Quantile Regression in R: For Fin and Fun

Roger Koenker

University of Illinois at Urbana-Champaign

R in Finance: 25 April 2009

What is Quantile Regression?

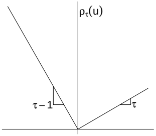
- Quantiles Describe Marginal Distributions
  - Proportion τ of portfolio managers perform better than the τth quantile. (Except in Lake Woebegone.)
- Regression Quantiles Describe Conditional Distributions
  - Given characteristics X, proportion τ of portfolio managers of type X perform better than τth conditional quantile.
- Quantiles minimize asymmetric linear loss
  - Sorting can be replaced by optimization.
- Regression Quantiles also minimize asymmetric linear loss
  - Optimization generalizes nicely to the regression setting.

Sample Quantiles via Optimization

The τth sample quantile can be defined as any solution to:

$$\hat{\alpha}(\tau) = \text{argmin}_{a \in \Re} \sum_{i=1}^{n} \rho_\tau(y_i - a)$$

where $\rho_\tau(u) = (\tau - I(u < 0))u$ as illustrated below.



Biases the argmin toward making the lower cost error; like forecasting flood crest levels.

The Least Squares Meta-Model

The unconditional mean solves

$$\mu = \min_m E(Y - m)^2$$

The Least Squares Meta-Model

The unconditional mean solves

$$\mu = \min_m \mathrm{E}(Y - m)^2$$

The conditional mean $\mu(x) = \mathrm{E}(Y|X = x)$ solves

$$\mu(x) = \min_m \mathrm{E}_{Y|X=x}(Y - m(x))^2.$$

The Least Squares Meta-Model

The unconditional mean solves

$$\mu = \min_m E(Y - m)^2$$

The conditional mean $\mu(x) = E(Y|X = x)$ solves

$$\mu(x) = \min_m E_{Y|X=x}(Y - m(x))^2.$$

Similarly, the unconditional τth quantile solves

$$\alpha_\tau = \min_a E\rho_\tau(Y - a)$$

The Least Squares Meta-Model

The unconditional mean solves

$$\mu = \min_m E(Y - m)^2$$

The conditional mean $\mu(x) = E(Y|X = x)$ solves

$$\mu(x) = \min_m E_{Y|X=x}(Y - m(x))^2.$$

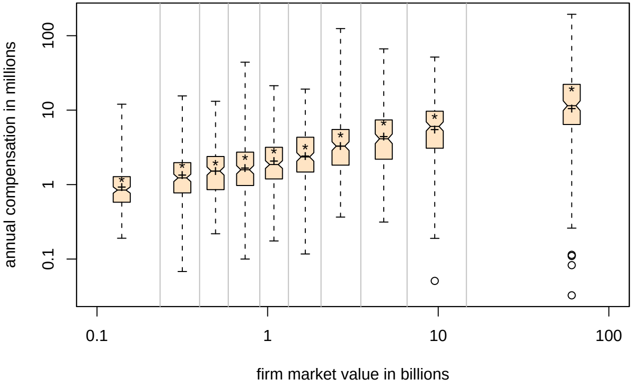
Similarly, the unconditional τth quantile solves

$$\alpha_\tau = \min_a E\rho_\tau(Y - a)$$

and the conditional τth quantile solves

$$\alpha_\tau(x) = \min_q E_{Y|X=x}\rho_\tau(Y - q(x))$$

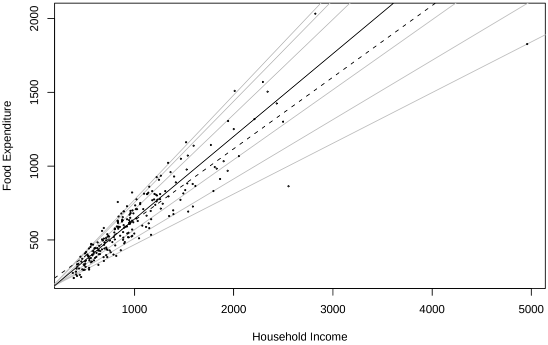
Boxplot of CEO Pay by Firm Size



Three Applications

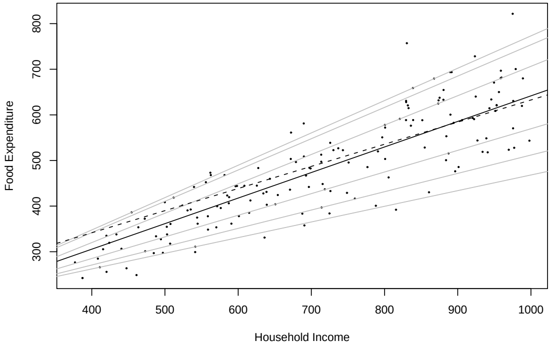
- Engel's Law: A Classical Economic Example
- Infant Birthweight: A Public Health Example
- Melbourne Daily Temperature: A Time Series Example

Engel's Food Expenditure Data



Engel Curves for Food: This figure plots data taken from Engel's (1857) study of the dependence of households' food expenditure on household income. Seven estimated quantile regression lines for $\tau \in \{.05, .1, .25, .5, .75, .9, .95\}$ are superimposed on the scatterplot. The median $\tau = .5$ fit is indicated by the darker solid line; the least squares estimate of the conditional mean function is indicated by the dashed line.

Engel's Food Expenditure Data

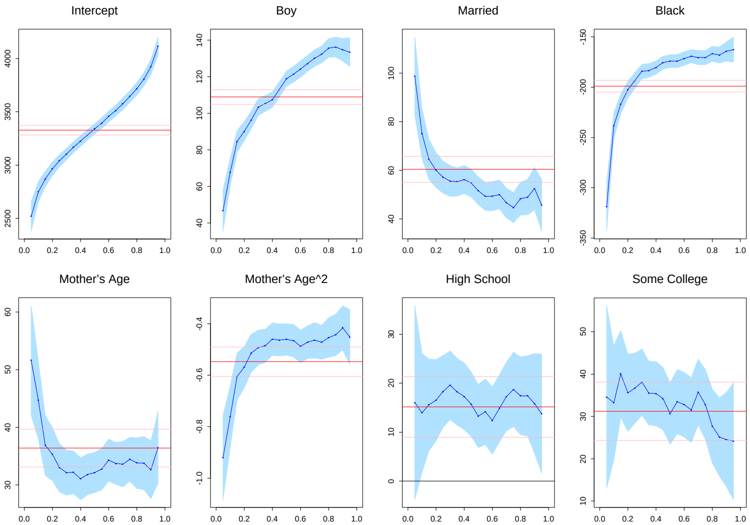


Engel Curves for Food: This figure plots data taken from Engel's (1857) study of the dependence of households' food expenditure on household income. Seven estimated quantile regression lines for $\tau \in \{.05, .1, .25, .5, .75, .9, .95\}$ are superimposed on the scatterplot. The median $\tau = .5$ fit is indicated by the darker solid line; the least squares estimate of the conditional mean function is indicated by the dashed line.

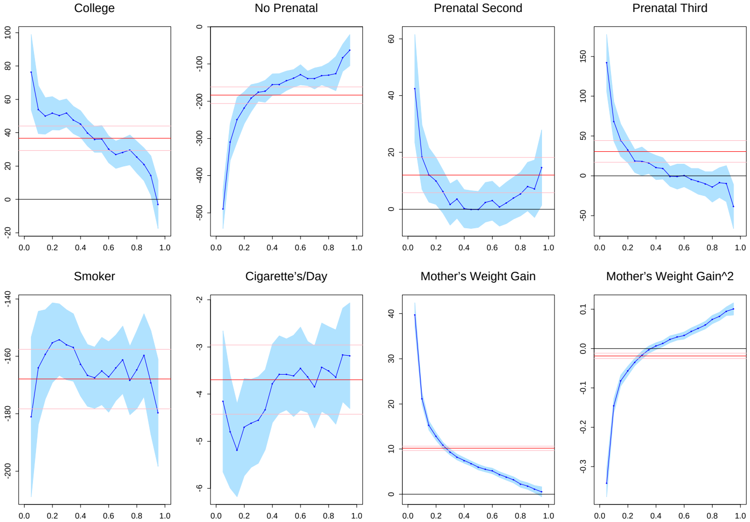
A Model of Infant Birthweight

- Reference: Abrevaya (2001), Koenker and Hallock (2001)
- Data: June, 1997, Detailed Natality Data of the US. Live, singleton births, with mothers recorded as either black or white, between 18-45, and residing in the U.S. Sample size: 198,377.
- Response: Infant Birthweight (in grams)
- Covariates:
  - Mother's Education
  - Mother's Prenatal Care
  - Mother's Smoking
  - Mother's Age
  - Mother's Weight Gain

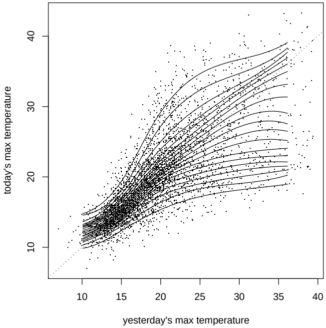
Quantile Regression Birthweight Model I



Quantile Regression Birthweight Model II

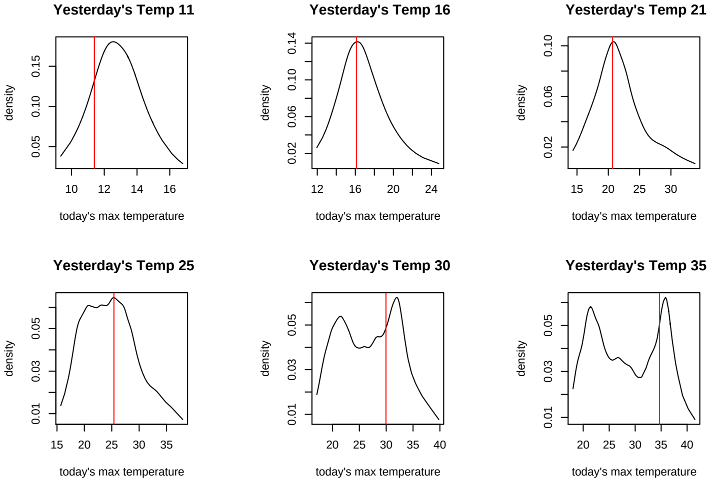


AR(1) Model of Melbourne Daily Temperature



The plot illustrates 10 years of daily maximum temperature data for Melbourne, Australia as an AR(1) scatterplot. Superimposed are estimated conditional quantile functions for $\tau \in \{.05, .10, ..., .95\}$. parameterized via B-splines.

Conditional Densities of Melbourne Daily Temperature



Quantile Autoregression and Irrational Exuberance

- Simple linear QAR models

$$Q_{Y_t|Y_{t-1}}(\tau|y_{t-1}) = \alpha(\tau) + \beta(\tau)y_{t-1}$$

  can exhibit strong unit-root or even explosive episodic tendencies, but still be stationary, and mean reverting, provided that $\beta(\tau)$ is square integrable.

- Copulas offer a rich source of convenient *nonlinear* specifications of QAR models.
- Similar methods yield more flexible GARCH type models.

Pessimistic Risk and Portfolio Selection

- Classical measures of risk like standard devation, variance, and value at risk have some serious logical difficulties,

Pessimistic Risk and Portfolio Selection

- Classical measures of risk like standard devation, variance, and value at risk have some serious logical difficulties,
- Axiomatics of Artzner et al (1999) suggest a class of pessimistic risk measures in which conditional expectation is replaced by tail expectation:

$$R_{\nu_\alpha}(Y) = -\alpha^{-1} \int_0^\alpha F_Y^{-1}(t)dt$$

when $\alpha = 1$ this is just the usual expectation, but for $\alpha < 1$ it is the conditional expectation given that returns are *below* the αth quantile.

Pessimistic Risk and Portfolio Selection

- Classical measures of risk like standard devation, variance, and value at risk have some serious logical difficulties,
- Axiomatics of Artzner et al (1999) suggest a class of pessimistic risk measures in which conditional expectation is replaced by tail expectation:

$$R_{\nu_\alpha}(Y) = -\alpha^{-1} \int_0^\alpha F_Y^{-1}(t)dt$$

when $\alpha = 1$ this is just the usual expectation, but for $\alpha < 1$ it is the conditional expectation given that returns are *below* the αth quantile.

- The class can be expanded to include anything of the form:

$$R_\nu(Y) = -\int_0^1 F_Y^{-1}(t)d\nu(t)$$

where $\nu : [0, 1] \rightarrow [0, 1]$ is a concave function.

Pessimistic Risk and Portfolio Selection

- Classical measures of risk like standard devation, variance, and value at risk have some serious logical difficulties,
- Axiomatics of Artzner et al (1999) suggest a class of pessimistic risk measures in which conditional expectation is replaced by tail expectation:

$$R_{\nu_\alpha}(Y) = -\alpha^{-1} \int_0^\alpha F_Y^{-1}(t)dt$$

  when $\alpha = 1$ this is just the usual expectation, but for $\alpha < 1$ it is the conditional expectation given that returns are *below* the αth quantile.
- The class can be expanded to include anything of the form:

$$R_\nu(Y) = -\int_0^1 F_Y^{-1}(t)d\nu(t)$$

  where $\nu : [0, 1] \to [0, 1]$ is a concave function.
- These pessimistic expectations inflate the probability of unfavorable events and deflate the probabilities of favorable events.

Choquet Expectations and Quantile Regression

- There is a close link between the simplest pessimistic risk measure and the quantile regression problem:

$$R_{\nu_\alpha}(Y) = -\min_{\xi} E\rho_\alpha(Y - \xi) - \alpha\mu(Y)$$

so up to a shift by a multiple of the expected return, they are the same.

Choquet Expectations and Quantile Regression

- There is a close link between the simplest pessimistic risk measure and the quantile regression problem:

$$R_{\nu_\alpha}(Y) = -\min_{\xi} E\rho_\alpha(Y - \xi) - \alpha\mu(Y)$$

  so up to a shift by a multiple of the expected return, they are the same.
- The general class of pessimistic risk measures can be approximated by approximating a general concave ν by a piecewise linear version.

Choquet Expectations and Quantile Regression

- There is a close link between the simplest pessimistic risk measure and the quantile regression problem:

$$R_{\nu_\alpha}(Y) = -\min_{\xi} E\rho_\alpha(Y - \xi) - \alpha\mu(Y)$$

  so up to a shift by a multiple of the expected return, they are the same.
- The general class of pessimistic risk measures can be approximated by approximating a general concave ν by a piecewise linear version.
- When $\nu(t) = t$ we revert to expected return.

Pessimistic Portfolios I

Now let $X = (X_1, \ldots, X_p)$ denote a vector of potential portfolio asset returns and $Y = X^\top \pi$, the returns on the portfolio with weights π. Consider

$$\min_{\pi} R_{\nu_\alpha}(Y) - \lambda \mu(Y)$$

Minimize α-risk subject to a constraint on mean return.

Pessimistic Portfolios I

Now let $X = (X_1, \ldots, X_p)$ denote a vector of potential portfolio asset returns and $Y = X^\top \pi$, the returns on the portfolio with weights π. Consider

$$\min_\pi R_{\nu_\alpha}(Y) - \lambda \mu(Y)$$

Minimize α-risk subject to a constraint on mean return.
This problem can be formulated as a linear quantile regression problem

$$\min_{(\beta,\xi) \in \mathcal{R}^p} \sum_{i=1}^n \rho_\alpha (x_{i1} - \sum_{j=2}^p (x_{i1} - x_{ij}) \beta_j - \xi) \quad \text{s.t.} \quad \bar{x}^\top \pi(\beta) = \mu_0,$$

where $\pi(\beta) = (1 - \sum_{j=2}^p \beta_j, \beta^\top)^\top$.

Pessimistic Portfolios II

Any pessimistic risk measure may be approximated by

$$R_\nu(Y) = \sum_{k=1}^{m} \varphi_k R_{\nu_{\alpha_k}}(Y)$$

where $\varphi_k > 0$ for $k = 1, 2, ..., m$ and $\sum \varphi_k = 1$.

Pessimistic Portfolios II

Any pessimistic risk measure may be approximated by

$$R_\nu(Y) = \sum_{k=1}^{m} \varphi_k R_{\nu_{\alpha_k}}(Y)$$

where $\varphi_k > 0$ for $k = 1, 2, ..., m$ and $\sum \varphi_k = 1$.
Portfolio weights can be estimated for these risk measures by solving linear programs that are weighted sums of quantile regression problems:

$$\min_{(\beta,\xi)\in\mathcal{R}^p} \sum_{k=1}^{m} \sum_{i=1}^{n} \nu_k \rho_{\alpha_k}(x_{i1} - \sum_{j=2}^{p} (x_{i1} - x_{ij})\beta_j - \xi_k) \quad \text{s.t.} \quad \bar{x}^\top \pi(\beta) = \mu_0,$$

Pessimistic Portfolios II

Any pessimistic risk measure may be approximated by

$$R_\nu(Y) = \sum_{k=1}^{m} \varphi_k R_{\nu_{\alpha_k}}(Y)$$

where $\varphi_k > 0$ for $k = 1, 2, ..., m$ and $\sum \varphi_k = 1$.
Portfolio weights can be estimated for these risk measures by solving linear programs that are weighted sums of quantile regression problems:

$$\min_{(\beta,\xi)\in\mathcal{R}^p} \sum_{k=1}^{m} \sum_{i=1}^{n} \nu_k \rho_{\alpha_k}(x_{i1} - \sum_{j=2}^{p} (x_{i1} - x_{ij})\beta_j - \xi_k) \quad \text{s.t.} \quad \bar{x}^\top \pi(\beta) = \mu_0,$$

Software in R is available from my webpages.

Quantile Regression Bracketology: Or How to Bet on College Basketball, (If You Must)

In the classical paired comparison model, let Y_{ijg} denote the score of team i playing team j in game g and suppose:

$$EY_{ijg} = \alpha_i - \delta_j + \gamma D_g$$

where $D_g = I(\text{game } g \text{ is played on team } i\text{'s home court})$, so γ denotes the home court advantage. This model is estimable provided that there is sufficient overlap in scheduling of the observed games.

Quantile Regression Bracketology: Or How to Bet on College Basketball, (If You Must)

In the classical paired comparison model, let Y_{ijg} denote the score of team i playing team j in game g and suppose:

$$EY_{ijg} = \alpha_i - \delta_j + \gamma D_g$$

where $D_g = I(\text{game } g \text{ is played on team } i\text{'s home court})$, so γ denotes the home court advantage. This model is estimable provided that there is sufficient overlap in scheduling of the observed games.

Critique of Least Squares Estimation of the Paired Comparison Model

- Presumes Gaussian "errors," so extreme scores (blowouts) can exert "too much" influence on ratings,
- But binary response versions sacrifice too much information
- Ignores possible dependence in scores between and within games.

The Quantilesque Paired Comparison Model

Suppose instead of postulating a model for mean scores:

$$Q_{Y_{ijg}}(\tau) = \alpha_i(\tau) - \delta_j(\tau) + \gamma(\tau) D_g$$

- Median version $(\tau = 1/2)$ is quite similar to mean model,
- Except that it is less sensitive to extreme scores,
- For general τ we permit much richer class of rankings
- Some teams can be very consistent others very erratic
- Teams can have different locations, scales and shapes for their offensive and defensive ratings functions.

Prediction in the QPCM

Suppose teams i and j meet at a neutral site, the result is modeled by the quantile functions for the two scores:

$$(Q_{Y_{ig}}(\tau), Q_{Y_{jg}}(\tau)) = (\alpha_i(\tau) - \delta_j(\tau), \alpha_j(\tau) - \delta_i(\tau))$$

We can simulate the probability of team i winning by Δ.

$$\pi_{ij} = P(Q_{Y_{ig}}(U) > Q_{Y_{jg}}(V) + \Delta).$$

where U and V are independent (??) uniforms, provided we know the α's and δ's. This is quite like the Melbourne temperature model.

Estimation of the QPCM

Estimation is just a (very sparse) quantile regression problem:

$$\min_{(\alpha,\delta,\gamma)} \sum_{g} \rho_\tau(y_{ig} - \alpha_i + \delta_j - \gamma D_{ig}) + \rho_\tau(y_{jg} - \alpha_j + \delta_i - \gamma D_{jg})$$

Estimation of the QPCM

Estimation is just a (very sparse) quantile regression problem:

$$\min_{(\alpha,\delta,\gamma)} \sum_{g} \rho_\tau(y_{ig} - \alpha_i + \delta_j - \gamma D_{ig}) + \rho_\tau(y_{jg} - \alpha_j + \delta_i - \gamma D_{jg})$$

or,

$$\min_{\theta} \|y - X\theta\|_\tau,$$

where $\|u\|_\tau \equiv \sum \rho_\tau(u_i)$, $y = (y_i, y_j)$ denotes a stacked vector of scores, $\theta = (\alpha, \delta, \gamma)$ and X is an extremely sparse matrix; no row of X has more than 3 non-zero entries. Sparse linear algebra and interior point linear programming methods make estimation very efficient in R.

Estimation of the QPCM

Estimation is just a (very sparse) quantile regression problem:

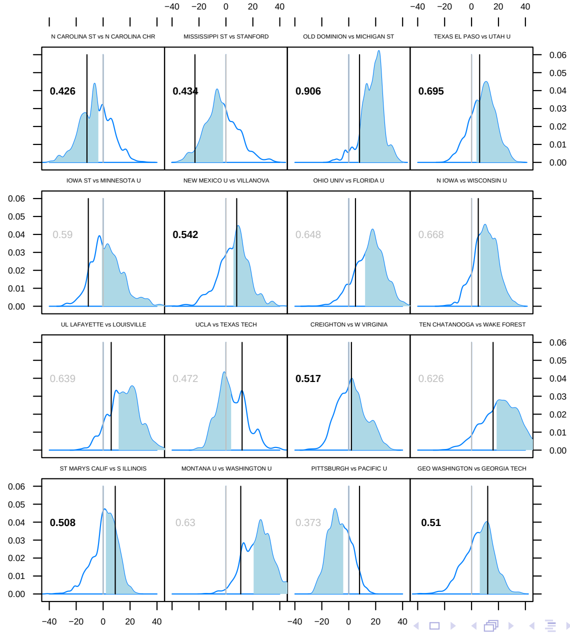
$$\min_{(\alpha,\delta,\gamma)} \sum_g \rho_\tau(y_{ig} - \alpha_i + \delta_j - \gamma D_{ig}) + \rho_\tau(y_{jg} - \alpha_j + \delta_i - \gamma D_{jg})$$

or,

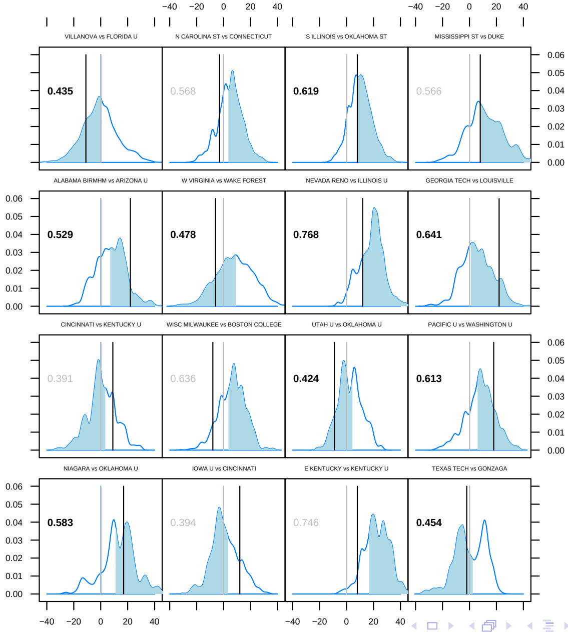
$$\min_\theta \|y - X\theta\|_\tau,$$

where $\|u\|_\tau \equiv \sum \rho_\tau(u_i)$, $y = (y_i, y_j)$ denotes a stacked vector of scores, $\theta = (\alpha, \delta, \gamma)$ and X is an extremely sparse matrix; no row of X has more than 3 non-zero entries. Sparse linear algebra and interior point linear programming methods make estimation very efficient in R.
The model was estimated on a sample of 2940 games involving 232 Division I NCAA college basketball teams for the 2004-05 regular season. The estimated model was then used to predict the outcomes of the 2005 NCAA basketball tournament.

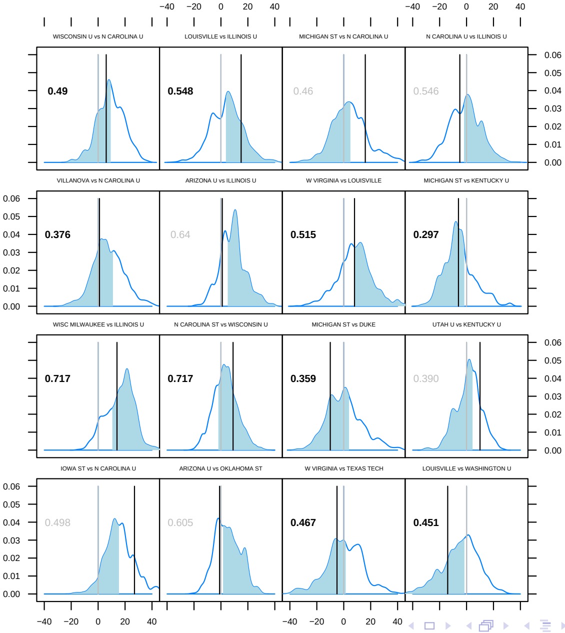
Predictive Densities for 2005 Tournament



Predictive Densities for 2005 Tournament



Predictive Densities for 2005 Tournament



Betting on the Pointspread

How well would we have done betting on the Las Vegas pointspreads in the 48 tournament games we have illustrated?

- Bet on the team with best probability of beating the pointspread,
- In 27 out of 47 games we have bet correctly,
- One game was a "push" so the bet is refunded.
- It costs $110 to place a $100 bet, so
- We have an mean gain of $10.50 on a $100 bet.

Should We Quit Our Day Jobs?

Probably not:

- 48 games is a rather small sample, but

Should We Quit Our Day Jobs?

Probably not:

- 48 games is a rather small sample, but
- More fun than picking up nickels in front of a steamroller,

Should We Quit Our Day Jobs?

Probably not:

- 48 games is a rather small sample, but
- More fun than picking up nickels in front of a steamroller,
- There are many possible refinements:
  - Shrinkage to control variability of the profligate model specification,
  - Weighting to accentuate the import of more recent games,
  - Introduction of prior season performance
  - Introduction of other covariates
- But evidence for the Hayek hypothesis that aggregation of market bets yields accurate probability assessment, is rather weak.

Some References

- Koenker, R. (2005) *Quantile Regression*, Cambridge U. Press.
- Koenker, R. and Z. Xiao (2006) Quantile Autoregression, *JASA*.
- Chen, X., R. Koenker, and Z. Xiao (2009) Copula-Based Nonlinear Quantile Autoregression, *Econometrics Journal*.
- Koenker, R. and Z. Xiao (2009) Conditional Quantile Methods for GARCH Models, preprint.
- Bassett, G., R. Koenker and G. Kordas, (2005) Pessimistic Portfolio Allocation and Choquet Expected Utility, *J. of Fin. Econometrics*
- Bassett, G., and R. Koenker (2009) March Madness, Quantile Regression Bracketology and the Hajek Hypothesis, *J. of Bus. & Econ. Statistics*.

Slides will be available from my webpage.