Lecture 5 : Methods for Regression

Anova 2 factors

K. Meziani

Dauphine | PSL
UNIVERSITÉ PARIS

Example under R

Atherosclerosis is the leading cause of death for men after age 35 and for women after age 45 in most developed countries. It is a thickening and a loss of elasticity of the internal walls of the arteries, one of the consequences of which is myocardial infarct. The arterial wall consists of three layers respectively from the arterial lumen: the intima, the media and the adventitia. The thickness of the intima-media is a recognized marker of atherosclerosis. It was measured ultrasonically on a sample of 110 subjects in 1999 at the Bordeaux University Hospital. Information on the main risk factors was also collected, including on smoking and alcohol consumption among patients:

- `Smoking status` is measured in 3 modalities: `0="do not smoke"`, `"1=quit smoking"`, `2="smoke"`.
- `Consumption of alcohol` is measured in 3 modalities: `0="do not drink"`, `1="drink occasionally"`, `2="drink regularly"`.

We want to conduct an analysis of the influence of these factors on the thickness of the intima-media.

Packages

```
library(carData)
library(car)
library(knitr)
library (survival)
library(MASS)
library(TH.data)
library(mvtnorm)
library(multcomp)
```

Section 1

I. Upload the dataset and Descriptive analysis

Upload the dataset

Consider in this section an Anova two factors model. Consider the example of the influence of the `Consumption of alcohol` and the `Smoking status` on the thickness of the intima-media.

We recall that the Consumption of alcohol :`alcool` has $J = 3$ modalities

"0" ="do not drink"

"1" ="drink occasionally"

"2" ="drink regularly"

We recall that the smoking status :`tabac` has $K = 3$ modalities

"0" ="do not smoke"

"1" ="quite smoke"

"2" ="smoke"

Read the data and select a subsample

First load and read the dataset.

```
Marqueur = read.table("Intima_Media.txt", header=T,
                      sep=" ", dec=",")
names(Marqueur)
```

```
## [1] "SEXE"   "AGE"    "taille" "poids"  "tabac"  "paqan"  "SPORT"  "mesure"
## [9] "alcool"
```

```
marqueur=Marqueur[,c(5,8,9)]
names(marqueur)
```

```
## [1] "tabac"  "mesure" "alcool"
```

Rename some modalties

We did rename

```
marqueur$alcool=replace(marqueur$alcool,marqueur$alcool==0,"NotDrink")
marqueur$alcool=replace(marqueur$alcool,marqueur$alcool==1,"DrinkOcc")
marqueur$alcool=replace(marqueur$alcool,marqueur$alcool==2,"DrinkReg")
marqueur$alcool=as.factor(marqueur$alcool)
```

For sake of simplicity in the interpretation, we change also the name of the modalities of the variable `tabac` and declare it as a factor.

```
marqueur$tabac=replace(marqueur$tabac,marqueur$tabac==0,"NotSmoke")
marqueur$tabac=replace(marqueur$tabac,marqueur$tabac==1,"QuitSmoke")
marqueur$tabac=replace(marqueur$tabac,marqueur$tabac==2,"Smoke")
marqueur$tabac=as.factor(marqueur$tabac)
```

Check the nature of the features

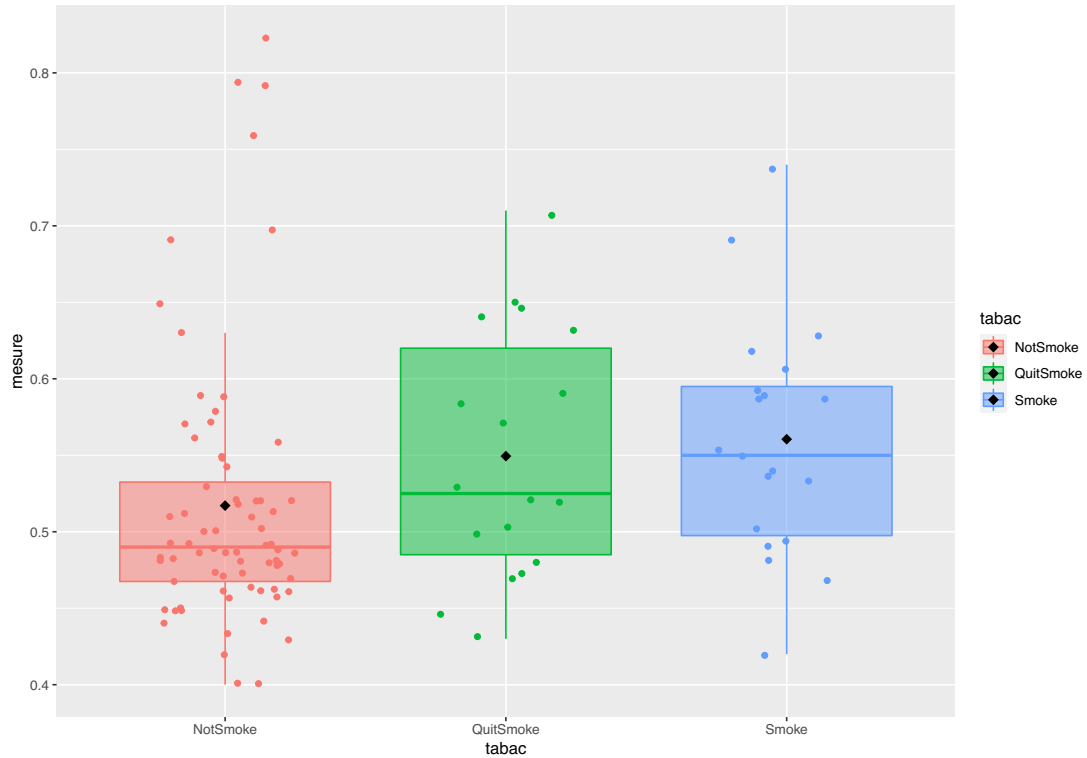
The variables have been correctly defined.

```
str(marqueur)
```

```
## 'data.frame':    110 obs. of  3 variables:
##  $ tabac : Factor w/ 3 levels "NotSmoke","QuitSmoke",..: 2 3 2 2 1 3 1 2 1 1
##  $ mesure: num  0.52 0.42 0.65 0.48 0.45 0.49 0.42 0.45 0.65 0.52 ...
##  $ alcool: Factor w/ 3 levels "DrinkOcc","DrinkReg",..: 1 1 3 1 1 1 1 1 2 1 .
```

Boxplots of the mesure per modality

```
library(cowplot); library(ggplot2)
ggplot(marqueur, aes(y=mesure, x=tabac,colour=tabac ,fill=tabac))+
geom_boxplot(alpha=0.5, outlier.alpha=0)+geom_jitter(width=0.25)+
stat_summary(fun=mean, colour="black", geom="point",shape=18, size=3)
```



Some resumes of the dataset

Display the number of modalities K of the factor `tabac` and J of the factor `alcool`

```
J =length(levels(marqueur$alcool))
K =length(levels(marqueur$tabac))
print(paste("K=",K," and J=",J))
```

```
## [1] "K= 3  and J= 3"
```

Display the n_{jk} , $j = 1, \cdots, J$ and $k = 1 \cdot, K$ the number of observations of the modality (j, k). Note that, in this dataset, the plan is unbalanced.

```
n_jk =table(marqueur$tabac,marqueur$alcool)
knitr::kable(n_jk)
```

	DrinkOcc	DrinkReg	NotDrink
NotSmoke	45	9	18
QuitSmoke	14	1	3
Smoke	12	6	2

Section 2

II. Empirical means (EM)

Display $\overline{Y}_{.jk}$ EM per cell (j, k)

$$\overline{Y}_{.jk} = \frac{1}{n_{jk}} \sum_{i=1}^{n_{jk}} Y_{ijk}$$

```
EM=tapply(marqueur$mesure,list(Tabac=marqueur$tabac,
          Alcool=marqueur$alcool),mean,na.rm=TRUE)
knitr::kable(EM)
```

	DrinkOcc	DrinkReg	NotDrink
NotSmoke	0.5208889	0.5600000	0.4861111
QuitSmoke	0.5450000	0.6400000	0.5400000
Smoke	0.5491667	0.5766667	0.5800000

Display other EM

Display EM of all the EM per cell (j, k)

$$\overline{\overline{Y}}_{...} = \frac{1}{JK} \sum_{j=1}^{J} \sum_{k=1}^{K} \overline{Y}_{\cdot jk} = 0.5553148$$

```
mean(EM)
```

```
## [1] 0.5553148
```

Display EM of all obsefvations Y_{ijk}

$$\overline{Y}_{...} = \frac{1}{n} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{i=1}^{n_{jk}} Y_{ijk} = 0.5302727$$

```
mean(marqueur$mesure)
```

```
## [1] 0.5302727
```

Display $\overline{\overline{Y}}_{.j.}$

The EM of the EM $\overline{Y}_{.jk}$ having modality j is denoted

$$\overline{\overline{Y}}_{.j.} = \frac{1}{K}\sum_{k=1}^{K} \overline{Y}_{.jk}$$

```
EM_EM_j=colMeans(EM)
EM_EM_j=as.data.frame(EM_EM_j)
names(EM_EM_j)="EM of the EM per cell (j,k) having modality j"
kable(EM_EM_j)
```

	EM of the EM per cell (j,k) having modality j
DrinkOcc	0.5383519
DrinkReg	0.5922222
NotDrink	0.5353704

Display $\overline{\overline{Y}}_{..k}$

The EM of the EM $\overline{Y}_{.jk}$ having modality k is denoted

$$\overline{\overline{Y}}_{..k} = \frac{1}{J}\sum_{j=1}^{J} \overline{Y}_{.jk}$$

```
EM_EM_k=rowMeans(EM)
EM_EM_k=as.data.frame(EM_EM_k)
names(EM_EM_k)=" EM of the EM per cell (j,k) having modality k"
kable(EM_EM_k)
```

	EM of the EM per cell (j,k) having modality k
NotSmoke	0.5223333
QuitSmoke	0.5750000
Smoke	0.5686111

Display $\overline{Y}_{.j.}$

The EM of the observations $Y_{.ijk}$ having modality j is denoted

$$\overline{Y}_{.j.} = \frac{1}{n_{j.}} \sum_{k=1}^{K} \sum_{i=1}^{n_{jk}} Y_{.ijk}$$

```
EM_j=tapply(marqueur$mesure,list(Alcool=marqueur$alcool),mean,na.rm=TRUE)
EM_j=as.data.frame(EM_j)
names(EM_j)="EM of the observations having modality j"
kable(EM_j)
```

	EM of the observations having modality j
DrinkOcc	0.5304225
DrinkReg	0.5712500
NotDrink	0.5013043

Display $\overline{Y}_{..k}$

The EM of the observations $Y_{.ijk}$ having modality j is denoted

$$\overline{Y}_{..k} = \frac{1}{n_{.k}} \sum_{j=1}^{J} \sum_{i=1}^{n_{jk}} Y_{.ijk}$$

```
EM_k=tapply(marqueur$mesure,list(Tabac=marqueur$tabac),mean,na.rm=TRUE)
EM_k=as.data.frame(EM_k)
names(EM_k)="EM of the observations having modality k"
kable(EM_k)
```

	EM of the observations having modality k
NotSmoke	0.5170833
QuitSmoke	0.5494444
Smoke	0.5605000

Section 3

III. Model anova two factors

Model anova two factors

Let define the following anova 2 factors model

$$Y = \mu \mathbb{1}_n + A\alpha + B\beta + C\gamma + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(O_n, \sigma^2 \mathbb{I}_n),$$

where α is the main effect of the factor `alcool`, β the main effect of the factor `tabac` and γ represents the interaction between the 2 factors. We choose the sums constraints

$$\sum_{j=1}^{J} \alpha_j = \sum_{k=1}^{K} \beta_k = \sum_{j=1}^{J} \gamma j k' = \sum_{k=1}^{K} \gamma j' k = 0, \quad \forall j', k'.$$

```
MOD1=lm(mesure~alcool*tabac,
        contrasts=list(tabac="contr.sum",alcool="contr.sum"),
        data=marqueur)
```

Recall that R renames the modality.

Output

```
summary(MOD1)
```

```
##
## Call:
## lm(formula = mesure ~ alcool * tabac, data = marqueur, contrasts = list(tabac = "contr.sum",
##     alcool = "contr.sum"))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.12917 -0.05089 -0.02003  0.03389  0.29911
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.5553148  0.0145005  38.296   <2e-16 ***
## alcool1         -0.0169630  0.0160592  -1.056   0.2934
## alcool2          0.0369074  0.0235409   1.568   0.1201
## tabac1          -0.0329815  0.0161587  -2.041   0.0438 *
## tabac2           0.0196852  0.0242560   0.812   0.4190
## alcool1:tabac1   0.0155185  0.0180743   0.859   0.3926
## alcool2:tabac1   0.0007593  0.0263577   0.029   0.9771
## alcool1:tabac2  -0.0130370  0.0263375  -0.495   0.6217
## alcool2:tabac2   0.0280926  0.0417098   0.674   0.5022
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08525 on 101 degrees of freedom
## Multiple R-squared:  0.1033, Adjusted R-squared:  0.03224
## F-statistic: 1.454 on 8 and 101 DF,  p-value: 0.1837
```

Resume of the output

Name	**R** outputs	OLSE1	In terms of empirical means
`Intercept`	0.5553148	$\widehat{\mu}$	$= \overline{\overline{\overline{Y}}}_{...}$
`alcool1`	-0.0199444	$\widehat{\alpha}_1$	$= \overline{\overline{Y}}_{.1.} - \overline{\overline{\overline{Y}}}_{...}$
`alcool2`	0.0369074	$\widehat{\alpha}_2$	$= \overline{\overline{Y}}_{.2.} - \overline{\overline{\overline{Y}}}_{...}$
`tabac1`	-0.0329815	$\widehat{\beta}_1$	$= \overline{\overline{Y}}_{..1} - \overline{\overline{\overline{Y}}}_{...}$
`tabac2`	0.0196852	$\widehat{\beta}_2$	$= \overline{\overline{Y}}_{..2} - \overline{\overline{\overline{Y}}}_{...}$
`alcool1:tabac1`	-0.0162778	$\widehat{\gamma}_{11}$	$= \overline{Y}_{.11} + \overline{\overline{\overline{Y}}}_{...} - \overline{\overline{Y}}_{.1.} - \overline{\overline{Y}}_{..1}$
`alcool2:tabac1`	0.0007593	$\widehat{\gamma}_{21}$	$= \overline{Y}_{.21} + \overline{\overline{\overline{Y}}}_{...} - \overline{\overline{Y}}_{.2.} - \overline{\overline{Y}}_{..1}$
`alcool1:tabac2`	0.0150556	$\widehat{\gamma}_{12}$	$= \overline{Y}_{.12} + \overline{\overline{\overline{Y}}}_{...} - \overline{\overline{Y}}_{.1.} - \overline{\overline{Y}}_{..2}$
`alcool2:tabac2`	0.0280926	$\widehat{\gamma}_{22}$	$= \overline{Y}_{.22} + \overline{\overline{\overline{Y}}}_{...} - \overline{\overline{Y}}_{.2.} - \overline{\overline{Y}}_{..2}$

Recall

With our notations and if one calculate the OLSE in the setting of unbalanced plan with the constraints sums: some coefficients have to be calculated by hand

$\widehat{\alpha}_3 = -(\widehat{\alpha}_1+\widehat{\alpha}_2)$, $\widehat{\beta}_3 = -(\widehat{\beta}_1+\widehat{\beta}_2)$, $\widehat{\gamma}_{13} = (\widehat{\gamma}_{11}+\widehat{\gamma}_{12})$ and $\widehat{\gamma}_{23} = -(\widehat{\gamma}_{21}+\widehat{\gamma}_{22})$

Section 4

IV. Model selection

Impact of the order of the features on commands anova(.) and Anova(.)

To highlight the limit of the anova and Anova commands, let's introduce our model in two different ways: change the order of the factor in the lm command

```
MOD=lm(mesure~tabac*alcool,data=marqueur)
MODbis=  lm(mesure~alcool*tabac,data=marqueur)
```

Recall that the anova command compares nested models by introducing one by one the factors.

Impact of the order of the features with command anova(.)

When the factor `tabac` is introduce first, the command concludes that no factor has an impact on the variable `mesure`.

```
anova(MOD)
```

```
## Analysis of Variance Table
##
## Response: mesure
##               Df  Sum Sq   Mean Sq F value  Pr(>F)
## tabac          2 0.03741 0.0187074  2.5743 0.08120 .
## alcool         2 0.03531 0.0176530  2.4292 0.09324 .
## tabac:alcool   4 0.01180 0.0029509  0.4061 0.80389
## Residuals    101 0.73397 0.0072670
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Impact of the order of the features with command anova(.)

On the other hand, when the `alcool` factor is entered first, the command concludes that only the `alcool` factor has an impact on the variable `mesure`.

```
anova(MODbis)
```

```
## Analysis of Variance Table
##
## Response: mesure
##               Df  Sum Sq   Mean Sq F value  Pr(>F)
## alcool         2 0.04617 0.0230843  3.1766 0.04593 *
## tabac          2 0.02655 0.0132762  1.8269 0.16619
## alcool:tabac   4 0.01180 0.0029509  0.4061 0.80389
## Residuals    101 0.73397 0.0072670
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Impact of the order of the features with command `Anova(.)`

Recall that the `Anova` command compares nested models by removing one of the two factors in the models without interaction (the 2 first tests).

```
Anova(MOD)
```

```
## Anova Table (Type II tests)
##
## Response: mesure
##               Sum Sq  Df F value  Pr(>F)
## tabac         0.02655   2  1.8269 0.16619
## alcool        0.03531   2  2.4292 0.09324 .
## tabac:alcool 0.01180   4  0.4061 0.80389
## Residuals     0.73397 101
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Impact of the order of the features with command `Anova(.)`

Whatever the order, in our case the command fails to highlight the impact of the factor `alcool`.

```
Anova(MODbis)
```

```
## Anova Table (Type II tests)
##
## Response: mesure
##               Sum Sq  Df F value  Pr(>F)
## alcool       0.03531   2  2.4292 0.09324 .
## tabac        0.02655   2  1.8269 0.16619
## alcool:tabac 0.01180   4  0.4061 0.80389
## Residuals    0.73397 101
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model selection : Step-by-step method

```
library(MASS)
MOD0=lm(mesure~1,data=marqueur)
MOD=lm(mesure~tabac*alcool,data=marqueur)
step(MOD,direction='backward',trace=F)
```

```
##
## Call:
## lm(formula = mesure ~ alcool, data = marqueur)
##
## Coefficients:
##    (Intercept)  alcoolDrinkReg  alcoolNotDrink
##        0.53042         0.04083        -0.02912
```

Here,the 3 methods give the same final model which is the model anova single factor study in the file (ANOVA1). We refer to it to finish the sudy (validation of the model).

```
#step(MOD0,mesure~tabac*alcool,direction='forward',trace=F)
#step(MOD0,mesure~tabac*alcool,direction='both',trace=F)
```