Analysis of covariance

Statistical modelling: theory and practice

Gilles Guillot

gigu@dtu.dk

October 2, 2013

The travel "stylized data" (from Sheather)

"A small travel agency has retained your services to help them better understand two important customer segments. The first segment, which we will denote by A, consists of those customers who have purchased an adventure tour in the last twelve months. The second segment, which we will denote by C, consists of those customers who have purchased a cultural tour in the last twelve months. Data are available on 925 customers (i.e. on 466 customers from segment A and 459 customers from segment C). Note that the two segments are completely separate in the sense that there are no customers who are in both segments. Interest centres on identifying any differences between the two segments in terms of the amount of money spent in the last twelve months . In addition, data are also available on the age of each customer, since age is thought to have an effect on the amount spent. "

EDA

```
travel=read.table(
"http://www.stat.tamu.edu/~sheather/book/docs/datasets/travel.txt",
                      header=TRUE)

plot(travel$Age,travel$Amount,
col=travel$C+1,xlab="Age",ylab="Amount")

legend(col=c(1,2),pch=1,legend=c("Adventure","Culture"),
       x="top")
```

A model

We need a linear regression with segment-specific coefficients or one-way ANOVA with continuous effect.
Notation: y_{ij} spending of j-th customer in segment i ($i = 1, 2$)
x_{ij} age of j-th customer in segment i with $x_{ij} = 0$

ANCOVA model without interaction:

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

- μ
- α_i segment-specific departure from intercept
- β effect of age
- $\varepsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

Identifying constraint: $\alpha_1 = 0$.
For a I-level factor, we have $1 + (I - 1) + 1 + 1 = I + 2$ unknown independent parameters.

Parameter inference

Business as usual: LSE.
As seen with other models, the interpretation and expression of entries in $\boldsymbol{\theta}$ depend on identifying constraints.

Parameter inference

Inference with R. Example with travel spending data:

```
## ANCOVA without interaction:
lm.noint <- lm(Amount ~ 1 + Segment + Age,data=travel)
## same as call above: lm(Amount ~ .,data=travel)
summary(lm.noint)

plot(travel$Age,travel$Amount,
     col=travel$C+1,xlab="Age",ylab="Amount")
legend(col=c(1,2),pch=1,legend=c("Adventure","Culture"),
       x="top")
## fitted lines
abline(sum(lm.noint$coefficients[1:2]),
       lm.noint$coefficients[3],col=2)
abline(lm.noint$coefficients[1],
       lm.noint$coefficients[3],col=1)
```

In the above, Segment has to be a factor. Intrepretation of results....

ANCOVA model with interaction:

$$y_{ij} = \mu + \alpha_i + \beta x_{ij} + \gamma_i x_{ij} + \varepsilon_{ij}$$

- μ
- α_i segment-specific departure from intercept
- β common effect of age
- γ_i segment-specific effect of age
- $\varepsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

Identifying constraint: $\alpha_1 = 0$ and $\gamma_1 = 0$.
For a I-level factor, we have $1 + (I-1) + 1 + (I-1) + 1 = 2I + 1$ unknown independent parameters

Parameter inference with interaction

Example with travel spending data:

```
lm.int <- lm(Amount ~ 1 + Segment + Age +  Segment*Age,data=travel)
## same as call above
## lm(Amount ~ Segment*Age,data=travel)
summary(lm.int)

anova(lm.int)

## fitted lines
abline(sum(lm.int$coefficients[1:2]),
       sum(lm.int$coefficients[3:4]),col=2,lty=2)
abline(lm.int$coefficients[1],
       lm.int$coefficients[3],col=1,lty=2)
```

ANCOVA model in matrix form

ANCOVA model without interaction:

With $\boldsymbol{\theta} = (\mu, \alpha_1, ..., \alpha_I, \beta)^t$, the model $y_{ij} = \mu + \alpha_i + \beta x_{ij} + \varepsilon_{ij}$ can be written

$$Y = X\boldsymbol{\theta} + \boldsymbol{E}$$

where X is an $n \times (I + 2)$ matrix a bit cumbersome to write in LaTeX.

Suggested reading

Sheather, section 5.3, pdf [here]