Introduction à la modélisation statistique

Résumé

Avant d'entrer dans le cœur de notre sujet, le modèle linéaire gaussien général, nous situons tout d'abord, dans ce chapitre d'introduction, la modélisation statistique au sein de la modélisation mathématique. Nous indiquons ensuite quelles sont les principales méthodes de modélisation statistique et nous précisons, parmi ces dernières, les méthodes traitées dans ce cours. Nous rappelons également les pré-traitements des données qui sont indispensables avant toute modélisation statistique. Enfin, nous donnons une formalisation plus mathématique de ce qu'est la modélisation statistique.

Retour au plan du cours

1 Notion de modélisation mathématique

Une grande partie des mathématiques appliquées consiste, d'une certaine façon, à faire de la modélisation, c'est-à-dire à définir un (ou plusieurs) modèle(s), de nature mathématique, permettant de rendre compte, d'une manière suffisamment générale, d'un phénomène donné, qu'il soit physique, biologique, économique ou autre.

De façon un peu schématique, on peut distinguer la modélisation déterministe (au sein d'un modèle déterministe, on ne prend pas en compte de variations aléatoires) et la modélisation stochastique (qui prend en compte ces variations aléatoires en essayant de leur associer une loi de probabilité).

Les outils classiques de la modélisation déterministe sont les équations différentielles ordinaires (EDO) et les équations aux dérivées partielles (EDP), qui prennent en compte les variations d'un phénomène en fonction de facteurs tels que le temps, la température... Ces équations ont rarement des solutions explicites et leur résolution nécessite, le plus souvent, la mise en œuvre d'algorithmes numériques plus ou moins sophistiqués, permettant d'obtenir une solution, éventuellement approchée. C'est le champ d'application de ce que l'on appelle aujourd'hui le calcul scientifique.

La modélisation stochastique a pour but essentiel de préciser des lois de probabilité rendant compte des variations aléatoires de certains phénomènes, variations dues à des causes soit inconnues, soit impossible à mesurer (par exemple, parce qu'elles sont à venir).

Au sein de la modélisation stochastique, la modélisation probabiliste a surtout pour but de donner un cadre formel permettant, d'une part de décrire les variations aléatoires dont il est question ci-dessus, d'autre part d'étudier les propriétés générales des phénomènes qui les régissent. Plus appliquée, la modélisation statistique consiste essentiellement à définir des outils appropriés pour modéliser des données observées, en tenant compte de leur nature aléatoire.

Il faut noter que le terme de modélisation statistique est très général et que, à la limite, toute démarche statistique en relève. Toutefois, ce qui est traité dans ce cours est relativement précis et constitue une partie spécifique de la modélisation statistique.

2 Principales méthodes de modélisation statistique

Les méthodes de modélisation statistique sont, en fait, très nombreuses. Nous citons ci-dessous les principales, sachant que la croissance considérable des masses de données enregistrées dans différents secteurs (internet, biologie à haut débit, marketing...), le besoin d'exploiter ces données sur le plan statistique, ainsi que les outils modernes de calcul ont donné naissance ces dernières années (disons depuis le début du XXI^{e} siècle) à de nombreuses méthodes, de plus en plus sophistiquées et, dans le même temps, de plus en plus "gourmandes" en temps calcul.

Dans les méthodes décrites ci-dessous, il y a presque toujours une variable privilégiée, en général appelée variable à expliquer, ou variable réponse, et notée Y (il s'agit d'une variable aléatoire). Le but est alors de construire un modèle permettant d'expliquer "au mieux" cette variable Y en fonction de variables explicatives observées sur le même échantillon.

Le modèle linéaire (gaussien) de base

À la fois le plus simple, le plus ancien et le plus connu des modèles statistiques, il englobe essentiellement la régression linéaire, l'analyse de variance et l'analyse de covariance. Dans ce modèle, les variables explicatives (régresseurs ou facteurs) ne sont pas aléatoires (elles sont à effets fixes). Pour pouvoir être exploité pleinement, ce modèle nécessite l'hypothèse de normalité des erreurs, donc de la variable à expliquer (hypothèse gaussienne). Ce modèle est présenté en détail dans le chapitre 2.

Le modèle linéaire généralisé

Il généralise le précédent à deux niveaux : d'une part, la loi des erreurs, donc de la variable réponse, n'est plus nécessairement gaussienne, mais doit appartenir à l'une des lois de la famille exponentielle ; d'autre part, la liaison linéaire entre l'espérance de la variable réponse et les variables explicatives se fait à travers une fonction particulière appelée fonction lien (spécifiée a priori). Ce modèle englobe différentes méthodes telles que la régression logistique, la régression Poisson, le modèle log-linéaire ou certains modèles de durée de vie.

Les modèles non linéaires

De façon très générale, il s'agit de modèles permettant d'expliquer la variable réponse (aléatoire) au moyen des variables explicatives (non aléatoires dans les modèles usuels), à travers une fonction quelconque, inconnue (on est donc en dehors du cadre du modèle linéaire généralisé). Cette classe de modèles est très vaste et relève, en général, de la statistique non paramétrique. Citons, à titre d'exemple, la régression non paramétrique, les *GAM (Generalized Additive Models)* et les réseaux de neurones.

Les modèles mixtes

On désigne sous ce terme des modèles permettant d'expliquer la variable aléatoire réponse au moyen de diverses variables explicatives, certaines étant aléatoires (on parle en général de facteurs à effets aléatoires) et intervenant dans la modélisation de la variance du modèle, d'autres ne l'étant pas (on parle de facteurs à effets fixes) et intervenant dans la modélisation de la moyenne. On trouve ainsi des modéles linéaires gaussiens mixtes, des modèles linéaires généralisés mixtes et des modèles non linéaires mixtes. Les premiers d'entres eux (les modéles linéaires gaussiens mixtes) seront introduits au chapitre 6 et utilisés encore au chapitre 7 de ce cours.

Les modèles pour données répétées

On appelle données répétées, ou données longitudinales, des données observées au cours du temps sur les mêmes individus (en général, il s'agit de personnes ou d'animaux suivis dans le cadre d'une expérimentation médicale ou biologique). De façon claire, il est nécessaire de prendre en compte dans ces modèles une certaine dépendance entre les observations faites sur un même individu à différents instants. Les modèles linéaires ou linéaires généralisés, qu'ils soient standards ou mixtes, sont utilisés dans ce contexte ; nous aborderons les modèles linéaires mixtes pour données répétées au chapitre 7.

Les modèles pour séries chronologiques

Les séries chronologiques sont les observations, au cours du temps, d'une certaine grandeur représentant un phénomène économique, social ou autre. Si données répétées et séries chronologiques ont en commun de rendre compte de l'évolution au cours du temps d'un phénomène donné, on notera que ces deux types de données ne sont pas réellement de même nature (dans une série chronologique, ce sont rarement des personnes ou des animaux que l'on observe). Pour les séries chronologiques, on utilise des modèles spécifiques : modèles AR (*Auto-Regressive*, ou auto-régressifs), MA (*Moving Average*, ou moyennes mobiles), ARMA, ARIMA (I pour *Integrated*)...

L'analyse discriminante et la classification

S'il est plus courant d'utiliser ces méthodes dans un contexte d'exploration des données plutôt que dans un contexte de modélisation, l'analyse discriminante et la classification peuvent tout de même être utilisées dans la phase de recherche d'un modèle permettant d'ajuster au mieux les données considérées. C'est en particulier le cas lorsque la variable réponse du modèle envisagé est de nature qualitative.

Les modèles par arbre binaire de régression et de classification

Ces méthodes (plus connues sous le nom de *CART*, pour *Classification And Regression Trees*) consistent à découper une population en deux parties, en fonction de celle des variables explicatives et du découpage en deux de l'en-

semble de ses valeurs ou modalités qui expliquent au mieux la variable réponse. On recommence ensuite sur chaque sous-population ainsi obtenue, ce qui permet de définir, de proche en proche, un arbre binaire et de classer les variables explicatives selon l'importance de leur liaison avec la variable réponse (on parle d'arbre de régression en présence d'une variable réponse quantitative et d'arbre de classification en présence d'une variable réponse qualitative). De telles méthodes peuvent constituer un complément intéressant au modèle linéaire ou au modèle linéaire généralisé.

Quelques autres modèles

Concernant les méthodes de modélisation statistique, on ne saurait être exhaustif dans cette introduction. Parmi les méthodes récentes, faisant un usage intensif de l'ordinateur, citons, pour mémoire, la régression *PLS (Partial Least Squares)*, les méthodes d'agrégation, ou de combinaison, de modèles (*bagging, boosting, random forests*), les méthodes de régularisation et les SVM (*Support Vector Machines*).

Dans ce cours, nous n'aborderons qu'un petit nombre de modèles parmi ceux évoqués ci-dessus. En fait, tous les modèles qui seront abordés relèvent du modèle linéaire gaussien : le modèle de base dans les chapitres 2 et 3 ; le cas particulier des plans d'expériences au chapitre 4 et celui de l'analyse de variance multidimensionnelle au chapitre 5 ; les modèles mixtes au chapitre 6 et les modèles pour données répétées au chapitre 7.

On trouvera d'intéressants développements sur d'autres modèles statistiques dans Saporta (2006) ainsi que dans cette vignette.

3 Préliminaires à toute modélisation statistique

Quel que soit le modèle, ou le type de modèles, envisagé face à un jeu de données, quel que soit le problème qu'il s'agit de traiter, une modélisation statistique ne peut sérieusement s'envisager que sur des données "propres", c'est à dire pré-traitées, afin de les débarasser, autant que faire se peut, de tout ce qui peut nuire à la modélisation : codes erronés, données manquantes, données aberrantes, variables inutiles, variables redondantes... C'est cet ensemble de pré-traitements que nous décrivons dans ce paragraphe.

On notera que cette phase est parfois appelée *datamanagement*, autrement dit "gestion des données".

3.1 "Nettoyage" des données

Avant toute chose, il faut disposer d'un fichier informatique contenant les données dans un format exploitable (texte ou excel, par exemple), les individus étant disposés en lignes et les variables en colonnes. Avec ce fichier, il faut essayer de repérer d'éventuels codes interdits ou aberrants : chaîne de caractères pour une variable numérique ; code "3" pour la variable sexe ; valeur 153 pour l'âge d'un groupe d'individus, etc. Une fois repérés, ces codes doivent être corrigés si possible, supprimés sinon.

Dans cette phase, il faut également essayer de repérer des données manquantes en grande quantité, soit sur une colonne (une variable), soit sur une ligne (un individu). Si quelques données manquantes ne sont pas vraiment génantes dans la plupart des traitements statistiques, il n'en va pas de même lorsque cela concerne un fort pourcentage des observations d'une variable ou d'un individu. Dans ce cas, il est préférable de supprimer la variable ou l'individu (dont la colonne, ou la ligne, serait, de toutes façons, inexploitable).

3.2 Analyses univariées

Cette phase, souvent fastidieuse, consiste à étudier chaque variable l'une après l'autre, afin d'en connaître les principales caractéristiques et d'en repérer, le cas échéant, certaines anomalies.

Pour les variables quantitatives, on pourra faire un histogramme ou un diagramme en boîte et déterminer des caractéristiques telles que le minimum, le maximum, la moyenne, l'écart-type, la médiane et les quartiles. Cela peut conduire à supprimer une variable (si elle présente très peu de variabilité), à la transformer (par exemple, en prenant son logarithme si elle est à valeurs positives et très dissymétrique), ou encore à repérer des valeurs très particulières (que l'on devra, éventuellement, corriger ou éliminer).

Pour les variables qualitatives, on pourra faire un diagramme en colonnes des modalités et déterminer les effectifs et les fréquences de ces dernières. Cela pourra encore conduire à supprimer une variable (si tous les individus,

ou presque, présentent la même modalité), ou à en regrouper des modalités "proches" (si certains effectifs sont trop faibles).

Ces analyses univariées permettent également de prendre connaissance des données et de fournir certaines indications pour la phase ultérieure de modélisation. Toutefois, il faut noter que ces analyses peuvent être inenvisageables avec des données "fortement multidimensionnelles", c'est-à-dire comportant des centaines, voire des milliers, de variables ; on rencontre aujourd'hui de telles données dans certains contextes particuliers.

3.3 Analyses bivariées

Ces analyses ont pour but d'étudier d'éventuelles liaisons existant entre couples de variables. Il peut s'agir de deux variables explicatives, dont on soupçonne qu'elles sont fortement corrélées, dans le but d'éliminer l'une des deux. Il peut aussi s'agir d'étudier les liens entre la variable à expliquer et chaque variable explicative (de façon systématique), pour avoir une première idée des variables explicatives susceptibles de jouer un rôle important lors de la modélisation. Enfin, ces analyses peuvent aussi permettre de repérer des points aberrants (ou extrêmes) qui n'ont pas pu l'être avec les analyses univariées.

Rappelons que, pour étudier la liaison entre deux variables quantitatives, on dispose, comme graphique, du nuage de points (ou diagramme de dispersion) et, comme indicateur de liaison, du coefficient de corrélation linéaire. Dans le cas d'une variable quantitative et d'une variable qualitative, on dispose du diagramme en boîtes parallèles et du rapport de corrélation. Enfin, dans le cas de deux variables qualitatives, on utilise en général un diagramme en colonnes de profils (profils-lignes ou profils-colonnes selon ce que l'on souhaite mettre en évidence) et des indicateurs de liaison liés au khi-deux (coefficients de Tschuprow ou de Cramér).

3.4 Analyses multivariées quantitatives

Elles consistent à déterminer la matrice des corrélations entre toutes les variables quantitatives considérées, notamment la variable à expliquer, lorsque celle-ci est quantitative. Cela peut permettre encore de supprimer des variables très corrélées, par exemple afin d'éviter de faire une régression sur de telles variables, dont on sait que les résultats seraient très instables, voire sans aucune signification. Cela permet aussi de prendre connaissance de la structure de corrélation entre les variables considérées, ce qui est toujours utile dans le cadre d'une modélisation.

On peut également envisager, à ce niveau, de réaliser une analyse en composantes principales (A.C.P.) de toutes ces variables, afin de préciser davantage, de façon globale, leurs relations linéaires.

3.5 Analyses multivariées qualitatives

C'est le pendant des analyses ci-dessus, cette fois pour les variables qualitatives. On peut, tout d'abord, déterminer la matrice des coefficients de Tschuprow (ou celle des coefficients de Cramér) et l'analyser comme une matrice de corrélations. Toutefois, il est bien connu que, dans la pratique, ces coefficients sont systématiquement petits : pratiquement toujours inférieurs à 0.5 et le plus souvent compris entre 0.1 et 0.3. Leur interprétation est donc, en général, assez délicate. Ils permettent néanmoins de repérer les liaisons les plus importantes, même si elles sont de l'ordre de 0.3, 0.4 ou 0.5.

Il est d'autant plus important d'envisager, dans ces analyses préliminaires, de réaliser une analyse des correspondances multiples (A.C.M.) entre variables qualitatives. Celle-ci permettra, le cas échéant, de confirmer une liaison forte entre certains couples de variables et, si nécessaire, d'en éliminer quelques-unes. L'A.C.M. permet également de regrouper certaines modalités d'une même variable lorsque celles-ci apparaissent proches dans l'ensemble des résultats et, par suite, de simplifier les données. Enfin, le tableau de Burt, fourni avec les résultats de l'A.C.M., permet de repérer des occurences très faibles pour certains croisements de modalités et d'envisager encore d'autres regroupements.

3.6 Bilan

Une fois réalisées toutes les étapes préliminaires décrites ci-dessus, on dispose de données "mises au propre", simplifiées, et dont on commence à connaître certaines caractéristiques. On peut, à partir de ce moment là, envisager leur modélisation.

Les modèles susceptibles d'être adaptés aux données considérées, parmi tous ceux décrits dans le paragraphe précédent, sont nécessairement limités à ce stade là. Ils sont fonction de la nature des données ainsi que des questions posées par l'utilisateur, autrement dit de ses objectifs.

Insistons ici sur le fait que des données sont toujours recueillies (produites) par un utilisateur (biologiste, informaticien, gestionnaire...) dans un but bien précis. La modélisation statistique doit avoir pour objectif premier de répondre aux questions que s'est posé cet utilisateur lorsqu'il a décidé de recueillir les données. Une collaboration entre utilisateur et statisticien est donc, à ce niveau là, absolument indispensable.

4 Formalisation de la notion de modèle statistique

Même si nous ne l'utilisons que fort peu dans la suite de ce cours, nous donnons, dans ce dernier paragraphe, une formalisation de ce qu'est un modèle statistique, afin de relier cette notion au formalisme habituellement utilisé en calcul des probabilités.

La notion de modèle statistique correspond à la modélisation d'une succession d'expériences aléatoires, chacune associée à une observation de l'échantillon considéré. Ainsi, considérons n variables aléatoires réelles (v.a.r.) Y_i, chacune associée à une expérience aléatoire dont le résultat est la valeur observée de Y_i (en fait, on suppose ici que l'expérience considérée est quantitative, par exemple le résultat d'une certaine mesure ; cela étant, ce qui suit se généralise sans difficulté au cas qualitatif).

On suppose donc, au départ, que les v.a.r. Y_i sont définies sur un certain espace probabilisé $(\Omega, \mathcal{A}, \Pi)$ et sont à valeurs dans $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$. Si l'on appelle Q la loi de probabilité conjointe des v.a.r. $(Y_1, \ldots, Y_n)$, soit encore la loi induite sur $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$ par $Y = (Y_1, \ldots, Y_n)$, alors le modèle statistique associé à l'expérience considérée est, par définition :

$$(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n}, Q).$$

C'est donc l'espace probabilisé qui va rendre compte des expériences aléatoires réalisées. Ainsi, préciser le modèle (faire des hypothèses...) reviendra à préciser la loi de probabilité Q.

La première hypothèse que l'on fait généralement dans la pratique est celle de l'**indépendance** des différentes expériences, autrement dit l'indépendance mutuelle des v.a.r. Y_i, $i = 1, \ldots, n$. Si l'on appelle P_i la loi de probabilité induite par Y_i sur $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$, le modèle statistique peut alors se mettre sous la forme suivante :

$$(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n}, \prod_{i=1}^{n} P_i).$$

On retiendra que c'est ce cadre général qui est celui du modèle linéaire et du modèle linéaire généralisé, l'hypothèse de linéarité concernant, dans les deux cas, la relation entre $\mathbb{E}(Y_i)$ et les variables explicatives.

Une autre hypothèse, souvent faite dans la pratique, est que les Y_i ont toutes la même loi de probabilité (elles sont **identiquement distribuées**). Dans ce cas, on a $P_i = P, \forall i = 1, \ldots, n$, et le modèle devient :

$$(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n}, P^n).$$

On a coutume de le noter $(\mathbb{R}, \mathcal{B}_{\mathbb{R}}, P)^{\otimes n}$ ou, plus simplement, $(\mathbb{R}, \mathcal{B}_{\mathbb{R}}, P)^n$. C'est ce qu'on appelle le **modèle d'échantillonnage** qui suppose les v.a.r. **indépendantes et identiquement distribuées** (i.i.d.). On notera que ce modèle ne peut servir de cadre au modèle linéaire que pour la loi des erreurs (les v.a.r. Y_i n'ont pas toutes, dans le modèle linéaire, la même espérance).

Dans la pratique, un modèle statistique n'est réellement opérationnel que si l'on précise la loi de probabilité P (cas i.i.d.) ou les lois P_i (cas seulement indépendant ; dans ce dernier cas, les P_i sont en général choisies dans une même famille de lois : normale, binomiale...). Après avoir ainsi précisé la loi de probabilité (ou la famille de lois de probabilité) du modèle, il reste d'abord à faire des tests, d'une part pour essayer de simplifier le modèle retenu, d'autre part pour tester la significativité de ce dernier, ensuite à en estimer les paramètres. C'est tout ce travail – choix de la loi de probabilité ou de la famille de lois, tests, choix du modèle, estimation des paramètres du modèle retenu, validation du modèle – qui constitue la modélisation statistique. ;