Contrôle des connaissances du cours « Introduction aux méthodes de traitement des données » du DEA de Strasbourg.

Énoncé et Corrigé

Le mardi 14 janvier 2002 de 8h à 10h.

Le contrôle est noté sur 20, le barème est indiqué à la fin du problème. Le polycopié et les notes de cours sont les seuls documents autorisés.

Vous remarquerez que certaines questions sont démontrées dans le cours mais que l'exercice propose d'autres voies. Il ne s'agit donc pas pour vous de simplement recopier le cours mais de répondre suivant la méthode indiquée ici.

Élimination de points aberrants lors d'une régression.

Le but de cet exercice est de montrer comment il est possible d'éliminer des points qui influencent trop fortement le résultat d'une régression. On aborde donc ici le problème de l'ajustement fiable (robuste) d'un nuage de points par la méthode des moindres-carrés.

Pour cette étude, on se place dans le cadre de l'estimation de k paramètres β à partir de n observations $\boldsymbol{y}$ par la méthode des moindres-carrés. Selon cette méthode, dans le cas linéaire le modèle de régression peut s'écrire sous la forme matricielle :

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\,,$$

où :

- $\boldsymbol{y}$ est le vecteur $(n \times 1)$ des observations ;
- $\boldsymbol{X}$ est la matrice $(n \times k)$ du modèle linéaire ;
- $\boldsymbol{\epsilon}$ est le vecteur $(n \times 1)$ des erreurs ;
- enfin $\boldsymbol{\beta}$ est le vecteur $(k \times 1)$ des paramètres à estimer.

Notez que l'écriture $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ au lieu de $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$ sous entend généralement que les $\boldsymbol{\beta}$ sont en réalité des *fonctions à estimer* c'est-à-dire que la matrice $\boldsymbol{X}$ est de rang k. On continuera cependant à dire « les paramètres $\boldsymbol{\beta}$ » au lieu de dire « les fonctions à estimer $\boldsymbol{\beta}$ ». Cela signifie en pratique que vous n'avez pas à tenir compte d'une éventuelle matrice $\boldsymbol{C}$ des combinaisons linéaires des $\boldsymbol{\theta}$. Vous pouvez considérer que ce travail a déjà été fait et que $\boldsymbol{X}$ est effectivement de rang k.

① La méthode des moindres-carrés conduit à estimer les $\boldsymbol{y}$ par $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ et conduit à une estimation linéaire à partir des observations suivant la formule : $\widehat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$.
Donnez l'expression de la matrice $\boldsymbol{H}$.

② On rappelle que les opérations d'inversion et de transposition commutent, c'est-à-dire : $(\boldsymbol{A}^t)^{-1} = (\boldsymbol{A}^{-1})^t$, et que par ailleurs : $(\boldsymbol{A}\boldsymbol{B})^t = \boldsymbol{B}^t\boldsymbol{A}^t$.
Montrez alors que la matrice $\boldsymbol{H}$ est un projecteur orthogonal, c'est-à-dire qu'elle est idempotente ($\boldsymbol{H}^2 = \boldsymbol{H}$) et qu'elle est symétrique.

③ Si les h_{ij} désignent les éléments de $\boldsymbol{H}$, montrez que l'on peut écrire :

$$h_{ii} = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2 \,.$$

④ En déduire que les éléments diagonaux de $\boldsymbol{H}$ sont compris dans l'intervalle $[0, 1]$.

⑤ L'expression « $\operatorname{trace} \boldsymbol{A}$ » désigne la trace de la matrice $\boldsymbol{A}$, c'est-à-dire la somme de ces éléments diagonaux.

Sur la remarque que $\operatorname{trace}(\boldsymbol{AB}) = \operatorname{trace}(\boldsymbol{BA})$, en déduire que $\operatorname{trace} \boldsymbol{H} = k$. Autrement dit, montrez que la trace de la matrice $\boldsymbol{H}$ est égale aux nombre de paramètres à estimer.

(Indication : décomposez astucieusement la matrice $\boldsymbol{H}$ en un produit de deux matrices $\boldsymbol{AB}$.)

⑥ L'expression $\widehat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$ montre que l'élément diagonal h_{ii} mesure l'influence de l'observation y_i sur l'estimation $\widehat{y}_i$. En outre, la propriété $\operatorname{trace} \boldsymbol{H} = k$ montre que cette influence s'exerce au détriment des autres estimations $\widehat{y}_j$, $j \neq i$. La quantité h_{ii} s'appelle le *levier* du point i.

Bien qu'une étude objective soit relativement complexe, il est cependant assez simple d'établir un critère qui permette d'éliminer une mesure y_i sur la base que son levier h_{ii} est trop grand. Pouvez-vous établir un tel critère ? La question est évidemment : « trop grand par rapport à quoi ? »

⑦ Un point éliminé de la façon décrite ci-dessus est appelé un *point de levier*. L'objet des questions suivantes est d'illustrer la méthode à l'aide d'un exemple.

On dispose d'un ensemble de six mesures (x_i, y_i) données par le tableau suivant :

x	50	52	55	75	57	58
y	6	8	9	7	8	10

Les variables x_i sont connues alors que les y_i sont aléatoires et constituent le vecteur $\boldsymbol{y}$ évoqué dans cet exercice. On a donc :

$$\boldsymbol{y} = \begin{bmatrix} 6 \\ 8 \\ 9 \\ 7 \\ 8 \\ 10 \end{bmatrix} .$$

On se propose d'étudier la dépendance des y en fonction des x suivant le modèle affine : $y = \beta_1 + \beta_2 x$. S'agit-il d'un modèle linéaire ? Que vaut la matrice $\boldsymbol{X}$?

⑧ Reportez les points (x_i, y_i) dans un diagramme cartésien xOy et tracez à la main la droite qui passe au mieux dans ce nuage de points.

Tous calculs fait, la matrice $\boldsymbol{H}$ vaut :

$$\boldsymbol{H} = \begin{bmatrix} 0.33 & 0.28 & 0.22 & -0.17 & 0.18 & 0.16 \\ 0.28 & 0.25 & 0.21 & -0.08 & 0.18 & 0.16 \\ 0.22 & 0.21 & 0.19 & 0.04 & 0.17 & 0.17 \\ -0.17 & -0.08 & 0.04 & 0.90 & 0.13 & 0.17 \\ 0.18 & 0.18 & 0.17 & 0.13 & 0.17 & 0.17 \\ 0.16 & 0.16 & 0.17 & 0.17 & 0.17 & 0.17 \end{bmatrix} .$$

Cette matrice montre-t-elle l'existence d'un point de levier et si oui lequel ? Supprimez le point de levier éventuel et retracez la droite qui passe au mieux dans ce nouveau nuage de points. Conclure.

($1+1+1+2+1+2+1+1=10$ points.)

Corrigé.

❶ On a $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t$ d'où : $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t$.

❷ Les calculs sont immédiats.
Idempotence : $\boldsymbol{H}^2 = \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t = \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t$.
Symétrie : $\boldsymbol{H}^t = (\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t)^t = \boldsymbol{X}((\boldsymbol{X}^t\boldsymbol{X})^{-1})^t\boldsymbol{X}^t = \boldsymbol{X}((\boldsymbol{X}^t\boldsymbol{X})^t)^{-1}\boldsymbol{X}^t = \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t$.

❸ De $\boldsymbol{H}^2 = \boldsymbol{H}$ on tire : $h_{ii} = \sum_{l=1}^n h_{il}h_{li}$ et, par symétrie $h_{ii} = \sum_{l=1}^n h_{il}^2$. D'où la solution : $h_{ii} = h_{ii}^2 + \sum_{j\neq i} h_{ij}^2$.

❹ De $h_{ii} = h_{ii}^2 + \sum_{j\neq i} h_{ij}^2$ il vient $h_{ii} \geq h_{ii}^2$ et la solution demandée (la fonction $y = x$ n'est supérieure ou égale à $y = x^2$ que sur l'intervalle $[0, 1]$)

❺ Il suffit de poser $\boldsymbol{A} = \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}$ et $\boldsymbol{B} = \boldsymbol{X}^t$ et d'utiliser la formule. Il vient $\operatorname{trace}\boldsymbol{H} = \operatorname{trace}(\boldsymbol{X}^t\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}) = \operatorname{trace}\boldsymbol{I}$. La matrice $\boldsymbol{I}$ est la matrice identité au format $(k \times k)$, il s'ensuit $\operatorname{trace}\boldsymbol{I} = k$.

❻ En moyenne les h_{ii} valent k/n un point i sera un point de levier si le levier correspondant est significativement plus grand que cette valeur. Par significatif on peut entendre si $h_{ii} \geq 2k/n$ par exemple à la condition que $n \geq 2k$.

❼ Le modèle est affine en x mais linéaire en β. La matrice du modèle vaut :

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \\ 1 & x_5 \\ 1 & x_6 \end{bmatrix} = \begin{bmatrix} 1 & 50 \\ 1 & 52 \\ 1 & 55 \\ 1 & 75 \\ 1 & 57 \\ 1 & 58 \end{bmatrix}.$$

❽ Le point de levier est le point 4, son élimination rend le régression beaucoup plus significative comme illustré sur les figures 1 et 2. Il est probable que la saisie de la valeur x du point 4 est en cause, il fallait sans doute saisir 57 et non 75.

Bibliographie.

Pour en savoir plus sur ce sujet, vous pouvez consulter :

Huber, P.J. (1981), *Robust Statistics*, John Wiley & Sons, New York

Mosteller, F. and Tukey, J.W. (1977), *A Second Course in Statistics*, Addison-Wesley, Reading Mass.

Staudte, R.G. and Sheather, S.J. (1990), *Robust Estimation and Testing*, John Wiley & Sons, New York

Recherche d'une source gamma

On cherche à détecter un pulsar radio de période connue dans le domaine gamma. On dispose d'un détecteur qui compte les photons provenant d'une région du ciel, sans capacité d'imagerie. Au signal détecté, se superpose un bruit de fond important provenant en particulier de l'environnement radioactif du détecteur. Pour déterminer ce bruit, on vise une région du ciel qui ne comporte pas de source gamma pendant un temps t_{b}.

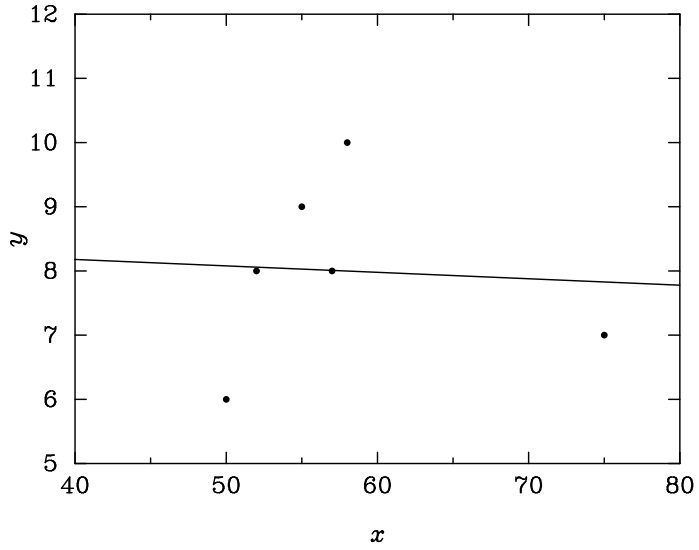

FIG. 1 – Ajustement d'une droite des moindres-carrés en tenant compte de tous les points.

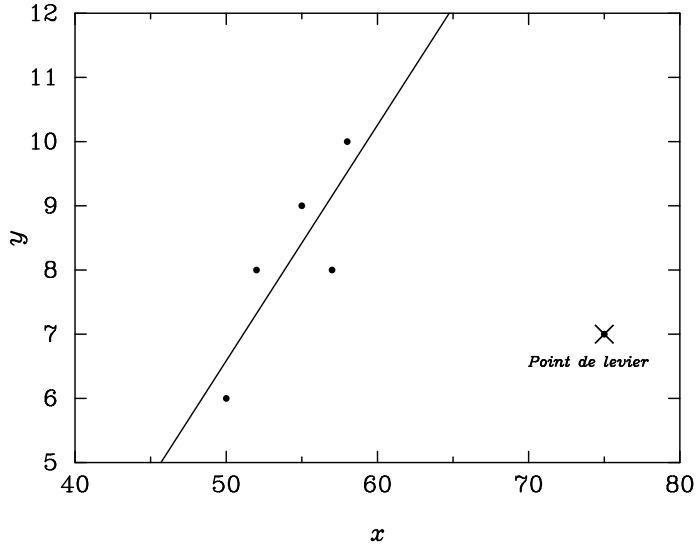

FIG. 2 – Ajustement d'une droite des moindres-carrés en supprimant le point de levier.

① Si on appelle Φ_b le flux correspondant au bruit de fond, Φ_s le flux de la source, et t_s le temps d'observation de la source, à quelles lois de probabilité obéissent les nombres de photons détectés lors de l'observation de la région du ciel ne contenant aucune source et celle contenant la source, N_b et N_s ?

② Quel doit être le rapport $t_\mathrm{s}/t_\mathrm{b}$ pour optimiser l'observation, c'est-à-dire obtenir un rapport signal sur bruit maximum, le temps total disponible $t_\mathrm{s} + t_\mathrm{b}$ étant fixé, et lorsque $\Phi_\mathrm{s} \ll \Phi_\mathrm{b}$?

③ Avec des temps d'observation égaux, on a obtenu $N_\mathrm{s} = 1243$ et $N_\mathrm{b} = 1180$. Peut-on affirmer que la source est détectée ?

④ On veut tenir compte du fait que la source est périodique. Pour ce faire, on calcule la phase des temps d'arrivée des photons détectés pendant l'observation de la source (la phase, temps de détection modulo la période, divisé par la période, est un nombre sans dimension compris entre 0 et 1). La période du pulsar est très petite devant le temps d'observation. Ceci permet de calculer une courbe de lumière, qu'on caractérise ici par le rapport du nombre de photons détectés a des phases inférieures et supérieures à 0.5. On mesure, dans les mêmes conditions que plus haut, un rapport de 3. Ce rapport est-il statistiquement significatif ? Quelle conclusion faut-il en tirer ?

($2 + 3 + 3 + 2$ points)

Corrigé.

❶ N_b et N_s suivent des lois de Poisson de moyenne $\Phi_\mathrm{b} t_\mathrm{b}$ et $(\Phi_\mathrm{s} + \Phi_\mathrm{b})t_\mathrm{s}$ respectivement

❷ On mesure la quantité $N_\mathrm{s} - N_\mathrm{b} t_\mathrm{s}/t_\mathrm{b}$ dont la variance est $(\Phi_\mathrm{s}+\Phi_\mathrm{b})t_\mathrm{s}+\Phi_\mathrm{b} t_\mathrm{b}(t_\mathrm{s}/t_\mathrm{b})^2$. Le carré du rapport signal sur bruit vaut alors :

$$\frac{S^2}{B^2} = \frac{\Phi_\mathrm{s}^2 t_\mathrm{s}^2}{(\Phi_\mathrm{s} + \Phi_\mathrm{b})t_\mathrm{s} + \Phi_\mathrm{b} t_\mathrm{b}(t_\mathrm{s}/t_\mathrm{b})^2} \tag{1}$$

Qu'on peut encore écrire, pour $\Phi_\mathrm{s} \ll \Phi_\mathrm{b}$:

$$\frac{S^2}{B^2} = \frac{\Phi_\mathrm{s}^2}{\Phi_\mathrm{b}} \frac{t_\mathrm{s}(T - t_\mathrm{s})}{T} \tag{2}$$

où T est la durée totale de l'observation. Cette grandeur est maximale pour $t_\mathrm{s} = T/2$, soit $t_\mathrm{s}/t_\mathrm{b}$ = 1.

❸ Le signal net est de $N_\mathrm{s} - N_\mathrm{b} = 63$, alors que l'écart type attendu est de $(2 \times 1180)^{1/2} = 48$, ce qui donne un rapport signal/bruit de 1.3, très insuffisant pour affirmer la réalité d'une détection

❹ Sur les 1243 photons détectés pendant l'observation de la source, 311 correspondent à une phase inférieure à 0.5 et 932 à une phase supérieure à 0.5. Cette différence est très significative ; en effet, dans l'hypothèse nulle, le nombre de photons de phase inférieure à 0.5 devrait suivre une loi binomiale de paramètres $p = 0.5$ et $N = 1243$. La variance de la loi binomiale est $p(1-p)N$, soit un écart type de 17.6, très petit devant l'écart entre les valeur mesurées (311) et attendues (621.5). Cet écart est d'ailleurs si important qu'il ne peut provenir du pulsar ; en effet, la mesure directe donne une limite supérieure à 3 sigmas du nombre de photons de la source de 144, plus faible que la différence entre les deux phases. Il faut donc conclure que la périodicité mesurée est d'origine instrumentale.