Institut Polytechnique de Paris
École Nationale Supérieure de Techniques Avancées

Contrôle de la teneur en sucre de biscuits à partir de spectroscopie infrarouge

Réalisé par :
Marwen Bahri
Louay El Melki

2ème année Techniques Avancées

16 mai 2021

Table des matières

Table des figures

Introduction

Notre étude porte sur le contrôle de qualité de biscuits, plus particulièrement la teneur en sucre des biscuits. Le contrôle qualité pourrait s'effectuer dans un laboratoire en utilisant des procédés chimiques mais celles-ci étant très couteuses, il était décidé d'utiliser un spectromètre en proche infrarouge (NIR). L'appareil envisagé mesure l'absorbance c'est-à-dire les spectres dans les longueurs d'ondes du proche infra-rouge. On utilisera pour ce projet un jeu de données fourni par Osbone et al. (1984) [1]. Le jeu de donnée est divisé en deux, une partie consacrée à l'apprentissage des modèles statistiques qu'on développera et une partie test pour tester les différents modèles. Les deux jeux de donnés sont de tailles 40 et 32 respectivement et tous les deux comportent 700 observations donc cette étude se déroule dans un contexte de grande dimension. En ce qui concerne l'organisation du travail, on commence par une étude théorique en élaborant plusieurs formules qui seront utiles ultérieurement lors de l'implémentation sur machine des modèles statistiques. Ensuite on procède à une analyse exploratoire pour bien représenter les données. Enfin on développe plusieurs modèles de régression : régression ridge et régression logistique pénalisé.

1. B. G. Osborne, T. Fearn, A. R. Miller et S. Douglas, Application of Near Infrared Reflectance spectroscopy to the compositional analysis of biscuits and biscuit doughs, J. Sci. Food Agric. 35 (1984), 99 - 105

Chapitre 1

Un peu de théorie

1.1 Régression linéaire et régression ridge (Question 1)

Le nombre p de variables explicatives étant très largement supérieur à n, il existe nécessairement des variables très corrélées (multicolinéarité) et donc la matrice $X'X$ a des valeurs propres très proches de zéro, ce qui implique que les valeurs propres de la matrice $(X'X)^{-1}$ sont très grandes. Par conséquent, si l'on modélise une régression linéaire, la variance de l'estimateur des moindres carrés $\widehat{\theta}_{MC}$ qui est égale à $var(\widehat{\theta}_{MC}) = (X'X)^{-1}$ sera très grande ce qui résulte en des problèmes de précision du modèle. La régression ridge vient dans le cadre de grande dimension où $n < p$, et elle permet de pallier le problème du non inversibilité de la matrice $X'X$ en lui apportant le terme $\kappa \mathbb{1}_p$ ce qui va assurer que toutes ses valeurs propres soient positives et donc réduit la variance de l'estimateur θ.

1.2 Calcul d'estimateurs (Question 2)

Dans la régression ridge on ajoute un terme de pénalisation à la somme des carrés résiduels ce qui va rétrécir les coefficients de θ. La SCR de la régression ridge est définie par :

$$SCR(\theta) = \sum_{i=1}^{n} \left(Y_i - \theta_0 - \sum_{j=1}^{p} \theta_j x_{ij} \right)^2 + \kappa \sum_{j=1}^{p} \theta_j^2$$

qui l'on peut aussi écrire sous cette forme :

$$SCR(\theta) = \|Y - \theta_0 1_n - X\theta\|^2 + \kappa \|\theta\|^2$$

avec 1_n est le vecteur colonne dont toutes les composantes sont égales à 1. On obtient donc la définition de l'estimateur ridge suivant :

$$\hat{\theta}_{\text{ridge}}\,(\kappa) = \arg\min_{\theta \in \mathbb{R}^P} \|Y - \theta_0 1_n - X\theta\|^2 + \kappa \|\theta\|^2$$

que l'on peut aussi écrire sous forme d'un problème d'optimisation sous contrainte l_2 avec son lagrangien associé suivants :

$$\tilde{\theta} = \arg\min_{\theta \in \mathbb{R}^p ; \|\theta\|^2 \leq \delta} \|Y - \theta_0 1_n - X\theta\|^2$$

$$L(\theta, \kappa) = \|Y - \theta_0 1_n - X\theta\|^2 + \kappa\|\theta\|^2$$

δ doit être petit car si non $\tilde{\theta}$ est l'EMC ordinaire. En dérivant le lagrangien on obtient :

$$-2X'\left(Y - \theta_0 1_n - X\theta\right) + 2\kappa\theta = 0$$

d'où $\tilde{\theta} = \hat{\theta}_{\text{ridge}} = \left(X'X + \kappa Id_p\right)^{-1}\left(X'Y - X'\theta_0 1_n\right)$

Soit maintenant la matrice $\bar{X}$ tels que $X_{ij} = \bar{X}_j$, $1 \leq i \leq n, 1 \leq j \leq p$. $\bar{X}_j$ est la valeur moyenne de la variable explicative j. la paramétrisation $\tilde{\theta}$ quand les variables ont été préalablement centrées est définie par :

$$\tilde{\theta} = \left((X - \bar{X})'(X - \bar{X}) + \kappa Id_p\right)^{-1}\left((X - \bar{X})'Y - (X' - \bar{X})\theta_0 1_n\right)$$

$$\tilde{\theta} = \left((X - \bar{X})'(X - \bar{X}) + \kappa Id_p\right)^{-1}\left((X'Y - \bar{X}'Y) - (X' - \bar{X})\theta_0 1_n\right)$$

Or $X'Y = (X'X + \kappa Id_p)\theta + X'\theta_0 1_n$ donc :

$$\tilde{\theta} = \left((X - \bar{X})'(X - \bar{X}) + \kappa Id_p\right)^{-1}\left((X'X + \kappa Id_p)\theta - \bar{X}'Y + \bar{X}\theta_0 1_n\right)$$

Si on revient maintenant à l'expression du lagrangien et on la dérivant par rapport à θ_0, on obtient :

$$1_n'(Y - \theta_0 1_n - X\theta) = 0 = \sum_{i=1}^{n}\left(y_i - \theta_0 - \sum_{j=1}^{p}\left(x_{ij}\theta_j\right)\right)$$

ce qui donne :

$$\hat{\theta_0} = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}\theta_j\right)$$

dans le cas où les variables explicatives ont été centrées le terme de la somme sur x_{ij} sera nul. En effet, soit $\overline{X}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$:

$$\sum_{i=1}^{n}\sum_{j=1}^{p}\left(x_{ij} - \overline{X}_j\right)\theta_j = \sum_{j=1}^{p}\sum_{i=1}^{n}\left(x_{ij} - \overline{X}_j\right)\theta_j = \sum_{j=1}^{p}\left(\sum_{i=1}^{n} x_{ij} - n\overline{X}_j\right)\theta_j = 0$$

on obtient donc :

$$\tilde{\theta_0} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Finalement si la variable cible est centrée aussi, alors $\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y}) = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\overline{y}\right) = 0$ et par conséquent :

$$\tilde{\theta_0} = 0$$

1.3 Limite d'une matrice (Question 3)

Soit X une matrice de dimension $n \times p$. L'objectif de cette question est de trouver la limite de la matrice $A_\lambda = \left(X'X + \lambda Id_p\right)^{-1} X'$ quand λ tend vers 0 dans le cas où X n'est pas injective. Soit $\sum_{j=1}^{r}\sigma_j u_j v_j'$ une décomposition en valeurs singulières de la matrice X, où $r = \text{rang}(X)$, σ_j^2 sont les valeurs propres non nulles de la

matrice $X'X$ et $\{u_j\}$ et $\{v_j\}$ sont deux familles orthonormales de $\mathbb{R}^n$ et $\mathbb{R}^p$ telles que : $XX'u_j = \sigma_j^2 u_j$ et $X'Xv_j = \sigma_j^2 v_j$.
on a en un premier lieu :

$$\begin{aligned}
X'X &= \left(\sum_{j=1}^{r} \sigma_j u_j v_j'\right)' \left(\sum_{i=1}^{r} \sigma_i u_i v_i'\right) \\
X'X &= \left(\sum_{j=1}^{r} \sigma_j v_j u_j'\right) \left(\sum_{i=1}^{r} \sigma_i u_i v_i'\right) \\
X'X &= \sum_{j=1}^{r}\sum_{i=1}^{r} \sigma_j \sigma_i v_j u_j' u_i v_i' \\
X'X &= \sum_{j=1}^{r}\sum_{i=1}^{r} \sigma_j \sigma_i v_j \delta_{ij} v_i' \\
X'X &= \sum_{j=1}^{r} \sigma_j^2 v_j v_j' \quad (1)
\end{aligned}$$

avec δ_{ij} le symbole de kronecker dans la quatrième ligne car $\{u_j\}$ famille orthonormale. D'après le résultat (1) on peut déduire la formule (2) suivante :

$$(X'X + \lambda Id)^{-1} = \sum_{j=1}^{r} \frac{1}{\sigma_j^2 + \lambda} v_j v_j' \quad (2)$$

Alors pour montrer la formule (2) on peut montrer :

$$(X'X + \lambda Id) \sum_{j=1}^{r} \frac{1}{\sigma_j^2 + \lambda} v_j v_j' = Id$$

On a alors :

$$\begin{aligned}
(X'X + \lambda Id) \sum_{j=1}^{r} \frac{1}{\sigma_j^2 + \lambda} v_j v_j' &= X'X \sum_{j=1}^{r} \frac{1}{\sigma_j^2 + \lambda} v_j v_j' + \lambda Id \sum_{j=1}^{r} \frac{1}{\sigma_j^2 + \lambda} v_j v_j' \\
(X'X + \lambda Id) \sum_{j=1}^{r} \frac{1}{\sigma_j^2 + \lambda} v_j v_j' &= \sum_{i=1}^{r} \sigma_i^2 v_i v_i' \sum_{j=1}^{r} \frac{1}{\sigma_j^2 + \lambda} v_j v_j' + \sum_{j=1}^{r} \frac{\lambda}{\sigma_j^2 + \lambda} v_j v_j' \\
(X'X + \lambda Id) \sum_{j=1}^{r} \frac{1}{\sigma_j^2 + \lambda} v_j v_j' &= \sum_{j=1}^{r}\sum_{i=1}^{r} \frac{\sigma_i^2}{\sigma_j^2 + \lambda} v_i v_i' v_j v_j' + \sum_{j=1}^{r} \frac{\lambda}{\sigma_j^2 + \lambda} v_j v_j' \\
(X'X + \lambda Id) \sum_{j=1}^{r} \frac{1}{\sigma_j^2 + \lambda} v_j v_j' &= \sum_{j=1}^{r}\sum_{i=1}^{r} \frac{\sigma_i^2}{\sigma_j^2 + \lambda} v_i \delta_{ij} v_j' + \sum_{j=1}^{r} \frac{\lambda}{\sigma_j^2 + \lambda} v_j v_j' \\
(X'X + \lambda Id) \sum_{j=1}^{r} \frac{1}{\sigma_j^2 + \lambda} v_j v_j' &= \sum_{j=1}^{r} \frac{\sigma_j^2}{\sigma_j^2 + \lambda} v_j v_j' + \sum_{j=1}^{r} \frac{\lambda}{\sigma_j^2 + \lambda} v_j v_j' \\
(X'X + \lambda Id) \sum_{j=1}^{r} \frac{1}{\sigma_j^2 + \lambda} v_j v_j' &= \sum_{j=1}^{r} \frac{\sigma_j^2 + \lambda}{\sigma_j^2 + \lambda} v_j v_j' \\
(X'X + \lambda Id) \sum_{j=1}^{r} \frac{1}{\sigma_j^2 + \lambda} v_j v_j' &= \sum_{j=1}^{r} v_j v_j'
\end{aligned}$$

Or on a pour $V = \sum_{j=1}^{r} v_j v_j'$, $V^2 = V$:

$$\begin{aligned} V^2 &= \sum_{j=1}^{r} v_j v_j' \sum_{i=1}^{r} v_i v_i' \\ V^2 &= \sum_{j=1}^{r}\sum_{i=1}^{r} v_j v_j' v_i v_i' \\ V^2 &= \sum_{j=1}^{r}\sum_{i=1}^{r} v_j \delta_{ij} v_i' \\ V^2 &= \sum_{j=1}^{r} v_j v_j' \end{aligned}$$

Puisque V^2 est symétrique définie positive on peut dire que la matrice V est inversible donc on obtient :

$$\begin{aligned} V^2 &= V \\ V^2 - V &= 0 \\ V\,(V - Id) &= 0 \end{aligned}$$

or V inversible donc :

$$\begin{aligned} (V - Id) &= 0 \\ V &= Id \end{aligned}$$

Ainsi on obtient :

$$\begin{aligned} (X'X + \lambda Id)\sum_{j=1}^{r} \frac{1}{\sigma_j^2 + \lambda} v_j v_j' &= Id \\ (X'X + \lambda Id)^{-1} &= \sum_{j=1}^{r} \frac{1}{\sigma_j^2 + \lambda} v_j v_j' \end{aligned}$$

on a $A_\lambda = (X'X + \lambda I_p)^{-1} X'$, donc :

$$A_\lambda = \sum_{i=1}^{r} \frac{1}{\sigma_j^2 + \lambda} v_j v_j' \sum_{i=1}^{r} \sigma_j v_j u_j' = \sum_{j=1}^{r}\sum_{k=1}^{r} \frac{\sigma_k}{\sigma_j^2 + \lambda} v_j v_j' v_k u_k'$$

$$A_\lambda = \sum_{j=1}^{r} \frac{\sigma_j}{\sigma_j^2 + \lambda} v_j u_j'$$

donc :

$$\lim_{\lambda \to 0} A_\lambda = \sum_{j=1}^{r} \frac{1}{\sigma_j} v_j u_j'$$

Chapitre 2

Analyse exploratoire

2.1 Traitement et étude initiale des données (Question 1)

On trace le boxplot des données pour obtenir la courbe suivante :

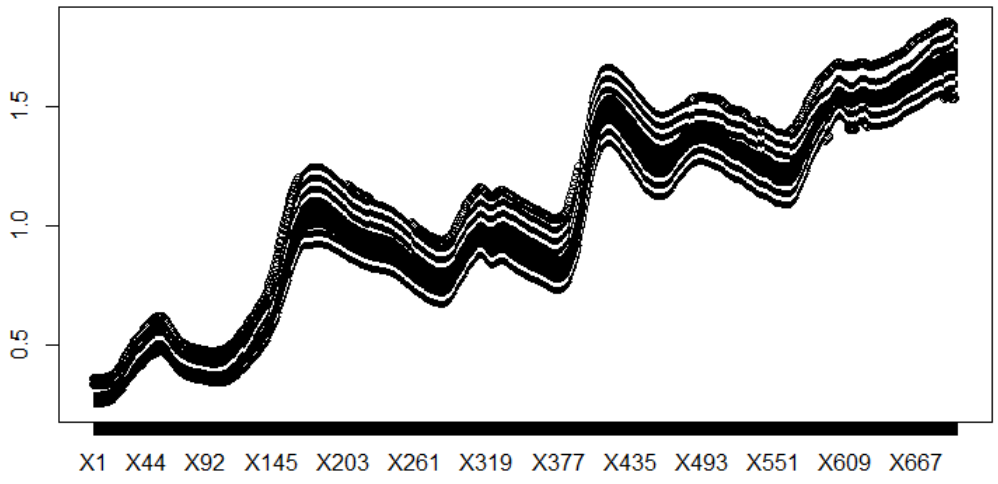


FIGURE 2.1 – boxplot des variables explicatives

Si on fait les boxplots des 200 premières variables explicatives avec chaque 50 variables dans un boxplot séparé :

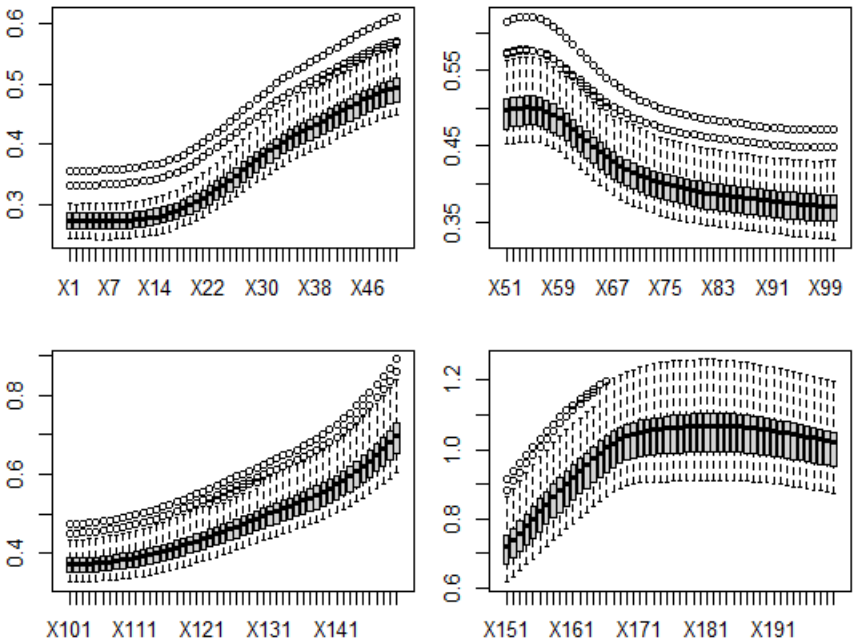


FIGURE 2.2 – les boxplots des variables 1 jusqu'à 200

Pour étudier la corrélation on calcul la matrice de corrélation des données et ses valeurs propres.

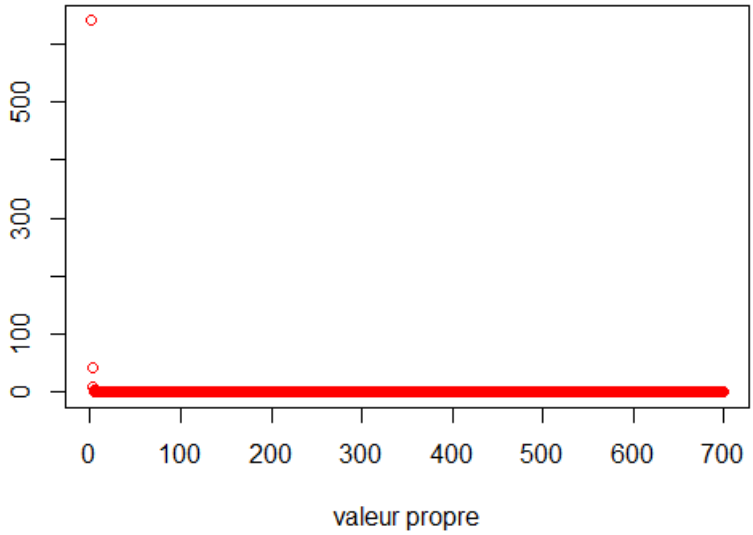


FIGURE 2.3 – valeurs propres de la matrice des corrélations de xtrain

On remarque que la grande majorité des valeurs propres de la matrice des corrélations du jeu d'apprentissage sont très proches de zéro ce qui implique qu'il existe

une multicolinéarité entre les variables explicatives, une chose attendue vu que nous somme dans le cadre de grande dimension $n = 40 < p = 700$.

2.2 Analyse en composantes principales (Question 2)

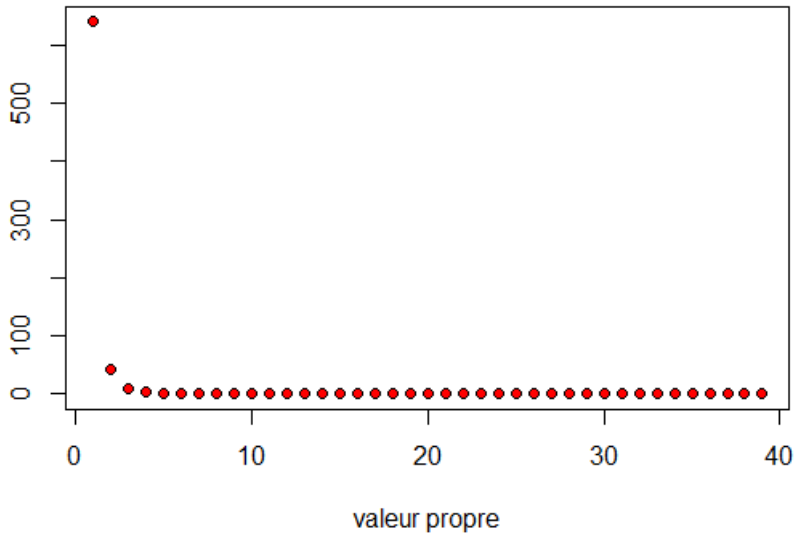


FIGURE 2.4 – valeurs propres de l'ACP

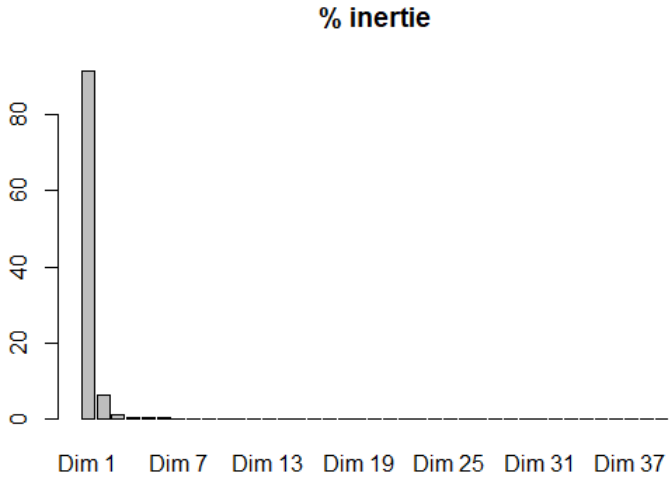


FIGURE 2.5 – pourcentages d'intertie des valeurs propres de l'ACP

Le nombre de valeurs propres est égale à 39 << p. Ceci vient du fait qu'il y a une forte corrélation entre plusieurs variables explicatives.

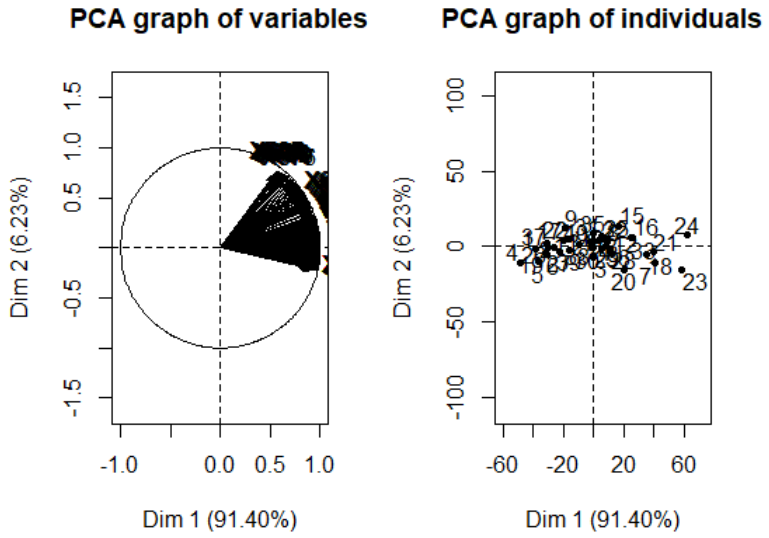


FIGURE 2.6 – premier plan principal de l'ACP

Le premier axe principale représente 91.40% de la variabilité totale du nuage. On voit aussi qu'il existe beaucoup de variables corrélées et que toutes les variables sont bien représentées dans le premier plan principale.

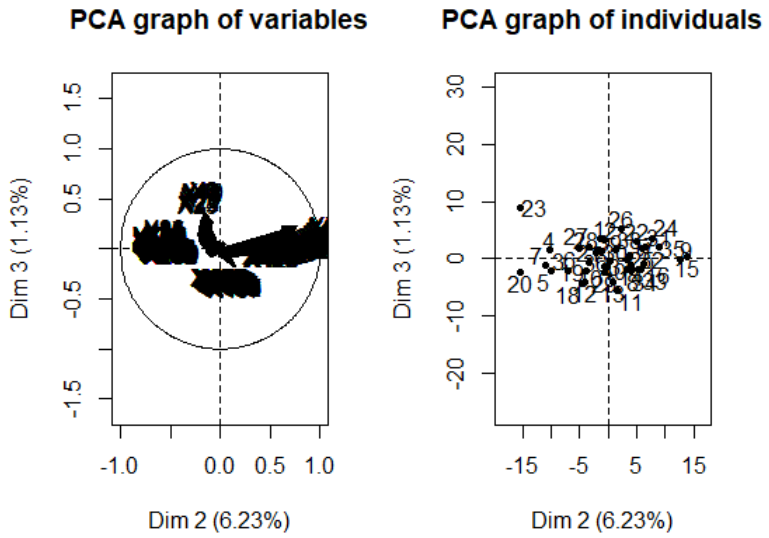


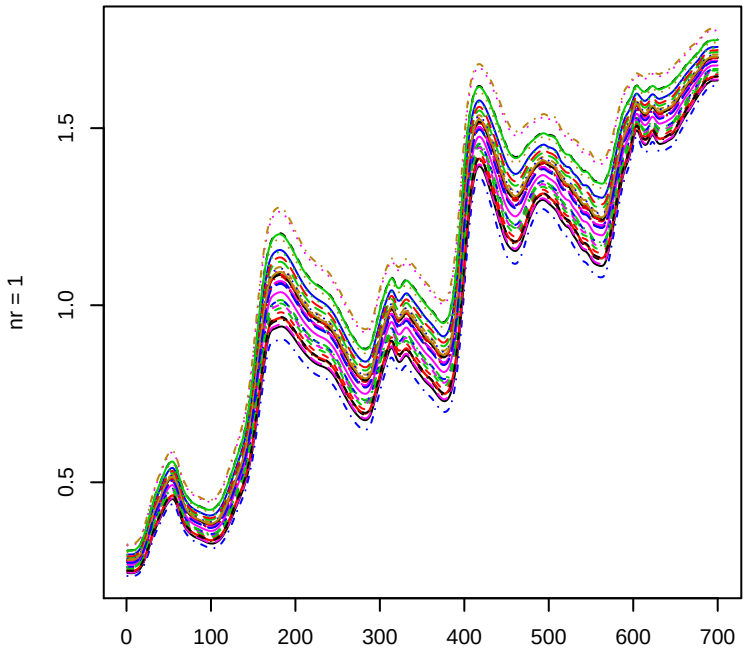
FIGURE 2.7 – deuxième plan principal de l'ACP

Les variables ne sont pas bien représentées sur le deuxième plan principal à cause de la faible inertie des deux axes principaux 2 et 3.

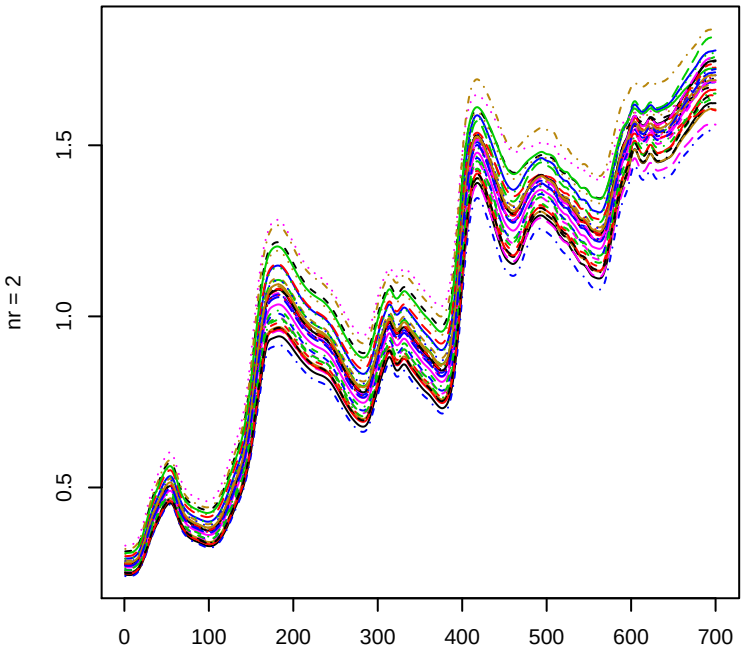
2.3 Reconstruction des données (Question 3)

Pour reconstruire les données à partir des n_r premiers axes de l'ACP on utilise la formule suivante : $X^* = \sum_{i=1}^{n_r} X u_s u_s' = \sum_{i=1}^{n_r} F_s u_s'$ avec F_s la composante principale associée à l'axe u_s. Or l'ACP utilise les données centrées réduites donc la formule ci-dessus nous donne ainsi le nuage reconstruit centré réduit X^*. Pour obtenir le nuage initial on multiplie X^* par l'écart type de chaque colonne de la matrice des données et on ajoute la moyenne de chaque colonne, on obtient ainsi la formule suivante : $X = Xsd \times X^* + Xm$ avec Xsd vecteur des écarts types et Xm vecteur des moyennes.

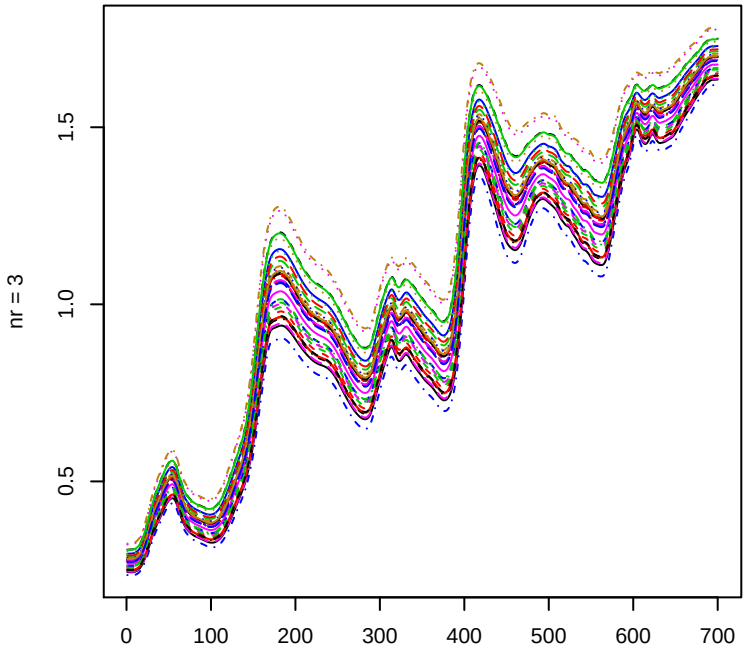
RMSE= 0.01687 MAE= 0.01131



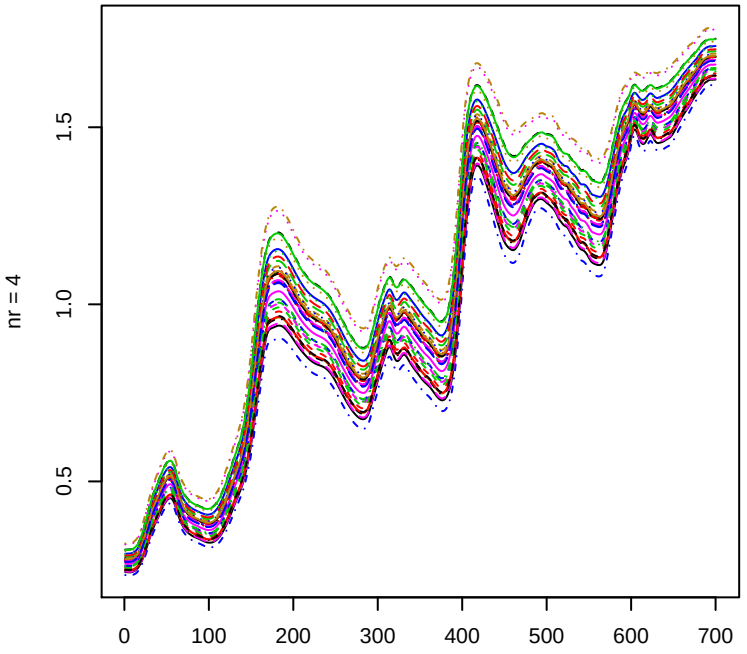
RMSE= 0.00916 MAE= 0.00648



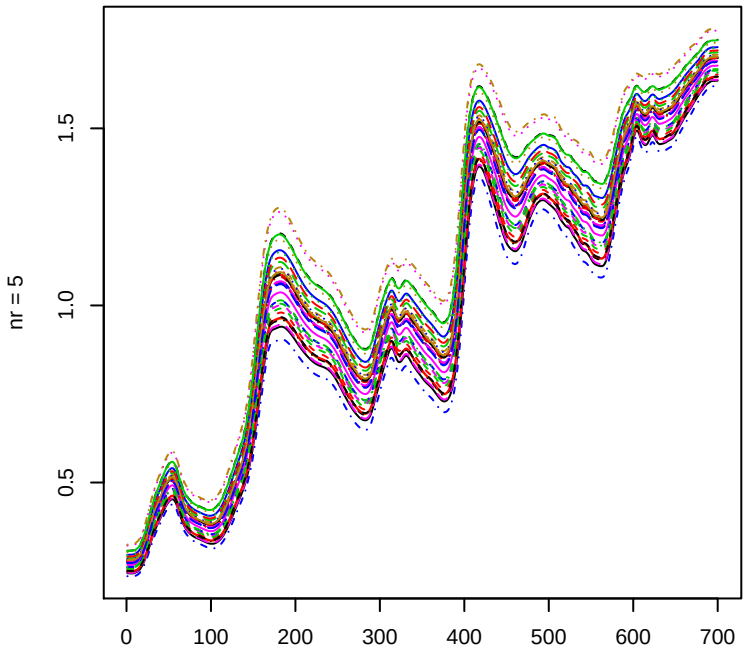
RMSE= 0.00743 MAE= 0.00503



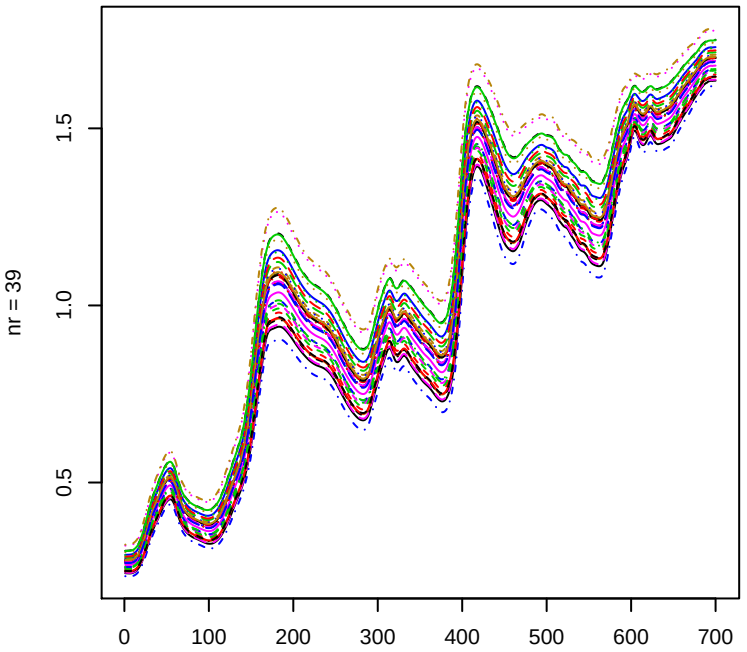
RMSE= 0.00493 MAE= 0.00361



RMSE= 0.00372 MAE= 0.0027



RMSE= 0.00083 MAE= 0.00061



La RMSE (Racine de l'Ecart Quadratique Moyen) et la MAE (Erreur Absolue Moyenne) diminuent lorsque le nombre d'axe principaux utilisés pour la reconstruction du nuage augmente.

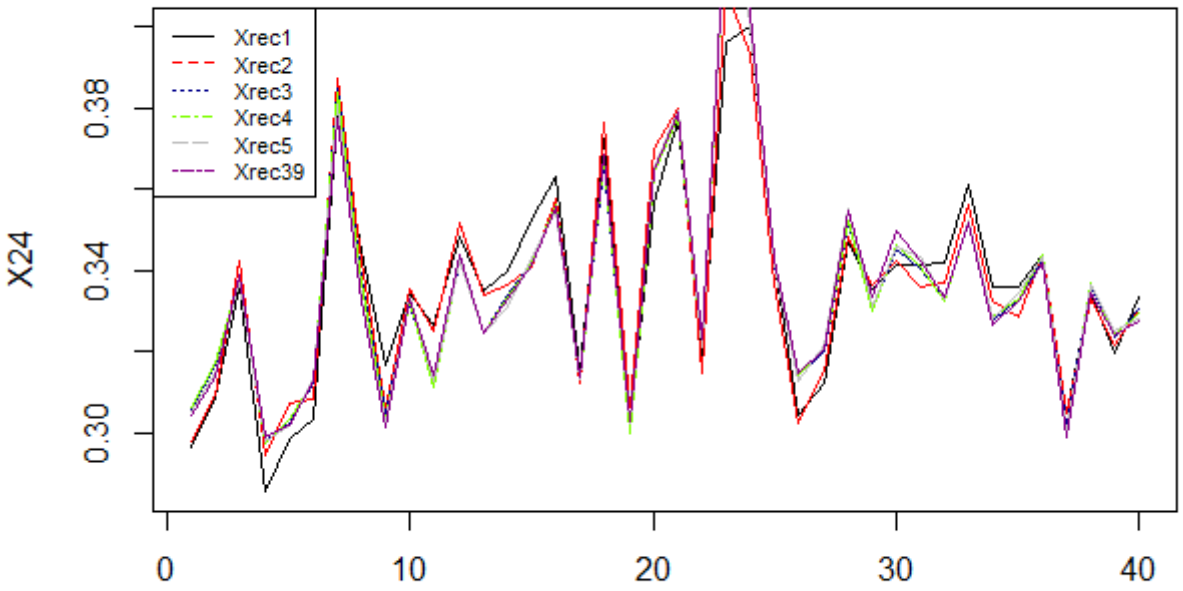


FIGURE 2.8 – courbes de la variable X24 pour chaque niveau

On n'observe pas une grande variation entre les courbes de la variable X_{24} pour les différent niveaux, ceci peut être expliqué par le fait que le premier axe principal représente 91.4% de l'inertie totale du nuage.

Chapitre 3

Régression pénalisée

3.1 Régression ridge et étude de la valeur de l'intercept (Question 1)

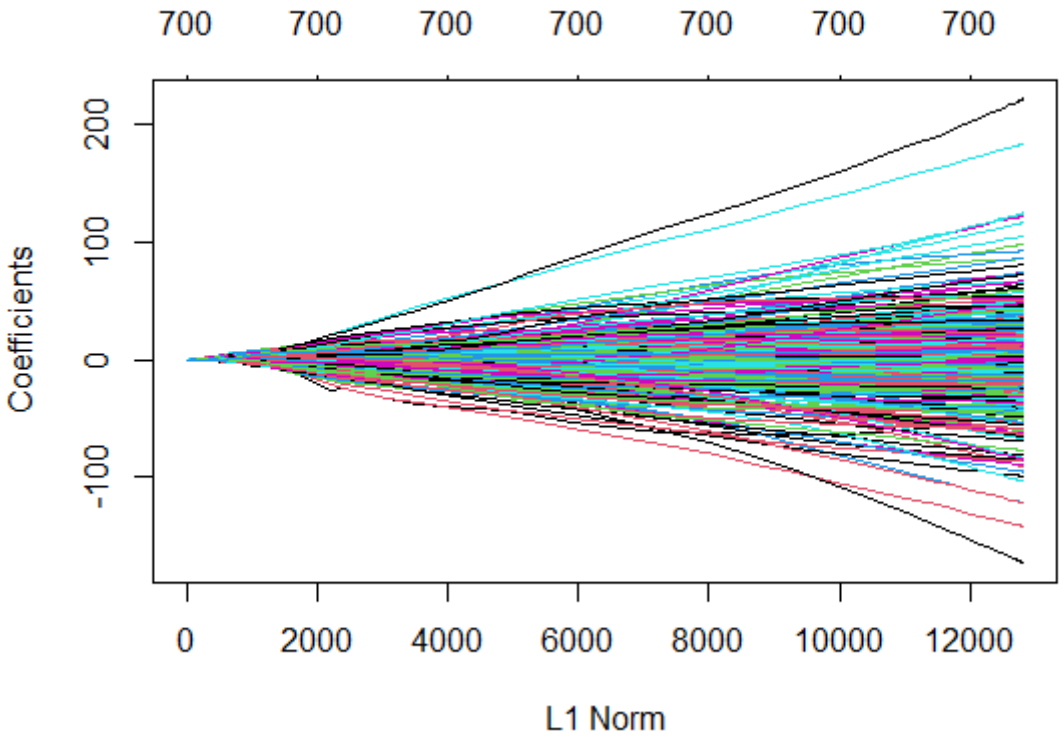


FIGURE 3.1 – les coefficients de θ obtenues par la régression ridge

Il est mentionné dans la vignette écrite par les auteurs de **glmnet** que le plot de la fonction glmnet n'affiche pas les coéfficients en fonction de κ mais plutôt le rapport du coefficient sur la norme l_1 du vecteur de coefficient entier en fonction de κ ce qui explique la forme des courbes.

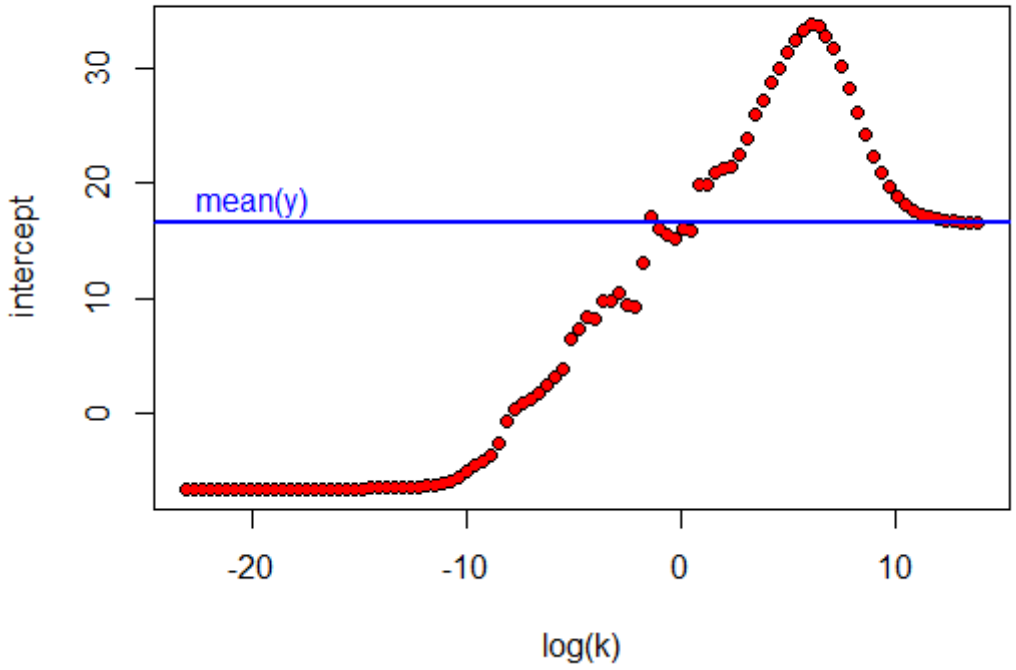


FIGURE 3.2 – l'intercept en fonction de $\log(\kappa)$

On remarque que pour les grandes valeurs de κ la valeur de l'intercept est proche de la valeur moyenne de y. Ceci peut être expliqué en revenant à l'expression de $\hat{\theta}_0$ dans le cas où les données n'ont pas été centrées :

$$\hat{\theta}_0 = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}\theta_j\right)$$

on sait que lorsque la valeur de la pénalisation est très grande les coéfficients θ_j sont très proches de zéro, donc le deuxième terme dans l'expression de $\hat{\theta}_0$ est presque égale à zéro. D'où $\hat{\theta}_0 \approx \overline{y}$

On remarque aussi que pour les petites valeurs de κ l'intercept est négatif, on peut expliquer ceci en exploitant une autre fois l'expréssion de $\hat{\theta}_0$. Lorsque la pénalisation est petite, les coéfficients θ_j peuvent prendre des grandes et donc le deuxième terme dans l'expression de $\hat{\theta}_0$ peut avoir une valeur supérieure à $\overline{y}$, d'où :

$$\hat{\theta}_0 < 0$$

Si on centre **ytrain** seulement, $\hat{\theta}_0$ aura comme expression la suivante :

$$\hat{\theta}_0 = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}\theta_j$$

donc pour les grandes valeurs de pénalisation $\hat{\theta}_0$ sera proche de zéro, et pour les très petites valeurs il sera plus négatif que dans le cas où ytrain n'été pas centrée.

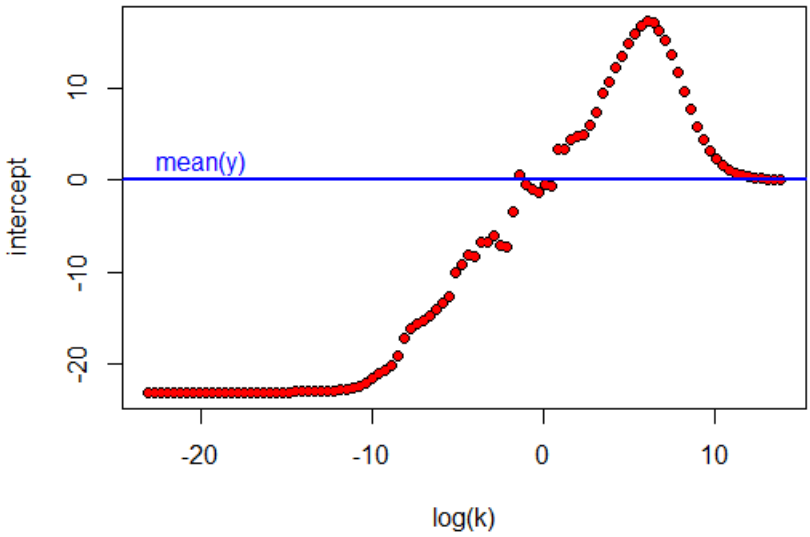


FIGURE 3.3 – l'intercept quand ytrain est centrée

Si on centre **xtrain** seulement on obtient $\hat{\theta}_0 = \overline{y}$ et donc $\hat{\theta}_0$ ne dépend pas de la valeur de κ d'où le graph suivant :

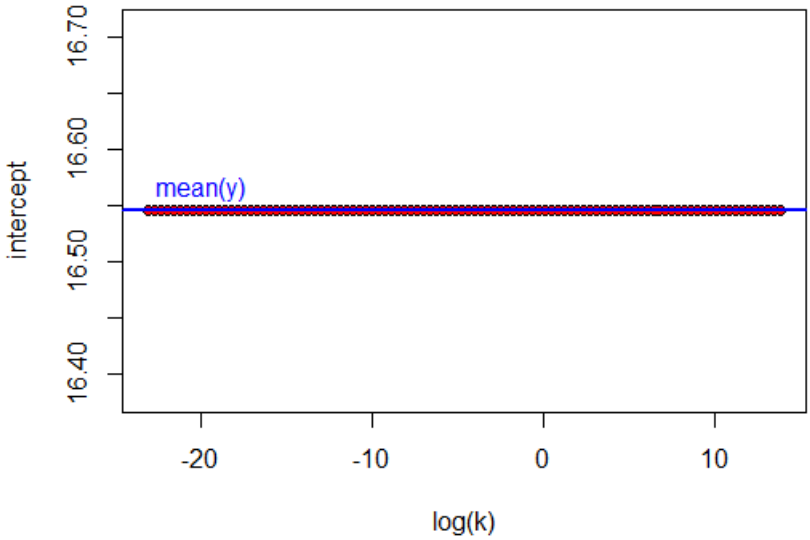


FIGURE 3.4 – l'intercept quand xtrain est centrée

finalement, si on centre **ytrain** et **xtrain**, $\hat{\theta}_0 = 0$ pour toutes les valeurs de κ, mais en pratique on trouve qu'il prend des valeurs très très faibles mais pas égales à zéro pour les très petites valeurs de κ.

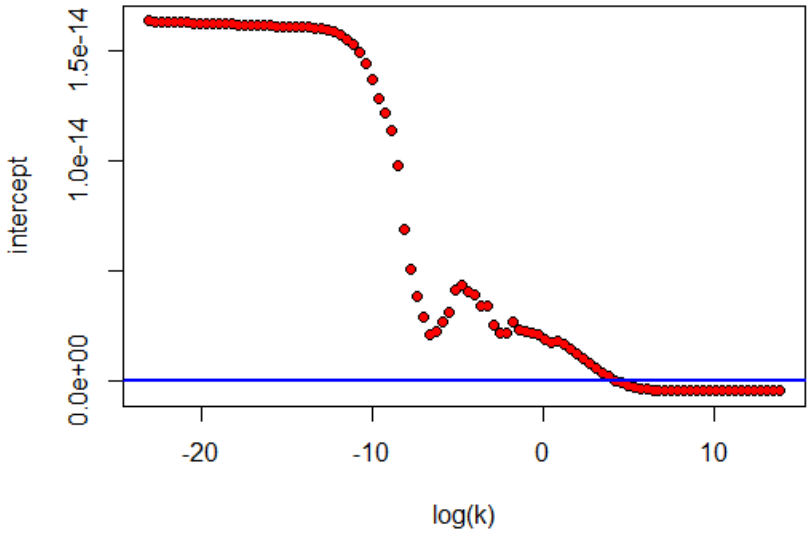


FIGURE 3.5 – l'intercept quand xtrain et ytrain sont centrées

Dans le cas où **ytrain** et **xtrain** on été centrées réduites, $\hat{\theta}_0 = 0$ et $\hat{\theta}_\lambda$ a pour expréssion la suivante :

$$\hat{\theta}_\lambda = \left(((X - \bar{X})\sigma_x)'((X - \bar{X})\sigma_x) + \lambda I_p\right)^{-1} ((X - \bar{X})\sigma_x)'(Y - \bar{Y})/\sigma_y$$

avec $\sigma_x = diag(\frac{1}{\sigma_j})$, σ_j^2 est la variance de la variable j de la matrice X et σ_y^2 est la variance de Y on pose :

$$\tilde{X} = (X - \bar{X})\sigma_x$$

et

$$\tilde{Y} = (Y - \bar{Y})/\sigma_y$$

donc l'expression de θ devient :

$$\tilde{\theta}_\lambda = \left(\tilde{X}'\tilde{X} + \lambda I_p\right)^{-1} \tilde{X}'\tilde{Y}$$

en utilisant les résultats de la question 1.3 de la première partie on a :

$$\lim_{\lambda \to 0} \tilde{\theta}_\lambda = \left(\sum_{j=1}^{r} \frac{1}{\delta_j} v_j u_j'\right) \tilde{Y}$$

où δ_j^2 sont les valeurs propres non nulles de la matrice $\tilde{X}'\tilde{X}$, v_j sont les vecteurs propres de la matrice $\tilde{X}'\tilde{X}$ et u_j sont les vecteurs propres de la matrice $\tilde{X}\tilde{X}'$ et r est le rang de la matrice $\tilde{X}'\tilde{X}$.

3.2 Comparaison de méthode (Question 2)

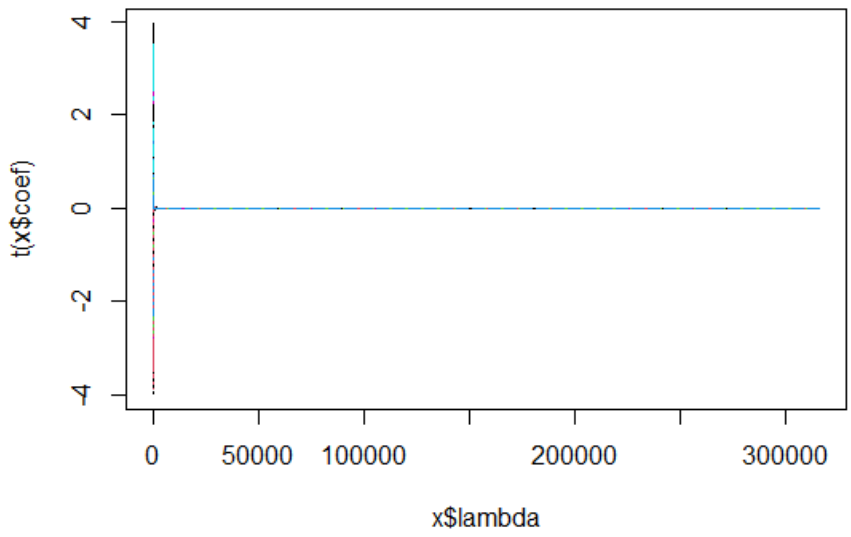


FIGURE 3.6 – les coefficients de θ obtenues par la fonction lm.ridge

On n'obtient pas les mêmes coefficients. et si on fait un zoom sur le graphe donnée par la fonction **lm.ridge**, on voit que les coefficients de θ commence par être non nulles lorsque la valeur de la pénalisation est petite et tendent vers zéro au fur et à mésure que κ augmente. La nature du graphe donné par cette fonction est différente de la nature de celui donné par **glmnet** car ici on trace la valeur du coefficient en fonction de la pénalisation.

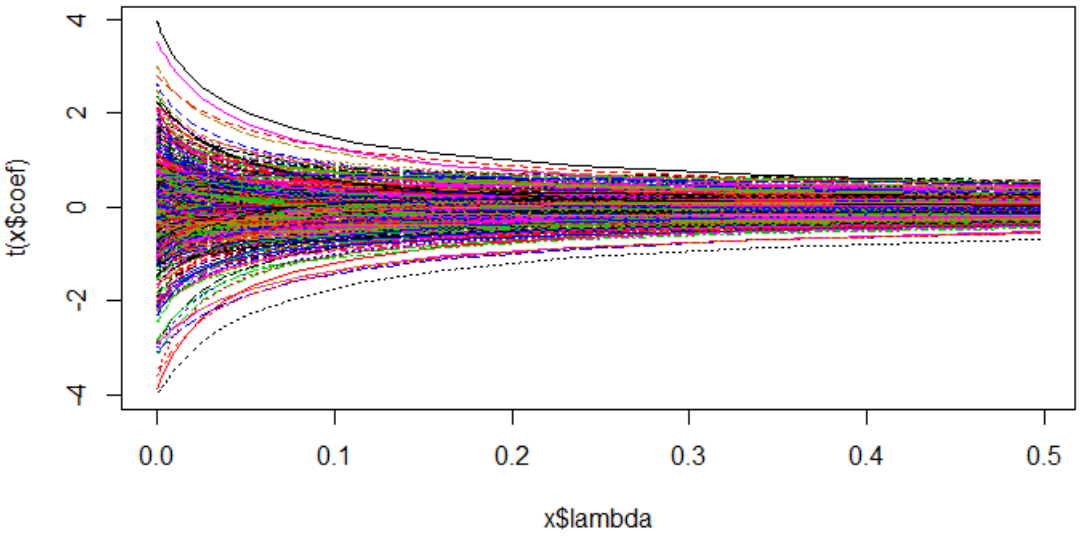


FIGURE 3.7 – zoom sur les coefficients de θ obtenues par la fonction lm.ridge

On a

$$\tilde{\theta} = \left((X - \bar{X})'(X - \bar{X}) + \kappa Id_p\right)^{-1} \left((X - \bar{X})'Y - (X' - \bar{X})\theta_0 1_n\right)$$

Or lorsque X et Y ont été centrées, $\hat{\theta}_0 = 0$ donc

$$\tilde{\theta} = \left((X - \bar{X})'(X - \bar{X}) + \kappa Id_p\right)^{-1} \left((X - \bar{X})'(Y - \overline{Y})\right)$$

et on peut maintenant calculer $\tilde{\theta}$. Or

$$\tilde{\theta} = \left((X - \bar{X})'(X - \bar{X}) + \kappa Id_p\right)^{-1} ((X'X + \kappa Id_p)\theta - X'\overline{Y} - \bar{X}'(Y - \overline{Y}))$$

donc l'estimateur lorsque les données n'ont pas été centrées est donné par l'expression suivante en fonction de $\tilde{\theta}$:

$$\theta = (X'X + \kappa Id_p)^{-1} \left(\left((X - \bar{X})'(X - \bar{X}) + \kappa Id_p\right)\tilde{\theta} + X'\overline{Y} + \bar{X}'(Y - \overline{Y})\right)$$

maintenant en utilisant la formule de la question 2 de la partie théorique on peut calculer θ_0 pour chaque valeur de κ :

$$\hat{\theta}_0 = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}\theta_j\right)$$

après calcul on obtient des coefficients de θ différentes de celles données par les fonction **glmnet** et**lm.ridge**. Pour l'intercept, on obtient la courbe suivante :

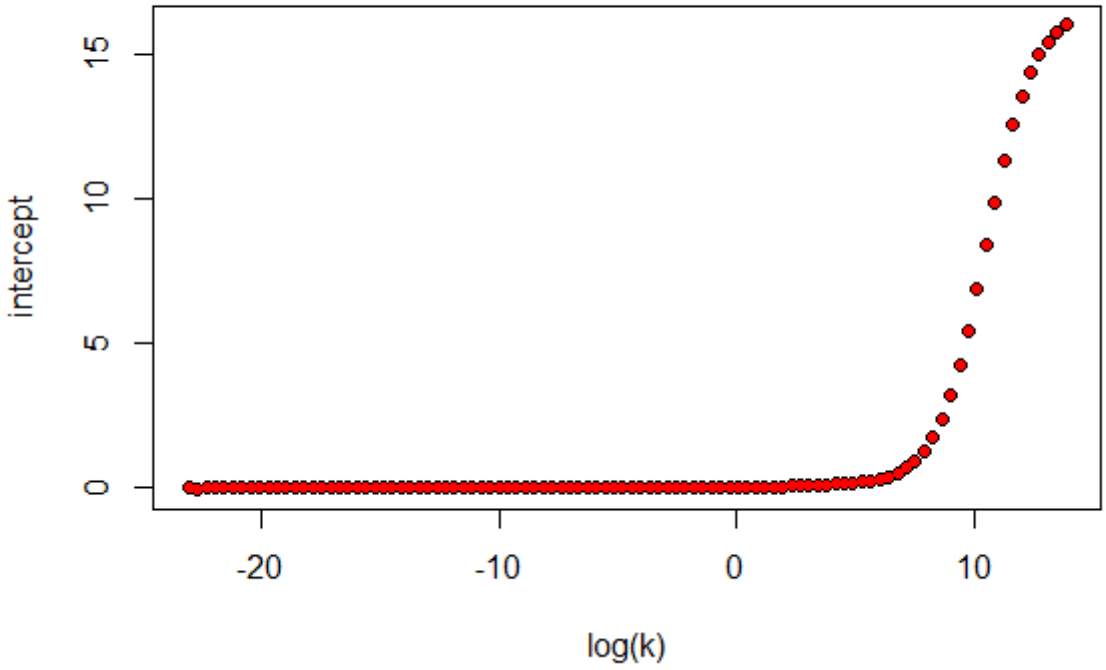


FIGURE 3.8 – intercept calculé manuellement

3.3 Validation Croisée (Question 3)

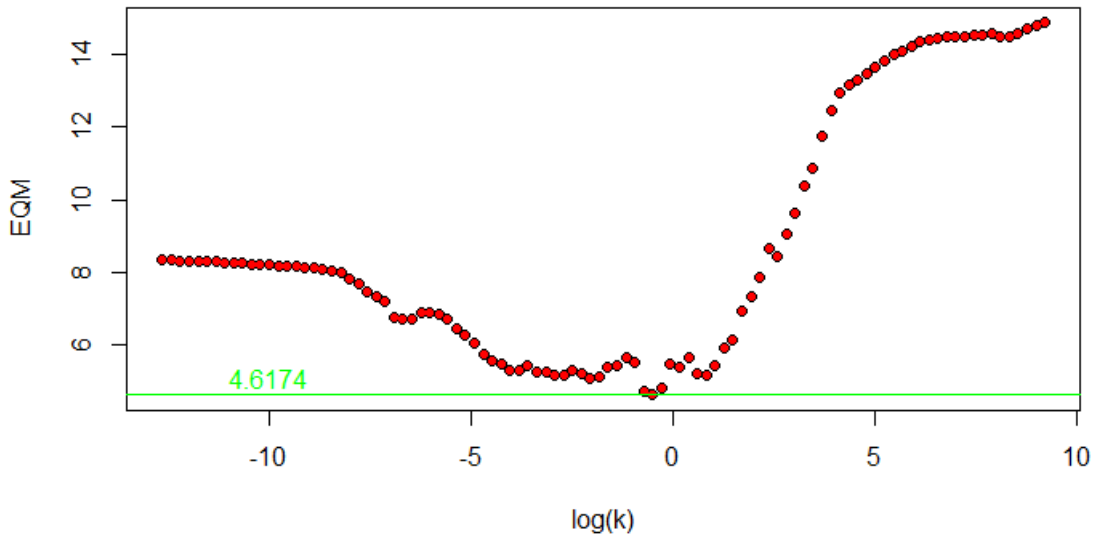


FIGURE 3.9 – l'Erreur quadratique moyenne pour la validation croisée manuelle

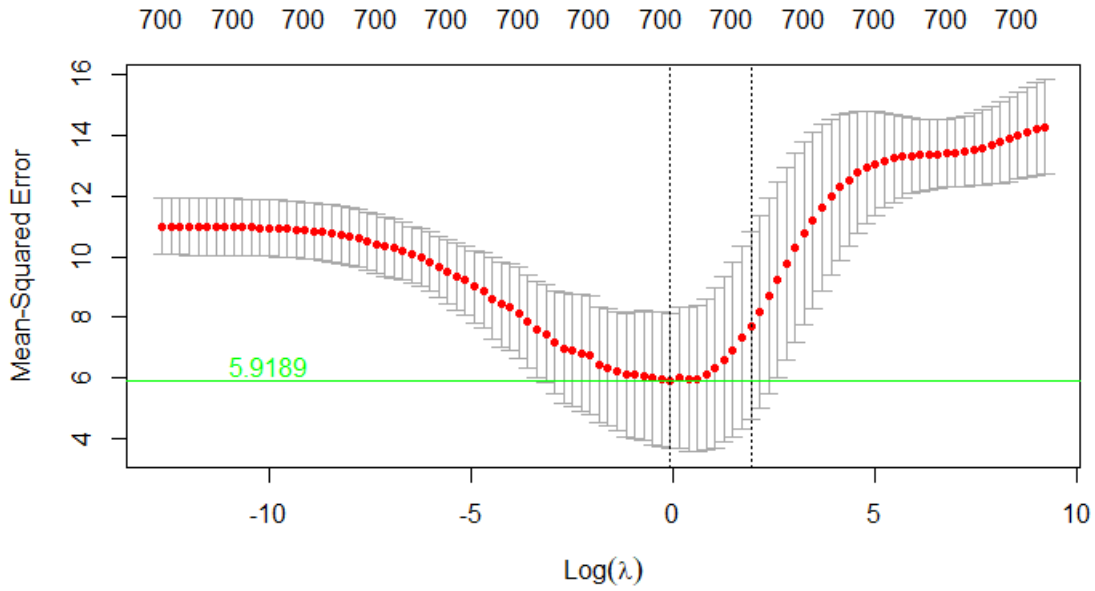


FIGURE 3.10 – l'EQM des κ donnée par la fonction **cv.glmnet**

L'EQM donnée par la validation croisée que nous avons implémenté est inférieure à celle donnée par la fonction **cv.glmnet**. On va donc l'utiliser pour faire

l'apprentissage sur la totalité du jeu de données.

Chapitre 4

Régression logistique pénalisée

Dans cette partie de notre étude ce n'est pas la teneur en sucre qu'on va étudier mais plutôt le fait qu'elle dépasse le seuil de 18 ou pas.

4.1 Hypothèses de la régression logistique (Question 1)

les hypothèses de la régression logistique sont :

— le modèle des variables est un modèle de variables indépendantes de loi binomiale

— la variable réponse est à deux niveaux (0 ou 1) : $Y_i \sim \mathcal{B}(1, p(X_i))$

— l'espérance des variables explicatives est une fonction non linéaire d'un régresseur linéaire : $logit(p(X_i)) = X_i\theta$

Le jeu d'apprentissage et le jeu test ne sont pas équilibrés.En effet, un jeu de données équilibré est un jeu de données qui contient autant de positives que négatives donc autant de 0 que des 1 dans notre cas. Or le jeu d'apprentissage, qui contient 40 observations, a seulement 16 observations qui valent 1 et le jeu de données test, qui contient 32 observations, contient seulement 13 valent 1.

4.2 Courbes ROC et adéquation (Questions 2&3)

Dans cette partie on élabore deux régression logistique pénalisées, une régression logistique pénalisée en ridge et l'autre en lasso. après calcul on obtient les courbes suivantes :

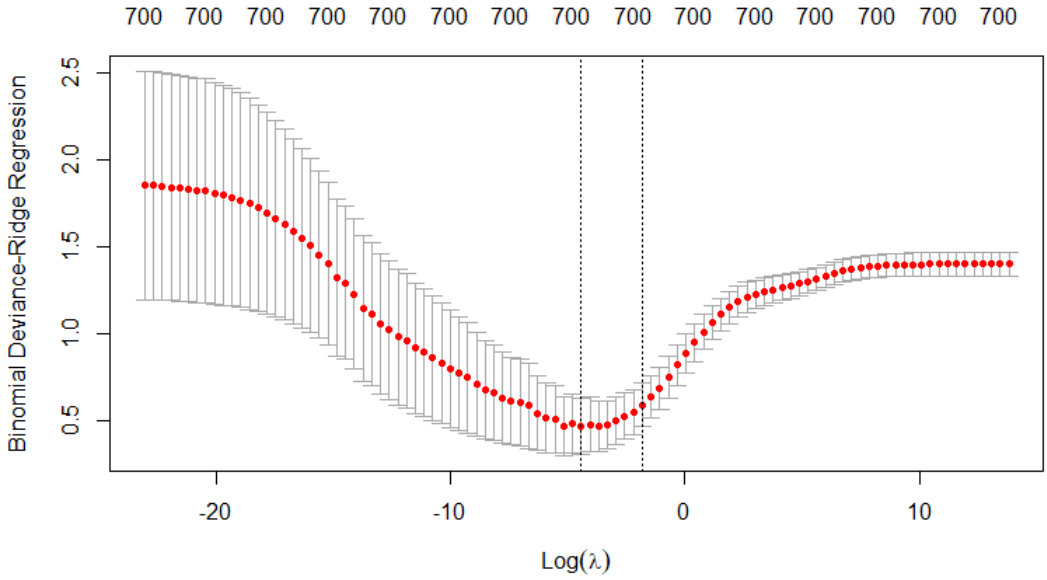


FIGURE 4.1 – Régression logistique Ridge

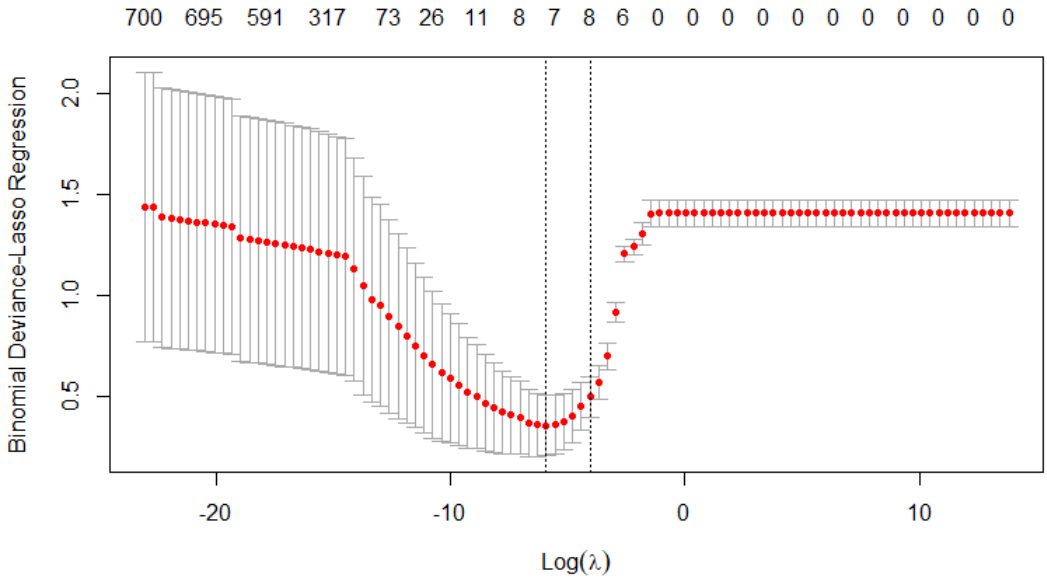


FIGURE 4.2 – Régression logistique Lasso

D'après les courbes on remarque que les valeurs de λ_{min} sont différents pour les deux types de régression. En effet pour la régression ridge $\lambda_{min}^{ridge} = 0.01205$ et pour la régression lasso $\lambda_{min}^{lasso} = 0.002719$.
Les courbes ROC (Receiver Operating Curve) sont des graphiques représentants les performances d'un modèle de classification pour tous les seuils de classification.

On tracera les courbes ROC pour les deux modèles de régression et on obtient les courbes suivantes :

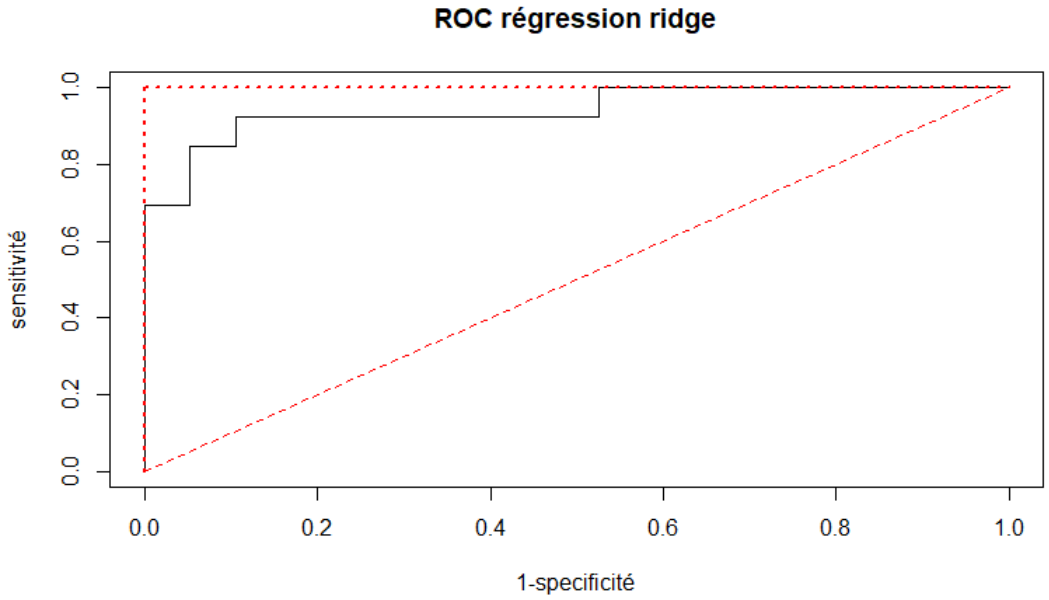


FIGURE 4.3 – ROC régression logistique Ridge

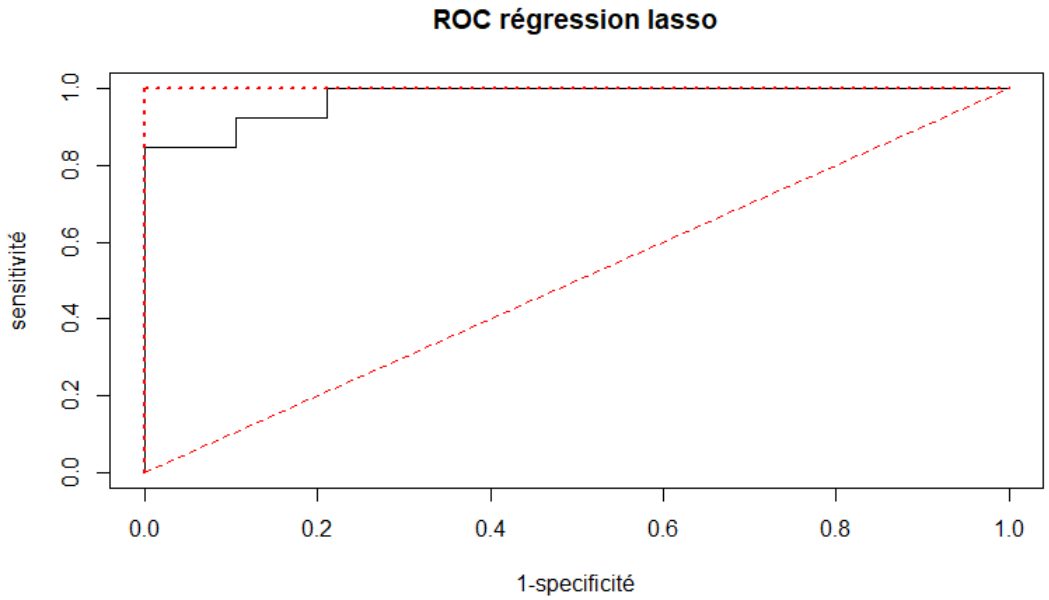


FIGURE 4.4 – ROC régression logistique Lasso

On peut tester l'adéquation des deux modèles en calculant l'AUC qui fournit une mesure de performance. En un premier lieu on utilise le prédicteur de Bayes

pour faire des prédiction en utilisant nos modèles puis on calcul la moyenne de la conformité des prédictions avec les données test et on obtient :

— 1.25% de mauvaises réponses pour le modèle ridge
— 0.625% de mauvaises réponses pour le modèle lasso

Conclusion

Lors de notre étude on a pu développer plusieurs modèles statistiques : un modèle de régression ridge et deux régression logistique pénalisée en ridge et en lasso. Notre objectif initial était le control de la teneur en sucre à partir de la spectroscopie infrarouge des biscuits ou du mélange de préparation des biscuits, donc on peut modéliser ce control par la détermination du taux du sucre directement à partir du spectre (les variables explicatives) et on a obtenu des erreurs relativement petites dans notre modélisation ou bien on peut modéliser le control par le fait que la teneur dépasse un certain taux ou pas ceci en effectuant la régression pénalisée et le modèle lasso nous donne une erreur de mauvais choix de 0.625% qui est un taux assez petit donc on a obtenu une grande performance pour notre modèle.

Annexe A

Analyse exploratoire

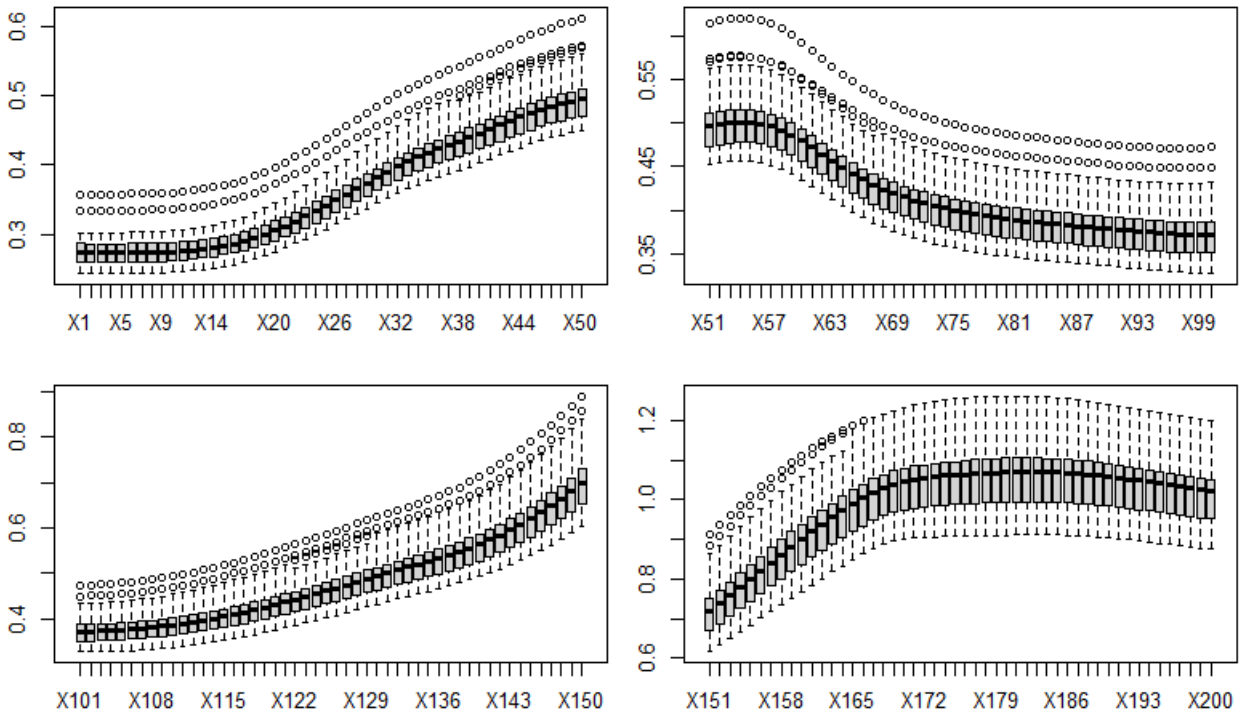


FIGURE A.1 – les boxplots des variables 1 jusqu'à 200

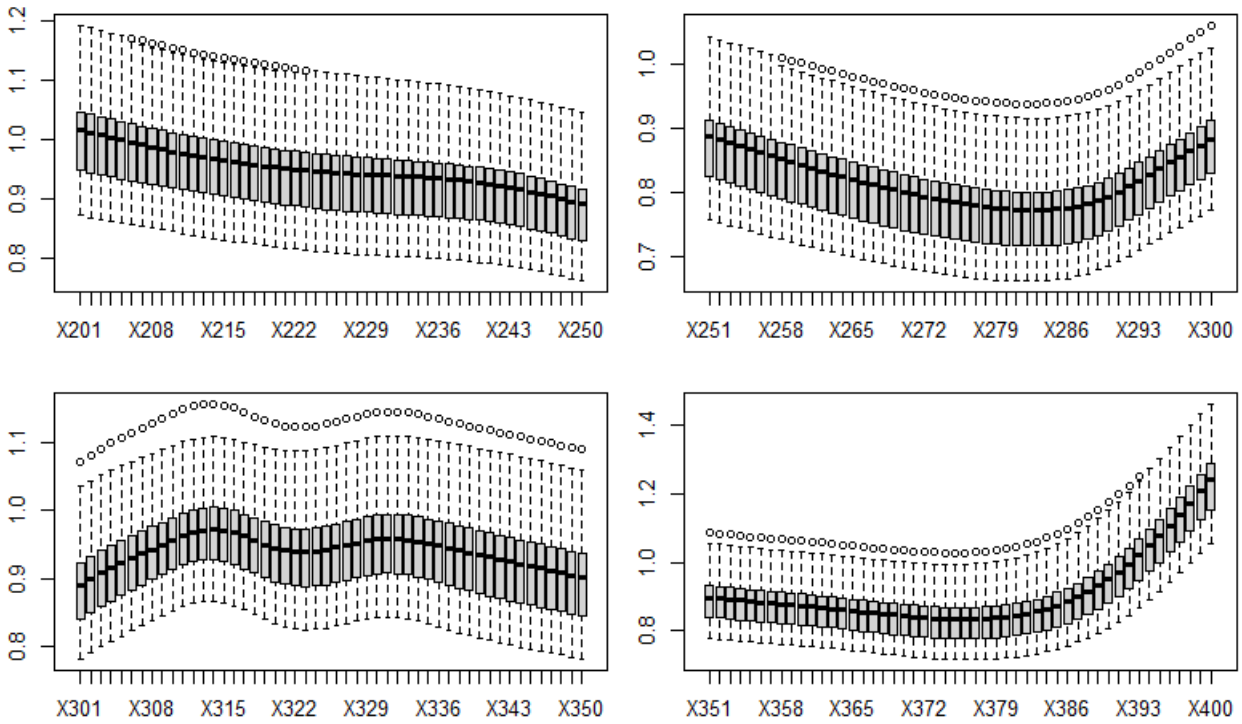


FIGURE A.2 – les boxplots des variables 201 jusqu'à 400

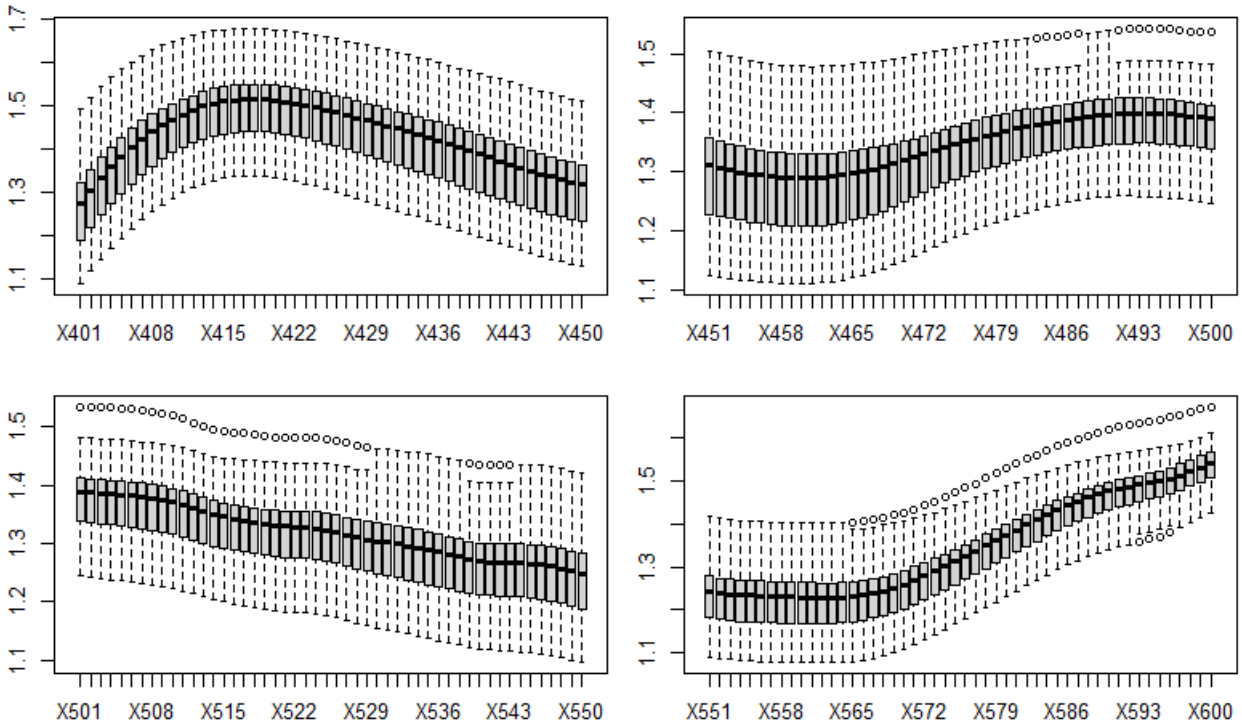


FIGURE A.3 – les boxplots des variables 401 jusqu'à 600

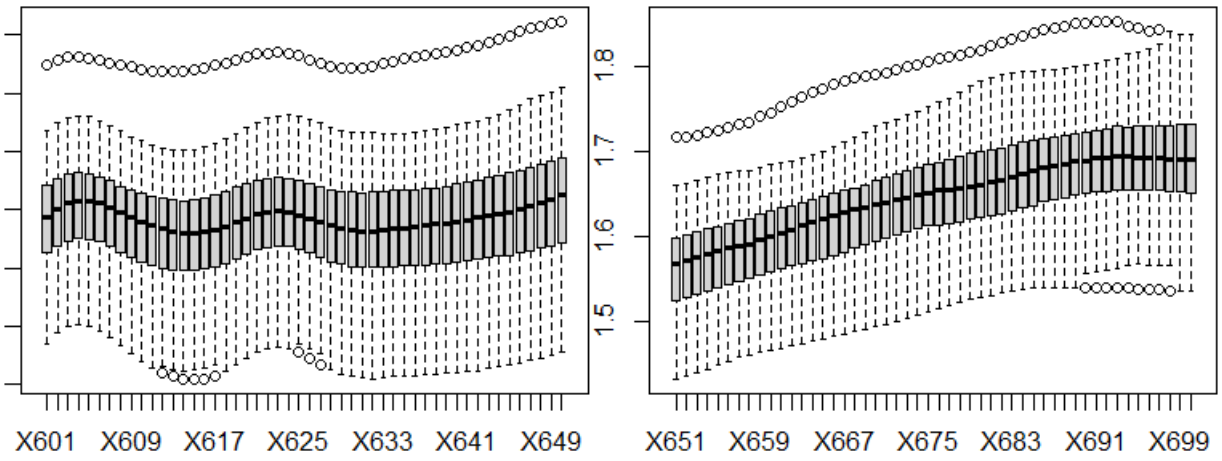


FIGURE A.4 – les boxplots des variables 601 jusqu'à 700

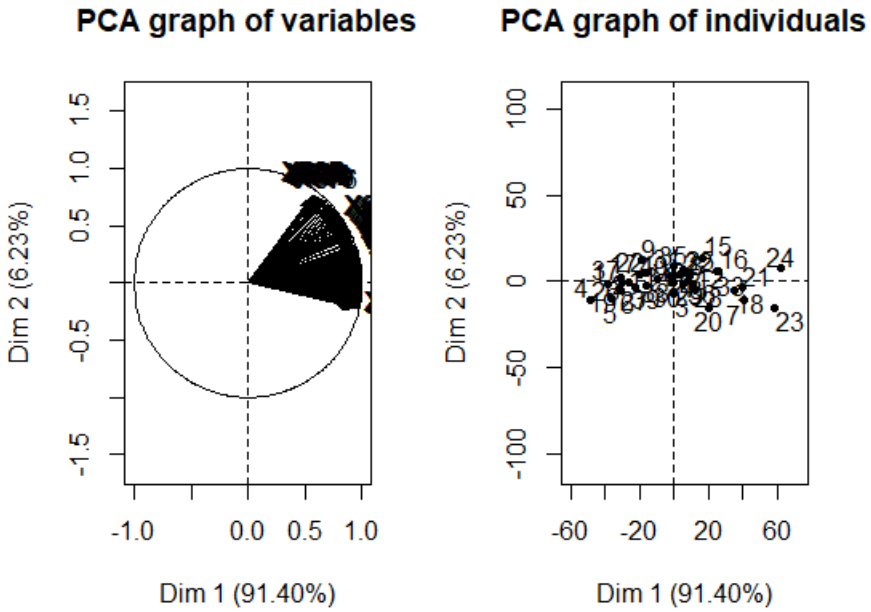


FIGURE A.5 – Premier plan principal de l'ACP

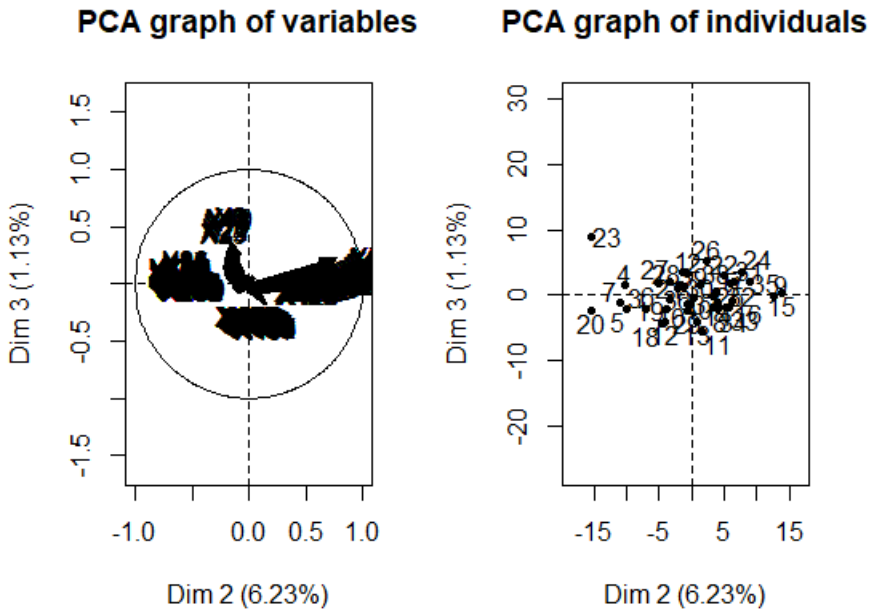


FIGURE A.6 – Deuxième plan principal de l'ACP

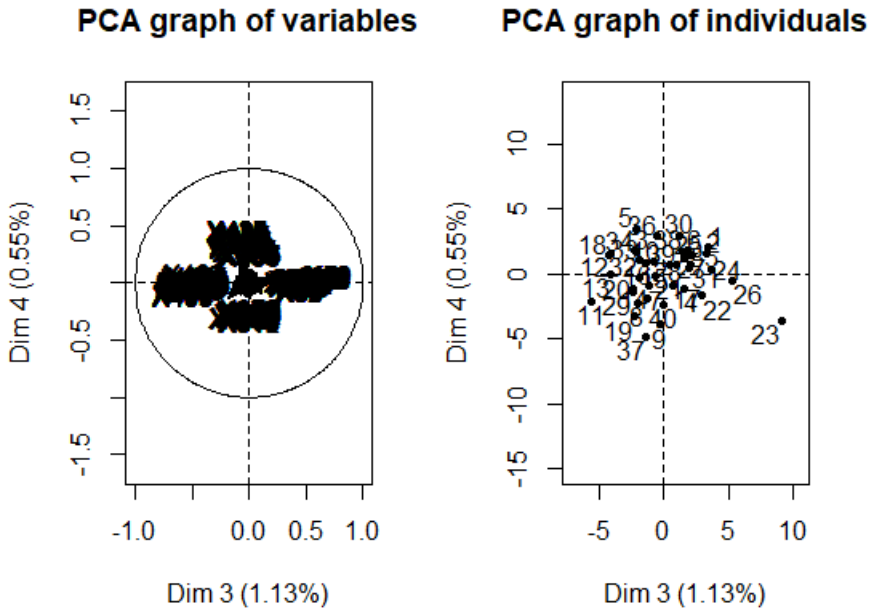


FIGURE A.7 – Troisième plan principal de l'ACP

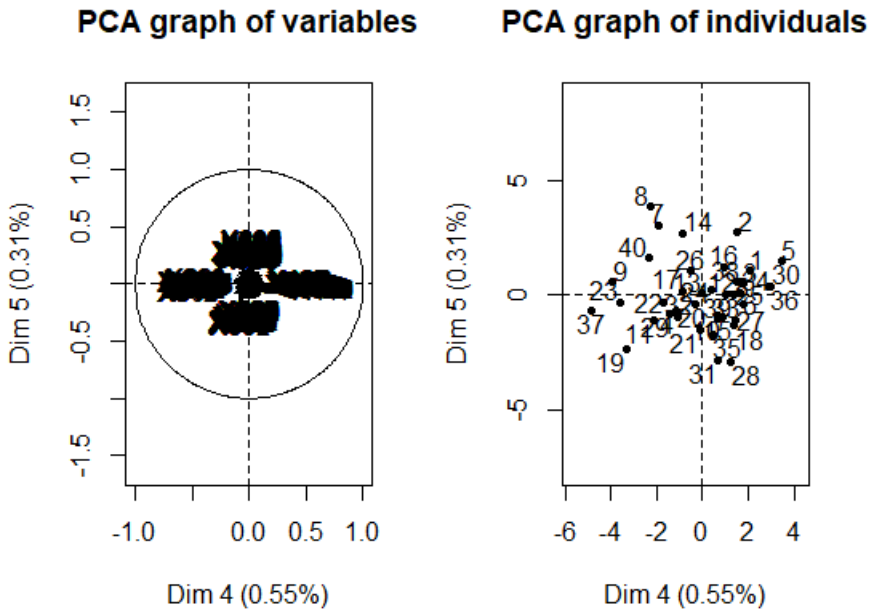


FIGURE A.8 – Quatrième plan principal de l'ACP

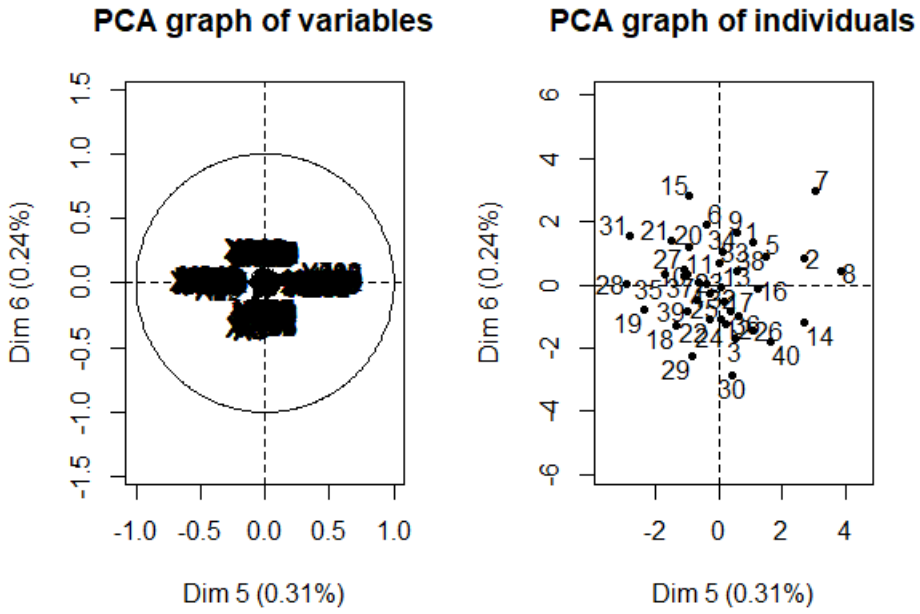


FIGURE A.9 – Cinquième plan principal de l'ACP