Analysis of variance

Gilles Guillot

gigu@dtu.dk

September 30, 2013

Doughnuts data

The doughnuts data contain the quantity of fat absorbed by doughnuts during cooking for various fat types:

```
Fat1 Fat2 Fat3 Fat4
164 178 175 155
172 191 193 166
168 197 178 149
177 182 171 164
156 185 163 170
195 177 176 168
```

Does the fat type induce any *significant* difference in the quantity of fat absorbed?

The mean testing problem as a regression problem

Testing equality of means is ubiquitous in medicine, marketing, quality control etc...

Setting:

- y a quantitative variable
- dependig on a categorical variable with I possible values
- y_{ij} j-th observation receiving "treatment" i

The doughnuts data reformatted

	Quantity	Fat.type
1	164	Fat1
2	178	Fat2
3	175	Fat3
4	155	Fat4
5	172	Fat1
6	191	Fat2
7	193	Fat3
8	166	Fat4
9	168	Fat1
10	197	Fat2
11	178	Fat3
12	149	Fat4
13	177	Fat1

...

One-way ANOVA

The one-way ANOVA model:

- y_{ij} j-th observation having received treatment i
- $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$
- μ deterministic unkown parameter
- α_i deterministic unknown effect of treatment i
- $(\varepsilon_{ij})_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma_e^2)$ unobserved random noise

Alternative notation: $y_i = \mu + \alpha(i) + \varepsilon_i$

Model rationale illustrated in R: `boxplot()`

Linear algebra interpretation

We denote by X the $n \times (I+1)$ matrix with first column containing ones, column j containing 1 for observations that received treatment j. X is the design matrix. Then

$$Y = X\theta + E$$

with $\theta = (\mu, \alpha_1, ..., \alpha_I)$ and $E = (\varepsilon)_{i=1,...,I}$

To see the dummy variables in R:

```
library(dummies)
dummy(x=...)
```

Identifiability issues and parameters interpretation I

The identifiability issue:

Model $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ is equivalent to $y_{ij} = \mu' + \alpha'_i + \varepsilon_{ij}$
with $\mu' = \mu + \delta$ and $\alpha'_i = \alpha_i - \delta$

- There are infinitely many parameter combinations providing the same fit to the data.
- Parameters are meaningless without additional constraints.

Common identifiability constraints under model $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$:

- $\alpha_1 = 0$ (or $\alpha_I = 0$)
  α_i is the departure from mean value under treatment 1 (or I) induced by treatment i.
- $\mu = 0$
  α_i is the mean value under treatment i
- $\sum_i \alpha_i = 0$
  α_i is the departure from the mean induced by treatment i

In R, the default constraint is $\alpha_1 = 0$. The parameter μ is the average response un treatment 1.

Parameter estimation in the one-way ANOVA model I

$I + 2$ parameters: $(\mu, \alpha_1, \alpha_2, ..., \alpha_I)$ and σ
Only $I + 1$ linearly independent parameters

Least square principle:

We seek $(\mu, \alpha_1, \alpha_2, ..., \alpha_I)$ that minimize

$$S(\mu, \alpha_1, \alpha_2, ..., \alpha_I) = \sum_{ij} (y_{ij} - (\mu + \alpha_i))^2$$

Finding $(\hat{\mu}, \hat{\alpha_1}, \hat{\alpha_2}, ..., \hat{\alpha_I})$:

With the constraint $\alpha_1 = 0$, at the optimum:

$\frac{\partial S}{\partial \mu}(\hat{\mu}, \hat{\alpha_1} = 0, \hat{\alpha_2}, ..., \hat{\alpha_I}) = 0$ and

$\frac{\partial S}{\partial \alpha_i}(\hat{\mu}, \hat{\alpha_1} = 0, \hat{\alpha_2}, ..., \hat{\alpha_I}) = 0$ for $i = 2, ..., I$

$(I + 1)$ linear equations, $(I + 1)$ parameters

The explicit form of $(\hat{\mu}, \hat{\alpha_1}, \hat{\alpha_2}, ..., \hat{\alpha_I})$ depends on the identifying constraint chosen.

Finding $(\hat{\mu}, \hat{\alpha_1}, \hat{\alpha_2}, ..., \hat{\alpha_I})$ under an arbitrary constraint $\phi(\mu, \alpha_1, \alpha_2, ..., \alpha_I) = 0$:

With the constraint $\phi(\mu, \alpha_1, \alpha_2, ..., \alpha_I) = 0$, at the optimum:

- $\frac{\partial}{\partial \mu}[S(\hat{\mu}, \hat{\alpha_1}, \hat{\alpha_2}, ..., \hat{\alpha_I}) + \hat{\lambda}\phi(\hat{\mu}, \hat{\alpha_1}, \hat{\alpha_2}, ..., \hat{\alpha_I})] = 0$ and
- $\frac{\partial}{\partial \alpha_i}[S(\hat{\mu}, \hat{\alpha_1}, \hat{\alpha_2}, ..., \hat{\alpha_I}) + \hat{\lambda}\phi(\hat{\mu}, \hat{\alpha_1}, \hat{\alpha_2}, ..., \hat{\alpha_I})] = 0$ for $i = 2, ..., I$
- $\phi(\hat{\mu}, \hat{\alpha_1}, \hat{\alpha_2}, ..., \hat{\alpha_I}) = 0$

Variable λ above knows as Lagrange multiplier.

Model fitting in R

- For a `data.frame` with
  - response variable `y` of class `numeric`
  - explanatory variable `x` of class `factor`
- `lm(y ~ x)` will fit a one-way ANOVA model

A common mistake: explanatory variable `x` coded as an `integer` with class `numeric`. R will perform a simple linear regression and return all (meaningless) outputs without any warning.

Model fit in R on doughnuts data

```
> summary(lm.res)
Call:
lm(formula = Quantity ~ Fat.type, data = dough)
Residuals:
   Min     1Q Median     3Q    Max
-16.00  -7.00   0.00   5.25  23.00
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   172.000      4.101  41.943   <2e-16 ***
Fat.typeFat2   13.000      5.799   2.242   0.0365 *
Fat.typeFat3    4.000      5.799   0.690   0.4983
Fat.typeFat4  -10.000      5.799  -1.724   0.1001
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 10.04 on 20 degrees of freedom
Multiple R-squared:  0.4478, Adjusted R-squared:  0.365
F-statistic: 5.406 on 3 and 20 DF,  p-value: 0.006876
```

Geometric interpretation (cont')

With matrix notation introduced earlier $Y = X\theta + E$ and under the least square estimation principle we have:

ANOVA model fit as an orthogonal projection:

$\hat{Y} = X\hat{\theta}$ is an element of $Span(X)$ that minimizes $\|Y - \hat{Y}\|^2$
$\hat{Y}$ is the orthogonal projection of Y on $Span(X)$.

Towards models with two explanatory variables: the coke data

Response variable: processing time
Explanatory variables: temperature of oven, size of oven.

```
> coking
   width temp time
1      4 1600  3.5
2      4 1600  3.0
3      4 1600  2.7
4      4 1900  2.2
5      4 1900  2.3
6      4 1900  2.4
7      8 1600  7.1
8      8 1600  6.9
9      8 1600  7.5
10     8 1900  5.2
11     8 1900  4.6
12     8 1900  6.8
13    12 1600 10.8
14    12 1600 10.6
15    12 1600 11.0
16    12 1900  7.6
17    12 1900  7.1
18    12 1900  7.3
```

Remark on the experimental design of the Coke data.

Notation: n_{ij} nb. of observations having received level i of factor 1 and level j of factor 2.

- All combination of factors are observed $n_{ij} > 0 \ \ \forall i, j$
  full factorial design
- Each combination of factors is observed several times $n_{ij} > 1 \ \ \forall i, j$
  experimental design with replicates
- Same number of replications in each combination $n_{ij} = n_{i'j'}$
  balanced design

This is a best case scenario. A lot of statistical theory can be required to handle neatly other types of experimental designs.

Two-way ANOVA without interaction

Notation: y_{ijk} k-th observation having received level i of factor 1 and level j of factor 2.

Definition of the two-way ANOVA model without interaction:

- $y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$
- μ unobserved mean value
- α_i unobserved deterministic effect of level i of factor 1
- β_j unobserved deterministic effect of level j of factor 2
- ε_{ijk} unobserved random residual assumed to be i.i.d $\mathcal{N}(0, \sigma^2)$

Model above subject to identifiability issue as before.

Examples of identifiability constraints under model $y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$:

- $\mu = 0$ and $\alpha_1 = 0$
  or
- $\alpha_1 = 0$ and $\beta_1 = 0$
  or
- $\sum_i \alpha_i = 0$ and $\sum_j \beta_j = 0$

In R the default set of constraints is $\alpha_1 = 0$ and $\beta_1 = 0$

Model fitting in R

For `x1` and `x2` two factors,

```
lm(y ~ x1 + x2)
```

Output of lm on coking data

```
summary(lm.no.int)
Call:
lm(formula = time ~ temp + width, data = coking)
Residuals:
    Min      1Q  Median      3Q     Max
-0.9889 -0.6181 -0.1667  0.5847  1.4278

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.6611     0.3785   9.671 1.41e-07 ***
temp1900     -1.9556     0.3785  -5.166 0.000143 ***
width8        3.6667     0.4636   7.909 1.56e-06 ***
width12       6.3833     0.4636  13.768 1.57e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.803 on 14 degrees of freedom
Multiple R-squared:  0.9396, Adjusted R-squared:  0.9266
F-statistic: 72.55 on 3 and 14 DF,  p-value: 9.004e-09
```

Output of lm on coking data, cont'

```
> anova(lm.no.int)
Analysis of Variance Table

Response: time
          Df  Sum Sq Mean Sq F value    Pr(>F)
temp       1  17.209  17.209  26.687 0.0001432 ***
width      2 123.143  61.572  95.483 6.936e-09 ***
Residuals 14   9.028   0.645
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Two-way ANOVA with interactions

Back to the Coke data:
Is the effect of increased temperature the same on all ovens?
The previous two-way ANOVA model says:

$$\text{time}_{ij} = \begin{cases} \text{mean} + \text{oven\_effect}_i & \text{if temp=1600} \\ \text{or} & \\ \text{mean} + \text{oven\_effect}_i + \text{temp\_effect} & \text{if temp=1900} \end{cases}$$

The decrease of time induced by a change from 1600 to 1900 degrees is not of the same magnitude on all ovens.

Two-way ANOVA with interaction:

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

The extra set of parameters (γ_{ij}) is known as the interaction between factor 1 and 2.

Identifiability constraints:

$\gamma_{1j} = 0 \ \forall j$ and $0 \ \gamma_{i1} = 0 \ \forall i$ (default in R)

Or

$\sum_i \gamma_{ij} = 0 \ \forall j$ and $0 \ \sum_j \gamma_{ij} = 0 \ \forall i$

Model fitting in R:

For `x1` and `x2` two factors,

`lm(y ~ x1 + x2 + x1*x2)` or `lm(y ~ . + x1*x2)` or `lm(y ~ x1*x2)`

NB:

- The dot . alone does not include interactions.
- In a multiple regression on quantitative variables, terms such as `x1*x2` are syntaxically correct in R but have a different interpretation. There are used only for non-linear regressions.

Output of lm fit with interaction

```
summary(lm.int)
Call:
lm(formula = time ~ temp + width + temp * width, data = coking)
Residuals:
    Min      1Q  Median      3Q     Max
-0.9333 -0.2250 -0.0500  0.1750  1.2667
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.0667     0.3040  10.088 3.26e-07 ***
temp1900          -0.7667     0.4299  -1.783 0.099819 .
width8             4.1000     0.4299   9.537 5.96e-07 ***
width12            7.7333     0.4299  17.989 4.79e-10 ***
temp1900:width8   -0.8667     0.6080  -1.426 0.179501
temp1900:width12  -2.7000     0.6080  -4.441 0.000805 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.5265 on 12 degrees of freedom
Multiple R-squared:  0.9777, Adjusted R-squared:  0.9685
F-statistic: 105.4 on 5 and 12 DF,  p-value: 1.738e-09
```

Output of lm fit with interaction, cont'

```
> anova(lm.int)
Analysis of Variance Table

Response: time
            Df  Sum Sq Mean Sq F value    Pr(>F)
temp         1  17.209  17.209  62.076 4.394e-06 ***
width        2 123.143  61.572 222.102 3.312e-10 ***
temp:width   2   5.701   2.851  10.283  0.002504 **
Residuals   12   3.327   0.277
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Testing in ANOVA models

Design matrix in multiple regression vs. design matrix in ANOVA

The model $Y = X\theta + E$ is formally the same for a multiple regression and an ANOVA.

- In a multiple regression the variables in X result from the experiment which are often only partially controlled. The columns of X are almost always non-orthogonal (cf. caterpillar dataset)
- In an ANOVA, the values in X result from a choice of the scientist (nb. of replicates for each condition).
  - The columns of X can be linearly independant (easy case)
  - or linearly dependent (unbalanced cases) which complicates the analyses and interpretations

Testing global significance in ANOVA models

Does the model explain anything at all?

$H_0 : \alpha_i = \beta_j = \gamma_{ij} = 0 \ \ \forall i, j$

The test of H_0 is a Fisher test. Value returned in the last line of `summary()`.

Test of a specific level

Does level i differ from others?

This is a Student test returned by `summary(lm())`.

Testing effects in ANOVA models

Does factor 1 explain anything at all?

H_0 can be tricky to write explicity

The test of H_0 is a Fisher test. Value not returned in `summary()`.

Can be obtained by `anova(lm())`.

Recommended reading

Bingham and Fry, Regression, Chapter 2 (ANOVA) [pdf here]
Pages 42-56 are the most relevant.