



EC 651 et 56/53 : Modélisation statistique

Projet : régression linéaire

Guillaume Franchi

Année universitaire 2025-2026

■ Présentation du projet

Dans le cadre d'une étude qualité d'une pâte à tartiner, un service R&D souhaite mieux comprendre les facteurs influençant la viscosité du produit final. Celle-ci est une caractéristique essentielle car elle impacte :

- ▶ la « tartinabilité »,
- ▶ la perception sensorielle,
- ▶ la stabilité au stockage,
- ▶ le comportement lors du conditionnement industriel.

Afin d'identifier les variables influençant cette propriété, 250 essais expérimentaux ont été réalisés en faisant varier :

- la formulation (ingrédients),
- les paramètres de fabrication,
- certaines caractéristiques physico-chimiques,
- les lots de noisettes.

L'objectif est de construire un modèle statistique explicatif de la viscosité et d'identifier les variables réellement influentes.

Remarque : les données utilisées ici ont été simulées de façon à des fins pédagogiques, et ne reflètent pas nécessairement la réalité industrielle.

Par binôme, vous devrez rendre sur connect pour le **vendredi 3 avril 23h59** :

- un compte-rendu (*format CR_nombinome1_nombinome2.pdf*),
- un script R (*format script_nombinome1_nombinome2.R*),

permettant de répondre au problème posé. Vous vous appuyerez sur les méthodes présentées en cours.

■ Présentation des données

Le jeu de données « [pates_tartiner.csv](#) », disponible sur connect, contient donc les résultats de 250 expériences sur différentes formulations de pâtes à tartiner. On présente ci-dessous les variables présentes dans le dataset (*les valeurs sont données pour 100g de produit fini*).

- ▷ Variable réponse :
 - *viscosite* : viscosité mesurée du produit final (en Pa.s) ;
- ▷ Variables de formulation :
 - *sucres* : quantité de sucre utilisée (en g/100g) ;
 - *noisettes* : quantité de noisettes utilisée (en g/100g) ;
 - *cacao* : quantité de cacao utilisée (en g/100g) ;
 - *poudre_lait* : quantité de lait en poudre utilisée (en g/100g) ;
 - *proteines* : teneur en protéines (en g/100g) ;
 - *fibres* : quantité de fibres dans la pâte (en g/100g) ;
 - *eau* : quantité d'eau utilisée (en cl/100g) ;
 - *émulsifiant* : quantité d'émulsifiant utilisée (en g/100g) ;
 - *sel* : quantité de sel utilisée (en g/100g) ;
 - *huile_palme* : quantité d'huile de palme utilisée (en cL/100g) ;
 - *huile_colza* : quantité d'huile de colza utilisée (en cL/100g) ;
- ▷ Variables de procédé :
 - *temperature_fabrication* : température lors du mélange (en °C) ;
 - *temps_melange* : durée du mélange (en min) ;
 - *vitesse_melange* : vitesse d'agitation (en tr/min)
 - *pression_homogeneisation* : pression d'homogénéisation (en bar)
 - *lot_noisettes* : numéro du lot de provenance des noisettes ;
- ▷ Variables physico-chimiques :
 - *pH* : pH du produit ;
 - *granulometrie* : taille moyenne des particules (en μm) ;
 - *humidite_stockage* : humidité relative du lieu de stockage (en %).

■ Problème

Ajuster un modèle statistique permettant d'expliquer et de prédire la viscosité d'une pâte à tartiner en fonction des autres variables.

Pour ce faire, on ajustera un modèle linéaire le plus judicieux possible. On trouvera ci-dessous une liste non exhaustive d'éléments attendus dans cette étude.

- Etude des corrélations.
- Modèle linéaire et tests de significativité.
- Sélection de variables.
- Analyse des résidus et validation du modèle.

Le jeu de données « [pates_tartiner_extra](#) », disponible sur connect, contient les données pour 20 pâtes à tartiner supplémentaires. On pourra tester la qualité prédictive du modèle ajusté sur ces données (*en étudiant le MSE par exemple*).

■ Une indication de code...

Afin de simplifier le traitement de la variable qualitative *lot_noisettes*, on pourra se servir du code ci-dessous.

```
library(dplyr)
pates_tartiner <- pates_tartiner %>%
  mutate(val=1) %>%
  pivot_wider(names_from = lot_noisettes,
              values_from = val,
              values_fill =0) %>%
  select(-Lot_1)
```