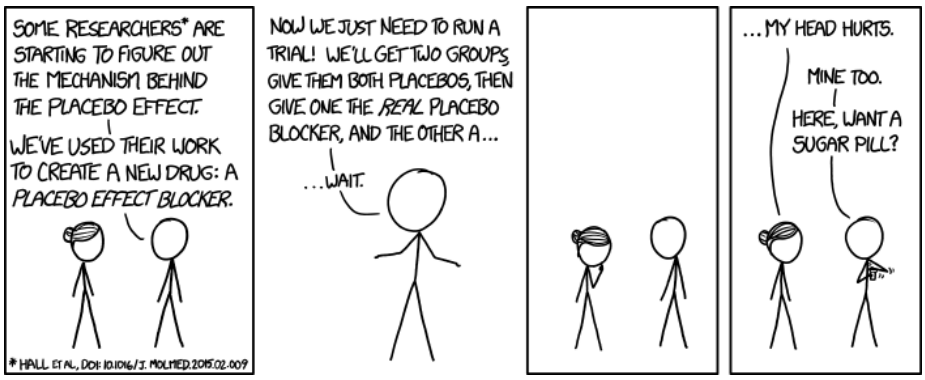
The results: 34% of the 82 patients in the treatment group improved versus 38% of the 78 patients in the control group.

Blinding isn't always possible...



(xkcd.com/1462/, Creative Commons license)

Blinding isn't always possible...



(`xkcd.com/1526/`, Creative Commons license)

9.6 Replication

Replication

Greater precision is achieved by **replication**.

We replicate treatment combinations to minimise the variance of our estimators. Doing this optimally leads to **balanced designs**.

9.7 Types of design

Types of design

Suppose we have a factor whose effect is of interest — the **treatment** — and zero or more confounding factors, which are dealt with by blocking. Any lurking factors are dealt with by randomisation.

- Completely randomised design (CRD)
  no confounding factors (one-way classification)
- (Randomised) complete block design (CBD)
  one confounding/blocking factor (two-way classification)
- Latin squares
  two confounding/blocking factors (three-way classification)
  more efficient than combining the confounding factors
- Balanced incomplete block design (BIBD)
  one confounding factor but CBD impossible

10 Completely randomised design (CRD)

10.1 Examples

Completely randomised design (CRD)

We use a completely randomised design for a single factor (treatment) with k levels.

For example:

- no drug; small amount of drug A; large amount of drug A;
- no drug; drug A; drug B;
- fertiliser A; fertiliser B.

Given n_i test units for factor level i, we assign them randomly.
To analyse, we use a one-way classification model:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} = \mu_i + \varepsilon_{ij}.$$

10.2 How to choose at random

Choosing test units at random

```
n <- c(5,6,4)
nsum <- sum(n)
x <- sample(nsum, nsum)
(j1 <- x[1:n[1]])

## [1]  2 12  6  9  3

(j2 <- x[(n[1]+1):(n[1]+n[2])])

## [1] 11  1  7 10  5  4

(j3 <- x[(n[1]+n[2]+1):nsum])

## [1] 13 15 14  8
```

Choosing test units at random

$$\begin{bmatrix} y_2 \\ y_{12} \\ y_6 \\ y_9 \\ y_3 \\ y_{11} \\ y_1 \\ y_7 \\ y_{10} \\ y_5 \\ y_4 \\ y_{13} \\ y_{15} \\ y_{14} \\ y_8 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_2 \\ \varepsilon_{12} \\ \varepsilon_6 \\ \varepsilon_9 \\ \varepsilon_3 \\ \varepsilon_{11} \\ \varepsilon_1 \\ \varepsilon_7 \\ \varepsilon_{10} \\ \varepsilon_5 \\ \varepsilon_4 \\ \varepsilon_{13} \\ \varepsilon_{15} \\ \varepsilon_{14} \\ \varepsilon_8 \end{bmatrix}$$

$$\mathbf{y} \quad = \quad X \quad \boldsymbol{\beta} \quad + \quad \boldsymbol{\varepsilon}$$

Completely randomised design (CRD)

In a one-factor model, we have already seen that $\hat{\mu}_i = \bar{y}_i$, the average of the responses in group i.

What is its variance? We have (for example)

$$\begin{aligned} \mu_1 &= \begin{bmatrix} 1 & 1 & 0 & \dots & 0 \end{bmatrix}^T \boldsymbol{\beta} \\ \text{var } \hat{\mu}_1 &= \begin{bmatrix} 1 & 1 & 0 & \dots & 0 \end{bmatrix} (X^T X)^c \begin{bmatrix} 1 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \sigma^2 \end{aligned}$$

Completely randomised design (CRD)

$$\begin{aligned} \text{var } \hat{\mu}_1 &= \begin{bmatrix} 1 & 1 & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & \frac{1}{n_1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{n_k} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \sigma^2 \\ &= \begin{bmatrix} 1 & 1 & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} 0 \\ \frac{1}{n_1} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \sigma^2 \\ &= \frac{\sigma^2}{n_1}, \end{aligned}$$

and this is just the formula for the variance of a sample mean. Equally var $\mu_i = \frac{\sigma^2}{n_i}$.

10.3 Optimality

Completely randomised design (CRD)

Theorem 6.11. *In a completely randomised design with n test units, the allocation of test units to factor levels which minimises the total variance for all the estimates of the group means:*

$$\sum_{i=1}^{k} var\ \hat{\mu}_i = \sigma^2 \sum_{i=1}^{k} \frac{1}{n_i}$$

is

$$n_i = \frac{n}{k}$$

(assuming n is a multiple of k).

Completely randomised design (CRD)

Proof: We use Lagrangian multipliers to deal with the constraint in sample size. Take

$$f(n_1, \ldots, n_k, \lambda) = \sigma^2 \sum_{i=1}^{k} \frac{1}{n_i} + \lambda \left(\sum_{i=1}^{k} n_i - n \right).$$

We minimise this function with respect to all variables; the equation $\frac{\partial f}{\partial \lambda} = 0$ ensures that the total sample size is constrained to n.

Completely randomised design (CRD)

The remaining equations give

$$\begin{aligned} \frac{\partial f}{\partial n_i} &= -\frac{\sigma^2}{n_i^2} + \lambda = 0 \\ n_i^2 &= \frac{\sigma^2}{\lambda}. \end{aligned}$$

This does not depend on i, so in order to satisfy these equations we must choose all n_i equal.

The same method can be used to calculate *optimal allocations* when we wish to minimise the variance of other statistics.

Completely randomised design (CRD)

Example. Suppose we have 4 treatments and want to study the treatment contrasts $\tau_2 - \tau_1$, $\tau_3 - \tau_1$ and $\tau_4 - \tau_1$. We have

$$\text{var } \widehat{\tau_i - \tau_1} = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_i} \right),$$

so we minimise

$$f(n_1, n_2, n_3, n_4, \lambda) = \sigma^2 \left(\frac{3}{n_1} + \frac{1}{n_2} + \frac{1}{n_3} + \frac{1}{n_4} \right) + \lambda \left(\sum_{i=1}^{4} n_i - n \right).$$

Completely randomised design (CRD)

We get (for $i \neq 1$):

$$\begin{aligned}
\frac{\partial f}{\partial n_1} &= -3\frac{\sigma^2}{n_1^2} + \lambda = 0 \\
n_1^2 &= 3\frac{\sigma^2}{\lambda} \\
\\
\frac{\partial f}{\partial n_i} &= -\frac{\sigma^2}{n_i^2} + \lambda = 0 \\
n_i^2 &= \frac{\sigma^2}{\lambda} \\
n_1^2 &= 3n_i^2 \\
n_1 &= \sqrt{3}n_i.
\end{aligned}$$

Completely randomised design (CRD)

Therefore

$$\begin{aligned}
n_1 + 3\frac{n_1}{\sqrt{3}} &= n \\
n_1 &= \frac{n}{1+\sqrt{3}} \\
n_i &= \frac{n}{3+\sqrt{3}}.
\end{aligned}$$

For example if we have $n = 20$, then rounding gives $n_1 = 8$ and $n_2 = n_3 = n_4 = 4$.

This is a reflection of the fact that all the contrasts involve τ_1, so it is more important to estimate that accurately.