Université de Lille

Master 1 MIASHS

DM ECONOMETRIE

Souadou DIALLO
Ablaye SOW

November 8, 2025

November 8, 2025

1 Exercice1

Dans cet exercice, on cherche à comprendre comment le comportement d'un individu face à un événement peut être affecté par sa consommation de boissons : aucune consommation, consommation d'alcool seul, de café seul, ou bien des deux simultanément.

1.1 Question1: Modélisation du temps de réaction en fonction de la consommation d'alcool et de café.

Proposons une formulation dans laquelle une constante est incluse.

On note :

- $x_{1i} = 1$ si l'individu i n'a consommé ni alcool ni café, 0 sinon,
- $x_{2i} = 1$ si l'individu i a consommé de l'alcool mais pas de café, 0 sinon,
- $x_{3i} = 1$ si l'individu i a consommé du café mais pas d'alcool, 0 sinon,
- $x_{4i} = 1$ si l'individu i a consommé à la fois de l'alcool et du café, 0 sinon,
- Y_i : temps de réaction de l'individu i.

On constate que les quatre variables indicatrices vérifient :

$$x_{1i} + x_{2i} + x_{3i} + x_{4i} = 1 \quad \forall i,$$

Donc les variables ont l'air collinéaire. Il est donc nécessaire de supprimer une des variables pour estimer le modèle.
On va donc supprimer la première variable dans le modèle avec constante(groupe ni alcool, ni café) .

1. Modèle avec constante
Le modèle s'écrit alors :

$$Y_i = \beta_0 + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \varepsilon_i.$$

Interprétation :

- β_0 : moyenne du temps de réaction pour le premier groupe , ici le groupe de référence (ni alcool ni café), ordonnée à l'origine, c'est-à-dire la valeur moyenne prédite de Y quand toutes les variables explicatives Xj=0,
- β_2 : différence moyenne entre le groupe alcool seul et le groupe de référence(le groupe ni alcool , ni café),
- β_3 : différence moyenne entre le groupe café seul et le groupe de référence,
- β_4 : différence moyenne entre le groupe alcool et café et le groupe de référence,
- ε_i : les erreurs ou résidus théoriques du modéle.

2. Modèle sans constante

Dans le modèle sans constante, on peut inclure les quatre variables indicatrices :

$$Y_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{3i} + \alpha_4 x_{4i} + \varepsilon_i.$$

Interprétation :

- Chaque coefficient α_j représente directement la moyenne du temps de réaction pour le groupe j.
- ε_i : les erreurs ou résidus théoriques du modéle.

1.2 Question 2 :

1.2.1 Estimons chacun des deux modèles et commentons nos résultats

Modèle avec constante

En langage `R`, ce modèle est estimé par :

```
model_1 <- lm(Y ~ A.seul + C.seul + A.et.C, data = df)
summary(model_1)
```

Les résultats obtenus sont les suivants :

	Estimate	Std. Error	t value	Pr($\geq \lvert t \rvert$)
(Intercept)	1.3576	0.0895	15.18	$<$ 2e-16 ***
A.seul	1.7875	0.1397	12.80	$<$ 2e-16 ***
C.seul	0.6456	0.1573	4.11	0.000106 ***
A.et.C	1.1080	0.1239	8.94	3.03e-13 ***

Résidu standard : 0.429 sur 71 degrés de liberté.
$R^2 = 0.7152$, $R^2_{ajusté} = 0.7031$, $F(3, 71) = 59.42$, $p < 2.2 \times 10^{-16}$.

Interprétation:

- L'intercept $\hat{\beta}_0 = 1.3576$ représente la moyenne du temps de réaction pour le groupe *ni alcool ni café* , qui sert de référence,
- Le coefficient $\hat{\beta}_2 = 1.7875$ indique que le groupe *alcool seul* a en moyenne un temps de réaction supérieur de 1.7875 unités à celui du groupe 1,
- Le coefficient $\hat{\beta}_3 = 0.6456$ montre que le groupe *café seul* présente un temps de réaction plus élevé de 0.6456 unités en moyenne,
- Le coefficient $\hat{\beta}_4 = 1.1080$ signifie que le groupe *alcool et café* a un temps de réaction supérieur de 1.1080 unités à celui du groupe de référence.

Tous les coefficients sont hautement significatifs ($p < 0.001$), ce qui indique que les différences observées entre les groupes sont statistiquement significatives. Le coefficient de détermination $R^2 = 0.7152$ montre que le modèle explique environ 71,5% de la variabilité totale du temps de réaction, ce qui traduit un bon ajustement global.

Modèle sans constante

En langage `R`, ce modèle est estimé par :

```
model_1 <- lm(Y ~ 0 + ni.A.ni.C + A.seul + C.seul + A.et.C, data = df)
summary(model_2)
```

Les résultats obtenus sont les suivants :

	Estimate	Std. Error	t value	Pr($>\|t\|$)
ni.A.ni.C	1.3576	0.0895	15.18	$<$ 2e-16 ***
A.seul	3.14504	0.10724	29.33	$<$ 2e-16 ***
C.seul	2.00314	0.12934	15.49	$<$ 2e-16 ***
A.et.C	2.46555	0.08579	28.14	$<$ 2e-16 ***

Résidu standard : 0.429 sur 71 degrés de liberté.
$R^2 = 0.9681$, $R^2_{ajusté} = 0.9663$, $F(4, 71) = 539$, $p < 2.2 \times 10^{-16}$.

Interprétation.

- Le coefficient $\hat{\alpha}_1 = 1.3576$ correspond à la moyenne du temps de réaction pour le groupe *ni alcool ni café* .
- Le coefficient $\hat{\alpha}_2 = 3.1450$ correspond à la moyenne du groupe *alcool seul* .
- Le coefficient $\hat{\alpha}_3 = 2.0031$ correspond à la moyenne du groupe *café seul* .
- Le coefficient $\hat{\alpha}_4 = 2.4656$ correspond à la moyenne du groupe *alcool et café* .

On constate par ailleurs que les gens qui boivent de l'alcool prennnent plus de temps pour réagir lorsqu'un évènement se produit.

Toutes les estimations sont bien significatives ($p < 0.001$), montrant ainsi que les moyennes des différents groupes diffèrent de manière significative. Le coefficient de détermination $R^2 = 0.9681$ indique que 96,8% de la variabilité du temps de réaction est expliquée par le modèle, ce qui sous entend une très bonne estimation.

1.2.2 Etablissons des liens entre les estimations des paramètres dans un modéle et dans l'autre.

Il semble que les coefficients des deux modèles soient directement liés parce que, dans le modèle avec constante les β_j représentent les différences entre le groupe j et le groupe de référence (groupe 1), et que, dans le modèle sans constante, les α_j représentent les moyennes de chaque groupe.
On peut donc écrire les relations suivantes qui peuvent être obtenues même en égalant les expressions des deux modèles.

$$\alpha_1 = \beta_0, \quad \alpha_2 = \beta_0 + \beta_2, \quad \alpha_3 = \beta_0 + \beta_3, \quad \alpha_4 = \beta_0 + \beta_4,$$

et inversement :

$$\beta_0 = \alpha_1, \quad \beta_2 = \alpha_2 - \alpha_1, \quad \beta_3 = \alpha_3 - \alpha_1, \quad \beta_4 = \alpha_4 - \alpha_1.$$

Toutes ces relations peuvent être vérifiées dans R et sont disponible dans le fichier R.

1.3 Question 3 :

1.3.1 Pourquoi la différence de coefficient de détermination dans les deux modèles?

Le coefficient de détermination R^2 mesure la proportion de la variance totale de la variable dépendante Y expliquée par le modèle. Il est défini par :

$$R^2 = \frac{\text{Ve}(\hat{Y})}{\text{Ve}(\hat{Y}) + \text{Ve}(\hat{\epsilon})}$$

où :

- $\text{Ve}(\hat{Y})$ correspond à la variance empirique de la partie de Y expliquée par le modèle.
- $\text{Ve}(\hat{\epsilon})$ correspond à la variance empirique de la partie de Y non expliquée par le modèle.

Dans le modèle **avec constante**, la somme des carrés totale est calculée par rapport à la moyenne globale $\overline{Y}$. Alors que, dans le modèle **sans constante**, la décomposition de la variance n'est plus faite autour de cette moyenne globale mais autour de zéro.
Ce qui conduit généralement à un R^2 plus élevé, car la variance totale autour de zéro est plus grande que celle autour de la moyenne $\overline{Y}$.

Autrement dit, le modèle sans constante surestime la part de variance expliquée, tout en sachant que les résidus sont identiques dans les deux cas.

Dans le modèle avec constante, la première variable (ni.A.ni.C) est incluse dans l'intercept , ce qui diminue le nombre de variables que le modéle prend en compte d'ou le fait qu'il n'explique que 71.52% de la variance de Y.
Tandis que, dans le modèle sans constante, cette catégorie est explicitement modélisée avec ni.A.ni.C et l'intercept est retiré, ce qui permet au modèle d'expliquer presque toute la variance de Y, d'où un R2 beaucoup plus élevé(96.81%).

1.3.2 Qu'est ce qui'il faudrait faire pour interpréter correctement ce coefficient de determination?

Il semble que dans le modèle avec constante, on peut interpréter correctement R^2 comme la part de la variabilité de Y expliquée par les variables explicatives, tandis que, dans le modèle sans constante, on peut dire que le R^2 ne peut pas être interprété comme une proportion de variance expliquée , tendant souvent à être élevé et donc pas comparable avec le modéle avec constante.
Il faudrait plûtot se fier aux valeurs ajustées et aux résidus pour juger de la qualité du modèle.

1.4 Question 4:

1.4.1 Si on considère le modèle sans constante, exprimons la matrice X utilisée dans l'expression de l'estimateur MC des coefficients.

Dans le modèle sans constante :

$$Y = X\alpha + \varepsilon$$

avec

$$X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}_{75\times 4}$$

Sur R, on a :

```
# Matrice X
X = as.matrix(cbind(df$ni.A.ni.C, df$A.seul, df$C.seul, df$A.et.C))
```

1.4.2 En effectuant un calcul à la main, démontrons que les coefficients des estimations obtenues dans le modèle sans constante doivent nécessairement coincider avec des moyennes empiriques.

L'estimateur des moindres carrés est :

$$\hat{\alpha} = (X'X)^{-1}X'Y.$$

Or, la matrice $X'X$ est diagonale puisque les colonnes de X sont disjointes :

$$X'X = \begin{pmatrix} n_1 & 0 & 0 & 0 \\ 0 & n_2 & 0 & 0 \\ 0 & 0 & n_3 & 0 \\ 0 & 0 & 0 & n_4 \end{pmatrix}$$

$$(X'X)^{-1} = \begin{pmatrix} 1/n_1 & 0 & 0 & 0 \\ 0 & 1/n_2 & 0 & 0 \\ 0 & 0 & 1/n_3 & 0 \\ 0 & 0 & 0 & 1/n_4 \end{pmatrix}$$

car det(X'X) $= n_1 \cdot n_2 \cdot n_3$
où n_j est le nombre d'individus dans chaque modalité.
De plus :

$$X'Y = \begin{pmatrix} \sum_{i\in \text{ni.A.ni.C}} Y_i \\ \sum_{i\in \text{A.seul}} Y_i \\ \sum_{i\in \text{C.seul}} Y_i \\ \sum_{i\in \text{A.et.C}} Y_i \end{pmatrix}.$$

Comme l'estimateur de α est donnée par $\hat{\alpha} = (X'X)^{-1} \cdot X' \cdot Y$, on peut écrire :

$$\hat{\alpha}_j = \frac{1}{n_j} \sum_{i \in modalité_j} Y_i,$$

c'est-à-dire que chaque coefficient estimé correspond à la **moyenne empirique de Y** dans le groupe j.

Ce qu'il fallait démontrer!
Sur R, on a
`solve(t(X) %*% X) %*% t(X) %*% Y` qui correspond exactement aux valeurs des coefficients trouvés récemment dans l'estimation du modèle sans constante (les moyennes des temps de réactions pour les différents groupes).

Les coefficients du modèle sans constante sont donc les moyennes observées dans chaque modalité.

1.4.3 Commentons ce résultat!

On a que montré que les coefficients du modèle sans constante sont les moyennes observées dans chaque modalité.

$$\hat{\alpha}_j = \bar{Y}_j.$$

Ce résultat signifie que la régression linéaire sans constante, avec des variables indicatrices, revient à estimer les moyennes empiriques des groupes observés. Cela explique également pourquoi le coefficient de détermination R^2 est très élevé.

1.5 Question 5:

1.5.1 Dans chacune des deux formulations(avec et sans constante) testons l'hypothèse que les consommations d'alcool ou de café n'ont aucun effet sur le temps de réaction.

On souhaite tester l'hypothèse que les consommations d'alcool et de café n'ont aucun effet sur le temps de réaction.

1. Hypothèses

- Hypothèse nulle :

$$H_0 : \beta_{\text{A.seul}} = \beta_{\text{C.seul}} = \beta_{\text{A.et.C}} = 0$$

- Hypothèse alternative :

$$H_1 : \text{au moins un des coefficients } \beta \text{ est différent de } 0$$

2. Statistique de test

Pour répondre à cette question dans la meilleure des manières, nous allons utilisé les tests de contraintes linéaires sur β :
On pose :

$$R = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$r = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Ainsi on a $R \cdot \beta = r$

Comment interpréter?

- Si p-value < 0.05, on rejette H_0 : cela signifie que la consommation d'alcool ou de café a un effet significatif sur le temps de réaction.
- Si p-value ≥ 0.05, on ne rejette pas H_0 : Pas de preuve statistique que la consommation d'alcool ou de café influence le temps de réaction.

Application aux modèles

- **Modèle avec intercept :** $Y \sim A.seul + C.seul + A.et.C$
- **Modèle sans intercept :** $Y \sim 0 + ni.A.ni.C + A.seul + C.seul + A.et.C$

Dans les deux cas, on teste simultanément les coefficients $\beta_{\text{A.seul}}, \beta_{\text{C.seul}}, \beta_{\text{A.et.C}}$ pour évaluer l'effet des consommations.
Dans le deuxième modéle, on omet la variable indicatrice ni.A.ni.C parce qu'on veut tester l'effet de la consommation d'alcool ou de café, donc il n'est pas intéressant de la tester.

Voici le code R utilisé pour le test de l'effet des consommations

```
# Charger le package
library('car')

# Modèle avec constante
linearHypothesis(model_1, c("A.seul = 0", "C.seul = 0", "A.et.C = 0"))

# Modèle sans constante
linearHypothesis(model_2, c("A.seul = 0", "C.seul = 0", "A.et.C = 0"))
```

Voici les résultats et l'interprétation

1.5.2 Modéle 1:

- Statistique F : $F = 59.42$
- p-value : $< 2.2 \times 10^{-16}$
- La p-value très faible permet de rejeter l'hypothèse nulle $H_0 : \beta_{\text{A.seul}} = \beta_{\text{C.seul}} = \beta_{\text{A.et.C}} = 0$. Cela signifie qu'au moins une des consommations (alcool ou café) a un effet significatif sur le temps de réaction par rapport au groupe de référence.
  C'est-à-dire que, comparé aux personnes ne consommant ni alcool ni café, certaines consommations ont un effet significatif sur le temps de réaction.
- Une statistique de test F grande est observée et favorise cet effet significatif.

1.5.3 Modèle 2 :

- Statistique F : $F = 641.92$
- p-value : $< 2.2 \times 10^{-16}$
- Interprétation : La p-value significativement plus petite que 0.05 permet de rejeter l'hypothèse nulle $H_0 : \beta_{\text{A.seul}} = \beta_{\text{C.seul}} = \beta_{\text{A.et.C}} = 0$. Le modèle sans intercept explique presque toute la variance de Y, et les consommations d'alcool et/ou de café ont un effet significatif sur le temps de réaction.

2 Exercice 2

Cette partie a pour objectif d'analyser la relation entre la qualité gustative du cheddar et ses caractéristiques chimiques, tout en tenant compte du mode de production, à l'aide de données issues d'observations expérimentales.

2.1 Question 1: Modèle adéquat

$$TASTE_i = \beta_0 + \beta_1 ACETIC_i + \beta_2 H2S_i + \beta_3 LACTIC_i + \beta_4 ORG_i + \varepsilon_i.$$

avec $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, où $TASTE_i$ désigne la note moyenne de dégustation du cheddar i, et $ORG_i = 1$ si le cheddar est certifié biologique, 0 sinon.

2.2 Question 2: Effet du mode de production (BIO ou pas)

L'effet du mode production (BIO ou pas) se mesure à travers le coefficient β_4 de notre modèle qui représente la différence moyenne de qualité gustative entre les cheddars BIO et non BIO, toutes propriétés chimiques égales.

- Si cette valeur est négative(et significativement nulle), on peut dire que la qualité gustative des cheddars bio est jugée moins bonne par par rapport aux cheddars non bio.
- Si elle est nulle, on peut dire que le mode de production n'a aucun effet sur la qualité gustative connaissant ses autres propriétés chimiques.
- Si elle est grande, on conclut que sa variabilité apporte d'importantes informations sur la qualité gustative (c'est-à dire que les cheddars bio ont une meilleure note gustative en moyenne).
- Si elle proche de zéro, on conclut le mode de production n'affecte presque pas la qualité gustative du cheddar.

2.2.1 Mesure de l'effet du mode de production:

Pour procèder à sa mesure, on peut réaliser un test d'hypothèses de Fisher pour voir si l'effet est significatif ou pas.

2.3 Question 3: Commandes R

2.3.1 Paramètres du modèles

```
MODELE = lm(formula = TASTE ~ ACETIC + H2S + LACTIC + ORG, data = cheddar)
```

```
MODELE$coefficients
```

nous donne la valeurs estimée de β c.à.d $\hat{\beta}$

Les résultats sont regroupés dans la matrice suivante : $\hat{\beta} = \begin{pmatrix} -3.334601 \\ -2.423550 \\ 2.7469 \\ 13.663624 \\ 19.381454 \end{pmatrix}$

- $\widehat{\beta}_0 = -3.33$: On va ignorer cette valeur de l'intercept qui n'est valable que si les propriétés chimiques du cheddar non bio sont nulles , ce qui n'est jamais vrai. Cette valeur n'a pas de sens concret ici (puisqu'il n'existe pas de cheddar avec toutes les teneurs nulles). Elle sert quand même de référence pour les autres coefficients.
- $\widehat{\beta}_1 = -2.42$: lorsque la teneur en acide acétique (`ACETIC`) augmente d'une unité, la note gustative moyenne (`TASTE`) diminue de 2.42 points, car la valeur étant négative.
- $\widehat{\beta}_2 = 2.75$: lorsque la teneur en hydrogène sulfuré (`H2S`) augmente d'une unité, la note gustative augmente de 2.75 points.
- $\widehat{\beta}_3 = 13.66$: une augmentation d'une unité de la teneur en acide lactique (`LACTIC`) correspond à une hausse moyenne de 13.66 points de la note gustative: On note une influence forte du `LACTIC` sur la qualité gustative.
- $\widehat{\beta}_4 = 19.38$: les cheddars BIO (`ORG`=1) obtiennent en moyenne une note gustative supérieure de 19.38 points à celle des cheddars non-BIO (`ORG`=0), avec les mêmes teneurs chimiques.

2.3.2 L'estimation de la variance de Y

```
S_MODELE =  summary(MODELE)
```

$$\hat{\sigma}^2 = S_{\text{MODELE}}\sigma \times S_{\text{MODELE}}\sigma$$
$$\hat{\sigma}^2 = 48.77175$$

Cette valeur apparait sur le sommaire de notre modèle. Elle correspond à la valeur de residuals standard error élevée au carré.

2.3.3 Interprétation de F-statistic

$S\_MODELE\$fstatistic[1]$ nous permet d'avoir la valeur de la statistique du test global c'est à dire $H_O : b1 = ... = b_p = 0$ contre H_1 : il existe $i \in 1, ..., p$ telque $b_i \neq 0$.
Cette valeur nous permet de tirer des conclusions à propos des hypothèses formulées. Si elle est grande, ça suffira largement pour rejeter l'hypothèse H_0.
Dans notre cas, cette valeur vaut 33.02933. Donc, on peut retenir l'hypothèse H_1 qui considère que les coefficients ont un impact significatif c'est-à-dire que les variations des propriétés chimiques impactent la qualité gustative du cheddar.

2.4 Question4

2.4.1 Estimons par moindres carrés un modéle dans lequel on a retiré les variables explicatives dont les coefficients ne sont pas significativement différents de 0.

On définit le nouveau modèle par :

$$\text{TASTE} = \beta_0 + \beta_1 \cdot \text{H2S} + \beta_2 \cdot \text{LACTIC} + \beta_3 \cdot \text{ORG} + \epsilon$$

2.4.2 Comment pouvons nous expliquer que le coefficient de détermination est plus faible dans l'estimation de ce modéle que dans le modéle initial.

Nous savons que lorsqu'on diminue la taille de notre modèle, la variabilité de la partie de TASTE non expliquée augmente et celle de la partie expliquée diminue. Or, ce coefficient de détermination est définie par $R^2 = \frac{\text{Ve}(\hat{Y})}{\text{Ve}(\hat{Y})+\text{Ve}(\hat{\epsilon})}$ ce qui confirme bien le résultat observé (c'est-à-dire que le coefficient de détermination du modèle complet est généralement plus élevé que celui du modèle réduit).
avec :

- $\text{Ve}(\hat{Y})$ correspond à la variance empirique de la partie de Y expliquée par le modèle.
- $\text{Ve}(\hat{\epsilon})$ correspond à la variance empirique de la partie de Y non expliquée par le modèle.

De plus, ce coefficient permet de mesurer la fiabilité de notre modèle, or, nous savons de façon intuitive qu'un modèle linéaire devient plus robuste losqu'on augmente le nombre de variables hexogènes(variables explicatives).

2.4.3 Est-ce une constatation propre au jeu de données utilisées ou est ce une propriété qui est valable pour n'importe quelle jeu de données ?

En vue de la formule qui permet de calculer ce coefficient, on peut dire que ce constat est général pour n'importe quel jeu de données.

D'ailleurs, l'exercice 1 en est une preuve.
Si on retire une variable , le coefficient de détermination baisse et donc plus le nombre de variable est élevé, plus le modèle s'ajuste correctement aux données d'ou l'augmentation du R^2.

2.5 Question 5 :

2.5.1 Estimons par moindres carrés le modéle avec toutes les variables explicatives, mais sans terme constant.

On définit ce modèle par :
$\text{TASTE} = \beta_1 \cdot \text{ACETIC} + \beta_2 \cdot \text{H2S} + \text{LACTIC} + \text{ORG} + \epsilon$
En langage `R`, ce modèle est estimé par :

```
modele_3 <- lm(TASTE ~ 0 + ACETIC + H2S + LACTIC + ORG, data = data)
summary(model_3)
```

2.5.2 Constat à propos du coefficient de détermination.

Le coefficient de détermination $R^2 = 0.9525$. Nous constatons que cette valeur est proche de 1 ce qui est une bonne nouvelle pour la robustesse du nouveau modèle. Il est beaucoup plus élevé que dans le modèle avec constante. Mais, on ne peut pas dire que ce modéle est mieux.

2.5.3 Interprétation

Le R^2 sera artificiellement élevé car il est calculé par rapport à l'origine et non à la moyenne. Cette valeur n'est donc pas comparable au R^2 du modèle avec constante car, la valeur élevée n'est dû qu'à une différence de calcul et non à une réelle amélioration de la quaité d'ajustement.

2.6 Question 6 :

2.6.1 Pour chaque variable explicative, testons au niveau 2%, si ses variations provoquent des variations de la variable dépendante.

Dans cette partie, nous allons utiliser un test de student sur les coefficients pour voir si les variables agissent ou pas sur la variabilité de la qualité gustative.

Pour ce faire, on définit le test par :
$H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$
La statistique du ième test est donné par :
$t_j = \frac{\widehat{\beta}_j}{\widehat{\sigma}_j}$

La commande permettant d'obtenir le sommaire du modèle de départ est écrite à la question 5. On peut maintenant récupérer le vecteur des valeurs statistiques qui correspond à la troisième colonne du tableau des coefficients.

$$t\_values = \begin{pmatrix} -0.2317446 \\ -0.7778947 \\ 3.0976865 \\ 2.2585549 \\ 5.4508928 \end{pmatrix}$$ et le quantile vaut : 2.499867

Ici, seules les variables ORG et H2S ont une valeur de statistique supérieure au quantile en valeur absolue. D'ailleurs, ce sont les seules aussi qui ont une p-value inférieure à 0.02. Donc, on rejette l'hypothèse H_0 pour ces deux propriétés chimiques.
Ainsi, on peut conclure qu'au niveau alpha = 2%, seules ces deux variables agissent sur la variation de la qualité gustative du cheddar et positivement(leur augmentation améliore le goût du cheddar).

2.6.2 Testons au niveau 1%, l'hypothèse qu'une augmentation de la teneur en acide acétique a exactement le même effet sur la qualité gustative qu'une baisse identique de la teneur en hydrogène sulfuré.

Dans cette partie on prend comme hypothèse :
$H_0 : \beta_1 = -\beta_2$ contre $H_1 : \beta_1 \neq \beta_2$
Pour mieux modéliser cette hypothèse, nous faisons recours au test de fisher. Pour cela on a utilisé la librairie (car) afin de répondre à cette question posée. Ici les matrices R = (0,1,1,0,0) et r = (0).
Les résulats obtenus après application sont :
$F = 0.01198526$
$p\_value = 0.913698$
Comme le p_value est supérieur au seuil donné(alpha = 0.01), donc on peut conclure que l'hypothèse H_0 est accepté c'est à dire qu'une augmentation de la teneur en acide acétique a exactement le même effet sur la qualité gustative qu'une baisse identique de la teneur en hydrogène sulfuré.

D'ailleurs , les résultats de l'estimation avaient montré ce résultat :

$$\widehat{TASTE}_i = -3.06\, ACETIC_i + 2.80\, H2S_i + 13.52\, LACTIC_i + 19.65\, ORG_i,$$

On voit très bien l'effet positif d'une augmentation du H2S sur la qualité gustative et à l'opposé, l'effet négatif d'une hausse de l'acide ACETIC qui risque de dégrader ce goût du cheddar.

2.6.3 Avec le même niveau, testons que le carré de la somme des coefficients de variables chimiques (ACETIC,LACTIC,H2S) est égale à 10 fois le coefficient de la variable ORG.

On définit l'hypothèse par :
$H_0 : (\beta_1 + \beta_2 + \beta_3)^2 = 10\,\beta_4$ contre $H_1 : (\beta_1 + \beta_2 + \beta_3)^2 \neq 10\,\beta_4$

Pour une réponse beaucoup plus explicite, on utilise la forme de test générale définie dans le cours(test Wald) en posant $g(\beta) = (\beta_1 + \beta_2 + \beta_3)^2 - 10\,\beta_4$
Après application, on trouve que la valeur de la statistique vaut 0.0001169535 qui est inférieure au quantile d'ordre $1 - \alpha$ de la loi de Khi2 à 1 degré de liberté ($quantil\_Khi2 = 6.634897$).Donc, on peut conclure que l'hypothèse H_0 est accepté.