TP de Modélisation Stochastique n^0 7 : Modèle linéaire gaussien

On rappelle les résultats suivants:

1. **Régression simple:** On suppose que le modèle est: $y_i = a + bx_i + \varepsilon_i$ pour $i = 1, ..n$ et que les ε_i sont les réalisations de n v.a. i.i.d. $\mathcal{N}(0, \sigma^2)$. On note $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - \bar{x}^2)$ et $\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i y_i - \bar{x}\bar{y})$.

   (a) Les estimateurs par moindres carrés ordinaires (distance euclidienne) de a, b et σ^2 sont:
   $\hat{b} = \frac{\sigma_{xy}}{\sigma_x^2}$, $\quad \hat{a} = \bar{y} - \hat{b}\bar{x}$ et $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - (\hat{a} + \hat{b}x_i)^2$.

   (b) $\hat{a} \stackrel{\mathcal{L}}{=} \mathcal{N}\left(a, \frac{\sigma^2}{n}(1 + \frac{\bar{x}^2}{\sigma_x^2})\right)$, $\quad \hat{b} \stackrel{\mathcal{L}}{=} \mathcal{N}\left(b, \frac{\sigma^2}{n\sigma_x^2}\right)$, $\quad \hat{a} + \hat{b}x_0 \stackrel{\mathcal{L}}{=} \mathcal{N}\left(a + bx, \frac{\sigma^2}{n}(1 + \frac{(x_0 - \bar{x})^2}{\sigma_x^2})\right)$, ce qui donne des intervalles de confiance pour a et b et sur une prévision d'une valeur y_0 à partir de x_0 lorsque σ est connue.

   (c) $\frac{\sqrt{n}(\hat{a} - a)}{\hat{\sigma}\sqrt{1 + \frac{\bar{x}^2}{\sigma_x^2}}} \stackrel{\mathcal{L}}{=} t_{n-2}$, $\quad \frac{\sqrt{n}\sigma_x(\hat{b} - b)}{\hat{\sigma}} \stackrel{\mathcal{L}}{=} t_{n-2}$, $\quad \frac{\sqrt{n}(y_0 - (\hat{a} + \hat{b}x_0)}{\hat{\sigma}\sqrt{n + 1 + \frac{(x_0 - \bar{x})^2}{\sigma_x^2}}} \stackrel{\mathcal{L}}{=} t_{n-2}$ ce qui donne des intervalles de confiance pour a et b et sur une prévision d'une valeur y_0 à partir de x_0 lorsque σ est inconnue.

   (d) On utilise aussi $\frac{\sqrt{n}(\mathbb{E}(y_0 | x = x_0) - (\hat{a} + \hat{b}x_0)}{\hat{\sigma}\sqrt{1 + \frac{(x_0 - \bar{x})^2}{\sigma_x^2}}} \stackrel{\mathcal{L}}{=} t_{n-2}$, ce qui donne des intervalles de confiance sur la valeur moyenne de la prédiction y_0.

   (e) $\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \stackrel{\mathcal{L}}{=} \chi^2_{n-2}$, ce qui donne des intervalles de confiance pour σ^2.

   (f) $\frac{\sum_{i=1}^{n}(\bar{y} - (\hat{a} + \hat{b}x_i))^2}{\hat{\sigma}^2} \stackrel{\mathcal{L}}{=} F(1, n-2)$, ce qui permet de faire un test sur l'hypothèse $b = 0$, donc sur le bien-fondé de la régression.

2. **Régression multilinéaire:** On suppose que le modèle est: $y_i = b_0 1 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_p x_{ip} + \varepsilon_i$ pour $i = 1, ..n$ et que les ε_i sont les réalisations de n v.a. i.i.d. $\mathcal{N}(0, \sigma^2)$. On note:

$$X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ : & : & & : \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ : \\ y_n \end{pmatrix} \quad \text{et} \quad B = \begin{pmatrix} b_0 \\ : \\ b_p \end{pmatrix}.$$

   (a) Les estimateurs par moindres carrés ordinaires (distance euclidienne) de B et de σ^2 sont
   $\hat{B} = ({}^tX.X)^{-1}.{}^tX.Y$ et $\hat{\sigma}^2 = \frac{{}^t(Y - X.\hat{B}).(Y - X.\hat{B})}{n - p - 1}$.

   (b) $\hat{B} \stackrel{\mathcal{L}}{=} \mathcal{N}_{p+1}(B, ({}^tX.X)^{-1}.\sigma^2)$, $\quad X_0.\hat{B} \stackrel{\mathcal{L}}{=} \mathcal{N}(X_0.B, {}^tX_0.({}^tX.X)^{-1}.X_0.\sigma^2)$, ce qui donne des intervalles de confiance pour B et sur une prévision d'une valeur y_0 à partir de X_0 lorsque σ est connue.

   (c) En notant $\hat{b}_j$ la $(j+1)$-ième coordonnée de $\hat{B}$, et A_{jj} le terme (j, j) de la matrice A, on a:
   $\frac{\sqrt{n-p-1}(\hat{b}_j - b_j)}{\hat{\sigma}\sqrt{({}^tX.X)^{-1}_{jj}}} \stackrel{\mathcal{L}}{=} t_{n-p+1}$, $\quad \frac{y_0 - X_0.\hat{B}}{\hat{\sigma}\sqrt{1 + {}^tX_0.({}^tX.X)^{-1}.X_0}} \stackrel{\mathcal{L}}{=} t_{n-p+1}$ ce qui donne des intervalles de confiance pour b_j et sur une prévision d'une valeur y_0 à partir de X_0 lorsque σ est inconnue.

   (d) On utilise aussi $\frac{\mathbb{E}(y_0 | X = X_0) - X_0.\hat{B}}{\hat{\sigma}\sqrt{{}^tX_0.({}^tX.X)^{-1}.X_0}} \stackrel{\mathcal{L}}{=} t_{n-p+1}$, ce qui donne des intervalles de confiance sur la valeur moyenne de la prédiction y_0.

   (e) $\frac{(n-p+1)\hat{\sigma}^2}{\sigma^2} \stackrel{\mathcal{L}}{=} \chi^2_{n-p+1}$, ce qui donne des intervalles de confiance pour σ^2.

(f) On définit R^2 le coefficient de détermination de la régression: il quantifie le ratio de Y expliqué par la régression. R est le coefficient de corrélation multiple entre Y et les variables explicatives. On a:

$$R^2 = \frac{{}^t(\hat{B}.X - \bar{y}.1).(\hat{B}.X - \bar{y}.1)}{{}^t(Y - \bar{y}.1).(Y - \bar{y}.1)}.$$

(g) $\dfrac{\sum_{i=1}^n (\bar{y} - X.\hat{B})^2}{p.\hat{\sigma}^2} \stackrel{\mathcal{L}}{=} F(p, n-p-1)$, ce qui permet de faire un test sur l'hypothèse $b_1 = \ldots = b_p = 0$, donc sur le bien-fondé de la régression avec les p variables explicatives.

(h) Plus généralement, pour tester l'hypothèse H_0: $H.B = B_0$, où B_0 est un vecteur de $\mathbb{R}^{p+1}$ et H une matrice de projection de rang q, on utilise la statistique:

$$\frac{X.({}^tX.X)^{-1}.{}^tH.(H.({}^tX.X)^{-1}.{}^tH)^{-1}.(H.\hat{B} - B_0)}{q.\hat{\sigma}^2} \stackrel{\mathcal{L}}{=} F(q, n-p-1).$$

(i) Enfin, une méthode pratique d'observer s'il existe des valeurs "abérrantes" lors d'une régression et le tracé des résidus studentisés: $\dfrac{y_i - (\hat{B}.X)_i}{\hat{\sigma}\sqrt{1 - (X({}^tX.X)^{-1}.{}^tX)_{ii}}}$. On demande à ces résidus d'être compris entre -2 et $+2$.

Les données et outils de StixboxA111

StixboxA111 possède 10 bases de données empiriques. Elles sont contenues dans les fichiers $datas@$, ou $@ = 1, 2, .., 10$. Pour les utiliser, et obtenir une information sur leur provenance, on peut taper dans MATLAB: $x = getdata;$. Consultez ces données.
On trouve aussi plusieurs outils permettant d'étudier les modèles linéaires:

- *linreg* permet d'obtenir la plupart des éléments intéressants pour une régression linéaire **simple** ou une régression polynomiale et un tracé.
- *lsfit* permet d'obtenir la plupart des éléments intéressants pour une régression linéaire **multiple**.
- *lsfit* permet d'obtenir des informations suivant différents critères sur le nombre optimal de variables explicatives et les estimations des coefficients pour une régression linéaire **multiple**.
- *cmpmod* quantifie la différence entre un modèle et un sous-modèle pour une régression linéaire **multiple**.

Un exemple de régression

clf	
$clear$	
$heure = [73, 50, 128, 170, 87, 108, 130, 74, 88, 139];$	10 données d'heures
$taille = [20, 20, 60, 80, 40, 50, 60, 30, 40, 60];$	10 données de taille
$identify(taille, heure);$	permet de repérer avec la souris le numéro de la donnée Sortir en cliquant sur le bouton central de la souris
$[b, Ib, res, sig, Isig] = linreg(heure, taille);$	Trace la régression.
$disp('L''equation\ de\ regression\ est\ :\ ')$	
$'heure\ =\ '$	
$b(1)$	
$'+'$	
$b(2)$	
$'*taille'$	

Calculer l'estimateur de la variance des résidus. Faire un test sur la légitimité de cette régession. Tracer les résidus studentisés. Conclusions.

Exercices sur le modèle linéaire gaussien

1. On pose $n = 20$, $a = -2$, $b = 3$, $\sigma^2 = 0.1$. Choisir n valeurs $x_1, x_2, \ldots, x_n$ entre 0 et 1 (soit en les simulant, soit en les fixant arbitrairement). Générer $y_1, y_2, \ldots y_n$ avec le modèle suivant :

$$y_i = bx_i + a + \varepsilon_i$$

   où les ε_i sont les réalisations de n v.a. i.i.d $\mathcal{N}(0, \sigma^2)$. Représenter graphiquement le nuage de points formé par les (x_i, y_i). Calculer l'estimateur des moindres carrés de (a, b) et tracer la droite de régression ainsi obtenue. Représenter la droite $y = bx + a$. Estimer la variance de ε. Donner un intervalle de confiance approché de risque 5% pour a et b. Reprendre les questions précédentes avec différentes valeurs de n et de σ^2 (on pourra donner à n les valeurs 5, 10, 20, 50, 100 et à σ^2 les valeurs 0.01, 1, 10). Comparer les résultats obtenus.

2. On considère maintenant le modèle

$$y_i = b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + b_4 x_{i4} + \varepsilon_i$$

   Choisir les x_{ij} (en les simulant ou en les fixant arbitrairement) et simuler $y_1, y_2, \ldots, y_n$. Calculer l'estimateur des moindres carrés de $\boldsymbol{b} = (b_1, b_2, b_3, b_4)$ et estimer σ^2. Calculer la matrice de Variance-Covariance de $\boldsymbol{b}$ et en déduire des intervalles de confiance approchés pour chaque b_k. Représenter graphiquement les y_i et les $\hat{y}_i = \hat{b}_1 x_{i1} + \hat{b}_2 x_{i2} + \hat{b}_3 x_{i3} + \hat{b}_4 x_{i4}$

3. Faire des régessions sur les données fournies par StixboxA111.

4. Reprendre le premier exercice en supposant que les ε_i ne sont plus indépendants, mais: $\sigma_{ij} = \mathrm{cov}(\varepsilon_i, \varepsilon_j) = \dfrac{1}{((j-i)^2 + 1)}$ (le modèle reste gaussien). On note $\Sigma = (\sigma_{ij})$. Faire une régession. En considérant la projection Σ^{-1}-orthogonale au lieu de la projection I_n-orthogonale, retrouver les formules d'une régression par moindres carrés généralisés, et donner les estimation de (a, b). Conclusion. (voir par exemple, Saporta G., *Probabilités, analyse des données et statistique*, Technip, p. 375-401.)