TP 11 (Contrôle continu n°2) : Simulation, estimation, model fitting, intervalles de confiance

Luca Ganassali

Bonjour à tous ! Aujourd'hui, même format, même concept : vous pouvez utiliser votre travail des TPs précédents, l'aide de R ainsi qu'internet.

Consignes

Le TP suivant vous est donné sous la forme d'un fichier .pdf, contenant le sujet, ainsi qu'un notebook qui tient lieu de "fiche réponse" : sur celui-ci, des zones sont laissées libres pour inclure votre code et vos réponses écrites. A la fin du contrôle, vous devrez enregistrer votre notebook complété par vos soins, le renommer sous la forme prenom_nom.Rmd, et me l'envoyer par mail à l'adresse suivante : `luca.ganassali@inria.fr`.

Bon courage ! :)

Exercice 1 : régression logistique, détection de profils et prédiction (env. 10 points) *Pour cet exercice, on rappelle que renormaliser un vecteur, au sens statistique, consiste à effectuer une opération sur ce vecteur pour qu'il devienne de moyenne nulle et d'écart-type* 1. Un laboratoire cherche à comprendre le risque de développer une pathologie X chez un patient en fonction de deux caractéristiques qu'il pensent pertinentes : l'âge a renormalisé du patient, et son degré d'exposition au soleil dans les six derniers mois, renormalisé, noté s. On a réalisé un sondage sur 20 individus qui recense ces caractéristiques (âge et exposition au soleil), ainsi qu'un variable Y qui vaut 1 si la personne souffre de la maladie X, et 0 sinon. On modélise la loi de Y de la façon suivante : pour un individu d'âge renormalisé a et d'exposition au soleil renormalisée s, on a

$$\mathbb{P}(Y = 1 \,|\, a, s) = \frac{e^{b_0+b_1 a+b_2 s}}{1 + e^{b_0+b_1 a+b_2 s}},$$

et logiquement,

$$\mathbb{P}(Y = 0 \,|\, a, s) = \frac{1}{1 + e^{b_0+b_1 a+b_2 s}}.$$

C'est le modèle de *régression logistique.*

1. Donner une interprétation des paramètres b_1 et b_2 de ce modèle. A quoi sert le paramètre b_0 ?

2. Ouvrir le fichier `data.csv` (qui se situe dans le même dossier que votre fiche réponse). Extraire les données `malade` (Y) et `age,exposition` qui correspondent, puis renormaliser les variables `age,exposition` pour créer les vecteurs a et s. Afficher tous les individus dans le plan a/s. Ces deux variables vous semblent-elles corrélées ? On pourra justifier graphiquement, par un argument de bon sens, et/ou par un calcul numérique sous R.

3. Il est aisé de montrer que la vraisemblance (dépendant des données $Y_1, a_1, s_1, \dots, Y_n, a_n, s_n$) s'écrit:

$$L(b_0, b_1, b_2; Y_1, a_1, s_1, \dots, Y_n, a_n, s_n) = \prod_{i=1}^{n} \left(\frac{e^{b_0+b_1 a_i+b_2 s_i}}{1 + e^{b_0+b_1 a_i+b_2 s_i}} \right)^{Y_i} \left(\frac{1}{1 + e^{b_0+b_1 a_i+b_2 s_i}} \right)^{1-Y_i},$$

qu'on peut réécrire sous la forme

$$L(b_0, b_1, b_2; Y_1, a_1, s_1, \ldots, Y_n, a_n, s_n) = \exp\left(\sum_{i=1}^{n} Y_i(b_0 + b_1 a_i + b_2 s_i) - \log(1 + e^{b_0+b_1 a_i + b_2 s_i})\right).$$

Ecrire une fonction `L(b0,b1,b2,Y,a,s)` qui prend en argument `b0,b1,b2`, les données Y, a, s (sous forme de vecteurs), et qui renvoie la vraisemblance associée.

4. On cherche à maximiser numériquement cette vraisemblance pour les données réelles observées, en `b0,b1,b2`. Pour ce faire on fait une 'grid search': on fait parcourir à b_0, b_1, b_2 l'intervalle $[-2, 2]$, discrétisé (découpé) en $k = 120$ morceaux de taille égale *(on utilisera `seq` en réglant le paramètre `length.out`)*. On calcule ainsi le maximum (approché) de cette vraisemblance, et on renvoie le maximum de vraisemblance (noté `MV`) ainsi que le triplet (b_0^*, b_1^*, b_2^*) de paramètres pour lequel il est atteint. Ecrire le code nécessaire à cette opération. *Le code met une dizaine de secondes à tourner, normalement.*

5. Interpréter les résultats : quel(s) est (sont) le(s) facteur(s) dominant(s) qui expliquent l'apparition de la maladie X chez un patient ?

6. Un patient a un âge renormalisé de 0 (c'est à dire un âge moyen) et présente une exposition au soleil renormalisée de 0.6. D'après nos estimations, quelle est la probabilité que le patient développe la maladie X?

Fin de l'exercice 1 : partie bonus. *Dans ces deux dernières questions, vous aurez peut-être besoin de nouvelles commandes... A vous de chercher, de toute façon ce n'est que du bonus !*

7. Au vu des résultats précédents, on admet que $b_0^* \sim -0.62$. On cherche juste à visualiser de façon plus fine la vraisemblance L à $b_0 = b_0^*$ fixé, en fonction des variables b_1 et b_2 variant autour de b_1^* et b_2^*. Pour ce faire, on va parcourir l'intervalle $[1, 3]$ pour b_1 et $[-1, 0]$ pour b_2, en découpant ces deux intervalles en $k = 300$ parts égales. On stockera dans des vecteurs `b1s90, b2s90` les valeurs de b_1 et b_2 atteignant au moins 90% du maximum de vraisemblance calculé précédemment. On fera de même avec `b1s95, b2s95` (pour 95%) et `b1s99, b2s99` (pour 99%). On répsentera ensuite dans le plan (b_1, b_2) les zones correspondant au seuils $90, 95$ et 99%. On tâchera de mettre des couleurs différentes. On pourra aussi afficher le point correspondant au maximum calculé précédemment.

8. Reprendre le graphique des individus dans le plan a/s, mais cette fois-ci en coloriant les points différemment selon que l'individu est sain ou malade. Quel résultat retrouve-t-on ?

Exercice 2 : la ruine du joueur. (env. 6 points) Dans cet exercice, on étudie le problème suivant : un joueur possède initialement $a > 0$ euros (a entier). A chaque partie, il mise 1 euro, et gagne 2 fois sa mise (donc $+1$ au total) avec probabilité p, ou perd sa mise (-1 au total) avec probabilité $q = 1 - p$. On supposera bien sûr que les parties sont indépendantes, on notera S_n la somme détenue par le joueur après la partie n. Par convention, $S_0 = a$.

On fixe un autre paramètre entier $b > 0$. Le jeu s'arrête après la partie T (ou au temps T), où T est défini par:

$$T := \inf \{n \in \mathbb{N},\ S_n \in \{0, a + b\}\}.$$

1. A quoi correspond T pour le joueur ? Quel nom donneriez-vous au paramètre b ? Que peut valoir S_T par définition ?

2. Dans cette question, on prendra $p = 0.3$, $a = 100$ et $b = 25$. Ecrire un code qui simule une trajectoire $S = (S_0, S_1, \ldots, S_T)$. *Attention, cette trajectoire devra être arrêtée au temps T. On pourra utiliser une boucle `while` et on pourra utiliser la commande `v = c(v, x)` pour ajouter un terme `x` à la fin d'un vecteur `v`.* Représenter graphiquement la trajectoire dans le temps.

Refaites tourner ce code plusieurs fois : qu'arrive-t-il au joueur avec ces paramètres ?

3. Refaites tourner le code précédent avec $a = 25$, $b = 25$ et $p = 0.5$, et plusieurs fois. Avec ces nouveaux paramètres, et au vu des trajectoires et de votre intuition, quelle est la probabilité que $S_T = 0$? Expliquer.

4. Ecrire une fonction `time(p,a,b)` qui simule et renvoie le temps T pour des paramètres p, a, b.

Utiliser cette fonction pour représenter un histogramme (convaincant) de T lorsque $p = 0.5$ et $a = b = 10$. Commenter *(bonus : pouvez-vous expliquer la valeur moyenne ?).*

Exercice 3 : do it yourself (env. 4 points) Cette semaine, $n_0 = 1074309$ personnes ont effectué un test pour savoir si elles portaient un virus précis. Parmi ces personnes, $S_0 = 66671$ étaient positives. La semaine précédente, $n_1 = 1041279$ avaient été testées et $S_1 = 56227$ étaient positives. L'augmentation du taux d'incidence (c'est-à-dire du taux de positifs) est-elle significative ?

Dans cet exercice, on attend deux intervalles de confiance jusitifiés, que l'on pourra calculer avec R. Toute présentation de résultat, même incomplète, sera comptabilisée.