Auto-concordance pour la vraisemblance empirique

Art B. Owen, Stanford University

Motivation

Dylan Small et Dan Yang (2012) ont trouv un cas o mon ancienne itération Levenberg-Marquardt a échoué. Une simple réduction d'itration fonctionne mieux.

La nouvelle optimisation est

1) de faible dimension
2) convexe
3) sans contrainte
4) **auto-concordante**

Le nouvel ingrédient est l'auto-concordance (décrite ci-dessous).

Cela donne des garanties mathématiques de convergence.

Avant la convergence, ça nous permet de borner la sous-optimalité.

Également

Une log-vraisemblance quartique (Corcoran 1998) est aussi auto-concordante.

Vraisemblance empirique

Permet des inférences de vraisemblance sans supposer une famille paramétrique.

Pour des données $X_i \overset{\text{iid}}{\sim} F$

$$L(F) = \prod_{i=1}^{n} F(\{X_i\}) \qquad \text{Vraisemblance}$$

$$\hat{F} = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i} \qquad \text{EMVNP}$$

$$R(F) = \prod_{i=1}^{n} n w_i, \quad w_i \equiv F(\{X_i\}) \qquad \text{Rapport de vraisemblance empirique}$$

Si $L(F) > 0$ alors $w_i > 0$. Il est aussi pratique de supposer que $\sum_{i=1}^{n} w_i = 1$ aussi.

Nous obtenons alors une distribution multinomiale sur n éléments $X_1, \ldots, X_n$.

Propriétés de la VE

La vraisemblance empirique hérite de plusieurs propriétés des vraisemblances paramétriques.

- Distribution limite χ^2 du style Wilks
- Sélection de forme automatique pour les régions de confiance
- Correction Bartlett DiCiccio, Hall, Romano
- Très grande puissance Kitamura et Lazar & Mykland

Suppositions statistiques: indépendance et moments bornés.

Curieusement

Avoir $n-1$ paramètres pour n observations n'entraine pas de difficultés.

Vraisemblance empirique pour la moyenne

$$\mathcal{R}(\mu) = \max\Big\{\prod_{i=1}^{n} nw_i \mid w_i > 0 \sum_{i=1}^{n} w_i X_i = \mu, \sum_{i=1}^{n} w_i = 1\Big\}$$

Semblable à Wilks: $-2\log(\mathcal{R}(\mu_0)) \xrightarrow{\mathrm{d}} \chi^2_{(d)}$ permet des tests et régions de confiance

Équations estimantes $\mathbb{E}(m(X,\theta)) = 0$

$m(X,\theta) = X - \theta$	moyenne
$m(X,\theta) = 1_{X<\theta} - 0.5$	médiane
$m(X,Y,\theta) = (Y - X^{\mathsf{T}}\theta)X$	régression
$m(X,\theta) = \frac{\partial}{\partial\theta}\log(f(X,\theta))$	estimateur EMV

Calculs

Maximiser $\sum_{i=1}^{n} \log(nw_i)$ où $\sum_i w_i = 1$ et $\sum_i w_i Z_i = 0$

Ici $Z_i = X_i - \mu_0$ ou $Z_i = m(X_i, \theta)$.

L'enveloppe

Si 0 n'est pas dans l'enveloppe convexe de Z_i alors $\log(\mathcal{R}(\cdot)) = -\infty$

Lagrangien

$$G = \sum_{i=1}^{n} \log(nw_i) - n\lambda^\mathsf{T} \sum_{i=1}^{n} w_i Z_i + \delta\left(\sum_{i=1}^{n} w_i - 1\right)$$

$$\frac{\partial G}{\partial w_i} = \frac{1}{w_i} - n\lambda^\mathsf{T} Z_i + \delta$$

$$0 = \sum_{i=1}^{n} w_i \frac{\partial G}{\partial w_i} = n - 0 + \delta$$

Donc

$$w_i = \frac{1}{n} \frac{1}{1 + \lambda^\mathsf{T} Z_i}, \quad \text{pour } \lambda \in \mathbb{R}^d$$

Trouver λ

$$w_i = \frac{1}{n}\frac{1}{1+\lambda^\mathsf{T} Z_i}, \quad \text{où} \quad \sum_{i=1}^n w_i(\lambda) Z_i = 0 \in \mathbb{R}^d.$$

Nous devons résoudre

$$\frac{1}{n}\sum_{i=1}^n \frac{Z_i}{1+\lambda^\mathsf{T} Z_i} = 0$$

Le dual

$$\mathbb{L}(\lambda) = -\sum_{i=1}^n \log(1+\lambda^\mathsf{T} Z_i)$$

Cette fonction est convexe en λ et,

$$\frac{\partial \mathbb{L}}{\partial \lambda} = \frac{1}{n}\sum_{i=1}^n \frac{Z_i}{1+\lambda^\mathsf{T} Z_i}.$$

Minimiser le dual maximise la vraisemblance.

n contraintes

Rappel: $\mathbb{L}(\lambda) = -\sum_{i=1}^{n} \log(1 + \lambda^{\mathsf{T}} Z_i)$

Le minimiseur doit avoir $1 + \lambda^{\mathsf{T}} Z_i > 0, \quad i = 1, \ldots, n$

Cela provient de $w_i > 0$.

Plus précisément

$$w_i < 1 \implies \frac{1}{n} \frac{1}{1 + \lambda^{\mathsf{T}} Z_i} < 1$$

Donc

$$1 + \lambda^{\mathsf{T}} Z_i > \frac{1}{n}, \quad i = 1, \ldots, n$$

Éliminer les contraintes

Remplacer $\log(x)$ par

$$\log_*(x) = \begin{cases} \log(x), & x \geqslant 1/n \\ Q(x), & x < 1/n \end{cases}$$

où Q est quadratique avec

$$\begin{aligned} Q(1/n) &= \log(1/n) \\ Q'(1/n) &= \log'(1/n) \quad \text{and} \\ Q''(1/n) &= \log''(1/n) \end{aligned}$$

$$Q(x) = \log(1/n) - 3/2 + 2nx - (nx)^2/2$$

Maintenant minimiser

$$\mathbb{L}_* = -\sum_{i=1}^{n} \log_*(1 + \lambda^\mathsf{T} Z_i)$$

Même optimum que $\mathbb{L}$. Aucune contrainte. Toujours fini.

Les itérations de Newton

Le gradient est $g(\lambda) \equiv \frac{\partial}{\partial \lambda} \mathbb{L}_*(\lambda)$.

La hessienne est $H(\lambda) \equiv \frac{\partial^2}{\partial \lambda \partial \lambda^\mathsf{T}} \mathbb{L}_*(\lambda)$

L'itération de Newton est

$$\lambda \leftarrow \lambda + s \quad \text{où} \quad s = -H^{-1} g$$

Analyse plus approfondie

Notre H est de la forme $J^\mathsf{T} J$ et $g = J^\mathsf{T} \eta$

Donc l'itération de Newton peut être résolue par la méthode des moindres carrés (plus stable d'un point de vue numérique).

Diminution des itérations

Les itérations de Newton nécessitent encore une méthode de réduction des itérations. S'il n'y a pas eu suffisamment de progrès vers le minimum, utiliser un plus petit multiple de s.

Levenberg-Marquardt: Si une itération devient trop petite, choisir des directions plus près de $-g$.

Exemple de Small et Yang

$$0 = \mathbb{E}(Z_1(Y - \beta_1 W - \alpha_1))$$
$$0 = \mathbb{E}(Y - \beta_1 W - \alpha_1)$$
$$0 = \mathbb{E}(Z_2(Y - (\beta_1 + \delta)W - \alpha_2))$$
$$0 = \mathbb{E}(Y - (\beta_1 + \delta)W - \alpha_2)$$

Résidus $Y - \beta_1 W - \alpha_1$ et $Y - (\beta_1 + \delta)W - \alpha_2$.

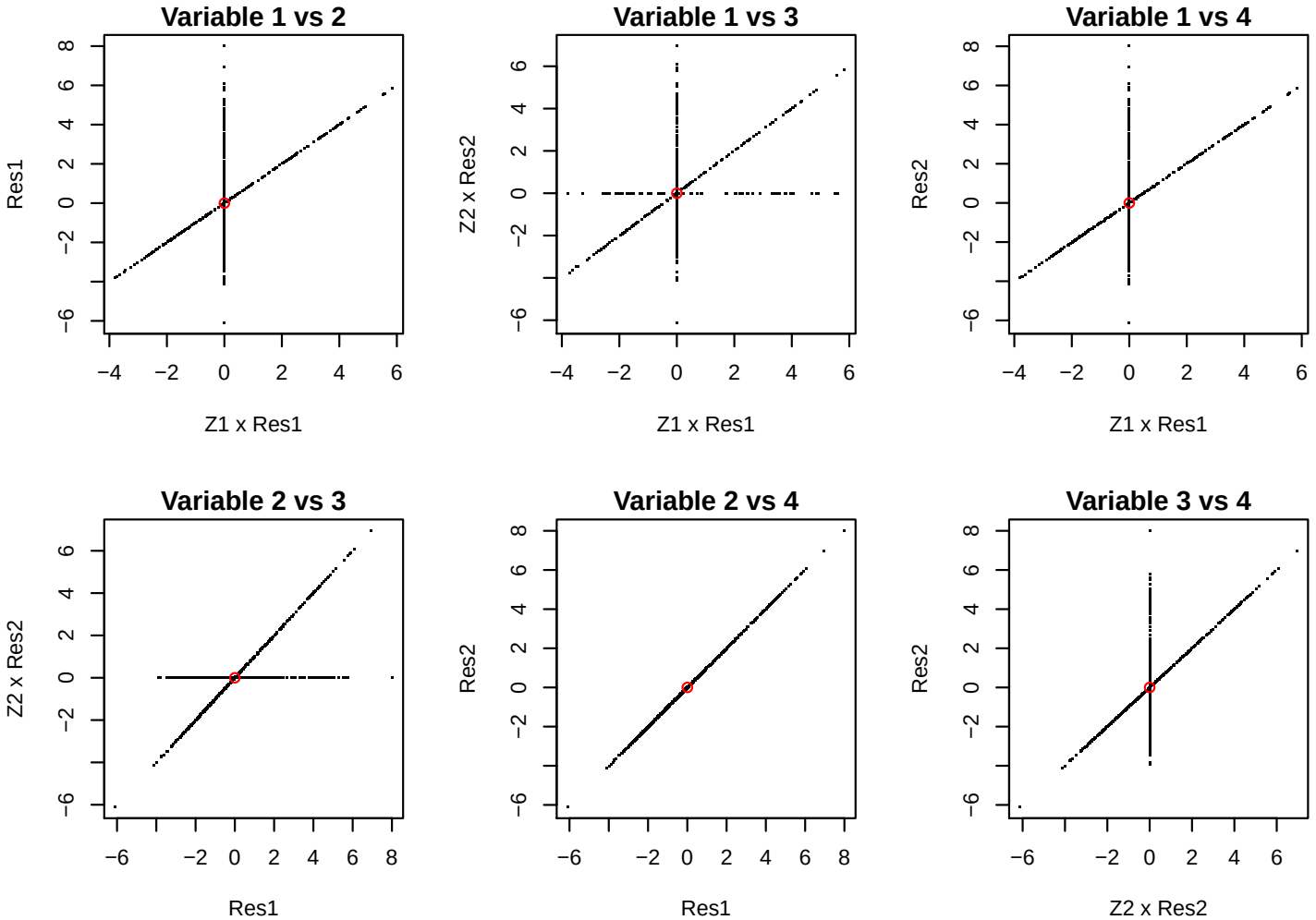
Variables instrumentales $Z_1, Z_2 \in \{0, 1\}$

Le problème est survenu dans un échantillon bootstrap.

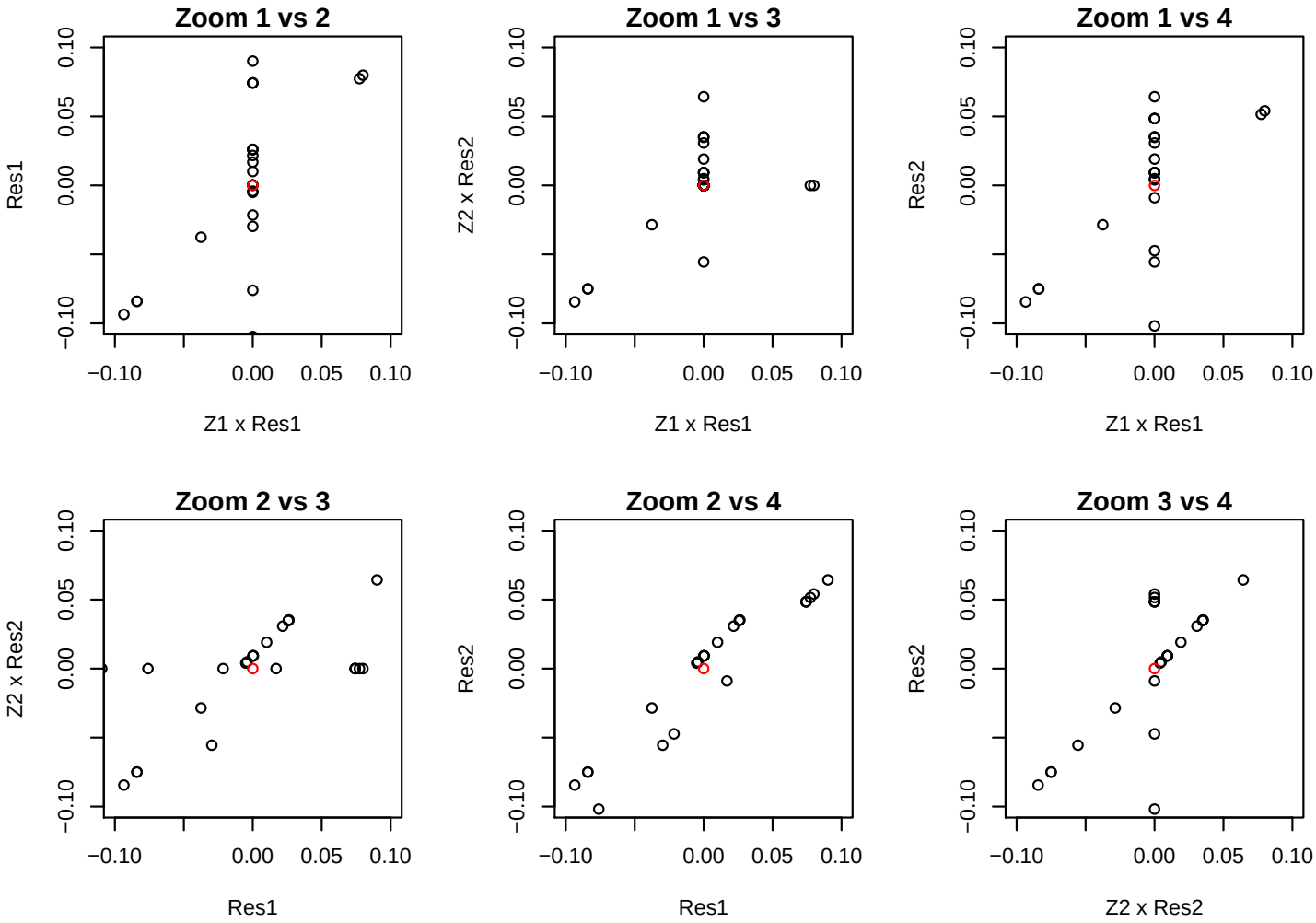
Exemple de Small et Yang

Ils devaient tester la moyenne de 1000 points dans $\mathbb{R}^4$.

Ce problème particulier est survenu dans un contexte de variables instrumentales.



Zoom avant


Zoom 1 vs 2
Res1
Z1 x Res1
Zoom 1 vs 3
Z2 x Res2
Z1 x Res1
Zoom 1 vs 4
Res2
Z1 x Res1
Zoom 2 vs 3
Z2 x Res2
Res1
Zoom 2 vs 4
Res2
Res1
Zoom 3 vs 4
Res2
Z2 x Res2
−0.10
0.00
0.05
0.10

Vraie log-vraisemblance empirique

$$\mathcal{R}(0) = -399.6937$$

L'ancien algorithme s'est bloqué; a taille des itérations est devenue petite et les réductions ad hoc de Levenberg-Marquardt n'aidaient pas.

Ils ont plutôt utilisé une recherche linéaire qui réduit les itérations.

Auto-concordance

Une fonction convexe g de $\mathbb{R}$ à $\mathbb{R}$ est **auto-concordante** si

$$|g'''(x)| \leqslant 2g''(x)^{3/2} \qquad \text{N.B. } g'' \geqslant 0$$

Nesterov & Nemirovskii (1994) Boyd & Vandeberghe (2004)

Une fonction convexe de g de $\mathbb{R}^d$ à $\mathbb{R}$ est auto-concordante si

$$g(\boldsymbol{x}_0 + t\boldsymbol{x}_1)$$

est une fonction auto-concordante de $t \in \mathbb{R}$.

Implications

La hessienne de $g(\boldsymbol{x})$ auto-concordante ne peut changer rapidement avec $\boldsymbol{x}$.

Les révisions de Newton avec la recherche linéaire qui réduit les itérations vont nécessairement converger.

Aussi, le décrément de Newton (ci-dessous) limite la sous-optimalité.

Le 2 n'est pas essentiel

Si $|g'''(x)| \leqslant Cg''(x)^{3/2}$ alors $\frac{C^2}{4}g$ est auto-concordante.

Newton avec retour

1) Choisir un point de départ $\boldsymbol{x}$
2) Repéter jusqu'à ce que le décrément de Newton $\nu(\boldsymbol{x})$ soit sous le seuil de tolérance
   a) $\boldsymbol{s} \leftarrow -H(\boldsymbol{x})^{-1} g(\boldsymbol{x}), \quad t \leftarrow 1$
   b) Pendant que $f(\boldsymbol{x} + t\boldsymbol{s}) > f(\boldsymbol{x}) + \alpha t \boldsymbol{s}^{\mathsf{T}} g$
      i) $t \leftarrow t \times \beta$
3) $\boldsymbol{x} \leftarrow \boldsymbol{x} + t\boldsymbol{s}$

Convergence garantie si

$\alpha \in (0, 1/2)$, $\beta \in (0, 1)$, f bornées inférieurement, l'ensemble sous-niveau de $\boldsymbol{x}$ est fermé

Décrément de Newton

$$\nu(\boldsymbol{x}) = (g(\boldsymbol{x})^{\mathsf{T}} H(\boldsymbol{x})^{-1} g(\boldsymbol{x}))^{1/2}$$

Si f est strictement convexe et auto-concordante et si $\nu(\widetilde{\boldsymbol{x}}) \leqslant 0.68$ alors

$$\inf f(\boldsymbol{x}) \geqslant f(\widetilde{\boldsymbol{x}}) - \nu(\widetilde{\boldsymbol{x}})^2$$

Chen, Sitter, Wu

- Biometrika (2002)
- Utilisent la recherche de ligne par retour en arrière avec réduction de moitié des étapes quand l'objectif ne s'améliore pas
- Démontrent la convergence par les résultats dans Polyak (1987)
- Débutent la $k$ième recherche à la taille $t = (k+1)^{-1/2}$.
- Commencer avec $t < 1$ ralentira Newton sous la convergence quadratique. Ils observent que commencer avec $t = 1$ fonctionne.

De retour à $\mathbb{L}_*$

$$\mathbb{L}_*(\lambda) = -\sum_{i=1}^{n} \log_*(1 + \lambda^\mathsf{T} Z_i) \qquad \text{où} \quad \log_*(x) = \begin{cases} \log(x), & x \geqslant 1/n \\ Q(x), & x < 1/n \end{cases}$$

$\log_*$ est auto-concordant sur $(-\infty, 1/n)$ et sur $(1/n, \infty)$.

Mais il lui manque une troisième dérivée à $1/n$

Donc n'est pas auto-concordant.

Approximations d'ordre supérieur

$$-\log_{(k)}(x) = \begin{cases} -\log(x), & x \geqslant \epsilon > 0 \\ h_k(x - \epsilon) & x < \epsilon \end{cases}$$

Approximation de Taylor de $-\log$ à ϵ

$$h_k(y) = h_k(y; \epsilon) = -\sum_{t=0}^{k} \log^{(t)}(\epsilon) \frac{y^t}{t!}$$

$k = 2$	Convexe mais pas auto-concordante (échoue à ϵ)	$-\log_{(2)} = -\log_*$
$k = 3$	Même pas convexe	
$k = 4$	Convexe et auto-concordante	$\underset{\smile}{\mathbf{0}\,\mathbf{0}}$

De retour à l'exemple

La version auto-concordante donne aussi $\log \mathcal{R}() = -399.6937$

Décrément de Newton

$$\eta \equiv (g^\mathsf{T} H^{-1} g)^{-1/2} = 6.74277 \times 10^{-16}$$

L'estimation donne $\log(\mathcal{R})$ à moins de η^2 du vrai optimum.

C.-à-d. bon à l'intérieur d'une précision donnée.

Ébauche de preuve

Nous devons démontrer que $h_4(y)$ est auto-concordante sur $(-\infty, 0]$.

- c.-à-d., $|h_4'''| \leqslant 2(h_4')'^{3/2}$
- Il suffit de démontrer que $h_4(\epsilon \times \cdot)$ est auto-concordante
- $h_4'''(t\epsilon) = \epsilon^{-3}(-2 + 6t)$
- $h_4''(t\epsilon) = \epsilon^{-2}((1-t)^2 + t^2)$
- $\rho(t) \equiv \dfrac{|h_4'''(t\epsilon)|}{h_4''(t\epsilon)^{3/2}} = \dfrac{2-6t}{(t-1)^2 + t^2}$ sur $t \leqslant 0$.
- $\rho(0) = 2$
- $\rho'(t) \geqslant 0$ pour $t \leqslant 0$

Donc le rapport ρ augmente à 2 quand $t \uparrow 0$

Log-vraisemblance quartique

Utiliser $\mathcal{R}_Q = -\sum_{i=1}^{n} \widetilde{\log}(nw_i)$

$$\widetilde{\log}(1+z) = z - \frac{1}{2}z^2 + \frac{1}{3}z^3 - \frac{1}{4}z^4$$

Propriétés

Correction Bartlett Corcoran (1998)

Fait correspondre 4 dérivées et 4 moments

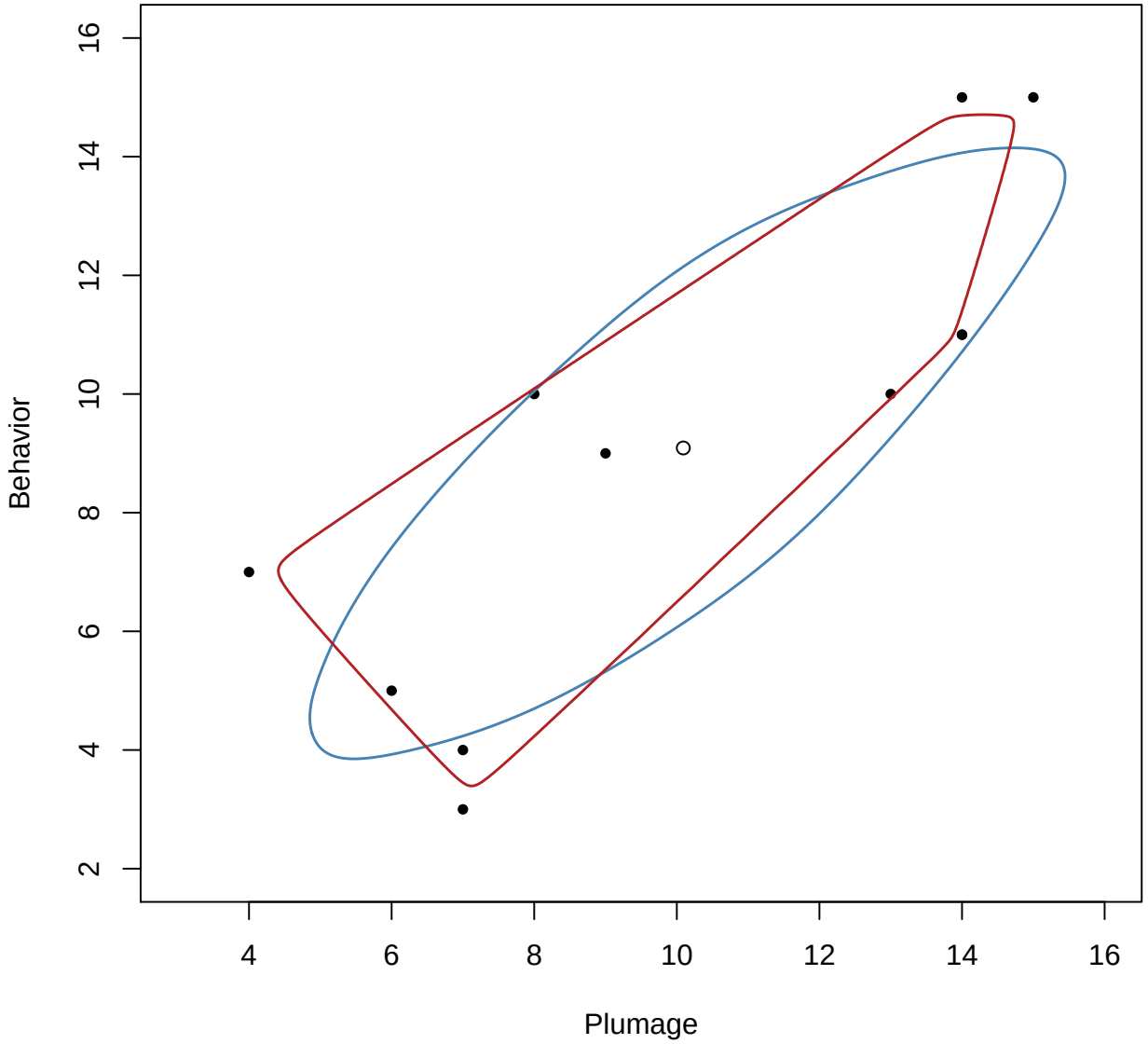
Auto-concordant O (2013) [$C = 3.92$ à la place de $C = 2$]

Régions de confiance convexes pour la moyenne O (2013)

Le multiplicateur de Lagrange pour $\sum w_i = 1$ ne peut être éliminé.

Algorithme primal-dual de Boyd & Vandeberghe disponible

Données sur les canards



Région de confiance extrême. Rouge $\mathcal{R}$; bleu $\mathcal{R}_Q$

Larsen & Marx (1986)

Considérations futures

Il n'est peut-être pas nécessaire d'imposer $1 + \lambda^\mathsf{T} Z_i > 1/n$

Éviter complètement les pseudo-logarithmes par partie

La réduction d'itération garde $1 + \lambda^\mathsf{T} Z_i > 0$

$-\sum_{i=1}^{n} \log(1 + \lambda^\mathsf{T} Z_i)$ est aussi auto-concordante

Plus simple, mais

$\log(z)$ peut être légèrement moins bien conditionné que z^4

Maximiser sur des paramètres de nuisance pourrait être plus simple sans la contrainte linéaire λ

Si le temps le permet . . .

Quelques défis computationnels.

Modèle pour la régression

Maximiser $\sum_{i=1}^{n} \log(nw_i)$ où $w_i \geqslant 0$ $\sum_i w_i = 1$

$$\sum_i w_i(Y_i - \boldsymbol{x}_i^\mathsf{T}\beta)\boldsymbol{x}_i = 0$$

et $\beta_j = \beta_{j0}$.

Une optimisation pas tout à fait convexe

Les variables libres sont β_k pour $k \neq j$ ainsi que $w_1, \ldots, w_n$.

Le défi computationnel provient de la **bilinéarité** de la contrainte.

Si β est fixe, la contrainte de l'équation normale est linéaire en w et vice versa.

VE pour échantillons multiples

Le chapitre 11.4 du livre “Empirical likelihood” s’intéresse à des situations d’échantillons multiples. Observations $\boldsymbol{X}_i \overset{\text{iid}}{\sim} F$ pour $i = 1, \ldots, n$ indépendants de $\boldsymbol{Y}_j \overset{\text{iid}}{\sim} G$ pour $j = 1, \ldots, m$. Le rapport de vraisemblance est

$$\prod_{i=1}^{n}\prod_{j=1}^{m}(nu_i)(mv_j)$$

avec $u_i \geqslant 0$, $v_j \geqslant 0$, $\sum_i u_i = 1$, $\sum_j v_j = 1$ et

$$\sum_i \sum_j u_i v_j h(\boldsymbol{x}_i, \boldsymbol{y}_j, \theta) = 0 \tag{1}$$

Par exemple: $h(X, Y, \theta) = 1_{X-Y>\theta} - 1/2$. Le problème computationnel est un défi. La log-vraisemblance est convexe mais la contrainte (1) est bilinéaire. Donc le calcul est difficile.

Encore de la régression

$$Y \approx \boldsymbol{x}^{\mathsf{T}}\beta, \quad \boldsymbol{x} \in \mathbb{R}^d \quad y \in \mathbb{R}$$

Équations estimantes*

$$\mathbb{E}\big((Y - \boldsymbol{x}^{\mathsf{T}}\beta)\boldsymbol{x}\big) = 0$$

Équations normales

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^{\mathsf{T}}\beta)\boldsymbol{x}_i = 0 \in \mathbb{R}^d$$

En principe nous laissons $\boldsymbol{z}_i = \boldsymbol{z}_i(\beta) \equiv (y_i - \boldsymbol{x}_i^{\mathsf{T}}\beta)\boldsymbol{x}_i \in \mathbb{R}^d$, adjoignons $\boldsymbol{z}_{n+1}$ et $\boldsymbol{z}_{n+2}$, et continuons.

*Les résidus $\varepsilon = y - \boldsymbol{x}^{\mathsf{T}}\beta$ ne sont pas corrélés avec $\boldsymbol{x}$.
Ils ont aussi une moyenne de zéro quand $\boldsymbol{x}$, comme d'habitude, contient une constante.

Condition pour l'enveloppe en régression

$$\mathcal{R}(\beta) = \sup\left\{ \prod_{i=1}^{n} n w_i \;\middle|\; w_i \geqslant 0, \sum_{i=1}^{n} w_i = 1, \sum_{i=1}^{n} w_i (y_i - \boldsymbol{x}_i^\mathsf{T}\beta)\boldsymbol{x}_i = 0 \right\}$$

$$\mathcal{P} = \mathcal{P}(\beta) = \{\boldsymbol{x}_i \mid y_i - \boldsymbol{x}_i^\mathsf{T}\beta > 0\} \qquad \boldsymbol{x} \text{ avec résidus positifs}$$

$$\mathcal{N} = \mathcal{N}(\beta) = \{\boldsymbol{x}_i \mid y_i - \boldsymbol{x}_i^\mathsf{T}\beta < 0\} \qquad \boldsymbol{x} \text{ avec résidus négatifs}$$

Condition de l'enveloppe convexe O (2000)

$$\text{chull}(\mathcal{P}) \bigcap \text{chull}(\mathcal{N}) \neq \varnothing \implies \beta \in C(0)$$

Pour $\boldsymbol{x}_i = (1, t_i)^\mathsf{T} \in \mathbb{R}^2$ $\mathcal{P}$ et $\mathcal{N}$ sont des intervalles en $\{1\} \times \mathbb{R}$.

Inverse

Supposer que $\tau \notin \{t_1, \ldots, t_n\}$ et

$$\text{Sign}(y_i - \beta_0 - \beta_1 t_i) = \begin{cases} 1, & t_i > \tau \\ -1, & t_i < \tau \end{cases}$$

Supposer aussi que

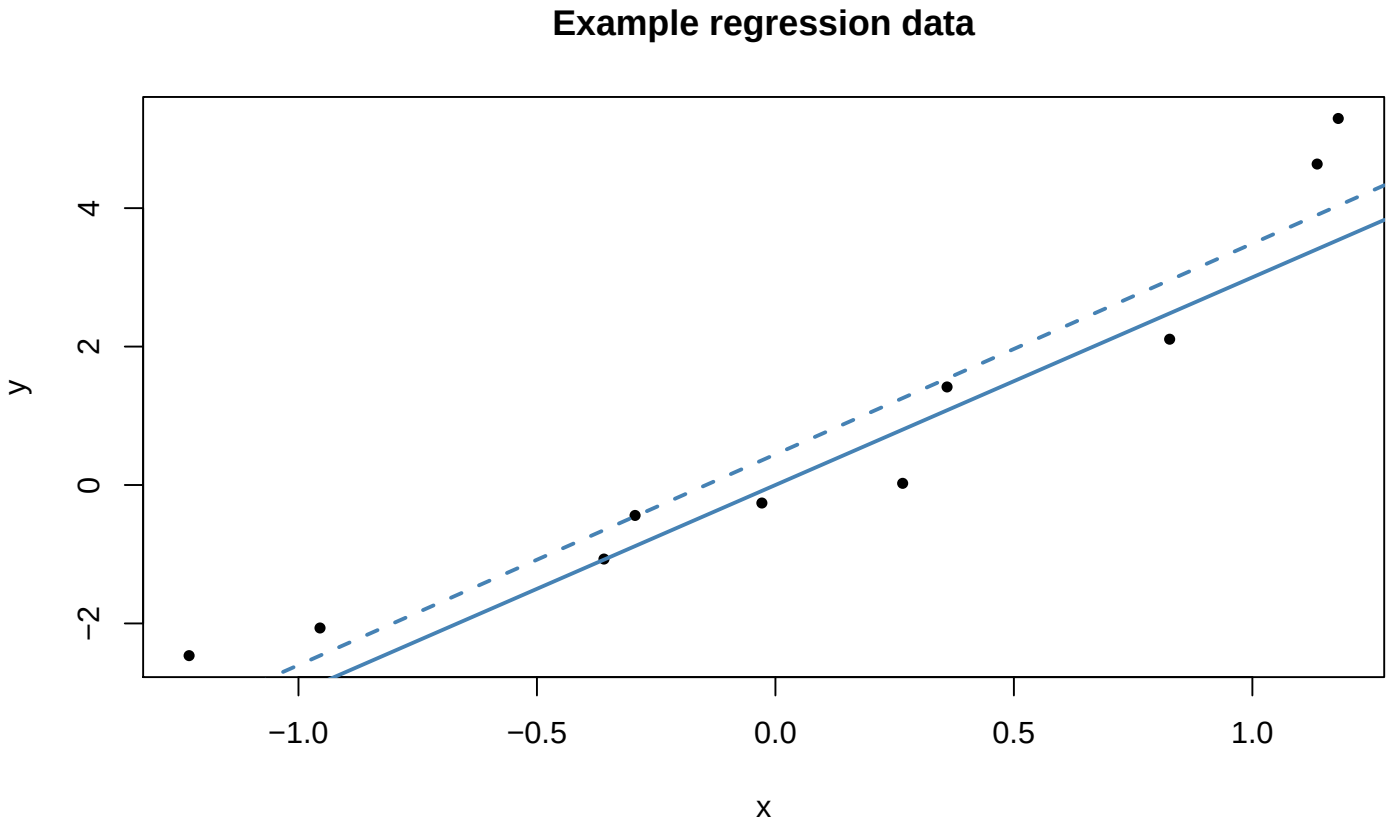
$$\sum_i w_i \begin{pmatrix} 1 \\ t_i \end{pmatrix} (y_i - \beta_0 - \beta_1 t_i) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Alors

$$\sum_i w_i (y_i - \beta_0 - \beta_1 t_i)(t_i - \tau) = 0$$

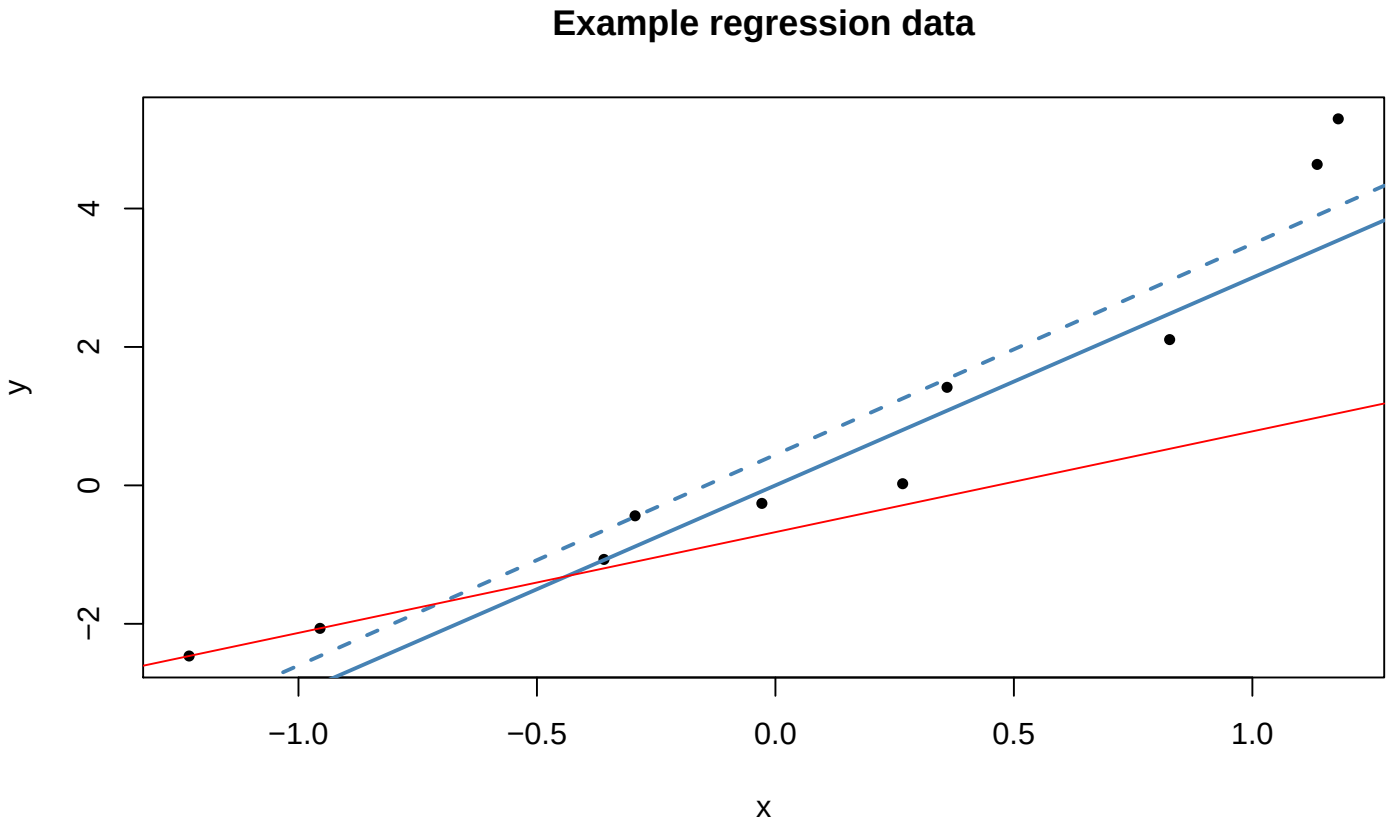
Mais $(y_i - \beta_0 - \beta_1 t_i)(t_i - \tau) > 0 \; \forall i$

Donc la condition de l'enveloppe est nécessaire.

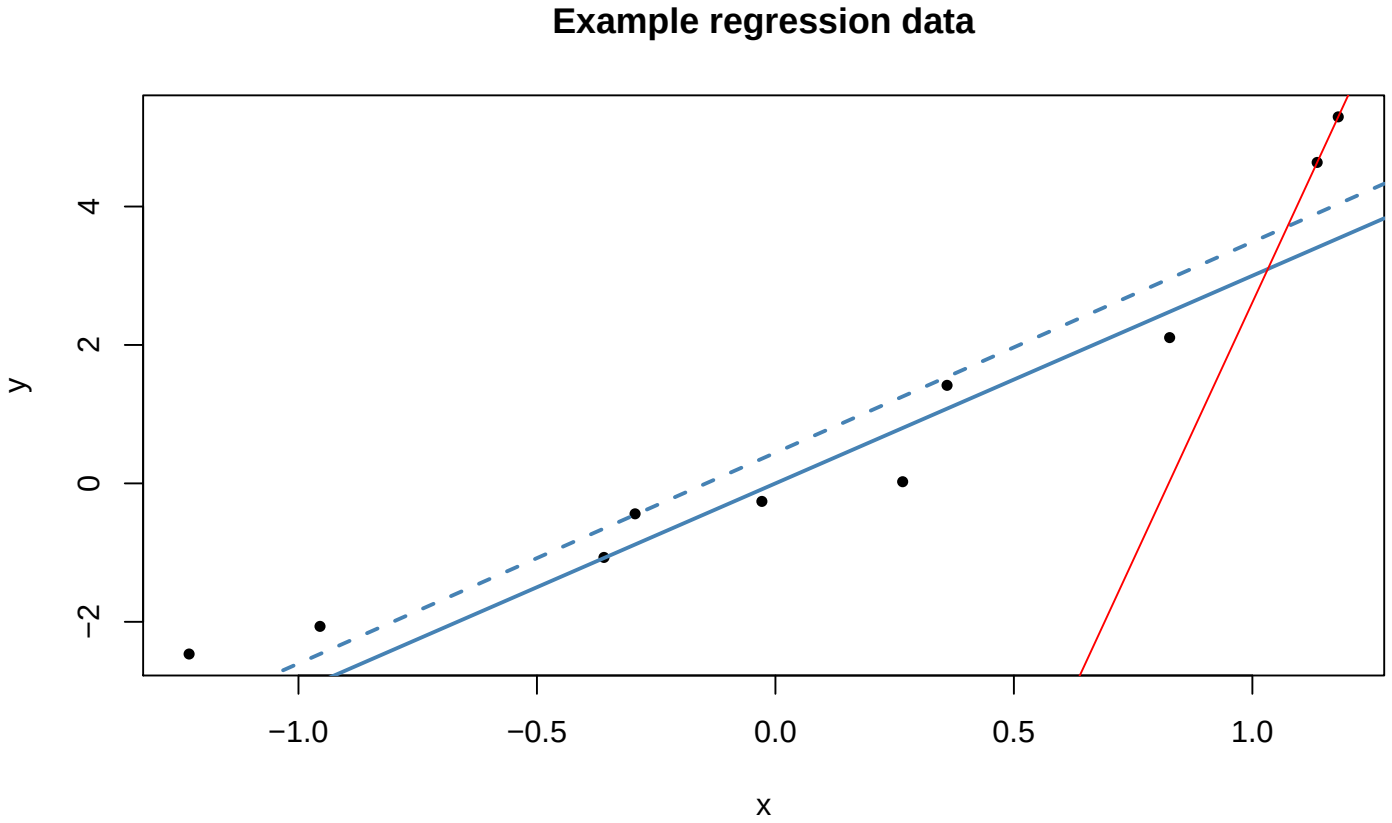


$Y = \beta_0 + \beta_1 X + \sigma\varepsilon \quad \beta = (0, 3)^{\mathsf{T}}, \sigma = 1$

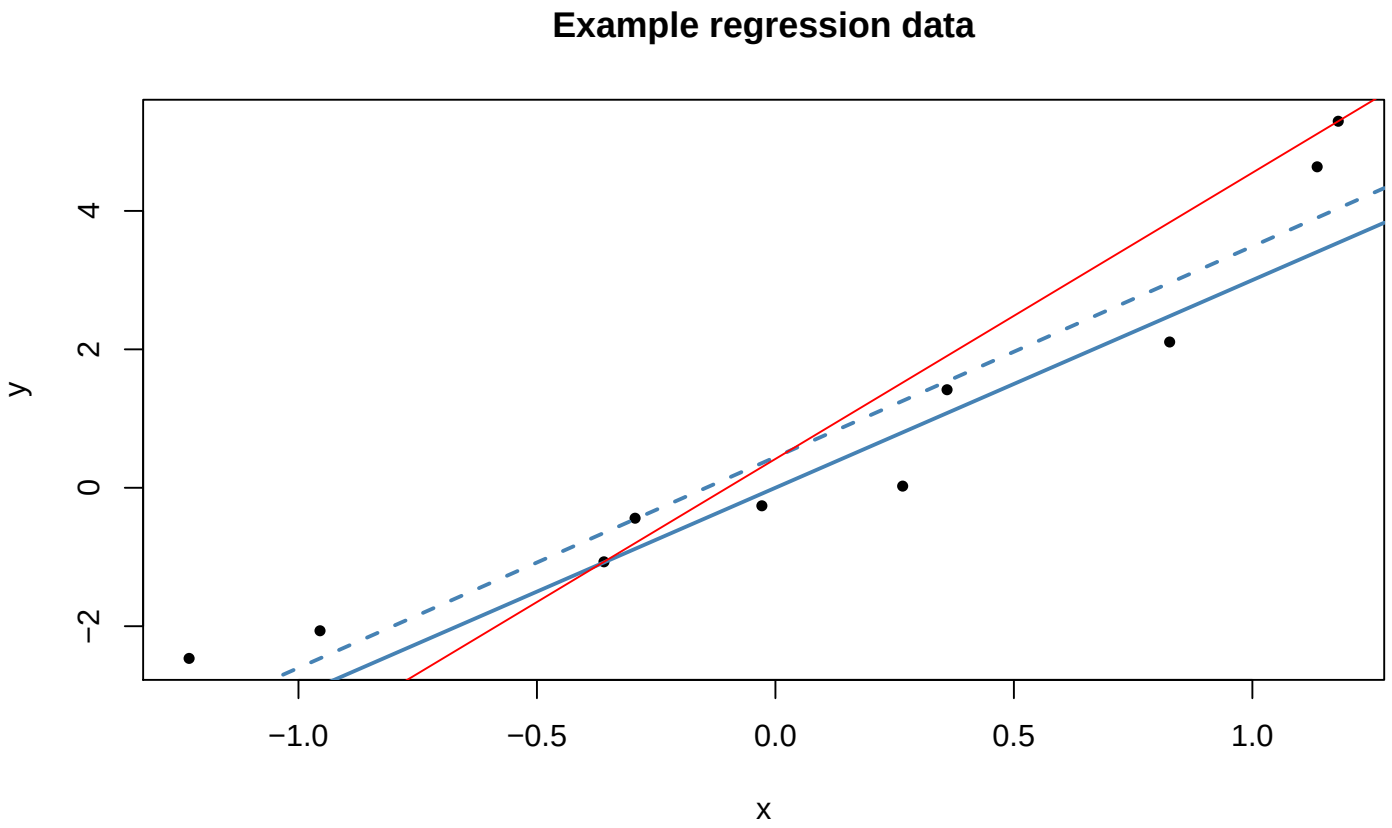
β continue $\hat{\beta}$ pointillé



La droite rouge est sur la frontière de l'ensemble (β_0, β_1) avec une vraisemblance empirique positive.



Une autre droite frontière.

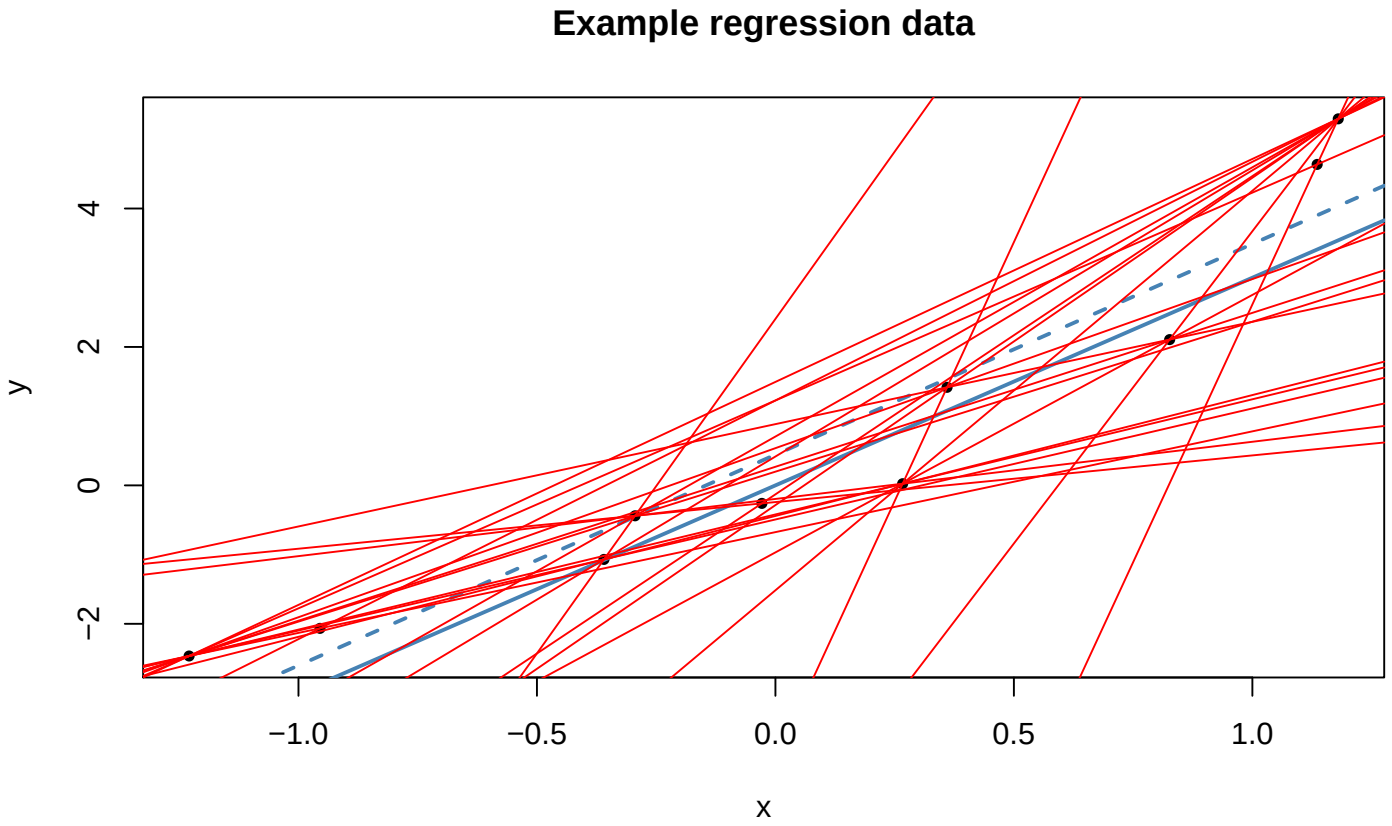


Encore une autre droite frontière.

Le côté gauche a des résidus positifs; le côté droit des résidus négatifs.

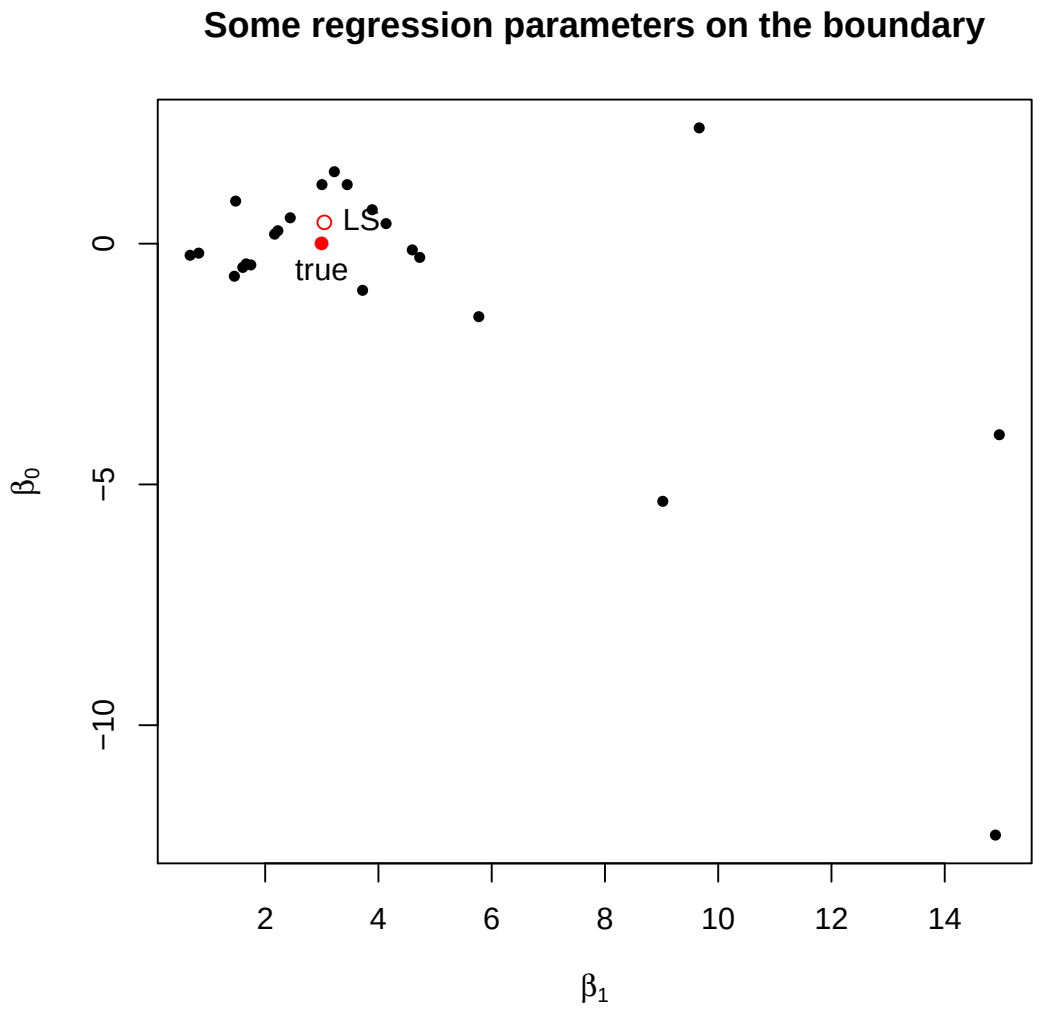
Bougez-la vers le haut et point 3 obtient un résidu négatif $\Longrightarrow$ ok.

Bougez vers le bas $\Longrightarrow$ PAS ok.



Toutes les droites frontières qui interpolent deux données.

Elles sont un sous-ensemble de la frontière.



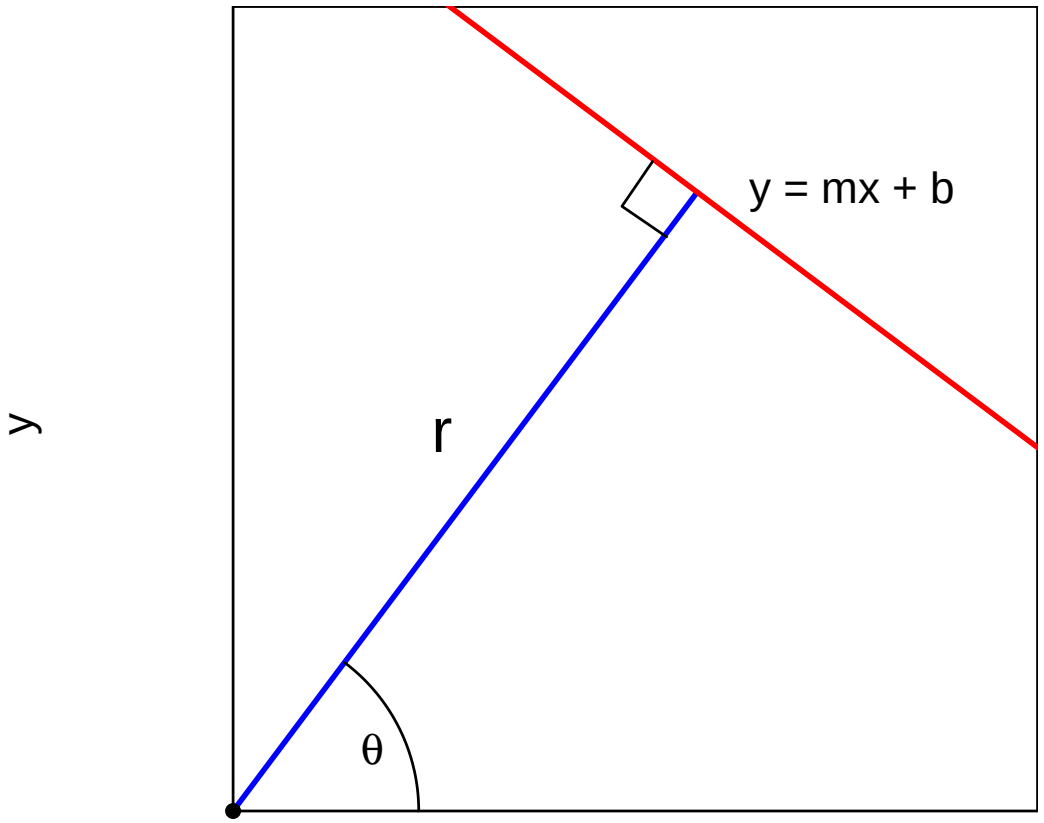
Les points frontières (β_0, β_1). La région n'est pas convexe.

Elle **est** convexe en β_0 (vertical) pour β_1 fixe (horizontal).

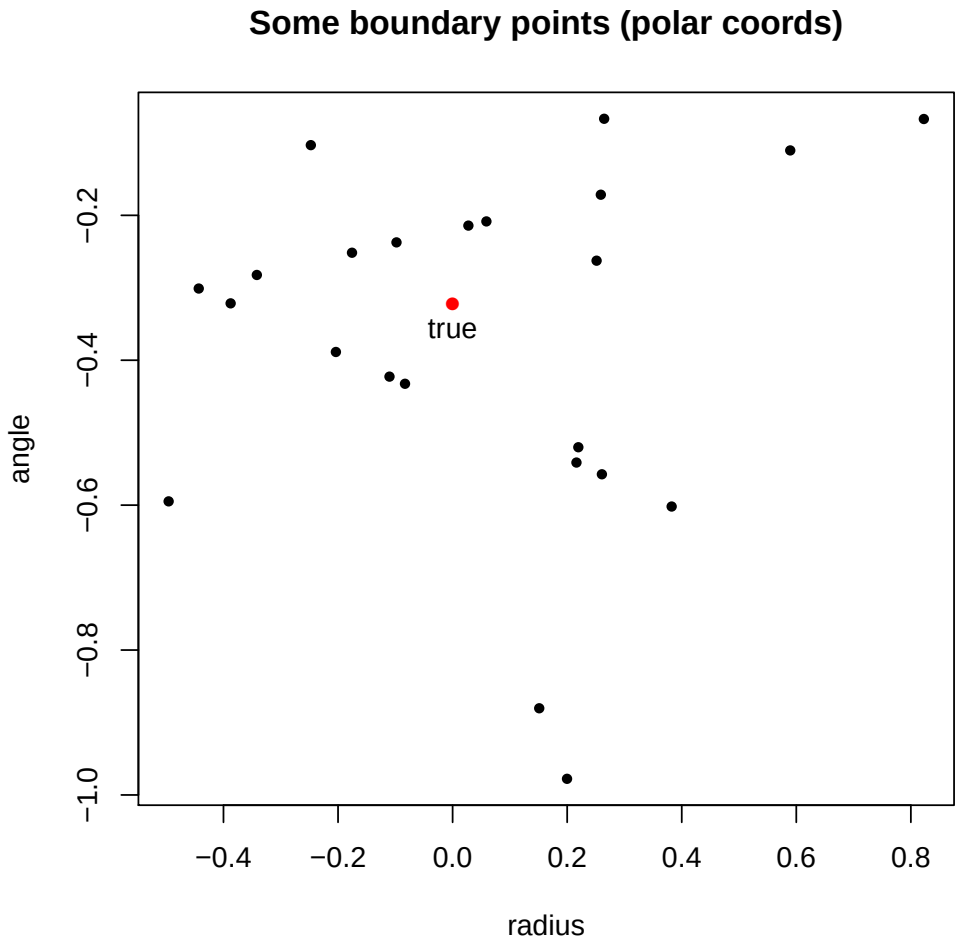
Qu'est-ce qu'un ensemble de droites convexes?

- ensemble convexe de (β_0, β_1)?
- ensemble convexe de (ρ, θ)? (coordonnées polaires)
- ensemble convexe de (a, b) $(ax + by = 1)$?

Coordonnées polaires d'une droite



Points frontières en coordonnées polaires



Pas convexe ici non plus.

Convexité intrinsèque

Il existe une notion géométriquement intrinsèque pour un ensemble d'hyperplans linéaires.

J. E. Goodman (1998) "Quand un ensemble de droites dans l'espace est-t-il convexe" Peut-être que $\cdots$ cela peut aider quelques calculs.

Définition duale

L'ensemble d'hyperplans qui intercepte un ensemble convexe $C \subset \mathbb{R}^d$ est un ensemble convexe d'hyperplans.

Il en est de même de l'ensemble d'hyperplans qui intercepte **tous les** $C_1, \ldots, C_k \subset \mathbb{R}^d$ pour C_j convexe.

Fonction convexes

Cette notion d'ensemble convexe ne semble pas encore avoir de notion correspondante de fonction convexe. Il pourrait y avoir des fonctions quasi-convexes, celles où les lignes de niveau sont convexes. Mais une quasi-convexité est beaucoup moins puissante au niveau computationnel qu'une convexité.

Remerciements

1) Dylan Small et Dan Yang

2) Jiahua Chen, partageant une première version d'un article

3) La revue Canadienne de statistique

4) Changbao Wu et le comité de la RCS

5) Caroline Petit-Turcotte & Pierre-Jérôme Bergeron, traduction

6) NSF DMS-0906056