Lecture 6 : Methods for Regression

Anocva 1 factor

K. Meziani

Dauphine | PSL
UNIVERSITÉ PARIS

Example under R :The dataset Cepages

- **pH** : pH of the wine.
- **Origine**: factor which admits $I = 2$ modalities : **Bordeaux** and **Bourgogne**.
- **Couleur** : factor which admits $I = 2$ modalities : **Blanc** and **Rouge**.
- **Alcool** : It is the alcohol content of the wine.
- **Malique** : Malic acid that reflects greenness / biting wine (green apple).
- **Tartrique** : Tartaric acid that reflects hardness / structure of the wine (the acid most present in the grapes).
- **Citrique** : Citric acid that reflects freshness of the wine (lemony taste).
- **Acetique** : Acetic acid is a natural organic acid, the main constituent of the volatile acidity of a wine.
- **Lactique** : Lactic acid is an organic acid that plays a role in various biochemical processes.
- **AcTot** : Total acidity.

We will only consider here an Ancova single factor model. We want to explain $Y =$**pH** by the factor **Couleur** and the covariate **AcTot**.

Packages

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

Section 1

I. Upload the dataset

Cepages dataset

```
Cepages =read.csv2("CepagesB.csv")
Cepages= Cepages[,-(3)] # do not consider the column "Libelle"
names(Cepages)
```

```
##  [1] "Origine"   "Couleur"   "Alcool"    "pH"        "AcTot"     "Tartrique"
##  [7] "Malique"   "Citrique"  "Acetique"  "Lactique"
```

Here, our aim is to explain Y =**pH** by the factor Couleur and the covariate/regressor **AcTot**.

Check the nature of the dataset

```
str(Cepages)
```

```
## 'data.frame':    36 obs. of  10 variables:
##  $ Origine  : chr  "Bordeaux" "Bordeaux" "Bordeaux" "Bordeaux" ...
##  $ Couleur  : chr  "Blanc" "Blanc" "Blanc" "Blanc" ...
##  $ Alcool   : num  12 11.5 14.6 10.5 14 13.2 11.2 15.4 13.4 11.4 ...
##  $ pH       : num  2.84 3.1 2.96 3.1 3.29 2.94 2.91 3.43 3.35 2.9 ...
##  $ AcTot    : int  89 97 99 72 76 83 95 86 76 103 ...
##  $ Tartrique: num  21.1 26.4 20.7 29.7 22.3 24.6 39.4 14.1 18.9 50 ...
##  $ Malique  : num  21 34.2 21.8 4.2 9.3 9.4 14.5 28.8 23 18 ...
##  $ Citrique : num  4.3 3.9 8.1 3.6 4.7 4.1 4.2 8.5 6.4 2.8 ...
##  $ Acetique : num  16.9 9.9 19.7 11.9 20.1 19.7 19.4 15 14.4 14.4 ...
##  $ Lactique : num  9.3 16 11.2 14.4 21.6 16.8 10.5 12.6 10.5 8.5 ...
```

Check the nature of the dataset

```
Cepages$Origine=as.factor(Cepages$Origine)
Cepages$Couleur=as.factor(Cepages$Couleur)
```

Section 2

2. Descriptive analysis

Table of counts

We have $n = 36$ observations and the plan is balanced.

```
knitr::kable(table(Cepages$Couleur),col.names =c("Couleur","Counts"))
```

Couleur	Counts
Blanc	18
Rouge	18

Display the table of empirical means by cell.

```
EM=tapply(Cepages$pH,list(Coul=Cepages$Couleur),mean);
knitr::kable(EM,col.names =c("Empirical means"))
```

	Empirical means
Blanc	3.040556
Rouge	3.414444

As the plan is balanced, the empirical mean of the **pH** and the empirical mean of all the empirical means are equal and are equal to

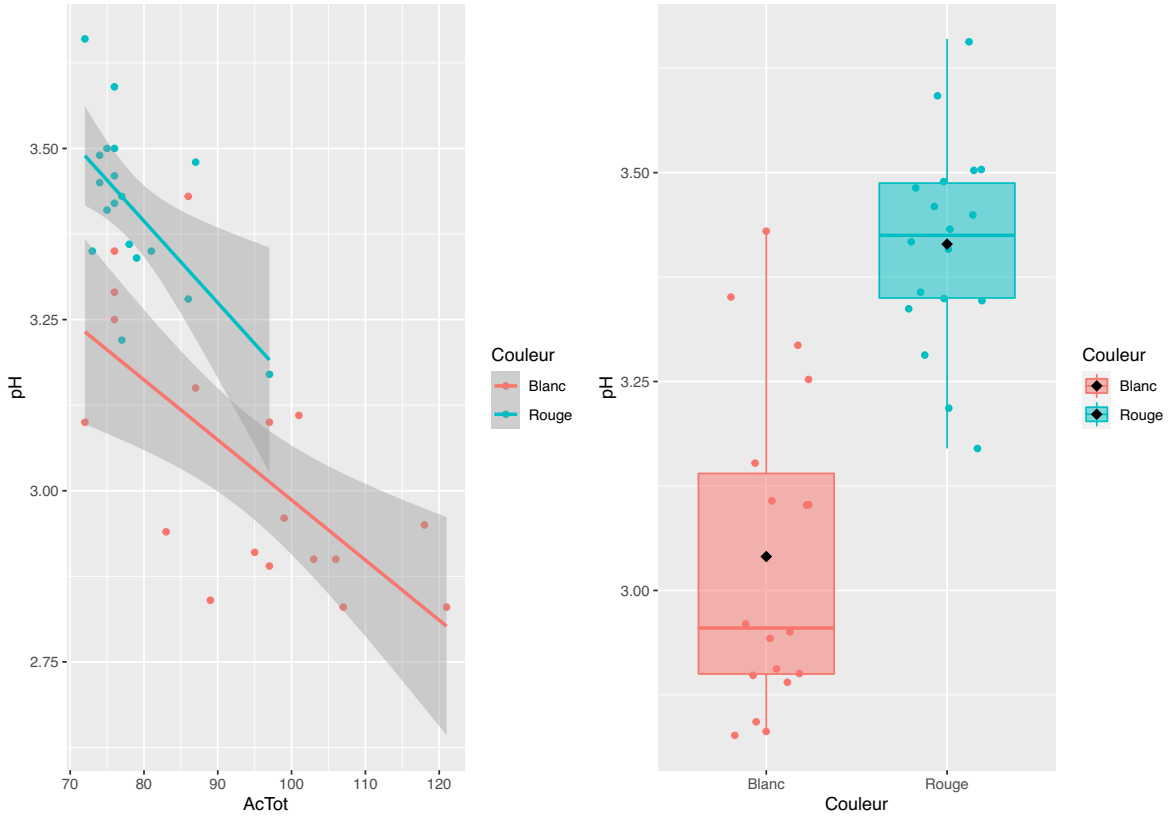
```
mean(Cepages$pH)
```

```
## [1] 3.2275
```

Plot dataset

```
library(cowplot)
library(ggplot2)

AcTot=Cepages[,"AcTot"]
pH=Cepages[,"pH"]
Couleur= as.factor(Cepages[,"Couleur"])


PlotCouleur1=ggplot(Cepages, aes(x = AcTot, y =pH,color=Couleur)) +
geom_point()+geom_smooth(method = "lm")

PlotCouleur2=ggplot(Cepages, aes(y=pH, x=Couleur,colour=Couleur ,fill=Couleur))
+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)

plot_grid(PlotCouleur1,PlotCouleur2,ncol=2,nrow=1)
```

Plot dataset



Comments

It seems that the **`Couleur`** factor has an impact on the variable **`pH`**. The regression lines are dfifferent with respect to the choosen modality.

Section 3

III. Ancova Single factor model

Ancova Single factor model

Let define the following Ancova 1 factor model

$$Y = \mu \mathbb{1}_n + A\alpha + bx + \text{[illegible]} + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(O_n, \sigma^2 \mathbb{I}_n),$$

Ancova 1 factor Model with default constraint

We use here the constraint by default under **R**.

$$\alpha_1 = c_1 = 0$$

```
modAncova=lm(pH~Couleur*AcTot)
```

Test the influence of the regressor with anova()

```
anova(modAncova)
```

```
## Analysis of Variance Table
##
## Response: pH
##               Df  Sum Sq Mean Sq F value    Pr(>F)
## Couleur        1 1.25814 1.25814  80.475 3.015e-10 ***
## AcTot          1 0.35643 0.35643  22.798 3.820e-05 ***
## Couleur:AcTot  1 0.00543 0.00543   0.347    0.5599
## Residuals     32 0.50029 0.01563
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test the influence of the regressor with Anova()

```
Anova(modAncova)
```

```
## Anova Table (Type II tests)
##
## Response: pH
##                Sum Sq Df F value    Pr(>F)
## Couleur       0.31151  1  19.926 9.368e-05 ***
## AcTot         0.35643  1  22.798 3.820e-05 ***
## Couleur:AcTot 0.00543  1   0.347    0.5599
## Residuals     0.50029 32
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comments

Whatever the tests (type I or type II), the interaction have no impact. Then, we select the following model without interaction

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

```
modAncovaWI=lm(pH~Couleur+AcTot)
```

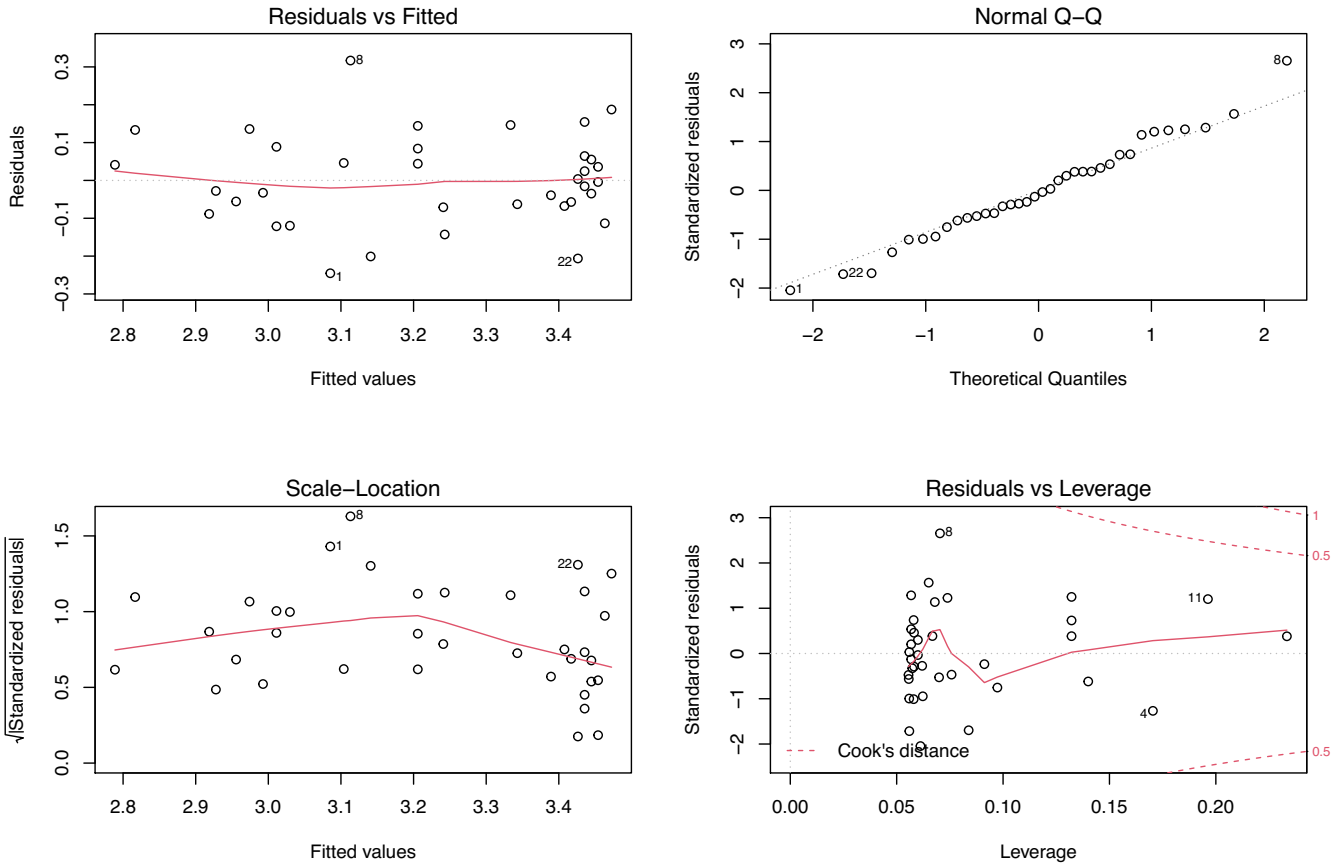
Summary

```
summary(modAncovaWI)
```

```
##
## Call:
## lm(formula = pH ~ Couleur + AcTot)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24535 -0.06855 -0.00982  0.06938  0.31685
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.910142   0.182656  21.407  < 2e-16 ***
## CouleurRouge  0.229730   0.050953   4.509 7.79e-05 ***
## AcTot        -0.009267   0.001922  -4.823 3.11e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1238 on 33 degrees of freedom
## Multiple R-squared:  0.7615, Adjusted R-squared:  0.747
## F-statistic: 52.68 on 2 and 33 DF,  p-value: 5.357e-11
```

The postulates are validated. No outliers to remove.

```
par(mfrow=c(2,2))
plot(modAncovaWI)
```



Section 4

IV. What about the full dataset ?

Model Declaration

```
MOD=lm(pH~.,data=Cepages)
# plot(MOD) :Model seems to be valid.
```

What about model selection ?

Forward method

```
MOD0=lm(pH~1,data=Cepages)
modForw=stepAIC(MOD0, pH~Origine+Couleur+Alcool+AcTot+Tartrique+Malique+
     Citrique+Acetique+Lactique,trace=F,
     direction=c('forward'))
modForw
```

```
##
## Call:
## lm(formula = pH ~ AcTot + Lactique + Couleur + Malique + Origine +
##     Acetique, data = Cepages)
##
## Coefficients:
##      (Intercept)             AcTot          Lactique      CouleurRouge
##         4.150399          -0.018998          0.010127          0.234626
##          Malique  OrigineBourgogne          Acetique
##         0.016363          0.089375          0.009695
```

Forward method

```
MOD0=lm(pH~1,data=Cepages)
modForw=stepAIC(MOD0, pH~Origine+Couleur+Alcool+AcTot+Tartrique+Malique+
     Citrique+Acetique+Lactique,trace=F,
     direction=c('forward'))
modForw
```

```
##
## Call:
## lm(formula = pH ~ AcTot + Lactique + Couleur + Malique + Origine +
##     Acetique, data = Cepages)
##
## Coefficients:
##      (Intercept)             AcTot          Lactique      CouleurRouge
##         4.150399          -0.018998          0.010127          0.234626
##          Malique  OrigineBourgogne          Acetique
##         0.016363          0.089375          0.009695
```

Selection of the following features : **AcTot**,**Lactique**, **Malique**, **Acetique**

Couleur and **Origine**.

Backward method

```
modBack=stepAIC(MOD,~.,trace=F,direction=c("backward"))
modBack
```

```
##
## Call:
## lm(formula = pH ~ Origine + Couleur + AcTot + Tartrique + Malique +
##     Citrique + Acetique + Lactique, data = Cepages)
##
## Coefficients:
##      (Intercept)  OrigineBourgogne      CouleurRouge             AcTot
##         3.982360          0.057932          0.278029         -0.021457
##        Tartrique           Malique          Citrique          Acetique
##         0.005425          0.019198          0.018160          0.013890
##         Lactique
##         0.012482
```

Backward method

```
modBack=stepAIC(MOD,~.,trace=F,direction=c("backward"))
modBack
```

```
##
## Call:
## lm(formula = pH ~ Origine + Couleur + AcTot + Tartrique + Malique +
##     Citrique + Acetique + Lactique, data = Cepages)
##
## Coefficients:
##      (Intercept)  OrigineBourgogne      CouleurRouge             AcTot
##         3.982360          0.057932          0.278029         -0.021457
##        Tartrique           Malique          Citrique          Acetique
##         0.005425          0.019198          0.018160          0.013890
##         Lactique
##         0.012482
```

Selection of the following features : **AcTot**,**Tartrique**,**Malique**, **Citrique**, **Acetique**, **Lactique**, **Origine** and **Couleur**.

Both method

```
modBoth=stepAIC(MOD0,pH~Origine+Couleur+Alcool+AcTot+Tartrique+Malique
   +Citrique+Acetique+Lactique,trace=F,direction=c("both"))
modBoth
```

```
##
## Call:
## lm(formula = pH ~ AcTot + Lactique + Couleur + Malique + Origine +
##     Acetique, data = Cepages)
##
## Coefficients:
##      (Intercept)             AcTot          Lactique      CouleurRouge
##         4.150399         -0.018998          0.010127          0.234626
##          Malique  OrigineBourgogne          Acetique
##         0.016363          0.089375          0.009695
```

Both method

```
modBoth=stepAIC(MOD0,pH~Origine+Couleur+Alcool+AcTot+Tartrique+Malique
   +Citrique+Acetique+Lactique,trace=F,direction=c("both"))
modBoth
```

```
##
## Call:
## lm(formula = pH ~ AcTot + Lactique + Couleur + Malique + Origine +
##     Acetique, data = Cepages)
##
## Coefficients:
##      (Intercept)             AcTot          Lactique      CouleurRouge
##         4.150399         -0.018998          0.010127          0.234626
##          Malique  OrigineBourgogne          Acetique
##         0.016363          0.089375          0.009695
```

Selection of the following features : **AcTot**, **Lactique**, **Malique**, **Acetique**, **Origine** and **Couleur**.

Selection with anova(.)

```
anova(MOD)
```

```
## Analysis of Variance Table
##
## Response: pH
##            Df  Sum Sq Mean Sq  F value    Pr(>F)
## Origine     1 0.04202 0.04202   6.4259 0.0176049 *
## Couleur     1 1.25814 1.25814 192.3766 1.587e-13 ***
## Alcool      1 0.08123 0.08123  12.4212 0.0015943 **
## AcTot       1 0.25300 0.25300  38.6845 1.400e-06 ***
## Tartrique   1 0.13579 0.13579  20.7637 0.0001083 ***
## Malique     1 0.03171 0.03171   4.8487 0.0367376 *
## Citrique    1 0.00030 0.00030   0.0466 0.8308253
## Acetique    1 0.01504 0.01504   2.3004 0.1414030
## Lactique    1 0.13299 0.13299  20.3352 0.0001227 ***
## Residuals 26 0.17004 0.00654
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Selection with anova(.)

```
anova(MOD)
```

```
## Analysis of Variance Table
##
## Response: pH
##            Df  Sum Sq Mean Sq  F value    Pr(>F)
## Origine     1 0.04202 0.04202   6.4259 0.0176049 *
## Couleur     1 1.25814 1.25814 192.3766 1.587e-13 ***
## Alcool      1 0.08123 0.08123  12.4212 0.0015943 **
## AcTot       1 0.25300 0.25300  38.6845 1.400e-06 ***
## Tartrique   1 0.13579 0.13579  20.7637 0.0001083 ***
## Malique     1 0.03171 0.03171   4.8487 0.0367376 *
## Citrique    1 0.00030 0.00030   0.0466 0.8308253
## Acetique    1 0.01504 0.01504   2.3004 0.1414030
## Lactique    1 0.13299 0.13299  20.3352 0.0001227 ***
## Residuals 26 0.17004 0.00654
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Selection of the following features : **Alcool**, **AcTot**, **Tartrique**, **Malique**, **Lactique**, **Couleur** and **Origine**.

```
modanov=lm(pH~Origine+Couleur+Alcool+AcTot+Tartrique+Malique+Lactique,
  data=Cepages)
```

Selection with Anova(.)

```
Anova(MOD)
```

```
## Anova Table (Type II tests)
##
## Response: pH
##             Sum Sq Df F value    Pr(>F)
## Origine   0.009593  1  1.4668 0.2367400
## Couleur   0.191182  1 29.2329 1.152e-05 ***
## Alcool    0.000530  1  0.0810 0.7782193
## AcTot     0.277870  1 42.4880 6.555e-07 ***
## Tartrique 0.009793  1  1.4975 0.2320373
## Malique   0.120321  1 18.3977 0.0002191 ***
## Citrique  0.009836  1  1.5039 0.2310593
## Acetique  0.041836  1  6.3970 0.0178351 *
## Lactique  0.132992  1 20.3352 0.0001227 ***
## Residuals 0.170039 26
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Selection with Anova(.)

```
Anova(MOD)
```

```
## Anova Table (Type II tests)
##
## Response: pH
##             Sum Sq Df F value    Pr(>F)
## Origine   0.009593  1  1.4668 0.2367400
## Couleur   0.191182  1 29.2329 1.152e-05 ***
## Alcool    0.000530  1  0.0810 0.7782193
## AcTot     0.277870  1 42.4880 6.555e-07 ***
## Tartrique 0.009793  1  1.4975 0.2320373
## Malique   0.120321  1 18.3977 0.0002191 ***
## Citrique  0.009836  1  1.5039 0.2310593
## Acetique  0.041836  1  6.3970 0.0178351 *
## Lactique  0.132992  1 20.3352 0.0001227 ***
## Residuals 0.170039 26
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Selection of the following features : **AcTot**,**Malique**,**Acetique**, **Lactique** and **Couleur**

```
modAnov=lm(pH~Couleur+AcTot+Malique+Acetique+Lactique,data=Cepages)
```

Which model? ⇒ Extract AIC

```
AIC=rbind(extractAIC(modForw),extractAIC(modBack),extractAIC(modBoth),
     extractAIC(modanov),extractAIC(modAnov),extractAIC(modAncovaWI))
Names=c('Forward','Backwar','Both','anova','Anova','AncovaWI')
AIC=cbind(Names,AIC)
AICmod=as.data.frame(AIC)
names(AICmod)=c('Method','size','AIC')
kable(AICmod)
```

Method	size	AIC
Forward	7	-175.060956402565
Backwar	9	-174.676893987726
Both	7	-175.060956402565
anova	8	-168.66622730311
Anova	6	-170.566353354271
AncovaWI	3	-147.551101576071

Model with Forward method (the same as Both) has the smallest AIC.

Validation of the model

```
par(mfrow=c(2,2)); plot(modForw)
```