Université de Nantes
UFR des Sciences et Techniques
Département de Mathématiques

TP Statistique en grande dimension.
M2 Ingénierie Statistique
Frédéric Lavancier

Exercice 1. On étudie le jeu de données `Hitters` disponible dans la libraire `ISLR` de R.

1. Que contient ce jeu de données?
2. On désire modéliser le salaire `Salary` en fonction des variables disponibles. Ajuster un modèle de régression linéaire en utilisant toutes les variables à disposition. Analyser la qualité de cet ajustement.
3. On désire trouver le meilleur sous-modèle possible. Plusieurs outils sont disponibles :
   - Mettre en oeuvre une méthode de sélection automatique exhaustive (fonction `regsubsets` de la libraire `leaps`) et observer l'évolution des SCR pour les modèles retenus en fonction de leur taille (disponibles dans le `summary` de la sortie précédente).
   - Déduire de ces SCR la valeur des R^2, R^2_a, AIC, BIC et C_p correspondants. Comparer avec les valeurs fournies dans le `summary` de `regsubsets`. Tracer leur évolution en fonction de la taille du modèle.
   - Mettre en oeuvre une sélection stepwise backward basée sur les mêmes critères (fonction `regsubsets` avec `method="backward` ou fonction `step`)
   - Mettre en oeuvre une sélection stepwise forward
   - Mettre en oeuvre une sélection stepwise backward hybride et une sélection stepwise forward hybride
4. Quel modèle suggérez-vous de retenir au final?

Exercice 2. On considère le sous-jeu de données composé des 18 premières lignes sans valeur manquante de `Hitters`.

1. Effectuer la régression de `Salary` en fonction de toutes les variables à disposition. Analyser la qualité de cet ajustement.

2. A l'aide du modèle précédent, prédire le salaire des autres joueurs de la table `Hitters` (ceux n'ayant pas servi à ajuster le modèle) et analyser la qualité des prévisions.

3. Mettre en oeuvre des méthodes de sélection automatique classiques pour réduire le nombre de variables explicatives. Qu'obtient-on?

4. Permuter de façon aléatoire les salaires des 18 joueurs. Quel lien est à attendre entre ces nouveaux salaires et les variables explicatives? Ajuster un modèle de régression expliquant ces salaires à l'aide des toutes les variables. Commenter.

5. Reprendre le jeu de données `Hitters` complet et permuter tous les salaires de façon aléatoire. Ajuster le meilleur modèle de régression possible pour expliquer les salaires en fonction des autres variables.

Exercice 3.

1. Générer sous `R` les vecteurs suivants : `x=rnorm(1000)`, `y=x-2*x^2+rnorm(1000)`. Ecrire le modèle linéaire qui a permis de générer ces données. Que valent n et p?

2. On suppose avoir observé les deux vecteurs x et y précédents, sans connaître le lien théorique précédent qui lie x et y. On cherche à estimer le lien entre ces deux variables. Calculer la corrélation et tracer le nuage de points entre x et y. Commenter.

3. Ajuster les modèles suivants :

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1}$$
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon \tag{2}$$
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon \tag{3}$$
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \epsilon \tag{4}$$

4. Calculer différents critères de sélection pour chacun des modèles précédents.

5. Créer deux fonctions permettant d'estimer l'erreur test par une validation croisée LOO (Leave-one-out) pour un modèle ajusté par la fonction `lm` : la première en utilisant le principe général de cette méthode qui nécessite donc l'estimation de n modèles différents; la seconde en utilisant la formule adaptée à la régression linéaire donnant directement le risque LOO à partir de la seule estimation du modèle complet (on pourra utiliser la fonction `hatvalues`).

6. Appliquer les deux fonctions précédentes aux quatre modèles ci-dessus. Comparer les résultats des deux fonctions. Comparer les résultats des quatre modèles.

7. La fonction `cv.glm` de la librairie `boot` permet d'estimer l'erreur test par validation croisée K-fold. Pour quelle valeur de K cela correspond-il à une validation croisée LOO? Vérifier que vous obtenez les mêmes résultats dans ce cas que dans la question précédente. Il vous faudra pour cela ré-estimer vos modèles avec la fonction `glm` (format de modèle reconnu par `cv.glm`) à la place de `lm`. Calculer enfin une estimation de l'erreur test de chaque modèle avec $K = 10$.

8. Affiner au maximum le modèle retenu et conclure : selon votre choix, quel est le lien estimé entre y et x? Comparer au lien théorique.

Exercice 4. On considère le jeu de données `Caravan` de la librairie `ISLR` de R. Ce jeu de données contient, pour 5822 clients d'une assurance, 86 variables décrivant leur profil. La dernière variable `Purchase` indique si le client a souscrit une assurance pour caravane ou non.

1. Ouvrir le jeu de données et l'aide associée pour comprendre ce que représentent les variables disponibles.

2. Quel est le pourcentage de clients ayant souscrit une assurance caravane?

3. Ajuster un modèle de régression logistique modélisant la probabilité de souscrire une assurance caravane en fonction de toutes les autres variables à disposition. Calculer les VIF (variance inflation factor) de chaque variable explicative et commenter.

4. Effectuer une sélection automatique des variables (stepwise au choix) en utilisant le critère AIC, puis le critère BIC. Calculer les VIF pour les modèles retenus.

5. L'objectif de l'assureur est de démarcher des clients de manière ciblée pour leur faire souscrire une assurance caravane. S'il démarchait les clients de façon aléatoire, sans tenir compte de leurs caractéristiques, quel serait environ son taux de réussite?

6. On souhaite utiliser l'un des 3 modèles estimés ci-dessus (le global, celui sélectionné par AIC et celui par BIC) pour cibler les clients à démarcher. Si l'on choisissait de démarcher tous les clients ayant une probabilité de souscrire l'assurance supérieure à 0.5, quel pourcentage de clients cela représenterait-il pour chacun des 3 modèles estimés? Quel seuil faudrait-il choisir à la place de 0.5 pour que ce pourcentage corresponde à environ 6% des clients? On décide dans la suite de fixer ce seuil à 0.2 et on cherche à sélectionner le meilleur modèle parmi les 3 précédents.

7. Estimer le taux de réussite du démarchage (c'est à dire le nombre de vrais positifs par rapport au nombre de positifs prédits) sur l'échantillon d'apprentissage pour chaque modèle. Quel problème pose cette manière d'estimer le taux de réussite?

8. Estimer le taux de réussite de chaque modèle par validation croisée. Motiver la méthode de validation croisée choisie. On pourra utiliser la fonction `cv.glm` de la librairie `boot` en définissant bien la fonction de coût souhaitée (option `cost`).

9. Estimer de même le taux de réussite pour chaque modèle lorsque le seuil varie de 0.10 à 0.30 par pas de 0.01. Tracer la courbe des taux estimés pour chaque modèle. Quel modèle semble préférable? Quel seuil semble préférable? Recommencer cette démarche 5 fois et superposer les nouvelles courbes. Conclure.

Exercice 5.

1. Simuler un jeu de données dont la première colonne y contient 100 réalisations indépendantes d'une loi de Bernoulli de paramètre 0.5, tandis que les autres variables, au nombre de 5000, sont toutes des réalisations indépendantes de 100 valeurs issues d'une loi normale standard. Le jeu de données ainsi créé contient donc 100 individus et 5001 variables.

2. Quel est le lien attendu entre y et les variables explicatives?

3. Une méthode de prévision de y basée sur les variables explicatives $X_1, \ldots, X_{5000}$ s'exprime nécessairement sous la forme $\hat{y} = f(X_1, \ldots, X_{5000})$. La forme de la fonction f est généralement obtenue grâce à une estimation sur un échantillon d'apprentissage. On s'intéresse au taux d'erreur de classification "test", c'est à dire à la probabilité que $\hat{y}$ soit différent de y, lorsque y et $X_1, \ldots, X_{5000}$ sont indépendants de l'échantillon d'apprentissage. Montrer que quelle que soit la méthode utilisée, le taux d'erreur sera de 50%.

Dans la suite de l'exercice, on suppose qu'on a observé le jeu de données simulé ci-dessus, sans connaître les liens théoriques entre les variables. On souhaite ajuster un modèle expliquant au mieux y en fonction des variables à disposition, et estimer le taux d'erreur des prévisions associées. Etant donné le très grand nombre de variables explicatives, un statisticien décide de ne garder que les 5 variables les plus corrélées avec y afin d'ajuster un modèle de régression logistique.

4. Pour quantifier la corrélation entre la variable y, qui est binaire, et une variable quantitative, deux possibilités s'offrent a priori au statisticien : (i) soit utiliser la corrélation de Pearson $\hat{\rho}$ (comme si y était une variable quantitative), (ii) soit utiliser le "rapport de corrélation" quantifiant le lien entre une variable qualitative et une variable quantitative :

$$\hat{\eta}^2 = \frac{S^2_{\text{inter}}}{S^2_{\text{total}}},$$

où S^2_{inter} est la somme des carrés "inter-classes" et S^2_{total} est la somme des carrés total. Montrer que dans notre cas

$$\hat{\eta}^2 = \hat{\rho}^2.$$

5. Quelles sont les 5 variables les plus corrélées (en valeur absolue) avec y?

6. Ajuster un modèle de régression logistique faisant intervenir ces 5 variables et analyser la qualité du modèle.

7. Afin d'estimer le taux d'erreur de classification associé à ce modèle, le statisticien décide de procéder à une validation croisée K-fold avec $K = 10$. Estimer de cette manière le taux d'erreur.

8. Recommencer 10 fois la démarche précédente, c'est à dire les questions 1., 5., 6. et 7, et stocker les 10 taux d'erreurs estimés. Que valent ces estimations. Quelle est leur moyenne? Ces résultats sont-ils conformes à la valeur théorique attendue?

Le problème précédent est appelé biais de sélection. Lorsqu'on effectue la validation croisée, l'échantillon test n'est certes pas utilisé pour l'estimation du modèle, mais il a été utilisé initialement pour sélectionner les variables. Ainsi lorsqu'on confronte les prévisions à l'échantillon test, ces dernières sont trop optimistes car elles ont déjà utilisé l'information contenue dans cet échantillon test. Pour estimer convenablement le taux d'erreur, il faut que l'échantillon test ne soit utilisé *à aucun moment* dans la procédure de modélisation.

9. Mettre en oeuvre une validation croisée valide, respectant la règle précédente, pour estimer le taux d'erreur associé à la démarche de modélisation choisie par le statisticien. Recommencer la démarche 10 fois, comme dans la question 7., afin d'obtenir 10 estimation du taux d'erreur. Comparer avec la valeur théorique attendue. On pourra utiliser la fonction `cvsegments` de la librairie `pls` pour découper l'échantillon aléatoirement en K parties de taille égale.

Le résultat précédent est cohérent avec la question 3. Néanmoins, il peut paraître étonnant que dans le modèle estimé ci-dessus, les variables soient toutes significatives au niveau $\alpha = 5\%$. De même (c'est lié), ces variables semblent corrélées à y, d'après la valeur de $\hat{\rho}$, alors qu'elles sont censées être indépendantes de y. Les questions suivantes éclaircissent ce phénomène.

10. On admet qu'en vertu du théorème limite central, la corrélation de Pearson $\hat{\rho}$ entre deux variables non-corrélées suit approximativement, lorsque n est grand, une loi normale centrée de variance $1/n$. Si on se réfère à ce résultat, à partir de quelle valeur $|\hat{\rho}|$ peut-on considérer que deux variables ont une corrélation significativement non-nulle au seuil $\alpha = 5\%$?

11. Comparer les corrélations des variables dans le modèle précédent avec ce seuil?

12. Parmi 5000 variables théoriquement non corrélées avec y, environ combien de variables observées sur $n = 100$ individus présenteront une corrélation empirique significativement non nulles avec y au seuil $\alpha = 5\%$? Le vérifier sur le jeu de données simulé.

Les corrélations empiriques précédentes sont appelées "faux positifs" : elles semblent significatives alors qu'en théorie aucune corrélation n'existe. Ce phénomène est propre à la grande dimension : à force de chercher des corrélations parmi un grand nombre de variables, on en trouve toujours, à cause des fluctuations de la corrélation empirique, même si aucune corrélation théorique n'existe. La validation croisée, correctement appliquée, permet de détecter ce phénomène.

Exercice 6. On considère le jeu de données `Hitters` de la librairie `ISLR`. On souhaite modéliser la variable `salary` en fonction des autres variables disponibles.

1. Créer un jeu de données "train" contenant 3/4 des individus sans valeurs manquantes de `Hitters`, tirés aléatoirement. Le reste du jeu de données composera l'échantillon "test".

2. Effectuer une régression PCR sur l'échantillon "train" en sélectionnant le nombre de composantes par une validation croisée K-fold où $K = 10$. On utilisera les fonctions `pcr` et `validationplot` de la librairie `pls`. Observer les coefficients estimés.

3. Effectuer une régression PLS sur l'échantillon "train" en sélectionnant le nombre de composantes par une validation croisée K-fold où $K = 10$. On utilisera les fonctions `plsr` et `validationplot` de la librairie `pls`. Observer les coefficients estimés.

4. Prédire le salaire des joueurs de l'échantillon test à l'aide des deux modèles précédents. Lequel conduit à une erreur de prévision minimale?

Exercice 7. Continuer l'exercice précédent en ajustant un modèle ridge sur l'échantillon d'apprentissage, ainsi qu'un modèle Lasso. Que valent les erreurs de prévision sur l'échantillon test pour ces deux modèles?

Exercice 8. On considère le jeu de données `cookies` de la libraire `fdm2id` qui contient, pour 72 cookies, leur spectre par proche infrarouge (les 700 premières variables, chacune correspondant à une longueur d'onde) ainsi que la mesure de 4 ingrédients (variables 701 à 704). On souhaite prédire le taux de sucre (variable 702) en fonction du spectre. Les 40 premiers cookies formeront l'échantillon d'apprentissage et les 32 derniers l'échantillon test.

1. Représenter sur un même graphique les 40 spectres de l'échantillon d'apprentissage.

2. Effectuer une régression linéaire avec sélection stepwise et critère BIC pour expliquer le taux de sucre à l'aide des spectres. Quelles sont les longueurs d'onde retenues? Les mettre en évidence sur le graphique précédent.

3. Quelle est l'erreur de prévision issue de ce modèle sur l'échantillon test?

4. Effectuer une régression PCR pour le même objectif et calculer de même l'erreur test.

5. Même question avec une régression PLS.

6. Même question avec une régression Ridge.

7. Même question avec une régression Lasso. Mettre en évidence les longueurs d'onde retenues sur le graphique précédent.

8. Est-il opportun d'envisager une amélioration avec l'estimateur Gauss-Lasso? Effectuer l'estimation et calculer l'erreur de prévision test.

9. Effectuer une régression Lasso adaptative, évaluer l'erreur test et mettre en évidence les longueurs d'onde sélectionnées.

10. Même question avec une méthode Elastic Net de paramètre $\alpha = 0.5$.

11. Analyser les résultats, à la fois en terme de qualité de prévision et d'identification des longueurs d'onde importantes.

Exercice 9. On souhaite réaliser une petite étude par simulation pour évaluer les qualités respectives de 4 méthodes d'estimation d'un modèle de régression linéaire. On s'intéresse pour chacune d'elle à ses qualités de sélection de variables et à ses qualités prédictives. Le programme "SimusReg.R" permet de réaliser cette étude. Il contient deux fonctions, `Simudata` et la fonction principale `fun`, et un exemple d'utilisation en fin de programme.

1. Quel modèle génère la fonction `Simudata`? Combien de variables explicatives sont générées? Parmi elles, lesquelles sont pertinentes pour la modélisation? Ecrire l'équation du modèle.

2. Identifier les 4 méthodes d'estimation mises en oeuvre dans la fonction `fun`.

3. Détailler les différentes sorties proposées par la fonction `fun`.

4. Remplacer la valeur des options `names` et `title` du boxplot réalisé dans l'exemple par les bonnes informations.

5. Réaliser une étude comparative des méthodes lorsque $n = 50$ et $p = n/10$, $p = n$, $p = 2n$, $p = 10n$. Pour chaque situation, on considèrera 100 simulations afin de calculer les différents critères. On synthétisera les résultats en terme de qualité de sélection, nombre de variables sélectionnées, erreurs de prévision et temps de calcul.

6. Réaliser la même étude pour $n = 100$ et $p = n/10$, $p = n$, $p = 2n$, toujours basée sur 100 simulations dans chaque cas. Considérer de plus le cas $p = 10n$ en ne faisant qu'une seule simulation afin d'en évaluer le temps de calcul. Une fois ce temps analysé, lancer 100 simulations pour $p = 10n$ mais en omettant la méthode la plus couteuse en temps de calcul.

7. Conclure sur les mérites respectifs de chaque méthode dans le contexte de l'étude.

8. Quelles autres types de simulations pourrait-on envisager pour confirmer ou affiner ces conclusions?

Exercice 10. On reprend le jeu de données `Caravan` traité dans l'exercice 4. On rappelle que l'objectif est de prédire la probabilité qu'un client souscrive une assurance caravane.

1. Créer un jeu de données d'apprentissage contenant la moitié des clients ayant souscrit une assurance caravane et la moitié des clients n'en ayant pas, le tout tiré aléatoirement. Le reste des données formera l'échantillon test.

2. Le modèle retenu dans l'exercice 4 était une régression logistique issue d'une sélection stepwise forward hybride. Réestimer ce modèle sur l'échantillon d'apprentissage et en déduire les probabilités de souscrire une assurance caravane pour les clients de l'échantillon test.

3. Représenter la courbe ROC issue de ces prévisions. On rappelle que si `pi` désigne les probabilités prédites et `r` les vraies réponses, ceci peut se faire grâce à la librairie `ROCR` en utilisant la fonction `pr=prediction(pi,r)` puis `performance(pr,measure = "tpr", x.measure = "fpr")` suivi d'une représentation graphique du résultat.

4. Calculer l'AUC, c'est à dire l'aire sous la courbe ROC : `performance(pr,measure="auc")`.

5. Ajuster à présent un modèle logistique Lasso sur l'échantillon d'apprentissage. On choisira le paramètre de régularisation λ, par validation croisée, qui maximise l'AUC (option `type.measure="auc"` dans la fonction `cv.glmnet`).

6. En déduire les prévisions sur l'échantillon test, la courbe ROC associée et l'AUC.

7. Comparer les courbes ROC et l'AUC des deux modèles précédents, ainsi que les variables sélectionnées.

8. Essayer d'améliorer les qualités prédictives et/ou d'interprétation des variables sélectionnées grâce au(x) modèle(s) de votre choix.

Exercice 11.

1. Créer une fonction qui prend en entrée n, m, $m_0 \leq m$ et μ_1, et

   - simule m_0 échantillons contenant chacun n réalisations d'une loi normale centrée réduite, et $m - m_0$ échantillons de taille n d'une loi normale d'espérance μ_1 et de variance 1.
   - effectue pour chacun des m échantillons un test de Student de nullité de la moyenne.
   - retourne les m p-values associées à ces tests

2. On considère dans la suite $n = 100$ et $m = 1000$. Pour chacune des situations suivantes, relever le nombre de positifs, de vrais-positifs, de faux-positifs et la proportion de faux-positifs lorsqu'on applique chacun des m tests précédents au niveau $\alpha = 0.05$ sans correction, lorsqu'on applique une procédure de Bonferroni associée à $FWER \leq 0.05$ et lorsqu'on applique une procédure de Benjamini-Hochberg associée à $FDR \leq 0.05$. Il y a donc pour chaque situation 3 méthodes et 4 scores à calculer par méthode :

   - Pas de positifs : $m_0 = 1000$
   - Peu de positifs, facilement identifiables : $m_0 = 950$, $\mu_1 = 1$
   - Peu de positifs, difficilement identifiables : $m_0 = 950$, $\mu_1 = 0.3$

- Pas mal de positifs, facilement identifiables : $m_0 = 800$, $\mu_1 = 1$
- Pas mal de positifs, difficilement identifiables : $m_0 = 800$, $\mu_1 = 0.3$
- Beaucoup de positifs, facilement identifiables : $m_0 = 200$, $\mu_1 = 1$
- Beaucoup de positifs, difficilement identifiables : $m_0 = 200$, $\mu_1 = 0.3$

3. Discuter les avantages et inconvénients respectifs de chaque méthode à partir des résultats de ces simulations.

Exercice 12. On étudie le jeu de données `singh2002` de la librairie `sda`. Ce jeu de données est une liste qui contient, pour l'élément `singh2002$x`, une matrice de taille 102x6033 représentant l'expression de 6033 gènes chez 102 patients. Parmi ces patients, les 50 premiers sont sains et les 52 autres sont atteints du cancer de la prostate. Cette labellisation est disponible dans l'élément `singh2002$y`. On souhaite savoir quels gènes s'expriment différemment entre les patients sains et les patients malades.

1. Pour chaque gène, effectuer un test d'égalité des moyennes de leur expression entre les patients sains et les patients malades. On obtient donc 6033 p-values, chacune associée au résultat du test sur chaque gène.

2. Représenter ces p-values

3. Si on se donne un niveau $\alpha = 0.05$, combien de p-values sont rejetées? Parmi ces rejets, combien de faux positifs peut-on s'attendre à avoir?

4. Combien de rejets donne la procédure de Bonferroni associée à $FWER \leq 0.05$? Quels sont les gènes identifiés? Quelle est la probabilité qu'ils soient tous des vrais positifs?

5. Combien de rejets donne la procédure de Benjamini-Hochberg associée à $FDR \leq 0.05$? Quels sont les gènes identifiés? Parmi ces gènes, combien de faux positifs peut-on s'attendre à avoir?

6. Tester les autres procédures de tests multiples proposées par la fonction `p.adjust` et analyser les résultats.