

# Self-concordance for empirical likelihood

Art B. Owen, Stanford University

# Motivation

Dylan Small and Dan Yang (2012) found a case where my old Levenberg-Marquardt iterations failed. Plain step reduction works better.

## New optimization is

- 1) low dimensional
- 2) convex
- 3) unconstrained
- 4) **self-concordant**

The new ingredient is self-concordance (described below)

It gives mathematical guarantees of convergence.

Prior to convergence it lets us bound sub-optimality

## Also

A quartic log likelihood Corcoran (1998) is also self-concordant.

# Empirical Likelihood

Provides likelihood inferences without assuming a parametric family

For data  $X_i \stackrel{\text{iid}}{\sim} F$

$$\begin{array}{ll} L(F) = \prod_{i=1}^n F(\{X_i\}) & \text{Likelihood} \\ \hat{F} = \frac{1}{n} \sum_{i=1}^n \delta_{X_i} & \text{Nonparametric MLE} \\ R(F) = \prod_{i=1}^n n w_i, \quad w_i \equiv F(\{X_i\}) & \text{Empirical likelihood ratio} \end{array}$$

If  $L(F) > 0$  then  $w_i > 0$ . Convenient to assume  $\sum_{i=1}^n w_i = 1$  too.

Then we get a multinomial distribution on  $n$  items  $X_1, \dots, X_n$ .

# EL properties

Empirical likelihood inherits many properties from parametric likelihoods.

- Wilks style  $\chi^2$  limit distribution
- automatic shape selection for confidence regions
- Bartlett correctability DiCiccio, Hall, Romano
- Very high power Kitamura and Lazar & Mykland

Statistical assumptions: independence and bounded moments.

## Oddly

Having  $n - 1$  parameters for  $n$  observations does not lead to trouble.

# Empirical likelihood for the mean

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

Wilks-like:  $-2 \log(\mathcal{R}(\mu_0)) \xrightarrow{d} \chi_{(d)}^2$  allows confidence regions and tests

Estimating equations  $\mathbb{E}(m(X, \theta)) = 0$

$m(X, \theta) = X - \theta$	Mean
$m(X, \theta) = 1_{X < \theta} - 0.5$	Median
$m(X, Y, \theta) = (Y - X^\top \theta) X$	Regression
$m(X, \theta) = \frac{\partial}{\partial \theta} \log(f(X, \theta))$	MLE estimand

# Computation

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

Here  $Z_i = X_i - \mu_0$  or  $Z_i = m(X_i, \theta)$ .

## The hull

If 0 is not in the convex hull of  $Z_i$  then  $\log(\mathcal{R}(\cdot)) = -\infty$

## Lagrangian

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

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

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

## Therefore

$$w_i = \frac{1}{n} \frac{1}{1 + \lambda^\top Z_i}, \quad \text{for some } \lambda \in \mathbb{R}^d$$

# Finding $\lambda$

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

We have to solve

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

The dual

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

This function is convex in  $\lambda$  and,

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

Minimizing the dual maximizes the likelihood.

# $n$ constraints

$$\text{Recall: } \mathbb{L}(\lambda) = - \sum_{i=1}^n \log(1 + \lambda^\top Z_i)$$

Minimizer must have  $1 + \lambda^\top Z_i > 0$ ,  $i = 1, \dots, n$

This comes from  $w_i > 0$ .

Sharper

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

Therefore

$$1 + \lambda^\top Z_i > \frac{1}{n}, \quad i = 1, \dots, n$$



# Removing the constraints

Replace  $\log(x)$  by

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

where  $Q$  is quadratic with

$$Q(1/n) = \log(1/n)$$

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

$$Q''(1/n) = \log''(1/n)$$

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

Now minimize

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

Same optimum as  $\mathbb{L}$ . No constraints. Always finite.

# Newton steps

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

The Hessian is  $H(\lambda) \equiv \frac{\partial^2}{\partial \lambda \partial \lambda^\top} \mathbb{L}_*(\lambda)$

The Newton step is

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

## Further analysis

Our  $H$  is of the form  $J^\top J$  and  $g = J^\top \eta$

So the Newton step can be solved by least squares (more numerically stable)

## Step reductions

Newton steps still require some kind of step reduction methods. If there is not enough progress to the minimum, take a smaller multiple of  $s$ .

Levenberg-Marquardt: if the step gets too small start picking directions more near to  $-g$ .

# Small and Yang's example

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

$$0 = \mathbb{E}(Y - \beta_1 W - \alpha_1)$$

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

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

Residuals  $Y - \beta_1 W - \alpha_1$  and  $Y - (\beta_1 + \delta)W - \alpha_2$ .

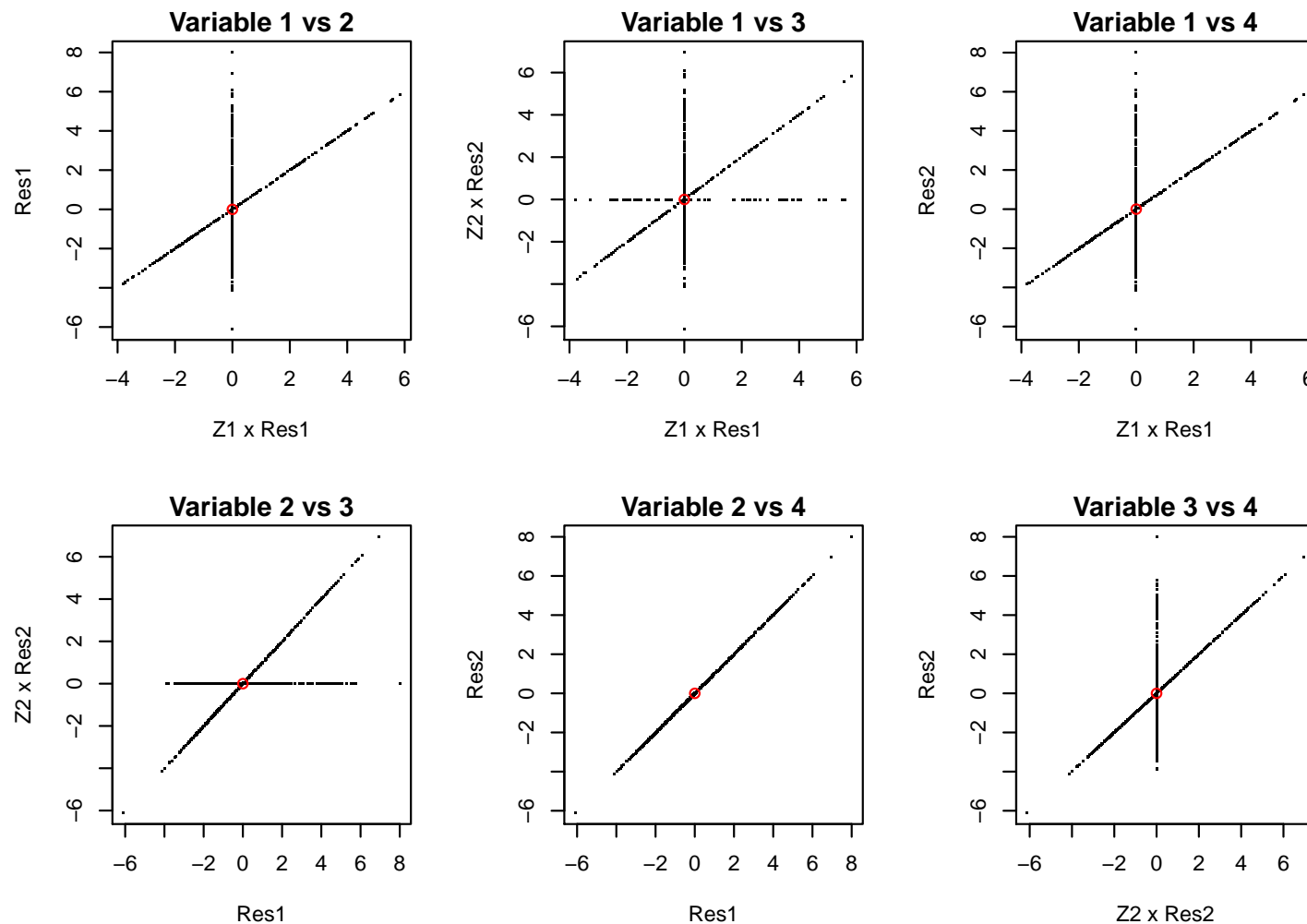
Instrumental variables  $Z_1, Z_2 \in \{0, 1\}$

Problem arose in a bootstrap sample.

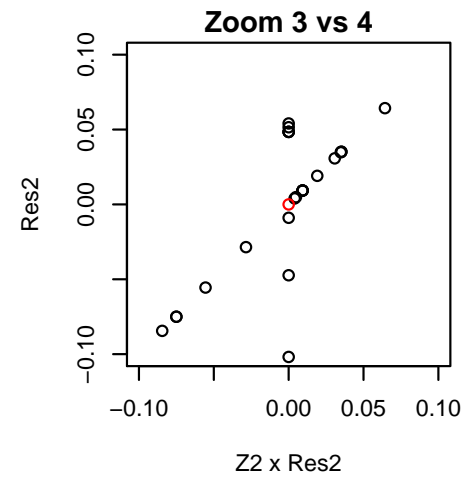
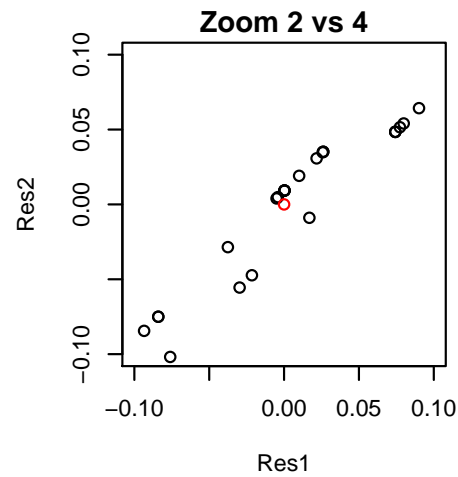
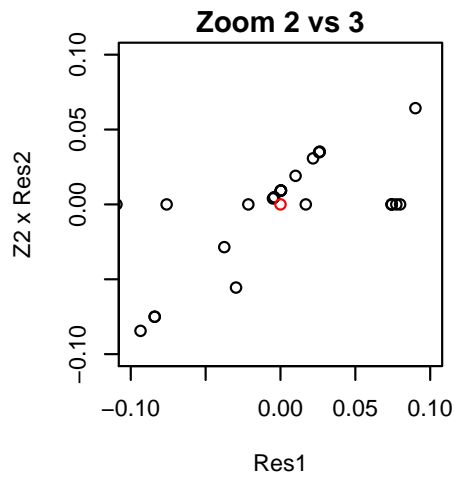
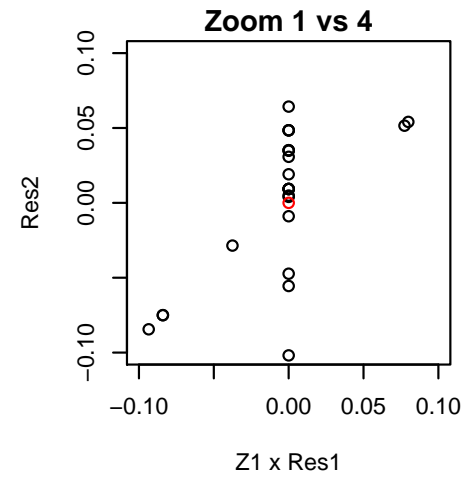
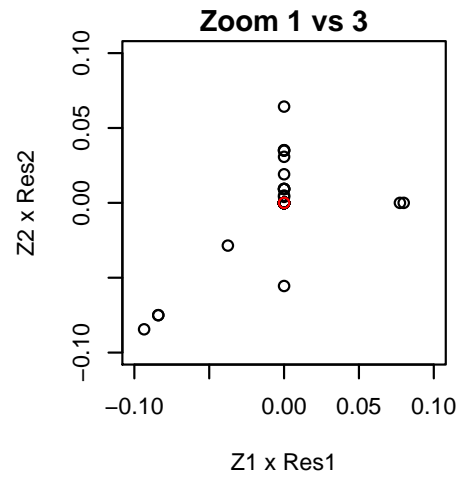
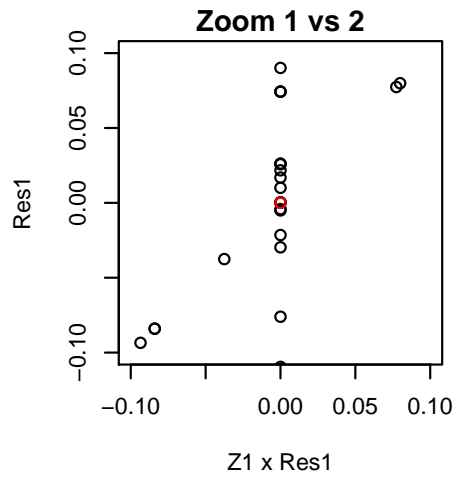
# Small and Yang's example

They needed to test the mean of 1000 points in  $\mathbb{R}^4$ .

The specific problem arose in an instrumental variables context.



# Zooming in



# True empirical log likelihood

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

Old algorithm got stuck; stepsize got small ad hoc Levenberg-Marquardt reductions did not help.

They used step reducing line search instead.

# Self-concordance

A convex function  $g$  from  $\mathbb{R}$  to  $\mathbb{R}$  is **self-concordant** if

$$|g'''(x)| \leq 2g''(x)^{3/2} \quad \text{N.B. } g'' \geq 0$$

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

A convex function  $g$  from  $\mathbb{R}^d$  to  $\mathbb{R}$  is self-concordant if

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

is a self-concordant function of  $t \in \mathbb{R}$ .

## Implications

The Hessian of self-concordant  $g(\mathbf{x})$  cannot change too rapidly with  $\mathbf{x}$ .

Newton updates with line search step-reduction are guaranteed to converge.

Also the Newton decrement (below) bounds the suboptimality.

## The 2 is not essential

If  $|g'''(x)| \leq Cg''(x)^{3/2}$  then  $\frac{C^2}{4}g$  is self-concordant.

# Backtracking Newton

- 1) Select starting point  $\boldsymbol{x}$
- 2) Repeat until Newton decrement  $\nu(\boldsymbol{x})$  below tolerance
  - a)  $\boldsymbol{s} \leftarrow -H(\boldsymbol{x})^{-1}g(\boldsymbol{x}), \quad t \leftarrow 1$
  - b) While  $f(\boldsymbol{x} + t\boldsymbol{s}) > f(\boldsymbol{x}) + \alpha t\boldsymbol{s}^\top g$ 
    - i)  $t \leftarrow t \times \beta$
- 3)  $\boldsymbol{x} \leftarrow \boldsymbol{x} + t\boldsymbol{s}$

Guaranteed convergence if

$\alpha \in (0, 1/2), \beta \in (0, 1), f$  bounded below, sublevel set of  $\boldsymbol{x}$  is closed

Newton decrement

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

If  $f$  is strictly convex self-concordant and  $\nu(\tilde{\boldsymbol{x}}) \leq 0.68$  then

$$\inf_{\boldsymbol{x}} f(\boldsymbol{x}) \geq f(\tilde{\boldsymbol{x}}) - \nu(\tilde{\boldsymbol{x}})^2$$



# Chen, Sitter, Wu

- Biometrika (2002)
- Use backtracking line search with step halving when objective not improved (i.e., improvement factor  $\alpha = 0$  and step factor  $\beta = 1/2$ )
- Show convergence via results in Polyak (1987)
- Starts  $k$ 'th search at size  $t = (k + 1)^{-1/2}$ .
- Starting with  $t < 1$  will slow Newton from quadratic convergence. They observe that starting at  $t = 1$  works.

## Back to $\mathbb{L}_*$

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

$\log_*$  is self-concordant on  $(-\infty, 1/n)$  and on  $(1/n, \infty)$ .

But it lacks a third derivative at  $1/n$

Hence not self-concordant.

# Higher order approximations

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

Taylor approx to  $-\log$  at  $\epsilon$

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

$k = 2$  Convex but not self-concordant (fails at  $\epsilon$ )  $-\log_{(2)} = -\log_*$

$k = 3$  Not even convex

$k = 4$  Convex and self-concordant



# Back to the example

Self-concordant version also gets  $\log \mathcal{R}() = -399.6937$

Newton decrement

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

Estimate has  $\log(\mathcal{R})$  within  $\eta^2$  of true optimum.

I.e. good to within given precision.

# Sketch of proof

We need to show that  $h_4(y)$  is self-concordant on  $(-\infty, 0]$ .

- i.e.,  $|h_4'''| \leq 2(h_4')^{3/2}$
- Suffices to show  $h_4(\epsilon \times \cdot)$  self-concordant
- $h_4'''(t\epsilon) = \epsilon^{-3}(-2 + 6t)$
- $h_4''(t\epsilon) = \epsilon^{-2}((1 - t)^2 + t^2)$
- $\rho(t) \equiv \frac{|h_4'''(t\epsilon)|}{h_4''(t\epsilon)^{3/2}} = \frac{2 - 6t}{(t - 1)^2 + t^2}$  on  $t \leq 0$ .
- $\rho(0) = 2$
- $\rho'(t) \geq 0$  for  $t \leq 0$

So the ratio  $\rho$  increases to 2 as  $t \uparrow 0$

# Quartic log likelihood

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

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

## Properties

Bartlett correctable [Corcoran \(1998\)](#)

Match 4 derivatives & match 4 moments

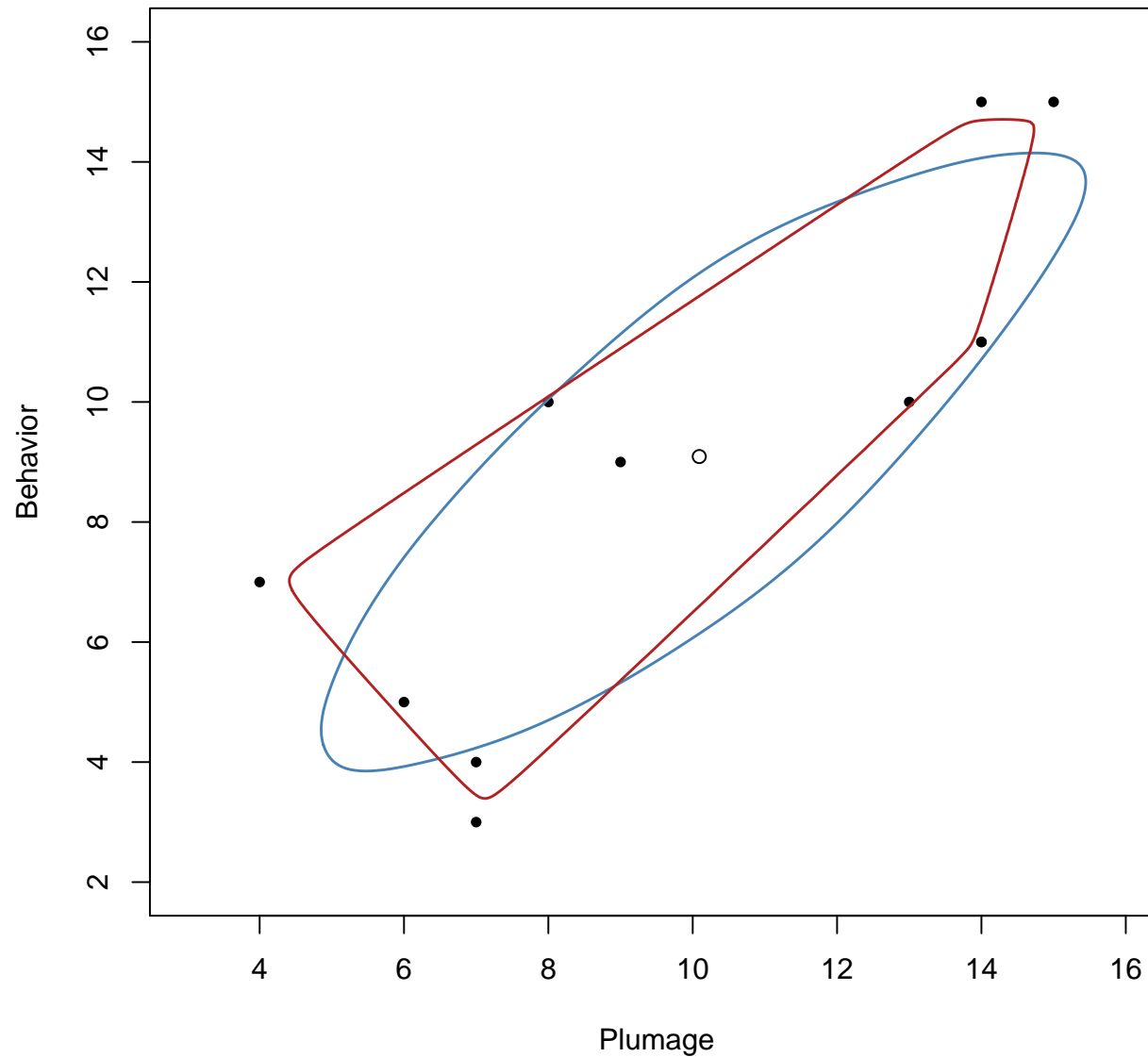
Self-concordant [O \(2013\)](#) [ $C = 3.92$  instead of  $C = 2$ ]

Convex confidence regions for the mean [O \(2013\)](#)

Lagrange multiplier for  $\sum w_i = 1$  cannot be eliminated.

Primal-dual algorithm in [Boyd & Vandenberghe](#) available

# Duck data



Extreme confidence region. Red  $\mathcal{R}$ ; Blue  $\mathcal{R}_Q$

Larsen & Marx (1986)

# Next thoughts

Maybe it is not necessary to enforce  $1 + \lambda^\top Z_i > 1/n$

Avoid piece-wise pseudo-logarithm altogether

Step reduction keeps  $1 + \lambda^\top Z_i > 0$

–  $\sum_{i=1}^n \log(1 + \lambda^\top Z_i)$  also self-concordant

Simpler, but

$\log(z)$  may be slightly worse conditioned than  $z^4$

Maximizing over nuisance parameters might be easier without linearly constraining  $\lambda$



# Time permitting . . .

Some computational challenges.

# Profiling for regression

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

$$\sum_i w_i (Y_i - \mathbf{x}_i^\top \beta) \mathbf{x}_i = 0$$

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

## Not quite convex optimization

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

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

If  $\beta$  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  $\mathbf{X}_i \stackrel{\text{iid}}{\sim} F$  for  $i = 1, \dots, n$  independent of  $\mathbf{Y}_j \stackrel{\text{iid}}{\sim} G$  for  $j = 1, \dots, m$ . The likelihood ratio is

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

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

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

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

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

Estimating equations\*

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

Normal equations

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

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

\*residuals  $\varepsilon = y - \mathbf{x}^T \boldsymbol{\beta}$  are uncorrelated with  $\mathbf{x}$ .

They have mean zero too, when as usual,  $\mathbf{x}$  contains a constant.

# Regression hull condition

$$\mathcal{R}(\beta) = \sup \left\{ \prod_{i=1}^n n w_i \mid w_i \geq 0, \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i (y_i - \mathbf{x}_i^\top \beta) \mathbf{x}_i = 0 \right\}$$

$$\mathcal{P} = \mathcal{P}(\beta) = \{ \mathbf{x}_i \mid y_i - \mathbf{x}_i^\top \beta > 0 \} \quad \mathbf{x} \text{ with pos resid}$$

$$\mathcal{N} = \mathcal{N}(\beta) = \{ \mathbf{x}_i \mid y_i - \mathbf{x}_i^\top \beta < 0 \} \quad \mathbf{x} \text{ with neg resid}$$

## Convex hull condition $\mathcal{O}$ (2000)

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

For  $\mathbf{x}_i = (1, t_i)^\top \in \mathbb{R}^2$      $\mathcal{P}$  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

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

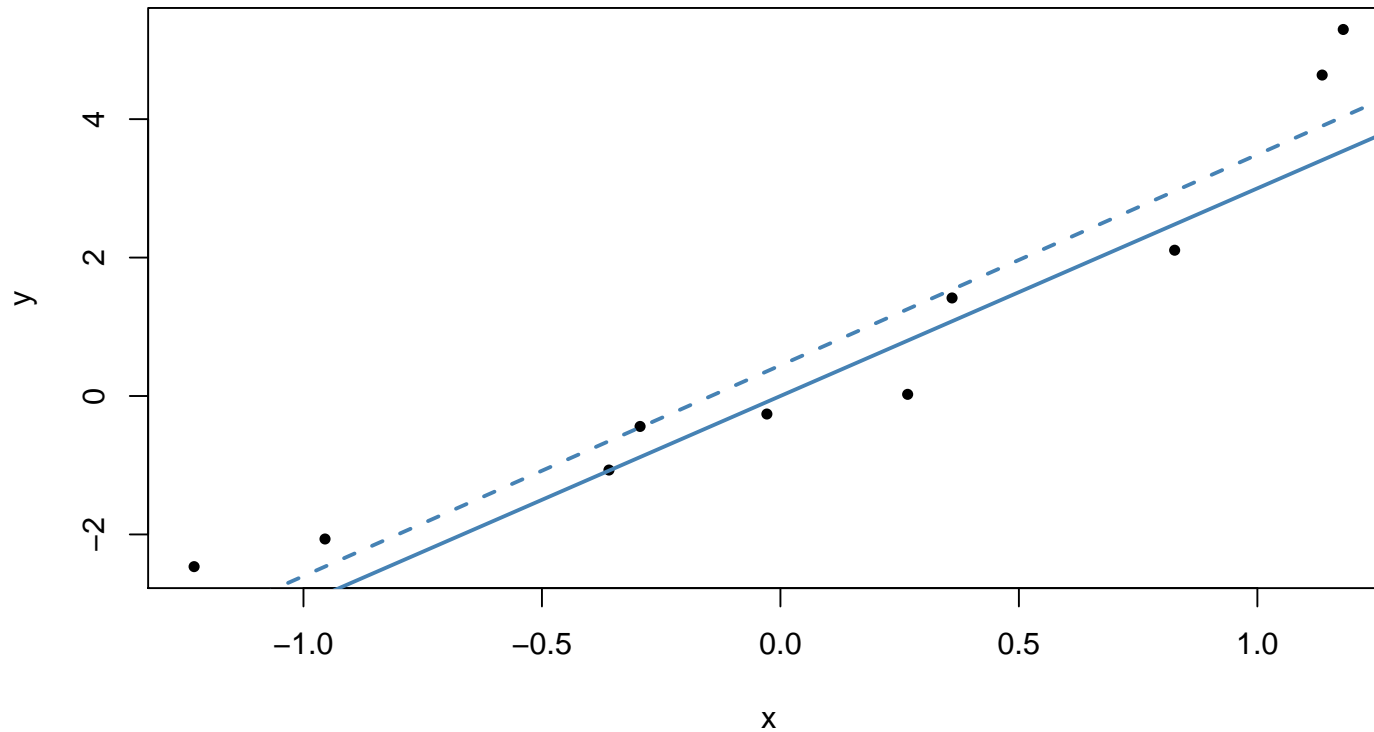
Then

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

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

Therefore the hull condition is **necessary**.

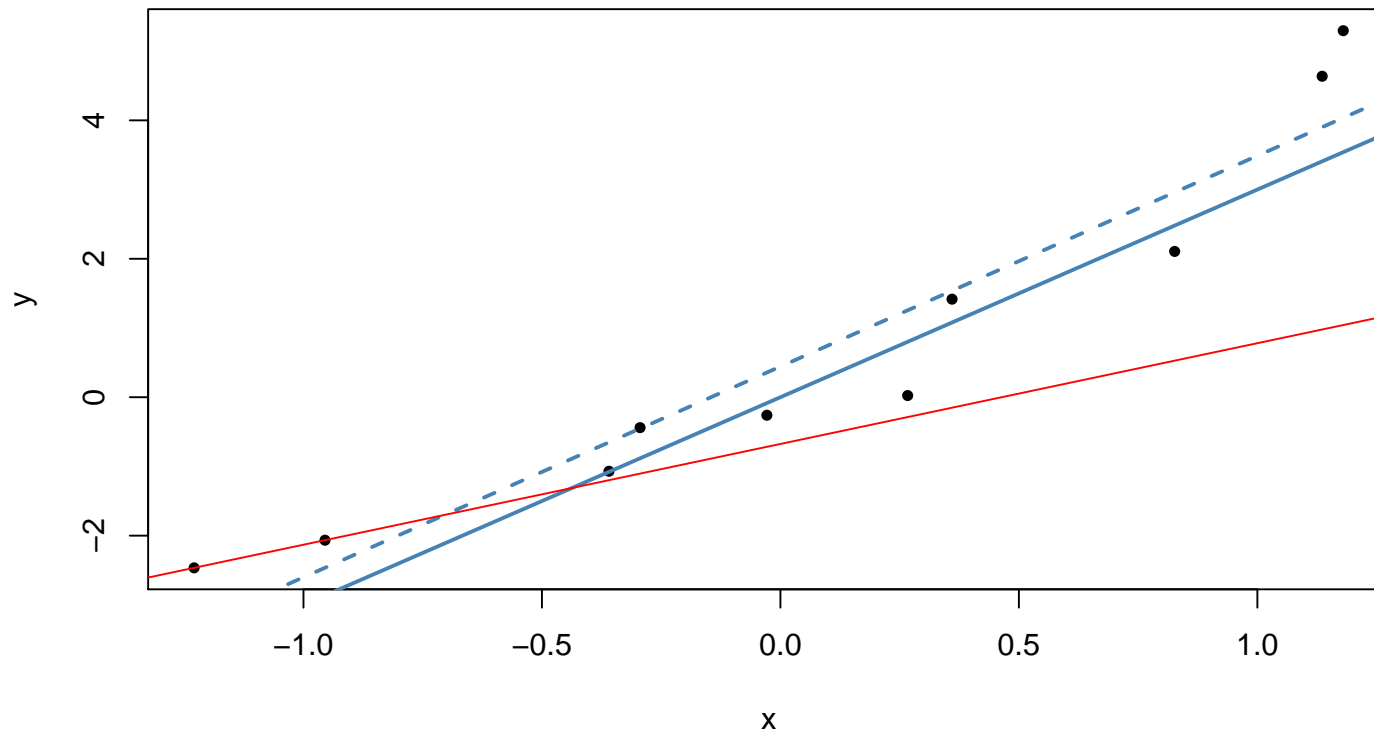
Example regression data



$$Y = \beta_0 + \beta_1 X + \sigma \varepsilon \quad \beta = (0, 3)^\top, \sigma = 1$$

$\beta$  solid    $\hat{\beta}$  dashed

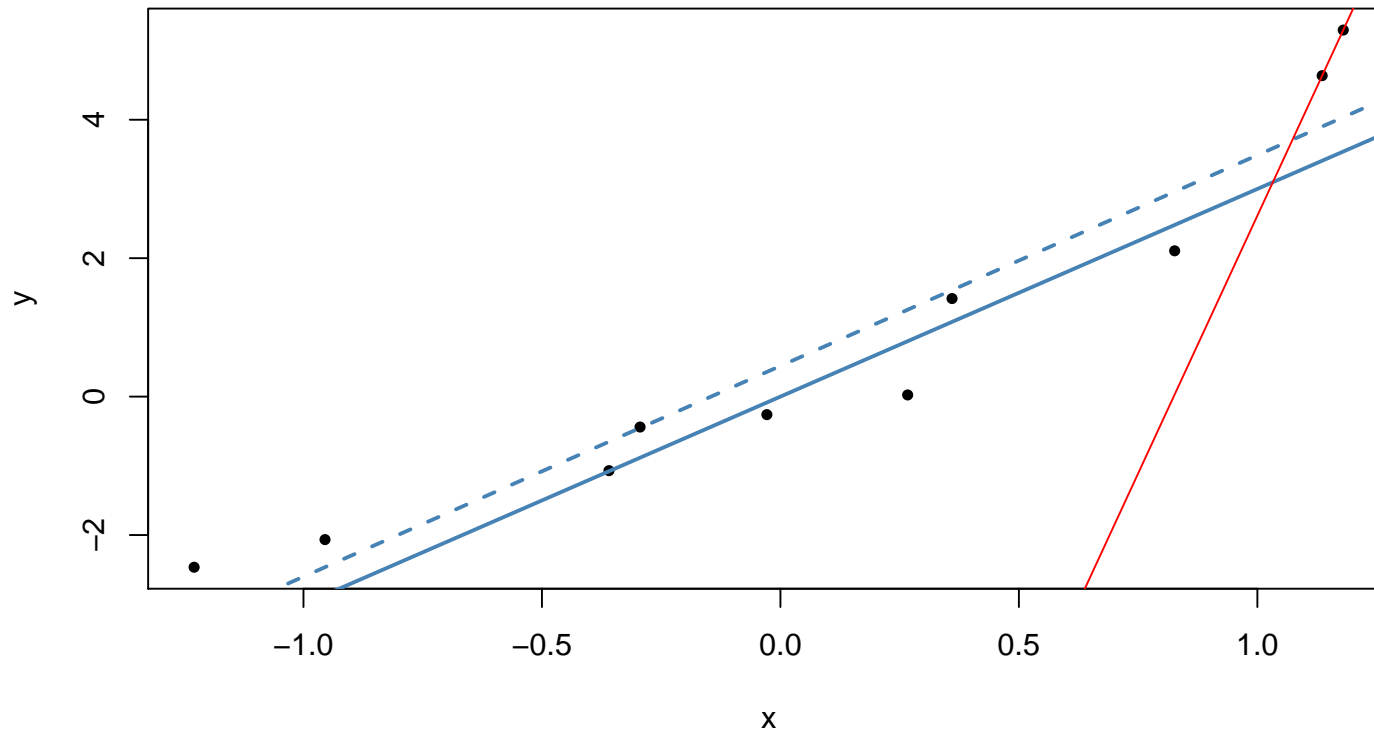
Example regression data



Red line is on boundary of set of  $(\beta_0, \beta_1)$  with positive empirical likelihood

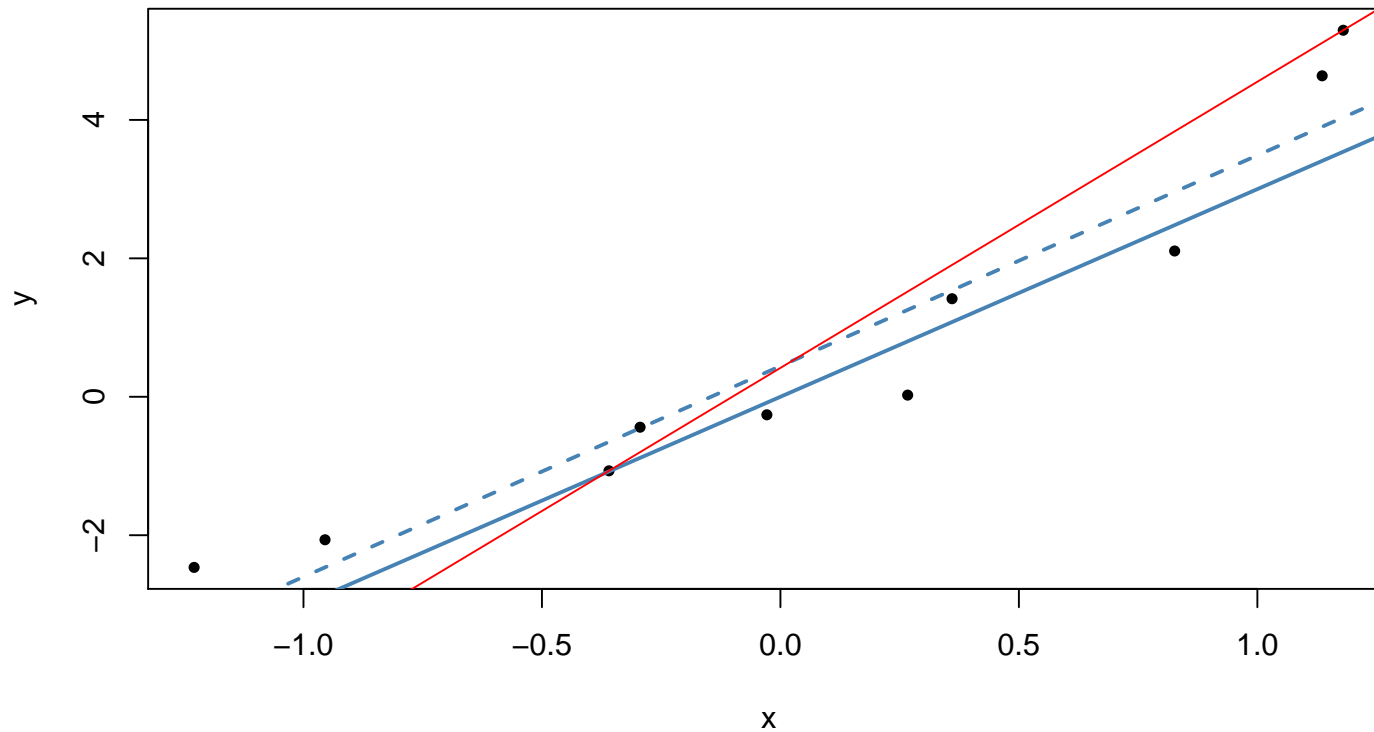


Example regression data



Another boundary line.

Example regression data



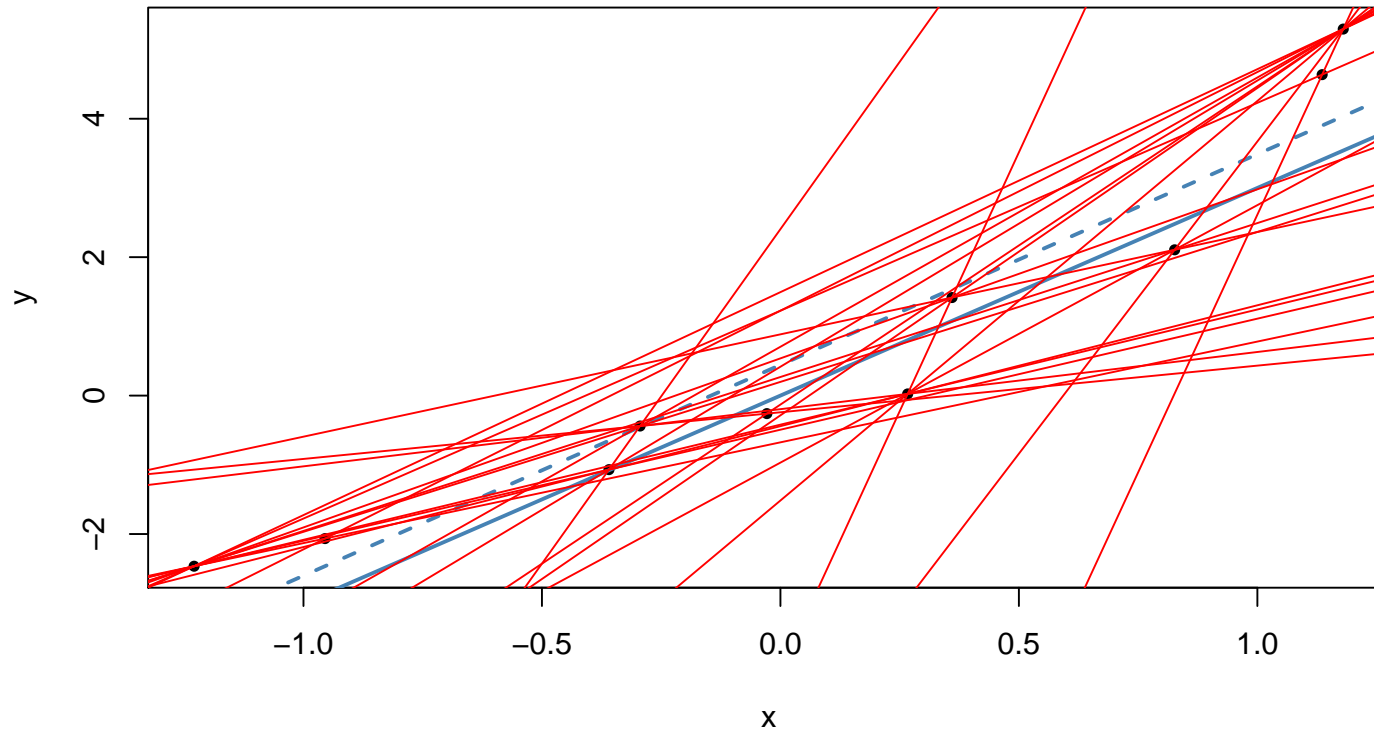
Yet another boundary line.

Left side has positive residuals; right side negative.

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

Wiggle down  $\implies$  NOT ok.

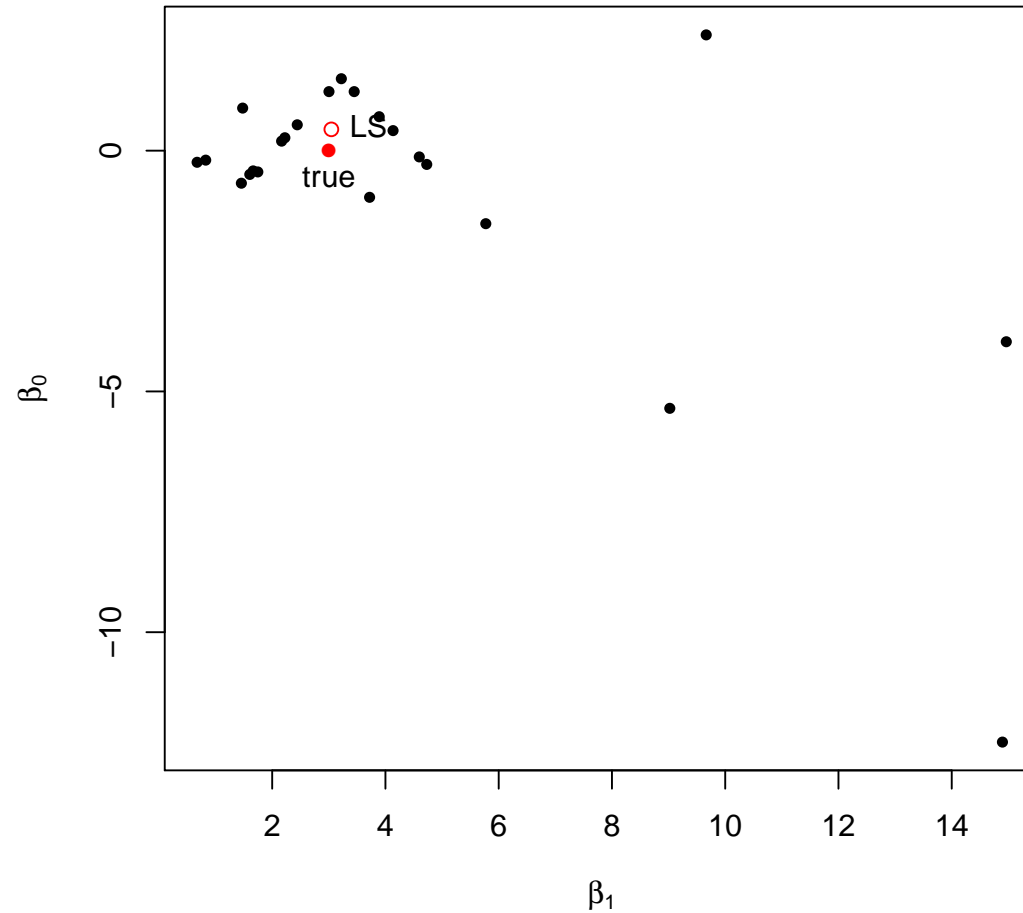
Example regression data



All the boundary lines that interpolate two data points.

They are a subset of the boundary.

## Some regression parameters on the boundary



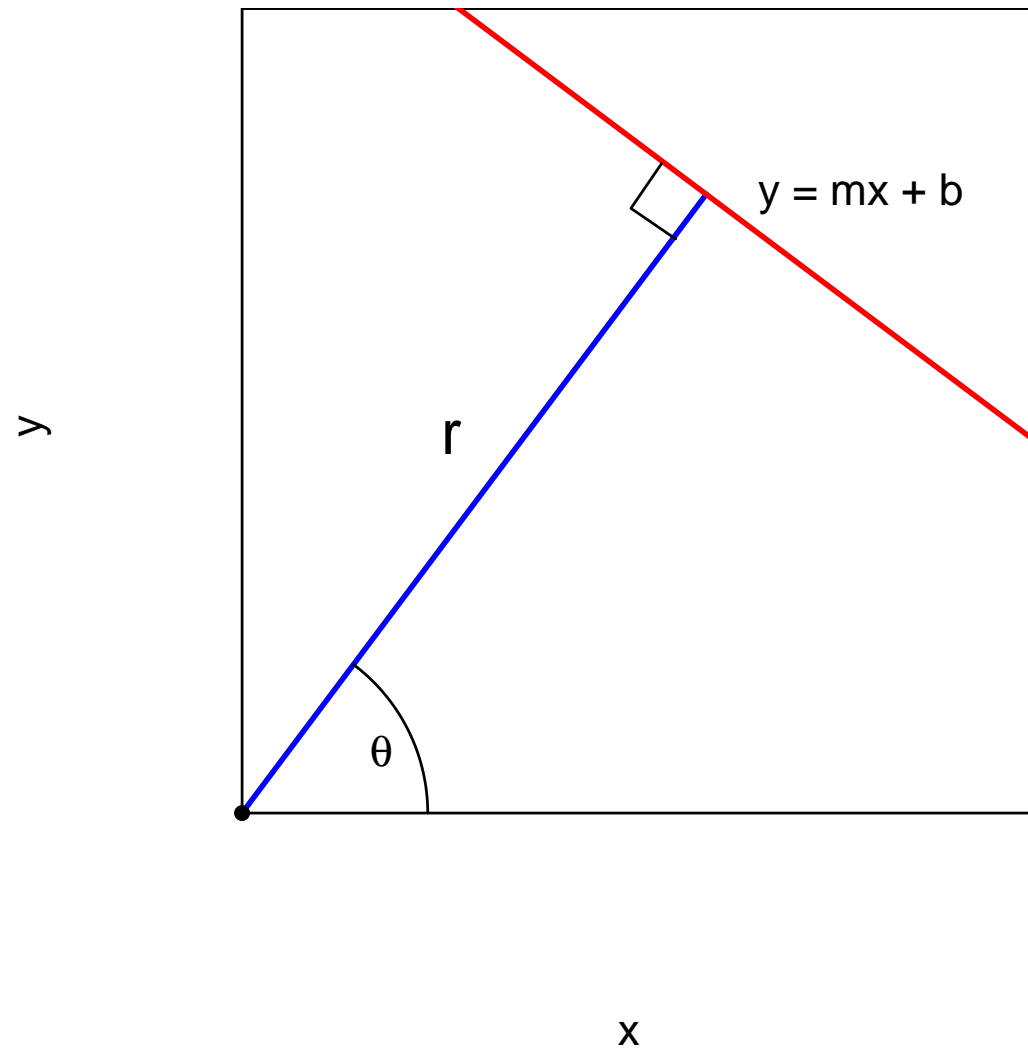
Boundary points  $(\beta_0, \beta_1)$ . Region is not convex.

It **is** convex in  $\beta_0$  (vertical) for fixed  $\beta_1$  (horizontal).

# What is a convex set of lines?

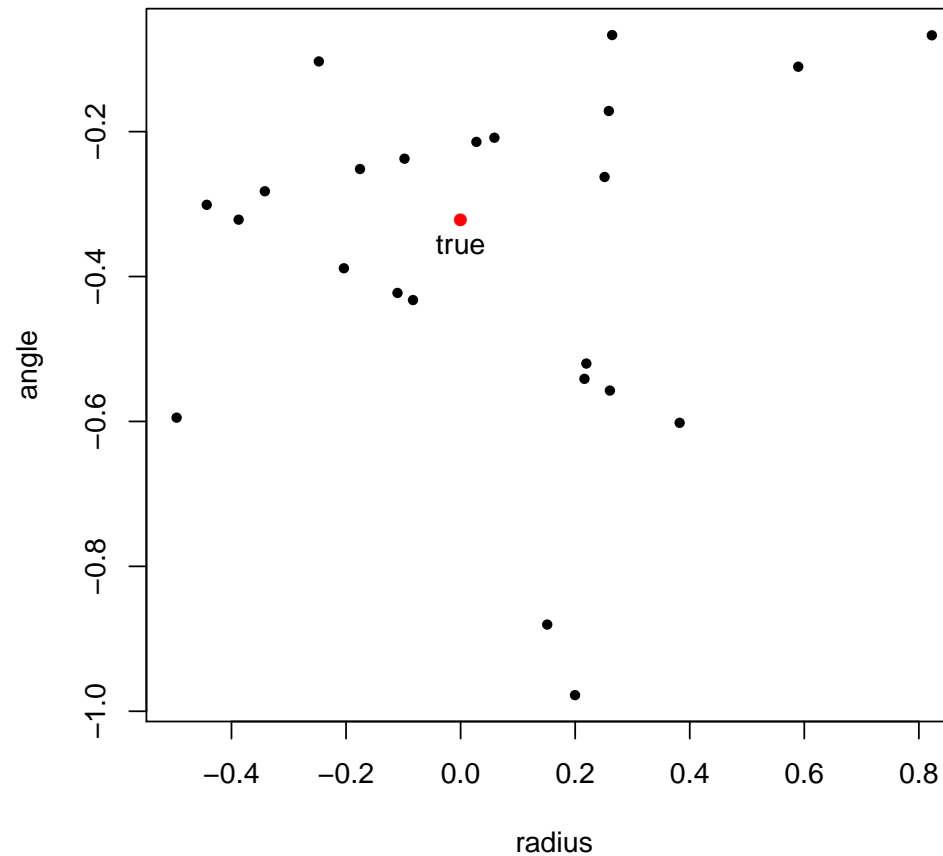
- convex set of  $(\beta_0, \beta_1)$ ?
- convex set of  $(\rho, \theta)$ ? (polar coordinates)
- convex set of  $(a, b)$  ( $ax + by = 1$ )?

# Polar coordinates of a line



# Boundary pts in polar coords

Some boundary points (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 . . . 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**  $C_1, \dots, C_k \subset \mathbb{R}^d$  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.



# Thanks

- 1) Dylan Small and Dan Yang
- 2) Jiahua Chen, sharing an early paper
- 3) Canadian Journal of Statistics
- 4) Changbao Wu & the CJS committee
- 5) Caroline Petit-Turcotte & Pierre-Jérôme Bergeron, translation
- 6) NSF DMS-0906056