stats546.02 Analyses of variance and covariance

Many

August 13, 2015

Contents

1 Analysis of variance one and two factors

1.1 Factors and levels

A factor is a classification (categorical) variable such as a farm, gender, color and so forth. The possible values which a factor can take on are called levels. For example color may be red, blue, green and so forth.

A factor is a classification (categorical) variable such as a farm, gender, color and so forth. The possible values which a factor can take on are called levels. For example color may be red, blue, green and so forth.

1.2 Classification variables - two groups

When comparing two means the basic model is

$y_i = \beta_1 + e_i,\ i = 1, \dots n$
$y_i = \beta_2 + e_i,\ i = n+1 \dots m$

Note that the **X**-matrix can be of arbitrary form. In particular one can define classification variables:

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} \begin{matrix} 1 \\ 1 \\ \vdots \\ n \\ n+1 \\ n+2 \\ \vdots \\ n+m \end{matrix}$$

i.e. $y = \mathbf{X}\beta + \mathbf{e}$ is equivalent to the above model, which concerns estimation or comparisons of two means.

The linear models, $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$ allow quite general special cases.

As an example, take the comparison of two groups and assume the basic model is

$y_i = \beta_1 + e_i,\ i = 1, \dots n$
$y_i = \beta_2 + e_i,\ i = n+1 \dots m$

Note that the **X**-matrix can be of any form. In particular the columns do not have to correspond to continuous measurements. It is therefore possible to define **categorical variables**:

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} \begin{matrix} 1 \\ 1 \\ \vdots \\ n \\ n+1 \\ n+2 \\ \vdots \\ n+m \end{matrix}$$

i.e. $y = \mathbf{X}\beta + \mathbf{e}$ is equivalent to the above model, which concerns estimation or comparisons of two means.

1.3 Classification variables - another representation

One could also write

$$
\begin{aligned}
y_i &= \mu + e_i & 1 \le i \le n \\
y_i &= \mu + \beta + e_i & n+1 \le i \le n+m
\end{aligned}
$$

and $H_0 : \mu_1 = \mu_2$ becomes $H_0 : \beta = 0$.

$$
\mathbf{X} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ \vdots & 0 \\ 1 & 1 \\ \vdots & 1 \\ 1 & 1 \end{bmatrix}, \mathbf{Z} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}
$$

One could also write

$$
\begin{aligned}
y_i &= \mu + e_i & 1 \le i \le n \\
y_i &= \mu + \beta + e_i & n+1 \le i \le n+m
\end{aligned}
$$

and the original null hypothesis, $H_0 : \mu_1 = \mu_2$ becomes $H_0 : \beta = 0$.

In matrix notation we have

$$
\mathbf{X} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ \vdots & 0 \\ 1 & 1 \\ \vdots & 1 \\ 1 & 1 \end{bmatrix}, \mathbf{Z} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}
$$

1.4 Simple analysis of variance

Several groups

$$
\begin{aligned}
y_{1j} &= \mu_1 + e_{1j} & j = 1, \ldots, J_1 \\
y_{2j} &= \mu_2 + e_{2j} & j = 1, \ldots, J_2 \\
&\vdots \\
y_{Ij} &= \mu_I + e_{Ij} & j = 1, \ldots, J_I,
\end{aligned}
$$

with a total of $n = J_1 + \ldots + J_I$ measurements.

In addition to simple comparisons of two means, i.e. tests of $H_0 : \mu_1 = \mu_2$ with data of the form

$$
\begin{aligned}
y_i &= \mu_1 + e_i & i = 1, \ldots, n \\
y_i &= \mu_2 + e_i & i = n+1, \ldots, n+m
\end{aligned}
$$

it is also of interest to compare several means.

Thus we want to consider data from several (I) groups.

$$
\begin{aligned}
y_{1j} &= \mu_1 + e_{1j} \quad j = 1, \ldots, J_1 \\
y_{2j} &= \mu_2 + e_{2j} \quad j = 1, \ldots, J_2 \\
&\vdots \\
y_{Ij} &= \mu_I + e_{Ij} \quad j = 1, \ldots, J_I,
\end{aligned}
$$

with a total of $n = J_1 + \ldots + J_I$ measurements.

In addition to simple comparisons of two means, i.e. tests of $H_0 : \mu_1 = \mu_2$ with data of the form

$$
\begin{aligned}
y_i &= \mu_1 + e_i \qquad i = 1, \ldots, n \\
y_i &= \mu_2 + e_i \quad i = n+1, \ldots, n+m
\end{aligned}
$$

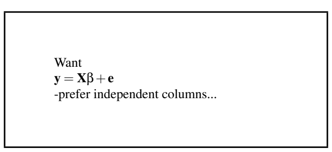
it is also of interest to compare several means.

Thus we want to consider data from several (I) groups.

$$
\begin{aligned}
y_{1j} &= \mu_1 + e_{1j} \quad j = 1, \ldots, J_1 \\
y_{2j} &= \mu_2 + e_{2j} \quad j = 1, \ldots, J_2 \\
&\vdots \\
y_{Ij} &= \mu_I + e_{Ij} \quad j = 1, \ldots, J_I,
\end{aligned}
$$

with a total of $n = J_1 + \ldots + J_I$ measurements.

1.5 Developing matrix notation



The models are set up using matrix notation,

- usually omit those columns in $\mathbf{X}$ which would make them linearly dependent (also set the corresponding elements of the β-vector to zero without further estimation).

The models are set up using matrix notation, $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$, usually omitting those columns in $\mathbf{X}$ which would make them linearly dependent (also set the corresponding elements of the β-vector to zero without further estimation).

1.6 Different versions of the same model

The model can be written in different ways, e.g.

$$
\begin{aligned}
y_{1j} &= \mu + \alpha_1 + e_{1j}, \quad j = 1, \ldots, J_1 \\
y_{2j} &= \mu + \alpha_2 + e_{2j}, \quad j = 1, \ldots, J_2 \\
&\vdots \\
y_{Ij} &= \mu + \alpha_I + e_{Ij}, \; j = 1, \ldots, J_I.
\end{aligned}
$$

Here, μ is an overall mean but α_i is the deviance of each group from the overall mean.

The model can be written in different ways, e.g.

$$\begin{aligned}
y_{1j} &= \mu + \alpha_1 + e_{1j}, \quad j = 1, \ldots, J_1 \\
y_{2j} &= \mu + \alpha_2 + e_{2j}, \quad j = 1, \ldots, J_2 \\
&\vdots \\
y_{Ij} &= \mu + \alpha_I + e_{Ij}, \; j = 1, \ldots, J_I.
\end{aligned}$$

Here, μ is an overall mean but α_i measures how much each group mean deviates from the overall mean.

1.7 Deviations from overall mean in matrix form

This model can be written using matrix notation as:

$$\mathbf{y} = \begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1J_1} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2J_2} \\ \vdots \\ y_{I1} \\ y_{I2} \\ \vdots \\ y_{IJ_I} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & & 0 \\ 1 & 1 & 0 & & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & & 0 \\ 1 & 0 & 1 & & 0 \\ 1 & 0 & 1 & & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 1 & & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & & 1 \\ 1 & 0 & 0 & & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_I \end{bmatrix} + \mathbf{e}$$

This model can be written using matrix notation as:

$$\mathbf{y} = \begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1J_1} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2J_2} \\ \vdots \\ y_{I1} \\ y_{I2} \\ \vdots \\ y_{IJ_I} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & & 0 \\ 1 & 1 & 0 & & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & & 0 \\ 1 & 0 & 1 & & 0 \\ 1 & 0 & 1 & & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 1 & & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & & 1 \\ 1 & 0 & 0 & & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_I \end{bmatrix} + \mathbf{e}$$

1.8 Null hypotheses, several means

The null hypothesis

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

is the same as

$H_0 : \alpha_1 = \ldots = \alpha_I = 0.$

The alternative hypothesis H_a is simply that H_0 is not correct.

We are interested in testing null hypotheses concerning the means.

The primary null hypothesis becomes

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

and this is the same as

$H_0 : \alpha_1 = \ldots = \alpha_I = 0.$

The alternative hypothesis H_a is simply that H_0 is not correct.

1.9 Dependent column vectors of X

Note now that the columns of $\mathbf{X}$ are dependent so that $(\mathbf{X}'\mathbf{X})^{-1}$ does not exist. Therefore columns must be dropped or some other conditions set in order to find a solution.

Note now that the columns of $\mathbf{X}$ are dependent so that $(\mathbf{X}'\mathbf{X})^{-1}$ does not exist. Therefore columns must be dropped or some other conditions set in order to find a solution.

It is simplest to drop columns. SAS and similar packages simply drop the columns "as they come".

1.10 Point estimates

One solution...

$$\mu_i = \mu + \alpha_i$$

$$\sum_i \alpha_i = 0$$

$$J_i = J$$

$$\hat{\mu}_i = \bar{y}_{i.}$$
$$\hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}$$

Assume the sample sizes are equal, $J_i = J$ and the model formulation is

$$\mu_i = \mu + \alpha_i.$$

In this case some restriction is needed in order to make the parameters estimable, or uniquely defined. When sample sizes are equal, the usual constraint is

$$\sum_i \alpha_i = 0$$

In this case the point estimates are easy to derive, e.g. using a Lagrangian.

$$\hat{\mu}_i = \bar{y}_{i.}$$

$$\hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}$$

1.11 The sum of squares is well-defined

$$SSE = \sum_{ij} (y_{ij} - \hat{y}_{ij})^2 = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (y_{ij} - \bar{y}_{i.})^2$$

where

$$\bar{y}_{i.} = \frac{1}{J_i} \sum_{j=1}^{J_i} y_{ij}.$$

We also know that

$$SSTOT = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (y_{ij} - \bar{y}_{..})^2$$

so the following variation is explained by the model

$$SSR = SSTOT - SSE = \ldots = \sum_{ij} (\bar{y}_{i.} - \bar{y}_{..})^2 \sum_i J_i (\bar{y}_{i.} - \bar{y}_{..})^2$$

Alternative estimates of the parameters can of course be obtained since the original problem was not uniquely defined. On the other hand, the values of $\hat{y}_{ij}$ will always be unique.

Therefore the sums of squares are well-defined. They are also easy to compute, regardless of how the matrix is simplified or a specific solution is found.

Upon estimation of the coefficients in the model the following variation is unexplained:

$$SSE = \sum_{ij} (y_{ij} - \hat{y}_{ij})^2 = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (y_{ij} - \bar{y}_{i.})^2$$

where

$$\bar{y}_{i.} = \frac{1}{J_i} \sum_{j=1}^{J_i} y_{ij}.$$

This is a relatively simple conclusion when considering the corresponding projections.

We also know that

$$SSTOT = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (y_{ij} - \bar{y}_{..})^2$$

so the following variation is explained by the model

$$SSR = SSTOT - SSE = \ldots$$

1.12 Components of sums of squares

The residuals add up and so do the sums of squares:

$$y_{ij} - \bar{y}_{..} = (y_{ij} - \bar{y}_{i.}) + (\bar{y}_{i.} - \bar{y}_{..})$$

$$\sum_{ij}(y_{ij}-\bar{y}_{..})^2=\sum_{ij}(y_{ij}-\bar{y}_{i.})^2+\sum_{ij}(\bar{y}_{i.}-\bar{y}_{..})^2$$

The orthogonality of the deviations implies that the corresponding sums of squares add up.

$$y_{ij}-\bar{y}_{..}=(y_{ij}-\bar{y}_{i.})+(\bar{y}_{i.}-\bar{y}_{..})$$

$$\sum_{ij}(y_{ij}-\bar{y}_{..})^2=\sum_{ij}(y_{ij}-\bar{y}_{i.})^2+\sum_{ij}(\bar{y}_{i.}-\bar{y}_{..})^2$$

This is not too hard to derive since the sums of products sum nicely to zero. Alternatively, note that the left hand side is SSTOT which corresponds to the model $E[y_{ij}]=\mu$ which is a submodel of $E[y_{ij}]=\mu_i$ which gives the second term, SSE, on the right hand side of the equation. The deviations themselves clearly correspond to the corresponding projections and hence they must be orthogonal.

1.13 One-way anova

The ANOVA table becomes

	df	SS	MS	F
Model	$I-1$	$SSR=\sum_{i=1}^{I}J_i(\bar{y}_{i.}-\bar{y}_{..})^2$	$MSR=SSR/(I-1)$	$F=MSR/MSE$
Error	$n-I$	$SSE=\sum_{i=1}^{I}\sum_{j=1}^{J_i}(y_{ij}-\bar{y}_{i.})^2$	$MSE=SSE/(n-I)$	
Total	$n-1$	$SSTOT=\sum_{i=1}^{I}\sum_{j=1}^{J_i}(y_{ij}-\bar{y}_{..})^2$		

We will reject H_0 if $F>F_{I-1,n-I,1-\alpha}$

The ANOVA table becomes

	df	SS	MS	F
Model	$I-1$	$SSR=\sum_{i=1}^{I}J_i(\bar{y}_{i.}-\bar{y}_{..})^2$	$MSR=SSR/(I-1)$	$F=MSR/MSE$
Error	$n-I$	$SSE=\sum_{i=1}^{I}\sum_{j=1}^{J_i}(y_{ij}-\bar{y}_{i.})^2$	$MSE=SSE/(n-I)$	
Total	$n-1$	$SSTOT=\sum_{i=1}^{I}\sum_{j=1}^{J_i}(y_{ij}-\bar{y}_{..})^2$		

We will reject H_0 if $F>F_{I-1,n-I,1-\alpha}$

Example: Suppose we have a small data set for testing a single factor (classification variable). There are 3 different values of the factor so in effect 3 means are being compared, i.e. the hypothesis to be tested is $H_0:\mu_1=\mu_2=\mu_3$.

Assume that the measurements are obtained from independent normally distributed random variables.

```
/*
* Example of using SAS for one-way ANOVA
*   The data
*
*     1     2     3
*  0.97 -1.16 -0.06
*  0.68 -2.08  1.89
*  0.41  1.19  0.32
*/
options linesize=120;
```

```
data;
  input f y;
  datalines;
 1  0.97
 2 -1.16
 3 -0.06
 1  0.68
 2 -2.08
 3  1.89
 1  0.41
 2  1.19
 3  0.32
proc glm;
  classes f;
  model y=f;
run:
```

The SAS run gives the following output:

```
                                  The SAS System       11:50 Thursday, November 1, 200

                                The GLM Procedure

                             Class Level Information

                          Class         Levels    Values

                          f                  3    1 2 3


                           Number of observations    9
                                  The SAS System       11:50 Thursday, November 1, 200

                                The GLM Procedure

Dependent Variable: y

                                                Sum of
        Source                      DF         Squares     Mean Square    F Value    Pr > F

        Model                        2      3.83780000      1.91890000       1.44    0.3079

        Error                        6      7.98140000      1.33023333

        Corrected Total              8     11.81920000


                         R-Square     Coeff Var      Root MSE        y Mean

                         0.324709      480.5656      1.153357      0.240000


        Source                      DF       Type I SS     Mean Square    F Value    Pr > F

        f                            2      3.83780000      1.91890000       1.44    0.3079
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
f	2	3.83780000	1.91890000	1.44	0.3079

2 Distributions and expectations in the one-way layout

2.1 Distributions

It is of interest to consider the distributions of various quantities, not only under $H_0 : \mu_1 = \ldots = \mu_I$ but also when H_0 does not hold. Assume, therefore that

$$y_{ij} \sim n(\mu_i, \sigma^2),\ 1 \le j \le J_i,\ 1 \le i \le I, \text{ i.i.d.}$$

In particular, y_{ij} independent with $Ey_{ij} = \mu_i$ and $Vy_{ij} = \sigma^2$.

We then have $\bar{y}_{i.} = \frac{\sum_j y_{ij}}{J_i}$ with expected value

$$E\left[\bar{y}_{i.}\right] = \mu_i$$

and variance

$$V\left[\bar{y}_{i.}\right] = \sigma^2/J_i$$

and under normality the estimators $\bar{y}_{i.}$ have the obvious properties

$$\bar{y}_{i.} \sim n(\mu_i, \sigma^2/J_i)$$

and these are independent.

It follows in particular that

$$E\left[\bar{y}_{i.}^2\right] = \mu_i^2 + \sigma^2/J_i,$$

which will be needed later.

Let

$$\mu = \frac{\sum_i J_i \mu_i}{n}$$

where $n = \sum_i J_i$.

Since $\bar{y}_{..}$ can be written as

$$\frac{\sum_i J_i \bar{y}_{i.}}{\sum_i J_i}$$

it follows trivially that

$$E\left[\bar{y}_{..}\right] = \mu$$

and

$$V\left[\bar{y}_{..}\right] = V\left[\frac{\sum_{ij} y_{ij}}{n}\right] = \sigma^2/n$$

We thus obtain

$$\bar{y}_{..} \sim n(\mu, \sigma^2/n)$$

but of course the values of the various expected values are different when H_0 is not true.

It is of interest to consider the distributions of various quantities, not only under $H_0 : \mu_1 = \ldots = \mu_I$ but also when H_0 does not hold. Assume, therefore that

$$y_{ij} \sim n(\mu_i, \sigma^2),\ 1 \le j \le J_i,\ 1 \le i \le I, \text{ i.i.d.}$$

In particular, y_{ij} independent with $Ey_{ij} = \mu_i$ and $Vy_{ij} = \sigma^2$.

We then have $\bar{y}_{i.} = \frac{\sum_j y_{ij}}{J_i}$ with expected value

$$E\left[\bar{y}_{i.}\right] = \mu_i$$

and variance

$$V\left[\bar{y}_{i.}\right] = \sigma^2 / J_i$$

and under normality the estimators $\bar{y}_{i.}$ have the obvious properties

$$\bar{y}_{i.} \sim n(\mu_i, \sigma^2 / J_i)$$

and these are independent.

It follows in particular that

$$E\left[\bar{y}_{i.}^2\right] = \mu_i^2 + \sigma^2 / J_i,$$

which will be needed later.

Let

$$\mu = \frac{\sum_i J_i \mu_i}{n}$$

where $n = \sum_i J_i$.

Since $\bar{y}_{..}$ can be written as

$$\frac{\sum_i J_i \bar{y}_{i.}}{\sum_i J_i}$$

it follows trivially that

$$E\left[\bar{y}_{..}\right] = \mu$$

and

$$V\left[\bar{y}_{..}\right] = V\left[\frac{\sum_{ij} y_{ij}}{n}\right] = \sigma^2 / n$$

so we obtain

$$\bar{y}_{..} \sim n(\mu, \sigma^2 / n).$$

It is important to remember that the values of the various expected values are different when H_0 is not true. For example, μ is a linear combination of **different** μ_i in this case.

2.2 The expected MSR

Can obtain

$$E\left[MSR\right] = \sigma^2 + \frac{\sum_i J_i (\mu_i - \mu)^2}{I - 1}$$

in one-way layout.

Can obtain $E[MSR]$ in one-way layout.

Note that we have y_{ij} independent with $Ey_{ij} = \mu_i$ and $Vy_{ij} = \sigma^2$.

Let
$$\mu = \frac{\sum_i J_i \mu_i}{n}$$
where $n = \sum_i J_i$.

Correspondingly $\bar{y}_{..}$ can be written as
$$\bar{y}_{..} = \frac{\sum_i J_i \bar{y}_{i.}}{\sum_i J_i}$$
and is thus a linear combination of the $\bar{y}_{i.}$. We can therefore find the mean and variance of $\bar{y}_{..}$.

Now look at
$$E\left[(\bar{y}_{i.} - \bar{y}_{..})^2\right]$$

first note that this is not just a simple quadratic corresponding to a sample variance since the $\bar{y}_{i.}$ are not i.i.d. Hence we need to square this and use the formulae for cross-products etc and then compute the corresponding expected values.

It follows that
$$E[MSR] = \sigma^2 + \frac{\sum_i J_i (\mu_i - \mu)^2}{I-1}$$

and one should note that this is equal to σ^2 when and only when the means are all the same, $\mu_i = \mu \forall i$.

From this it is also clear that $E[MSR]$ is uniformly larger than $E[MSE]$ unless the means are all equal, further justifying rejection of the null hypotheses only for large values of the F-statistic.

3 Topics in one-way analysis of variance

3.1 Plotting factor level means

Plotting factor level means

Plotting factor level means

3.2 Diagnostics

Diagnostics

Diagnostics

3.3 Error variance

Error variance

Error variance

3.4 Normality

Normality

Normality

4 The two-way layout

4.1 Two-way layout basics

Have two factors so factors make a table of level combinations μ_{ij}

Many possible two-way scenarios:

Two additive effects

Single observation per cell

Multiple observations + interactions in effects

Have two factors so factors make a table of level combinations μ_{ij}

Many possible two-way scenarios:

Two additive effects

Single observation per cell

Multiple observations + interactions in effects

Can in either 1-way or 2-way layout use plots of means to decide whether reg. fcn. is appropriate (if x's are quantitative but repeated).

See fig 17.6 p. 745 and fig. 20.5, p. 867 in book.

NB Can also do contour plots in 2-way layout.

Example: Simple generation of data for a two-way layout. Note how the "x" in this can be viewed either as a factor or a continuous variable.

```
set.seed(1)
x<-rep(1:4,c(6,6,6,6))
truvals<-c(1,4,2)
names(truvals)<-c("A","B","C")
w<-rep(truvals,8)
rbind(x,w)
f<-factor(rep(names(truvals),8))
n<-length(x)
y<-2+0.5*x+w+0.1*w*x+rnorm(n,0,0.1)
xf<-factor(x)
dat<-data.frame(y,xf,f)
```

It is useful to look at the layout, compute means etc before going further...

```
table(xf,f)
tapply(y,list(xf,f),mean)
```

Analysis of variance in the two-way laout is done with:

```
summary(aov(y~xf+f+xf:f,data=dat))
```

Example: Consider the two-way layout with one observation per cell,

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \; i = 1,\ldots,I, \text{ and } \; j = 1,\ldots,J$$

with $I = 2$ and $J = 3$ and corresponding X-matrix in R

```
> X
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    1    1    0    1    0    0
[2,]    1    1    0    0    1    0
[3,]    1    1    0    0    0    1
[4,]    1    0    1    1    0    0
[5,]    1    0    1    0    1    0
[6,]    1    0    1    0    0    1
```

where the model is now written $\mathbf{y} = \mathbf{X}\beta + \varepsilon$.

Adding constraints of the form $\alpha_1 + \alpha_2 = 0$ and$\beta_1 + \beta_2 + \beta = 0$ corresponds to $\mathbf{H}\beta = \mathbf{0}$ where

```
> H
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    0    1    1    0    0    0
[2,]    0    0    0    1    1    1
```

and the solutions will be based on inverting the matrix

$$\mathbf{G}'\mathbf{G} = \mathbf{X}'\mathbf{X} + \mathbf{H}'\mathbf{H}$$

which is

```
> t(X)%*%X+t(H)%*%H
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    6    3    3    2    2    2
[2,]    3    4    1    1    1    1
[3,]    3    1    4    1    1    1
[4,]    2    1    1    3    1    1
[5,]    2    1    1    1    3    1
[6,]    2    1    1    1    1    3
```

and the inverse (times 36) is

```
> solve(t(X)%*%X+t(H)%*%H)*36
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]   19   -9   -9   -4   -4   -4
[2,]   -9   15    3    0    0    0
[3,]   -9    3   15    0    0    0
[4,]   -4    0    0   16   -2   -2
[5,]   -4    0    0   -2   16   -2
[6,]   -4    0    0   -2   -2   16
```

It is now not too hard to see that the solution,

$$\hat{\beta} = \left(\mathbf{G}'\mathbf{G}\right)\mathbf{X}'\mathbf{y}$$

is the usual LS solution.

5 The single replicate two-way layout

5.1 Estimations in the two-way layout

Versions of two way ANOVA

Two-way analysis of variance, or two-factor anova, refers to the existence of two different factor or effects, A and B, which is some manner affect the mean of the response.

The model:

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

where the ε are assumed i.i.d. with mean zero and some variance, σ^2.

There is exactly one measurement for each combination of factor levels, hence the term single-replicate.

The effects α_i and β_j are called the **main effects**.

The usual restriction is $\sum \alpha_i = \sum \beta_j = 0$.

The LS estimates under the restriction are not difficult to obtain:

$$\hat{\mu} = \dots$$

$$\hat{\alpha}_i = \dots$$

$$\hat{\beta}_j = \dots$$

and the predicted values are

$$\hat{y}_{ij} = \dots$$

from which the *SSE* follows.

5.2 Slide number 10

There are in this case two hypotheses of interest,

$$H_{0A}$$

and

$$H_{0B}$$

The F-tests for each hypothesis can be derived based on considering the estimates under the corresponding reduced models and computing differences in sums of squares. The SSE under the reduced model for H_{0A} becomes

$$SSE(R^A) = \sum_{ij} (y_{ij} - \bar{y}_{.j})^2$$

since this is the residual sum of squares under the reduced model with $\alpha_i = 0$ and the resulting model is a one-way anova model.

One can then derive

$$SSA = SSE - SSE(R^A) = \dots$$

One can also write

$$y_{ij} - \bar{y}_{..} = (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}) + (\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{..})$$

and note that the corresponding sums of squares add up neatly and correspond to the above sums of squares:

$$\underbrace{\sum_{ij}(y_{ij}-\bar{y}_{..})^2}_{SSTOT} = \underbrace{\sum_{ij}(y_{ij}-y_{i.}-y_{.j}+\bar{y}_{..})^2}_{SSE} + \underbrace{\sum_{ij}(y_{i.}-\bar{y}_{..})^2}_{SSA} + \underbrace{\sum_{ij}(y_{.j}-\bar{y}_{..})^2}_{SSB}$$

since the cross-product terms vanish.

These sums of squares form the ANOVA tables with df n-1, (I-1)(J-1), I-1 and J-1.

Note that the residual vectors are of the form $\mathbf{A}_1\mathbf{y}$, $\mathbf{A}_2\mathbf{y}$ and $\mathbf{A}_3\mathbf{y}$ and the SSE's are the squared norms of these vectors, e.g.

$$SSA = ||\mathbf{A}_2\mathbf{y}||^2 = \mathbf{y}'\mathbf{A}_2'\mathbf{A}_2\mathbf{y}$$

Each of these matrices is a projection onto the corresponding subspace of $\mathbb{R}^n$.

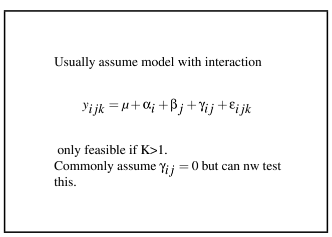
The fact that the cross-product terms vanish implies that e.g. $(\mathbf{A}_1\mathbf{y})\cdot(\mathbf{A}_2\mathbf{y}) = 0$, i.e. $\mathbf{y}'\mathbf{A}_1'\mathbf{A}_2\mathbf{y}) = 0$ for all data vectors $\mathbf{y}$ and hence $\mathbf{A}_1'\mathbf{A}_2 = \mathbf{0}$, i.e. all column vectors in each matrix are orthogonal to all column vectors in each of the other matrices.

Basically, $\mathbf{A}_1\mathbf{y}$, $\mathbf{A}_2\mathbf{y}$ and $\mathbf{A}_3\mathbf{y}$ are orthogonal vectors and hence have zero covariance, implying independence under normality[1].

It follows that the three sums of squares (SSE, SSA and SSB) are all independent.

6 Two-way layout with equal number of observations per cell

6.1 The model and estimates



The form of the interaction effect and constraints

When there are more observations in each cell one normally assumes a model with interaction

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

only feasible if K>1.

Commonly assume $\gamma_{ij} = 0$ but can now test this.

With the usual side conditions one obtains the estimates

$$\hat{\mu} = ...$$

$$\hat{\alpha} = ...$$

$$\hat{\beta} = ...$$

$$\hat{\gamma} = \bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}$$

[1]Warning: This is not a simple consequence of "For U, V with a joint multivariate normal distribution cov(U,v)=0 iff U and V are independent" since here the vectors in question do not have a joint multivariate normal distribution (the joint distribution is degenerate). Thus one needs to construct appropriate bases for the column space of each matrix and proceed from there.

That these are obvious estimators is best seen by looking at the corresponding theoretical quantities as functions of $\mu_{ij} := E\left[y_{ijk}\right]$

$$\mu = \bar{\mu}_{..}$$
$$\alpha := \bar{\mu}_{i.} - \bar{\mu}_{..}$$
$$\beta := \bar{\mu}_{.j} - \bar{\mu}_{..}$$

$$\gamma := \mu_{ij} - \alpha_i - \beta_j = \mu_{ij} - \mu_{i.} - \mu_{.j} + \bar{\mu}_{..}$$

As before one obtains deviations which add up and can go from there:

$$y_{ijk} - y_{ijk} = (y_{i..} - y_{...}) + (y_{.j.} - y_{...}) + (y_{ij.} - y_{i..} - y_{.j.} + y_{...}) + (y_{ijk} - y_{ik.})$$

The corresponding SSA, SSB, SSAB, SSE will add up to SSTOT and these have df of I-1, J-1, (I-1)(J-1) ... and n-1 respectively where n=IJK.

7 Analysis of covariance, including lack of fit tests

7.1 Analysis of covariance

Analysis of covariance:
Factor and continuous variables together
Special case of general linear model

Analysis of covariance

When a linear model includes both continuous and discrete independent variables, i.e. factors and regression variables, the analysis is called **analysis of covariance**.

Example: Consider simulated data with an x-variable and a factor as follows. The factor levels will be termed "A", "B" and "C", but the true effects associated with these levels will be 1, 4 and 2, respectively:

```
> set.seed(1)
> x<-rep(1:4,c(3,3,3,3))
> truvals<-c(1,4,2)
> names(truvals)<-c("A","B","C")
> w<-rep(truvals,4)
> f<-factor(rep(names(truvals),4))
> n<-length(x)
> y<-2+0.5*x+w+0.1*w*x+rnorm(n,0,0.1)
> dat<-data.frame(y,x,f)
```

Having generated the data, we can remove the original variables and just use the data frame.

These simulated data can now be used to test the various R commands and to understand the linear model, analysis of variance tables and so forth.

```
> rm(x,y,f,w)
> drop1(lm(y~x+f,data=dat),test="F")
> drop1(lm(y~f*x),test="F")
> drop1(lm(y~x+f+f:x),test="F")
> summary(aov(y~f))
> summary(lm(y~f))
```

7.2 Lack of fit tests

Simple linear regression: $y_i = \alpha + \beta x_i + e_i$

Want to test whether straight line is OK

Suppose have repeated measuresments at each (most) x-values: $y_{ij} = \alpha + \beta x_i + e_{ij}$

Can design new full model: $y_{ij} = \mu_i + e_{ij}$

Now test full vs reduced

For this SLR case we can write the table for the partitioned SSE (p. 119)

Simple linear regression: $y_i = \alpha + \beta x_i + e_i$

Want to test whether straight line is OK

Suppose have repeated measuresments at each (most) x-values: $y_{ij} = \alpha + \beta x_i + e_{ij}$

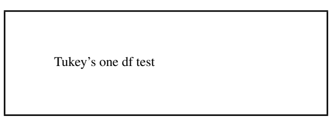
Can design new full model: $y_{ij} = \mu_i + e_{ij}$

Now test full vs reduced

For this SLR case we can write the table for the partitioned SSE (p. 119)

8 Topics

8.1 Slide number 00

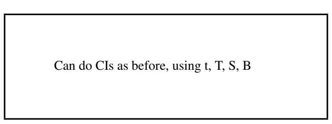


Test whether D=0 in

$$y_{ij} = \mu + \alpha_i + \beta_j + D\alpha_i\beta_j\varepsilon_{ij}$$

(See p. 882 in book)

8.2 Confidence bounds



Can do CIs as before, using t, T, S, B for $\mu_{i.}$ or $\mu_{.j}$ or μ_{ij} etc.

Note: CIs for main effects are NOT of interest in the presence of interactions. Then need to do CIs for $\mu_{ij} - \mu_{i'j'}$ (B, T or S)