MAP 433 : Introduction aux méthodes statistiques. Cours 9

M. Hoffmann

11 avril 2014

Aujourd'hui

1. Tests asymptotiques
2. Tests d'adéquation
   - Tests de Kolmogorov-Smirnov
   - Tests du χ^2
3. Compléments : p-valeur et liens entre tests et régions de confiance
4. Sélection de variables
   - Backward Stepwise Regression
5. Test du χ^2 d'indépendance

Le test de Wald : hypothèse nulle simple

- Situation la suite d'expériences $\left(\mathfrak{Z}^n, \mathcal{Z}^n, \{\mathbb{P}^n_\vartheta, \vartheta \in \Theta\}\right)$ est engendrée par l'observation Z^n, $\vartheta \in \Theta \subset \mathbb{R}$
- Objectif : Tester

$$H_0 : \vartheta = \vartheta_0 \quad \text{contre} \quad \vartheta \neq \vartheta_0.$$

- Hyopthèse : on dispose d'un estimateur $\widehat{\vartheta}_n$ asymptotiquement normal

$$\boxed{\sqrt{n}(\widehat{\vartheta}_n - \vartheta) \xrightarrow{d} \mathcal{N}\big(0, v(\vartheta)\big)}$$

en loi sous $\mathbb{P}^n_\vartheta$, $\forall \vartheta \in \Theta$, où $\vartheta \rightsquigarrow v(\vartheta) > 0$ est continue.

- Sous l'hypothèse (ici sous $\mathbb{P}^n_{\vartheta_0}$) on a la convergence

$$\sqrt{n}\frac{\widehat{\vartheta}_n - \vartheta_0}{\sqrt{v(\widehat{\vartheta}_n)}} \xrightarrow{d} \mathcal{N}(0, 1)$$

en loi sous $\mathbb{P}^n_{\vartheta_0}$.

Test de Wald (cont.)

- Remarque $\sqrt{v(\widehat{\vartheta}_n)} \leftrightarrow \sqrt{v(\vartheta_0)}$ ou d'autres choix encore...
- On a aussi

$$T_n = n\frac{(\widehat{\vartheta}_n - \vartheta_0)^2}{v(\widehat{\vartheta}_n)} \xrightarrow{d} \chi^2(1)$$

sous $\mathbb{P}^n_{\vartheta_0}$.

- Soit $q^{\chi^2}_{1-\alpha,1} > 0$ tel que si $U \sim \chi^2(1)$, on a $\mathbb{P}\left[U > q^{\chi^2}_{1-\alpha,1}\right] = \alpha$. On choisit la zone de rejet

$$\mathcal{R}_{n,\alpha} = \left\{ T_n \geq q^{\chi^2}_{1-\alpha,1} \right\}.$$

- Le test de zone de rejet $\mathcal{R}_{n,\alpha}$ s'appelle Test de Wald de l'hypothèse simple $\vartheta = \vartheta_0$ contre l'alternative $\vartheta \neq \vartheta_0$ basé sur $\widehat{\vartheta}_n$.

Propriétés du test de Wald

Proposition

Le test Wald de l'hypothèse simple $\vartheta = \vartheta_0$ contre l'alternative $\vartheta \neq \vartheta_0$ basé sur $\widehat{\vartheta}_n$ est

- *asymptotiquement de niveau α :*

$$\mathbb{P}^n_{\vartheta_0}\left[T_n \in \mathcal{R}_{n,\alpha}\right] \to \alpha.$$

- *convergent ou (consistant). Pour tout point $\vartheta \neq \vartheta_0$*

$$\mathbb{P}^n_{\vartheta}\left[T_n \notin \mathcal{R}_{n,\alpha}\right] \to 0.$$

Preuve

- Test asymptotiquement de niveau α par construction.
- Contrôle de l'erreur de seconde espèce : Soit $\vartheta \neq \vartheta_0$. On a

$$
\begin{aligned}
T_n &= \left(\sqrt{n}\frac{\widehat{\vartheta}_n - \vartheta}{\sqrt{v(\widehat{\vartheta}_n)}} + \sqrt{n}\frac{\vartheta - \vartheta_0}{\sqrt{v(\widehat{\vartheta}_n)}}\right)^2 \\
&=: T_{n,1} + T_{n,2}.
\end{aligned}
$$

  On a $T_{n,1} \xrightarrow{d} \mathcal{N}(0,1)$ sous $\mathbb{P}_\vartheta^n$ et

$$T_{n,2} \xrightarrow{\mathbb{P}_\vartheta^n} \pm\infty \ \text{ car } \ \vartheta \neq \vartheta_0$$

  Donc $T_n \xrightarrow{\mathbb{P}_\vartheta^n} +\infty$, d'où le résultat.
- Remarque : si $\vartheta \neq \vartheta_0$ mais $|\vartheta - \vartheta_0| \lesssim 1/\sqrt{n}$, le raisonnement ne s'applique pas. Résultat non uniforme en le paramètre.

Test de Wald : hypothèse nulle composite

- Même contexte : $\Theta \subset \mathbb{R}^d$ et on dispose d'un estimateur $\widehat{\vartheta}_n$ asymptotiquement normal :

$$\sqrt{n}\big(\widehat{\vartheta}_n - \vartheta\big) \xrightarrow{d} \mathcal{N}\big(0, V(\vartheta)\big)$$

où $V(\vartheta)$ est définie positive et continue en ϑ.

- But Tester $H_0 : \vartheta \in \Theta_0$ contre $H_1 : \vartheta \notin \Theta_0$, où

$$\boxed{\Theta_0 = \big\{\vartheta \in \Theta,\ \ g(\vartheta) = 0\big\}}$$

et

$$g : \mathbb{R}^d \to \mathbb{R}^m$$

$(m \leq d)$ est régulière.

Test de Wald cont.

- Hypothèse : la différentielle (de matrice $J_g(\vartheta)$) de g est de rang maximal m en tout point de (l'intérieur) de Θ_0.

Proposition

En tout point ϑ de l'intérieur de Θ_0 (i.e. sous l'hypothèse), on a, en loi sous $\mathbb{P}^n_\vartheta$:

- $$\sqrt{n}g(\widehat{\vartheta}_n) \xrightarrow{d} \mathcal{N}\big(0, J_g(\vartheta)V(\vartheta)J_g(\vartheta)^T\big),$$
- $$T_n = ng(\widehat{\vartheta}_n)^T\Sigma_g(\widehat{\vartheta}_n)^{-1}g(\widehat{\vartheta}_n) \xrightarrow{d} \chi^2(m)$$

  où $\Sigma_g(\vartheta) = J_g(\vartheta)V(\vartheta)J_g(\vartheta)^T$.

- Preuve : méthode « delta » multidimensionnelle.

Test de Wald (fin)

Proposition

Sous les hypothèses précédentes, le test de zone de rejet

$$\mathcal{R}_\alpha = \left\{ T_n \geq q^{\chi^2}_{1-\alpha,m} \right\}$$

avec $\mathbb{P}\left[U > q^{\chi^2}_{1-\alpha,m}\right] = \alpha$ *si* $U \sim \chi^2(m)$ *est*

- *Asymptotiquement de niveau* α *en tout point* ϑ *de (l'intérieur) de* Θ_0 *:*

$$\mathbb{P}^n_\vartheta\left[T_n \in \mathcal{R}_{n,\alpha}\right] \to \alpha.$$

- *Convergent : pour tout* $\vartheta \notin \Theta_0$ *on a*

$$\mathbb{P}^n_\vartheta\left[T_n \notin \mathcal{R}_{n,\alpha}\right] \to 0.$$

Tests d'adéquation

- Situation On observe (pour simplifier) un n-échantillon de loi F inconnu
$$X_1, \ldots, X_n \sim_{\text{i.i.d.}} F$$
- Objectif Tester
$$H_0 : F = F_0 \;\text{ contre }\; F \neq F_0$$
où F_0 distribution donnée. Par exemple : F_0 gaussienne centrée réduite.
- Il est très facile de construire un test asymptotiquement de niveau α. Il suffit de trouver une statistique $\phi(X_1, \ldots, X_n)$ de loi connue sous l'hypothèse.

Test d'adéquation : situation

- Exemples : sous l'hypothèse

$$\phi_1(X_1\ldots,X_n) = \sqrt{n}\overline{X}_n \sim \mathcal{N}(0,1)$$

$$\phi_2(X_1,\ldots,X_n) = \sqrt{n}\frac{\overline{X}_n}{s_n} \sim \text{Student}(n-1)$$

$$\phi_3(X_1,\ldots,X_n) = (n-1)s_n^2 \sim \chi^2(n-1).$$

- Le problème est que ces tests ont une faible puissance : ils ne sont pas consistants.
- Pas exemple, si $F \neq$ gaussienne mais $\int_{\mathbb{R}} x dF(x) = 0$, $\int_{\mathbb{R}} x^2 dF(x) = 1$, alors

$$\mathbb{P}_F\left[\phi_1(X_1,\ldots,X_n) \leq x\right] \rightarrow \int_{-\infty}^{x} e^{-u^2/2}\frac{du}{\sqrt{2\pi}},\ \ x \in \mathbb{R}\,.$$

(résultats analogues pour ϕ_2 et ϕ_3).
- La statistique de test ϕ_i ne caractérise pas la loi F_0.

Test de Kolmogorov-Smirnov

- Rappel Si la fonction de répartition F est continue,

$$\sqrt{n}\sup_{x\in\mathbb{R}}\left|\widehat{F}_n(x)-F(x)\right| \xrightarrow{d} \mathbb{B}$$

où la loi de $\mathbb{B}$ ne dépend pas de F.

Proposition (Test de Kolmogorov-Smirnov)

Soit $q_{1-\alpha}^{\mathbb{B}}$ tel que $\mathbb{P}\left[\mathbb{B} > q_{1-\alpha}^{\mathbb{B}}\right] = \alpha$. Le test défini par la zone de rejet

$$\mathcal{R}_{n,\alpha} = \left\{\sqrt{n}\sup_{x\in\mathbb{R}}\left|\widehat{F}_n(x)-F_0(x)\right| \geq q_{1-\alpha}^{\mathbb{B}}\right|\right\}$$

est asymptotiquement de niveau α : $\mathbb{P}_{F_0}\left[\widehat{F}_n \in \mathcal{R}_{n,\alpha}\right] \to \alpha$ et consistant :

$$\forall F \neq F_0 : \mathbb{P}_F\left[\widehat{F}_n \notin \mathcal{R}_{n,\alpha}\right] \to 0.$$

Test du Chi-deux

- X variables qualitative : $X \in \{1, \ldots, d\}$.

$$\mathbb{P}\left[X = \ell\right] = p_\ell,\ \ell = 1, \ldots d.$$

- La loi de X est caratérisée par $\mathbf{p} = (p_1, \ldots, p_d)^T$.
- Notation

$$\mathcal{M}_d = \left\{\mathbf{p} = (p_1, \ldots, p_d)^T,\ \ 0 \leq p_\ell, \sum_{\ell=1}^{d} p_\ell = 1\right\}.$$

- Objectif $\mathbf{q} \in \mathcal{M}_d$ donnée. A partir d'un n-échantillon

$$X_1, \ldots, X_n \sim_{\text{i.i.d.}} \mathbf{p},$$

tester $H_0 : \mathbf{p} = \mathbf{q}$ contre $H_1 : \mathbf{p} \neq \mathbf{q}$.

Construction « naturelle » d'un test

- Comparaison des fréquences empiriques

$$\widehat{p}_{n,\ell} = \frac{1}{n}\sum_{i=1}^{n} 1_{X_i=\ell} \quad \text{proche de} \quad q_\ell, \quad \ell = 1,\ldots,d\ ?$$

- Loi des grands nombres :

$$\left(\widehat{p}_{n,1},\ldots,\widehat{p}_{n,d}\right) \xrightarrow{\mathbb{P}_{\mathbf{p}}} (p_1,\ldots,p_d) = \mathbf{p}.$$

- Théorème central-limite ?

$$\mathbf{U}_n(\mathbf{p}) = \sqrt{n}\Big(\frac{\widehat{p}_{n,1}-p_1}{\sqrt{p_1}},\ldots,\frac{\widehat{p}_{n,d}-p_d}{\sqrt{p_d}}\Big) \xrightarrow{d} ?$$

- Composante par composante oui. Convergence globale plus délicate.

Statistique du Chi-deux

Proposition

Si les composantes de $\mathbf{p}$ *sont toute non-nulles*

- *On a la convergence en loi sous* $\mathbb{P}_{\mathbf{p}}$

$$\mathbf{U}_n(\mathbf{p}) \xrightarrow{d} \mathcal{N}\big(0, V(\mathbf{p})\big)$$

avec $V(\mathbf{p}) = \mathrm{Id}_d - \sqrt{\mathbf{p}}\big(\sqrt{\mathbf{p}}\big)^T$ *et* $\sqrt{\mathbf{p}} = \big(\sqrt{p_1}, \ldots, \sqrt{p_d}\big)^T$.

- *De plus*

$$\|\mathbf{U}_n(\mathbf{p})\|^2 = n \sum_{\ell=1}^{d} \frac{(\widehat{p}_{n,\ell} - p_\ell)^2}{p_\ell} \xrightarrow{d} \chi^2(d-1).$$

Preuve de la normalité asymptotique

- Pour $i = 1, \ldots, n$ et $1 \leq \ell \leq d$, on pose

$$Y_\ell^i = \frac{1}{\sqrt{p_\ell}}\left(1_{\{X_i=\ell\}} - p_\ell\right).$$

- Les vecteurs $\mathbf{Y}_i = (Y_1^i, \ldots, Y_d^i)$ sont indépendants et identiquement distribués et

$$\mathbf{U}_n(\mathbf{p}) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{Y}_i,$$

$$\mathbb{E}\left[Y_\ell^i\right] = 0,\ \mathbb{E}\left[(Y_\ell^i)^2\right] = 1 - p_\ell,\ \mathbb{E}\left[Y_\ell^i Y_{\ell'}^i\right] = -(p_\ell p_{\ell'})^{1/2}.$$

- On applique le TCL vectoriel.

Convergence de la norme au carré

- On a donc $\mathbf{U}_n(\mathbf{p}) \xrightarrow{d} \mathcal{N}\big(0, V(\mathbf{p})\big)$.
- On a aussi

$$\begin{aligned}\|\mathbf{U}_n(\mathbf{p})\|^2 &\xrightarrow{d} \|\mathcal{N}\big(0, V(\mathbf{p})\big)\|^2 \\ &\sim \chi^2\big(\mathrm{Rang}\big(V(\mathbf{p})\big)\big)\end{aligned}$$

par Cochran : $V(\mathbf{p}) = \mathrm{Id}_d - \sqrt{\mathbf{p}}\big(\sqrt{\mathbf{p}}\big)^T$ est la projection orthogonale sur $\mathrm{vect}\{\sqrt{\mathbf{p}}\}^\perp$ qui est de dimension $d-1$.

Test d'adéquation du χ^2

- « distance » du χ^2 :

$$\chi^2(\mathbf{p},\mathbf{q}) = \sum_{\ell=1}^{d} \frac{(p_\ell - q_\ell)^2}{q_\ell}.$$

- Avec ces notations $\|\mathbf{U}_n(\mathbf{p})\|^2 = n\chi^2(\widehat{\mathbf{p}}_n,\mathbf{p})$.

Proposition

Pour $\mathbf{q} \in \mathcal{M}_d$ le test simple défini par la zone de rejet

$$\mathcal{R}_{n,\alpha} = \left\{ n\chi^2(\widehat{\mathbf{p}}_n,\mathbf{q}) \geq q^{\chi^2}_{1-\alpha,d-1} \right\}$$

où $\mathbb{P}\left[U > q^{\chi^2}_{1-\alpha,d-1}\right] = \alpha$ si $U \sim \chi^2(d-1)$ est asymptotiquement de niveau α et consistant pour tester

$$H_0 : \mathbf{p} = \mathbf{q} \quad \text{contre} \quad H_1 : \mathbf{p} \neq \mathbf{q}.$$

Exemple de mise en oeuvre : expérience de Mendel

- Soit $d = 4$ et
$$\mathbf{q} = \Big(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}\Big).$$
- Répartition observée : $n = 556$
$$\widehat{\mathbf{p}}_{556} = \frac{1}{556}(315, 101, 108, 32).$$
- Calcul de la statistique du χ^2
$$556 \times \chi^2(\widehat{\mathbf{p}}_{556}, \mathbf{q}) = 0,47.$$
- On a $q_{95\%,3} = 0,7815$.
- Conclusion : Puisque $0,47 < 0,7815$, on accepte l'hypothèse $\mathbf{p} = \mathbf{q}$ au niveau $\alpha = 5\%$.

p-valeurs

- Exemple : on observe

$$X_1, \dots, X_n \sim_{\text{i.i.d.}} \mathcal{N}(\mu, \sigma^2), \quad \sigma^2 \ \text{connu.}$$

- Objectif : tester $H_0 : \mu = 0$ contre $H_1 : \mu \neq 0$.
- Au niveau $\alpha = 5\%$, on rejette si

$$|\overline{X}_n| > \frac{\phi^{-1}(1-\alpha/2)}{\sqrt{n}}$$

- Application numérique : $n = 100$, $\overline{X}_{100} = 0.307$. On a $\frac{\phi^{-1}(1-0.05/2)}{\sqrt{100}} \approx 0.196$. on rejette l'hypothèse....

p-valeur (cont.)

- Et pour un autre choix de α ?. Pour $\alpha = 0.01$, on a $\frac{\phi^{-1}(1-0.01/2)}{\sqrt{100}} \approx 0.256$. On rejette toujours... Pour $\alpha = 0.001$, on a $\frac{\phi^{-1}(1-0.001/2)}{\sqrt{100}} \approx 0.329$. On accepte H_0 !
- Que penser de cette petite expérience ?
  - En pratique, on a une observation une bonne fois pour toute (ici 0.307) et on « choisit » α... comment ?
  - On ne veut pas α trop grand (trop de risque), mais en prenant α de plus en plus petit... on va fatalement finir par accepter H_0 !
- Défaut de méthodologie inhérent au principe de Neyman (contrôle de l'erreur de première espèce).

p-valeur

- Quantité significative : non par le niveau α, mais le seuil de basculement de décision : c'est la p-valeur (p-value) du test.

Définition

Soit $\mathcal{R}_\alpha$ une famille de zones de rejet d'un test de niveau α pour une hypothèse H_0 contre une alternative H_1. Soit Z l'observation associée à l'expérience. On a $Z \in \mathfrak{Z}$ et $\mathcal{R}_0 = \mathfrak{Z}$. On appelle p-valeur du test la quantité

$$p - valeur(Z) = \inf\{\alpha, Z \in \mathcal{R}_\alpha\}.$$

Interprétation de la p-valeur

- Une grande valeur de la p-valeur s'interprète en faveur de ne pas vouloir rejeter l'hypothèse.
- « Ne pas vouloir rejeter l'hypothèse » peut signifier deux choses :
  - L'hypothèse est vraie
  - L'hypothèse est fausse mais le test n'est pas puissant (erreur de seconde espèce grande).
- Souvent : la p-valeur est la probabilité (sous H_0) que la statistique de test d'une expérience « copie » soit $\geq$ à la statistique de test observée.
- Exemple du test du χ^2 et de l'expérience de Mendel

Expérience de Mendel et p-valeur

- Sous l'hypothèse H_0

$$556 \cdot \chi^2(\widehat{\mathbf{p}}_{556}, \mathbf{q}) \sim \chi^2(3).$$

- Les données fournissent $556 \cdot \chi^2(\widehat{\mathbf{p}}_{556}, \mathbf{q}) = 0.47$ et $q^{\chi^2}_{1-0.05,3} = 0.7815$. On accepte l'hypothèse.
- Calcul de la p-valeur : pour $Z \sim \chi^2(3)$

$$p - \text{valeur} = \mathbb{P}_{\mathbf{q}}\left[Z > 0.47\right] = 0.93.$$

La « pratique » invite à ne pas rejeter H_0.

Régression linéaire multiple (=Modèle linéaire)

- La fonction de régression est $r(\vartheta, \mathbf{x}_i) = \vartheta^T \mathbf{x}_i$. On observe

$$(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_n, Y_n)$$

  avec

$$\boxed{Y_i = \vartheta^T \mathbf{x}_i + \xi_i, \quad i = 1, \ldots, n}$$

  où $\vartheta \in \Theta = \mathbb{R}^k, \quad \mathbf{x}_i \in \mathbb{R}^k$.

- Matriciellement

$$\boxed{\mathbf{Y} = \mathbb{M}\vartheta + \boldsymbol{\xi}}$$

  avec $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^T$ et $\mathbb{M}$ la matrice $(n \times k)$ dont les lignes sont les $\mathbf{x}_i$.

EMC en régression linéaire multiple

- Estimateur des moindres carrés en régression linéaire multiple : tout estimateur $\widehat{\vartheta}_n^{\,\mathtt{mc}}$ satisfaisant

$$\sum_{i=1}^n \big(Y_i - (\widehat{\vartheta}_n^{\,\mathtt{mc}})^T \mathbf{x}_i\big)^2 = \min_{\vartheta \in \mathbb{R}^k} \sum_{i=1}^n \big(Y_i - \vartheta^T \mathbf{x}_i\big)^2.$$

- En notations matricielles :

$$\begin{aligned} \|\mathbf{Y} - \mathbb{M}\,\widehat{\vartheta}_n^{\,\mathtt{mc}}\|^2 &= \min_{\vartheta \in \mathbb{R}^k} \|\mathbf{Y} - \mathbb{M}\,\vartheta\|^2 \\ &= \min_{v \in V} \|\mathbf{Y} - v\|^2 \end{aligned}$$

où $V = \mathrm{Im}(\mathbb{M}) = \{v \in \mathbb{R}^n : v = \mathbb{M}\,\vartheta,\ \vartheta \in \mathbb{R}^k\}$.
Projection orthogonale sur V.

Géométrie de l'EMC

- L'EMC vérifie

$$\boxed{\mathbb{M}\,\widehat{\vartheta}_{\mathsf{n}}^{\,\mathtt{mc}} = P_V \mathbf{Y}}$$

où P_V est le projecteur orthogonal sur V.

- Mais $\mathbb{M}^T P_V = \mathbb{M}^T P_V^T = (P_V \mathbb{M})^T = \mathbb{M}^T$. On en déduit les équations normales des moindres carrés :

$$\boxed{\mathbb{M}^T \mathbb{M}\,\widehat{\vartheta}_{\mathsf{n}}^{\,\mathtt{mc}} = \mathbb{M}^T \mathbf{Y}.}$$

Proposition

Si $\mathbb{M}^T \mathbb{M}$ (matrice $k \times k$) inversible, alors $\widehat{\vartheta}_{\mathsf{n}}^{\,\mathtt{mc}}$ est unique et

$$\boxed{\widehat{\vartheta}_{\mathsf{n}}^{\,\mathtt{mc}} = \left(\mathbb{M}^T \mathbb{M}\right)^{-1} \mathbb{M}^T \mathbf{Y}}$$

Exemple de données de régression

Données de diabète

Patient	age	sex	bmi	map	tc	ldl	hdl	tch	ltg	glu	Response
1	59	2	32.1	101	157	93.2	38	4	4.9	87	151
2	48	1	21.6	87	183	103.2	70	3	3.9	69	75
3	72	2	30.5	93	156	93.6	41	4	4.7	85	141
4	24	1	25.3	84	198	131.4	40	5	4.9	89	206
5	50	1	23.0	101	192	125.4	52	4	4.3	80	135
6	23	1	22.6	89	139	64.8	61	2	4.2	68	97
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
441	36	1	30.0	95	201	125.2	42	5	5.1	82	220
442	36	1	19.6	71	250	132.2	97	3	4.6	92	57

n=442, k=10

bmi = Body Mass Index

map = Blood Pressure

tc, ldl, tch, ltg, glu = Blood Serum Measurements

Response Y = a quantitative measure of disease progression 1 year after baseline

Résultats de traitement statistique initial

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	152.133	2.576	59.061	$< 2e-16 * **$
age	-10.012	59.749	-0.168	0.867000
sex	-239.819	61.222	-3.917	$0.000104 * **$
bmi	519.840	66.534	7.813	$4.30e-14 * **$
map	324.390	65.422	4.958	$1.02e-06 * **$
tc	-792.184	416.684	-1.901	0.057947
ldl	476.746	339.035	1.406	0.160389
hdl	101.045	212.533	0.475	0.634721
tch	177.064	161.476	1.097	0.273456
ltg	751.279	171.902	4.370	$1.56e-05 * **$
glu	67.625	65.984	1.025	0.305998

Propriétés de l'EMC : cadre gaussien

- Lois des coordonnées de $\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}}$:

$$(\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}})_j - \vartheta_j \sim \mathcal{N}(0, \sigma^2 b_j)$$

où b_j est le $j$ème élément diagonal de $\left(\mathbb{M}^T\,\mathbb{M}\right)^{-1}$.

$$\frac{(\widehat{\vartheta}_{\mathrm{n}}^{\,\mathtt{mc}})_j - \vartheta_j}{\widehat{\sigma}_n \sqrt{b_j}} \sim t_{n-k}$$

loi de Student à $n-k$ degrés de liberté.

$$t_q = \frac{\xi}{\sqrt{\eta/q}}$$

où $q \geq 1$ un entier, $\xi \sim \mathcal{N}(0,1)$, $\eta \sim \chi^2(q)$ et ξ indépendant de η.

Exemple de données de régression

Données de diabète

Patient	age	sex	bmi	map	tc	ldl	hdl	tch	ltg	glu	Response
1	59	2	32.1	101	157	93.2	38	4	4.9	87	151
2	48	1	21.6	87	183	103.2	70	3	3.9	69	75
3	72	2	30.5	93	156	93.6	41	4	4.7	85	141
4	24	1	25.3	84	198	131.4	40	5	4.9	89	206
5	50	1	23.0	101	192	125.4	52	4	4.3	80	135
6	23	1	22.6	89	139	64.8	61	2	4.2	68	97
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
441	36	1	30.0	95	201	125.2	42	5	5.1	82	220
442	36	1	19.6	71	250	132.2	97	3	4.6	92	57

n=442, k=10

bmi = Body Mass Index

map = Blood Pressure

tc, ldl, tch, ltg, glu = Blood Serum Measurements

Response Y = a quantitative measure of disease progression 1 year after baseline

Résultats de traitement statistique initial

	Estimate	Std. Error	t value	$\Pr(>\lvert t \rvert)$
(Intercept)	152.133	2.576	59.061	$< 2e-16 ***$
age	-10.012	59.749	-0.168	0.867000
sex	-239.819	61.222	-3.917	$0.000104 ***$
bmi	519.840	66.534	7.813	$4.30e-14 ***$
map	324.390	65.422	4.958	$1.02e-06 ***$
tc	-792.184	416.684	-1.901	0.057947
ldl	476.746	339.035	1.406	0.160389
hdl	101.045	212.533	0.475	0.634721
tch	177.064	161.476	1.097	0.273456
ltg	751.279	171.902	4.370	$1.56e-05 ***$
glu	67.625	65.984	1.025	0.305998

Questions statistiques

- **Sélection de variables.** Lesquelles parmi les 10 variables :

  `age,sex,bmi,map,tc,ldl,hdl,tch,ltg,glu`

  sont significatives ? Formalisation mathématique : trouver (estimer) l'ensemble $N = \{j : \vartheta_j \neq 0\}$.
- **Prévison.** Un nouveau patient arrive avec son vecteur des 10 variables $\mathbf{x}_0 \in \mathbb{R}^{10}$. Donner la prévison de la réponse Y $=$état du patient dans 1 an.

RSS (Residual Sum of Squares)

Modèle de régression

$$Y_i = r(\vartheta, \mathbf{x}_i) + \xi_i, \quad i = 1, \ldots, n.$$

- **Résidu :** si $\widehat{\vartheta}_n$ est un estimateur de ϑ,

$$\widehat{\xi}_i = Y_i - r(\widehat{\vartheta}_n, \mathbf{x}_i) \ \text{résidu au point } i.$$

- **RSS :** Residual Sum of Squares, somme résiduelle des carrés. Caractérise la qualité d'approximation.

$$\mathrm{RSS}(= \mathrm{RSS}_{\widehat{\vartheta}_n}) = \|\widehat{\xi}\|^2 = \sum_{i=1}^{n} \big(Y_i - r(\widehat{\vartheta}_n, \mathbf{x}_i)\big)^2.$$

- En régression linéaire : $\boxed{\mathrm{RSS} = \|\mathbf{Y} - \mathbb{M}\,\widehat{\vartheta}_n\|^2.}$

Sélection de variables : Backward Stepwise Regression

- On se donne un critère d'élimination de variables (plusieurs choix de critère possibles...).
- On élimine une variable, la moins significative du point de vue du critère choisi.
- On calcule l'EMC $\widehat{\vartheta}^{\mathrm{mc}}_{n,k-1}$ dans le nouveau modèle, avec seulement les $k-1$ paramétres restants, ainsi que le RSS :
$$\mathrm{RSS}_{k-1} = \|\mathbf{Y} - \mathbb{M}\,\widehat{\vartheta}^{\mathrm{mc}}_{n,k-1}\|^2.$$
- On continue à éliminer des variables, une par une, jusqu'à la stabilisation de RSS : $\mathrm{RSS}_m \approx \mathrm{RSS}_{m-1}$.

Données de diabète : Backward Regression

- **Sélection "naïve"** : $\{$`sex,bmi,map,ltg`$\}$
- **Sélection par Backward Regression** :
  Critère d'élimination : plus grande valeur de $\Pr(>|t|)$.

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
(Intercept)	152.133	2.576	59.061	$< 2e-16 ***$
age	-10.012	59.749	-0.168	0.867000
sex	-239.819	61.222	-3.917	$0.000104 ***$
bmi	519.840	66.534	7.813	$4.30e-14 ***$
map	324.390	65.422	4.958	$1.02e-06 ***$
tc	-792.184	416.684	-1.901	0.057947
ldl	476.746	339.035	1.406	0.160389
hdl	101.045	212.533	0.475	0.634721
tch	177.064	161.476	1.097	0.273456
ltg	751.279	171.902	4.370	$1.56e-05 ***$
glu	67.625	65.984	1.025	0.305998

Données de diabète : Backward Regression

Backward Regression : Itération 2.

Critère d'élimination : plus grande valeur de $\Pr(>|t|)$.

	Estimate	Std. Error	t value	$\Pr(>\lvert t\rvert)$
(Intercept)	152.133	2.573	59.128	$< 2e-16$
sex	-240.835	60.853	-3.958	0.000104
bmi	519.905	64.156	5.024	$8.85e-05$
map	322.306	65.422	4.958	$7.43e-07$
tc	-790.896	416.144	-1.901	0.058
ldl	474.377	338.358	1.402	0.162
hdl	99.718	212.146	0.470	0.639
tch	177.458	161.277	1.100	0.272
ltg	749.506	171.383	4.373	$1.54e-05$
glu	67.170	65.336	1.013	0.312

Données de diabète : Backward Regression

Backward Regression : Itération 5 (dernière).

Variables sélectionnées :

{sex,bmi,map,tc,ldl,ltg}

	Estimate	Std. Error	t value	$\Pr(>\lvert t\rvert)$
(Intercept)	152.133	2.572	59.159	$< 2e-16$
sex	-226.511	59.857	-3.784	0.000176
bmi	529.873	65.620	8.075	$6.69e-15$
map	327.220	62.693	5.219	$2.79e-07$
tc	-757.938	160.435	-4.724	$3.12e-06$
ldl	538.586	146.738	3.670	0.000272
ltg	804.192	80.173	10.031	$< 2e-16$

Sélection de variables : Backward Regression

Discussion de `Backward Regression` :

- Méthode de sélection purement empirique, pas de justification théorique.
- Application d'autres critères d'élimination en `Backward Regression` peut amener aux résultats différents.
  Exemple. Critère C_p de Mallows–Akaike : on élimine la variable j qui réalise

$$\min_j \left(\mathrm{RSS}_{m,(-j)} + 2\widehat{\sigma}_n^2 m\right).$$

Lien tests et régions de confiance

- $\mathcal{E} = \big(\mathfrak{Z}, \mathcal{Z}, \{\mathbb{P}_\vartheta, \vartheta \in \Theta\}\big)$, expérience statistique engendrée par l'observation Z avec $\Theta \subset \mathbb{R}^d$,.

Définition

Une région de confiance de niveau $1-\alpha$ pour $\vartheta \in \Theta$ est un sous-ensemble $\mathcal{C}_\alpha(Z)$ de $\mathbb{R}^d$ tel que

$$\forall \vartheta \in \Theta, \;\; \mathbb{P}_\vartheta\left[\vartheta \in \mathcal{C}_\alpha(Z)\right] \geq 1-\alpha.$$

Dualité tests – régions de confiance

Proposition

- *Si, pour tout $\vartheta_0 \in \Theta$, il existe un test de zone de rejet $\mathcal{R}_\alpha(\vartheta_0)$ pour tester $H_0 : \vartheta = \vartheta_0$ contre $\vartheta \neq \vartheta$, alors*

$$\mathcal{C}_\alpha(Z) := \left\{\vartheta \in \Theta,\ Z \in \mathcal{R}_\alpha^c\right\}$$

est une région de confiance pour ϑ de niveau $1-\alpha$.

- *Si $\mathcal{C}_\alpha(Z)$ est une région de confiance de niveau $1-\alpha$ pour $\vartheta \in \Theta$, alors le test défini par la région critique*

$$\mathcal{R}_\alpha := \left\{\vartheta_0 \in \mathcal{C}_\alpha^c\right\}$$

est de niveau α pour tester $H_0 : \vartheta = \vartheta_0$ contre $H_1 : \vartheta \neq \vartheta_0$.

Tests du χ^2

- Adéquation à une loi discrète (finie).
- Test du χ^2 avec paramètres estimés.
- Test d'indépendance.

Lois discrète finies

- X variables qualitative : $X \in \{1, \ldots, d\}$.

$$\mathbb{P}\big[X = \ell\big] = p_\ell,\ \ell = 1, \ldots d.$$

- La loi de X est caractérisée par $\mathbf{p} = (p_1, \ldots, p_d)^T$.
- Notation

$$\mathcal{M}_d = \Big\{\mathbf{p} = (p_1, \ldots, p_d)^T,\ \ 0 \leq p_\ell \leq 1, \sum_{\ell=1}^{d} p_\ell = 1\Big\}.$$

- Objectif $\mathbf{q} \in \mathcal{M}_d$ donnée. A partir d'un n-échantillon

$$X_1, \ldots, X_n \sim_{\text{i.i.d.}} \mathbf{p},$$

tester $H_0 : \mathbf{p} = \mathbf{q}$ contre $H_1 : \mathbf{p} \neq \mathbf{q}$.

Rappel : Test d'adéquation du χ^2

- « distance » du χ^2 :

$$\chi^2(\mathbf{p},\mathbf{q}) = \sum_{\ell=1}^{d} \frac{(p_\ell - q_\ell)^2}{q_\ell}.$$

Proposition

Pour $\mathbf{q} \in \mathcal{M}_d$ le test simple défini par la zone de rejet

$$\mathcal{R}_{n,\alpha} = \left\{ n\chi^2(\widehat{\mathbf{p}}_n, \mathbf{q}) \geq q^{\chi^2}_{1-\alpha,d-1} \right\}$$

où $q^{\chi^2}_{1-\alpha,d-1} > 0$ est défini par $\mathbb{P}\left[U > q^{\chi^2}_{1-\alpha,d-1}\right] = \alpha$ si $U \sim \chi^2(d-1)$ est asymptotiquement de niveau α et consistant pour tester

$$H_0 : \mathbf{p} = \mathbf{q} \quad \textit{contre} \quad H_1 : \mathbf{p} \neq \mathbf{q}.$$

Test du χ^2 avec paramètres estimés

- On observe $X_1, \ldots, X_n \sim_{\text{i.i ;d.}} \mathbf{p} \in \mathcal{M}_d$.
- On teste

$$H_0 : \mathbf{p} \in \left(\mathcal{M}_d\right)_0 \text{ contre } \mathbf{p} \in \mathcal{M}_d \setminus \left(\mathcal{M}_d\right)_0,$$

où la famille

$$\left(\mathcal{M}_d\right)_0 = \left\{\mathbf{p} = \mathbf{p}(\gamma),\ \gamma \in \Gamma\right\}$$

est régulière et $\Gamma \subset \mathbb{R}^d$ est « régulier » et de dimension $m < d - 1$.

EMV et paramètres estimés

Proposition

On a les estimateurs du maximum de vraisemblance suivants :

- *Pour la famille* $\mathcal{M}_d$ *: les fréquences empiriques*

$$\widehat{\mathbf{p}}_n^{\mathtt{mv}} = \big(\widehat{p}_{n,1}, \ldots, \widehat{p}_{n,d}\big)^T$$

- *Pour la famille restreinte* $\big(\mathcal{M}_d\big)_0$ *:*

$$\mathbf{p}(\widehat{\gamma}_n^{\mathtt{mv}}) = \arg\max_{\gamma \in \Gamma} \sum_{\ell=1}^{d} \widehat{p}_{n,\ell} \log p_\ell(\gamma).$$

- *Sous des hypothèses de régularité on a la convergence*

$$n\chi^2\big(\widehat{\mathbf{p}}_n^{\mathtt{mv}}, \mathbf{p}(\widehat{\gamma}_n^{\mathtt{mv}})\big) \xrightarrow{d} \chi^2(d - m - 1).$$

Application au test d'indépendance du χ^2

- On observe

$$(X_1, Y_1), \ldots, (X_n, Y_n) \sim_{\text{i.i.d.}} \mathbf{p} \in \mathcal{M}_{d_1,d_2}$$

où

$$\mathcal{M}_{d_1,d_2} = \big\{\mathbf{p} = \text{proba. sur } \{1, \ldots d_1\} \times \{1, \ldots, d_2\}\big\}.$$

- Objectif : tester l'indépendance entre X et Y, c'est-à-dire $\mathbf{p} = (p_{\ell,\ell'})$ de la forme

$$\boxed{p_{\ell,\ell'} = p_{\ell,\bullet}\, p_{\bullet,\ell'}}$$

où

$$p_{\ell,\bullet} = \sum_{\ell'=1}^{d_2} p_{\ell,\ell'}, \quad p_{\bullet,\ell'} = \sum_{\ell=1}^{d_1} p_{\ell,\ell'}.$$

EMV sur l'hypothèse nulle

- On note

$$\left(\mathcal{M}_{d_1,d_2}\right)_0 = \left\{\mathbf{p} = (p_{\ell,\ell'}),\ p_{\ell,\ell'} = p_{\ell,\bullet}p_{\bullet\ell'}\right\}.$$

Proposition

- $\left(\mathcal{M}_{d_1,d_2}\right)_0$ *est en correspondance avec* $\left\{\mathbf{p} = \mathbf{p}(\gamma),\ \gamma \in \Gamma\right\}$ $\Gamma \subset \mathbb{R}^m$ *de dimension* $m = d_1 + d_2 - 2$.
- *L'estimateur du maximum de vraisemblance restreint à* $\left(\mathcal{M}_{d_1,d_2}\right)_0$ *vaut*

$$\left(\widehat{p}_{n,0}^{\,\mathtt{mv}}\right)_{\ell,\ell'} = \frac{1}{n}\sum_{i=1}^{n} 1_{X_i=\ell} \times \frac{1}{n}\sum_{i=1}^{n} 1_{Y_i=\ell'}$$

i.e. le produit des fréquences empiriques.

Conclusion : test du χ^2 d'indépendance

- Objectif : Tester

$$H_0 : \mathbf{p} \in \big(\mathcal{M}_{d_1,d_2}\big)_0 \;\text{ contre }\; H_1 : \mathbf{p} \in \mathcal{M}_{d_1,d_2} \setminus \big(\mathcal{M}_{d_1,d_2}\big)_0.$$

- Sous l'hypothèse, on a la convergence

$$n\chi^2\big(\widehat{\mathbf{p}}_n^{\,\mathtt{mv}}, \widehat{\mathbf{p}}_{n,0}^{\,\mathtt{mv}}\big) \xrightarrow{d} \chi^2\big((d_1-1)(d_2-1)\big).$$

- En particulier, la statistique de test s'écrit

$$\boxed{n\chi^2\big(\widehat{\mathbf{p}}_n^{\,\mathtt{mv}}, \widehat{\mathbf{p}}_{n,0}^{\,\mathtt{mv}}\big) = n\sum_{\ell,\ell'} \frac{\Big((\widehat{\mathbf{p}}_n)_{\ell,\ell'} - \widehat{p}_{n,(\ell,\bullet)}\widehat{p}_{n,(\bullet,\ell')}\Big)^2}{\widehat{p}_{n,(\ell,\bullet)}\widehat{p}_{n,(\bullet,\ell')}}}$$