Empirical Likelihood

Art B. Owen

Department of Statistics
Stanford University

These lectures

I) Basics of empirical likelihood

II) Estimating equations

III) Research frontier✓

Today: Research topics

1) Hybrids with parametric likelihoods
2) Bayes and EL
3) Log concavity
4) Escaping from the hull
5) Sparse likelihoods
6) Convex objective; bilinear constraint
7) Regression and convexity

These are areas that are either new, have potential for new uses, or are ripe for improvement.

EL hybrids (mostly Jing Qin)

Part of the problem is parametric

We want to use that knowledge

The rest of the problem is non-parametric

One parametric sample, one not

$\boldsymbol{Y}$ well studied and has parametric distribution

$\boldsymbol{X}$ new and/or does not follow parametric distribution

$$\boldsymbol{X}_i \sim F, \quad i = 1, \dots, n$$
$$\boldsymbol{Y}_j \sim G(\boldsymbol{y}; \theta), \quad j = 1, \dots, m$$
$$0 = \int\int h(\boldsymbol{x}, \boldsymbol{y}, \phi) dF(\boldsymbol{x}) dG(\boldsymbol{y}; \theta)$$

e.g. $\phi = \mathbb{E}(\boldsymbol{Y}) - \mathbb{E}(\boldsymbol{X})$

Multiply the likelihoods

$$L(F, \theta) = \prod_{i=1}^n F(\{\boldsymbol{x}_i\}) \prod_{j=1}^m g(\boldsymbol{y}_j; \theta)$$
$$R(F, \theta) = L(F, \theta)/L(\widehat{F}, \hat{\theta})$$
$$\mathcal{R}(\phi) = \max_{F,\theta} R(F, \theta) \quad \text{such that}$$
$$0 = \sum_{i=1}^n w_i \int h(\boldsymbol{x}_i, \boldsymbol{y}, \phi) dG(\boldsymbol{y}; \theta)$$

Qin gets a χ^2 limit

Parametric model for data ranges

$$\boldsymbol{X} \sim \begin{cases} f(\boldsymbol{x}; \theta) & \boldsymbol{x} \in P_0 \\ \text{???} & \boldsymbol{x} \notin P_0 \end{cases}$$

Examples

- Extreme values, exponential tails on $P_0 = [T, \infty)$ something else below T
- Normal data on $P_0 = [-T, T]$ with outliers outside

$$L = \prod_{i=1}^n f(\boldsymbol{x}_i; \theta)^{\boldsymbol{x}_i \in P_0} w_i^{\boldsymbol{x}_i \notin P_0}$$

Define $\mathcal{R}$ using

$$1 = \int_{P_0} dF(\boldsymbol{x}; \theta) + \sum_{i=1}^n w_i 1_{\boldsymbol{x} \notin P_0}$$

Qin & Wong get a χ^2 limit for means

More hybrids

Parametric	Nonparametric	
$g(\boldsymbol{y} \mid \boldsymbol{x}; \theta)$	$\boldsymbol{X} \sim F$	
$\boldsymbol{X} \sim f(\boldsymbol{x}; \theta)$	$\boldsymbol{Y} \mid \boldsymbol{X} = \boldsymbol{x} \sim G_{\boldsymbol{x}}$	Few $\boldsymbol{x}$ vals
$\boldsymbol{X} \sim f(\boldsymbol{x}; \theta)$	$(\boldsymbol{Y} - \mu(\boldsymbol{x}))/\sigma(\boldsymbol{x}) \sim G$	

Bayesian empirical likelihood (Lazar)

Prior $\theta \sim \pi(\theta)$
$x \sim F$ nonparametric

Posterior $\propto \pi(\theta)\mathcal{R}(\theta)$

Here we have informative prior nonparametric likelihood
Reverse of a common practice

Posterior regions asymptotically properly calibrated

Maybe it can be justified via least favorable families

Schennach (2005) multiplies an exponential likelihood by a prior.

Approximate Bayesian Computation

ABC is used in problems where the likelihood cannot be computed.

For example, suppose we have a model with parameter θ for how biological populations may have evolved over a long time period. But we only have data on the present. There may be no good way to evaluate the probability of the present as a function of θ.

In ABC we sample $\theta_1, \dots, \theta_N$ from the prior distribution on θ and then data $\boldsymbol{X}$ from its distribution given θ. If $\boldsymbol{X}_i$ is close to the observed value $\boldsymbol{X}^*$ then we retain θ_i and give it a 'weight' that is inversely proportional to some $\mathrm{dist}(\boldsymbol{X}_i, \boldsymbol{X}^*)$.

The normalized weights are interpreted as a posterior distribution on θ. There are many versions.

Mengersen, Pudlo & Robert (2013) use empirical likelihood for an ABC-like algoirthm, when the parameter is defined by estimating equations.

Log concavity

There is an MLE for the problem of maximizing $\prod_{i=1}^n f(\boldsymbol{x}_i)$ where f is a log concave density on $\mathbb{R}^d$.

Suppose now that we maximize this likelihood subject to

$$\int_{\mathbb{R}^d} \boldsymbol{x} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \mu, \quad \text{or} \quad \int_{\mathbb{R}^d} m(\boldsymbol{x},\theta) f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 0$$

Will the result yield a χ^2 calibration?

How will we compute it?

The MLE density $\hat{f}$ is supported on the convex hull of $\boldsymbol{x}_i$ and so the hull issue (below) will be relevant when d is large

Probability μ_0 in the hull

$$\mathcal{H} = \Big\{ \sum_{i=1}^n w_i \boldsymbol{x}_i \mid w_i \ge 0, \sum_{i=1}^n w_i = 1 \Big\}$$

Wendel (1962)

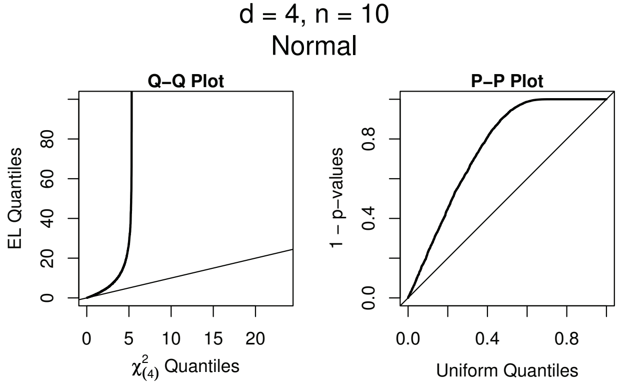
If distn of $\boldsymbol{X}_i$ symmetric about μ then

$$\begin{aligned}\Pr(\mu \notin \mathcal{H}) &= \sum_{k=0}^{d-1} \binom{n-1}{k} \Big(\frac{1}{2}\Big)^{n-1} \\ &= \Pr(\mathrm{Bin}(n-1, 1/2) < d)\end{aligned}$$

$d-1$ or fewer heads in $n-1$ trials

NB: a set of $n-1$ independent coin toss events corresponding to this result has yet to be exhibited.

Plain EL under-coverage (extreme case)



Emerson & O (2009)

Vertical asymptote from atom at $+\infty$ for $-2\log\mathcal{R}(\mu_0)$.

Growing dimension

Hjort, McKeague & Van Keilegom (2009)

Consider EL for dimension p growing with n

Bounded $\boldsymbol{X}_{n,i}$ IID mean 0 variance Σ_n with eigenvalues in $[A,B]\subset(0,\infty)$

Key condition for χ^2 limit is $\frac{p^3}{n}\to 0$

For $q>2$ moments $\frac{p^{3+6/(q-2)}}{n}\to 0$

Penalized EL

Bartolucci (2007) gives 15 points in $\mathbb{R}^4$ from $\chi^2_{(1)}$. The mean is not in the hull.

Bootstrapping: $\bar{\boldsymbol{x}}$ is not in the hull of resampled data 30% of the time.

relax the constraint

$$L^\dagger(\mu,h)=\max_w\prod_{i=1}^n w_i\times e^{-n\delta(\nu-\mu)/(2h^2)}$$

where $\nu=\sum_i w_i\boldsymbol{x}_i$ and $\delta(\nu-\mu)=(\nu-\mu)^\mathsf{T}V^{-1}(\nu-\mu)$ for V positive definite (eg sample covariance)

This favors ν close to μ but does not enforce it. There's a χ^2 limit if $h=O(n^{-1/2})$

Lahiri & Mukhopadhyay (2012) avoid using a sample covariance

extend to very large p including some $p>n$

Escape from the hull

Idea: extend the sample to ensure that $\mu\in\mathcal{H}$

If we knew a support set for F we could use it.

Or, add an artificial point (undata) $\boldsymbol{x}_{n+1}$. Now,

$$T(F)=\sum_{i=1}^{n+1}w_i\boldsymbol{x}_i,\quad\text{and,}$$

$$L(F)=\prod_{i=1}^n w_i,\quad\text{or,}$$

$$L(F)=\prod_{i=1}^{n+1}w_i.$$

The second version is easier computationally and asymptotically the same (if $\|\boldsymbol{x}_{n+1}\|$ reasonable).

Chen, Variyath & Abraham (2008) originate this approach.

Adjusted empirical likelihood

Chen, Variyath & Abraham (2008) use

$$\boldsymbol{x}_{n+1} = \mu - a_n(\bar{\boldsymbol{x}} - \mu), \quad a_n = \log(n)/2$$

$$a_n = o_p(n^{2/3}) \quad \text{preserves 1st order asymptotics}$$

Note: new point $\boldsymbol{x}_{n+1}$ depends on μ

Now μ is between $\bar{\boldsymbol{x}}$ and $\boldsymbol{x}_{n+1}$:

$$\mu = \frac{\boldsymbol{x}_{n+1} + a_n\bar{\boldsymbol{x}}}{1 + a_n}$$

Hull of $\boldsymbol{x}_1, \dots, \boldsymbol{x}_{n+1}$ contains μ

Not all is well yet

Let $\mathcal{R}^*$ be adjusted profile empirical likelihood. Then we can show:

$$-2\log\mathcal{R}^*(\mu) \le -2\left[n\log\left(\frac{(n+1)a_n}{n(a_n+1)}\right) + \log\left(\frac{n+1}{a_n+1}\right)\right]$$

which is bounded, even if $\|\mu\| \to \infty$.

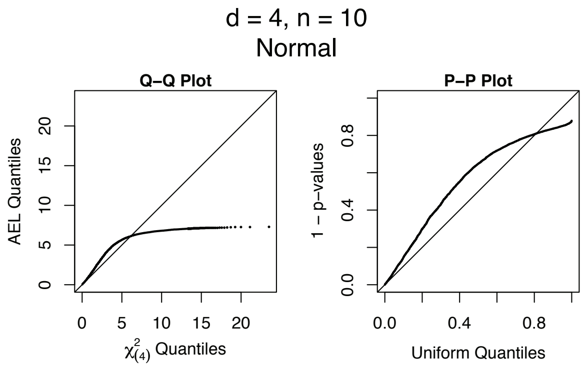
Opposite problem from $\log\mathcal{R}(\mu)$ which diverged at finite $\|\mu\|$.

Instead of a bounded 100% region we can get all of $\mathbb{R}^d$ at less than 100% confidence.

Extreme example ctd.

$n = 10$, $d = 4$, 88.1% region is $\mathbb{R}^4$.

Adjusted EL coverage (extreme case)



Emerson & O (2009)

Balanced adjusted empirical likelihood

Dissertation: Emerson (2009)

1) Add 2 points $\boldsymbol{x}_{n+1}$ and $\boldsymbol{x}_{n+2}$
2) $(\boldsymbol{x}_{n+1} + \boldsymbol{x}_{n+2})/2 = \bar{\boldsymbol{x}}$ (preserving sample mean)
3) farther new points if $\mu - \bar{\boldsymbol{x}}$ is a direction where the sample varies a lot

Add points

$$\boldsymbol{x}_{n+1} = \mu - sc_{u^*}u^*$$

$$\boldsymbol{x}_{n+2} = 2\bar{\boldsymbol{x}} - \mu + sc_{u^*}u^*$$

where

$$u^* = \frac{\bar{\boldsymbol{x}} - \mu}{\|\bar{\boldsymbol{x}} - \mu\|} \qquad c_{u^*} = ({u^*}^{\mathsf{T}}S^{-1}u^*)^{-1/2}$$

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^{\mathsf{T}} \qquad s \approx 1.9$$

Choice of s

Choice of s is based on empirical work. The best s depends (weakly) on d eg $s = 1.7$ for $d = 2$ to $s = 2.4$ for $d = 20$

Animation

Show some slides of S. Emerson illustrating how $\boldsymbol{x}_{n+1}$ and $\boldsymbol{x}_{n+2}$ move with μ

Related

Independently Liu & Chen (2009) also added 2 points.

Their 2 points were designed to improve Bartlett correction.

Ours were tuned to give good small sample coverage in high dimensions.

Invariance

Let $A \in \mathbb{R}^{d\times d}$ be non-singular.

Set $\widetilde{\boldsymbol{x}}_i = A\boldsymbol{x}_i$ and $\widetilde{\mu} = A\mu$.

Let C be the balanced adjusted empirical likelihood region for μ_0 based on $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$.

Let $\widetilde{C}$ be the balanced adjusted empirical likelihood region for $\widetilde{\mu}_0 = A\mu_0$ based on $\widetilde{\boldsymbol{x}}_1, \dots, \widetilde{\boldsymbol{x}}_n$.

Then $\mu \in C \iff \widetilde{\mu} \in \widetilde{C}$.

Emerson & O (2009) Proposition 4.1.

Hotelling's T^2 and the original EL are also invariant this way.

Avoiding the boundedness

Recall $-2\log\mathcal{R}^*$ was bounded.

The new criterion $-2\log\mathcal{R}^{**}$ is unbounded.

The ultimate cause is that

$\|\boldsymbol{x}_{n+1} - \mu\|$ is proportional to $\|\bar{\boldsymbol{x}} - \mu\|$ in AEL but is of constant order in BAEL

The larger $\|\boldsymbol{x}_{n+1} - \mu\|$ in AEL means that less weight needs to go there.
Less weight there $\cdots$ allows more weight on the other n points and a higher likelihood.

Connection to T^2

Recall

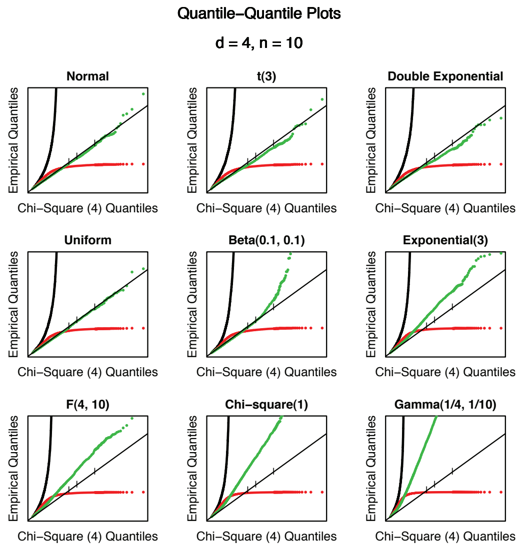
$$\boldsymbol{x}_{n+1} = \mu - sc_{u^*}u^* \qquad \boldsymbol{x}_{n+2} = 2\bar{\boldsymbol{x}} - \mu + sc_{u^*}u^*, \quad \text{where}$$

$$u^* = \frac{\bar{\boldsymbol{x}} - \mu}{\|\bar{\boldsymbol{x}} - \mu\|} \quad \text{and} \quad c_{u^*} = (u^{*\mathsf{T}} S^{-1} u^*)^{-1/2}.$$

Theorem 4.2

$$\lim_{s\to\infty} \frac{2ns^2}{(n+2)^2}\Big(-2\log \mathcal{R}^{**}(\mu)\Big) = T^2(\mu)$$

Emerson & O (2009)



Emerson & O (2009)

Comments

1) More examples in the article
2) Good calibration for distributions with shorter tails
3) High kurtosis is harder
4) Even there the calibration is almost linear so a Bartlett correction could help a lot
5) Exact nonparametric CI.s for the mean are unobtainable Bahadur & Savage (1956)

Infinitely many estimating equations

Symmetry:

$$\mathbb{E}(|X-\mu|^k \text{sign}(X-\mu)) = 0, \quad \forall k \geq 1$$

Independence:

$$\mathbb{E}(\phi(X)\psi(Y)) = \mathbb{E}(\phi(X))\mathbb{E}(\psi(Y)), \quad \forall \phi(\cdot), \psi(\cdot)$$

EL with sparse likelihoods

Replacing $-2\sum_{i=1}^n \log(nw_i)$ by some multiple of $\sum_{i=1}^n |nw_i - 1|$ should lead to many data points with $w_i = 1/n$ exactly. The exceptions may be interpretable.

L_∞ version

$$\max_{1\le i\le n} |nw_i - 1|$$

Using this criterion should often lead to a subset of observations with w_i at some maximal level and another subset at a minimal level. That pattern may be revealing.

Profiling for regression

Maximize $\sum_{i=1}^n \log(nw_i)$ subject to $w_i \ge 0$ $\sum_i w_i = 1$

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

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

Not quite convex optimization

The free variables are β_k for $k \ne j$ as well as $w_1, \dots, w_n$.

The computational challenge comes from **bilinearity** of the constraint.

If β is held fixed the normal equation constraint is linear in w and vice versa.

Multisample EL

Chapter 11.4 of the text “Empirical likelihood” looks at a multi-sample setting. Observations $\boldsymbol{X}_i \overset{\text{iid}}{\sim} F$ for $i = 1, \dots, n$ independent of $\boldsymbol{Y}_j \overset{\text{iid}}{\sim} G$ for $j = 1, \dots, m$. The likelihood ratio is

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

with $u_i \ge 0$, $v_j \ge 0$, $\sum_i u_i = 1$, $\sum_j v_j = 1$ and

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

For example: $h(X, Y, \theta) = 1_{X-Y>\theta} - 1/2$. The computational problem is a challenge. The log likelihood is convex but constraint (1) is bilinear. So computation is awkward.

Regression again

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

Estimating equations*

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

Normal equations

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

In principle we let $\boldsymbol{z}_i = \boldsymbol{z}_i(\beta) \equiv (y_i - \boldsymbol{x}_i^\mathsf{T}\beta)\boldsymbol{x}_i \in \mathbb{R}^d$, adjoin $\boldsymbol{z}_{n+1}$ and $\boldsymbol{z}_{n+2}$, and carry on.

*residuals $\epsilon = y - \boldsymbol{x}^\mathsf{T}\beta$ are uncorrelated with $\boldsymbol{x}$.
They have mean zero too, when as usual, $\boldsymbol{x}$ contains a constant.

Regression hull condition

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

$$\mathcal{P} = \mathcal{P}(\beta) = \{\boldsymbol{x}_i \mid y_i - \boldsymbol{x}_i^\mathsf{T}\beta > 0\} \qquad \boldsymbol{x} \text{ with pos resid}$$

$$\mathcal{N} = \mathcal{N}(\beta) = \{\boldsymbol{x}_i \mid y_i - \boldsymbol{x}_i^\mathsf{T}\beta < 0\} \qquad \boldsymbol{x} \text{ with neg resid}$$

Convex hull condition O (2000)

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

For $\boldsymbol{x}_i = (1, t_i)^\mathsf{T} \in \mathbb{R}^2$ $\quad \mathcal{P}$ and $\mathcal{N}$ are intervals in $\{1\} \times \mathbb{R}$.

Converse

Suppose that $\tau \notin \{t_1, \dots, t_n\}$ and

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

Suppose also that

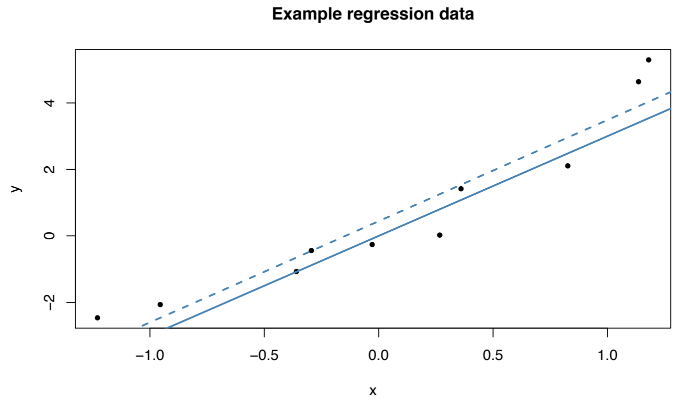
$$\sum_i w_i \begin{pmatrix} 1 \\ t_i \end{pmatrix} (y_i - \beta_0 - \beta_1 t_i) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Then

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

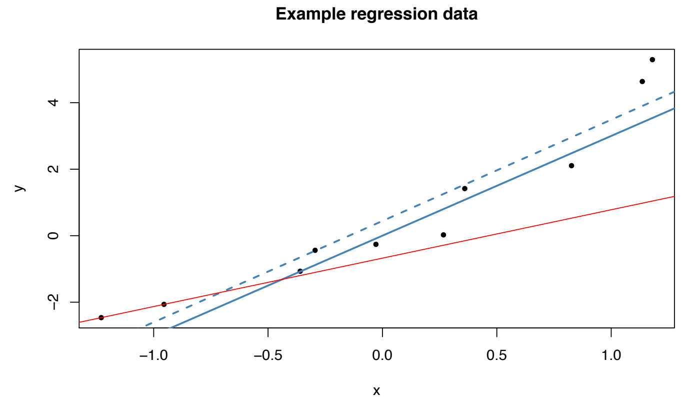
But $(y_i - \beta_0 - \beta_1 t_i)(t_i - \tau) > 0\ \forall i$

Therefore the hull condition is necessary.

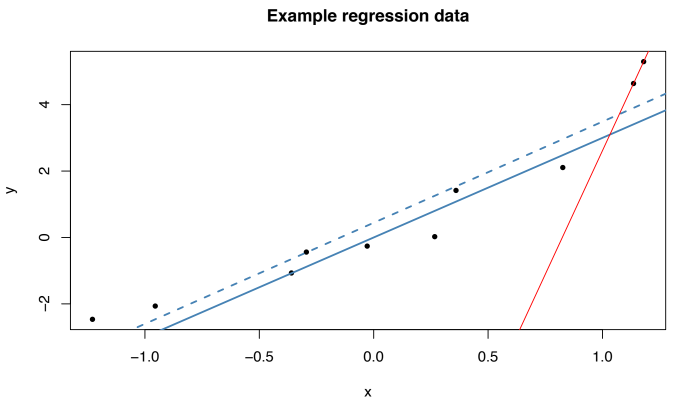


$Y = \beta_0 + \beta_1 X + \sigma\epsilon \quad \beta = (0, 3)^\mathsf{T}, \sigma = 1$

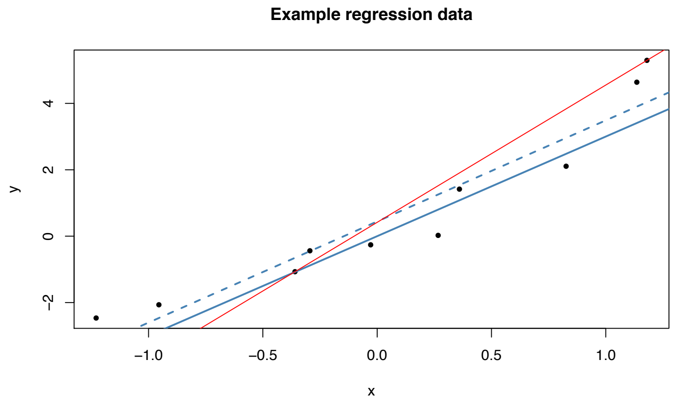
β solid $\quad \hat\beta$ dashed



Red line is on boundary of set of (β_0, β_1) with positive empirical likelihood



Another boundary line.

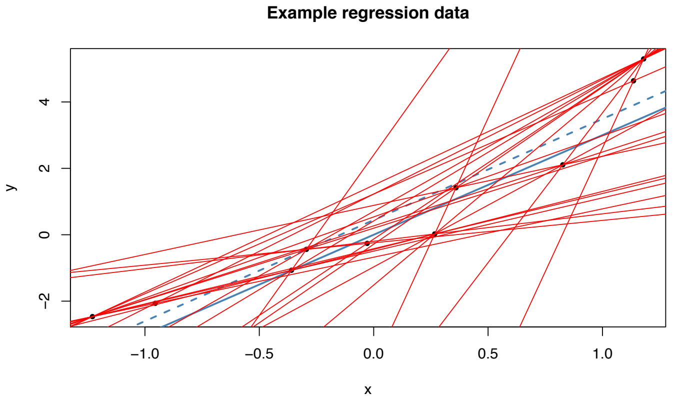


Yet another boundary line.

Left side has positive residuals; right side negative.

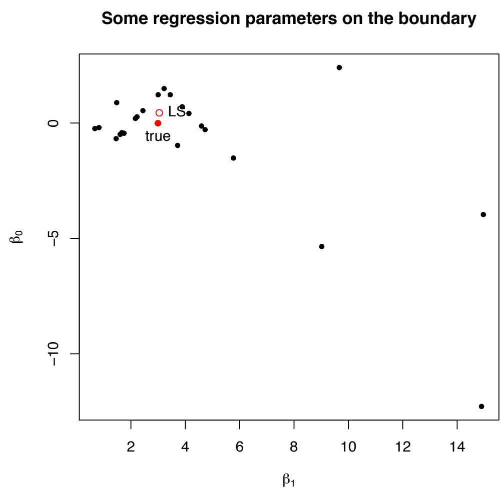
Wiggle it up and point 3 gets a negative residual $\Longrightarrow$ ok.

Wiggle down $\Longrightarrow$ NOT ok.



All the boundary lines that interpolate two data points.

They are a subset of the boundary.



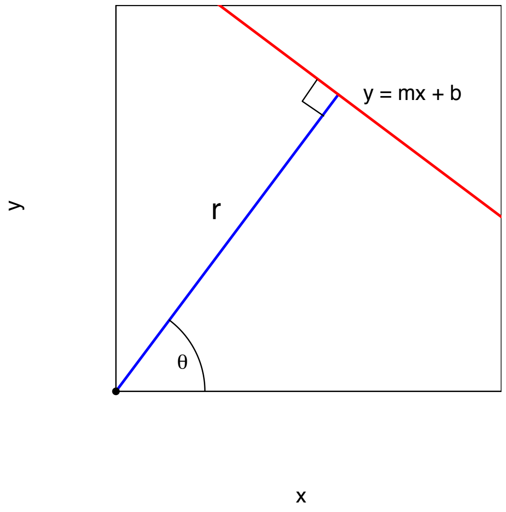
Boundary points (β_0, β_1). Region is not convex.

It **is** convex in β_0 (vertical) for fixed β_1 (horizontal).

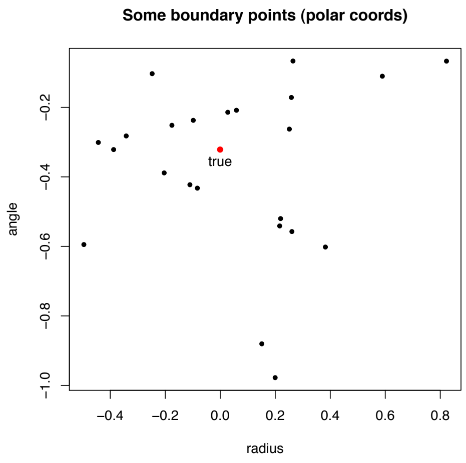
What is a convex set of lines?

- convex set of (β_0, β_1)?
- convex set of (ρ, θ)? (polar coordinates)
- convex set of (a, b) $(ax + by = 1)$?

Polar coordinates of a line



Boundary pts in polar coords



Not convex here either.

Intrinsic convexity

There is a geometrically intrinsic notion for a convex set of linear flats.

J. E. Goodman (1998) "When is a set of lines in space convex?"

Maybe $\cdots$ that can support some computation.

Dual definition

The set of flats that intersects a convex set $C \subset \mathbb{R}^d$ is a convex set of flats.

So is the set of flats that intersect **all of** $C_1, \ldots, C_k \subset \mathbb{R}^d$ for convex C_j.

Convex functions

This notion of convex set does not yet seem to have a corresponding notion of convex function. There could be quasi-convex functions, those where the level sets are convex. But quasi-convexity is much less powerful computationally than convexity.

Acknowledgments

1) Sarah Emerson and Jiahua Chen for discussions

2) National Science Foundation for support

3) Univ.s Geneva, Lausanne, Neuchatel, Fribourg, Bern, EPFL

4) Valérie Chavez-Demoulin

5) Elvezio Ronchetti

6) Mervet Cluzeau

Merci et au revoir!