Appendix A: A Limited Information Quantile Aggregation Method drawing on Mosteller (1946)

Rachael Meager

November 11, 2020

1 Asymptotic Quantile Models

Consider the task of aggregating sets of quantile treatment effects and assessing their generalizability. First recall that the uth quantile of some outcome is the value of the inverse CDF at u:

$$Q_Y(u) = F_Y^{-1}(u). \tag{1.1}$$

Performing quantile regression for some quantile u in site k when the only regressor is the binary treatment indicator T_{nk} requires estimating:

$$Q_{y_{nk}|T}(u) = \beta_{0k}(u) + \beta_{1k}(u)T_{nk} \tag{1.2}$$

For a single quantile u, the treatment effect is the univariate parameter $\beta_{1k}(u)$. If there is only one quantile of interest, a univariate Bayesian hierarchical model can be applied, as in Reich et al (2011). But in the microcredit data, researchers estimated a set of 10 quantiles $\mathcal{U} = \{0.05, 0.15, ..., 0.95\}$ and interpolated the results to form a "quantile difference curve". This curve is constructed by computing the quantile regression at all points of interest:

$$Q_{y_{ik}|T} = \{Q_{y_{ik}|T}(u) = \beta_{0k}(u) + \beta_{1k}(u)T_{ik} \;\; \forall \; u \in \mathcal{U}\} \tag{1.3}$$

The results of this estimation are two $|\mathcal{U}|$-dimensional vectors containing intercept and slope parameters. For the microcredit data, I work with the following vector of

10 quantile effects:

$$\begin{aligned}\beta_{0k} &= (\beta_{0k}(0.05), \beta_{0k}(0.15), ...\beta_{0k}(0.95))\\ \beta_{1k} &= (\beta_{1k}(0.05), \beta_{1k}(0.15), ...\beta_{1k}(0.95))\end{aligned} \tag{1.4}$$

The quantile difference curve is the vector β_{1k}, often linearly interpolated. With a binary treatment variable, the parameters in a quantile regression are simple functions of unconditional outcome quantiles. Let $Q_{0k}(u)$ be the value of the control group's quantile u in site k, and let $Q_{1k}(u)$ be the value of the treatment group's quantile u in site k. Then:

$$\begin{aligned}Q_{0k} &= \{Q_{0k}(u) \;\; \forall \; u \in \mathcal{U}\}\\ Q_{1k} &= \{Q_{1k}(u) \;\; \forall \; u \in \mathcal{U}\}.\end{aligned} \tag{1.5}$$

Then the vectors of intercepts and slopes for the quantile regression curves can be reformulated as

$$\begin{aligned}\beta_{0k} &= Q_{0k}\\ \beta_{1k} &= Q_{1k} - Q_{0k}.\end{aligned} \tag{1.6}$$

Hence, while the quantile difference curve β_{1k} need not be monotonic, it must imply a monotonic Q_{1k} when combined with a monotonic β_{0k}. The fact that any inference done quantile-by-quantile may violate monotonicity of $(Q_1, Q_0, \{Q_{1k}, Q_{0k}\}_{k=1}^{K})$ is a well-understood problem (Chernozhukov et al. 2010). Partial pooling for aggregation can exacerbate this problem because even if every lower level Q_{1k} and Q_{0k} satisfies monotonicity, their "average" or general Q_1 and Q_0 may not do so. Thus, unlike quantile crossing within a sample, the crossing in this setting is not necessarily the result of an incorrect asymptotic assumption or an extrapolation to a poorly-covered region of the covariate space. Indeed, for binary treatment variables, the within-sample estimators always satisfy monotonicity, but the averaging and pooling of these estimators may introduce crossing where none existed.[1] Ideally, therefore, an aggregation model should fit all quantiles simultaneously, imposing the monotonicity constraint. Aggregating the quantile difference curves, $\{\beta_{1k}\}_{k=1}^{K}$, requires more structure than aggregating quantile-by-quantile, but permits the transmission of information across quantiles.

I propose a general methodology to aggregate reported information on quantile

[1] Yet even if quantile crossing does not arise, neighboring quantiles contain information about each other not just because of monotonicity but because smooth distributions have quantiles that tend to lie close to each other; using that information can improve the estimation and reduce posterior uncertainty.

difference functions building on the approach of Rubin (1981) and a classical result from Mosteller (1946) about the joint distribution of sets of empirical quantiles. Mosteller shows that if the underlying random variable is continuously distributed, then the asymptotic sampling distribution of a vector of its empirical quantiles is a multivariate Normal centered at the true quantiles and with a known variance-covariance structure. This implies that the difference of the empirical quantile vectors from two independent samples, $\beta_{1k} = (Q_{1k} - Q_{0k})$, is also asymptotically a multivariate Gaussian. The theorem offers a foundation for a hierarchical quantile treatment effect aggregation model using the knowledge that the sampling variation is approximately a multivariate Gaussian, and that as a result modelling the parent distribution as Gaussian will be both tractable and have attractive performance (Rubin 1981, Efron and Morris 1975). The resulting analysis requires only the limited information reported by each study (although it can be fit to the full data) and is applicable to any continuous distribution as long as there is sufficient data in each of the studies to make the asymptotic approximation reasonable.

For this model, the data are the vectors of sample quantile differences $\{\hat{\beta}_{1k}\}_{k=1}^{K}$ and their sampling variance-covariance matrices $\{\hat{\Xi}_{\beta_{1k}}\}_{k=1}^{K}$. Thus, the lower level $f(\mathcal{Y}_k|\theta_k) = f(\hat{\beta}_{1k}|\beta_{1k})$ is given by the expression:

$$\hat{\beta}_{1k} \sim N(\beta_{1k}, \hat{\Xi}_{\beta_{1k}}) \;\forall\; k \tag{1.7}$$

The upper level of the model $\psi(\theta_k|\theta)$ is therefore:

$$\beta_{1k} \sim N(\beta_1, \Sigma_1) \;\forall\; k. \tag{1.8}$$

However, the estimated $(\tilde{\beta}_1, \{\tilde{\beta}_{1k}\}_{k=1}^{K})$ from this likelihood may not respect the implied quantile ordering restriction when combined with the estimated control quantiles, even if $\hat{\beta}_{1k}$s do. We need to add the relevant constraints to this model, but these difference functions are not the primary objects on which the constraints operate. While $(\beta_1, \{\beta_{1k}\}_{k=1}^{K})$ need not be monotonic, they must imply monotonic $(Q_1, \{Q_{1k}\}_{k=1}^{K})$ when combined with $(Q_0, \{Q_{0k}\}_{k=1}^{K})$. Since the objects $(Q_1, Q_0, \{Q_{1k}, Q_{0k}\}_{k=1}^{K})$ define the constraints, they must appear in the model.

Once the quantiles $(Q_1, Q_0, \{Q_{1k}, Q_{0k}\}_{k=1}^{K})$ appear in the model, transforming them into monotonic vectors will fully impose the relevant constraint on $(\beta_1, \{\beta_{1k}\}_{k=1}^{K})$. This strategy exploits the fact that Bayesian inference treats unknown parameters as random variables, so applying the transformation of variables formula and then reversing the transform at the end of the procedure completely preserves the posterior probability mass, and hence correctly translates the uncertainty intervals.

The Bayesian approach here has an advantage in incorporating knowledge about the properties of quantiles and indeed on any arbitrary parameter θ, because it offers a natural mechanism for imposing constraints on parameters. If the parameter θ can only belong to some subset of the parameter space, $\mathcal{A}_\Theta \subset \Theta$, this produces the following restricted likelihood:

$$\mathcal{L}_{\mathcal{A}_\Theta}(\mathcal{Y}|\theta) = \mathcal{L}(\mathcal{Y}|\theta) \cdot \mathbb{1}\{\theta \in \mathcal{A}_\Theta\}. \tag{1.9}$$

Because Bayesian inference treats unknown parameters as random variables, a statistical transformation of variables can impose constraints throughout the entire inferential process. If θ is a multivariate random variable with PDF $p_\theta(\theta)$ then a new random variable $\theta^* = \mathrm{f}(\theta)$ for a differentiable one-to-one invertible function $\mathrm{f}(\cdot)$ with domain $\mathcal{A}_\theta$ has density

$$p(\theta^*) = p_\theta(\mathrm{f}^{-1}(\theta))|det(J_{\mathrm{f}^{-1}}(\theta))|. \tag{1.10}$$

Therefore to implement inference using $\mathcal{L}_{\mathcal{A}_\Theta}(\mathcal{Y}|\theta)$, leading to the correctly constrained posterior $f_{\mathcal{A}_\Theta}(\theta|\mathcal{Y})$, I specify the model as usual and then implement a transformation of variables from θ to θ^*. I then perform Bayesian inference using $\mathcal{L}(\mathcal{Y}|\theta^*)$ and $\mathcal{P}(\theta^*)$, derive $f(\theta^*|\mathcal{Y})$, and then reverse the transformation of variables to deliver $f(\theta|\mathcal{Y}) \cdot \mathbb{1}\{\theta \in \mathcal{A}_\Theta\}$.[2] Frequentist implementation of constraints typically must reckon with the constraints twice, first in point estimation and second in interval estimation, and it can be challenging to ensure coherence between the two or to extend the consequences to other parameters. The Bayesian implementation ensures coherence because the constraint is imposed on the parameter itself throughout the construction of the full joint posterior which is then used for both estimation and inference.

I proceed with a transform proposed for use in Stan (2016), but in theory any valid monotonizing transform will do, since it is always perfectly reversed.[3] Consider monotonizing the $|\mathcal{U}|$-dimensional vector β_0, with uth entry denoted $\beta_0[u]$. One can

[2] In fact, for all the transformations I use here, this procedure has been automatically implemented in the software package Stan, a free statistical library which calls C++ to fit Bayesian models from R or Python (Stan Development Team, 2017).

[3] While some transforms may perform better than others in certain cases, to my knowledge there is little research on this issue that presently permits us to choose between transforms.

map β_0 to a new vector β_0^* as follows:

$$\beta_0^*[u] = \begin{cases} \beta_0[u], & \text{if } u = 1 \\ log(\beta_0[u] - \beta_0[u-1]) & \text{if } 1 < u < |\mathcal{U}| \end{cases} \tag{1.11}$$

Any vector β_0 to which this transform is applied and for which inference is performed in the transformed space will always be monotonically increasing. For the rest of the paper, I denote parameters for which monotonicity has been enforced by performing inference on the transformed object as in equation 1.11 with a superscript m. Thus, by applying the transform, I work with β_0^m rather than an unconstrained β_0.

Employing a monotonizing transform is an appealing alternative to other methods used in the econometrics literature to ensure monotonicity during quantile regression. Restricting the Bayesian posterior to have support only on parameters which imply monotonic quantiles means that, for example, the posterior means are those values which are most supported by the data and prior information from the set which satisfy the constraint. Frequentist solutions such as rearrangement, smoothing or projection each prevent the violation of the constraint in one specific way chosen *a priori* according to the analyst's own preferences (He 1997, Chernozhukov et al. 2010). While each strategy performs well in terms of bringing the estimates closer to the estimand (as shown in Chernozhukov et al. 2010) the Bayesian transformation strategy can flexibly borrow from each of the strategies as and when the data supports their use. Imposing the constraint throughout the inference avoids the additional complications of choosing when during aggregation one should insert the constraint; for example, in the case of rearrangement, it would be hard to interpret the result of partially pooling information on the 25th quantile only to have some other quantile substituted in for certain studies ex-post.

Equipped with this monotonizing transform, it is now possible to build models with restricted multivariate Normal distributions which only produces monotonically increasing vectors. I propose the following model to perform aggregation in a hierarchical framework, taking in the sets of empirical quantiles $\{\hat{Q}_{1k}, \hat{Q}_{0k}\}_{k=1}^{K}$ and their sampling variance-covariance matrices $\{\hat{\Xi}_{1k}, \hat{\Xi}_{0k}\}_{k=1}^{K}$ as data. For this hierarchical quantile set model, the lower level $f(\mathcal{Y}_k|\theta_k)$ is:

$$\begin{aligned} \hat{Q}_{0k} &\sim N(\beta_{0k}^m, \hat{\Xi}_{0k}) \ \ \forall \ k \\ \hat{Q}_{1k} &\sim N(Q_{1k}^m, \hat{\Xi}_{1k}) \ \ \forall \ k \\ \text{where } \ Q_{1k} &\equiv \beta_{0k}^m + \beta_{1k} \end{aligned} \tag{1.12}$$

The upper level $\psi(\theta_k|\theta)$ is:

$$\begin{aligned} \beta_{0k}^m &\sim N(\beta_0^m, \Sigma_0) \;\; \forall \, k \\ \beta_{1k} &\sim N(\beta_1, \Sigma_1) \;\; \forall \, k \\ \text{where } \;\; \beta_1 &\equiv Q_1^m - \beta_0^m \end{aligned} \tag{1.13}$$

The priors $\mathcal{P}(\theta)$ are:

$$\begin{aligned} \beta_0^m &\sim N(0, 1000 * I_{10}) \\ \beta_1 &\sim N(0, 1000 * I_{10}) \\ \Sigma_0 &\equiv diag(\nu_0)\Omega_0 diag(\nu_0)' \\ \Sigma_1 &\equiv diag(\nu_1)\Omega_1 diag(\nu_1)' \\ \text{where } \;\; \nu_0, \nu_1 &\sim \text{halfCauchy}(0, 20) \text{ and } \Omega_0, \Omega_1 \sim LKJCorr(1). \end{aligned} \tag{1.14}$$

This formulation is convenient as the form of $\hat{\Xi}_{1k}$ is exactly derived in the Mosteller (1946) theorem, though the individual entries need to be estimated. The structure could be modified to take in the empirical quantile treatment effects $\{\hat{\beta}_{1k}\}_{k=1}^K$ and their standard errors instead of $\{\hat{Q}_{1k}\}$ if needed. The model imposes no structure on (Σ, Σ_0), other than the logical requirement of positive semi-definiteness. This complete flexibility is made possible by the discretization of the quantile functions; these matrices could not take unconstrained form if the quantile functions had been modelled as draws from Gaussian Processes.[4] Overall, this structure passes information across the quantiles in two ways: first, by imposing the ordering constraint, and second, via the functional form of $\hat{\Sigma}_k$ from the Mosteller (1946) theorem.

The above model implements partial pooling not only on the $\{\beta_{1k}\}_{k=1}^K$ parameters but also on the $\{\beta_{0k}\}_{k=1}^K$ parameters, that is, the control group quantiles. The technical reason for this is that one needs to define a notion of a general β_0 in order to define the constraint on the general β_1 and the predicted $\beta_{1,K+1}$. However, this structure also provides us with useful insight that allows us to better interpret the results of the partial pooling on $\{\beta_{1k}\}_{k=1}^K$. Suppose for example that we observe substantial pooling on $\{\beta_{1k}\}_{k=1}^K$, but we also observe this on $\{\beta_{0k}\}_{k=1}^K$; in that case, we observe similarities in the treatment effects perhaps only because we have studied

[4] Gaussian Processes in general are too flexible to fit at the upper level of these models for this application, and popular covariance kernels tend to have identification issues that limit their usefulness in the current setting. In particular, most tractable and popular kernels do not permit the separation of dispersion of points within the functional draws from dispersion of points across the functional draws.

places with similar control groups. In that case it will be hard to justify extrapolation to another setting with a substantially different value of β_{0k}. On the other hand, suppose that we observe substantial pooling on $\{\beta_{1k}\}_{k=1}^{K}$ but no pooling at all on $\{\beta_{0k}\}_{k=1}^{K}$. Then we have learned much more generalisable information, because we now know that the treatment effects can be similar even when the underlying control distributions are different.

1.1 Model Performance

To assess the performance of the model, I provide Monte Carlo simulations under a variety of data scenarios and report the coverage of the posterior intervals. Ideally, the 50% posterior interval should contain the true parameter 50% of the time, and the analogous property should hold for the 95% posterior interval. When the data is simulated from the model itself this property is guaranteed in Bayesian inference, however, in practice one typically does not have the luxury of fitting data that one knows originates from a particular model. Therefore, all the monte carlo simulations I provide here fit the model above to data generated from a somewhat different model. In particular, I always simulate data for which my priors are incorrect; as the priors are reasonably diffuse they should not compromise the inference, and nor do they.

The results in table 1 show that the model typically provides approximately nominal coverage on β_1, and often provides greater than nominal coverage, regardless of the generating process. However, inference on the β_0 and the covariation parameters Σ_0, Σ_1 is more sensitive to underlying conditions. When there is large variation in the data within sites, the model can have difficulty with achieving nominal coverage on the 50% interval for these parameters, although the 95% interval usually retains its coverage properties. The fact that the model has trouble when the data exhibits large variation reflects one of the conditions of the Mosteller (1946) theorem, namely that the underlying density that generates the data is not vanishing in the neighbourhood of the quantile. While this condition is formally satisfied in the simulations, the model seems to be affected regardless: the poor average performance in these cases is generated by difficulty characterising the extremal quantiles where the density is thinnest.

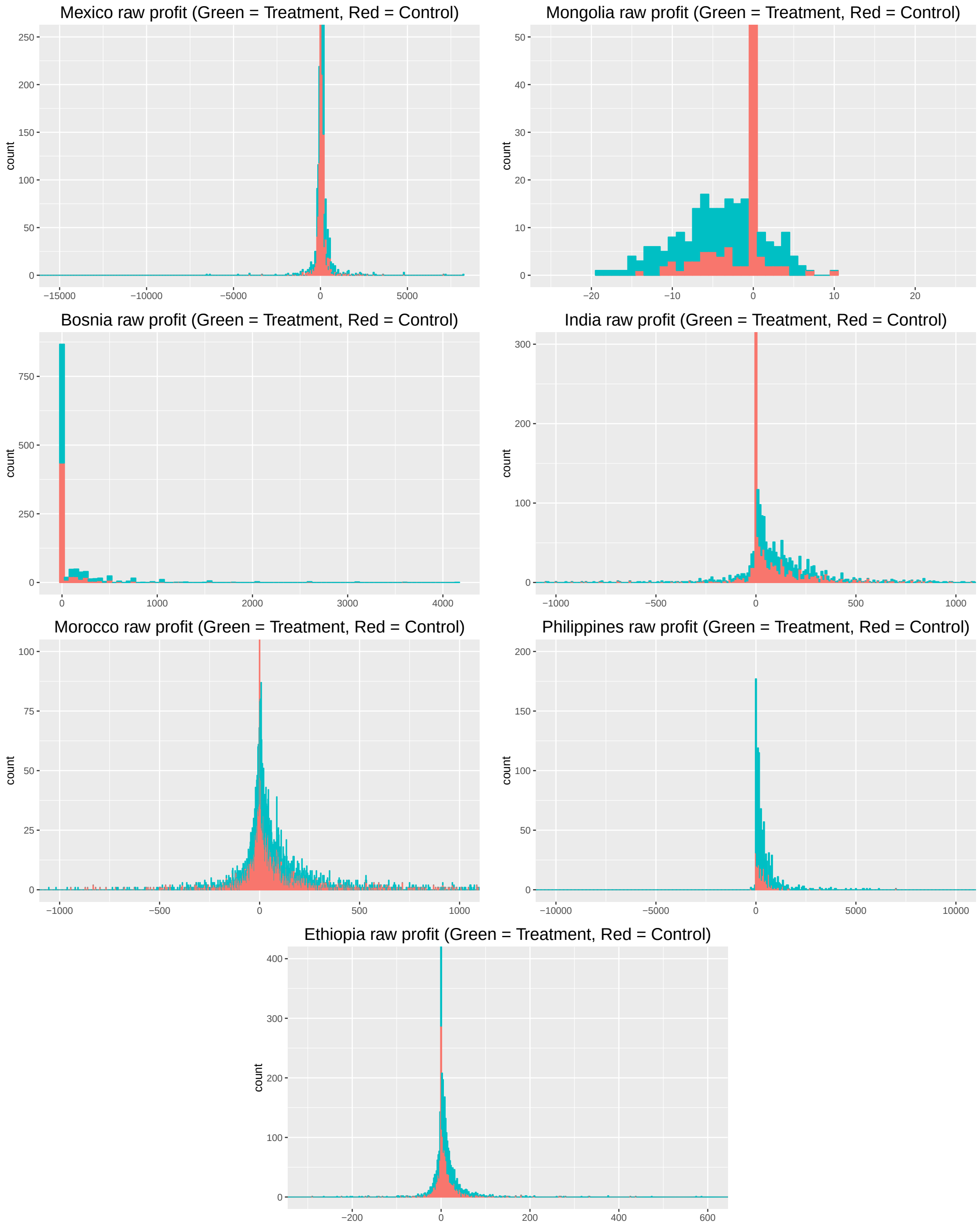
Encouragingly, when the data variation is moderate or small, the model does reasonably well on most parameters even when the full pooling or no pooling cases approximately hold. The no pooling case provides some trouble for inference on Σ_0, Σ_1 due to the extreme cross-site variation. Large within-site data variation seems to cause difficulties for the 50% intervals in the full pooling inference, but the 95% intervals retain their coverage properties even in this case. There are some results in the table that do not fit the broad patterns laid out here, but this may be due to the relatively small number of MC runs (due to the relatively long time it takes to run the model). The results point to overall good performance, although suggesting that caution should be applied when approaching data sets that have high variance or heavy tails even when the theoretical conditions for asymptotic normality are formally satisfied.

1.2 Limitations

The strength of the model based on the Mosteller (1946) theorem is that it works for any continuous outcome variable; its weakness is that it *only* works for continuous variables. In the microcredit data, this approach will work for household consumption, consumer durables spending and temptation goods spending. But household business profit, revenues and expenditures are not continuous because many households either did not own or did not operate their businesses in the month prior to being surveyed and therefore recorded zero for these outcomes. This creates large "spikes" at zero in the distributions, as shown in the histograms of the profit data for the sites (figure 1)). This spike undermines the performance of the Mosteller theorem and of the nonparametric bootstrap for standard error calculation. The Mexico data provides the cleanest example of this, shown in figure 2: the first panel is the result of using the Mosteller asymptotic approximation, and the second panel is the result of the nonparametric bootstrap applied to the standard errors on the same data. The former produces the dubious result that the uncertainty on the quantiles in the discrete spike is the same as the uncertainty in the tail; the latter produces the dubious result that the standard errors are exactly zero at most quantiles.

The potential for quantile regression techniques to fail when the underlying data is not continuous is a well-understood problem (Koenker and Hallock 2001; Koenker 2011). In some cases, "dithering" or "jittering" the data by adding a small amount of random noise is sufficient to prevent this failure and reliably recover the underlying parameters (Machado and Santos Silva, 2005). However this does not work for the microcredit setting: the reason for this is that the jittering method is intended to be used for count data, in which there are gaps between the integer values which can be filled in by the jitter while still maintaining the crucial one-to-one relationship between the quantiles of the count data and the quantiles of the new continuous data produced by the jitter. But in the business variables, the discrete spike at zero is accompanied by a continuous tail that has support right up until zero itself – even a small jitter applied to the spike at zero causes some of the "zeroes" to leapfrog some of the continuous data points, destroying the one-to-one relationship required for the jittering to be theoretically sound.[5]

[5] Author's correspondence with Machado and Santos Silva via email confirms this point. Correspondence available from the author on request.



Mexico raw profit (Green = Treatment, Red = Control)

Mongolia raw profit (Green = Treatment, Red = Control)

Bosnia raw profit (Green = Treatment, Red = Control)

India raw profit (Green = Treatment, Red = Control)

Morocco raw profit (Green = Treatment, Red = Control)

Philippines raw profit (Green = Treatment, Red = Control)

Ethiopia raw profit (Green = Treatment, Red = Control)

Figure 1: Histograms of the profit data in each site, in USD PPP per 2 weeks. Display truncated both vertically and horizontally in most cases. [Back to main]

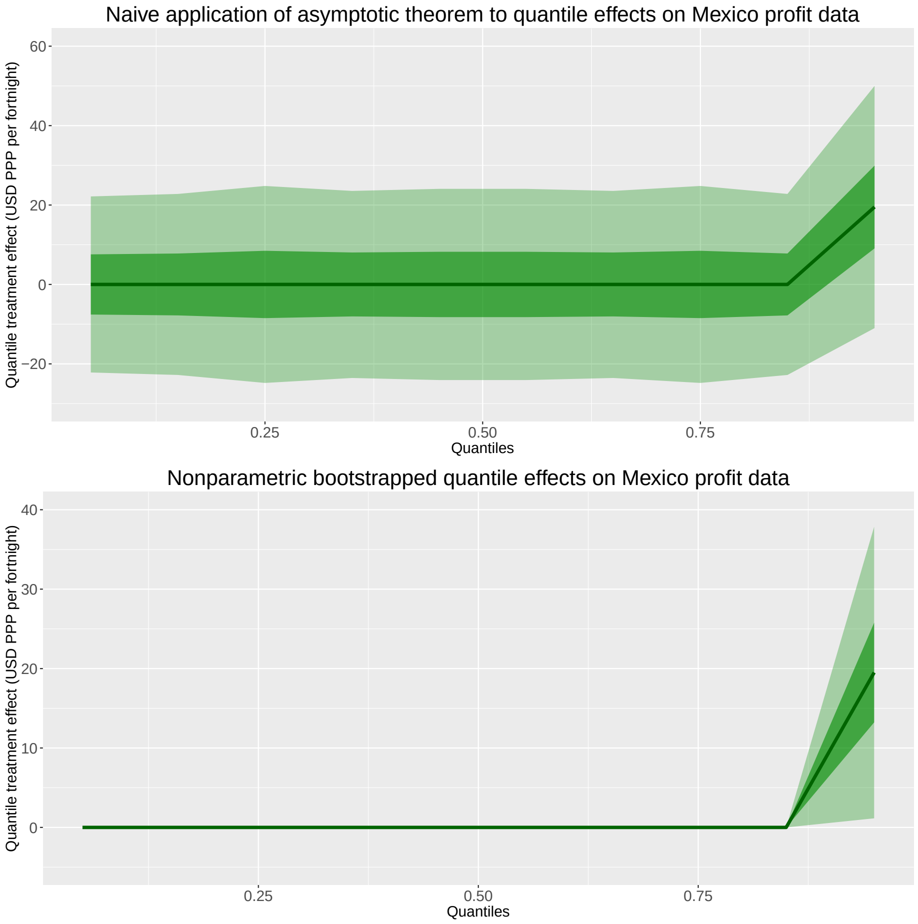
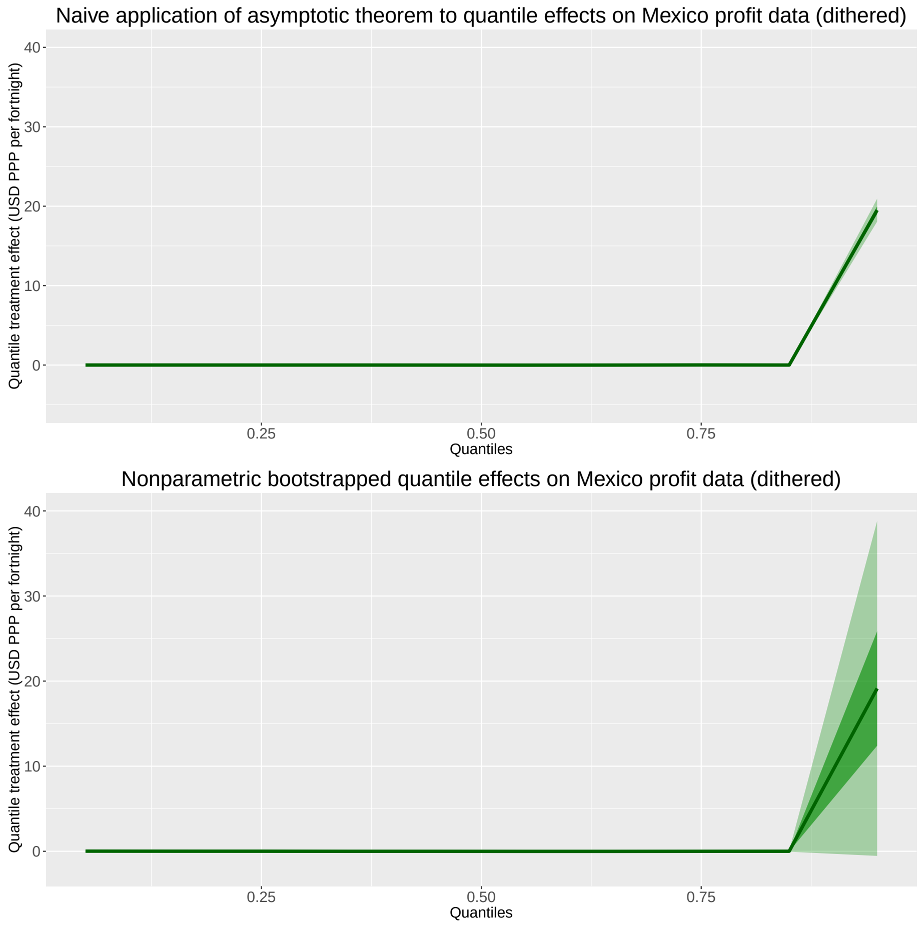


Figure 2: Quantile TEs for the Mexico profit data, Mosteller theorem approximation standard errors (above) and nonparametrically bootstrapped standard errors (below). The green line is the estimated effect, the opaque bands are the central 50% interval, the translucent bands are the central 95% interval. The output of these estimators should be similar if the Mosteller (1946) theorem holds, but it is not similar because profit is not in fact continuously distributed. This is due to a discrete probability mass at zero, reflecting numerous households who do not operate businesses. [Back to main]

Dithering is often an effective strategy for partially discrete data: In fact, a small amount of dithering is necessary for the microcredit data on consumer durables spending and temptation goods spending to conform to the Mosteller approximation, as this data is actually somewhat discrete. However, in the microcredit business data, the complications caused by these spikes at zero are not effectively addressed by dithering. The results in figure 3 show that applying the Mosteller theorem to the dithered profit data leads to inference that is too precise in the tail relative to the results of the bootstrap on the same data.


Naive application of asymptotic theorem to quantile effects on Mexico profit data (dithered)

Nonparametric bootstrapped quantile effects on Mexico profit data (dithered)

Figure 3: Quantile TEs for the dithered Mexico profit data, Mosteller theorem approximation standard errors (above) and nonparametrically bootstrapped standard errors (below). The green line is the estimated effect, the opaque bands are the central 50% interval, the translucent bands are the central 95% interval. Dithering is a simple strategy which can overcome problems associated with quantile regression on discrete distributions, recommended in Machado & Santos Silva (2005) and Koenker (2011). It has failed in this case. [Back to main]

Table 1: Simulation Results: Coverage of Limited-Information Quantile Model Posterior Inference under Cross-Site Variation (CSV) and Data Variation (DV)

Features	β_1: 50%	95%	β_0: 50%	95%
Little CSV, Little DV	0.524	0.976	0.560	0.952
Little CSV, Moderate DV	0.644	0.980	0.480	0.936
Large CSV, Moderate DV	0.532	0.956	0.448	0.928
Large CSV, Large DV	0.568	0.992	0.512	0.944
Moderate CSV, Large DV	0.632	0.988	0.480	0.940
Little CSV, Large DV	0.596	0.996	0.548	0.984
Moderate CSV, Little DV	0.512	0.952	0.488	0.912
Very Large CSV, Large DV	0.476	0.928	0.266	0.688
Approx No Pooling, Moderate DV	0.480	0.944	0.434	0.870
Approx Full Pooling, Moderate DV	0.668	0.988	0.668	0.996
Approx Full Pooling, Very Large DV	0.758	0.998	0.576	0.972
	Σ_1 (off diag): 50%	95%	Σ_1 (diag) :50%	95% (diag)
Little CSV, Little DV	0.903	1	0.880	0.996
Little CSV, Moderate DV	0.932	1	0.884	0.996
Large CSV, Moderate DV	0.845	1	0.616	0.996
Large CSV, Large DV	0.933	1	0.868	1
Moderate CSV, Large DV	0.887	1	0.810	1
Little CSV, Large DV	0.927	1	0.820	1
Moderate CSV, Little DV	0.807	1	0.520	0.992
Very Large CSV, Large DV	0.336	1	0.850	0.998
Approx No Pooling, Moderate DV	0.593	0.998	0.454	0.926
Approx Full Pooling, Moderate DV	1	1	0.420	0.996
Approx Full Pooling, Very Large DV	1	1	0.026	0.998
	Σ_0 (off diag): 50%	95%	Σ_0 (diag): 50%	95%
Little CSV, Little DV	0.857	1	0.476	0.996
Little CSV, Moderate DV	0.878	1	0.444	1
Large CSV, Moderate DV	0.255	1	0.304	0.984
Large CSV, Large DV	0.083	0.999	0.196	0.992
Moderate CSV, Large DV	0.101	1	0.198	0.994
Little CSV, Large DV	0.675	1	0.228	1
Moderate CSV, Little DV	0.800	1	0.420	0.980
Very Large CSV, Large DV	0.497	0.994	0.426	0.944
Approx No Pooling, Moderate DV	0.560	0.995	0.362	0.866
Approx Full Pooling, Moderate DV	1	1	0.684	0.996
Approx Full Pooling, Very Large DV	1	1	0.052	1

Notes: CSV is Cross-Site Variation, DV is Data Variation. Simulation runs kept small due to relatively long runtime for model fit, but results are relatively stable within run sets. For this exercise, the CSV in β_{0k} space is typically larger than that in β_{1k} space, to reflect the likely reality of comparing quite different places with plausibly similar effects.
[Back to main]

1.3 Interpretation

The goal of the hierarchical model is to estimate the central location and the dispersion in the distributions from which each of the observed $\{\beta_{0k}, \beta_{1k}\}$ are drawn. This permits analysts and policymakers to formulate expectations about what may be likely in future settings. The expectations are taken over the distribution of the vectors $\{\beta_{0k}, \beta_{1k}\}$, not over households; this point is crucial to avoid confusion about how to interpret quantiles in a hierarchical setting. The estimate of β_0 is an estimate of the expected marginal distributions of the control groups' outcomes in the set of sites exchangeable with the K sites at hand, subject to the constraint that these objects all be monotonic. The estimate of β_1 is an estimate of the expected differences between the marginal outcomes of the treatment and control groups in the set of exchangeable sites, subject to the known properties of each of the groups distributions. The monotonicity constraint complicates the interpretation relative to unconstrained Gaussian models - since, for example, the mean of a sum of ordered Gaussians may not necessarily obey the property of the mean of a sum of Gaussians. Yet the broad intuition is that the parameters β_0, β_1 provide an estimate of the centrality of the individual distributions over the vector spaces in which $\{\beta_{0k}, \beta_{1k}\}$ live.

Quantile effects are often subject to misinterpretation. Quantiles do not satisfy laws of iterated expectations, so the treatment effects at the quantiles of the outcome distribution (which is what is delivered here) are not the quantiles of the distribution of treatment effects.[6] Another source of confusion is that while unconditional quantiles of a distribution correspond to specific data points in the sample, these data points (and the individuals who produce them) are not meaningful in themselves. It does not make sense to think of quantile estimation as applying to specific households nor "tracking" them across time or place. As stressed in Koenker 2005, the fact that one can locate the specific data point that sits at a particular sample quantile does not mean that this datum is deeply related to the population quantile in some way: it happens to be the best estimate of that population quantile, nothing more. The population quantile is a *parameter*, not a "representative household"; the sample household whose outcome value is "selected" as a quantile estimate is not playing the role of an individual in the world but rather the role of a sample order statistic of similar general stature to a sample mean. If the sample had realised differently, a different household would have been selected as the estimate, but the

[6] While it would be nice to know the latter object, this is not estimable without considerable additional structure

population quantile remains the same.[7]

Koenker's point extends fully to the case of aggregation of quantiles. The hierarchical nature of the model does not imply that the information or outcome of any specific individual household is being passed (or not passed) up to the "general" level of the model. Rather, the model posits the potential existence of a general or meta-distribution which governs the K observed distributions and observed differences between the treatment and control distributions in each site. We then study the means and covariance matrices of these parent distributions in an attempt to understand how useful that structure is for prediction. In this context, the relative positions of e.g. the 25th quantile in site 1 and the 25th quantile in site 2, and how similar they are to the expected value of the 25th quantile taken across all the sites, is simply a question: how similar are the value of the 25th quantiles are in all the sites we have studied? If the answer is "not very similar" this is not a problem for interpretation of the quantiles, but simply a possible state of reality we have fully anticipated in the model and which will be expressed by a very large covariance matrix (or at least a large entry on the diagonal for that quantile).

In fact, the expected value of the quantile treatment effect in the hierarchical context does not find solutions corresponding to single households, but rather, takes weighted averages of the K solutions. Suppose for example that all consumption values in one site lie below all those in another site. This does not mean that the lower quantiles of the general distribution are all taken from the first site, nor that the upper quantiles of the general distribution are all taken from the second site. Instead, the procedure examines each site's data set quantile by quantile and asks "What is the average estimate of this quantile, and how similar are these quantiles across sites?". The expected value of a given quantile taken over all sites for example may not correspond to any household's value or any specific quantile effect in any site, any more than the expectation of a set of values would correspond to any one of the values: it wouldn't, except by chance. The expected median is not necessarily the median of the K medians and nor does it need to be such in order to be interpreted: the expectation is formulated over a posited distribution of medians, which corresponds to the question: "What should I expect the median to be in these kinds of places?"

Hence, β_0 and β_1 should not be interpreted as the quantiles or differences of

[7] Another common misunderstanding is that "only" the "chosen" household contributes to the inference at any given quantile. In fact, in quantile estimation, as in mean estimation, all the household data points together determine the best estimate of a point (or set of points) corresponding to a population object of interest.

some aggregated data set; rather, they are the *expected* quantiles and differences in any given site with this expectation taken across sites, subject to the monotonicity constraint in this case. The monotonicity constraint does make the situation more complex because the averages conditional on this constraint will tend to lie below the raw averages in each site, since when site effects are "drawn" from the distribution governed by this average, they will more often lie above it than below it. This complication aside, the hierarchical model does not attempt to arrange all the individual data points or quantile difference estimates in some kind of grand order (nor would it be clear how to interpret such an exercise). Quantile regression permits one to infer the shapes of distributions, not to track individuals specifically over time or over ranks of relative groups one could decide to place them in. The goal of the hierarchical quantile model is to infer a set of true differences that correspond to a population distribution's response to a treatment, and to understand how different these responses are across settings.

2 Results for Consumption

Table 2: Consumption: Comparison Of No Pooling, Partial Pooling and Full Pooling Results

Quantile:	5th	15th	25th	35th	45th	55th	65th	75th	85th	95th
No Pooling										
Bosnia	-5.2	-7.1	-4.7	-7.7	4.1	0.9	-16.3	-34.4	-64.5	104
	(-11.1,0.8)	(-16.9,2.7)	(-17.5,8.1)	(-22.8,7.3)	(-13.1,21.4)	(-20,21.7)	(-46.2,13.6)	(-74.9,6.1)	(-131.1,2.2)	(-77.4,285.5)
India	0.2	-1	-2.3	-1.2	-1.4	-2.6	2.2	4.6	8.2	40.1
	(-5.9,6.3)	(-6.3,4.3)	(-8.3,3.8)	(-7.4,4.9)	(-8.3,5.6)	(-10.1,4.8)	(-6.3,10.7)	(-6.3,15.6)	(-7.5,24)	(-4.5,84.7)
Mexico	-9.3	-1.5	-2	-2	-0.5	4.1	5.5	11	13.2	16.6
	(-13.7,-4.9)	(-5.9,2.9)	(-6.2,2.2)	(-6.5,2.5)	(-6,5)	(-1.6,9.7)	(0,11)	(2.7,19.2)	(1.8,24.7)	(-6.7,39.9)
Mongolia	12.3	6.5	5.1	8	-8.2	0.8	-0.9	-2.5	-12.8	87.4
	(-7,31.5)	(-9.4,22.4)	(-13.6,23.7)	(-13.7,29.8)	(-38.7,22.2)	(-18.5,20.1)	(-32.7,30.9)	(-42.1,37.1)	(-70.3,44.8)	(-40.6,215.4)
Morocco	1.6	5.3	4.1	-1.1	-2.6	3.6	3.7	0.4	-6.4	-54
	(-6.2,9.3)	(-0.9,11.5)	(-2.8,10.9)	(-7.3,5.1)	(-10.1,4.9)	(-4.3,11.5)	(-7.3,14.7)	(-12.4,13.2)	(-23.8,11)	(-104,-4)
Partial Pooling										
Bosnia	-1	7.7	-0.3	5.7	-0.3	5.3	-1	6.2	0.7	5.2
	(-4,2)	(1.7,13.9)	(-5.8,5.9)	(-6,15.5)	(-5.6,5.1)	(-8.8,15.8)	(-4.7,2.5)	(-0.9,13.1)	(-3.2,5.9)	(-3.2,12.4)
India	-1.5	8.3	1.6	3	-3	-1.3	-0.3	7.4	-2.2	3.1
	(-5.1,2.3)	(-0.7,17.9)	(-4.6,11.4)	(-22.9,22.1)	(-10.4,2.6)	(-37.8,17.7)	(-4.2,3.8)	(-3.4,18.8)	(-6.5,1.9)	(-10.9,14.8)
Mexico	-7.3	3.2	2.4	3	-3.5	4.4	-0.3	0.2	-1.1	3.4
	(-12,-2.5)	(-0.9,7.3)	(-7.4,15.9)	(-4,10.5)	(-9,0.8)	(-2.8,14.4)	(-5.5,5.6)	(-5.2,5.1)	(-6.6,5)	(-1.7,9)
Mongolia	-0.3	3.9	-0.1	4.2	-1.5	3.6	-0.3	4.2	1.7	4.6
	(-3.5,2.9)	(-0.7,8.7)	(-7,6.6)	(-3.9,12.2)	(-7.8,3.2)	(-5.6,11.1)	(-3.7,3.2)	(-0.8,9.4)	(-2.2,6.6)	(-0.8,10.4)
Morocco	-0.7	11.4	-5	76.6	4.7	89.9	0.2	36.7	-2.8	-31.9
	(-4.8,3.5)	(-10.2,33.3)	(-18.9,4.5)	(7.7,152.6)	(-2.8,13.8)	(-6.3,215.9)	(-4.6,4.9)	(-1.4,76.9)	(-7.9,1.9)	(-70.9,7.3)
Average	-2	0	-0.4	-1.1	-0.8	2.8	4.1	6	4.1	19.4
	(-9.7,7.2)	(-4.6,4.5)	(-4.4,4.1)	(-6.6,4.8)	(-9.1,6.8)	(-3,9.1)	(-1.8,9.9)	(-2.3,13.5)	(-14.2,16.8)	(-26.1,62.4)
Full Pooling										
Average	-3.9	0.2	-0.9	-1.8	-1.3	2.5	3.6	6.1	6.4	13.9
	(-6.8,-0.9)	(-2.4,2.9)	(-3.7,1.9)	(-5,1.4)	(-4.8,2.2)	(-1.4,6.3)	(-0.8,7.9)	(0.2,11.9)	(-1.8,14.6)	(-6.1,33.9)

Notes: All units are USD PPP per two weeks. Estimates are shown with their 95% uncertainty intervals below them in brackets. In this case the full pooling and no pooling models are frequentist, estimated using the quantreg package in R, per Koencker and Basset 1978 with the nonparametric bootstrap providing the standard errors.[Back to main]