Introduction : problèmes des méthodes statistiques classiques

Chargé du cours
Prof. Mustapha Rachdi

Université Grenoble Alpes
FR SHS, BP. 47, 38040 Grenoble cedex 09
France
Bureau : C08 Bât. Michel Dubois (BSHM)
e-mail : mustapha.rachdi@univ-grenoble-alpes.fr

Introduction

- Ce cours concerne l'analyse des données. Ces données pour lesquelles nous sommes tentés de prédire une réponse Y à partir d'un certain nombre de facteurs explicatifs (co-variables) $X_1, X_2, \ldots,$ dont certains peuvent ne pas être particulièrement utiles.
- Bien que les méthodes que nous discuterons puissent être utilisées uniquement pour la prédiction (c'est-à-dire comme une "boîte noire"), nous adopterons la perspective que nous aimerions que les méthodes statistiques soient interprétables et qu'elles permettent d'expliquer quelque chose sur la relation entre les X et Y.
- Les modèles de régression constituent un cadre attrayant pour aborder les problèmes de ce type, et dans la majorité de ce cours on se concentrera sur l'extension du modèle classique de régression pour traiter des données en grande dimension ou comportant le fléau de la dimension.

Données en grande dimension

Le calcul moderne a changé la façon dont la science est menée, et a permis aux chercheurs de collecter, stocker et accéder facilement aux données pour un grand nombre de co-variables (ci-dessous des exemples avec le nombre approximatif de co-variables entre parenthèses) :

- Les progrès de la technologie de l'information tels que REDCap ($\sim$ 100)
- Adoption de dossiers médicaux électroniques ($>$ 100)
- Technologies de biologie moléculaire telles que les puces d'ADN et RNA-Seq ($>$ 10000)
- Les progrès du génotypage et du séquençage génétique ($>$ 100000)

Données en grande dimension (suite)

- Ce type de données est connu sous le nom de données à haute dimension ou de grande dimension.

- Tout au long de ce cours, nous noterons :
  - par n la taille de l'échantillon (par exemple, le nombre de patients).
  - par p le nombre de caractères étudiés pour chaque individu

- Dans les données de grande dimension, p est grand par rapport à n :
  - Cela inclut certainement le cas où $p > n$.
  - Cependant, les idées que nous discuterons dans ce cours sont également pertinentes pour de nombreuses situations dans lesquelles $p < n$. Par exemple, si $n = 100$ et $p = 80$, il n'est probablement pas souhaitable d'utiliser la méthode des moindres carrés ordinaires

Introduction (suite)

- Nous utiliserons X pour désigner la matrice $n \times p$ contenant les variables, indépendantes, prédictives, l'élément x_{ij} correspond à la valeur de la $j$ème variable pour le $i$ème individu.
- Nous allons désigner par y une réalisation de la variable réponse Y : vecteur de longueur n.
- Par souci de simplicité, pour la majeure partie de ce cours, nous supposerons que Y est normalement distribuée, mais nous examinerons d'autres types de réponses.

Analyse uni-variée

- Une approche simple et largement utilisée pour analyser des données de grande dimension consiste à diviser le problème en un grand nombre de problèmes de faibles/petites dimensions.

- Par exemple, plutôt que d'essayer de régresser simultanément sur toutes les co-variables, nous pouvons effectuer p régressions à une seule variable séparémment (une pour chaque co-variable) :

$$y_i = \alpha_j + \beta_j x_{ij} + \epsilon_i \text{ où } \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2) \text{ sont i.i.d.}$$ [1]

Cette approche est connue sous le nom de *régression marginale*.

1. Indépendantes et Identiquement Distribuées

Analyse uni-variée : défis

- L'intérêt de cette approche est qu'elle est une régression classique. Elle peut être facilement appliquée aux analyses distinctes (uni-variées) afin de produire des estimations des coefficients ($\widehat{\beta}_j$), des intervalles de confiance et des tests d'hypothèses produisant des p-valeurs (p_j).

- Cependant, la complication majeure est que cette approche implique un grand nombre d'analyses séparées qui doivent en quelque sorte être combinées en un seul ensemble de résultats.

- Ainsi, pendant que les méthodes standard peuvent être utilisées pour les analyses initiales, il y a eu beaucoup d'innovations au cours des 30 dernières années en termes de combinaison de ces résultats. Nous discuterons ces innovations lors du traitement des "Tests à grande échelle : Large Scale Testing".

Modèles uni-variés : limitations

La régression marginale est simple, mais présente plusieurs inconvénients :

- Ne tient pas compte de la corrélation entre les co-variables.
- Ne fournit aucun moyen d'estimer l'effet d'indépendance d'une co-variable, alors que les autres co-variables restent inchangées.
- Puissance faible.
- Aucun "bon" moyen pour combiner les prédictions obtenues par les régressions séparées en une seule prédiction globale.
- Aucun moyen d'évaluer la proportion globale de la variabilité des réponses Y qui peut s'expliquer par les co-variables.

Modélisation conjointe

- Ces problèmes ne peuvent être résolus qu'en considérant un modèle conjoint de la relation entre y et l'ensemble des co-variables :

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i \quad \text{où} \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \text{ i.i.d.}$$

- L'approche du maximum de vraisemblance permet l'obtention de la valeur de β, connue sous le nom d'estimateur du maximum de vraisemblance (MLE) i.e., celui qui minimise la somme résiduelle des carrés (moindres carrés) :

$$||y - X\beta||^2$$

- Ici, $||v|| = \sqrt{\sum_i v_i^2}$ désigne la norme euclidienne. Nous utiliserons cette notation fréquemment tout au long de ce cours.

MCO=OLS comme Ordinary Least Square

- La solution est déterminée par le système linéaire, d'équations :

$$^tX\,X\widehat{\beta} = {}^tX\,y$$

- Etant donné que $^tX\,X$ est inversible, le système admet la solution unique :

$$\widehat{\beta} = (^tX\,X)^{-1}\,{}^tX\,y,$$

qui est connue sous le nom d'estimation des moindres carrés ordinaires (MCO ou OLS).

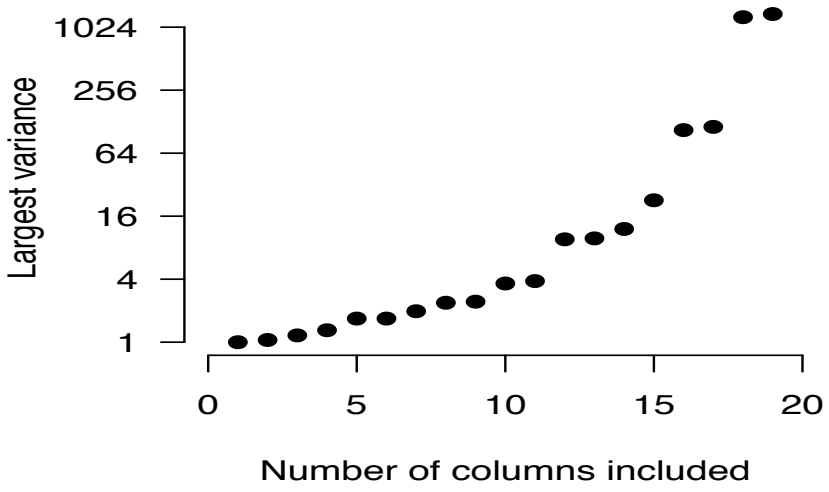
- L'estimation MCO résout tous les problèmes de la diapositive "Modèles uni-variés : limitations" et a de nombreux avantages bien reconnus tels qu'elle fournisse les meilleures estimations linéaires non biaisées de β.

Problèmes de MLE

- Cependant, il y a de nombreux inconvénients à utiliser le MLE pour estimer β quand p est grand.
- Plus dramatiquement, quand $p \geq n$ la matrice ${}^tX\,X$ n'est pas inversible, le MLE n'est pas unique.
- Cependant, même si ${}^tX\,X$ peut être inversée et un maximum unique identifié, comme p augmente et X approche la singularité, la surface de vraisemblance devient très plate.
- Cela signifie qu'une large gamme de valeurs de β est consistante avec les données, et de larges intervalles de confiance sont requis afin d'atteindre, disons, une confiance de 95%.

Un exemple vaut mieux qu'un long discours/que 100 démonstrations !

Considérons une matrice X avec $n = 20$ et dont les éléments sont constitués de nombres aléatoires indépendants et normalement distribués. La figure ci-dessous montre la plus grande variance des estimations $\widehat{\beta}_j$ en augmentant le nombre de colonnes de X :



quelques remarques

- Comme $p \to n$, $var(\widehat{\beta})$ croît sans limite (non bornée). L'augmentation devient substantielle lorsque p s'approche de n, et devient infinie lorsque $p \geq n$.

- Clairement, le MLE ne peut pas prendre en charge des données de grande dimension sans rencontrer de sérieux problèmes d'identifiabilité et d'inefficacité.

Modèle d'oracle

- Supposons, cependant, que de nombreuses co-variables ne sont pas liées à la réponse (dans le sens où $\beta_j = 0$), et seules quelques co-variables sont importantes.

- Si nous savions à l'avance quels éléments de β sont nuls/zéro et lesquels ne le sont pas, alors nous pourrions modifier le maximum de vraisemblance, sans l'abandonner complètement, et éviter tous les problèmes antérieurs.

- Plus précisément, nous pourrions appliquer le MLE uniquement aux variables pour lesquelles $\beta_j \neq 0$. Ceci est connu sous le nom de modèle d'oracle [2]

2. Les statistiques Oracle sont en fait la récupération de diverses informations concernant la volumétrie des tables, la distribution des différentes valeurs des champs indexés, la taille moyenne des tuples, ... Cet ensemble d'informations génèrera via un algorithme propre à Oracle un coût pour chaque plan d'exécution. En mode CBO (Cost Based Optimization), Oracle choisira, pour une requête donnée, le plan d'exécution le moins coûteux (le plus rapide). Ces différentes données (statistiques) sont stockées dans des tables du dictionnaire de données, et visibles sous les vues dba_tables, dba_indexes, ...

Sélection de modèle

- Évidemment, le modèle oracle est théorique, et non une approche réaliste de l'analyse de données, car il nécessiterait l'accès à un oracle qui pourrait nous dire quelles sont les co-variables qui sont liées à la réponse et lesquelles qui ne le sont pas.

- Dans le monde réel, nous devons utiliser les données afin de prendre des décisions empiriques sur les co-variables qui sont liées à la réponse et celles qui ne le sont pas. Ceci est connu sous le nom : la sélection du modèle.

Problème de sélection de modèle

- Malheureusement, l'utilisation des mêmes données à deux fins - pour sélectionner le modèle et aussi pour effectuer une inférence par rapport aux paramètres du modèle - introduit des biais substantiels et invalide les propriétés déductives que le MLE possède généralement.

- Pour illustrer ceci, considérons la simulation suivante :

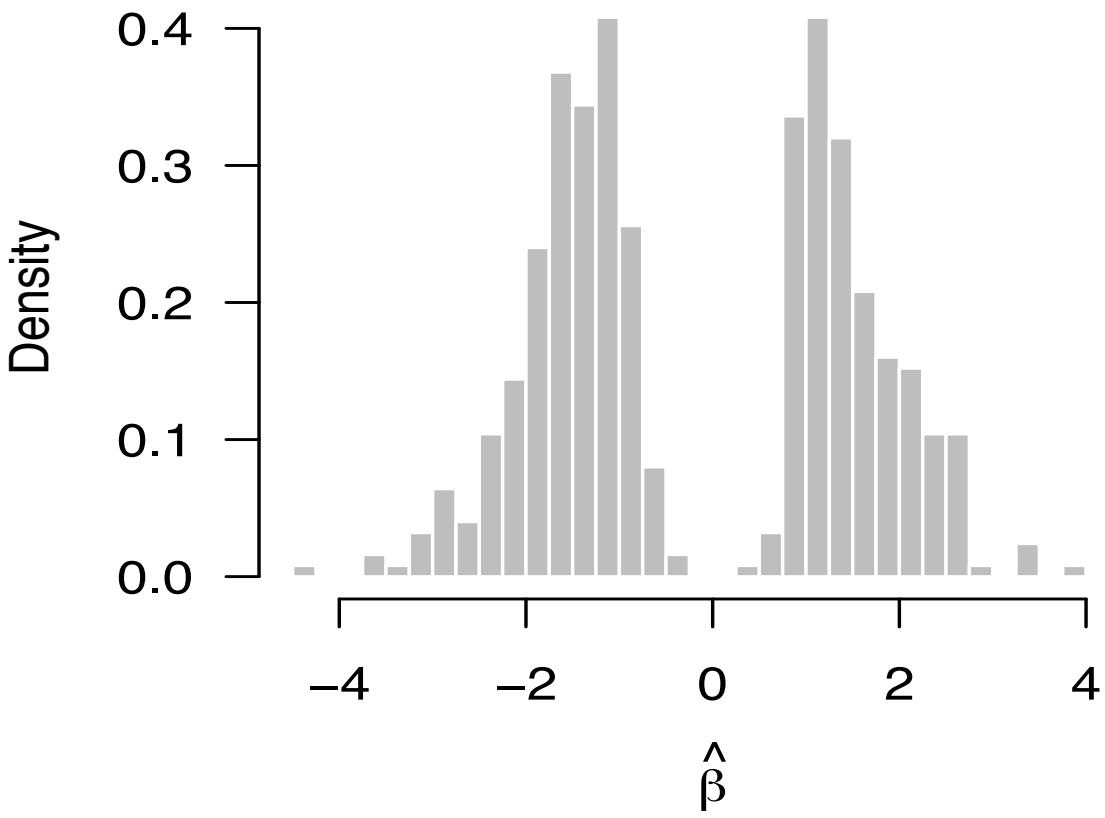
$$x_{ij} \overset{i.i.d.}{\sim} Unif(0,1) \text{ pour } j = 1, \ldots, 100$$

$$y_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1) \text{ pour } i = 1, \ldots, 25.$$

- Nous allons utiliser BIC pour sélectionner les 5 variables les plus importantes, puis utiliser MCO avec seulement ces variables, et répéter cela 100 fois.

Résultats

Un histogramme des 500 estimation $\widehat{\beta}_j$:



Remarques

- Comme nous le verrons, cette approche ne fonctionne pas bien.
- En utilisant l'ensemble des données pour la sélection du modèle ainsi que l'estimation et l'inférence, nous avons déformé grossièrement la distribution d'échantillonnage de $\widehat{\beta}$.
- Cela a des conséquences dramatiques en termes d'estimation, de prédiction, de sélection de variables et de validité de l'inférence

Estimation

- Le processus de sélection du modèle polarise fortement les estimations des coefficients de régression de zéro.
- Dans notre simulation, la plupart des estimations étaient d'environ ± 1.5 au lieu d'être proche de 0, qui est la vraie valeur.
- En particulier, l'EQM ou MSE moyenne est de 2.7, comparativement à 0.48 pour la régression marginale, soit 5 fois d'augmentation environ.
- Ce phénomène est parfois appelé la "malédiction des gagnants : winners' curse"

Sélection de variables

- Ici, nous avons imposé une limite supérieure de 5 sur le nombre de variables que nous avons permis d'être sélectionnées par le processus de sélection direct BIC-forward. Dans les 100 réplications, cette limite supérieure a été atteinte.

- De toute évidence, puisque le vrai modèle dans ce cas est le modèle nul (intercept seulement), le processus de sélection du modèle que nous avons employé ici entraîne un surapprentissage systématique.

- Alors qu'il est vrai que BIC choisira le vrai modèle avec une probabilité tendant vers 1, cet argument asymptotique repose sur le fait que p reste fixe alors que $n \to \infty$, ou en d'autres termes, sur le fait que $n >> p$.

- Clairement, BIC ne peut pas être invoqué pour une sélection précise des variables dans les problèmes de grande dimension.

La prédiction

- Pour évaluer la précision prédictive des modèles sélectionnés, nous calculons l'erreur quadratique moyenne de prédiction :

$$MSPE = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 ,$$

où pour la régression linéaire, $f(x_i) = {}^t x_i \widehat{\beta}$.

- En moyenne, les modèles sélectionnés ont obtenu une erreur quadratique moyenne de prédiction de 1.99, comparativement à une erreur de prédiction de $\sigma^2 = 1$ pour le modèle nul.

- Ainsi, en effectuant la sélection du modèle, nous avons réduit de moitié la précision prédictive du modèle (doubler son erreur)

Inférence

- Enfin, considérons la validité des inférences que nous obtenons du modèle de MCO après sélection :
  - La p-valeur médiane pour les tests $H_0 : \beta_j = 0$ était $p = 0.0013$
  - La confiance réelle obtenue en construisant des intervalles de confiance à 95% était inférieure à 5%

- Le fait de ne pas tenir compte des effets de la sélection lors de l'inférence post-sélection produit des conclusions beaucoup trop libérales, les erreurs réelles s'accumulant à un rythme beaucoup plus élevé que ne l'indiquent les approches inférentielles statistiques.

- En résumé, cette approche est extrêmement optimiste et trop confiante.

Remarques finales

- Ces problèmes sont largement reconnus. malheureusement, ils sont aussi largement ignorés.
- Le problème du développement de méthodes statistiques capables de sélection et d'inférence simultanées des variables a défié les statisticiens pendant des décennies, de Scheff (1953) à nos jours.
- L'un des principaux objectifs de ce cours est de démontrer dans quelle mesure les récents développements de la régression pénalisée corrigent et soulagent les préoccupations concernant la sélection et la déduction simultanées que nous avons soulevées aujourd'hui.