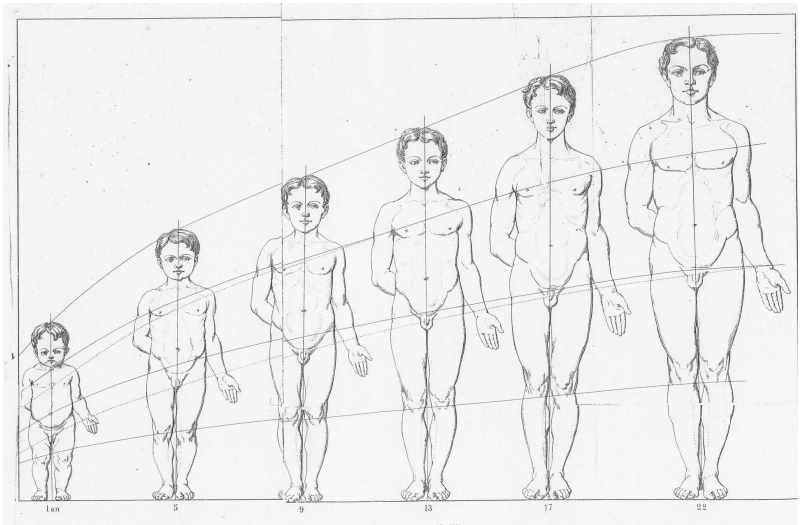
Beyond the Average Man: The Art of Unlikelihood

Roger Koenker

University of Illinois, Urbana-Champaign

FEMES/Taipei: July, 2007



An Outline

- A Brief Biography of the Average Man
  - Science Begins with Measurement (Kelvin)
- Convergence versus Diversity
  - Heterogeneity and the Regression Fallacy
- Skepticism and Unlikelihoods
  - Robustness: Local versus Global Models
  - L_1 Regularization as Occam's (New) Razor
- Progress in (Quantile) Regression: Four Directions
  - Additive Non-parametric Models
  - Longitudinal (Panel) Data
  - Nonlinear Filtering
  - Multivariate Conditional Quantiles

The Father of the Average Man

Adolphe Quetelet (1796 - 1874)

Quetelet's (1835) *Sur l'Homme (On Man)* invented the "Average Man".

Through systematic measurement and relentless averaging Quetelet sought to extract man's essential qualities: social, economic, aesthetic, and moral.

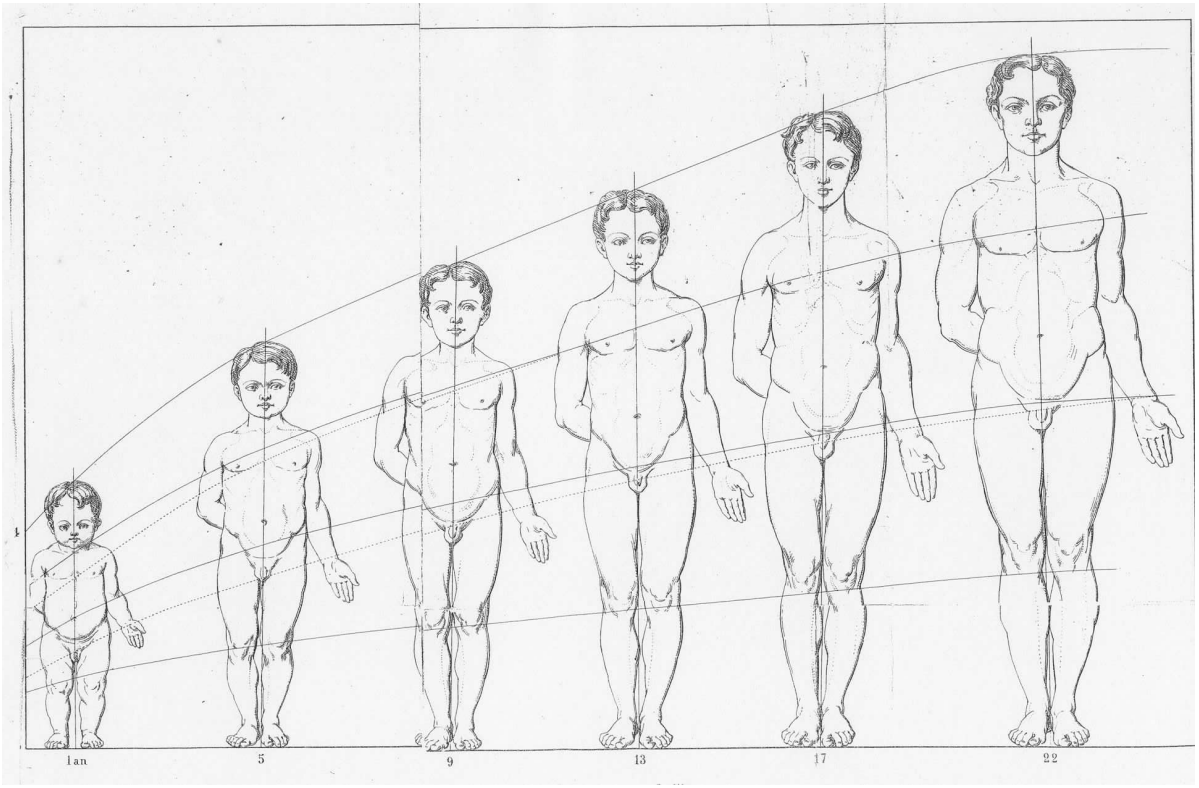
The Mother of the Average Man

Florence Nightingale (1820 - 1910)

Heroine of the Crimean War, Patron Saint of Nurses, admirer of Quetelet, and champion of the scientific, i.e. statistical, study of society.

To Nightingale every piece of legislation was an experiment in the laboratory of society deserving study and demanding evaluation.

Portrait of the Average Man



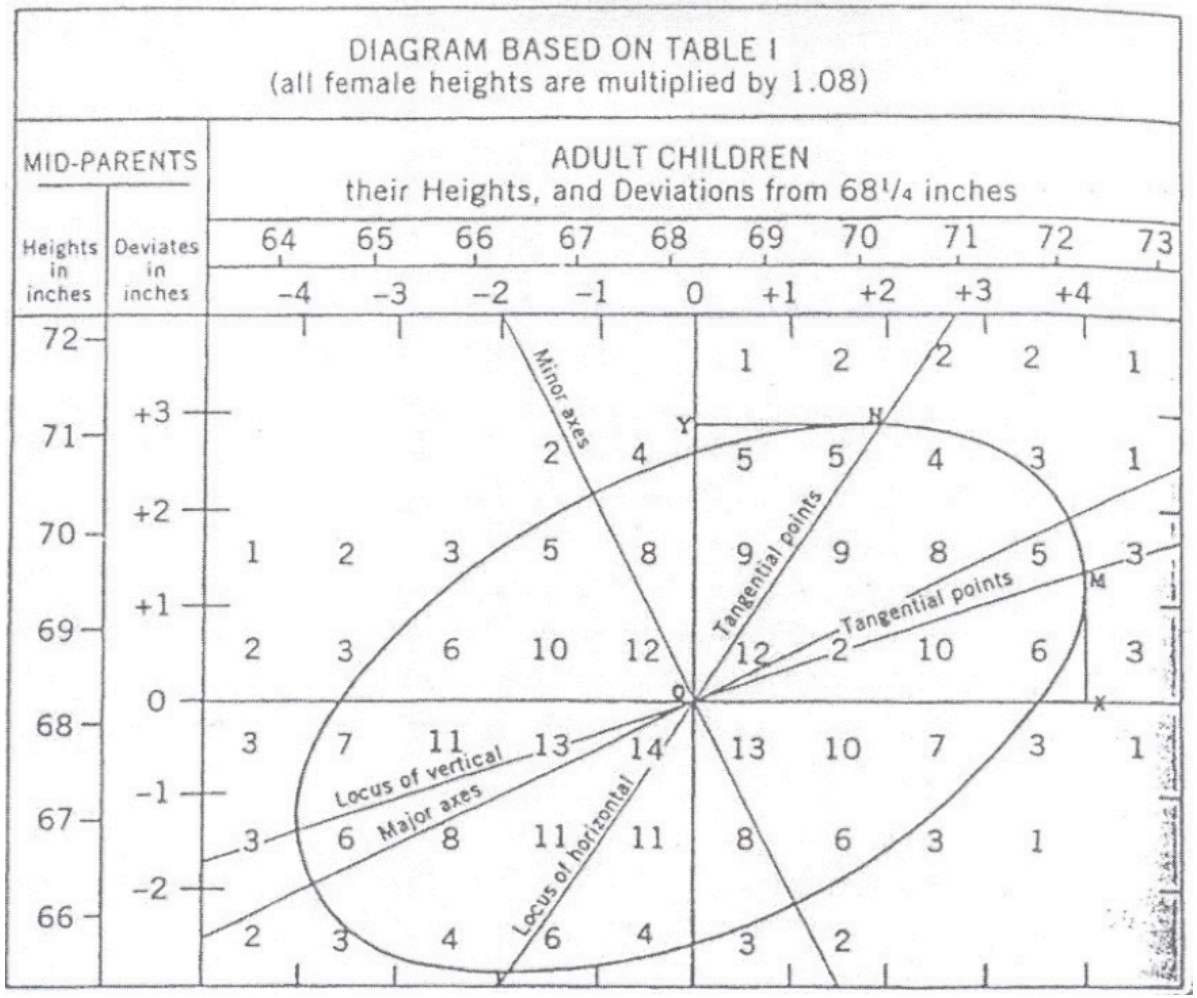
Quetelet's (1871) Anthropométrie

The Schoolmaster of the Average Man

Francis Galton (1822 - 1911)

Victorian adventurer, polymath and inveterate data collector and analyst. Galton's (1885) discovery of regression and correlation paved the way for the Average Man as a scientific construct.

Galton's Regression to Mediocrity



Galton's Discovery of Regression (1885)

Galton's "Most Likely" Children

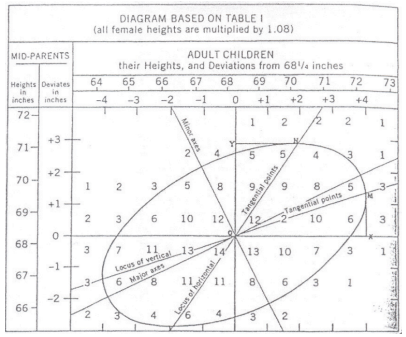
Children are more mediocre than their parents:

- Children of midparents 3 inches taller than average are most likely to be 2 inches taller than average.
- Children of midparents 3 inches shorter than average are most likely to be 2 inches shorter than average.

But parents are also more mediocre than their children:

- Midparents of children 3 inches taller than average are most likely to be 2 inches taller than average.
- Midparents of children 3 inches shorter than average are most likely to be 2 inches shorter than average.

Galton's Bivariate Gaussian Sandbox



Given the bivariate Gaussian form of the joint distribution of heights:

- Conditional mean heights are linear in midparents' height.
- Conditional density of height is symmetric, so
- Means, medians and modes (most likely heights) coalesce.
- Conditional scale and shape are invariant.

It all seems a little "too good to be true" – and even for heights it is, at best, only a decent first approximation.

Convergence versus Diversity

Galton's beautiful figure should come with a health warning:

Prolonged exposure to regression
may
cause confusion about convergence.

Quetelet already exhibits symptoms of this malady:

> *"The more knowledge is diffused, so much the more do the deviations from the average disappear; and the more, consequently, do we tend to approach that which is beautiful, that which is good." [On Man, p.108]*

Secrist and the Regression Fallacy

Stigler's (1996) paper "The History of Statistics in 1933" provides a classic economic case study:

- In 1933 Horace Secrist published *The Triumph of Mediocrity* that advanced the alarming thesis that American business was inexorably tending toward mediocrity: Groups of highly profitable firms, over time, on average lost profitability, less profitable firms gained.
- Harold Hotelling wrote a devastating review in JASA explaining that this was simply Galton's regression effect: firms were not – nor were human heights – converging to their respective means.
- Milton Friedman, a student of Hotelling's in 1933, felt compelled to write an essay for the JEL in 1992 called "Will Old Fallacies Ever Die?" to combat similar foolishness in contemporary economics.

The Demise of the Average Man

Robert Doisneau's (1952) "[Representative] Consumer"

The Average Man, rehabilited in recent years as The Representative Agent has not aged well. Here he is – down and out – living in Paris.

We have converged – not the the ideal world of Quetelet's Average Man, but to the realization that we need to take heterogeneity more seriously.

The Representative Agent – An Inconvenient Fiction

Heterogeneity is not the inessential cloud obscuring the ideal average man, it is precisely this diversity that we need to better understand.

- Diversity of tastes, beliefs, endowments is crucial – without it much of economic life comes to a standstill.
- Aggregation does not preserve economic rationality, the average man is not *homo economicus*: Sonnenschein, Debreu ...
- Treatment effects are often heterogeneous, and this is crucial to policy evaluation: Heckman, ...

Normal Skepticism and Unlikelihood

Before his conversion to the Gaussian faith, Galton was a confirmed nonparamet-nik. His obsessive data analysis relied almost exclusively on:

- The median as a measure of location,
- Half the interquartile range as "probable deviation."

The modern version of this is Tukey's boxplot and 5-Quantile summary:

$$Q(\tau) : \tau \in \{0, .25, .50, .75, 1\}$$

which serve to roughly summarize the form of the observed distribution.

Sufficiency of the mean and standard deviation require a dangerous leap of Gaussian faith, rarely justified when samples are moderately large.

Data Analysis versus Analysis of Data

Two principles compete for econometric attention:

- **Panzar Principle:** Are there conditions, however far-fetched, under which a given statistical analysis would have been optimal?
- **Tukey Principle:** Better an approximate answer to the right question, than an exact answer to the wrong question.

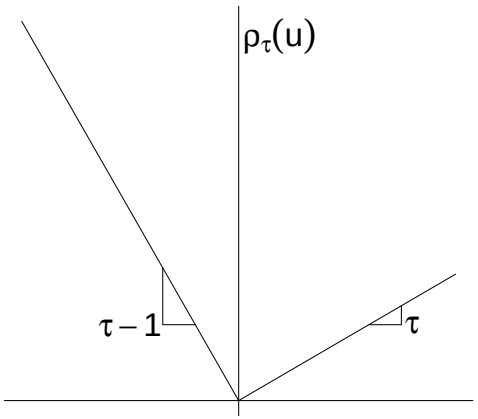
There should always be a healthy tension between good structural estimation and good descriptive data analysis. (Koopmans (1947))

This tension is, in effect, the interplay between Tukey's exploratory and confirmatory data analysis.

Likehoods versus Unlikelihoods

Classical Bayesian and Fisherian statistical theory share a common faith:

- How do we know the likelihood? What makes a likelihood likely?
- Classical robustness critique scorns the Gaussian likelihood.
- Such doubts lead to partially specified, semiparametric models.
- We minimize unlikelihood instead of maximizing likelihood.



Quantile Unlikelihood

Quantile Unlikelihoods

Quantiles are an easily interpretable way to characterize distributions.

Defined via optimization, a wide range of generalizations are possible:

$$\hat{\alpha}(\tau) = \text{argmin}_{\alpha} \sum \rho_{\tau}(y_i - \alpha) \quad \tau\text{th sample quantile}$$

$$\hat{\beta}(\tau) = \text{argmin}_{\beta} \sum \rho_{\tau}(y_i - x_i^{\top}\beta) \quad \tau\text{th regression quantile}$$

$$\hat{\gamma}(\tau) = \text{argmin}_{\gamma} \sum \rho_{\tau}(y_i - g(x_i, \gamma)) \quad \tau\text{th nonlinear regression quantile}$$

Nonparametric Quantile Regression

One can easily do univariate locally polynomial QR by kernel weighting:

$$\min_{\gamma} \sum \rho_\tau(y_i - \gamma_0 - (x_i - x)\gamma_1 - \cdots - (x_i - x)^p\gamma_p)K_h(x - x_i),$$

But this becomes difficult with higher dimensional x. An alternative is

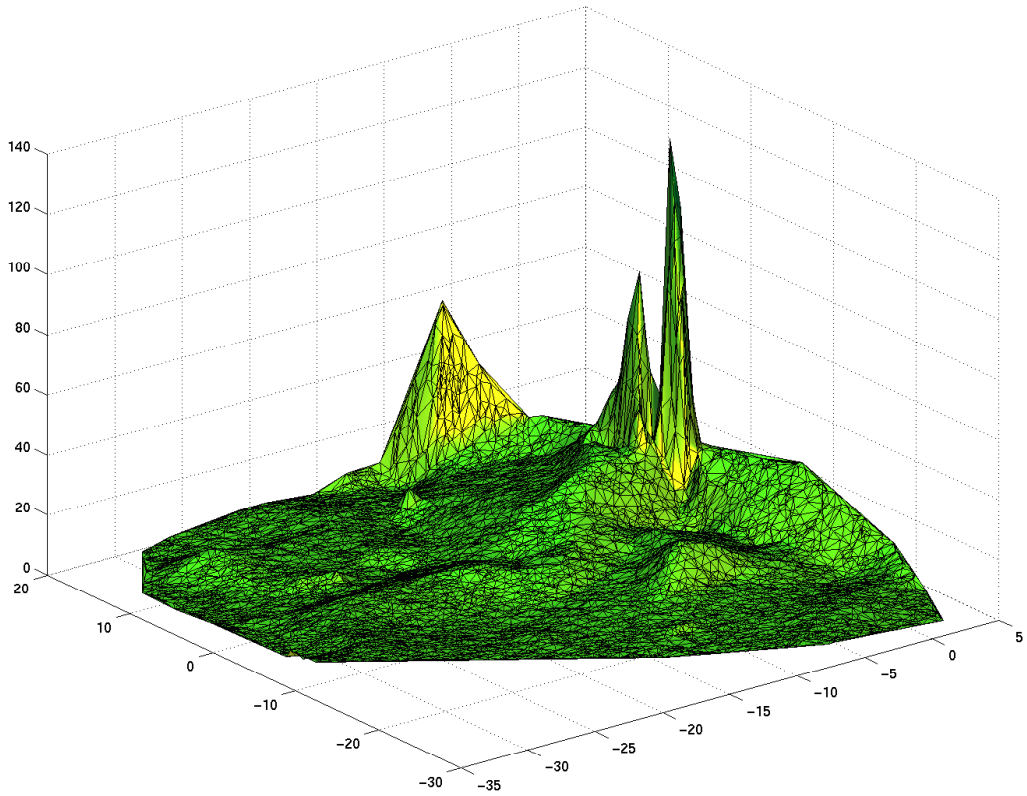
$$\min_{g} \sum \rho_\tau(y_i - g(x_i)) + \lambda \int \|\nabla^2 g\| dx.$$

Where the penalty term can be interpreted as total variation of ∇g.
Additive models of this type can also be easily estimated:

$$\min_{g} \sum_i \rho_\tau(y_i - \sum_j g_j(x_{ij})) + \sum_j \lambda_j \int \|\nabla^2 g_j\| dx.$$

In "R" such models can be estimated with `rqss` in my `quantreg` package.

Chicago Land Values via TV Regularization



Chicago Land Values: Based on 1194 vacant land sales and 7505 "virtual" sales introduced to increase the flexibility of the triangulation.

Quantile Regression for Panel Data

Time-series and panel data offer many interesting challenges, again regularization is a critical device. Consider the panel model,

$$Q_{Y_{it}}(\tau|x_{it}) = \alpha_i + x_{it}^\top \beta(\tau)$$

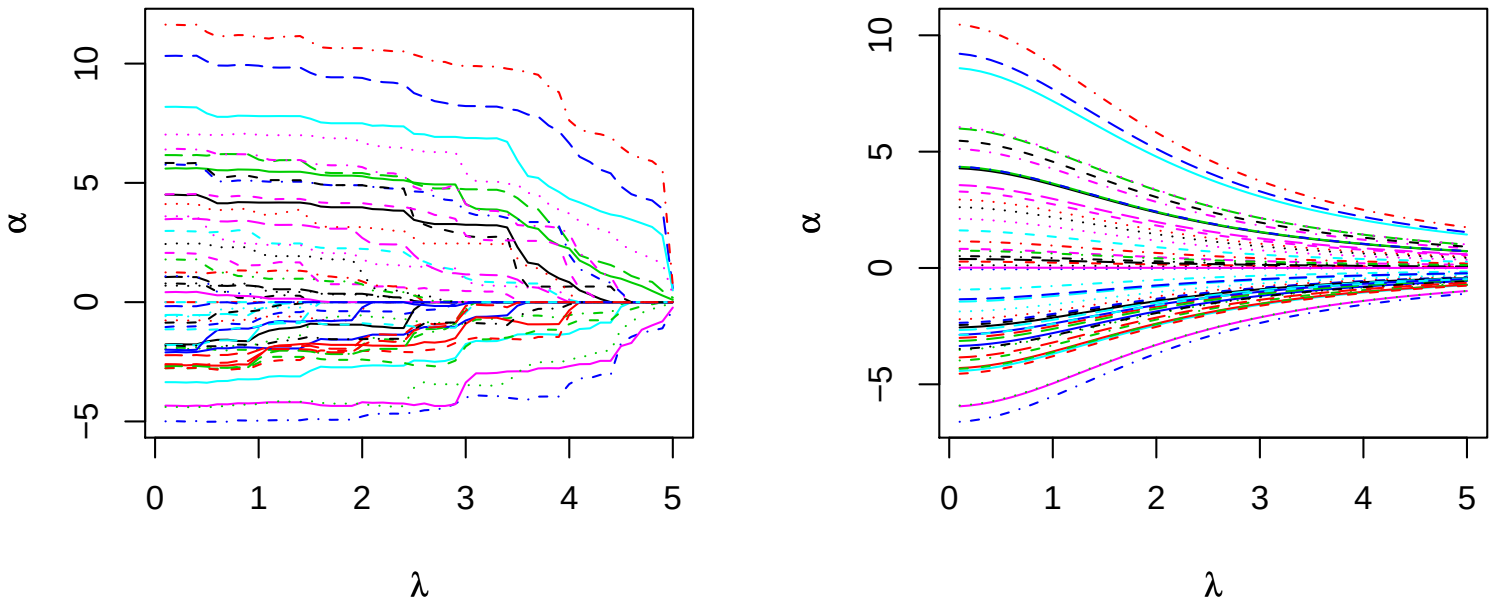
which can be estimated by solving,

$$\min \sum_{j=1}^{m} \sum_{i=1}^{n} \sum_{t=1}^{T} \rho_\tau(y_{it} - \alpha_i - x_{it}^\top \beta(\tau_j)) + \lambda \|D\alpha\|_1.$$

The regularization penalty shrinks the "fixed effects" toward a common value. Such problems are now feasible for large m, n, T thanks to recent developments in

- Barrier methods for linear programming, Frisch (1956), ...
- Sparse Linear Algebra

Shrinkage: ℓ_1 versus ℓ_2 Penalties



The ℓ_1 penalty tends to shrink most fixed effects to zero very quickly, while the ℓ_2 penalty imposes much more gradual shrinkage.

Regularization Penalties and Model Selection

Classical ℓ_2 regularization methods originating in the work of Tihkonov and Stein are too gentle; the polyhedral nature of ℓ_1 penalties are much better suited to model selection:

- Lasso/Lars of Tibshirani, Efron, ...
- Basis Pursuit of Donoho, ...
- Dantzig Selector of Candes and Tao.

ℓ_1 regularization is an emerging growth area in statistics; promising applications to imaging, signal processing and encryption. Strongly dependent on recent developments in convex optimization.

Whispering Down the Line

Problem: Transmit $x \in \mathbb{R}^n$ over a noisey channel.

Encoding: Send $y = Ax$ for $A \in \mathbb{R}^{m \times n}$, $m \gg n$, and receive either:

$$\tilde{y} = Ax + u \tag{1}$$

$$\tilde{y} = Ax + u + v \tag{2}$$

where $u \sim$ Gaussian, and v is (sparsely) arbitrarily bad.

Decoding: Set $Q = I - A(A^\top A)^{-1}A^\top$ and do either:

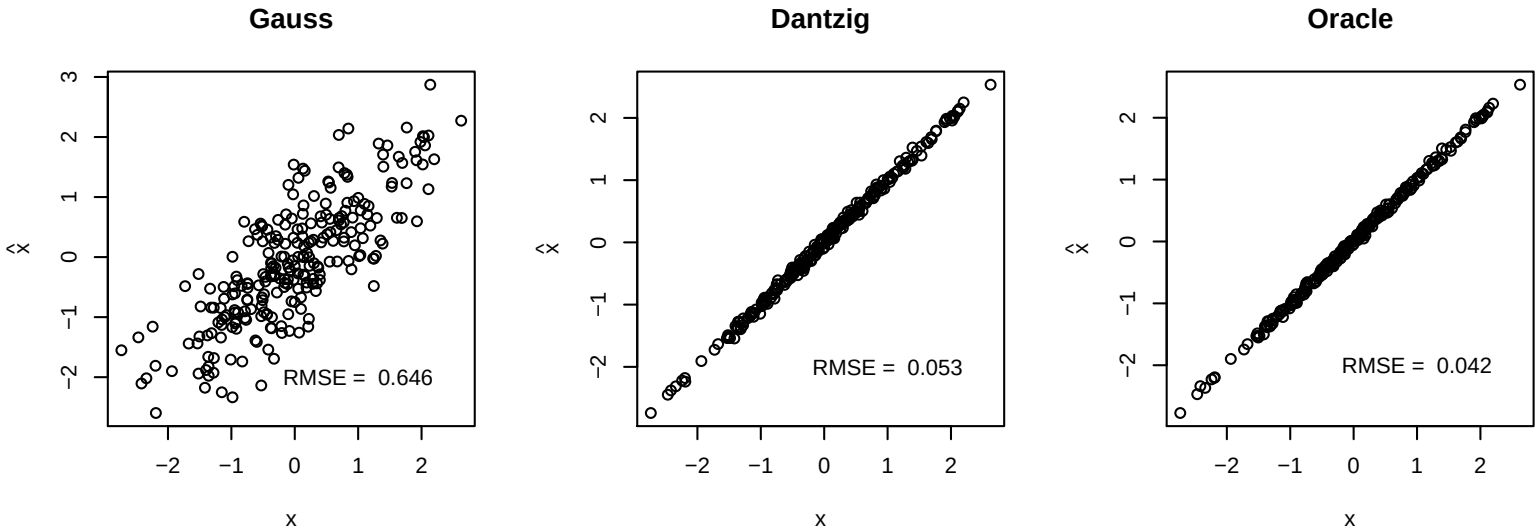
$$\hat{x} = (A^\top A)^{-1}A^\top \tilde{y} \tag{1}$$

$$\hat{x} = (A^\top A)^{-1}A^\top (\tilde{y} - \tilde{v}) \tag{2}$$

where: $\tilde{v} = \text{argmin}\{\|v\|_1 \text{ such that } \|Q(\tilde{y} - v)\|_\infty \leqslant K\}$

Dantzig Decoding: An Example

Let $n = 256$, $m = 512$, and entries in x, u and A be iid standard Gaussian, and let ν be the mixture: $\nu_i = 0.9\delta_0 + 0.1\delta_{-2y_i}$.



Dantzig decoding (2) achieves almost the same accuracy as if ν were known.

Multivariate Quantile Regression and Endogoneity

The search for a satisfactory notion of multivariate quantiles has become a quest for the statistical holy grail in recent years.

- Even the multivariate median has several competing definitions
- Quantile proposals by Chaudhuri, Koltchinskii, Liu and Singh, ...
- Wei (2007) offers a promising approach based on recursive conditioning.
- Antecedants in the debate over Wold's causal chain models, and
- Chesher's recent work on endogeneity in QR Models.

Generalized Growth Curves (Wei (2007))

Problem: Estimate a family of conditional quantile contours for height and weight of young children, conditioning on age, parental heights and weights, and possibly other covariates.

Model: Let x denote a vector of covariates and H and W denote height and weight:

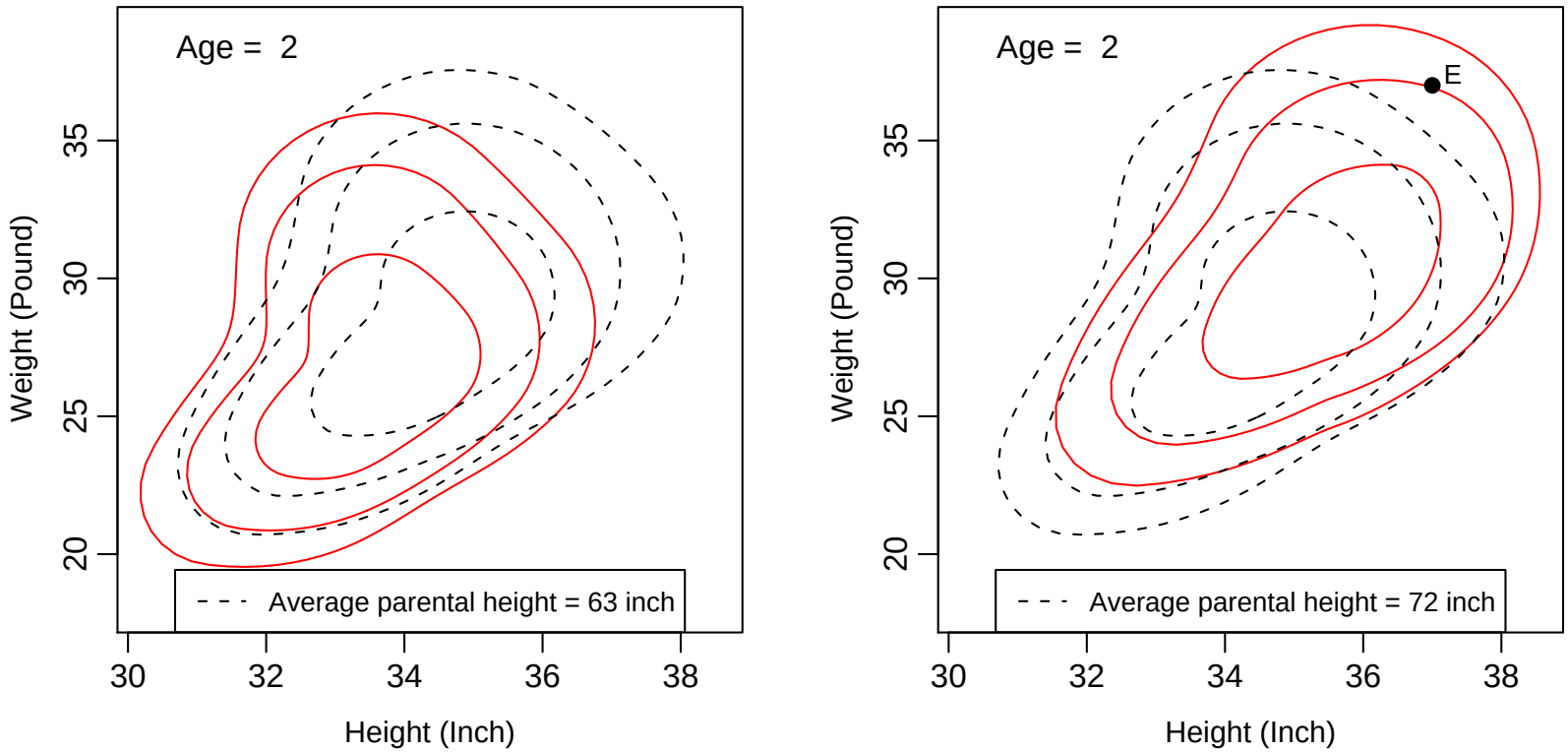
$$Q_H(\tau|x) = h(x; \gamma_H(\tau))$$

$$Q_W(\tau|H, x) = g(H, x; \gamma_W(\tau))$$

where, for convenience, g and h are linear in parameters, after expanding the covariates in B-spline basis functions.

Simulation: Fix x, and draw pairs (τ_H, τ_W) from $(U_1, U_2) \sim U[0,1]^2$ and evaluate $(Q_H(U_H|x), Q_W(U_W|Q_H(U_H|x), x))$ and now estimate density contours, Rosenblatt (1952).

Height-Weight Contours for Two Two-Year Olds



Dashed lines are reference quantile contours for $\tau \in \{0.5, 0.8, 0.9\}$ for two year olds. Red contours conditional on parents height. Ying Wei (2007).

Conclusions

The Average Man is alive, but unwell, drinking in Paris.

Fortunately, we aren't all converging to his unhappy state.

Heterogeneity is essential to most economic phenomena, but we need better econometric methods to explore it.

Empirical methods based on global likelihoods are often implausible and non-robust.

More flexible local (unlikelihood) methods are preferable for exploratory data analysis.

Some progress has been made, but there are many challenging problems that lie ahead.