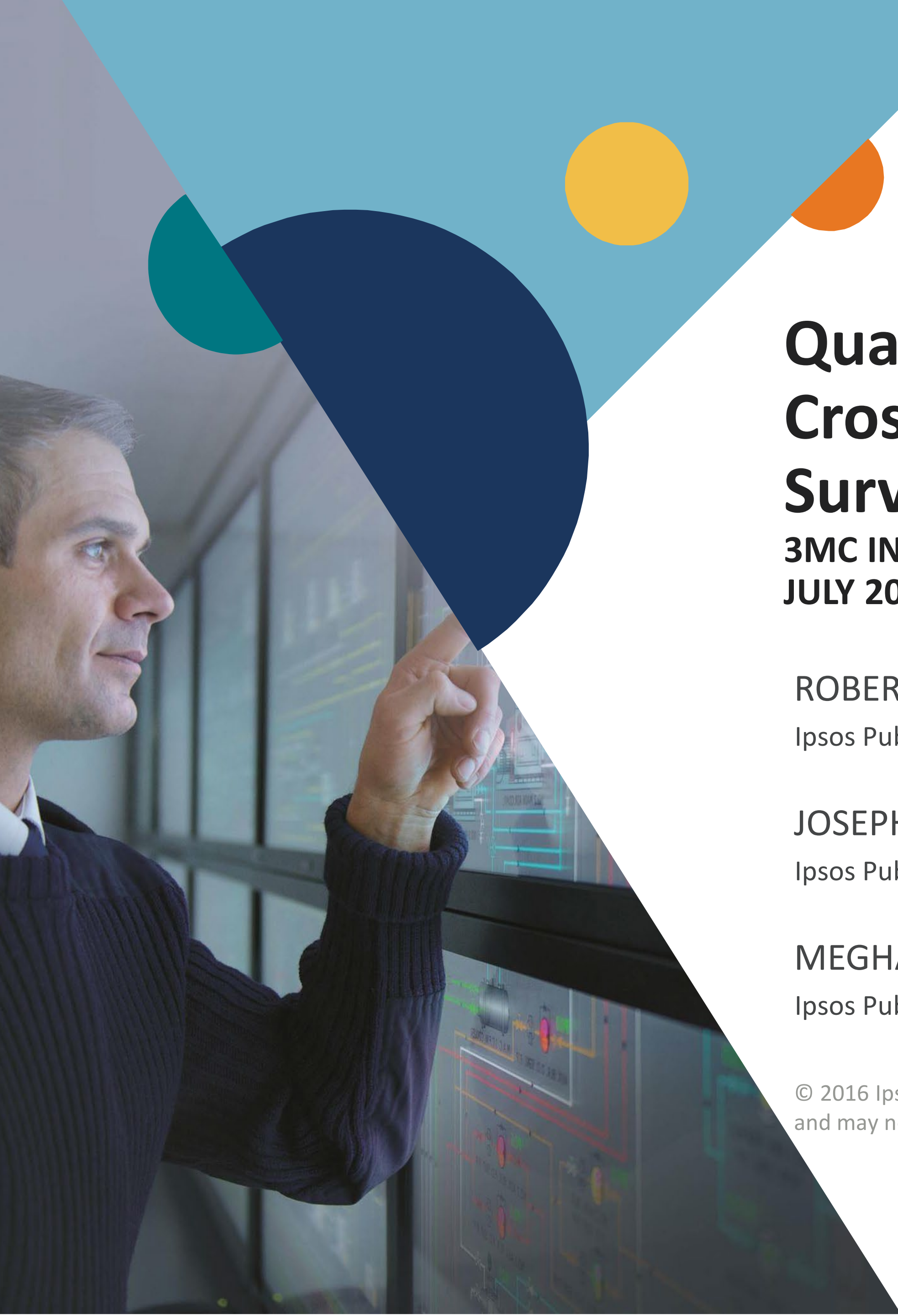


Ipsos Public Affairs

Quantile Regression as a Tool for Cross-National and Comparative Survey Research

3MC INTERNATIONAL CONVERENCE, CHICAGO ILLINOIS JULY 2016

ROBERT A. PETRIN, PH.D.
Ipsos Public Affairs

JOSEPH ZAPPA, M.S.
Ipsos Public Affairs

MEGHANA RAJA
Ipsos Public Affairs

GAME CHANGERS

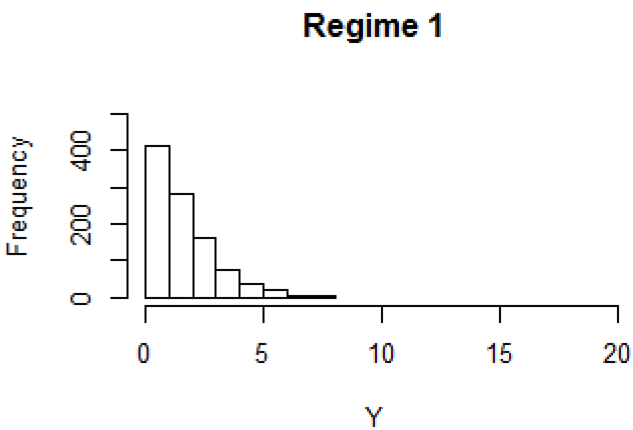
Contents

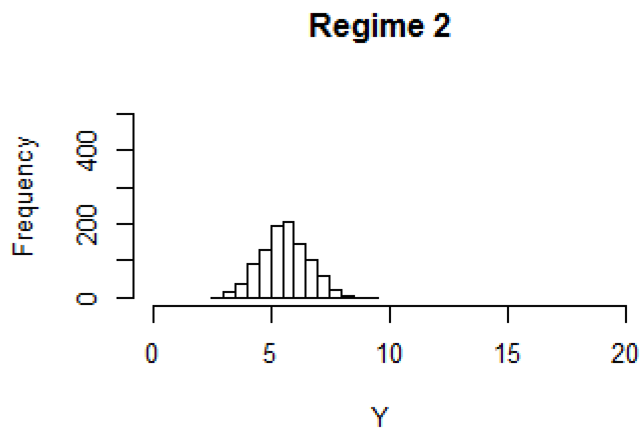
CONTEXT AND MOTIVATING CHALLENGES

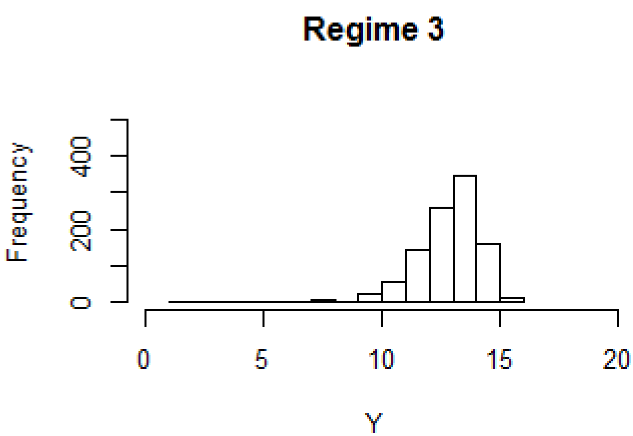
Qualitative and Quantitative Variation

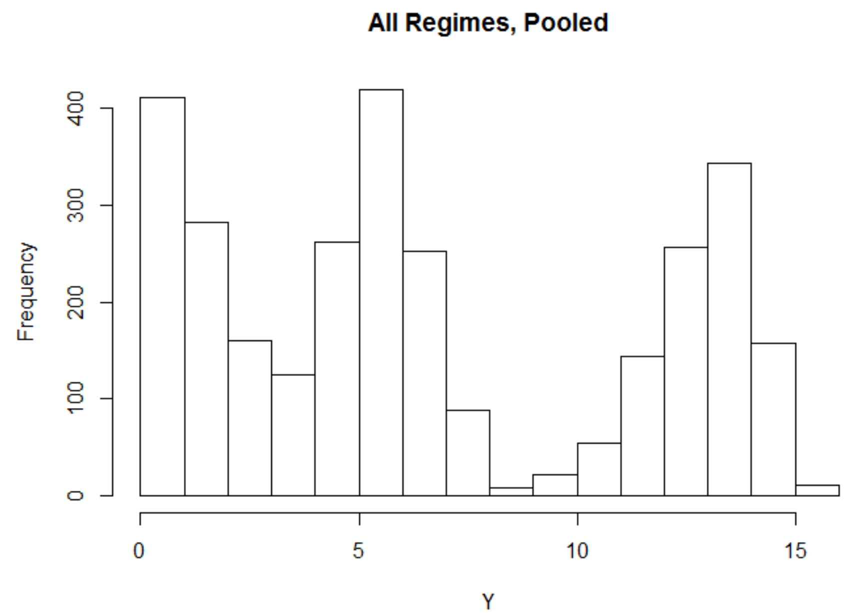
- Survey and evaluation research require researchers to take into account the particulars of populations and phenomena studied
  - **In agricultural surveys**, can often see qualitative variation in key measures across regions and countries, especially when subset on key groups
    - Distribution of arable soil can impact crop yields and crop quality dramatically across study regions
  - **In longitudinal evaluation research**, it is reasonable to expect that the intervention itself can alter both the qualitative and quantitative nature of the focal measures
    - **A business intervention which promotes shifting product mixes** to enhance profitability could lead to temporary re-alignment of goods and services sold and a thus temporary or longer-term shifts in the distribution of sales / profits
    - **An intervention which promotes financial record keeping** could lead to qualitative shifts in the distributions of reported income

Quantitative vs. Qualitative Variation









Other Considerations

- When setting up (parameterizing) quantitative analyses, have to be careful to be context-sensitive and not impose prior conceptions on what population phenomena look like
  - US overall income distribution (gamma-distributed) vs. income distribution among Sari-Sari store owners in the greater Manilla (Philippines) area (?-distributed)
    - **Estimating the distribution of Y could be an essential component of the study itself** to the extent that it either provides a better sense of a novel or unfamiliar phenomena, informs subsequent design, or enables researchers to better incorporate information on the error distribution into subsequent analyses
  - When designing studies which involve regionally-situated phenomena, have to recognize that “social” space and “social” geographies do not necessarily align with administrative space/geographies

Common Tacts

- Conventional statistical analyses are parametric, require researchers to <u>specify the distribution of the outcome Y</u>
  - Consequences of **misspecification** (bias, incorrect SEs / CIs, mis-interpretation of effects) can be substantial (e.g., Long, 1997), and not necessarily remedied by just “getting a larger sample”

- Common solutions:
  - **Transform Y** to better fit a conventional model (i.e., ln(Y), Y^{-1})
  - **GLMs** (McCullagh and Nelder, 1989) to model Y on it’s native scale, obtain correct SEs, reasonable-to-interpret effects (i.e., gamma regression, negative binomial regression, etc.)
    - However, as with transformations, <u>GLMs not allow differential-modeling of Y</u> distributions (can’t transform some groups but not others)

QUANTILE REGRESSION AND BAYESIAN QUANTILE REGRESSION

Quantile Regression

- Quantile regression (QR) offers a reasonable and highly flexible (more comprehensive) alternative to GLMs in many scenarios (Davino et al., 2014; Koenker, 2005)
  - GLMs: model conditional mean
  - QR: models conditional quantiles (median, 75^{th} percentile, etc.)

- OLS vs. QR loss functions
  - OLS: $\mu = \text{argmin}_c\ E[Y\text{-}c]^2$
  - QR for Me: $Me = \text{argmin}_c\ E|Y\text{-}c|$
  - QR for general quantile $\theta = P(Y \leq y)$: $q_\theta = \text{argmin}_c\ E|\rho_\theta(Y\text{-}c)|$

The ***ρ*** *are often referred to as* ***weights****, which are defined by a* ***check function***

$$\rho_\theta = [(1\text{-}\theta)\ I(y \leq 0)] + \theta\ I(y>0)]\ |y|$$

  - QR conditional on X for quantile θ: $b(\theta) = \text{argmin}_b\ E|\rho_\theta(Y\text{-}Xb)|$

Quantile Regression - II

- In the contexts described above, QR can be particularly attractive option (Davino et al., 2014; Koenker, 2005; McMillen, 2013)
  - **Inferences about Y | X relationships are distribution free**: method makes no assumption about the distribution of Y
    - Put differently, QR makes no assumptions about the error distribution for Y, and is thus robust to model misspecification (QR can in fact be used to **estimate the distribution of Y | X**)
  - QR readily amenable to **estimating percentile intervals in the data** (i.e., "80% of white male respondents with 16 years of education have incomes within the range $Y_L - Y_U$")
    - Potentially useful property for clients who seek an alternate way of understanding where substantively meaningful slices of the population fall
    - Relatedly, QR is amenable to **threshold analyses**

Going Bayesian

- Classical QR is not new per se; Bayesian QR is (relatively)
- In general, Bayesian methods allow researchers to model not just response Y, but also regression coefficients (Carlin and Louis, 2009):

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_i + e$$

$$\beta_{0j} = \gamma_{00} + \mu_{0j}$$

$$\beta_{1j} = \gamma_{10} + \mu_{1j}$$

- Treating Y and β_p as explicit random variables has numerous advantages in research where **contexts (geographic, social) are important** (Gelman and Hill, 2006)
  - **β_p can vary across aggregations** (such as regions or socio-political entities) in a way that adjusts for (allows evaluation of) contextual / cluster differences (Raudenbush and Bryk, 2001)
    - **Example from education research**: what are the characteristics of schools which have lower black/white test score differentials? How are school means correlated with black/white differentials?

Going Bayesian - II

- Bayesian paradigm provides principled mechanisms for **incorporating results from previous studies** or data collection efforts into analysis
  - Elicited priors (Gill and Walker, 2005; see also Rendell et al., 2009)
  - Bayesian updating for longitudinal studies (Carlin and Louis, 2009)
  - **WIPs** to stabilize estimation (Chung et al., 2015; Gelman et al., 2008)
- Suggests that payoffs to <u>integrating Bayesian methods with QR approach</u> could be substantial …

Bayesian Quantile Regression

- **Bayesan quantile regression (BQR)** attempts to achieve the same flexibility that HLMs / HGLMs enjoy – specification of priors for model coefficients, contextualizing models via random effects
  - Also, robustify quantile-specific estimates with **small samples**

- Bayesian extensions of QR have a couple of different flavors:
  - **Associated likelihood is approximate** and there are a handful of different ways to parameterize (e.g., Feng, 2015; Yang and He, 2012; Yu et al., 2001), some of which are slightly more robust than others in certain contexts
    - And not all of which allow joint estimation of quantiles…

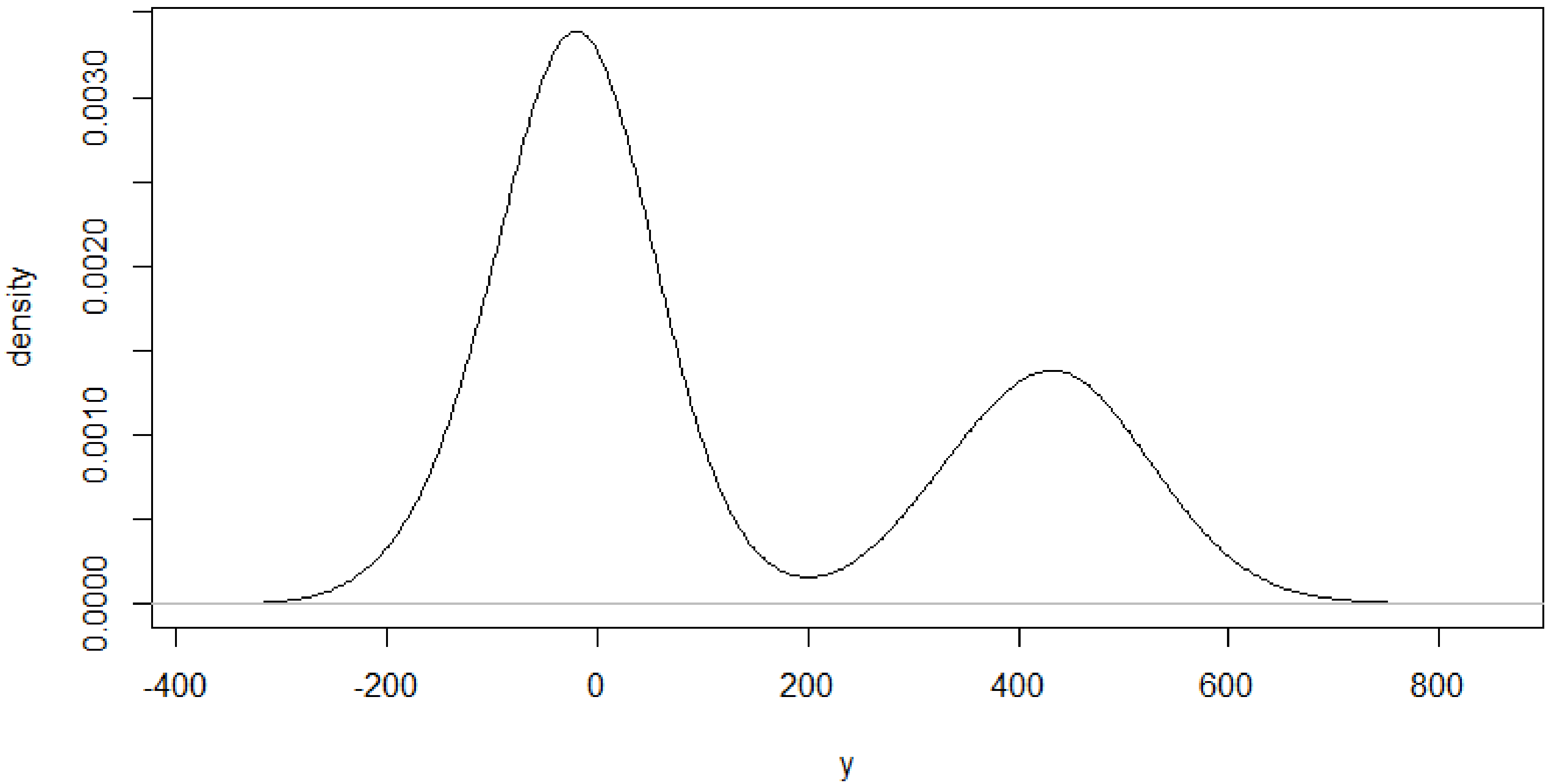
- Still, another payoff of to Bayesifying QR is that can potentially integrate **Bayesian approaches to incorporating sample design information** (Gelman et al., 2013; Si et al., 2015)

ILLUSTRATION

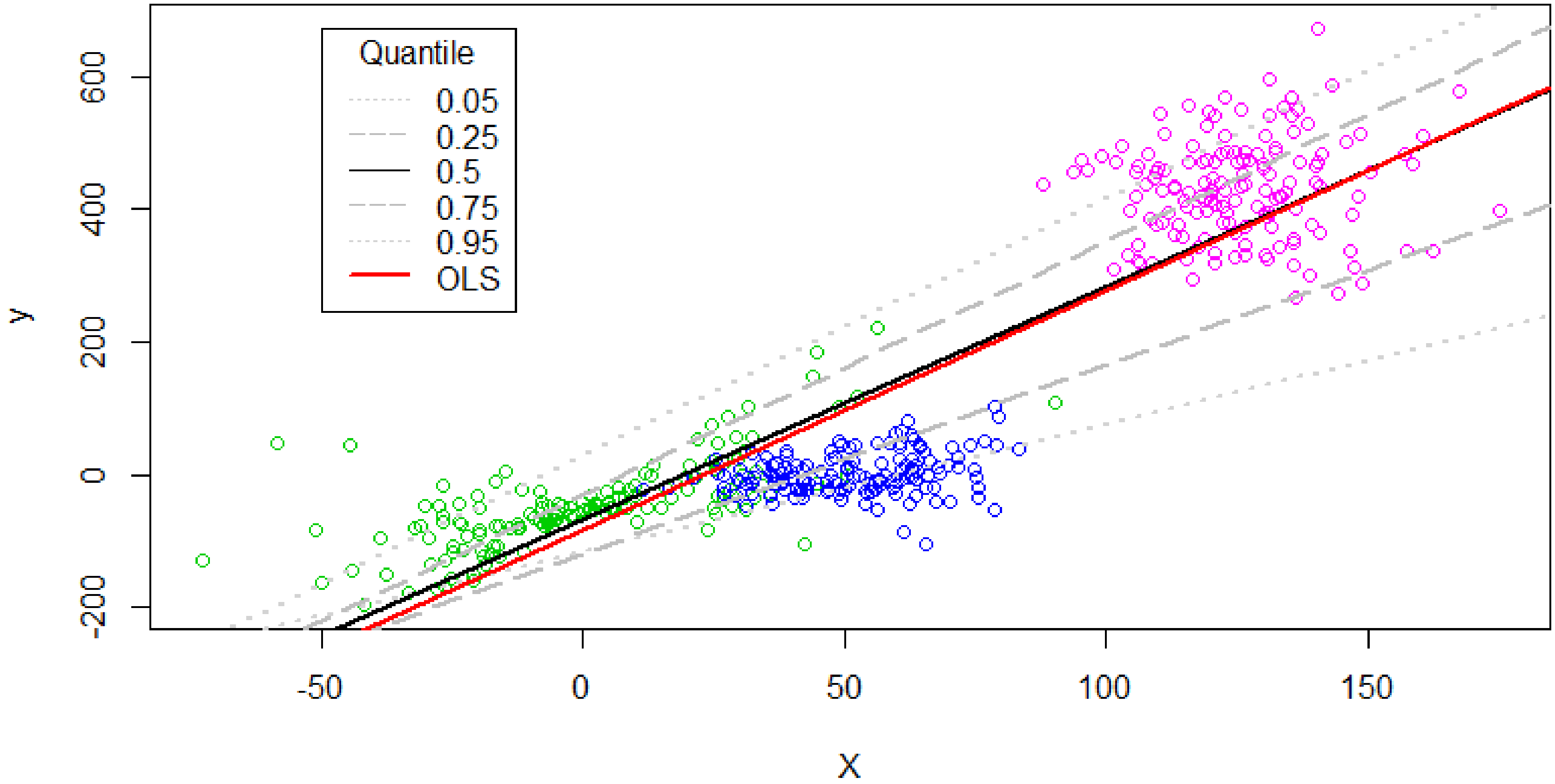
Simulation

- Data were generated from three different population distributions, N=10,000
  - Normal(0,25)
  - Gamma(10,5)
  - Chi Squared(125)
- A relationship with the response variable was applied (Y|X), and the data were combined to form a dependent variable with a single continuous independent variable
- Sample draws made of n=150 per distribution
- The error is heterosckedastic, non-normal, and differing distributions across the values of the independent variable
  - In other words, a worst case scenario for performing OLS

The Density of the Response



Best-Fit Lines (OLS, BQR with plat prior)



Why BQR vs. OLS?

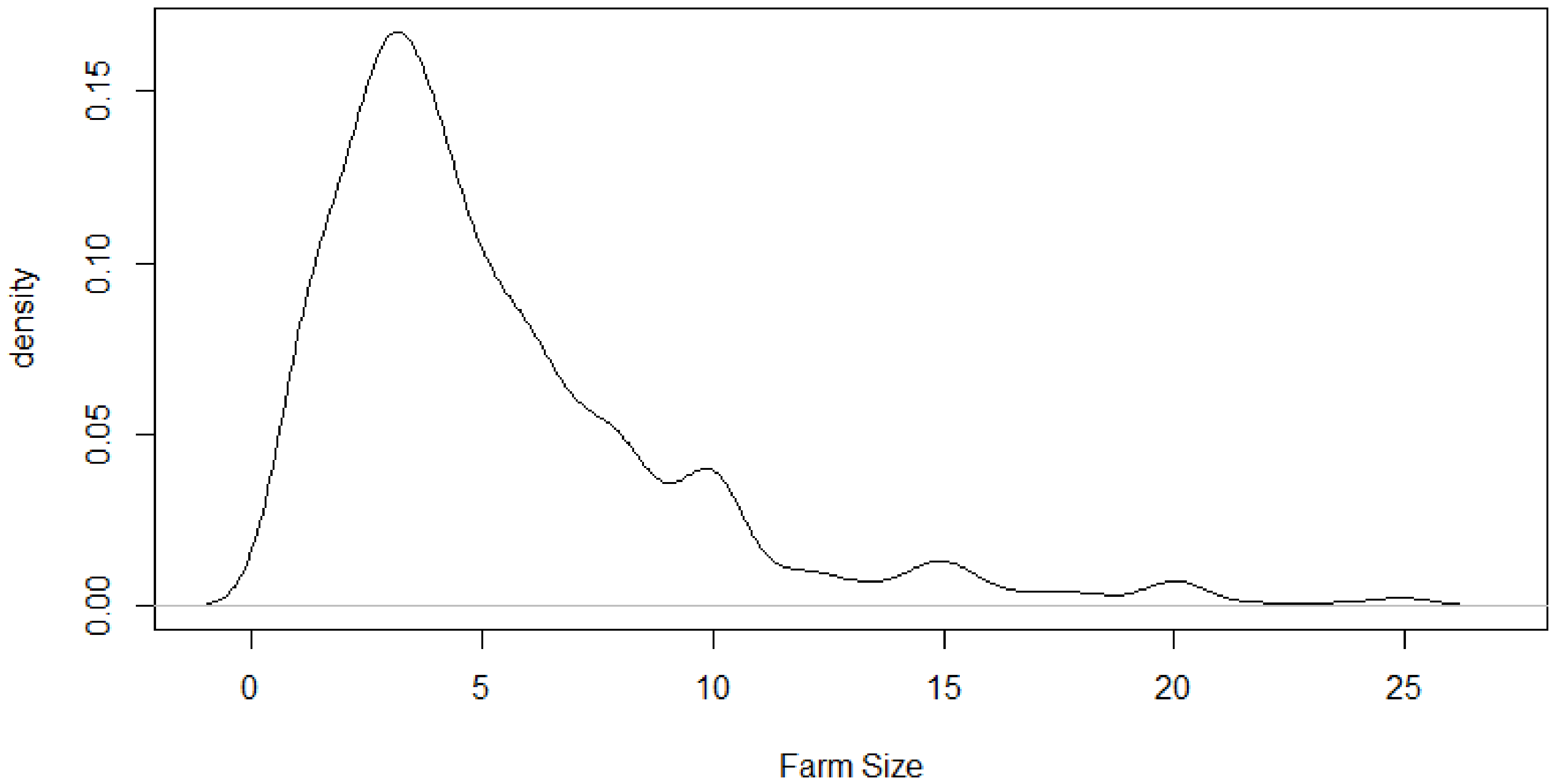
- The OLS line and the median regression line are very close, in spite of the asymmetry of Y … so why use quantile regression?

  - **It is impossible to get accurate estimates of the standard error for the OLS slope** without a significant increase in the complexity of the model (and would model interpretability be compromised?)
    - I.e., even though the β estimates are aligned (parametric vs. nonparametric), SEs for the OLS line are incorrect because the distribution of Y is not correctly specified (QR does not depend upon proper specification of Y, and so SEs are more robust)

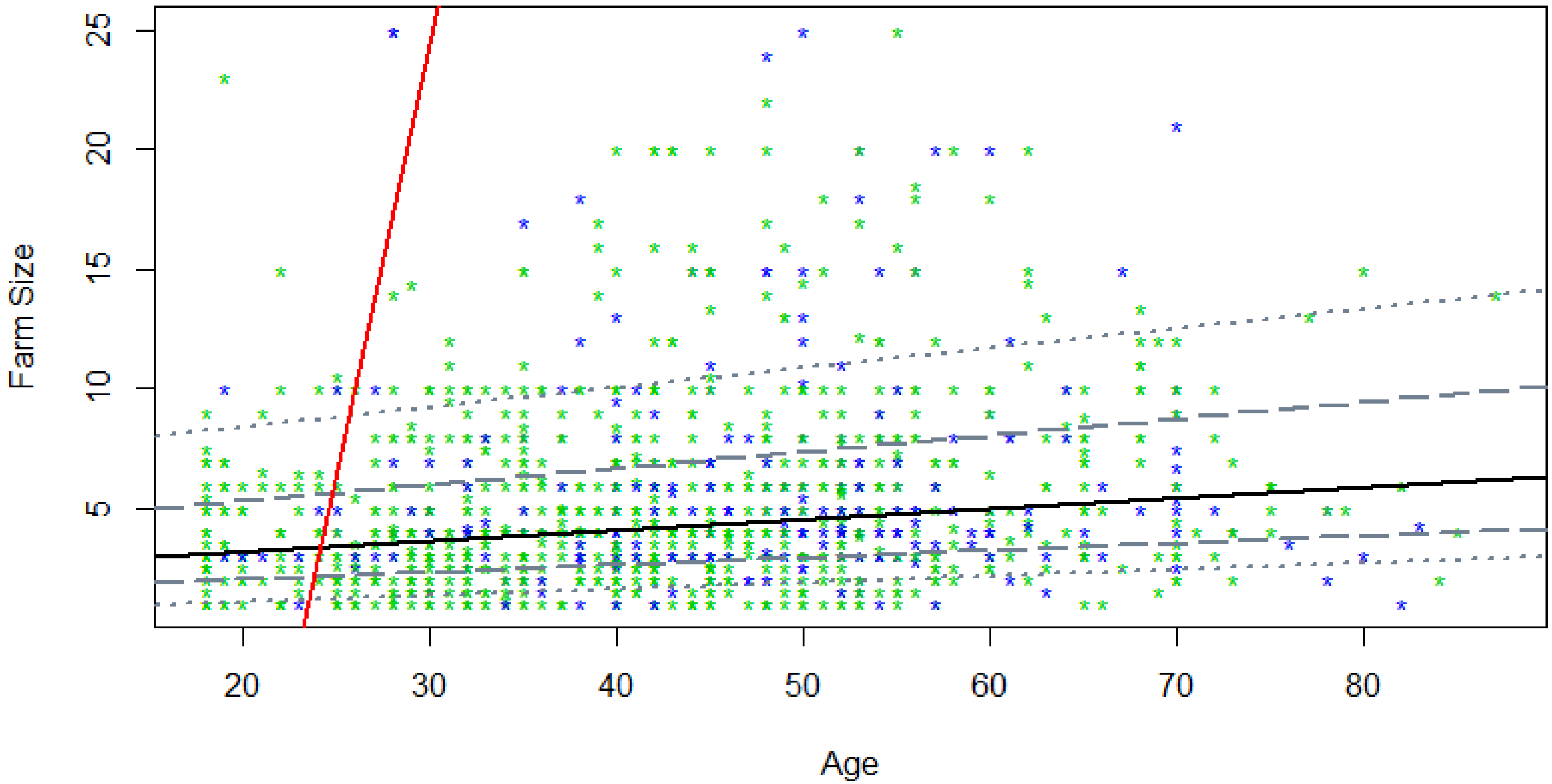
  - We can **reliably understand the spread of the response** with quantile regression because the distance between quantile lines is a measure of spread, e.g. interquartile range, θ differences, etc.

  - In this case, **flat prior means that BQR result is same as QR result**
    - Subsequent analyses show greater stability in upper/lower quantiles if alter informativeness

Actual Agricultural Example: Density of Y



Best-Fit Lines (OLS, BQR), Untrimmed Y



CONCLUSIONS AND NEXT STEPS

Concluding Thoughts

- QR is well-suited to a number of problems that arise in comparative survey research
  - BQR extensions have potential to allow flexible incorporation of survey information
- Additional simulations show that QR/BQR is reasonably robust to common model misspecifications
  - Error, X correlation, failure to include true region information
- In spite of this, full BQR is not as plug-and-play as, say, HLM
  - Conscientious implementation requires careful consideration of approximated likelihood, and reflection on prior (basic eBayes not as intuitive)

Contacts

Robert A. Petrin, Ph.D.
Vice President
Ipsos Public Affairs
Robert.Petrins@ipsos.com

Joseph Zappa, M.S.
Associate Statistician
Ipsos Public Affairs
Joseph.Zappa@ipsos.com

Meghana Raja
Assistant Statistician
Ipsos Public Affairs
Meghana.Raja@ipsos.com

CITATIONS AND REFERENCES

Citations and References

- Cai, T. 2013. "Investigation of Ways to Handle Sampling Weights for Multilevel Model Analyses" *Sociological Methodology* 43: 178-219.
- Carlin, B. P., and Louis, T. A. 2009. Bayesian Methods for Data Analysis. Third Edition. CRC Press.
- Congdon, P. 2005. Bayesian Models for Categorical Data. Wiley.
- Davino, C., Furno, M., and Vistocco, D. 2014. Quantile Regression: Theory and Applications. Wiley.
- Feng, Y., Chen, Y., and He, X. 2015. "Bayesian Quantile Regression with Approximate Likelihood" *Bernoulli* 21: 832-850.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. 2013. Bayesian Data Analysis. Chapman and Hall/CRC.
- Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. 2008. "A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models" *Annals of Applied Statistics* 4: 1360 – 1383.
- Gelman, A., and Hill, J. 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge.
- Gill, J. and Walker, L. D. 2005. "Elicited Priors for Bayesian Model Specifications in Political Scienc Research" *Journal of Politics* 67: 841-872.
- Koenker, R. 2005. Quantile Regression. Cambridge.
- Long, J. S. 1997. Regression Models for Categorical and Limited Dependent Variables. Sage.

Citations and References - II

- McCullagh, P., and Nelder, J. A. 1989. Generalized Liner Models. Second Edition. Chapman and Hall / CRC.
- McMillen, D. P. 2013. Quantile Regression for Spatial Data. Springer.
- Raudenbush, S. W., and Bryk, A. S. 2001. Hierarchical Linear Models. Sage.
- Rendell, M.S., Handcock, M. S., and Jonsson, S. H. 2009. “Bayesian Estimation of Hispanic Fertility Hazards from Survey and Population Data” *Demography*. 46: 65-83.
- Reich, B. J., Bondell, H. D., and Wang, H. 2010. “Flexible Bayesian Quantile Regression for Independent and Clustered Data” *Biostatistics* 11: 337-352.
- Si, Y., Pillai, N., and Gelman, A. 2015. “Bayesisan Nonparametric Weighted Sampling Inference” *Bayesian Analysis*. 1: 1-21.
- Yang, Y. and He, X. 2012. “Bayesian Empirical Likelihood for Quantile Regression” *Annals of Statistics* 40: 1102 – 1131.
- Yu, K., and Moyeed, R. A. 2001. “Bayesian Quantile Regression” *Statistics and Probability Letters*. 54: 437-447.

ABOUT IPSOS

Ipsos ranks third in the global research industry. With a strong presence in 87 countries, Ipsos employs more than 16,000 people and has the ability to conduct research programs in more than 100 countries. Founded in France in 1975, Ipsos is controlled and managed by research professionals. They have built a solid Group around a multi-specialist positioning – Media and advertising research; Marketing research; Client and employee relationship management; Opinion & social research; Mobile, Online, Offline data collection and delivery.

Ipsos is listed on Eurolist - NYSE-Euronext. The company is part of the SBF 120 and the Mid-60 index and is eligible for the Deferred Settlement Service (SRD).

ISIN code FR0000073298, Reuters ISOS.PA, Bloomberg IPS:FP
www.ipsos.com

GAME CHANGERS

At Ipsos we are passionately curious about people, markets, brands and society. We deliver information and analysis that makes our complex world easier and faster to navigate and inspires our clients to make smarter decisions.

We believe that our work is important. Security, simplicity, speed and substance applies to everything we do.

Through specialisation, we offer our clients a unique depth of knowledge and expertise. Learning from different experiences gives us perspective and inspires us to boldly call things into question, to be creative.

By nurturing a culture of collaboration and curiosity, we attract the highest calibre of people who have the ability and desire to influence and shape the future.

"GAME CHANGERS" - our tagline - summarises our ambition.