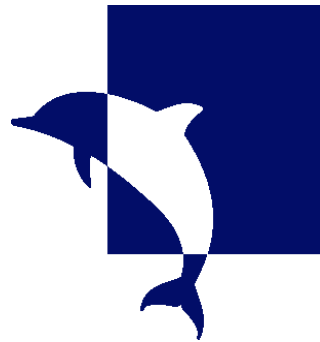


Multi-quantile Models for Small Area Estimation

Ray Chambers and Nikos Tzavidis
Southampton Statistical Sciences Research Institute
University of Southampton

Measuring neighbourhood effects, ESRC Research Methods Programme, RSS, 21 November 2003



Context: Estimation of totals, means and other parameters for small areas of interest e.g. neighbourhoods

Problem: Small sample sizes in small areas means that direct estimates based on the sample data alone can be very unstable

A Solution: Employ model-based methods to “borrow strength” from related areas

The Industry Standard: Multi-level Models that include Area (Neighbourhood) Effects

Concept: Include random area-specific effects to account for the between area variation beyond that explained by the variation in model covariates.

Notation: (i = area, j = individual)

Variable of interest: y_{ij}

Covariate information: \mathbf{x}_{ij}

Area level random effect: u_i

Random error: e_{ij}

Linear Random Intercepts Model

$$y_{ij} = \beta_0 + \mathbf{x}_{ij}^T \boldsymbol{\beta}_1 + u_i + e_{ij}$$

Assumptions: $u_i \sim NID(0, \sigma^2)$ $e_{ij} \sim NID(0, \omega^2)$ $u_i \perp e_{ij}$

Inference based on Random Intercepts model

- Area means of y estimated via

$$\hat{y}_i = \hat{\beta}_0 + \bar{\mathbf{x}}_i^T \hat{\boldsymbol{\beta}}_1 + \hat{u}_i$$

Linear Random Slopes Model

$$y_{ij} = \beta_0 + \mathbf{x}_{ij}^T \boldsymbol{\beta}_1 + u_i + \mathbf{z}_{ij}^T \boldsymbol{\gamma}_i + e_{ij}$$

Assumptions: $(u_i, \boldsymbol{\gamma}_i)^T \sim NID(\mathbf{0}, \boldsymbol{\Sigma})$ $e_{ij} \sim NID(0, \omega^2)$ $(u_i, \boldsymbol{\gamma}_i) \perp e_{ij}$

Inference based on Random Slopes model

- Area means of y estimated via

$$\hat{y}_i = \hat{\beta}_0 + \bar{\mathbf{x}}_i^T \hat{\boldsymbol{\beta}}_1 + \hat{u}_i + \bar{\mathbf{z}}_i^T \hat{\boldsymbol{\gamma}}_i$$

An Alternative Approach to Regression Modelling: Regression Models for Percentiles

Basic Idea (Hogg, 1974; Koenker and Bassett, 1978)

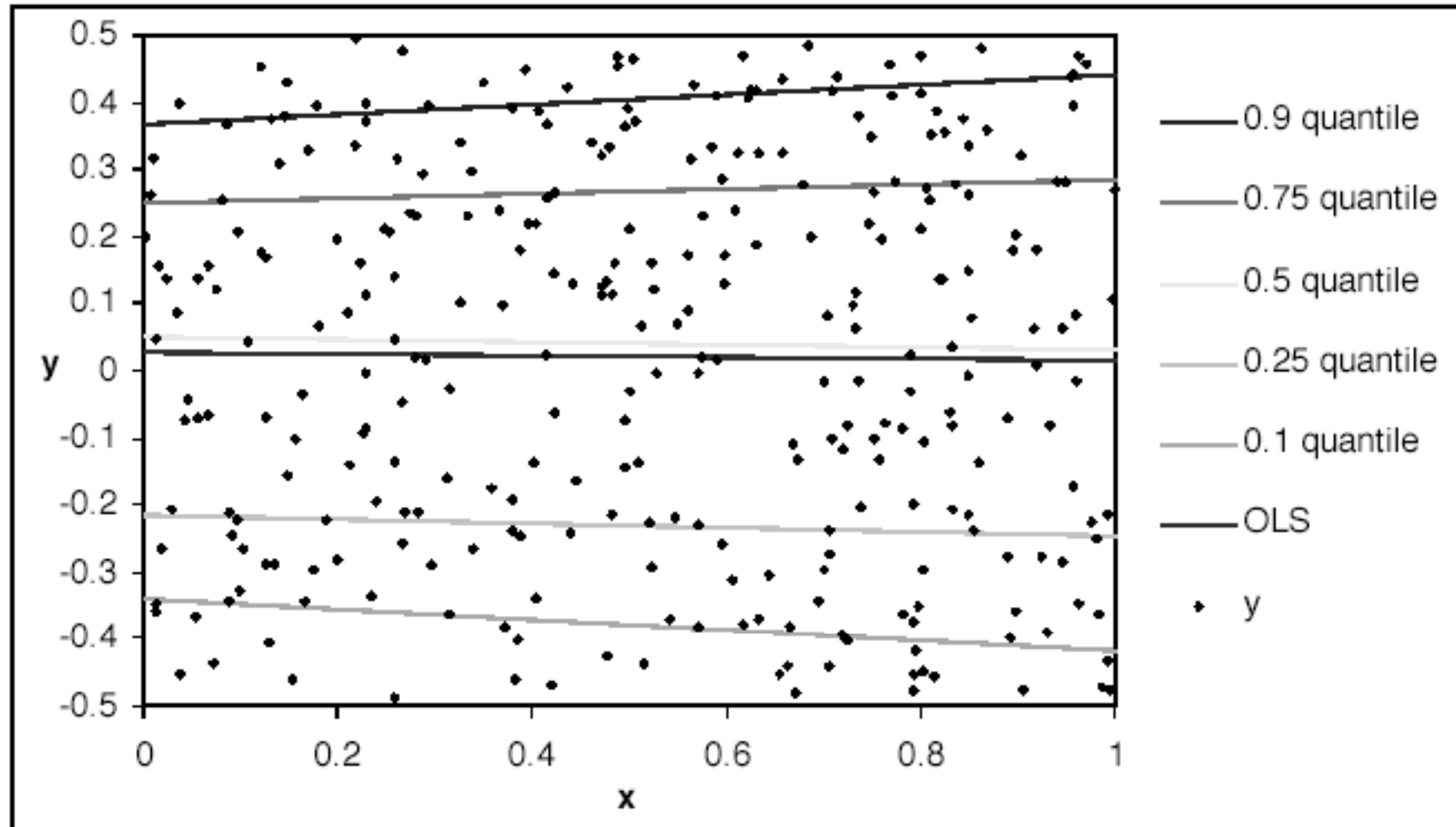
Rather than modelling the **expected value** of the conditional distribution of y given \mathbf{x} , model the **percentiles** of the conditional distribution of y given \mathbf{x}

In **linear case** leads to a family (or “ensemble”) of linear models indexed by the value of the corresponding percentile “coefficient”, $q \in (0,1)$

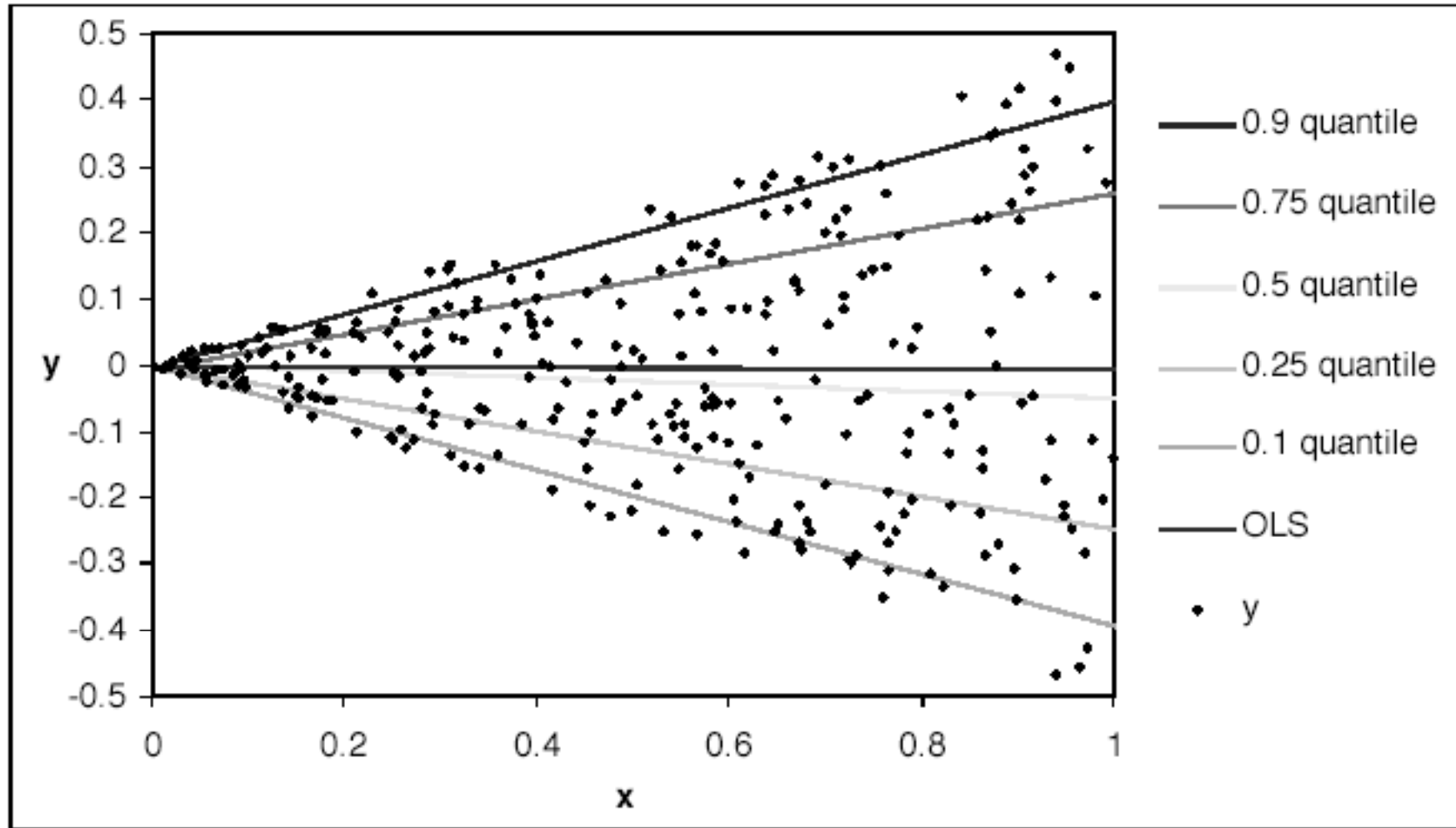
For each value of q , the corresponding model shows how the q^{th} percentile (quantile) of $f(y|\mathbf{x})$ varies with \mathbf{x}

- **$q = 0.5$ line** shows how the “middle” (median) of $f(y|\mathbf{x})$ changes with \mathbf{x}
- **$q = 0.1$ line** separates the “top” 90% of $f(y|\mathbf{x})$ from the “bottom” 10% - i.e. it represents the behaviour of units that are “better” than the “worst” 10% and “worse” than the “best” 90% ...
- $y = \beta_0 + \beta_1 x + e$, $e \sim NID(0,1) \Rightarrow \beta_0(q) = \beta_0 + \Phi^{-1}(q)$ and $\beta_1(q) = \beta_1$ (i.e. **symmetric** iid errors \Rightarrow parallel percentile regression lines)
- **Heteroskedasticity** and/or **asymmetry** in residuals (i.e. individual effects) will cause percentile regression lines to “spread out”

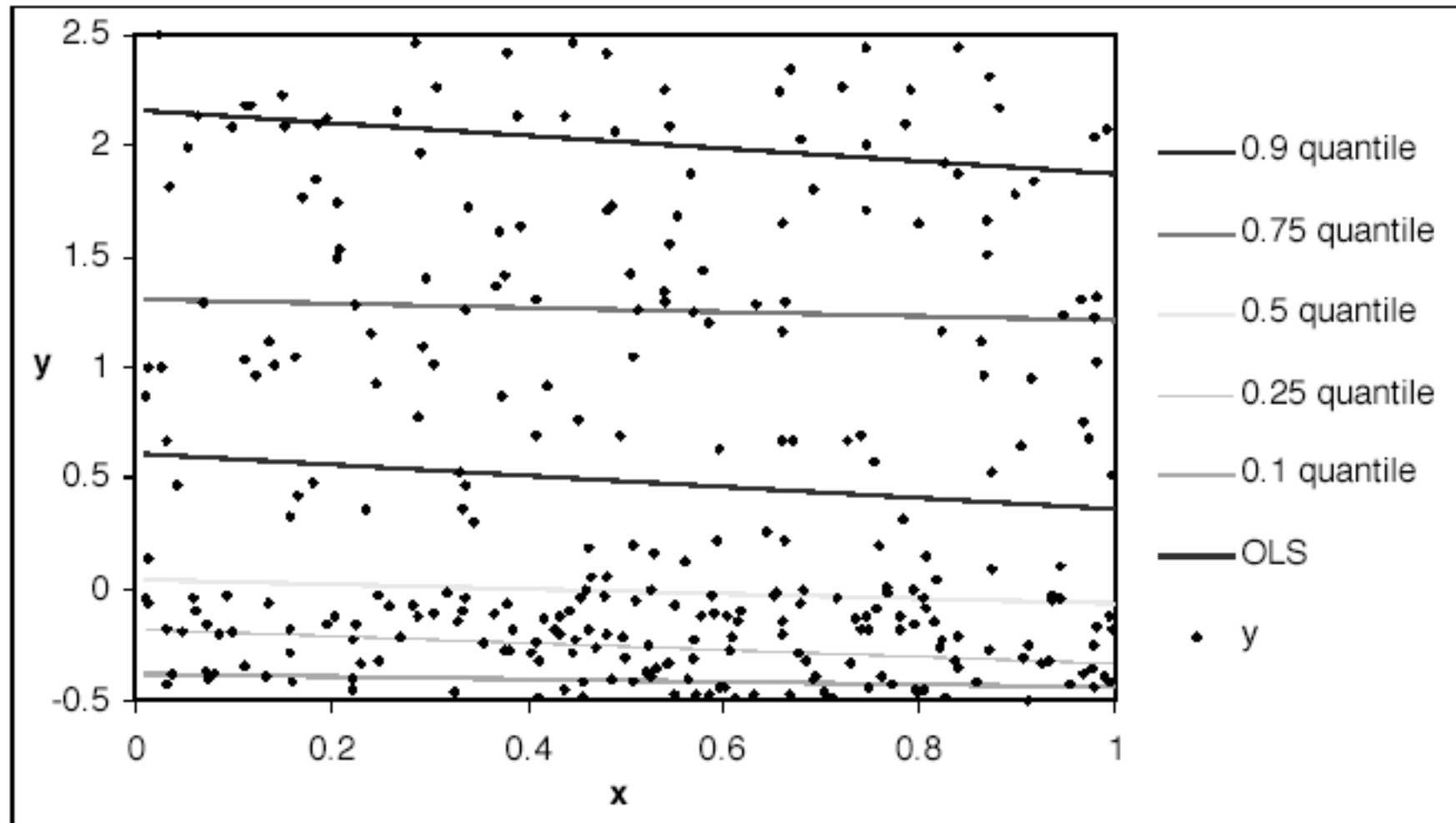
Simulated results (Melly, 2001) – homoskedastic case



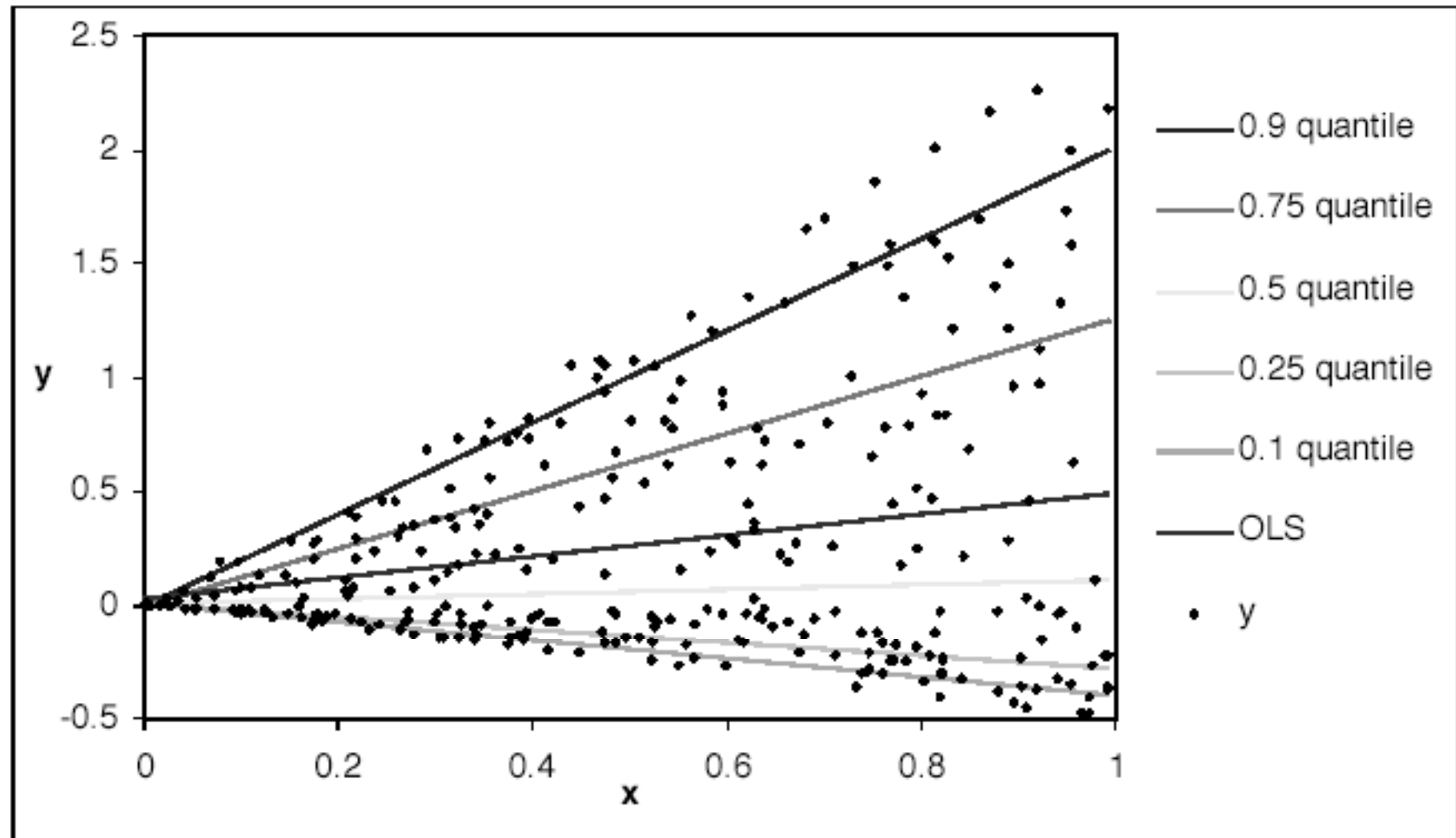
If there is heteroskedasticity



If there is homoskedasticity and asymmetry



If there is heteroskedasticity and asymmetry



Generalisation

Replace percentiles of $f(y|\mathbf{x})$ by **M-quantiles** of $f(y|\mathbf{x})$

- Breckling and Chambers (1988)
- includes the percentile as a special case, but also includes the **expectile** (percentile generalisation of expectation) and more generally any set of M-quantile “**percentile like**” values characterising $f(y|\mathbf{x})$
- fitting is very straightforward using iterated WLS, with positive residuals weighted by q and negative residuals weighted by $1 - q$

Estimating Equations for Linear M-quantile Models

The coefficients $\beta_0(q), \beta_1(q)$ of the linear model for the q^{th} M-quantile of the conditional distribution of y given \mathbf{x} are obtained by using IWLS to solve the “normal” equations

$$\sum_{k=1}^n \phi[r_k(q); q] \begin{pmatrix} 1 \\ \mathbf{x}_k \end{pmatrix} = \mathbf{0}$$

where $r_k(q) = y_k - \beta_0(q) - \mathbf{x}_k^T \beta_1(q)$ and

$$\phi[r_k(q); q] = \begin{cases} 2q\psi[s^{-1}r_k(q)] & \text{if } r_k(q) > 0 \\ 2(1-q)\psi[s^{-1}r_k(q)] & \text{if } r_k(q) < 0 \end{cases}$$

s = robust estimate of scale, ψ = M-quantile influence function

Choice of Influence Function

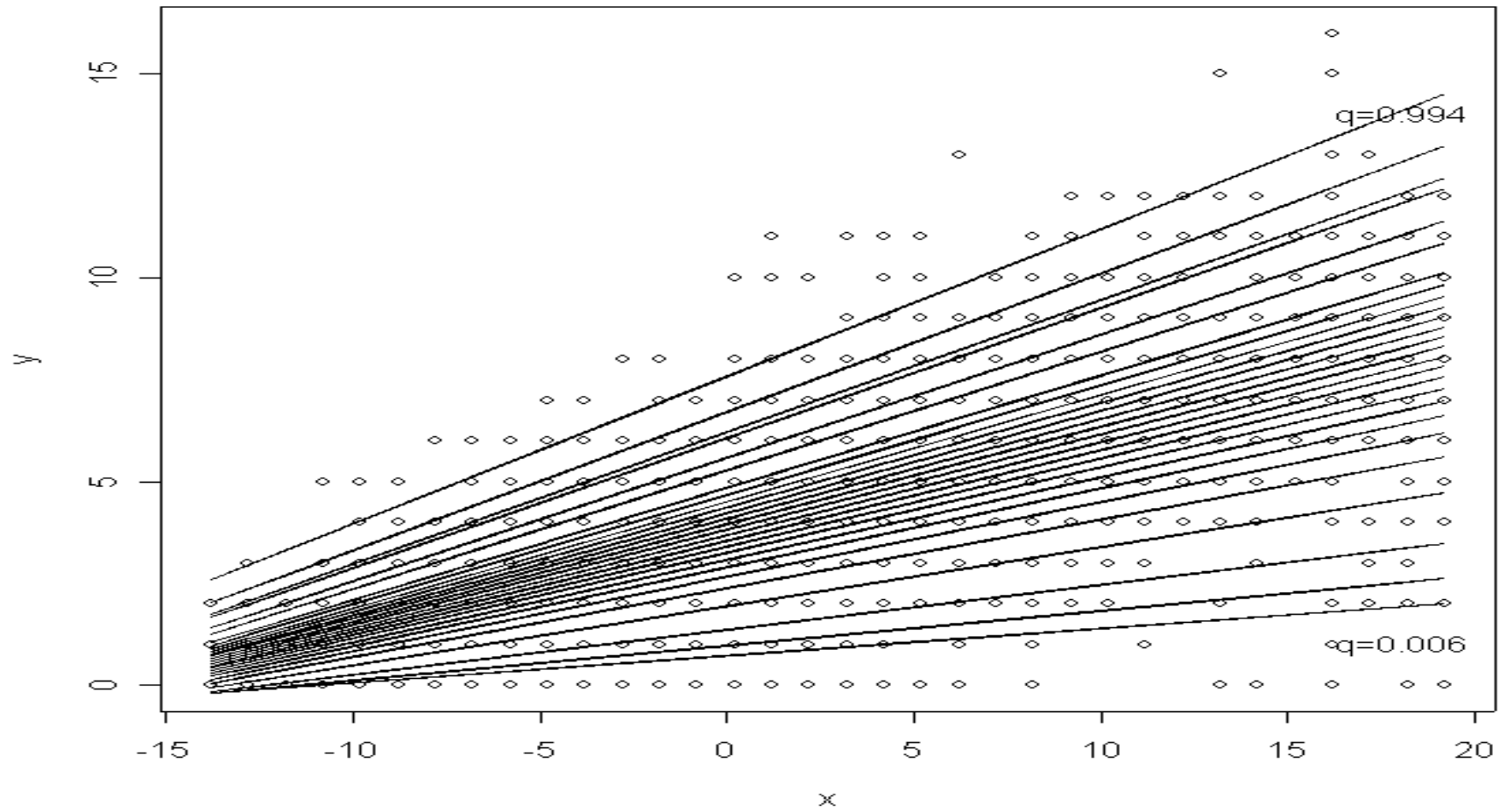
Throughout this presentation we will assume that ψ is the **Huber influence function** defined as

$$\psi(t) = \begin{cases} t & \text{if } -c < t < c \\ c & \text{if } |t| > c \end{cases}$$

where c is a “tuning” parameter.

As $c \downarrow 0$, robustness (and variability) \uparrow (we use $c = 2$)

An illustration (fertility survey data)



Using M-quantile Models to Measure Area Effects

- Assume that we have individual level data on y and \mathbf{x} . Each sample value of (x,y) will lie on one and only one M-quantile line. We refer to the q -value of this line as the **M-quantile coefficient** of the corresponding sample unit or its q value. Every sample unit will have an associated q value
- In order to estimate these unit specific q values, we define a fine grid of q -values, e.g. $g = (0.001, \dots, 0.999)$ that adequately “covers” the conditional distribution of y and \mathbf{x} . We fit an M-quantile line for each q -value in g and use linear interpolation to estimate a **unique** q value, q_k , for each individual k in the sample

Using M-quantile Models to Measure Area Effects (contd)

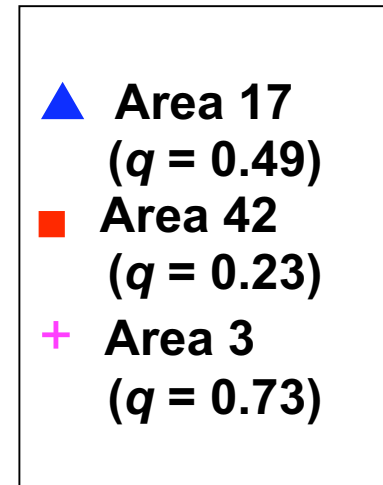
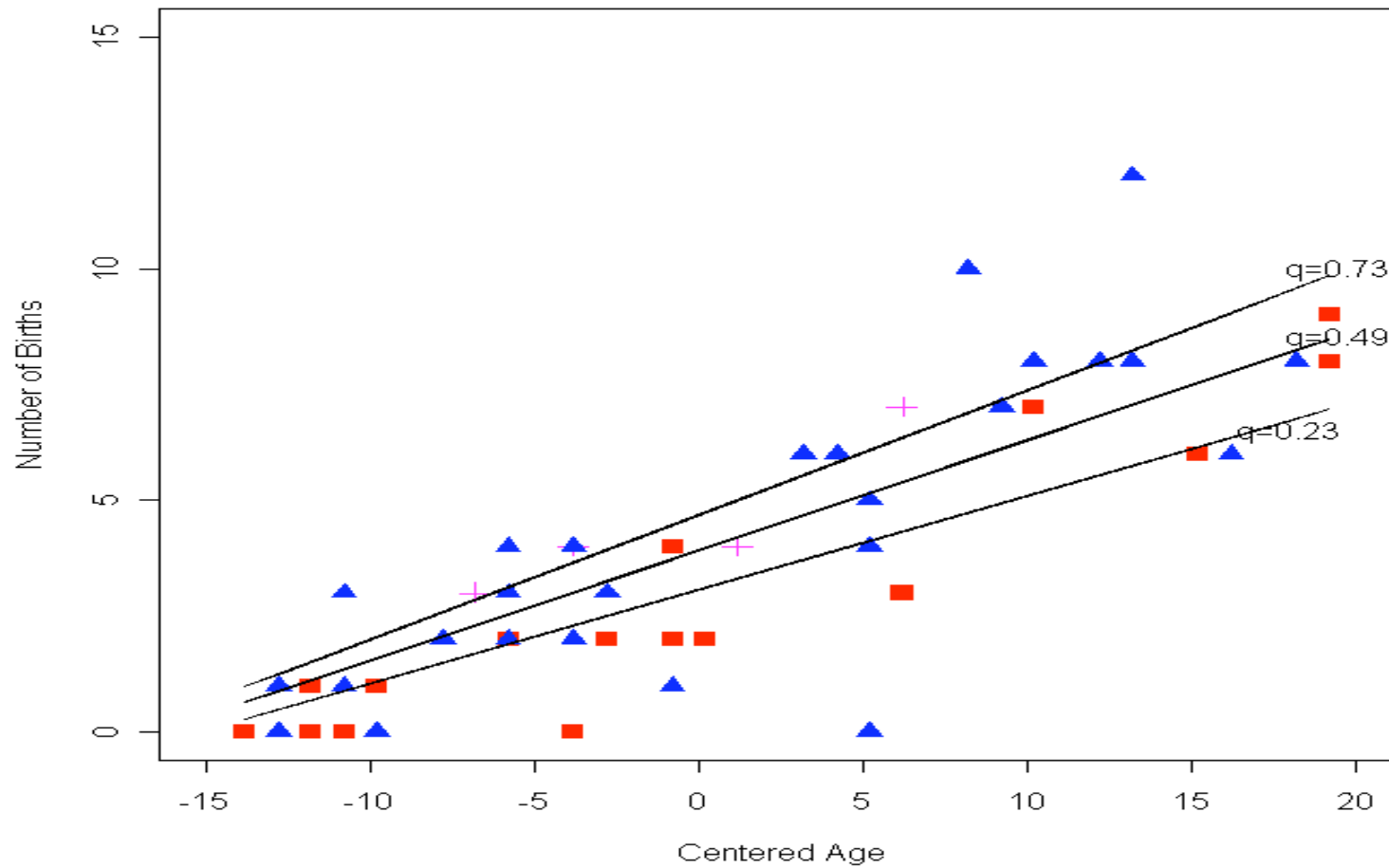
- Calculate an M-quantile coefficient for each area i by suitably **averaging** the q values of each sampled individual in that area. Denote this **area-specific q -value** by \bar{q}_i
- Estimate the area specific mean by

$$\hat{y}_i = \hat{\beta}_0(\bar{q}_i) + \bar{\mathbf{x}}_i^T \hat{\beta}_1(\bar{q}_i)$$

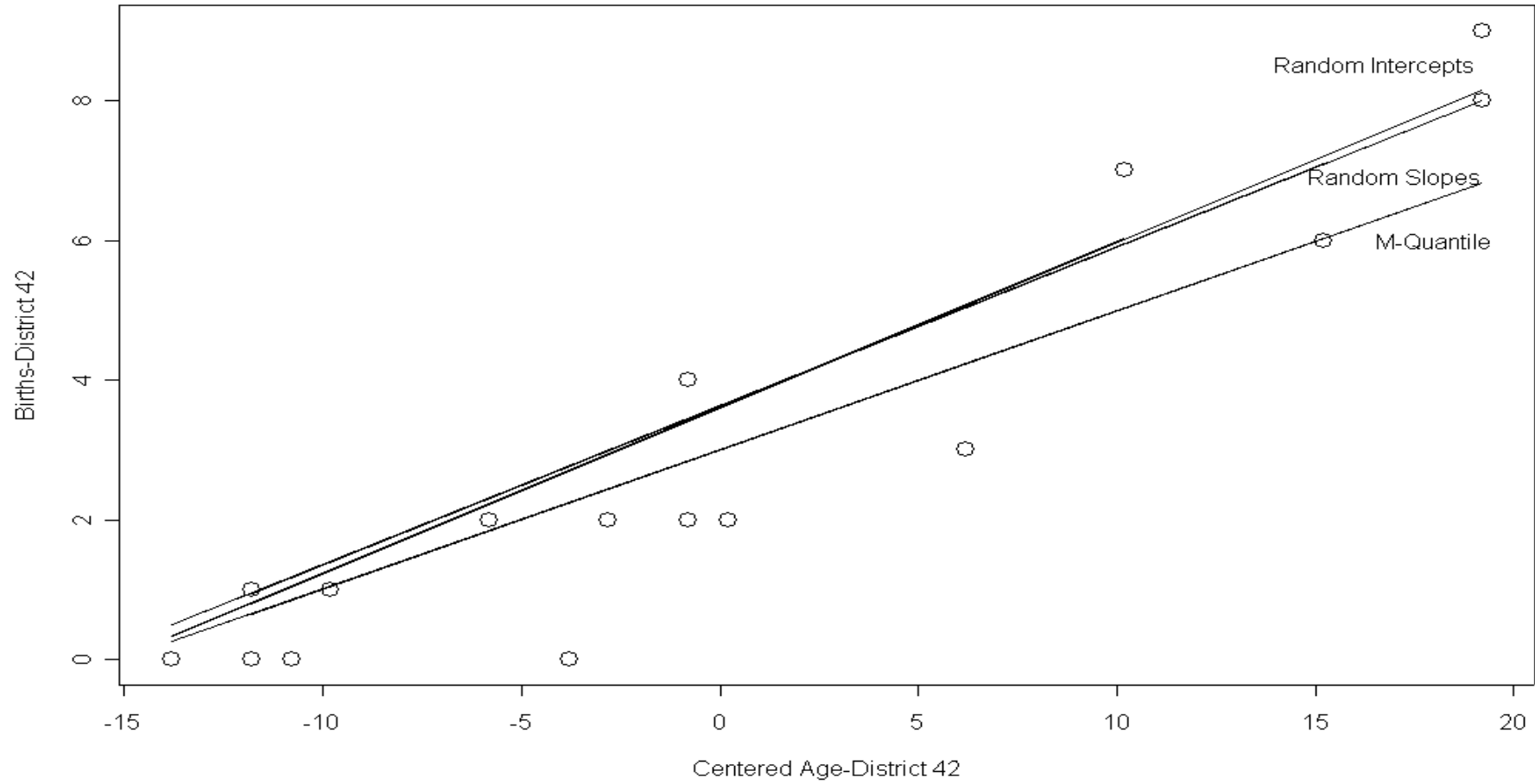
Comparing the Multi-quantile and Multi-level approaches

- Random Intercepts/Random Slopes model fitted in R using *nlme* (REML option)
- Random Slopes model rejected (change in deviance compared with chisquare on 2 df)
- M-quantile models fitted in R using IRLS with imposed monotonicity

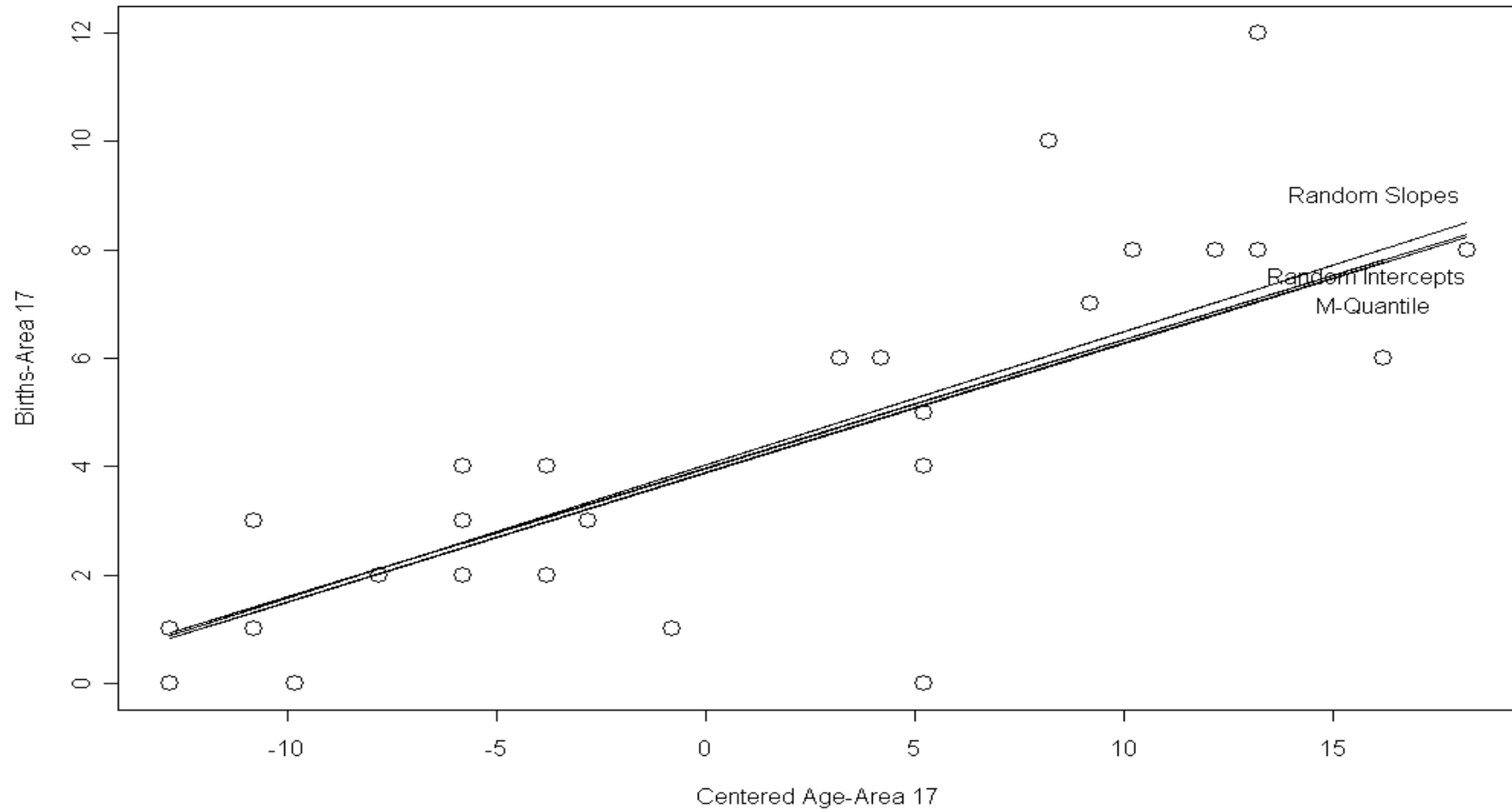
Application to Fertility Survey data



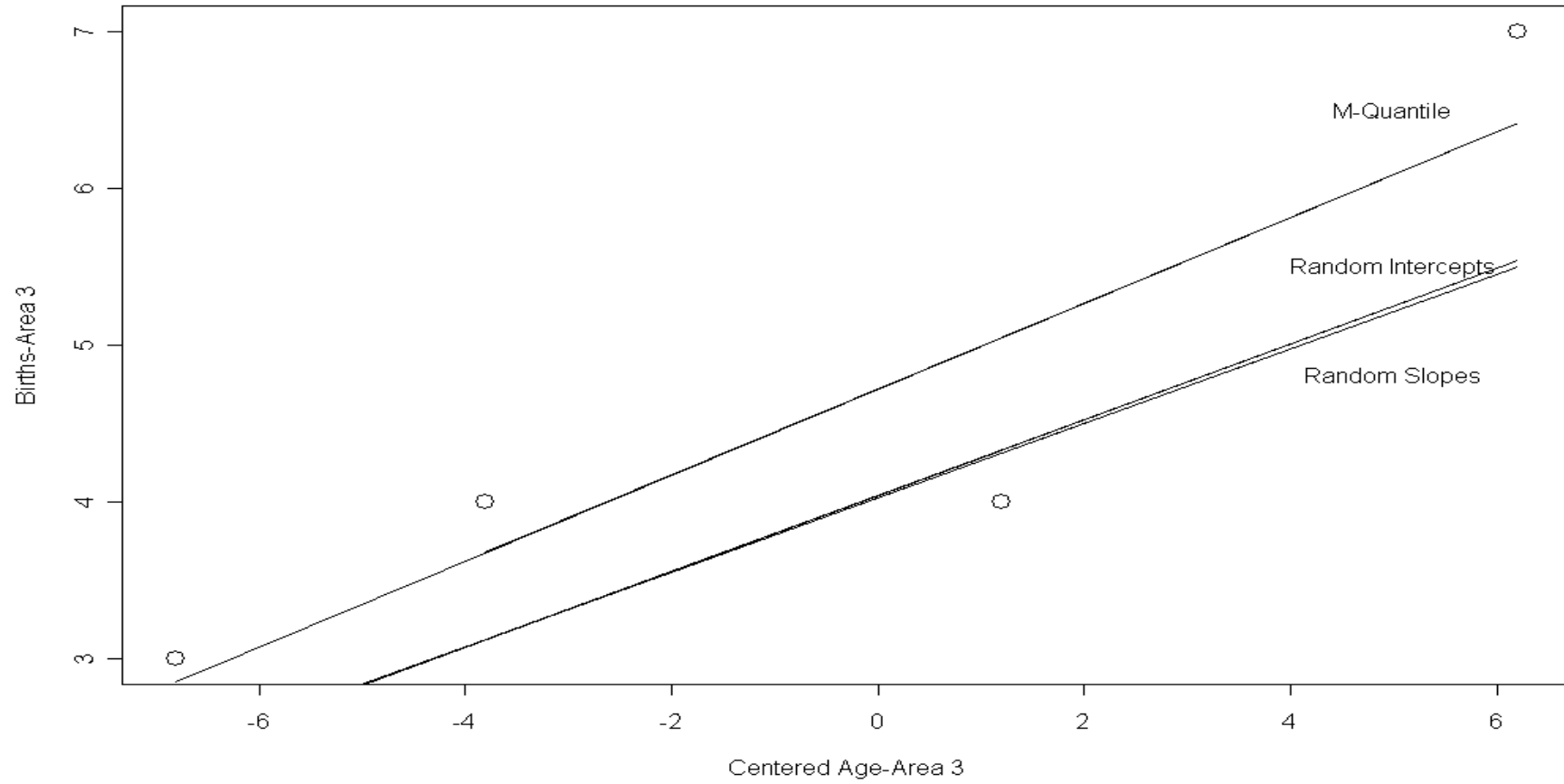
Area 42 ($q = 0.23$)



Area 17 ($q = 0.49$)



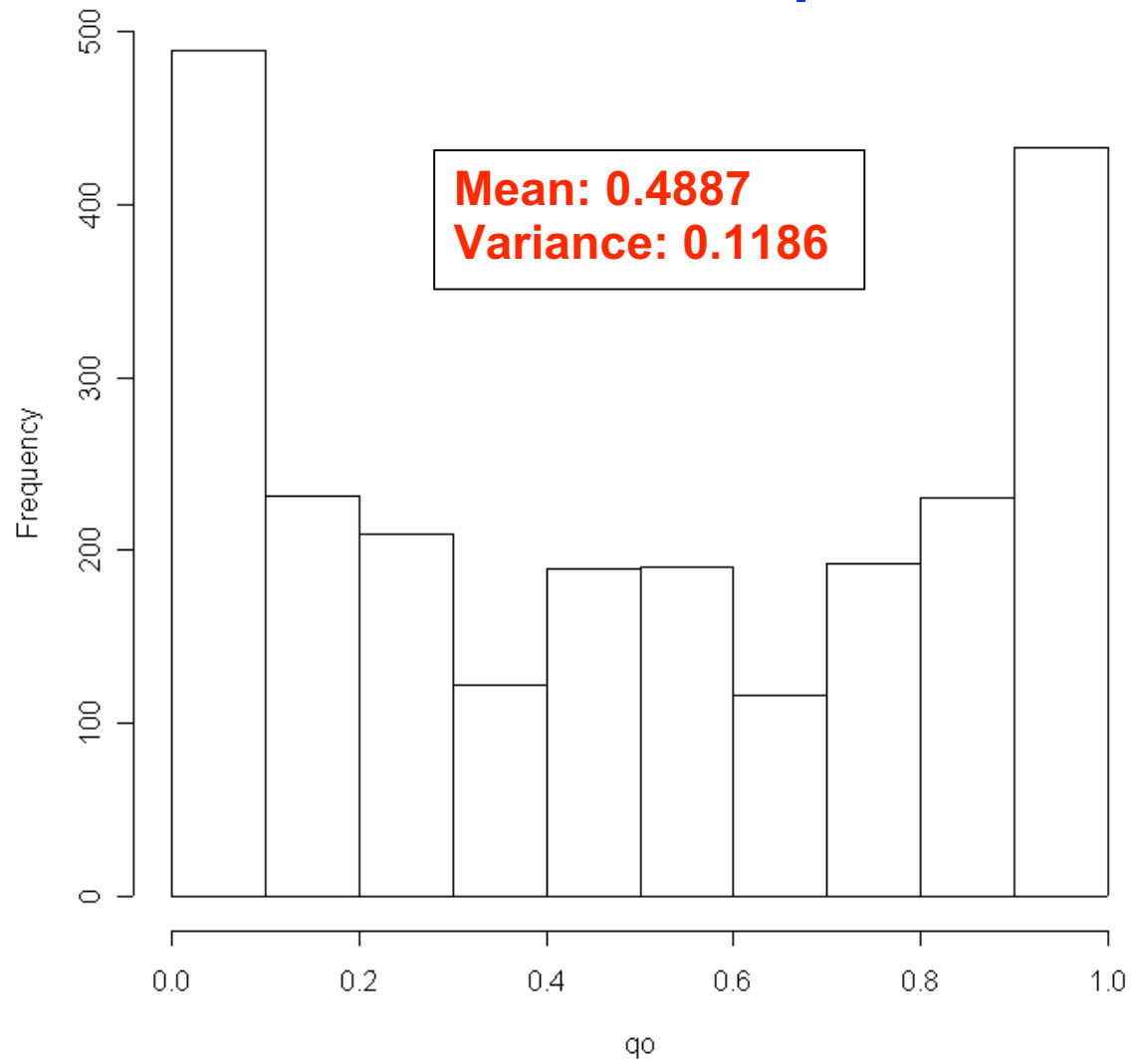
Area 3 ($q = 0.73$)



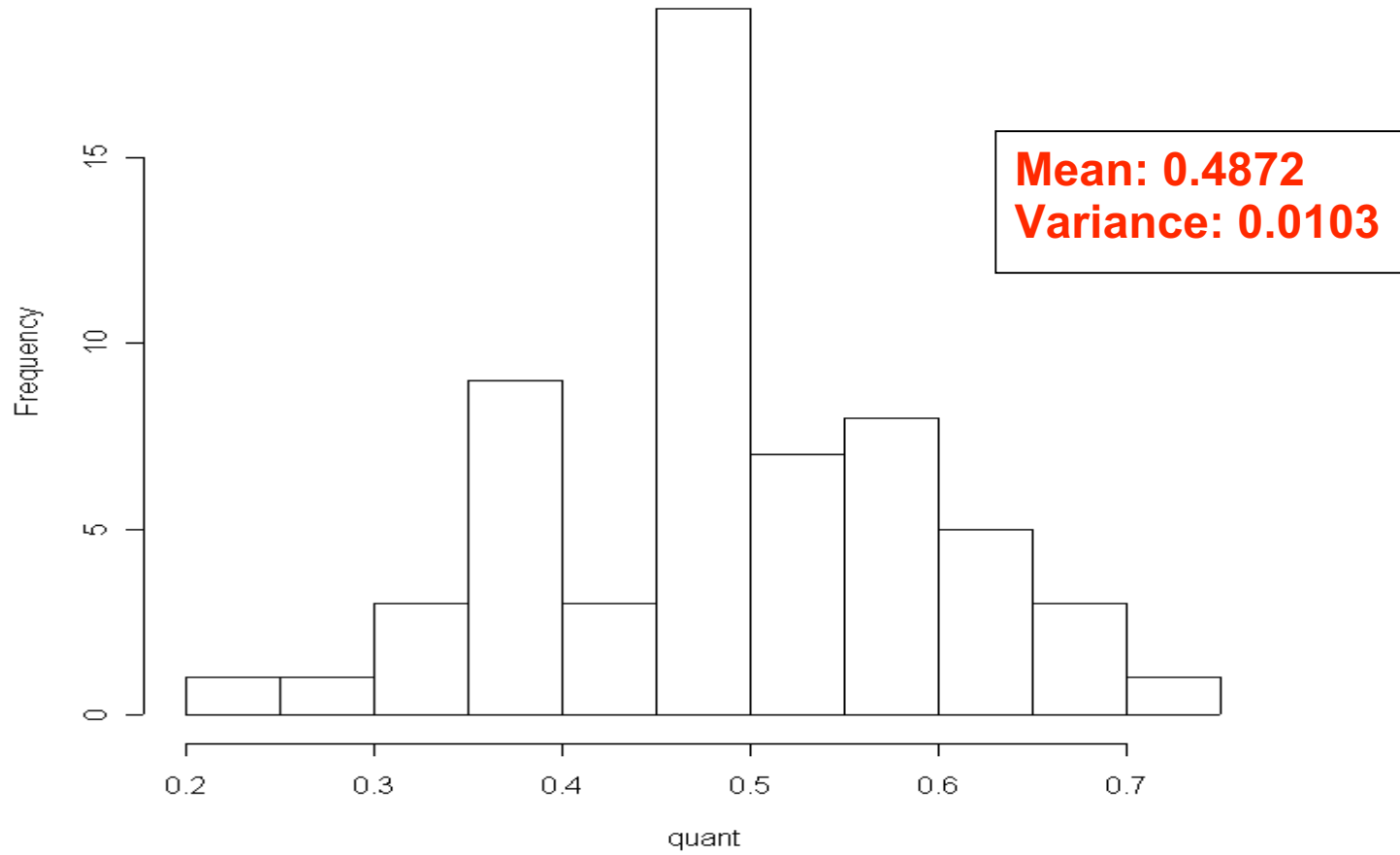
Relationship Between Multi-quantile and Multi-level Models

- The multi-level model uses **random area effects** to capture the dissimilarity between small areas – in contrast, the multi-quantile model captures this dissimilarity via **area-specific q values**
- The unit-specific q values are estimated without any reference to the areas and hence their within area averages (the area-specific q values) can be used as **diagnostics** for the existence of area (and other) effects
- How we average unit q values to get area q values is quite general. We can compute medians, **sample weighted means** etc etc ...

Distribution of unit q values

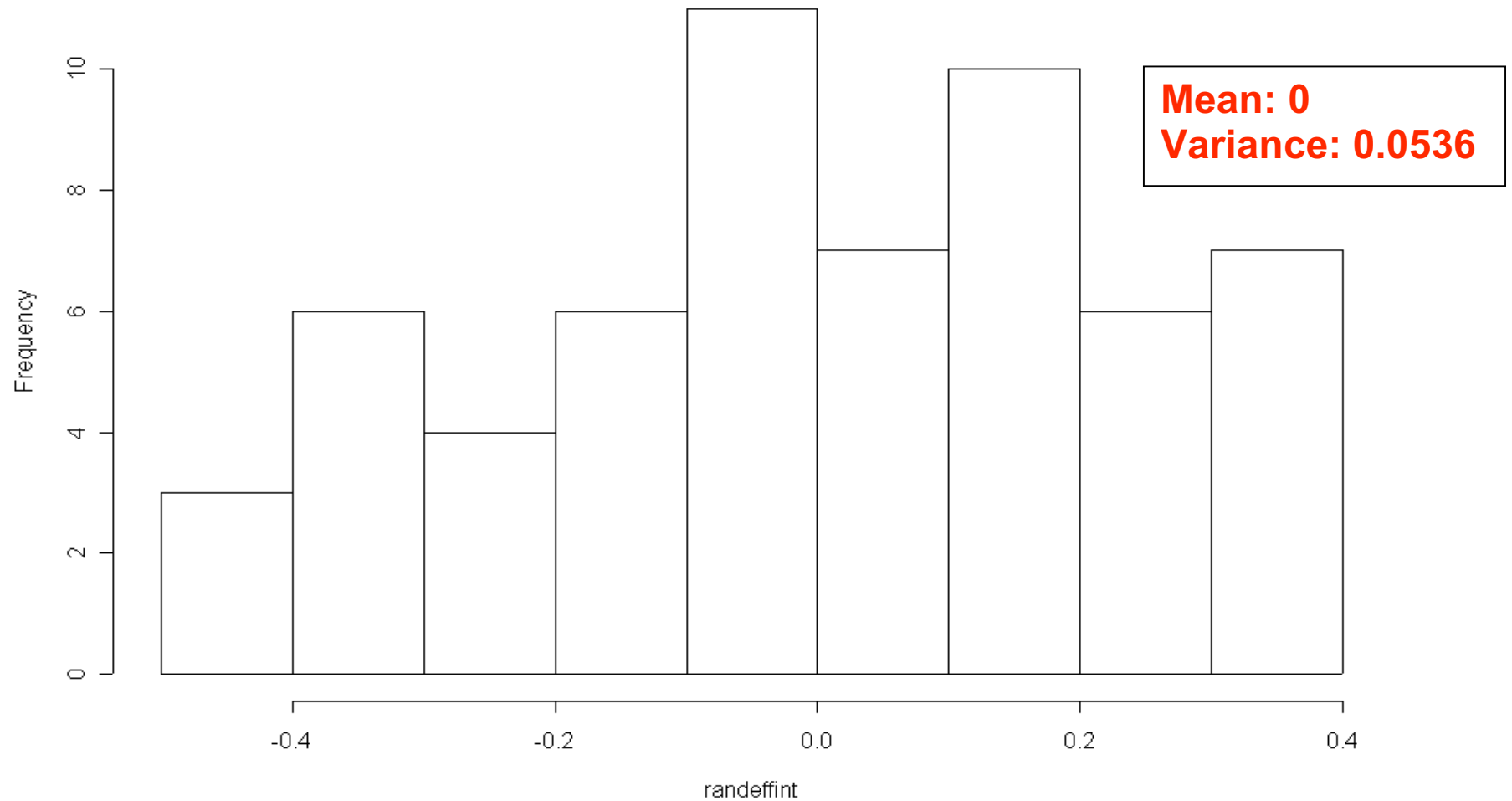


Distribution of area q values



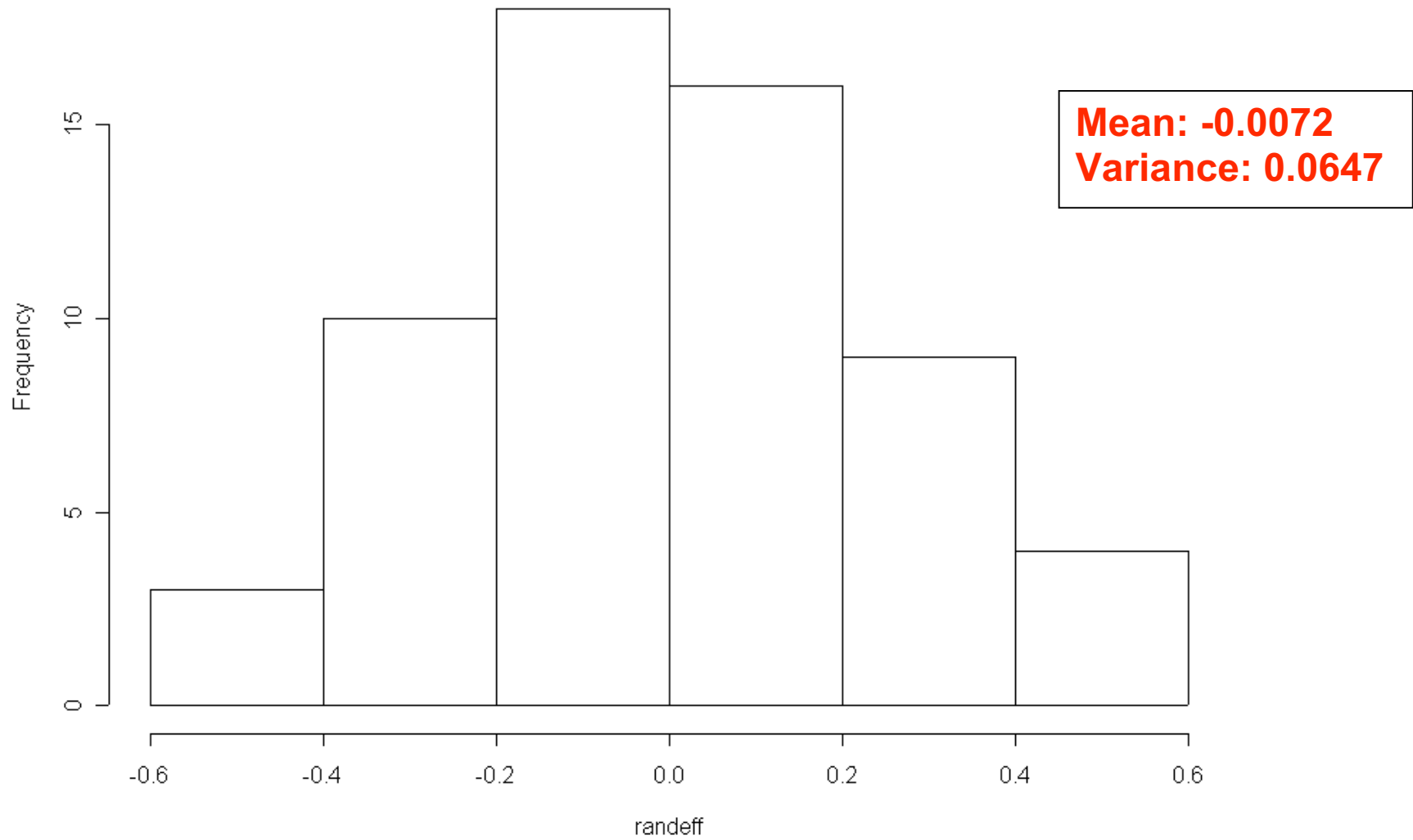
A Measure of Areal Homogeneity?
 $0.0103 / (0.1186) = 0.086$

Distribution of area effects (Random Intercepts model)



Estimated Intraclass Correlation:
 $0.1072 / (0.1072 + 3.48) = 0.0298$

Distribution of area effects (Random Slopes model)



Simulation Study

Data: Bangladesh Fertility Survey (DHS, 1988)

Variables: Number of births (y), age of the mother (x)

Sample Size: 2401 individuals, 60 districts

Design of Simulation Study:

- Create a population of size $N = 50,000$ by bootstrapping the original dataset
- From this population select a SRSWOR sample from each area of size equal to the size of the original area sample. Repeat 500 times

Simulation Study (contd)

- Small area estimates were calculated using: (a) the Direct estimator, (b) the Random Intercepts model, (c) the Random Slopes model and (d) the M-quantile model. **Note that the Random Slopes model is not supported by the data but it is included because of the obvious heteroskedasticity in the data**
- The grid for computing the individual q values was $g = (0.006, \dots, 0.994; \text{step} = 0.045)$
- The parameter of interest is the population mean number of births within each district

We evaluate the performance of the alternative modelling strategies using the following criteria:

MSE of area level means (area specific performance)

Bias in weighted average of estimated area level means
(average performance over areas)

MSE of weighted average of estimated area level means
(average performance over areas)

Distribution of Shrinkage Factors - shrinkage factor for a specific simulation defined by ratio of variance of direct estimates across areas to corresponding variance of model-based estimates

Results

Average correlation between the area-specific q values and estimated area-specific random effects

Random Intercepts model	Random Slopes model
0.89	0.81

Weighted overall analysis

Estimator	Relative % Bias	Relative Std Deviation	Relative Root MSE
M-quantile	-1.262	0.094	0.095
Random Intercepts	0.252	0.096	0.096
Random Slopes	0.252	0.099	0.099
Direct	0.252	0.110	0.110

*Weighted overall population mean = 3.96

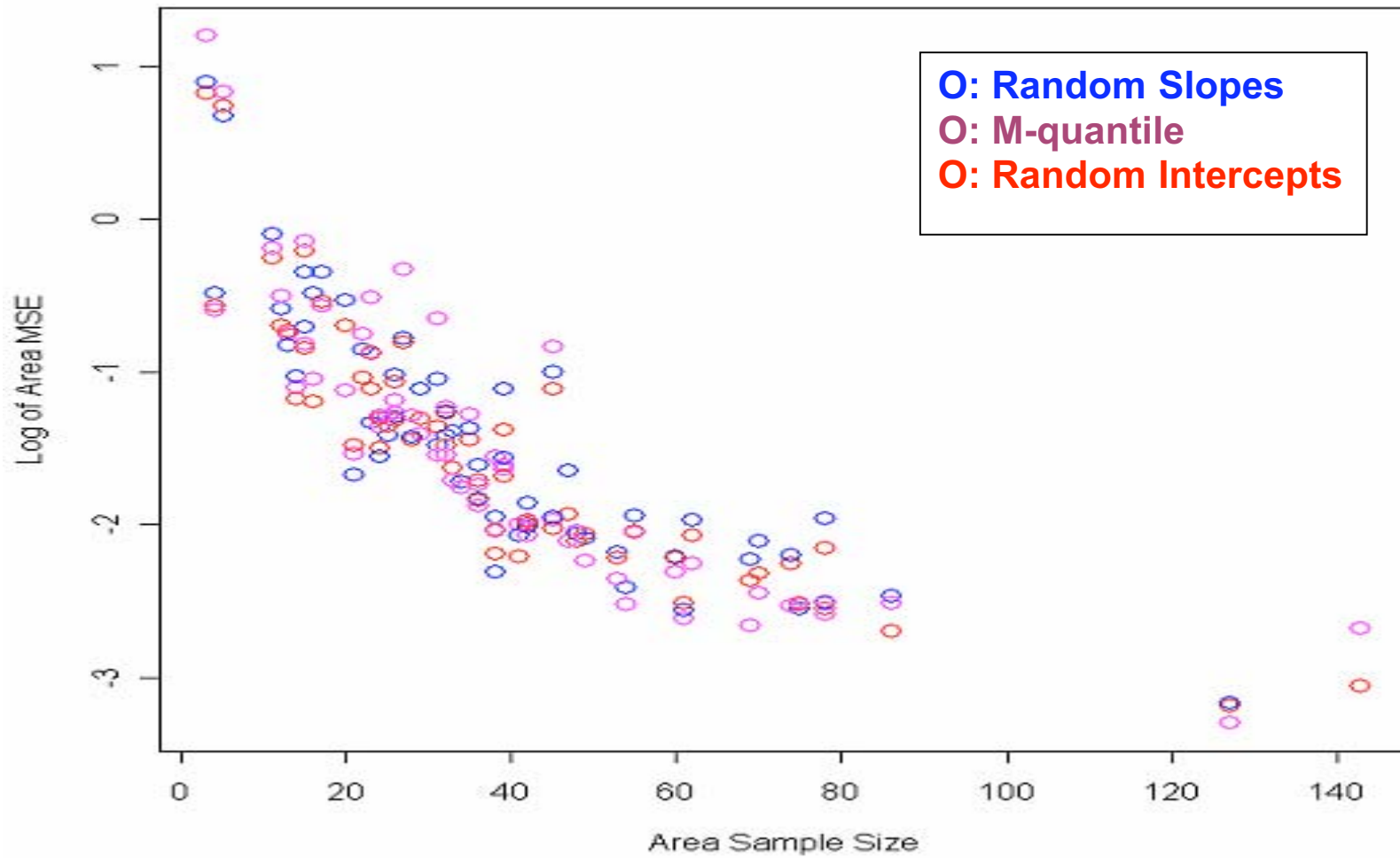
Relative efficiency of small area means using the different models

MSE(M-quantile) < MSE(Random Slopes) 36/60

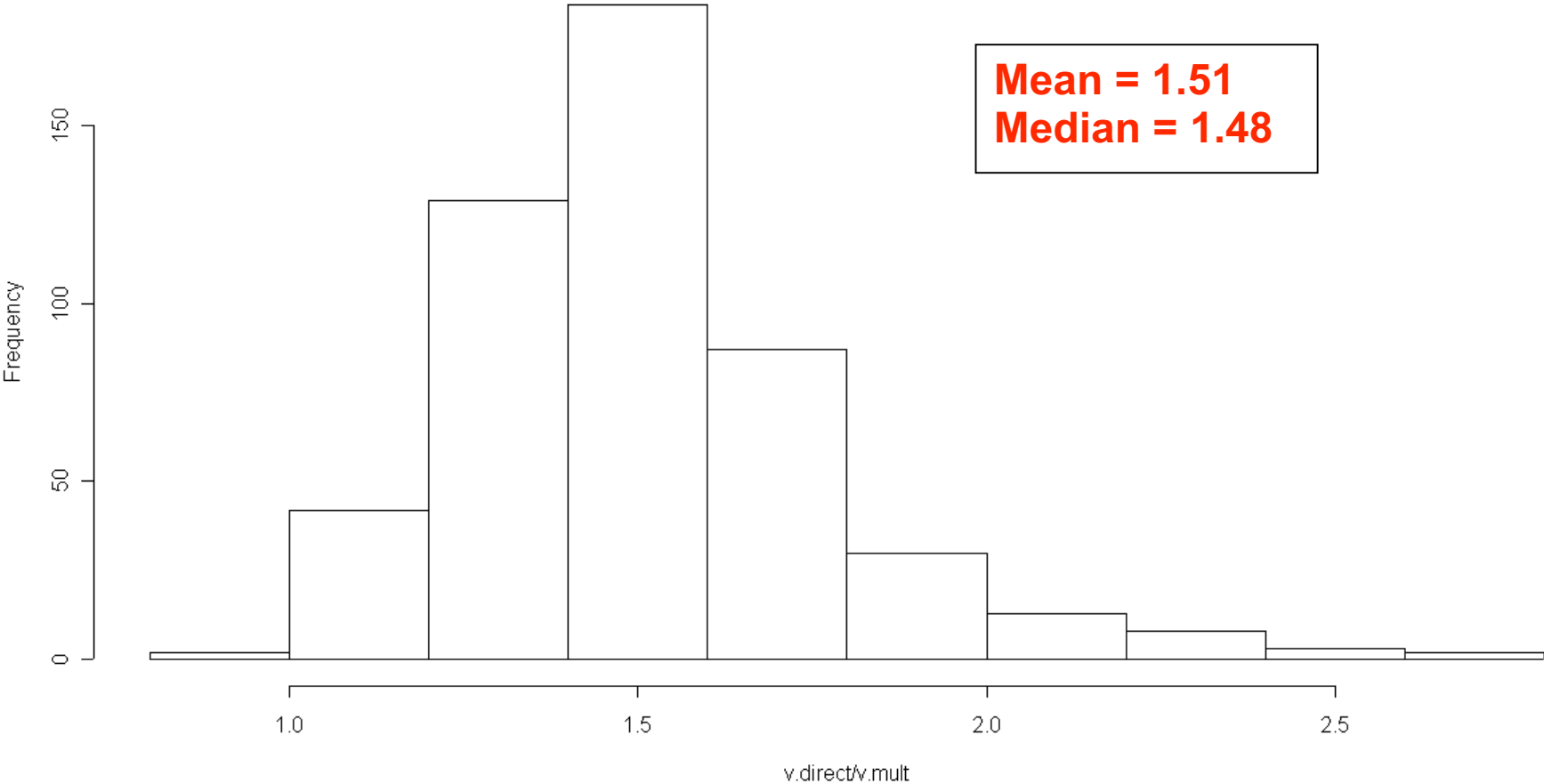
MSE(M-quantile) < MSE(Random Intercepts) 29/60

MSE(Random Intercepts) < MSE(Random Slopes) 46/60

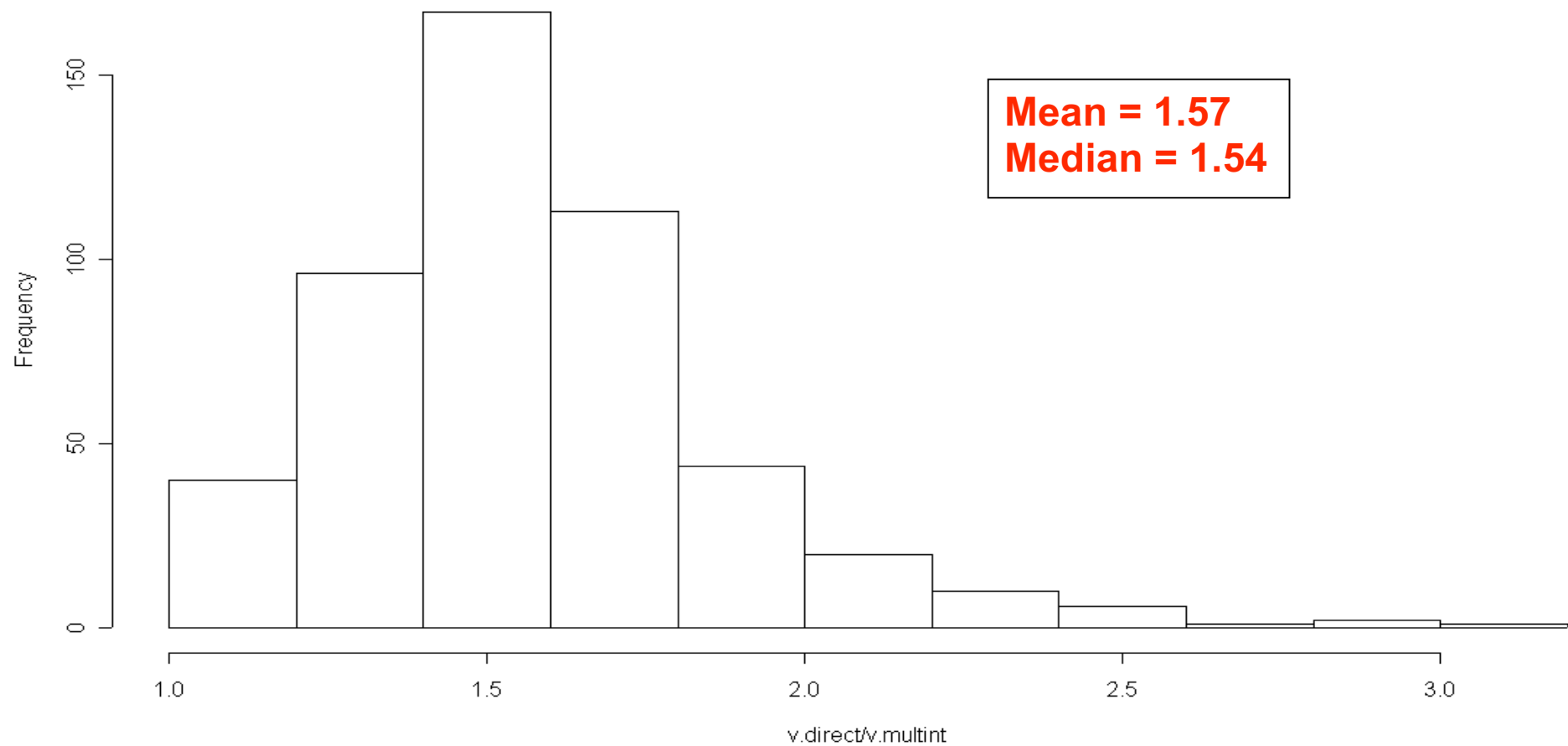
Area level MSEs



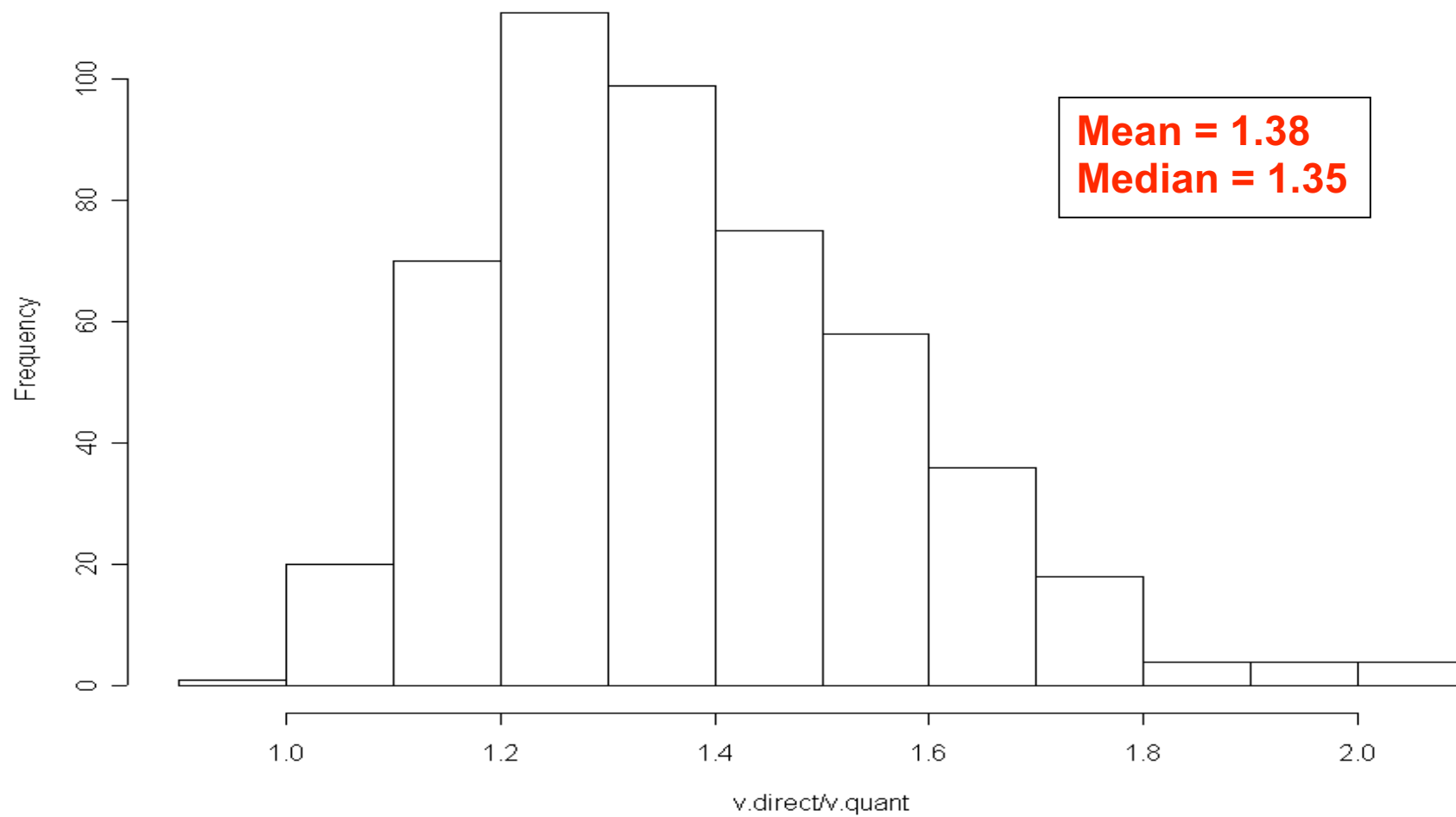
Distribution of Shrinkage Factors – Random Intercepts Model



Distribution of Shrinkage Factors – Random Slopes Model



Distribution of Shrinkage Factors – M-quantile Model



Some Initial Conclusions

- As far as average and area specific MSEs are concerned, the M-quantile and Random Intercepts models seem roughly equivalent for this population and both dominate the Random Slopes model
- The M-quantile model leads to significantly less shrinkage than both the Random Intercepts and Random Slopes models

Advantages of Multi-quantile Approach

- No distributional assumptions on the error term
- No modelling assumptions analogous to random intercepts or random slopes - the data guide the modelling process
- Sample weighting is straightforward
- Outlier robust inference is straightforward
- Estimation is straightforward (IWLS)
- Easily extended to non-parametric modelling
- Potential solution to the Modifiable Area Unit Problem (MAUP)? Changes in geography do not change individual level q values ...

Disadvantages and Unresolved Issues for M-quantile Models

- Inefficient if the multi-level model is in fact true
- Not appropriate if our aim is to compare between area variability with within area variability
- Need to ensure monotonicity (in q) of M-quantile lines
- Asymptotic theory? Standard errors?
- M-quantile models for discrete variables?
- Small areas with no sample? (problem for multi-level models as well)

References on M-quantiles

- Breckling, J. U. and Chambers, R. L. (1988). M-Quantiles. *Biometrika* **75**, 761 – 771.
- Kokic, P. N., Chambers, R. L., Breckling, J. U. and Beare, S. (1997). A measure of production performance. *Journal of Business and Economic Statistics* **10**, 419 - 435.
- Kokic, P., Chambers, R. and Beare, S. (2000). Microsimulation of business performance. *International Statistical Review* **68**, 259-275.

References on Quantile Regression

- Hogg, R. V. (1974). A Modification of the Brown-Mood Regression Estimate for Percentile Lines. *Technical Report No. 34*, Department of Statistics, University of Iowa.
- Koenker, R. and Bassett, G. (1978). Regression Quantiles. *Econometrica*, 46, 33-50.
- Koenker, R. and Hallock, K. F. (2001). Quantile Regression. *Journal of Economic Perspectives* **15**, 143 – 156.
- Melly, B. (2001). The Theory and Practice of Quantile Regression. *Diplomarbeit*, Universität St Gallen, Hochschule Für Wirtschafts-, Rechts- Und Sozialwissenschaften.