

Première Année Master M.A.E.F. 2025 – 2026

Econométrie II

Contrôle continu n°2, avril 2026

Examen de 1h30. Tout document ou calculatrice est interdit.

Exercice théorique (Sur 16 points)

Soit $0 \leq p \leq n$ et on suppose que l'on a observé (Y_1, \dots, Y_n) , qui s'écrit

$$Y_i = \theta_0 + \theta_1 X_i^{(1)} + \dots + \theta_p X_i^{(p)} + \varepsilon_i \quad \text{pour } 1 \leq i \leq n,$$

où on a observé la famille de réels $X = (X_i^{(j)})_{1 \leq i \leq n, 1 \leq j \leq p}$ telle que $\text{rang}(X) = p + 1$ et on note le paramètre inconnu $\theta = (\theta_j)_{0 \leq j \leq p} \in \mathbf{R}^{p+1}$. Enfin, $(\varepsilon_i)_{1 \leq i \leq n}$ forme une suite de variables aléatoires gaussiennes indépendantes centrées telles qu'il existe $\sigma^2 > 0$, inconnu, et $1 \leq K \leq n$, $(n_1, \dots, n_K) \in \{1, \dots, n\}^K$ connus vérifiant:

$$\text{var}(\varepsilon_i) = k \sigma^2 \quad \text{pour } n_0 + \dots + n_{k-1} + 1 \leq i \leq n_0 + \dots + n_k \text{ et } 1 \leq k \leq K$$

avec $n_0 = 0$ et $n_0 + \dots + n_K = n$.

1. Ecrire le modèle suivi par $Y = {}^t(Y_1, \dots, Y_n)$ sous forme matricielle (**0.5pts**). Déterminer les matrices de variance-covariance de $\varepsilon = {}^t(\varepsilon_1, \dots, \varepsilon_n)$ et de Y (**1pt**).
2. Déterminer la vraisemblance de Y par rapport aux paramètres inconnus θ et σ^2 (**1pt**). En déduire que l'estimateur par maximum de vraisemblance de (θ, σ^2) est $(\hat{\theta}, \hat{\sigma}^2)$ qui vérifie:

$$\begin{cases} \hat{\theta} &= \underset{\theta \in \mathbf{R}^{p+1}}{\text{argmin}} \left\{ {}^t(Y - X\theta) \Omega^2 (Y - X\theta) \right\} \\ \hat{\sigma}^2 &= \frac{1}{n} {}^t(Y - X\hat{\theta}) \Omega^2 (Y - X\hat{\theta}) \end{cases},$$

où Ω est une matrice que l'on précisera (**2.5pts**).

3. Montrer que $\hat{\theta} = ({}^tX \Omega^2 X)^{-1} {}^tX \Omega^2 Y$ après avoir expliqué pourquoi ${}^tX \Omega^2 X$ est inversible (**1pt**). Déterminer la loi de $\hat{\theta} = {}^t(\hat{\theta}_0, \dots, \hat{\theta}_p)$ (**0.5pts**). En déduire la loi de $\hat{\theta}_j$ (**0.5pts**).
4. On considère $\tilde{\theta}$ l'estimateur par moindres carrés ordinaires de θ . Déterminer la loi de $\tilde{\theta}$ (**0.5pts**) et comparer son risque quadratique avec celui de $\hat{\theta}$ (**1.5pts**).
5. On note $\tilde{X} = \Omega X$. Montrer que $n \hat{\sigma}^2 = \|(I_n - \tilde{X} ({}^t\tilde{X} \tilde{X})^{-1} {}^t\tilde{X}) \Omega \varepsilon\|^2$ (**0.5pts**). En déduire la loi de $\hat{\sigma}^2$ (**1pt**) ainsi qu'un estimateur $\tilde{\sigma}^2$ de σ^2 non biaisé, dont on précisera également la loi (**0.5pts**).
6. Soit le problème de test: $H_0: \theta_1 = 0$ contre $H_1: \theta_1 \neq 0$. Déduire des questions précédentes une statistique de test \hat{T}_1 dont on précisera la loi sous H_0 (**2pts**).
7. On suppose que $K = 2$ et $1 < n_1 < n$ inconnu. Reprendre la vraisemblance de Y avec θ , σ^2 et n_1 comme paramètre et en déduire que

$$\hat{n}_1 = \underset{1 < n_1 < n}{\text{argmin}} \left\{ 2^{-\frac{n_1}{n}} \left(\sum_{i=1}^{n_1} (Y_i - (X_i^{(j)})_{0 \leq j \leq p} \hat{\theta}(n_1))^2 + \frac{1}{2} \sum_{i=n_1+1}^n (Y_i - (X_i^{(j)})_{0 \leq j \leq p} \hat{\theta}(n_1))^2 \right) \right\} \quad (\mathbf{3pts}).$$

Exercice de TP utilisant le logiciel R (Sur 7 points)

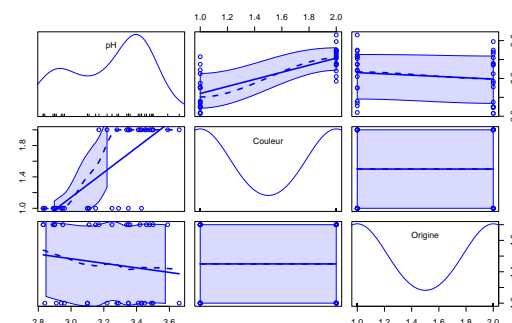
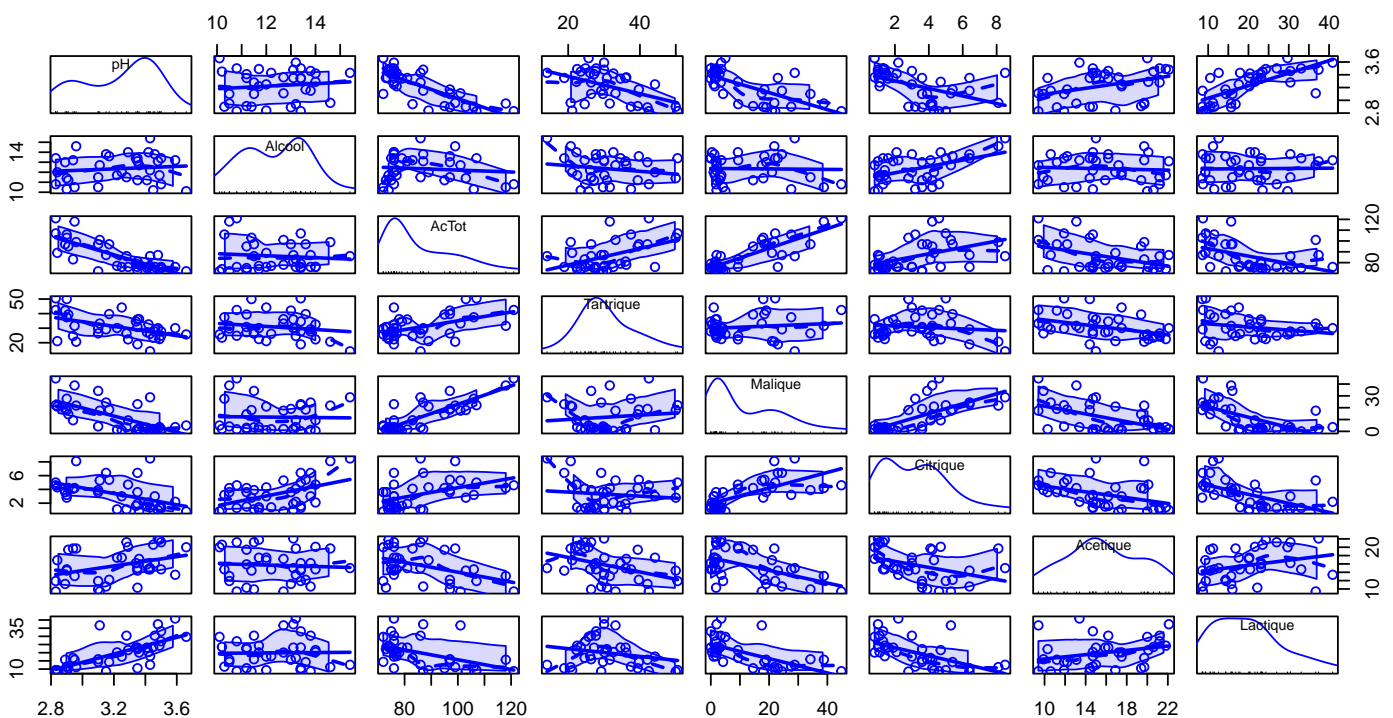
On considère la base de données **Cepage** composée de 36 observations et 10 variables. La variable que nous voulons expliquer est $Y=pH$, le pH du vin, à partir des autres variables qui sont:

- **Origine:** facteur avec 2 modalités: Bordeaux et Bourgogne.
- **Couleur:** facteur avec 2 modalités: Blanc et Rouge.
- **Alcool:** Taux d'alcool dans le vin.
- **Malique:** Acide malique .
- **Tartrique:** Acide tartarique (reflète la dureté du vin, comme le tanin).
- **Citrique:** Acide citrique (reflète la fraîcheur de vin).
- **Acétique:** Acide acétique.
- **Lactique:** Acide lactique.
- **AcTot:** Acidité totale.

1. On commence par analyser les différents liens entre les variables:

```
scatterplotMatrix(~pH + Alcool + AcTot + Tartrique + Malique +
                  Citrique + Acétique + Lactique,data=Cepages)
table(Cepages$Origine,Cepages$Couleur)
scatterplotMatrix(~pH +Couleur+Origine,data=Cepages)
```

Voici les graphes et résultats numériques obtenus:



```
> table(Cepages$Origine,Cepages$Couleur)
```

	Blanc	Rouge
Bordeaux	9	9
Bourgogne	9	9

Questions 1: Que peut-on déduire du premier graphe? Concernant le second graphe, couplé avec les données numériques, que dire des variables Origine et Couleur quant à leur influence sur pH (1pt)?

2. On tape ensuite les commandes:

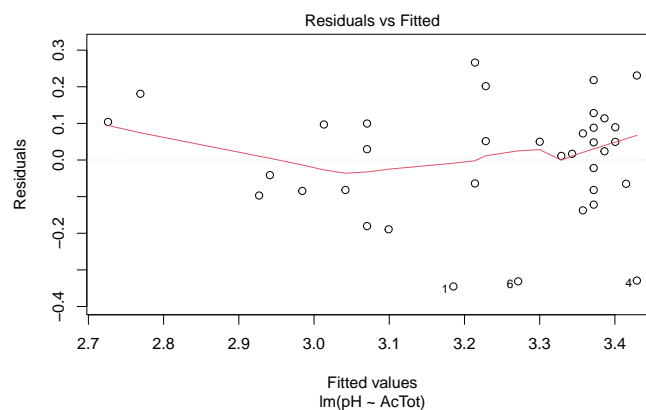
```
Reg1=lm(pH~AcTot,data=Cepages)
summary(Reg1); plot(Reg1,1)
```

Voici les résultats numériques et la figure obtenue:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.462394	0.169697	26.296	< 2e-16 ***
AcTot	-0.014350	0.001949	-7.363	1.56e-08 ***

Residual standard error: 0.155 on 34 degrees of freedom
 Multiple R-squared: 0.6146, Adjusted R-squared: 0.6032
 F-statistic: 54.21 on 1 and 34 DF, p-value: 1.558e-08



Questions 2: Que représentent formellement les valeurs -0.014350, 0.001949 et 54.21? Pourquoi a-t-on 34 DF? Que conclure des résultats numériques et du graphe? (2pts)

3. On a ensuite tapé les commandes:

```
Reg2=lm(pH~Origine*Couleur,data=Cepages); Anova(Reg2)
Reg21=lm(pH~Origine+Couleur,data=Cepages); Anova(Reg21)
Reg22=lm(pH~Couleur,data=Cepages); summary(Reg22)
```

Et à nouveau les résultats obtenus:

```
> Anova(Reg2)
```

	Sum Sq	Df	F value	Pr(>F)
Origine	0.04202	1	1.6961	0.2021
Couleur	1.25814	1	50.7768	4.367e-08 ***
Origine:Couleur	0.02722	1	1.0988	0.3024

```
> Anova(Reg21)
```

	Sum Sq	Df	F value	Pr(>F)
Origine	0.04202	1	1.691	0.2025
Couleur	1.25814	1	50.625	3.78e-08 ***

```
> summary(Reg22)
```

Coefficients:

```
Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.04056 0.03753 81.010 < 2e-16 ***
CouleurRouge 0.37389 0.05308 7.044 3.93e-08 ***
```

```
Residual standard error: 0.1592 on 34 degrees of freedom
Multiple R-squared: 0.5934, Adjusted R-squared: 0.5814
F-statistic: 49.62 on 1 and 34 DF, p-value: 3.932e-08
```

Questions 3: Expliquez ce qui a été fait et ce que l'on peut en conclure. Avec le modèle Reg22, quel serait la prédiction du pH pour un vin de Bourgogne blanc? (1.5pts)

4. On poursuit avec les commandes:

```
reg3=lm(pH ~ ., data=Cepages)
RegBIC=stepAIC(reg3,k=log(36), direction="both")
summary(RegBIC)
plot(RegBIC,1); plot(RegBIC,3)
```

Et voici une partie des résultats numériques obtenus:

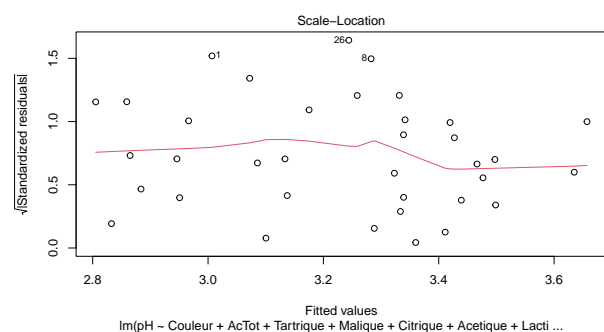
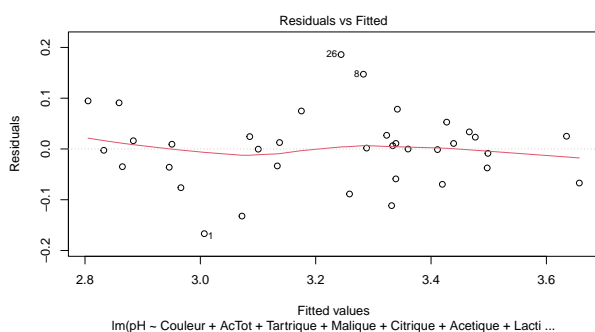
```
> Step: AIC=-161.71
pH ~ Couleur + AcTot + Tartrique + Malique + Citrique + Acetique +
Lactique
```

	Df	Sum of Sq	RSS	AIC
<none>			0.18182	-161.71
- Citrique	1	0.021727	0.20355	-161.23
+ Origine	1	0.011255	0.17057	-160.43
+ Alcoool	1	0.002191	0.17963	-158.56
- Acetique	1	0.037701	0.21952	-158.51
- Tartrique	1	0.038239	0.22006	-158.42
- Malique	1	0.148865	0.33069	-143.76
- Lactique	1	0.215726	0.39755	-137.13
- Couleur	1	0.223141	0.40496	-136.46
- AcTot	1	0.281547	0.46337	-131.61

> Coefficients:

```
Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.888267 0.190840 20.375 < 2e-16 ***
CouleurRouge 0.291017 0.049645 5.862 2.65e-06 ***
AcTot -0.021580 0.003277 -6.585 3.85e-07 ***
Tartrique 0.008215 0.003385 2.427 0.0219 *
Malique 0.019899 0.004156 4.788 4.95e-05 ***
Citrique 0.021942 0.011995 1.829 0.0780 .
Acetique 0.013192 0.005475 2.410 0.0228 *
Lactique 0.014053 0.002438 5.764 3.46e-06 ***
```

```
Residual standard error: 0.08058 on 28 degrees of freedom
Multiple R-squared: 0.9142, Adjusted R-squared: 0.8928
F-statistic: 42.64 on 7 and 28 DF, p-value: 2.791e-13
```



Questions 4: Expliquez ce qui a été fait, ce que représente -161.23 et ce qu'on en déduit. Que peut-on conclure de tout ceci (1.5pts)

5. On a enfin exécuté les commandes:

```
RegBIC2=lm(pH ~ Couleur+AcTot+Malique+Lactique,data=Cepages)
summary(RegBIC2)
anova(RegBIC2,RegBIC)
```

Avec pour résultats obtenus:

> Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.101534	0.155583	26.362	< 2e-16 ***
CouleurRouge	0.234663	0.046107	5.090	1.66e-05 ***
AcTot	-0.015853	0.002023	-7.835	7.66e-09 ***
Malique	0.012849	0.002692	4.772	4.12e-05 ***
Lactique	0.010982	0.002274	4.828	3.51e-05 ***

Residual standard error: 0.08656 on 31 degrees of freedom
 Multiple R-squared: 0.8904, Adjusted R-squared: 0.8763
 F-statistic: 62.99 on 4 and 31 DF, p-value: 1.925e-14

> anova(RegBIC2,RegBIC)

Analysis of Variance Table

Model 1: pH ~ Couleur + AcTot + Malique + Lactique

Model 2: pH ~ Couleur + AcTot + Tartrique + Malique + Citrique + Acétique + Lactique

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	31	0.23228				
2	28	0.18182	3	0.050456	2.59	0.07268 .

Questions 5: Expliquez ce qui a été fait et ce que l'on peut en conclure (1pt)