UNIVERSITÉ DE NANTES

Université de Nantes – UFR des Sciences et Techniques
Département de Mathématiques

Master 2 Ingénierie mathématique **2015-2016**

Travaux pratiques de Régression

Ex 1. *Lien entre 2 variables quantitatives*
Galilée a réalisé diverses expériences afin de comprendre la trajectoire des boulets de canon. Dans l'une d'entre elles, une balle est lâchée sur un plan incliné et quitte le plan incliné à une certaine hauteur du sol, on mesure ensuite la distance parcourue par la balle. Le tableau suivant résume les données mesurées.

Hauteur	1000	800	600	450	300	200	100
Distance	573	534	495	451	395	337	253

1) Effectuer une représentation graphique de ces données. Un lien apparait-il entre les deux variables ? Semble-t-il linéaire ?

2) Calculer le coefficient de corrélation linéaire entre les deux variables. Est-il significativement non nul ?

3) Même question pour les coefficient de corrélation de Spearman.

4) Quel coefficient quantifie le plus fidèlement le lien observé à la première question ?

Ex 2. *Lien entre 2 variables quantitatives*
Le fichier "lifeexp-TV.dat" contient l'espérance de vie dans différents pays ainsi que le nombre moyen de TV par habitants.

1) Effectuer une représentation graphique des données. Un lien apparait-il entre les deux variables ? Semble-t-il linéaire ?

2) Calculer le coefficient de corrélation linéaire entre les deux variables. Est-il significativement non nul ? Même question avec le coefficient de corrélation de Spearman et de Kendall.

3) Suggérer une transformation des variables rendant le lien "plus" linéaire. Vérifier votre démarche à l'aide d'une représentation graphique et du calcul de la corrélation linéaire.

4) Commenter l'affirmation suivante : "Pour augmenter l'espérance de vie des habitants, il suffit de leur fournir plus de TV".

Ex 3. *Lien entre 2 variables quantitatives*
Le fichier "motordata.txt" contient des enregistrements de l'accélération (en g) de la tête d'un motard après un choc en fonction du temps (en millisecondes).

1) Effectuer une représentation graphique de ces données. Un lien vous semble-t-il apparaître entre l'accélération et le temps écoulé après le choc ?

2) Calculer les coefficients de corrélation de Pearson, Spearman et Kendall, et tester leur significativité. Commenter.

Ex 4. *Lien entre 2 variables qualitatives*
La répartition des jeunes âgés de 16 à 25 ans sans diplômes et résidant en Bretagne en 1999 selon leur sexe et leur activité est la suivante (d'après le recensement INSEE) :

Type d'activité	Hommes	Femmes	Total
Actifs ayant un emploi	6 639	2 446	9 085
Chômeurs	2 544	1 850	4 394
Inactifs	1 201	1 872	3 073

1) Saisir ces données sous R et en proposer une représentation graphique synthétique.

2) D'après ces données, peut-on affirmer qu'il y a une dépendance significative entre le type d'activité et le sexe en Bretagne ?

Ex 5. *Lien entre 2 variables qualitatives*
Un site internet reçoit 113 457 visiteurs durant un mois. On désigne par X le navigateur internet utilisé et Y le système d'exploitation utilisé.

X/Y	Windows	Mac	Linux
Chrome	14103	1186	427
Firefox	30853	4392	3234
Internet Explorer	47389	23	0
Safari	668	6416	0
Autres	2974	40	1752

1) Saisir ces données sous R et en proposer une représentation graphique synthétique.

2) D'après ces données, peut-on affirmer qu'il y a une dépendance significative entre le navigateur et le système d'exploitation utilisés ?

Ex 6. *Lien variable quantitative/variable qualitative*
Le fichier "NO2_trafic.csv" contient différentes mesures de dioxyde d'azote (NO2) effectuées au sein de véhicules circulant dans la métropole parisienne. La variable "type" donne le type de routes principalement empruntées ("A" : Autoroute, "P" : Périurbain, "T" : Tunnel, "U" : Urbain, "V" : Voie rapide urbaine) et la variable "fluidite" donne les conditions de trafic (de "A" : fluide, à "D" : congestionné).

1) Proposer une représentation graphique de la variable "NO2" en fonction du type de routes empruntées.

2) Donner la moyenne du NO2 selon le type de routes empruntées et tester leur égalité.

3) Reprendre les deux questions précédentes pour étudier le lien entre la variable "NO2" et la variable "fluidite".

4) Constuire une nouvelle variable de type facteur contenant uniquement la modalité "D" (congestionné) de la variable "fluidite", toutes les autres étant regroupées sous le label "Autre".

5) Y a-t-il une différence significative de la concentration moyenne de NO2 lorsque le trafic est congestionné par rapport aux autres conditions de circulation ? On pourra utiliser un test d'égalité des moyennes à alternative bilatéral et/ou unilatéral.

Ex 7. On veut expliquer la hauteur des eucalyptus en fonction de leur circonférence à partir d'une régression linéaire simple. On dispose des mesures des hauteurs (ht) et circonférences (circ) de 1737 eucalyptus, qui se trouvent dans le fichier "eucalyptus.txt".

1. Extraire et représenter les données dans le plan.
2. Effectuer la régression $y = \beta_0 + \beta_1 x + \epsilon$ où y représente la hauteur et x la circonférence. Commenter les résultats.
3. Calculer un intervalle de confiance à 95% pour β_0 et β_1, en supposant la normalité des données.
4. Si le bruit ϵ ne suit pas une loi normale, les intervalles de confiance précédents restent-ils valables ?
5. Tracer l'estimateur de la droite de régression et un intervalle de confiance à 95% de celle-ci. Que déduisez-vous de la qualité de l'estimation ?
6. On veut à présent prédire la hauteur d'une nouvelle série d'eucalyptus de circonférences 50, 100, 150 et 200. Donner les estimateurs de la taille de chacun d'entre eux et les intervalles de prédiction à 95% associés, en supposant la normalité des données.
7. Si le bruit ϵ ne suit pas une loi normale, les intervalles de prédiction précédents restent-ils valables ?

Ex 8. *Convergence des estimateurs*

I) Lors d'une expérience chimique, on observe la teneur d'un certain produit à différents instants réguliers allant de 1 à n. Le résultat à l'instant i est noté y_i. On suppose le lien temporel suivant : $y_i = \beta_0 + \beta_1 i + \epsilon_i$, $i = 1, \ldots, n$, où les variables ϵ_i représentent les erreurs de mesures. On suppose que $\beta_0 = \beta_1 = 1$ et que les ϵ_i sont i.i.d suivant une loi $\mathcal{N}(0, 20^2)$.

1) Simuler 1000 valeurs de y_i pour i variant de 1 à 1000. Observer le nuage de points entre y et les instants de mesures. Pour k allant de 10 à 1000, effectuer la régression des k premières valeurs de y par rapport aux k premiers instants i. Observer graphiquement l'évolution de l'estimation de β_1 en fonction de k.

2) Répéter la simulation précédente 20 fois et superposer sur un même graphique l'évolution des 20 estimateurs de la pente de la régression. Commenter.

II) On suppose à présent que le lien temporel est $y_i = \beta_0 + \beta_1 \frac{1}{i} + \epsilon_i$ où $\beta_0 = \beta_1 = 1$ et les ϵ_i sont i.i.d suivant une loi $\mathcal{N}(0, 0.1^2)$. Effectuer le même type de simulations que dans la première partie et observer le comportement asymptotique de l'estimateur de β_1.

Ex 9. On tire aléatoirement $n = 50$ réalisations indépendantes x_i d'une variable aléatoire de loi normale centrée de variance 20. Le vecteur x va jouer le rôle de la variable explicative dans un modèle de régression et sera traité comme déterministe tout au long de l'exercice. On considère le modèle

$$y_i = 0.5 + 3x_i + \epsilon_i, i = 1, ..., n,$$

où les ϵ_i sont des variables aléatoires iid qui représentent le bruit. L'idée est de simuler $N = 10000$ fois ce modèle (sans changer la valeur de x) de manière à pouvoir estimer par la méthode de Monte-Carlo la loi des différents estimateurs et statistiques de test liées à ce modèle.

I) Cas Gaussien : on prend les ϵ_i iid de loi normale $\mathcal{N}(0, \sigma^2 = 1/3)$.

1. Tracer l'histogramme des N valeurs de $\hat{\beta}_0$ et le comparer avec sa loi théorique. Faire la même chose pour $\hat{\beta}_1$ et $\hat{\sigma}^2$.
2. Discuter au moyen d'un test statistique la corrélation entre $\hat{\sigma}^2$ et $\hat{\beta}$.
3. On considère maintenant le test de Student de $H_0 : \beta_1 = 6\beta_0$ contre $H_1 : \beta_1 \neq 6\beta_0$ au niveau $\alpha = 10\%$. A partir des N résultats du test obtenus, proposer un estimateur de α. Tracer l'histogramme des valeurs de la statistique de Student et la comparer avec la densité théorique de la statistique. Faire de même avec la statistique de Fisher. Commenter.
4. Comparer tous ces résultats avec ceux obtenus sur un échantillon de taille $n' = 5$ où on ne prend que les 5 premières valeurs de x_i.

II) Cas non Gaussien : on prend maintenant les ϵ_i iid de loi uniforme sur $[-1; 1]$ (donc également de variance $\sigma^2 = 1/3$). On cherche à vérifier la validité des tests de Student et Fisher dans le modèle avec bruits uniformes.

1. Simuler le même modèle que précédemment avec un bruit uniforme et répéter l'expérience $N = 10000$ fois.
2. Tracer l'histogramme des valeurs de $\hat{\beta}_0$ et le comparer avec la densité de la loi normale d'égales moyenne et variance. Faire la même chose pour $\hat{\beta}_1$ et $\hat{\sigma}^2$.
3. Discuter comme précédemment la corrélation entre $\hat{\sigma}^2$ et $\hat{\beta}$.
4. Estimer la densité de la statistique de Student associée au test de l'hypothèse $H_0 : \beta_1 = 6\beta_0$ contre $H_1 : \beta_1 \neq 6\beta_0$ au niveau $\alpha = 10\%$. Comparer avec la proportion de rejets de H_0.
5. Les conclusions sont-elles les mêmes si on reproduit la procédure sur un échantillon de taille $n' = 5$?

Ex 10. *Consommation de glaces*

On étudie la consommation de glaces aux Etats-Unis sur une période de 30 semaines du 18 Mars 1950 au 11 Juillet 1953. Les variables sont la période (de la semaine 1 à la semaine 30), et en moyenne sur chaque période : la consommation de glaces par personne ("Consumption", en 1/2 litre), le prix des glaces ("Price", en dollars), le salaire hebdomadaire moyen par ménage ("Income", en dollars), et la température ("Temp", en degré Fahrenheit). Les données sont disponibles dans le fichier "icecream-R.dat".

1) Extraire les données et représenter la consommation en fonction des différentes variables.

2) On propose de régresser linéairement la consommation sur les trois variables "Price", "Income" et "Temp", en supposant de plus qu'une constante est présente dans le modèle. On note la constante β_1 et les trois coefficients associés aux variables précédentes respectivement β_2, β_3 et β_4. Réaliser la phase d'estimation de cette régression et commenter le signe des coefficients estimés.

3) Tester la significativité globale du modèle proposé, i.e. $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$, à l'aide du test de Fisher global.

4) Tester la significativité de la variable "Price" dans ce modèle au seuil de 5%. Tester de même la significativité de "Income", puis de "Temp".

5) Comparer le modèle complet précédent et le modèle sans la variable "Price" à l'aide d'un test de Fisher :

1. En basant le calcul sur la somme des carrés résiduelle de chaque modèle ;
2. En basant le calcul sur le coefficient de détermination de chaque modèle ;
3. En utilisant la fonction `anova`.
Quel est la différence entre ce test et le test de Student de significativité de la variable "Price" ?

6) Comparer le modèle complet et le modèle sans la variable "Price" et sans la constante à l'aide d'un test de Fisher. Procéder selon les 3 manières décrites ci-dessus. Commenter.

7) On désire à présent prédire la consommation de glaces pour les données suivantes : $Price = 0.3$, $Income = 85$ et $Temp = 65$. Proposer la prévision qui vous semble la meilleure au vu des modèles étudiés précédemment. Donner un intervalle de prédiction au niveau 95% autour de cette prévision.

8) Sous quelle hypothèse l'intervalle de prédiction précédent est-il valable ? Vérifier-la en observant le QQ-plot des résidus de la régression et en effectuant un test statistique.

9) Vérifier les autres hypothèses en rappelant la définition et en calculant les VIF ("Variance Inflation Factor") de chaque variable explicative et en effectuant une analyse graphique des résidus.

10) Observer le nuage de points en 3 dimensions des variables "Consumption", "Income" et "Temp", et l'ajustement par le modèle linéaire, à l'aide de la fonction `scatter3d` de la librairie `car`. On pourra également utiliser le module `RCommander`.

Ex 11. *Emissions de Gaz à effet de serre*

Le jeu de données "EmissionsGES.txt" contient les émissions de Gaz à effet de serre (GES) pour l'année 2003 dans 16 pays industrialisés (en millions de tonnes équivalent CO2). Il contient également les objectifs d'émissions pour 2010 issus des accords de Kyoto (pour les pays ayant ratifié le protocole).

1) Ajuster un modèle régression linéaire expliquant les objectifs d'émissions de GES en 2010 en fonction des émissions mesurées en 2003. Analyser la qualité du modèle. D'après l'estimation, quel est le pourcentage moyen envisagé de réduction des émissions de GES ?

2) Ajuster un modèle de régression linéaire expliquant les émissions de GES en 2003 en fonction de la population des pays. Analyser la qualité du modèle, notamment l'influence des individus (c'est à dire les pays) sur l'estimation à l'aide de leur distance de Cook.

3) Effectuer le même ajustement que ci-dessus sans prendre en compte les USA. Analyser la qualité du modèle.

Ex 12. *Modélisation de la concentration maximale journalière en ozone*

Le jeu de données "ozone.txt" contient la concentration maximale d'ozone (maxO3) mesurée chaque jour de l'été 2006 à Rennes. Il contient également les températures, la nébulosité et la vitesse du vent mesurés à 9h, 12h et 15h (respectivement T9, T12, T15, Ne9, Ne12, Ne15 et Vx9, Vx12, Vx15), ainsi que la direction principale du vent et la présence ou non de pluie. On désire expliquer au mieux la concentration d'ozone à l'aide des variables disponibles dans le jeu de données.

1) Analyser le nuage de points et la corrélation linéaire entre maxO3 et chacune des variables quantitatives disponibles (c'est à dire T9, T12, T15, Ne9, Ne12, Ne15, Vx9, Vx12 et Vx15). Est-il raisonnable de supposer qu'il existe un lien linéaire entre maxO3 et ces variables ?

2) Ajuster le modèle de régression linéaire expliquant maxO3 en fonction de toutes les variables quantitatives précédentes. Tester la significativité de chacune des variables explicatives dans ce modèle. Le résultat est-il en accord avec les observations de la question précédente ?

3) Calculer les VIF (Variance Inflation Factor) pour chacune des variables explicatives du modèle précédent. En quoi ces valeurs expliquent les résultats des tests de Student effectués ci-dessus ?

4) On décide d'enlever des variables au modèle précédent. Quelles variables semble-t-il naturel d'enlever au vu de la question précédente ? Ajuster le ou les nouveaux modèles proposés et répéter les analyses effectuées dans les deux questions précédentes.

5) Sélectionner le meilleur modèle possible ayant toutes ses variables significatives et aucun problème de multicolinéarité. Le choix du modèle pourra reposer sur un critère de sélection de type BIC.

6) Mettre en oeuvre une sélection automatique du meilleur sous-modèle possible du "gros" modèle ajusté dans la question 2, selon le critère BIC. On pourra utiliser la fonction `regsubsets` dans la libraire `leaps` (puis `plot.regsubsets`) ou `stepAIC` dans la librairie `MASS`. Comparer le modèle retenu avec le modèle choisi à la question précédente.

7) Appliquer la sélection automatique précédente en vous basant sur d'autres critères que BIC. Les modèles retenus sont-ils les mêmes ? Si non, lequel semble préférable ?

8) Analyser résidus du modèle sélectionné à la question précédente par des représentations graphiques et en effectuant des tests d'homoscédasticité et de non-corrélation des résidus. Toutes les hypothèses d'un modèle linéaire semblent-elles vérifiées ?

9) Afin de résoudre le problème d'auto-corrélation des résidus, on propose d'ajouter la maximum d'ozone de la veille dans le modèle. Créer cette variable, que l'on nommera maxO3v et ajouter-la au jeu de données. Observe-t-on un lien linéaire entre maxO3 et maxO3v ?

10) Ajuster le modèle de régression contenant maxO3v comme variable explicative supplémentaire. Analyser les résultats de l'ajustement : les hypothèses d'un modèle linéaire sont-elles vérifiées ?

11) Comparer ce dernier modèle au modèle sans maxO3v à l'aide d'un test de Fisher et en comparant les différents critères de sélection (AIC, BIC, Cp de Mallows, R^2 ajusté).

Ex 13. *ANOVA à 1 facteur : effet du dosage d'un médicament*
Le fichier "chemical.txt" contient la concentration dans le sang (en ng/ml) d'un certain produit chimique chez 40 patients selon qu'ils ont absorbé un médicament dosé à 25, 50, 100 ou 200 mg de substance active (l'almitrine bismesylate). Les patients sont ainsi séparés en 4 groupes de 10 selon le dosage reçu.

1) Représenter les données à l'aide de boîtes à moustaches. Réordonner si nécessaire les niveaux du facteur "dose" pour avoir une représentation par ordre croissant du dosage. Les hypothèses nécessaires à une analyse de variance de l'effet dosage semblent-elles vérifiées ?

2) Effectuer un test d'égalité de variance de Bartlett (`bartlett.test`) et de Levene (`leveneTest` dans la librairie `car`) pour confirmer le problème vu précédemment.

3) On propose de s'intéresser à une transformation logarithmique de la variable "concent". Représenter les boîtes à moustaches des données transformées. Effectuer des tests d'égalité des variances pour vérifier la stabilité des variances après transformation.

4) Réaliser l'analyse de la variance et tester l'effet du dosage sur la concentration en produit chimique.

5) Analyser graphiquement les résidus du modèle pour valider la démarche.

6) On veut à présent comparer plus précisemment les effets des dosages entre eux. Combien de paires de dosages cela fait-il à comparer ? Effectuer pour chaque paire un test de Student de comparaison des moyennes au niveau $\alpha = 0.05$ (fonction `t.test`).

7) La sortie de la fonction `t.test` fournit également un intervalle de confiance pour la différence des espérances entre les 2 groupes testés. Peut-on affirmer qu'avec une probabilité de 0.95 toutes les différences d'espérances appartiennent aux intervalles de confiance obtenus ci-dessus par la fonction `t.test` au niveau $\alpha = 0.05$?

8) Modifier la démarche précédente en utilisant la correction de Bonferroni afin d'obtenir une probabilité de couverture simultanée des intervalles de confiance d'au moins 0.95.

9) Proposer de même des intervalles de confiance au niveau simultané 0.95 en utilisant la méthode de Tukey (fonction `glht` de la librairie `multcomp` ; si le modèle se nomme `reg` : `glht(reg, linfct = mcp(dose = "Tukey"))`) et représenter les (`plot` du résultat).

Ex 14. *ANOVA à 2 facteurs : alimentation des rats*
On veut étudier l'évolution du poids des rats selon 4 régimes alimentaires différents : combinaison de deux types de protéines différentes (Boeuf et Céréales) et de deux quantités différentes (élevée ou basse). Chaque traitement est effectué sur 10 rats choisis au hasard de telle sorte que notre échantillon est constitué de 40 rats. Les données se trouvent dans le fichier "poids-rats.txt".

1) Le plan est-il équilibré ?

2) Représenter les données à l'aide de 4 boîtes à moustaches croisant le poids et les deux facteurs. Les conditions semblent-elles remplies pour effectuer une ANOVA ?

3) Analyser l'effet de l'interaction des deux facteurs sur le poids à l'aide d'un graphique (utiliser la fonction `interaction.plot`) puis effectuer le graphique équivalent en échangeant le rôle joué par les facteurs. Commenter.

4) Réaliser l'analyse de la variance complet à deux facteurs (incluant les effets marginaux et l'interaction entre les facteurs) et tester la présence d'une interaction entre les facteurs. Comparer ce résultat avec les représentations précédentes.

5) On considère à présent le modèle additif sans interaction. Estimer ce modèle. Sous quelles contraintes sur les paramètres cette estimation est-elle effectuée ?

6) Tester l'effet de chacun des facteurs.

Ex 15. *ANCOVA : la taille des méduses*
Le fichier "jellyfish.txt" contient la largeur (Breadth) et la longueur (Length) en millimètres de 46 méduses réparties sur 2 sites en Australie (Dangar Island et Salamander Bay).

1) Vérifier que la variable "Site" est bien définie comme un facteur et représenter à l'aide de boxplots la largeur des méduses en fonction de leur site. Semble-t-il y avoir un effet site sur la largeur ? Confirmer l'analyse en effectuant une analyse de la variance.

2) Etudier de même l'effet du site sur la longueur des méduses.

3) Représenter à présent le nuage de points entre la largeur et la longueur en utilisant des couleurs différentes selon le site.

4) On désire modéliser la largeur des méduses en fonction de leur longueur et de leur site d'appartenance. Mettre en oeuvre le modèle complet avec interaction, puis tester l'égalité des pentes et l'égalité des ordonnées à l'origine. Superposer au nuage de points précédents les différentes droites prédites par les différents modèles testés.

5) Quel modèle final retenir ? Comparer ce résultat avec celui de la première question et commenter.

Ex 16. *ANCOVA : Retour à la modélisation de l'ozone*
On considère de nouveau les données "ozone.txt" étudiées dans l'exercice 12.

1) On reprend le modèle sélectionné dans l'exercice 12, soit "maxO3" en fonction de "T12", "Ne9", "Vx9" et "maxO3v" où "maxO3v" représente la concentration maximale en ozone de la veille (créer cette variable si besoin). Ajuster ce modèle sur les données.

2) Représenter graphiquement "maxO3" en fonction de la présence de pluie. Un lien semble-t-il présent ?

3) Ajouter au modèle de la première question la variable "pluie" de manière la plus générale possible (i.e. en incluant une interaction avec chaque variable en plus d'un effet sur la constante). Tester la significativité de ces ajouts en effectuant un test de Fisher de modèles emboités entre ce modèle et le modèle initial. Le résultat est-il en désaccord avec l'analyse graphique de la question précédente ?

4) Comparer la significativité de la variable "Vx9" dans le modèle initial (sans la variable "pluie") et dans le modèle incluant l'effet pluie. Comment expliquer cette différence ?

5) De même : représenter graphiquement "maxO3" en fonction de la direction du vent et étudier la pertinence d'inclure un effet vent dans le modèle initial.

Ex 17. *Application des régressions PLS, ridge et lasso sur données corrélées*
On considère le jeu de données "autos.txt". On désire expliquer au mieux la variable "PRIX" en fonction des autres variables du jeu de données.

1) Calculer la corrélation linéaire et observer les nuages de points entre "PRIX" et les autres variables du jeu de données.

2) Ajuster un modèle de régression linéaire expliquant "PRIX" par rapport à toutes les autres variables. Quelles sont les variables significatives dans ce modèle ? Comment expliquer ce résultat ?

3) Pour chaque k allant de 1 à n, prédire y_k (le prix de la voiture k) à l'aide de la régression linéaire effectuée sans l'individu k. En déduire le PRESS de la régression linéaire : $PRESS = \sum_{k=1}^{n}(y_k - \hat{y}_k)^2$.

4) Pourquoi des régressions de type PLS, ridge ou lasso sont-elles bien adaptées à cette situation ?

5) Mettre en oeuvre une modélisation PLS pour expliquer la variable PRIX et relever son erreur de prévision issue d'une modélisation croisée (i.e. son PRESS). On pourra adopter les étapes suivantes, en notant `y` la variable "PRIX" et `x` la matrice des variables explicatives :

i- On commence par calculer un nombre maximal de composantes PLS en spécifiant `ncomp` au maximum dans la fonction `plsr` de la librairie "pls" ; on indique également que l'on souhaite travailler avec les données centrées réduites (option `scale=TRUE`) et que l'on souhaite procéder à une validation croisée pour pouvoir choisir le nombre de composantes (option `validation='LOO'`) ce qui donne :
```
res=plsr(y~x,ncomp=6,scale=TRUE,validation='LOO')
```

ii- On calcule et on représente l'erreur quadratique moyenne de prévision (PRESS) issue de la validation croisée, en fonction du nombre de composantes :
```
plot(MSEP(res,estimate='CV'))
```

iii- L'étape précédente permet de choisir le nombre de composantes retenu. On peut à présent relancer la fonction `plsr` en spécifiant ce nombre choisi dans `ncomp` :
```
respls=plsr(y~x,ncomp=1,scale=TRUE)
```

iv- On peut enfin accéder aux coefficients estimés par `coef(respls)`. Ces coefficients sont déduits des coefficients de la régression PLS et s'appliquent aux données centrées réduites. Ils correspondent donc à l'estimation des coefficients β_i dans l'écriture

$$y - \bar{y} = \sum_{i=1}^{p} \beta_i \frac{x_i - \bar{x}_i}{\sqrt{var(x_i)}} + \epsilon$$

v- Les résidus du modèle sont accessibles dans `respls$residuals`.

vi- Pour chaque k allant de 1 à n, on estime le modèle PLS précédent sans l'individu k et en notant `respls` le résultat, on prédit y_k soit en utilisant `coef(respls)` et la formule précédente, soit en utilisant la commande `predict(respls,t(x[k,]))[1,1,1]`. On en déduit $PRESS = \sum_{k=1}^{n}(y_k - \hat{y}_k)^2$

6) Mettre en oeuvre une régression ridge pour expliquer la variable PRIX et relever son PRESS issu d'une validation croisée. On pourra adopter les étapes suivantes :

i- Choisir le paramètre de pénalisation par validation croisée à l'aide de la fonction `ridge.cv` de la librairie "parcor" : il suffit de proposer un vecteur `lambda` de valeurs et la fonction trace le PRESS obtenu par validation croisée en fonction de `lambda` ; elle renvoie également la valeur de `lambda` minimisant ce PRESS :
```
ridge.cv(x,y,lambda=seq(0,50,0.2),k=nrow(x),plot.it=TRUE)
```

ii- On peut ensuite lancer la régression ridge pour la valeur choisie :
```
resridge=lm.ridge(y~x,lambda=16.4)
```

iii- Les coefficients estimés sont accessibles dans `coef(resridge)`. Ils correspondent à l'estimation des coefficients β_i dans $y = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \epsilon$.

iv- Malheureusement, les résidus ne sont pas accessibles par `resridge$residuals`. La fonction `predict` n'est pas non plus utilisable avec les objets de la classe `ridgelm` : $\hat{y}_k$ se calcule par la formule précédente en utilisant les coefficients estimés.
Pour chaque k allant de 1 à n, on estime le modèle ridge précédent sans l'individu k et on prédit y_k comme expliqué ci-dessus. On en déduit le PRESS.

7) Mettre en oeuvre une régression lasso pour expliquer la variable PRIX et relever son PRESS issu d'une validation croisée. On pourra adopter les étapes suivantes :

i- Choisir le paramètre de pénalisation par validation croisée à l'aide de la fonction `cv.lars` de la librairie "lars". Le paramètre est testé entre 0 à 1 car il est par défaut normalisé par la norme ℓ_1 de l'estimateur MCO (cas limite de la régression lasso) :
```
rescv=cv.lars(x,y,K=nrow(x), type='lasso')
```
On peut récupérer la valeur `delta` qui minimise le PRESS grâce au résultat précédent :
```
delta=rescv$index[which.min(rescv$cv)]
```

ii- On lance la régression lasso par `reslasso=lars(x,y)` : elle calcule les coefficients associés à n'importe quelle valeur de `delta`.

iii- On peut récupérer les coefficients associés au `delta` choisi par :
```
coef(reslasso,s=delta,mode='fraction')
```
Ils s'appliquent aux données centrées et correspondent donc à l'estimation des coefficients β_i dans $y - \bar{y} = \sum_{i=1}^{p} \beta_i (x_i - \bar{x}_i) + \epsilon$.

iv- On peut enfin calculer les $\hat{y}_k$ associés au `delta` par :
```
predict(reslasso,x,s=delta,mode="fraction")
```
et en déduire les résidus.

v- Pour chaque k allant de 1 à n, on estime le modèle lasso précédent sans l'individu k, on prédit y_k comme ci-dessus et on en déduit le PRESS.

8) Comparer le PRESS des régressions précédentes (pour les paramètres choisis). Quelle méthode choisir pour ce jeu de données ?

Ex 18. *Régression sur un plus grand nombre de variables que d'individus*
On considère le jeu de données "cookies.txt" qui contient le taux de sucre de 40 cookies, ainsi que leur spectre d'absorbance mesuré sur 700 longueurs d'ondes.

1) Ajuster un modèle de régression linéaire expliquant le taux de sucre par rapport aux 700 valeurs du spectre d'absorbance. A quel problème est-on confronté ?

2) Ajuster un modèle de régression PLS pour expliquer le taux de sucre par rapport aux 700 valeurs du spectre d'absorbance, et relever son PRESS. Utiliser pour ce faire la démarche expliquée dans l'exercice précédent.

3) Ajuster de même un modèle de régression ridge (tester les valeurs du paramètre allant de 0 à 50) et relever son PRESS.

4) Ajuster de même un modèle de régression lasso (on ajoutera l'option `use.Gram=F` aux fonctions `cv.lars` et `lars`, ce qui est recommandé lorsque le nombre de variables est supérieur au nombre d'individus) et relever son PRESS.

5) D'après les PRESS, quelle modélisation choisir ?

6) On considère à présent un nouveau jeu de données de 32 cookies, présents dans "nouveaux_cookies.txt". Prédire le taux de sucre de ces 32 cookies à l'aide de leur spectre d'absorbance en utilisant le modèle retenu précédemment. Comparer l'erreur quadratique de prévision avec les erreurs associées aux autres modèles.