Online Appendix for Manuscript AER AER-2018-1811

Rachael Meager

December 1, 2021

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:

$$\hat{\beta}_{1k} \sim N(\beta_{1k}, \hat{\Xi}_{\beta_1 k}) \;\forall\; k \tag{1.7}$$

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:

$$\begin{aligned} \beta^m_{0k} &\sim N(\beta^m_0, \Sigma_0) \;\; \forall \, k \\ \beta_{1k} &\sim N(\beta_1, \Sigma_1) \;\; \forall \, k \\ \text{where} \;\; \beta_1 &\equiv Q^m_1 - \beta^m_0 \end{aligned} \tag{1.13}$$

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

$$\begin{aligned} \beta^m_0 &\sim N(0, 1000 * I_{10}) \\ \beta_1 &\sim N(0, 1000 * I_{10}) \\ \Sigma_0 &\equiv diag(\nu_0)\Omega_0 diag(\nu_0)' \\ \Sigma_1 &\equiv diag(\nu_1)\Omega_1 diag(\nu_1)' \\ \text{where} \;\; \nu_0, \nu_1 &\sim \text{halfCauchy}(0, 20) \text{ and } \Omega_0, \Omega_1 \sim LKJCorr(1). \end{aligned} \tag{1.14}$$

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 $\{\hat{\beta}_{1k}\}^K_{k=1}$ 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 $\{\beta_{1k}\}^K_{k=1}$ parameters but also on the $\{\beta_{0k}\}^K_{k=1}$ 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 $\{\beta_{1k}\}^K_{k=1}$. Suppose for example that we observe substantial pooling on $\{\beta_{1k}\}^K_{k=1}$, but we also observe this on $\{\beta_{0k}\}^K_{k=1}$; 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.

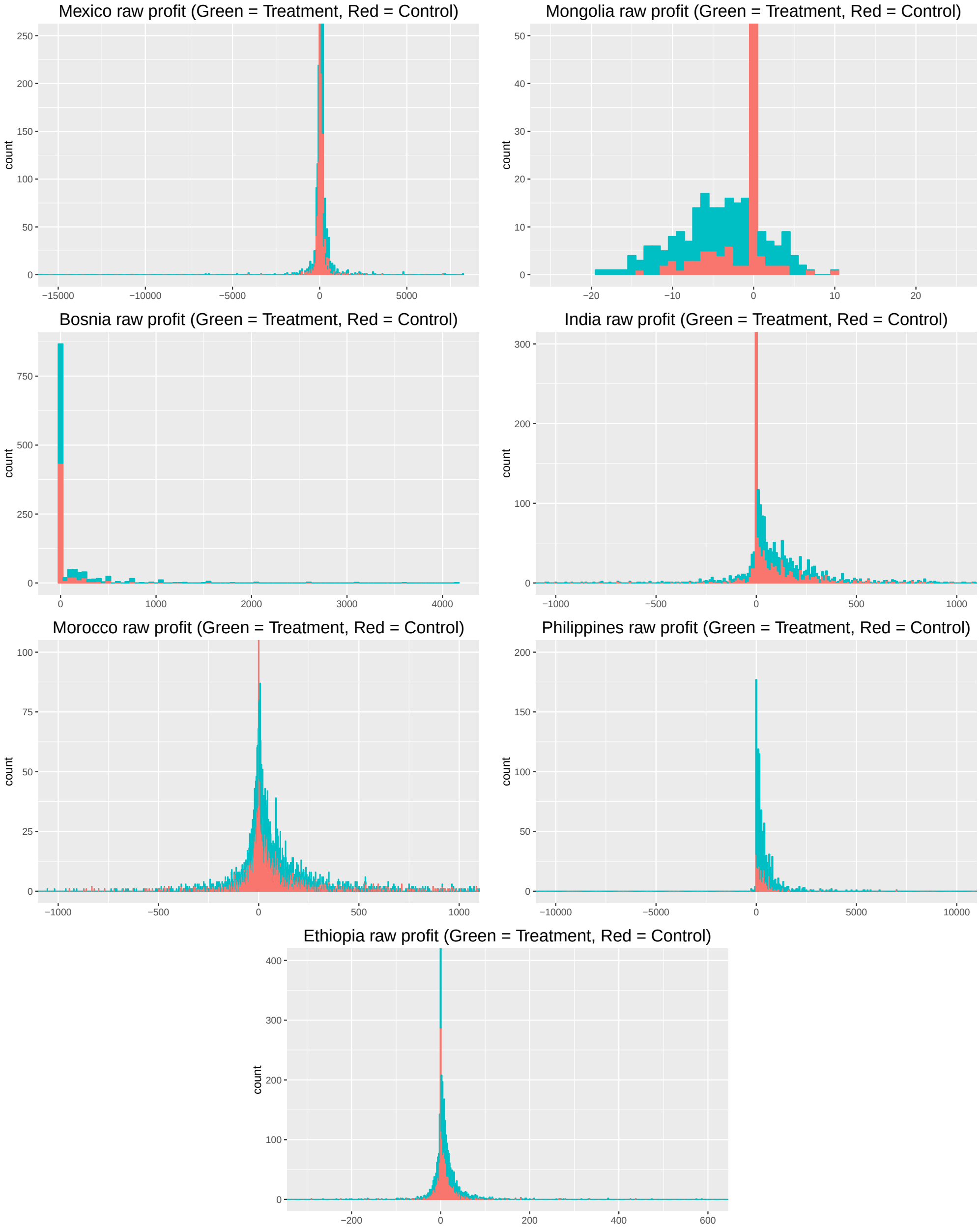
Encouragingly, when the data variation is moderate or small, the model does reasonably well on most parameters even when the full pooling or no pooling cases approximately hold. The no pooling case provides some trouble for inference on Σ_0, Σ_1 due to the extreme cross-site variation. Large within-site 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)

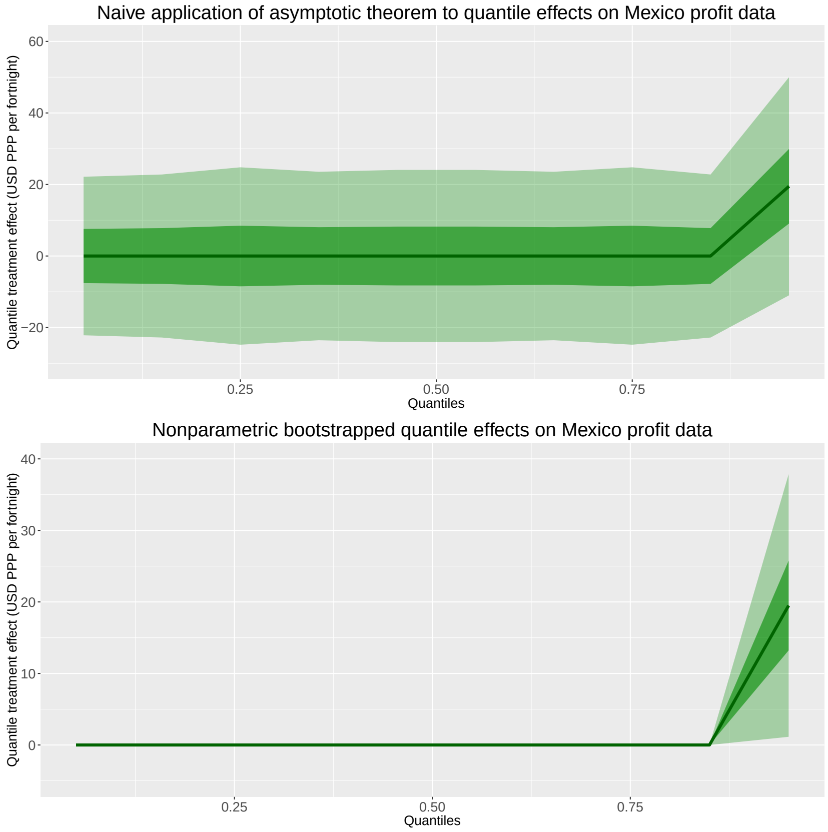
India raw profit (Green = Treatment, Red = Control)

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]



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

Nonparametric bootstrapped quantile effects on Mexico profit data

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.

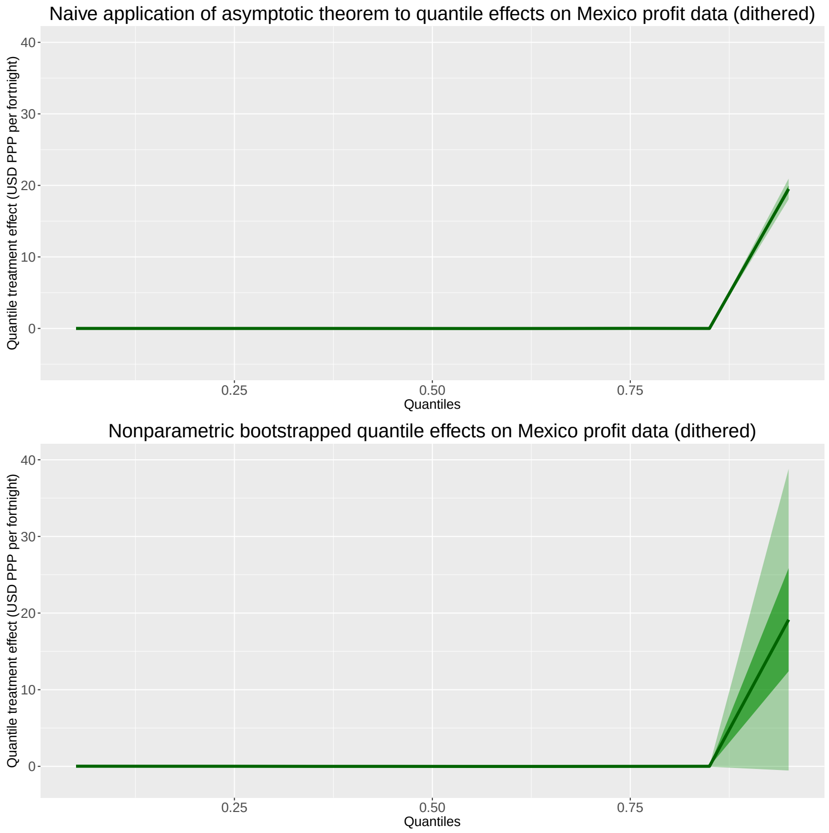


Figure 3: Quantile TEs for the dithered Mexico profit data, Mosteller theorem approximation standard errors (above) and nonparametrically bootstrapped standard errors (below). The green line is the estimated effect, the opaque bands are the central 50% interval, the translucent bands are the central 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]

Appendix B: Nonlinearity of Shrinkage Operations

Rachael Meager

November 11, 2020

1 The Set-up

Consider K parallel experiments each containing N participants randomized 50/50 into "treatment" and "control" groups (indicated by a binary variable $T_{ik} = 1$ if individual i in trial k is treated, else 0). Within a single experiment, the average treatment effect (ATE) is the difference of the means in the treatment and control groups, which is also the mean of the difference between the two groups due to the linearity of the expectation operator. Denote the ATE in experiment k by τ_k, estimated using the sample counterpart $\bar{\tau}_k = E[Y_{1ki} - Y_{0ki}] = E[Y_{1ki}] - [Y_{0ki}] = \bar{y_{1k}} - \bar{y_{0k}}$. This result relies only on the linearity of the expectations operator.

However, once the analyst places a hierarchical model on the data and jointly analyses all K experiments together, the updated expectations of these objects cease to obey this linear relationship. This is because shrinkage on any of these objects is a nonlinear operation in the unknown parameters, with particular issue caused by the unknown hypervariances. This appendix contains an illustration and proof of the problem, and concludes that analysts need to choose the object of interest and shrink directly on that object.

Consider the following Gaussian hierarchical model in which all the random parameters are independent to simplify the exposition.

$$\begin{aligned} y_{0k} &\sim N(y_0, \sigma_0^2) \\ \tau_k &\sim N(\tau, \sigma_\tau^2) \\ y_{1k} &\equiv y_{0k} + \tau_k \\ \therefore y_{1k} &\sim N(y_0 + \tau, \sigma_0^2 + \sigma_\tau^2) \\ \bar{y_{0k}} &\sim N(y_{0k}, \hat{se}_{y0}^2) \\ \bar{\tau_k} &\sim N(\tau_k, \hat{se}_{\tau k}^2) \\ \therefore \bar{y_{1k}} &\sim N(y_{0k} + \tau_k, \hat{se}_{0k}^2 + \hat{se}_{\tau k}^2) \end{aligned}$$

2 The Nonlinearity Result

This model generates new estimates of the parameters in each site k updated given the information in the other sites – that is, the model performs shrinkage.[1] Per Gelman et al (2004), if one knew the hyperparameters that govern $\{\tau_k\}_{k=1}^K$, i.e. (τ, σ_τ^2), one could manually compute the shrinkage on the observed $\bar{\tau_k}$ and thus the new posterior ATE $\tilde{\tau_k}$ for a given site k as follows:

$$\tilde{\tau_k} = \frac{\frac{1}{\hat{se}_{\tau k}^2}\bar{\tau_k} + \frac{1}{\sigma_\tau^2}\tau}{\frac{1}{\hat{se}_{\tau k}^2} + \frac{1}{\sigma_\tau^2}}.$$

Analogous objects exist for the y_{0k} and the y_{1k} if shrinkage is performed on them:

$$\tilde{y_{0k}} = \frac{\frac{1}{\hat{se}_{y_{0k}}^2}\bar{y_{0k}} + \frac{1}{\sigma_{y_0}^2}y_0}{\frac{1}{\hat{se}_{y_{0k}}^2} + \frac{1}{\sigma_{y_0}^2}}, \quad \tilde{y_{1k}} = \frac{\frac{1}{\hat{se}_{y_{1k}}^2}\bar{y_{1k}} + \frac{1}{\sigma_{y_1}^2}y_1}{\frac{1}{\hat{se}_{y_{1k}}^2} + \frac{1}{\sigma_{y_1}^2}}.$$

However, despite the fact that $\bar{\tau_k} = \bar{y_{1k}} - \bar{y_{0k}}$ and that $\tau = y_1 - y_0$, the following result holds:

Theorem 2.1. *Given the model above, $\tilde{\tau_k} \neq \tilde{y_{1k}} - \tilde{y_{0k}}$ and hence shrinkage is not a linear operation.*

[1] Notice that at most two of the triple (y_{0k}, y_{1k}, τ_k) can be independently distributed because the third object is constructed from the other two, and this third object will always be dependent on the components from which it is constructed. Here I have chosen y_{0k} and τ_k, which amounts to taking linear regression seriously as a model.

Proof. To see this, we begin with $\tilde{y_{1k}}$ and substitute in its components:

$$\begin{aligned}
\tilde{y_{1k}} &= \frac{\frac{1}{\hat{se}^2_{y_{1k}}}(\bar{y_{0k}} + \bar{\tau_k}) + \frac{1}{\sigma^2_{y_1}}(y_0 + \tau)}{\frac{1}{\hat{se}^2_{y_{1k}}} + \frac{1}{\sigma^2_{y_1}}} \\
&= \frac{\frac{1}{\hat{se}^2_{y_{0k}} + \hat{se}^2_{\tau k}}(\bar{y_{0k}} + \bar{\tau_k}) + \frac{1}{\sigma^2_{y_0} + \sigma^2_{\tau}}(y_0 + \tau)}{\frac{1}{\hat{se}^2_{y_{0k}} + \hat{se}^2_{\tau k}} + \frac{1}{\sigma^2_{y_0} + \sigma^2_{\tau}}}
\end{aligned}$$

Focusing only on the term that contains τ, we can see that it is being given the wrong weight in this expression:

$$\frac{\frac{1}{\sigma^2_{y_0} + \sigma^2_{\tau}}\tau}{\frac{1}{\hat{se}^2_{y_{0k}} + \hat{se}^2_{\tau k}} + \frac{1}{\sigma^2_{y_0} + \sigma^2_{\tau}}} \neq \frac{\frac{1}{\sigma^2_{\tau}}\tau}{\frac{1}{\hat{se}^2_{\tau k}} + \frac{1}{\sigma^2_{\tau}}}$$

Since there is no other instance of τ in the formula for $\tilde{y_{1k}}$, this issue cannot be rectified by the presence of some other term. □

Notice that the only time the shrinkage will be "correct" on τ_k is when both the sampling error and the cross-site heterogeneity in y_{0k} is exactly zero. This turns out to be the key to understanding why the two kinds of shrinkage do not coincide: it is because shrinking the composite object y_{1k} allows the noise from the control mean to corrupt the shrinkage on the treatment effects (and vice versa).

To make this clear, consider an example. Suppose that $y_0 = 50$, $\tau = 30$ and thus $y_1 = 80$. Suppose that $\sigma_{y0} = 50$ and $\sigma_\tau = 1$, so that $\sigma_{y1} = \sqrt{50^2 + 1^2} = 50.01$. Then suppose that $\bar{y_{0k}} = 10$, $\hat{se}_{y_{0k}} = 10$, $\bar{\tau_k} = 20$, $\hat{se}_{\tau k} = 8$, and that therefore $\bar{y_{1k}} = 30$ and $\hat{se}_{y_{1k}} = \sqrt{10^2 + 8^2} = 12.80625$. Applying the formulae above, we get:

$$\begin{aligned}
\tilde{y_{0k}} &= 11.53846 \\
\tilde{y_{1k}} &= 32.38005 \\
\tilde{\tau_k} &= 29.61538 \neq \tilde{y_{1k}} - \tilde{y_{0k}} = 21.53846.
\end{aligned}$$

In this case the discrepancy is about 30% of the underlying parameter's true magnitude. This occurs it is because the control means y_{0k} vary greatly across the K sites, but the treatment effects τ_k vary little across the K sites. Faced with the task of shrinking on the composite object y_{1k}, the treatment group's mean, the model compromises between the two patterns – but the huge variance in y_{0k} across the K sites dominates the weight, so we still see essentially zero shrinkage on y_{1k}. However, when we direct the hierarchical model to shrink directly on τ_k, we isolate it from the noise on y_{0k} and the model can therefore detect that the τ_k component of y_{1k} is very similar across the K sites and should be shrunk strongly towards the common mean of 30.

Appendix C: Tabular Results and Robustness Checks

Rachael Meager

February 4, 2021

1 Tabular results

Table 1: Lender and Study Attributes by Country

Country	Bosnia & Herzegovina	Ethiopia	India	Mexico	Mongolia	Morocco	The Philippines
Study Citation	Augsburg et al (2015)	Tarozzi et al (2015)	Banerjee et al (2015b)	Angelucci et al (2015)	Attanasio et al (2015)	Crepon et al (2015)	Karlan and Zinman (2011)
Treatment	Lend to marginally rejected borrowers	Open branches	Open branches	Open branches, promote loans	Open branches, target likely borrowers	Open branches	Lend to marginal applicants
Randomization Level	Individual	Community	Community	Community	Community	Community	Individual
Urban or Rural?	Both	Rural	Urban	Both	Rural	Rural	Urban
Target Women?	No	No	Yes	Yes	Yes	No	No
MFI already operates locally?	Yes	No	No	No	No	No	Yes
Microloan Liability Type	Individual	Group	Group	Group	Both	Group	Individual
Collateralized?	Yes	Yes	No	No	Yes	No	No
Any other MFIs competing?	Yes	No	Yes	Yes	Yes	No	Yes
Household Panel?	Yes	No	No	Partial	Yes	Yes	No
Interest Rate (Intended on Average)	22% APR	12% APR	24% APR	100% APR	24% APR	13.5% APR	63% APR
Sampling Frame	Marginal Applicants	Random Sample	Households with at least 1 woman age 18-55 of stable residence	Women ages 18-60 who own businesses or wish to start them	Women who registered interest in loans and met eligibility criteria	Random Sample plus Likely Borrowers	Marginal Applicants
Study Duration	14 months	36 months	40 months	16 months	19 months	24 months	36 months

Note: The construction of the interest rates here is different to the construction of Banerjee et al (2015a); they have taken the maximal interest rate, whereas I have taken the average of the intended range specified by the MFI. In practice the differences in these constructions are numerically small. This table was also printed in Meager (2018) which used the same studies.

Excess Kurtosis in LogNormal distributions is the extent to which tail indices are greater, and thus the extent to which the tails are heavier, than those of the Gaussian. For a LogNormal parameterised as

$$LogNormal(y|\mu,\omega) = \frac{1}{\sqrt{2\pi}\omega y}\exp\left(\frac{-(\log(y)-\mu)^2}{2\omega^2}\right)$$

the excess kurtosis is

$$\exp(4\omega^2) + 2\exp(3\omega^2) + 3\exp(2\omega^2) - 6.$$

I compute the kurtosis for the general control group based on the posterior mean values of μ and σ for this group in the tables below. The μ parameters are the same in the model and in the formula above, but the ω parameter requires some explanation. ω in the formula is the Lognormal scale. The σ parameters in the models are the log versions of this parameter. For example, the posterior mean scale parameter for the LogNormal in the "generalized" control group's positive tail is actually $\omega = \exp(\sigma_2^c)$, and this must be squared further to enter the formula above.The examples in the text are obtained using σ_2^c from profit and from consumption respectively, to give excess kurtosis values of 810.5 and 13.9 respectively (there will be a small rounding error if plugging in the values from the table here).

Table 2: All General-Level Posterior Marginals for the LogNormal Profit Model

	mean	MCMC error	sd	2.5%	25%	50%	75%	97.5%	# effective draws	$\hat{R}$
μ_1	3.200	0.008	0.732	1.722	2.784	3.200	3.615	4.698	$9,099$	1.000
μ_2	3.843	0.007	0.818	2.225	3.356	3.845	4.324	5.496	$15,000$	1.000
τ_1	0.094	0.001	0.094	-0.099	0.045	0.095	0.143	0.273	$6,719.600$	1.001
τ_2	0.077	0.0005	0.042	-0.007	0.054	0.078	0.102	0.157	$7,566.232$	1.000
σ_{μ_1}	1.659	0.008	0.654	0.867	1.227	1.514	1.923	3.302	$7,284.792$	1.000
σ_{μ_2}	2.033	0.006	0.677	1.153	1.574	1.889	2.332	3.711	$15,000$	1.000
σ_{τ_1}	0.117	0.004	0.128	0.005	0.035	0.079	0.154	0.459	$1,090.338$	1.003
σ_{τ_2}	0.055	0.001	0.052	0.002	0.020	0.043	0.075	0.183	$1,323.050$	1.004
σ_1^c	0.452	0.002	0.145	0.180	0.374	0.447	0.525	0.761	$7,205.404$	1.000
σ_2^c	0.225	0.001	0.101	0.022	0.167	0.225	0.284	0.428	$10,278.910$	1.000
σ_1^t	0.022	0.001	0.094	-0.162	-0.024	0.022	0.067	0.206	$6,128.028$	1.001
σ_2^t	0.017	0.0003	0.029	-0.043	0.001	0.017	0.032	0.072	$9,321.264$	1.000
$\sigma_{\sigma_1^c}$	0.302	0.002	0.164	0.122	0.196	0.262	0.357	0.724	$5,126.273$	1.001
$\sigma_{\sigma_2^c}$	0.242	0.001	0.100	0.125	0.176	0.220	0.280	0.499	$9,328.806$	1.000
$\sigma_{\sigma_1^t}$	0.163	0.002	0.116	0.034	0.089	0.134	0.201	0.467	$2,860.338$	1.001
$\sigma_{\sigma_2^t}$	0.046	0.001	0.037	0.002	0.020	0.038	0.062	0.140	$2,034.778$	1.002
β_{11}	-1.965	0.016	1.273	-4.525	-2.715	-1.958	-1.193	0.527	$6,334.358$	1.000
β_{12}	0.025	0.001	0.114	-0.187	-0.035	0.019	0.080	0.265	$6,957.068$	1.001
β_{21}	0.390	0.010	0.906	-1.379	-0.168	0.367	0.918	2.255	$7,964.995$	1.000
β_{22}	-0.067	0.001	0.104	-0.279	-0.124	-0.066	-0.012	0.143	$8,309.348$	1.001
$\sigma_{\beta_{11}}$	2.767	0.017	1.277	0.770	1.959	2.560	3.346	5.904	$5,636.316$	1.000
$\sigma_{\beta_{12}}$	0.128	0.002	0.125	0.005	0.047	0.096	0.168	0.446	$5,901.720$	1.001
$\sigma_{\beta_{21}}$	1.603	0.014	0.902	0.130	0.990	1.532	2.093	3.672	$3,987.814$	1.002
$\sigma_{\beta_{22}}$	0.146	0.002	0.114	0.007	0.065	0.124	0.197	0.432	$5,234.755$	1.001
$\sigma_{\beta_{31}}$	1.450	0.014	0.889	0.091	0.815	1.381	1.942	3.493	$3,896.658$	1.001
$\sigma_{\beta_{32}}$	0.117	0.002	0.109	0.004	0.041	0.089	0.161	0.390	$5,085.964$	1.002

Note: The β_3 parameters are normalized to be zero at the general level as required for multinomial logit models. The site-specific effects still have variation around this zero anchor as reported.

Table 3: All General-Level Posterior Marginals for the LogNormal Revenues Model

	mean	MCMC error	sd	2.5%	25%	50%	75%	97.5%	# effective draws	$\hat{R}$
μ_1	4.472	0.007	0.873	2.733	3.959	4.479	4.992	6.193	15,000	1.000
τ_1	0.083	0.001	0.068	-0.058	0.045	0.086	0.123	0.211	10,482.840	1.000
σ_{μ_1}	2.181	0.007	0.718	1.258	1.693	2.030	2.496	3.982	10,285.460	1.000
$\sigma_{\tau_1}]$	0.140	0.001	0.080	0.039	0.089	0.124	0.171	0.329	5,189.630	1.001
σ_1^c	0.213	0.001	0.136	-0.063	0.134	0.214	0.292	0.485	11,190.950	1.000
σ_1^t	-0.010	0.0003	0.031	-0.071	-0.028	-0.011	0.008	0.052	9,554.774	1.000
$\sigma_{\sigma_1^c}$	0.331	0.001	0.135	0.171	0.241	0.301	0.383	0.668	8,452.406	1.001
$\sigma_{\sigma_1^t}$	0.062	0.0004	0.033	0.020	0.040	0.055	0.075	0.146	6,447.524	1.000
β_{11}	0.011	0.008	0.734	-1.464	-0.424	-0.004	0.443	1.521	8,107.184	1.001
β_{12}	-0.063	0.001	0.081	-0.235	-0.101	-0.058	-0.020	0.091	6,772.048	1.001
$\sigma_{\beta_{11}}$	1.209	0.010	0.760	0.064	0.637	1.164	1.645	2.912	5,305.339	1.001
$\sigma_{\beta_{12}}$	0.095	0.001	0.091	0.003	0.032	0.071	0.129	0.327	5,418.020	1.001
$\sigma_{\beta_{21}}$	1.192	0.010	0.762	0.062	0.615	1.147	1.631	2.894	5,341.343	1.001
$\sigma_{\beta_{22}}$	0.095	0.001	0.091	0.003	0.033	0.071	0.130	0.328	5,944.329	1.000

Note: The β_3 parameters are normalized to be zero at the general level as required for multinomial logit models. The site-specific effects still have variation around this zero anchor as reported. Note also that σ_1^t can be negative as this is the effect specified on the exponential level.

Table 4: All General-Level Posterior Marginals for the LogNormal Expenditures Model

	mean	MCMC error	sd	2.5%	25%	50%	75%	97.5%	# effective draws	$\hat{R}$
μ_1	4.042	0.006	0.733	2.563	3.593	4.047	4.483	5.528	15,000	1.000
τ_1	0.103	0.001	0.048	0.005	0.076	0.104	0.132	0.198	8,840.624	1.000
σ_{μ_1}	1.867	0.005	0.624	1.061	1.449	1.735	2.135	3.449	15,000	1.001
σ_{τ_1}	0.078	0.001	0.060	0.004	0.035	0.067	0.106	0.226	1,919.668	1.002
σ_1^c	0.303	0.002	0.171	-0.037	0.204	0.304	0.401	0.649	8,974.738	1.001
σ_1^t	-0.008	0.001	0.045	-0.092	-0.033	-0.009	0.016	0.082	5,069.866	1.000
$\sigma_{\sigma_1^c}$	0.421	0.002	0.171	0.218	0.309	0.382	0.489	0.845	8,374.404	1.001
$\sigma_{\sigma_1^t}$	0.094	0.001	0.051	0.035	0.062	0.082	0.111	0.217	3,164.881	1.001
β_{11}	0.234	0.009	0.694	-1.177	-0.180	0.233	0.653	1.645	6,027.909	1.000
β_{12}	-0.116	0.001	0.117	-0.349	-0.177	-0.114	-0.053	0.112	7,262.210	1.000
$\sigma_{\beta_{11}}$	1.148	0.011	0.712	0.062	0.613	1.102	1.565	2.729	4,414.652	1.001
$\sigma_{\beta_{12}}$	0.157	0.002	0.125	0.007	0.071	0.132	0.209	0.465	5,601.528	1.000
$\sigma_{\beta_{21}}$	1.119	0.011	0.707	0.056	0.580	1.075	1.535	2.714	4,076.193	1.001
$\sigma_{\beta_{22}}$	0.159	0.002	0.124	0.007	0.074	0.136	0.212	0.463	5,427.373	1.001

Note: The β_3 parameters are normalized to be zero at the general level as required for multinomial logit models. The site-specific effects still have variation around this zero anchor as reported.Note also that σ_1^t can be negative as this is the effect specified on the exponential level.

For visual ease, the figures below graph the treatment effects and posterior predicted effects for each of the dimensions of change permitted in the model.

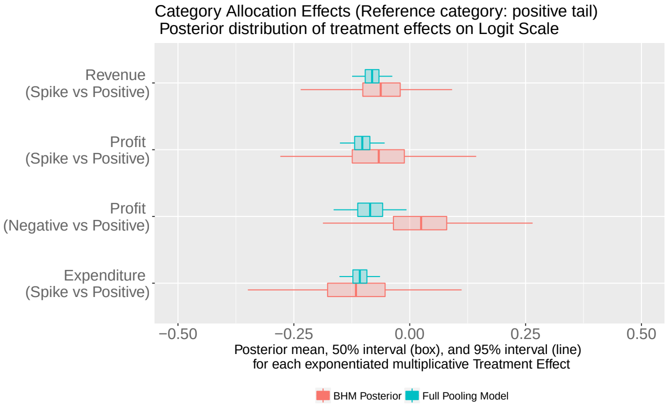


Figure 1: Posterior distributions for the logit treatment effects (π_j) on category assignment. These treatment effects are specified as an exponentiated multiplicative factor on the control group proportion of households in the category: if $\tilde{\pi}_j = 0$ the effect is zero, if $\tilde{\pi}_j < 0$ the treatment increases the proportion of households in the positive tail relative to other categories.

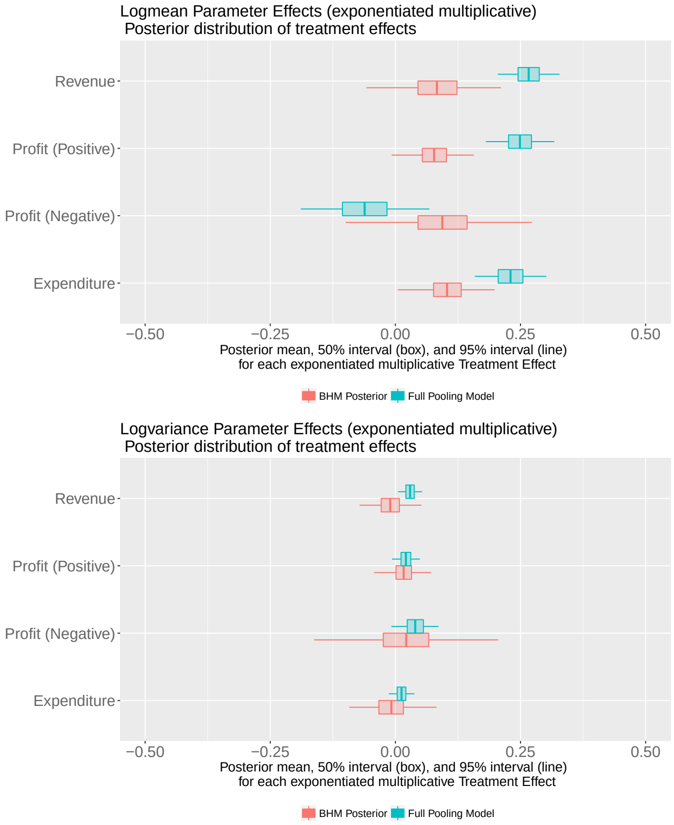
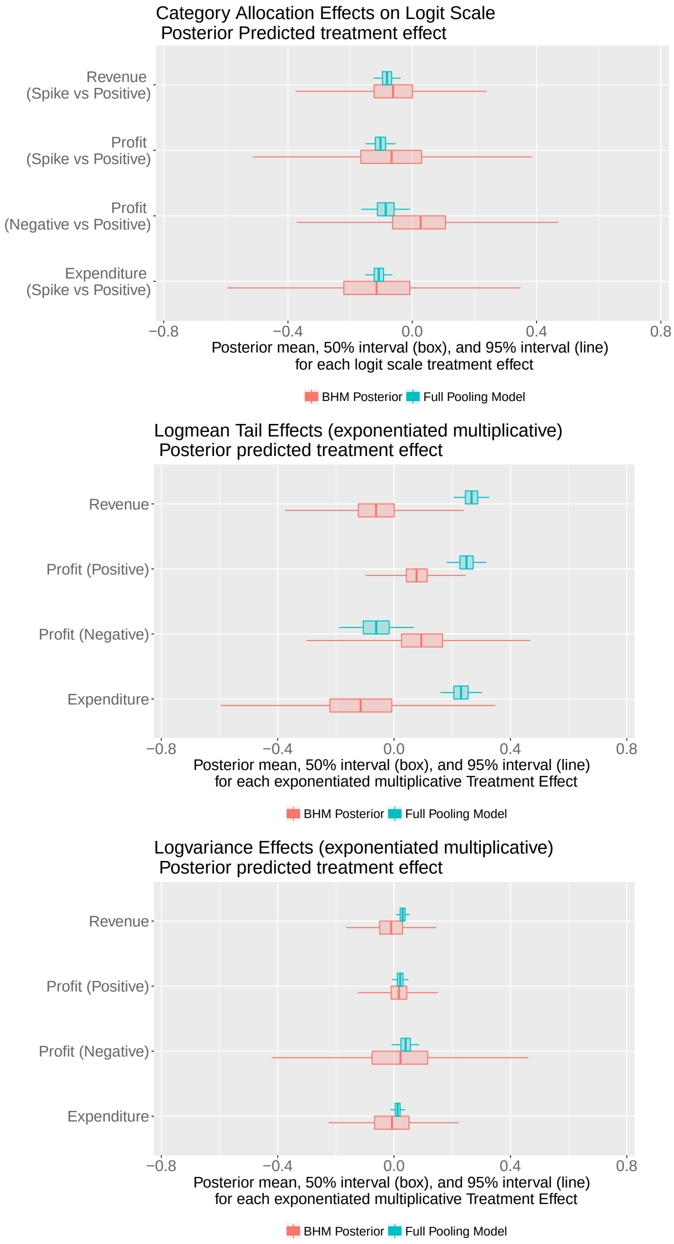


Figure 2: Posterior distributions for the location treatment effects (τ_j) and the scale treatment effects (σ_j^t).



Category Allocation Effects on Logit Scale
Posterior Predicted treatment effect

Logmean Tail Effects (exponentiated multiplicative)
Posterior predicted treatment effect

Logvariance Effects (exponentiated multiplicative)
Posterior predicted treatment effect

Figure 3: Posterior predicted distributions for the logit treatment effects on category assignment and tail shape effects.

2 Robustness Checks

2.1 Pareto Tail Models

If using the Pareto distribution for the continuous component, the tails are governed by a location parameter which controls the lower bound of the support and a scale parameter which controls the thickness of the tail. The location parameter ι_{jk} is exactly known because I have already defined the domain of each of the components by manually splitting the data. However the shape parameter is unknown and may be affected by treatment, which I model using a multiplicative exponential regression specification to impose a non-negativity constraint on the parameter. The shape parameter in mixture component j for household n in site k is therefore $\exp(\rho_{jk} + \kappa_{jk} T_{nk})$.

The lower level of the likelihood $f(\mathcal{Y}_k|\theta_k)$ is specified according to this mixture distribution. Let $j = 1$ denote the negative tail of the household profit distribution, let $j = 2$ denote the spike at zero, and let $j = 3$ denote the positive tail. Then the household's business profit is distributed as follows:

$$\begin{aligned} y_{nk}|T_{nk} \sim\ & \Lambda_{1k}(T_{nk}) Pareto(-y_{nk}|\iota_{1k}, \exp(\rho_{1k} + \kappa_{1k} T_{nk}) \\ & +\Lambda_{2k}(T_n)\delta_{(0)} \\ & +\Lambda_{3k}(T_n) Pareto(y_{nk}|\iota_{3k}, \exp(\rho_{3k} + \kappa_{3k} T_{nk})\ \forall\ k \\ \text{where}\ \ \Lambda_{jk}(T_{nk}) = & \frac{\exp(\alpha_{jk} + \pi_{jk} T_{nk})}{\sum_{j=1,2,3} \exp(\alpha_{jk} + \pi_{jk} T_{nk})} \end{aligned} \tag{2.1}$$

The quantiles are recovered thus using the Castellaci method:

$$\begin{aligned} Q(u) = -\text{Pareto}^{-1}\left(1 - \frac{u}{\Lambda_1(T_n)} \mid \iota_{1k}, \rho_{1k}(\exp(\kappa_{1k} T_n))\right) * \mathbb{1}\{u < \Lambda_1(T_n)\} \\ + 0 * \mathbb{1}\{\Lambda_1(T_n) < u < (\Lambda_1(T_n) + \Lambda_2(T_n)\} \\ + \text{Pareto}^{-1}\left(\frac{u - (1 - \Lambda_3(T_n))}{\Lambda_3(T_n)} \mid \iota_{3k}, \rho_{3k}(\exp(\kappa_{3k} T_n)\right) * \mathbb{1}\{u > (1 - \Lambda_3(T_n))\} \end{aligned} \tag{2.2}$$

The fit of this model to the microcredit data is not good. The table below shows the posterior predictive fit of this model and the LogNormal model.

Table 5: Posterior Predictive Comparison of LogNormal and Pareto Models

Control Group Quantiles	5%	15%	25%	35%	45%	55%	65%	75%	85%	95%
Revenues Data	0	0	0	0	0	0	4	41	154	622
Lognormal Prediction	0	0	0	0	0	12	37	77	154	408
Pareto Prediction	0	0	0	0	0	0	0	5	337	2,793,933
Expenditures Data	0	0	0	0	0	0	0	17	85	411
Lognormal Prediction	0	0	0	0	0	0	15	40	93	283
Pareto Prediction	0	0	0	0	0	0	0	1	94	1,172,324
Profit Data	-29	0	0	0	0	0	0	4	49	226
Lognormal Prediction	-2	0	0	0	0	0	4	21	56	173
Pareto Prediction	0	0	0	0	0	0	0	0	21	70,170

Notes: The posterior predictive distributions are generated by drawing samples of data from the likelihood averaged over the posterior probability of the unknown parameters. Because this data is itself fat tailed, I have compared the actual sample quantiles from the fully pooled control group against the posterior predicted median value of each quantile from each model. [Back to main]

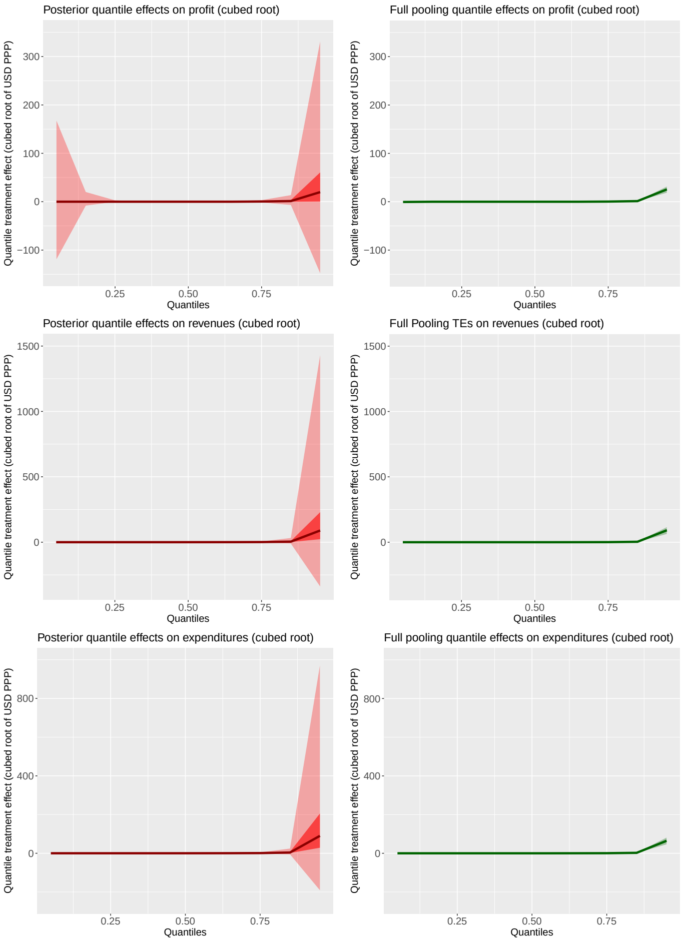


Figure 4: General Quantile Treatment Effect Curves (β_1) for business variables. The dark line is the median, the opaque bars are the central 50% interval, the translucent bands are the central 95% interval. Display is in cubed root of USD PPP due to the scale differences in the uncertainty at the right tail versus the rest of the distribution.

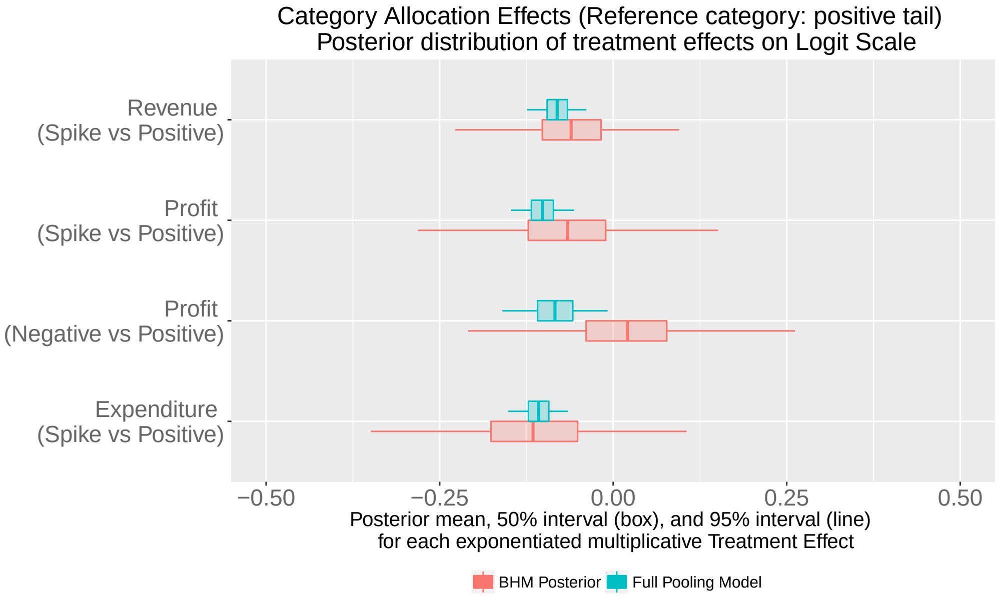


Figure 5: Posterior distributions for the logit treatment effects (π_j) on category assignment. These treatment effects are specified as an exponentiated multiplicative factor on the control group proportion of households in the category: if $\tilde{\pi}_j = 0$ the effect is zero, if $\tilde{\pi}_j < 0$ the treatment increases the proportion of households in the positive tail relative to other categories.

Table 6: Pooling Factors for Categorical Logit Parameters (Reference Category: Positive)

Outcome	Treatment Effects			Control Group Means		
	$\omega(\kappa_j)$	$\breve{\omega}(\kappa_j)$	$\lambda(\kappa_j)$	$\omega(\rho_j)$	$\breve{\omega}(\rho_j)$	$\lambda(\rho_j)$
Profit (Negative vs Positive)	0.378	0.721	0.907	0.144	0.421	0.240
Profit (Zero vs Positive)	0.137	0.476	0.688	0.013	0.379	0.487
Expenditures (Zero vs Positive)	0.084	0.612	0.783	0.010	0.498	0.570
Revenues (Zero vs Positive)	0.131	0.694	0.881	0.010	0.509	0.562

Notes: All pooling factors have support on [0,1], with 0 indicating no pooling and 1 indicating full pooling. The $\omega(\cdot)$ refers to the conventional pooling metric that scores signal strength at the general level against average signal strength at the local level. The $\breve{\omega}(\cdot)$ refers to the proximity-based "brute force" pooling metric that measures the geometric proximity of the partial pooling estimate to the no-pooling and full-pooling estimates. The $\lambda(\cdot)$ refers to the Gelman and Pardoe (2006) pooling metric that scores the posterior variation at the general level against the average posterior variation at the local level.

Table 7: Pooling Factors for Tail Shape Parameters

Outcome	Treatment Effects			Control Group Means		
	$\omega(\pi_j)$	$\breve{\omega}(\pi_j)$	$\lambda(\pi_j)$	$\omega(\alpha_j)$	$\breve{\omega}(\alpha_j)$	$\lambda(\alpha_j)$
Profit (Negative Tail)	0.389	0.855	0.991	0.284	0.346	0.494
Profit (Positive Tail)	0.219	0.785	0.988	0.036	0.074	0.089
Expenditures	0.175	0.756	0.987	0.019	0.061	0.050
Revenues	0.169	0.692	0.977	0.014	0.036	0.029

Notes: All pooling factors have support on [0,1], with 0 indicating no pooling and 1 indicating full pooling. The $\omega(\cdot)$ refers to the conventional pooling metric that scores signal strength at the general level against average signal strength at the local level. The $\breve{\omega}(\cdot)$ refers to the proximity-based "brute force" pooling metric that measures the geometric proximity of the partial pooling estimate to the no-pooling and full-pooling estimates. The $\lambda(\cdot)$ refers to the Gelman and Pardoe (2006) pooling metric that scores the posterior variation at the general level against the average posterior variation at the local level.

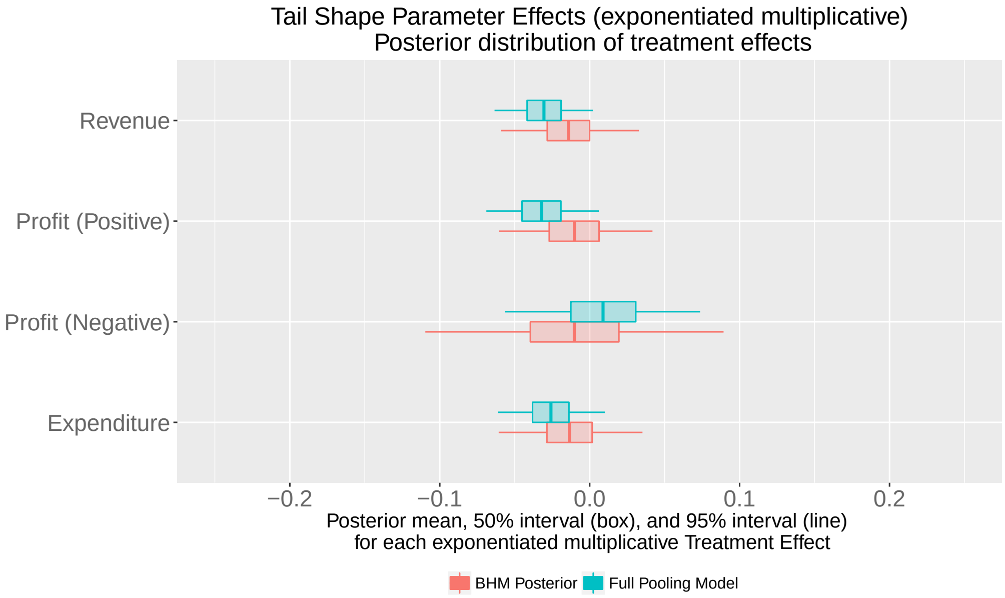


Figure 6: Posterior distributions for the Pareto shape treatment effects (κ_j) in each site. These treatment effects are specified as an exponentiated multiplicative factor on the control group scale parameter: if $\tilde{\kappa_j} = 0$ the effect is zero, if $\tilde{\kappa_j} = 0.7$ the effect is a 100% increase in the scale parameter.

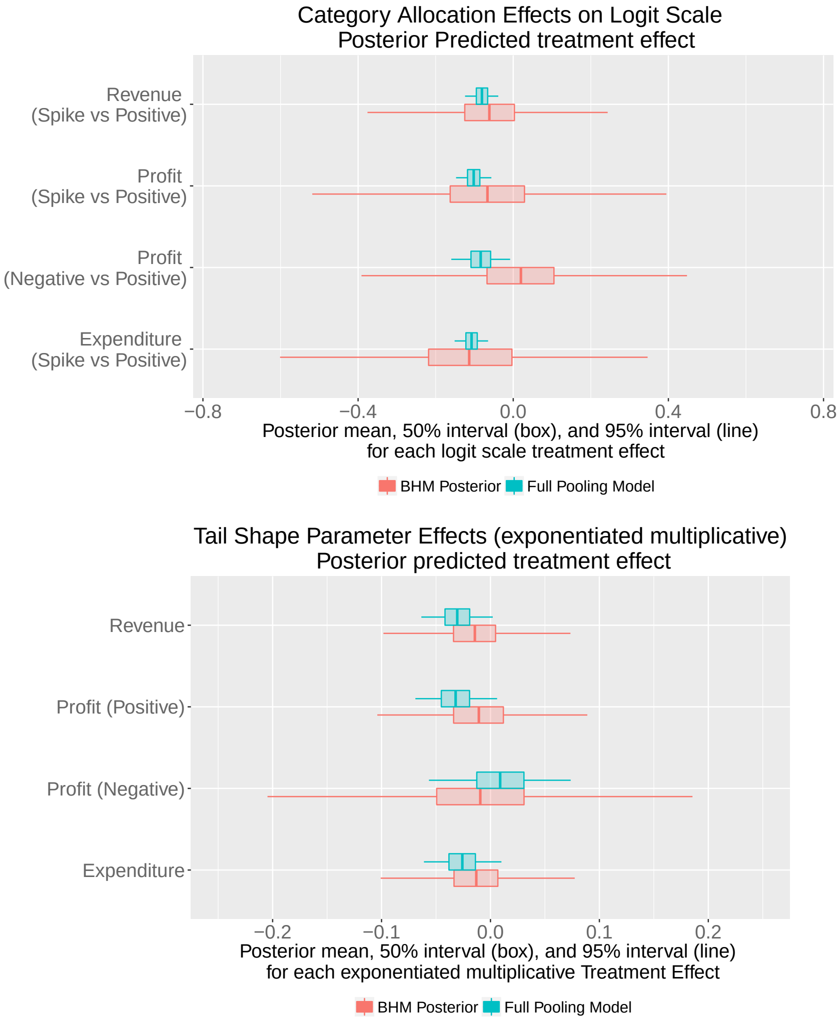


Figure 7: Posterior predicted distributions for the logit treatment effects on category assignment and tail shape effects. In each case this is the predicted treatment effect in a future exchangeable study site, with uncertainty intervals that account for the estimated generalizability (or lack of it).

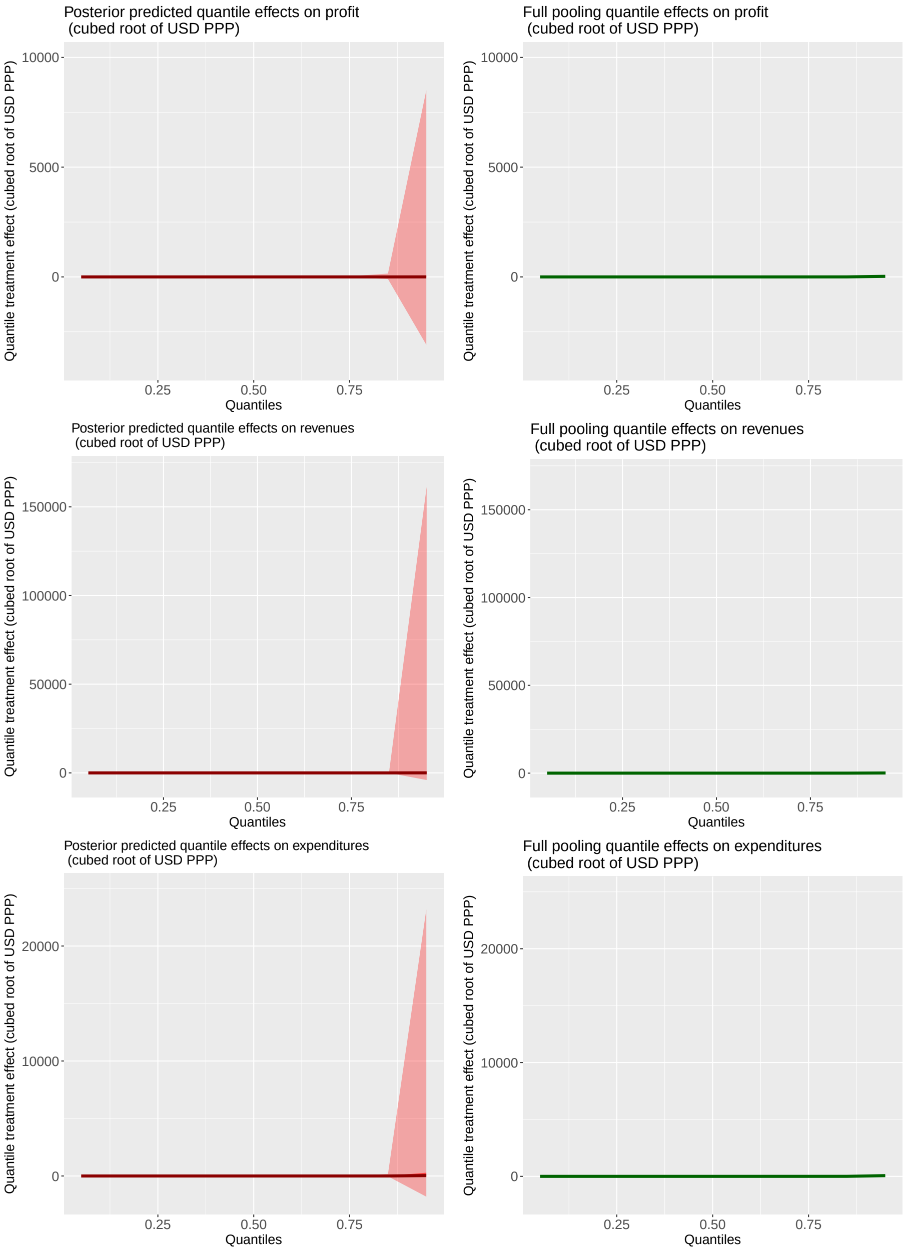


Figure 8: Posterior predicted quantile treatment effect Curves for Business Variables. The dark line is the median, the opaque bars are the central 50% interval, the translucent bands are the central 95% interval. Display is in cubed root of USD PPP due to the scale differences in the uncertainty at the right tail versus the rest of the distribution.

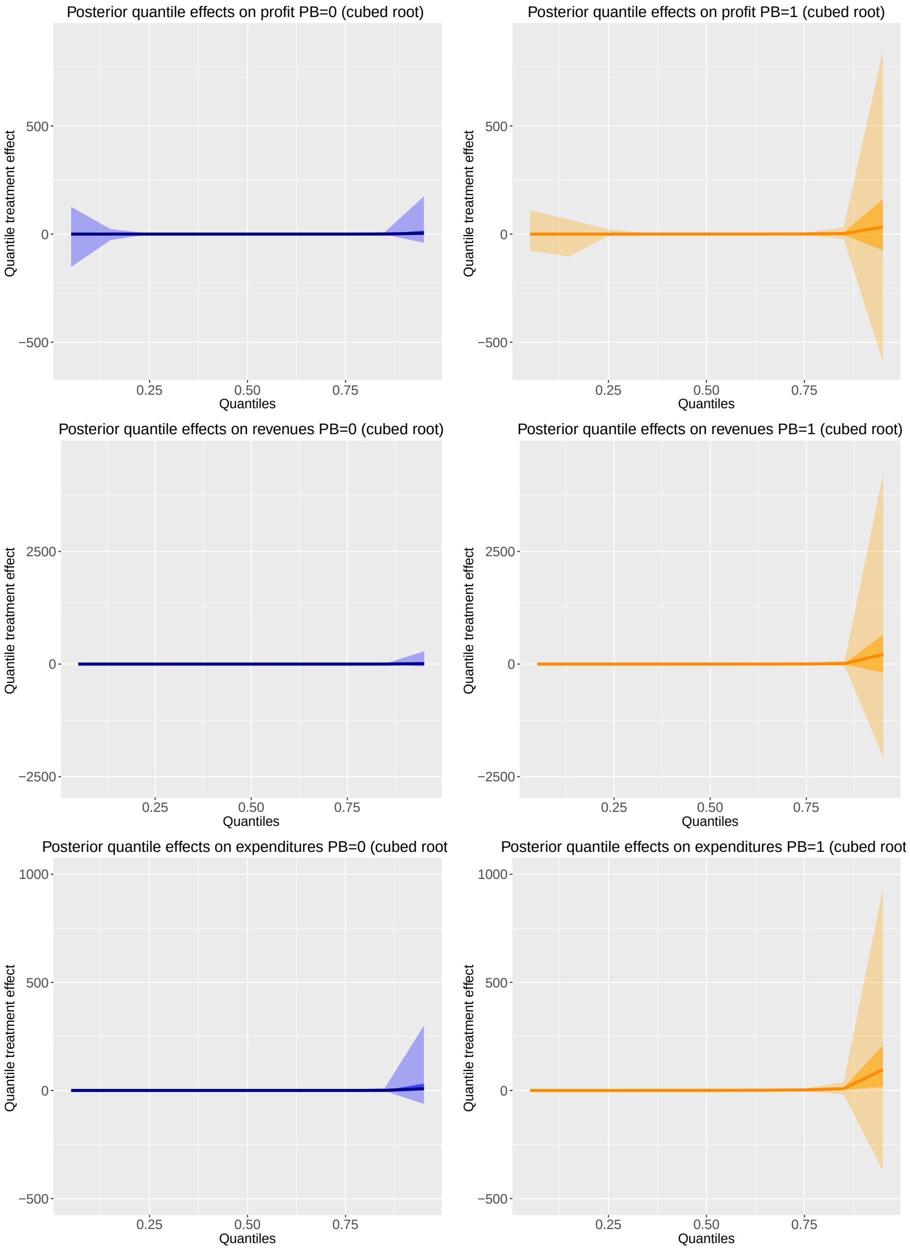


Figure 9: General Quantile Treatment Effect Curves (β_1) for business variables split by prior business ownership. The dark line is the median, the opaque bars are the central 50% interval, the translucent bands are the central 95% interval. Display is in cubed root of USD PPP due to the scale differences in the uncertainty at the right tail versus the rest of the distribution.

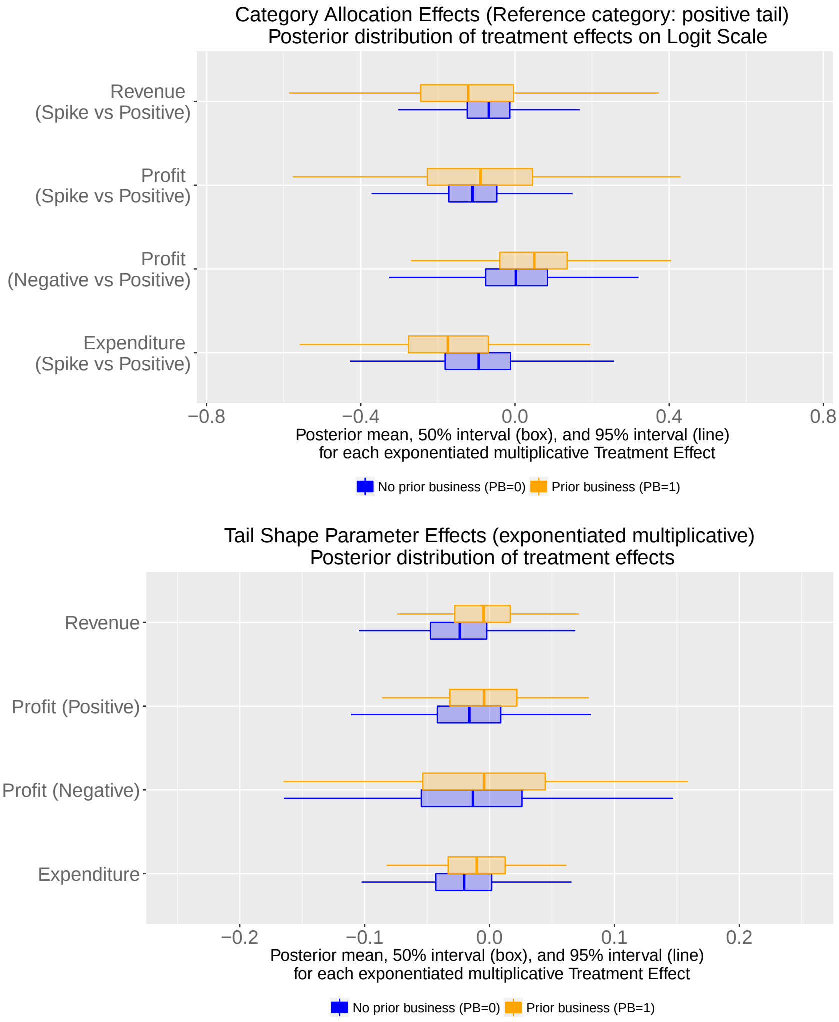


Figure 10: Upper panel: Posterior distributions for the logit treatment effects (π_j) on category assignment split by prior business ownership. These treatment effects are specified as an exponentiated multiplicative factor on the control group proportion of households in the category: if $\tilde{\pi_j} = 0$ the effect is zero, if $\tilde{\pi_j} < 0$ the treatment increases the proportion of households in the positive tail relative to other categories. Lower panel: Posterior distributions for the Pareto shape treatment effects (κ_j) in each site. These treatment effects are specified as an exponentiated multiplicative factor on the control group scale parameter: if $\tilde{\kappa_j} = 0$ the effect is zero, if $\tilde{\kappa_j} = 0.7$ the effect is a 100% increase in the scale parameter.

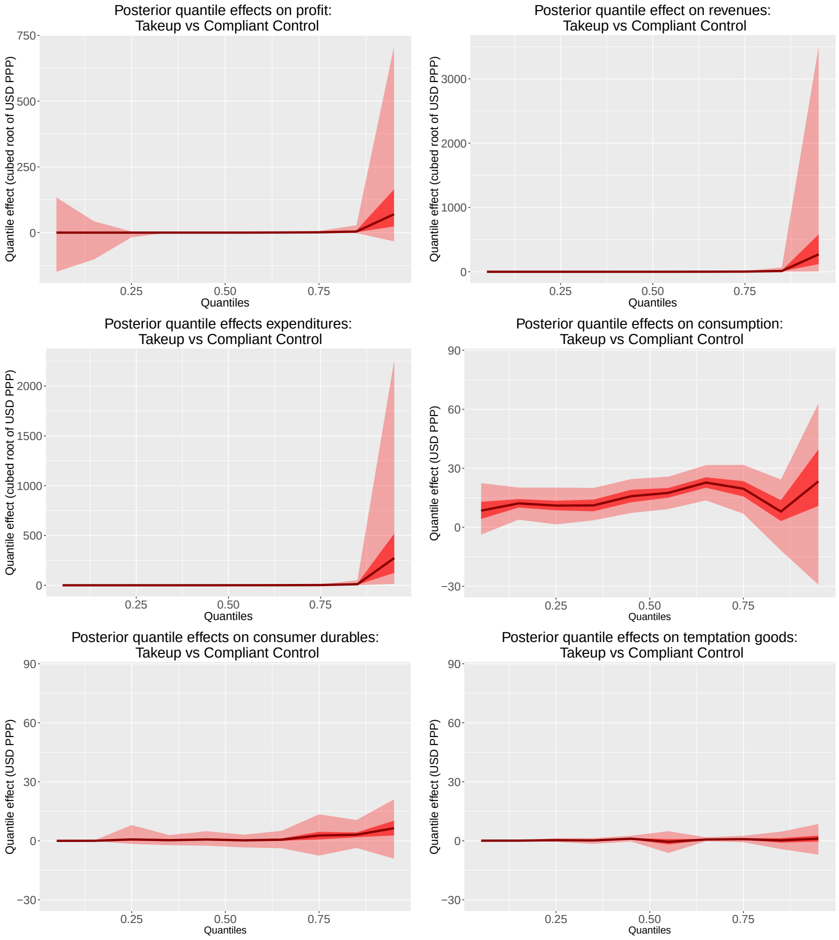


Figure 11: General Quantile Treatment Effect Curves for Business Outcomes: Treated households who took up vs Compliant control households who did not take up. This effect should overestimate the true impact of microcredit on those who take it up in a simple selection framework. Consumption variables are in USD PPP per two weeks, business variables are in cubed root of USD PPP per two weeks due to the scale differences in their uncertainty intervals. The dark line is the median, the opaque bars are the central 50% interval, the translucent bands are the central 95% interval.

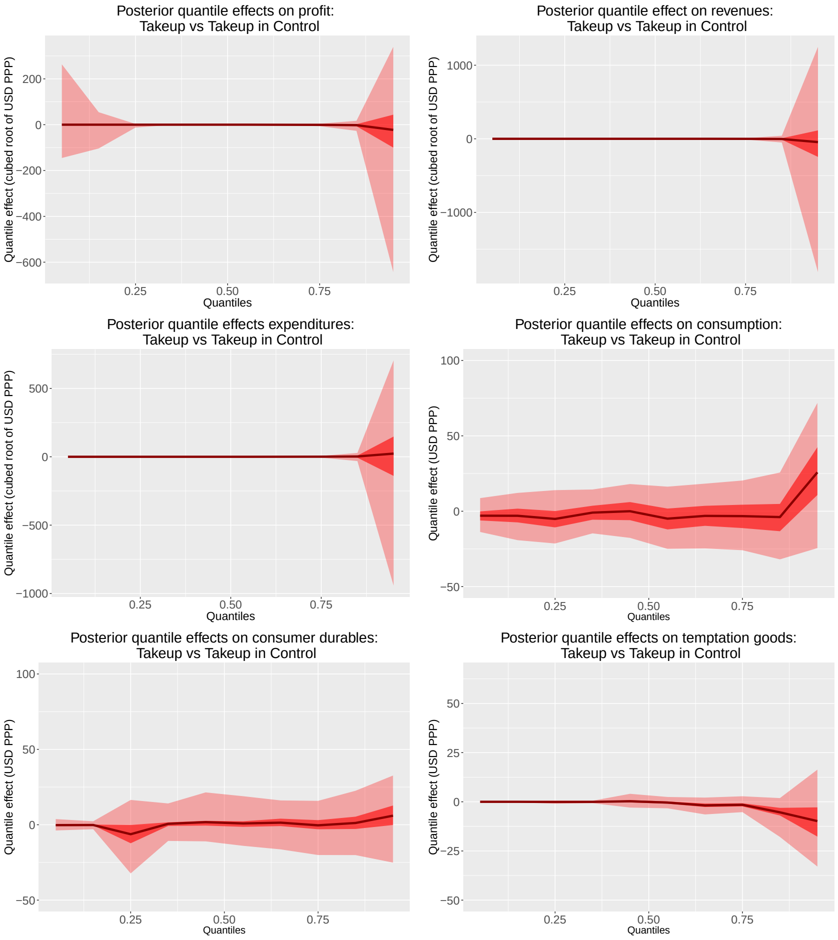


Figure 12: General Quantile Treatment Effect Curves for Business Outcomes: Treated households who took up vs Control households who took up. This effect should underestimate the true impact of microcredit on those who take it up in a simple selection framework. Consumption variables are in USD PPP per two weeks, business variables are in cubed root of USD PPP per two weeks due to the scale differences in their uncertainty intervals. The dark line is the median, the opaque bars are the central 50% interval, the translucent bands are the central 95% interval.

2.2 Flexible Tail Mixture Models

In this section I provide details of the two models fit with more flexible tail specifications than the simple Lognormal. First consider the Pareto-Lognormal model of Reed and Jorgensen (2004). Denote the Mills ratio of the standard Gaussian as $R(z) = (1 - \Phi(z))/\phi(z)$. The log of the Pareto-Lognormal is much more tractable computationally. By taking the log of equation 13 from Reed and Jorgensen (2004), I get the following likelihood function:

$$\ell(y|\alpha, \nu, \tau) := \log(\alpha) + log(\phi((y - \nu)/\tau)) + log(R(\alpha\tau - (y - \nu)/\tau) \tag{2.3}$$

Substituting this tail into the mixture model, for notational clarity given the main paper's notation I now denote these parameters by (A, N, t). Allowing microcredit to affect any of these distributional parameters in any way, I get the following:

$$\begin{aligned} y_{nk}|T_{nk} \sim\ & \Lambda_{1k}(T_{nk})\ell(-\log(y_{nk})|A_{1k} + TE_{A1k}T_{nk}, N_{1k} + TE_{N1k}T_{nk}, t_{1k} + TE_{t1k}T_{nk})) \\ & + \Lambda_{2k}(T_n)\delta_{(0)} \\ & + \Lambda_{3k}(T_n)\ell(\log(y_{nk})|A_{3k} + TE_{A3k}T_{nk}, N_{3k} + TE_{N3k}T_{nk}, t_{3k} + TE_{t3k}T_{nk})) \ \forall\ k \\ \text{where } \ \Lambda_{jk}(T_{nk}) = & \frac{\exp(\alpha_{jk} + \pi_{jk}T_{nk})}{\sum_{j=1,2,3}\exp(\alpha_{jk} + \pi_{jk}T_{nk}))} \end{aligned} \tag{2.4}$$

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

$$(\alpha_{1k}, \alpha_{2k}, \alpha_{3k}, \pi_{1k}, ...)' \equiv \zeta_k \sim N(\zeta, \Upsilon)\ \forall\ k \tag{2.5}$$

The priors for this model need to be strong to overcome the convergence issues noted in Reed and Jorgensen (2004), which on this data was particularly problematic on the parameter A. Following extensive testing and discussion with computational experts, I chose the following priors with a view to computational performance.[1]

$$\begin{aligned} A &\sim N(3,2) \\ \zeta \backslash A &\sim N(0,3) \\ \Upsilon &\equiv diag(\nu_\Upsilon)\Omega_\Upsilon diag(\nu_\Upsilon)' \\ \nu_\Upsilon &\sim \text{halfNormal}(0,3) \\ \Omega_\Upsilon &= I_{|\zeta|} \\ \alpha_{mk} &\sim N(0,5). \end{aligned} \tag{2.6}$$

Even with these quite specific priors, this model still represents a strict relaxation of the LogNormal model fit in the paper; while the priors are stronger conditional on the given parameters, the parameters themselves construct a weaker structure on the data. However, while the convergence issues are mitigated, they are not eliminated: the "Rhat" criteria statistics from this model are indeed further from 1 than those of the LogNormal tail model, indicating poorer convergence and less reliable posterior inference despite these priors (Gelman and Rubin 1992). For this reason I minimize focus on this model in the main paper.

To avoid the convergence issues without having to employ such strong and specific priors, it is possible to employ the original insight from the methods section again and split up the tail into two components with disjoint supports: a Lognormal for the component with support adjacent to zero, and a Pareto for the extremal tail component.[2] This leads to the following "composite tail" likelihood with the Pareto location parameter ι naturally taking the form of the breakpoint or cutoff location:

$$Composite(y|\iota,\rho,\mu,\sigma) := \mathbb{1}\{y \leq \iota\}Lognormal(y|\mu,\sigma) + \mathbb{1}\{y > \iota\}Pareto(y|\iota,\rho) \tag{2.7}$$

The challenge in practice is how to define the cutoff location ι in a hierarchical context, as one can no longer rely on the convenient scale-invariance of the cutoff location being zero. For tractability, in light of potential convergence issues, I do not allow the proportion of data in the two tail components to change at all in this model and I estimate it before the rest of the model; this two-step procedure is not ideal but it is computationally advantageous. I have defined the cutoff for the microcredit data as at the 80th quantile of the positive continuous tail and the 20th quantile

[1] I thank Dr Michael Betancourt in particular, as well as Dr Ben Goodrich and Professor Aki Vehtari, for their advice and assistance with this problem. A public record of our work can be found here https://discourse.mc-stan.org/t/double-pareto-lognormal-distribution-in-stan/10097/20

[2] Once again I think Ulrich Müller and Andriy Norets for this insight.

of the negative continuous tail within each site, which corresponds to a model in which 80% of the data in every tail takes a LogNormal form, and the most extremal 20% of draws take a Pareto shape. In practice this is quite easy to implement: one uses any well-behaved quantile estimator, frequentist or Bayesian, within each tail to generate $\hat{\iota}$, with no inferential problems as this data is continuous. Then, one fits the mixture model with the tails taking the form of the composite model above, with $\hat{\iota}$ treated as data. I use the original priors from the main model on all the hyperparameters. To recover the quantiles, one uses the same Castellaci (2012) method, noting that one must rescale the cutoff quantile by 0.2 in the negative tail and 0.8 in the positive tail to determine the "average" cutoff ι at the superpopulation level. Given the suboptimal two-step nature of this procedure I do not focus on this model in the main results.

2.3 Running the Rubin Model Quantile by Quantile

Table 8: Profit: Results of running the Rubin (1981) model quantile by quantile

Quantile:	94th	95th	96th
Partial Pooling			
Bosnia	280.5	255.6	251.7
	(39.1,524.6)	(65.7,445.3)	(-25.1,535.9)
India	-16.3	-16.9	-19.4
	(-53.6,21.5)	(-63.6,29.4)	(-58.4,20.2)
Mexico	-0.1	19.7	20.5
	(-15.6,15.2)	(-0.5,40)	(1.7,39)
Mongolia	0	-0.1	-0.5
	(-0.7,0.7)	(-1.4,1.1)	(-2.6,1.6)
Morocco	95.6	87.3	157.9
	(4.4,188)	(-43.6,217.3)	(4.2,311.6)
Ethiopia	5.5	3.6	4.3
	(-1.8,12.8)	(-7.2,14.4)	(-12.7,21.3)
Philippines	339.8	454.1	681.7
	(-24.9,705.2)	(-16.6,916.4)	(168.3,1208.1)
Average	2.3	7.9	9.5
	(-11.2,59.2)	(-13.3,99.2)	(-19.6,141.8)

Notes: All units are USD PPP per two weeks. Estimates are shown with their 95% uncertainty intervals below them in brackets. These models had difficulty converging and likely do not represent a good fit to the data. This may be because the Gaussian approximation to the sampling error is unlikely to hold for this data given its extreme kurtosis.

Table 9: Consumption: Results of running the Rubin (1981) model quantile by quantile

Quantile:	5th	15th	25th	35th	45th	55th	65th	75th	85th	95th
Partial Pooling										
Bosnia	-4.6	-2.2	-1.4	-2.2	-0.7	1.7	2.1	-1.6	-9.4	35.5
	(-10,0.8)	(-11.7,3.4)	(-10.7,5.3)	(-13.2,4.3)	(-8.9,9.9)	(-8.8,11)	(-19.1,11.2)	(-41.2,13.7)	(-77.1,18.5)	(-57,177)
India	-0.6	-0.6	-1.3	-1.5	-1.3	0.1	3.1	4.6	6.7	34.2
	(-6,5.3)	(-5.4,3.8)	(-6.8,3.5)	(-6.1,3.3)	(-6.3,4.1)	(-7.6,5.6)	(-4.4,8.9)	(-5.1,13.7)	(-7.3,20.8)	(-4.9,78.2)
Mexico	-8.4	-0.9	-1.4	-1.8	-1	2.9	4.5	8.9	11	16.3
	(-12.9,-3.9)	(-5,2.7)	(-5.5,2.3)	(-5.5,2)	(-5.4,3.7)	(-1.8,8.2)	(-0.1,9.5)	(1.5,17.5)	(0.3,22.7)	(-6.9,38.6)
Mongolia	1.8	0.9	-0.2	-1.1	-1.6	1.7	3.2	3.4	-0.4	40.3
	(-8.9,19.2)	(-6.4,12.8)	(-8.1,10.6)	(-8.1,11.7)	(-15.4,7.2)	(-8.7,10.7)	(-12.2,15.1)	(-22.5,22)	(-45.2,28.7)	(-40.5,157.8)
Morocco	0	2.5	1	-1.4	-1.7	2.4	3.5	2.4	-2.9	-37.4
	(-6.3,7.4)	(-2.4,9.5)	(-3.8,8)	(-5.9,3.2)	(-8,3.6)	(-3.4,8.8)	(-4.7,11.3)	(-9.5,12.4)	(-20.2,11.8)	(-89.8,14.4)
Average	-2.6	-0.2	-0.6	-1.6	-1.3	1.6	3.1	4	0.8	17.3
	(-11.2,7)	(-6.1,6.4)	(-6.8,5.3)	(-6.6,3.6)	(-9.2,4.5)	(-5.8,7.6)	(-7.5,11.2)	(-10,25.5)	(-27.9,19.7)	(-56.2,95.7)

Notes: All units are USD PPP per two weeks. Estimates are shown with their 95% uncertainty intervals below them in brackets.

2.4 Leaving Out Certain Studies

Across the different studies, both an eyeball test and the results of the main analysis show that the underlying data are on very different scales, and that the control groups look quite different. This makes the main conclusion that the quantile effects are quite similar for most of the distribution even more striking. However, one might be concerned that a single study with a particularly large or small scale is driving or unduly affecting the results. In this case Mongolia, with its much smaller scale than all other studies, or Bosnia with its unique lack of negative profit observations, are the main concerns.

I have re-run the analysis leaving out Bosnia and Mongolia respectively. The results are shown in the graph below with the main results for comparison. Leaving out Bosnia changes virtually nothing; leaving out Mongolia makes the results much more uncertain and somewhat more positive, but still displays the same fundamental pattern and substantive conclusions of the main analysis.

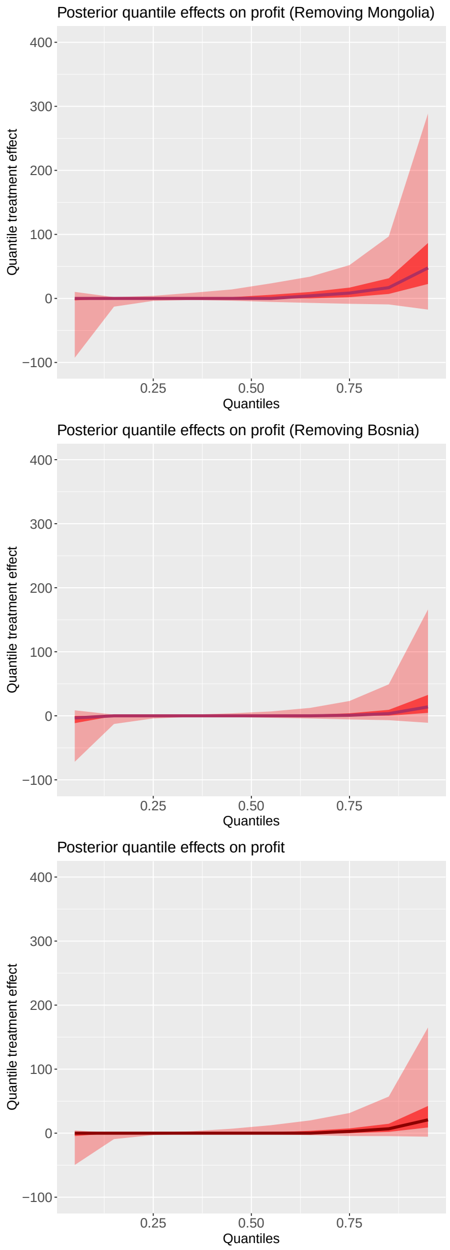


Figure 13: Average quantile treatment effects across all settings for profit (USD PPP per two weeks) without Mongolia (graph 1), without Bosnia (graph 2), and main results with all sites for comparison.

2.5 Trimming the data

With such extreme kurtosis values, it would be of interest to understand whether removing the largest 0.5% of values from the data set as a whole substantially impacts the inference. I examine the positive tail as this is the location of both the greatest uncertainty and greatest potential for positive effects. The table below shows the inference on the lognormal tail parameters for the profit data with these top positive values trimmed out. The posterior mean intervals on these parameters are reasonably stable across the original and trimmed data sets. While the lognormal scale parameters are slightly smaller, and τ_2 has most notably been reduced from approximately 0.077 to 0.057 indicating the important role of the extremal upper tail in generating even these results, this is within half a standard deviation of the original estimate.

Table 10: Profit Tail Inference from Trimmed Data (top 0.5% removed)

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
μ_1	3.228	0.025	0.810	1.733	2.747	3.214	3.679	4.828	$1,038.916$	0.999
μ_2	3.795	0.034	0.846	2.119	3.264	3.809	4.317	5.449	603.721	1.003
τ_1	0.096	0.003	0.092	-0.078	0.044	0.095	0.146	0.277	$1,265.345$	1.000
τ_2	0.057	0.001	0.047	-0.037	0.030	0.057	0.083	0.148	$1,946.863$	1.000
σ_{μ_1}	1.835	0.025	0.784	0.912	1.303	1.640	2.150	3.844	$1,002.582$	1.000
σ_{μ_2}	2.178	0.024	0.799	1.189	1.639	1.996	2.499	4.092	$1,094.257$	1.000
σ_{τ_1}	0.111	0.004	0.123	0.003	0.033	0.074	0.146	0.435	$1,082.284$	1.001
σ_{τ_2}	0.073	0.002	0.059	0.003	0.029	0.060	0.099	0.219	957.807	1.001
σ_1^c	0.453	0.005	0.144	0.181	0.373	0.446	0.526	0.759	837.234	1.004
σ_2^c	0.179	0.004	0.106	-0.049	0.116	0.180	0.242	0.388	868.749	1.003
σ_1^t	0.023	0.002	0.089	-0.154	-0.025	0.025	0.069	0.215	$1,370.381$	1.001
σ_2^t	0.001	0.001	0.026	-0.051	-0.014	0.001	0.015	0.054	$1,624.716$	1.001
$\sigma_{\sigma_1^c}$	0.307	0.006	0.169	0.126	0.199	0.264	0.364	0.737	912.078	1.004
$\sigma_{\sigma_2^c}$	0.268	0.004	0.112	0.142	0.195	0.242	0.309	0.530	992.346	1.003
$\sigma_{\sigma_1^t}$	0.164	0.004	0.122	0.037	0.091	0.136	0.202	0.448	$1,211.250$	1.000
$\sigma_{\sigma_2^t}$	0.038	0.001	0.032	0.002	0.016	0.031	0.051	0.118	$1,026.529$	1.002

Appendix D:
Bounds on Complier Effects

Rachael Meager

January 12, 2021

1 The role of take-up

One concern about the models presented in the main analysis is that they ignore the role of differential take-up in explaining the impact of microcredit. While the results of the analysis stand for themselves as group-level causal impacts, the economic interpretation of the results might differ if we knew, for example, that the zero impact along most of the outcome quantiles was entirely due to lack of take-up by most of the households in the sample. The main results contain suggestive evidence that the lack of impact at most quantiles is not solely due to lack of take-up: the 2 sites that randomized loan access rather than branch access and therefore had almost full take-up (Bosnia and the Philippines) displayed the same trend as all the other sites. Yet the observed pattern of zeroes could still be due to low differential in take-up between treatment and control group, which was recorded in most sites. It would be ideal to understand the effect of microcredit on those who are induced to take up loans by this random expansion of access (the "compliers" in the Neyman-Rubin causal framework).

The core challenge to an analysis of the impact on compliers is that the Stable Unit Treatment Value Assumption (SUTVA) is unlikely to hold for individual households within a village, such that there is no satisfactory way to identify the distributional treatment effect only on those households. Without SUTVA it may still be possible to infer certain average characteristics of the compliers as in Finkelstein and Notowidigdo (2018), but this exercise does not easily extend to quantiles and relies on zero effects for never-takers. Even if SUTVA did hold, the conventional LATE result is not available to us because there is two-sided non-compliance in these samples: I not only have treated households who do not take up loans, but

I also have control households who do manage to access loans from the MFI being studied. Finally, even if the above two complications were not present, we would still face the challenge of translating these results to a quantile effects framework which is nontrivial as quantile functions do not obey any law comparable to the law of iterated expectations.

As an alternative approach, I pursue a bounding exercise that provides suggestive evidence that loan take-up patterns are unlikely to be responsible for the precise zero results along most of the distribution. Ideally, the right comparison to make is between the group of households who took up microcredit only due to the random expansion of access, and the same group of households in the control group. This comparison estimates the distributional effect on the compliers. But we cannot identify those households in the control group, because they are indistinguishable from the "never taker" households. Nor can we separate the compliers from the "always takers" in the treatment group. However, under a set of broadly reasonable assumptions for the microcredit setting - that is, assuming SUTVA may be violated - it is possible to develop bounds on the changes in the compliers' distribution.

The bounds I propose can be intuitively described and justified using the following reasoning. First, in the style of an individual-rationality constraint, one assumes that the always-taker and complier groups who take up the microloans should see weakly positive effects from actually taking up these loans versus not taking them up, and that this should be the case even if other things are changing in the environment around these borrowers. Second one assumes that always-takers ought to do better than compliers from taking up, since they take up even if it is very costly to do so. Third, one assumes that the spillover effects or other consequences of any SUTVA violation ought generally to be smaller than the direct effects on the compliers or always-takers.

Under these assumptions, consider comparing the outcomes for individuals who took up in the treatment group (always takers and compliers) and subtracting the outcome of those who took up in the control group (always-takers). One would imagine that the difference in outcomes observed between these groups has to be smaller than the treatment effect on compliers, since it compares within a set of people who all took up loans, and if anything the compliers probably do worse than always takers in absolute terms in the take-up state. Hence, this comparison might form a sensible lower bound on the complier effect.

Now consider comparing the outcomes for individuals who took up in the treatment group (always takers, compliers) against individuals who did not take up in the control group (compliers, never-takers). One expects the always takers to do better from microcredit than compliers do, and if one's treatment effect is somewhat positively correlated with one's raw outcomes (as seems to be the case in the cross-country evidence) then one expects never-takers to fare no better than compliers on average. Therefore, one might expect that the outcomes for those who take up are an overestimate of the complier outcomes in the treatment state, while the outcomes of those who don't take up in the control state might be an underestimate of the outcomes of compliers in the control state. Thus, subtracting the latter from the former ought to produce a larger gap than the difference in the outcomes that compliers see from taking up versus not.

In the following section I derive a set of sufficient conditions under which these intuitive bounds hold even when we encounter violations of SUTVA and moderate rank re-ordering of households (even such that they can cross ranks with households from other groups).

1.1 Analytical Bounds Derivation Set-Up

Denote the three possible groups of households: always takers, compliers and never takers, $G \in \{AT, C, NT\}$, adopting the no defiers assumption standard from the LATE literature. Denote the quantiles of a group's outcome distribution $Q_G(TU, T)$ where TU is a binary indicator of taking up the loans, and T is any vector of assigned treatment status for the households in the given village or local area. I now derive a set of sufficient, though strong, conditions that justify the empirical bounds I propose above.

Assumption 1. *Quantile treatment effects on compliers and always-takers are weakly positive regardless of any effects of changes in the treatment assignment allocation (that is, regardless of the potential for SUTVA violations). Thus, pointwise for any u and ∀ T, T':*

$$\begin{aligned} Q_{AT}(1,T)(u) - Q_{AT}(0,T')(u) \geq 0 \\ Q_C(1,T)(u) - Q_C(0,T')(u) \geq 0 \end{aligned} \tag{1.1}$$

Assumption 1 would be implied by a similar ordering on the individual treatment effects, but as this present ordering does not imply a full ordering on the individual effects, it is more general. Some individuals in these groups can experience negative effects, but not so many nor so punitively that they outweigh the countervailing set of individuals who experience positive effects. In spirit, this is a quantiles version of the assumption of a positive expected return from taking up a loan.

If SUTVA holds, then the secondary argument is irrelevant and this collapses to an ordering of quantile treatment effects within any given treatment assignment. Once we accept that SUTVA is unlikely to hold, requiring stability of effects across treatment assignment regimes is natural, since by definition the compliers only take up when assigned to do so and therefore their "treatment effect" comparison always involves a counterfactual assignment status.

In addition, I use the following first order stochastic dominance assumption to generate a partial ordering of the quantiles of the three groups.

Assumption 2. *Pointwise for any u, ∀ T, T':*

$$\begin{aligned} &F_{AT}(1,T) \textit{ FOSD } F_C(1,T) \textit{ and} \\ &F_C(0,T') \textit{ FOSD } F_{NT}(0,T') \end{aligned} \tag{1.2}$$

When a particular CDF F FOSD another, let us say F', then F always lies to the right of F'. Hence, when they are transposed to quantile functions, the quantiles of F always lie above the quantiles of F'. Hence, Assumption 2 implies that for example $Q_{AT}(1,T)(u) \geq Q_C(1,T)(u)$ for any u. This is a strong assumption and makes the derivation of the quantile effect bounds quite simple; the bounds will still hold under moderate violations of this assumption.

Assumption 2 is also a little unusual in that it involves an ordering on absolute outcomes rather than the conventional ordering on treatment effects that one uses in monotonic selection models. Yet the specific assumption 2 here will be implied by the monotonic selection assumption if in addition the levels of the outcomes are somewhat positively correlated with this treatment effect. Assumption 2 will be unlikely to hold, even approximately, if either of these two patterns does not hold in the data. Fortunately, it seems likely that this is indeed the case for microcredit.

To see why, consider that as it generally takes a lot of time and effort to access microcredit, households are more likely to do so if their own treatment effects are larger. Further, while we cannot be sure this reflects the ordering within sites, the cross-site correlation between the average treatment effects and the control groups' levels of consumption, profit, etc. is generally positive (Meager, 2018). This at least provides suggestive evidence that a positive correlation between levels and effects may be present within sites as well, and that as a result, these bounds are reasonable here. Thus while this assumption and thus the bounds I derive may not be applicable to every situation, they do seem applicable to the microcredit data.

Finally I employ the following assumption on the SUTVA violation adjustments – also known as spillover effects across treatment assignment statuses – requiring them to be weakly smaller than the direct effects on compliers and always takers of actually *taking up* loans. This seems sensible since it is hard to imagine how any spillover could be larger than the direct effect occurring; if giving me a loan increased my neighbour's consumption more than my own, economic theory and practice would need to be quite different than it is now. Of course, this is likely not the case for the never takers who may not experience any effects – or even negative group effects – but fortunately my bounds do not require any assumption on the size of the spillovers on the nevertakers. This requirement for moderation in the effect of changing treatment regimes from T' to T on our complier and always-taker groups can be expressed in assumption 3:

Assumption 3. *For always-takers and compliers, spillover effects are always smaller than direct effects. Pointwise for any u, ∀ T, T', and ∀ G, G' $\in AT, C$,*

$$Q_G(1,T)(u) - Q_G(1,T')(u) \leq Q_{G'}(1,T)(u) - Q_{G'}(0,T')(u). \tag{1.3}$$

Armed with these conditions, I can derive bounds without assuming that there is no effect on the never-takers (in contrast to Imbens and Rubin 1997, Abadie Angrist and Imbens 2002, and Finkelstein and Notowidigdo 2018) which is fortunate because this is unlikely when SUTVA is violated in general, and particularly when other households in one's village are taking up new sources of credit.

In the following sections, I provide the upper and lower bounds for which the above assumptions are sufficient but not necessary. The value of the sufficiency conditions is that they provide some intuition for the situations in which the bounds are likely to hold. The bounds themselves are their own necessary conditions. The sufficiency conditions are nevertheless important because they allow us to develop an understanding of why and how we might expect these bounds to be relevant in any given study.

1.2 The Upper Bound

First, consider comparing the outcome of the households who take up in treatment versus those households who do not take up in control, as a potential upper bound on the distributional effects on compliers. These groups are composed of combinations of the three groups denoted above, so denote the quantiles of a combination of two groups by Q_{ATC} for the pooled set of always-takers and compliers, and Q_{NTC} for the pooled set of never-takers and compliers.

Theorem 1.1. *Under assumptions 1 and 2, the following* ***upper boundary condition on the complier quantile effects*** *holds pointwise for any quantile u.*

$$Q_{ATC}(1,T)(u) - Q_{NTC}(0,T')(u) \geq Q_C(1,T)(u) - Q_C(0,T')(u)\ \forall\ T,T'. \tag{1.4}$$

Proof. From assumption 2, we know $Q_{AT}(1,T)(u) \geq Q_C(1,T)(u)\ \forall\ u$. A random pooling of the two groups - such as occurs during a randomized trial - will generate an intermediate set of quantiles such that

$$Q_{AT}(1,T)(u) \geq Q_{ATC}(1,T)(u) \geq Q_C(1,T)$$

By a reverse argument, assumption 2 also implies that

$$Q_C(0,T)(u) \geq Q_{NTC}(0,T)(u) \geq Q_{NT}(0,T)$$

Hence,

$$Q_{ATC}(1,T)(u) - Q_{NTC}(0,T')(u) \geq Q_C(1,T)(u) - Q_C(0,T')(u).$$

□

1.3 The Lower Bound

Continuing with the same set-up, now consider comparing the outcome of the households who take up in treatment versus those households who take up loans in control. For this comparison to form a lower bound on the distributional effects for compliers, it must be that the following result holds.

Theorem 1.2. *Under assumptions 1, 2 and 3, the following* ***lower boundary condition on the complier quantile effects*** *holds pointwise for any quantile u.*

$$Q_{ATC}(1,T)(u) - Q_{AT}(1,T')(u) \leq Q_C(1,T)(u) - Q_C(0,T')(u)\ \forall\ T,T' \tag{1.5}$$

Proof. From assumption 2, we know $Q_{ATC}(1,T)(u) < Q_{AT}(1,T)(u)$, so

$$Q_{ATC}(1,T)(u) - Q_{AT}(1,T')(u) \leq Q_{AT}(1,T)(u) - Q_{AT}(1,T')(u).$$

The only difference in $Q_{AT}(1,T)(u)$ and $Q_{AT}(1,T')(u)$ is a spillover effect from the change in the treatment allocation regime. But for always takers and compliers, assumption 3 guarantees that this spillover effect must be smaller than the treatment effect at any quantile. Hence,

$$Q_{AT}(1,T)(u) - Q_{AT}(1,T')(u) \leq Q_C(1,T)(u) - Q_C(0,T')(u).$$

Combining these two statements produces the required bound. □

2 Empirical Bounds Results

I compute these bounds and I find that the posited lower bound does lie weakly below the posited upper bound in all cases (and strictly below in the case of consumption). However, the bounds are very close together and overall similar to the main distributional effect estimated by comparing treatment status itself (the "ITT" comparison, or the "access as treatment" comparison). Comparing the households who took up the loans in the treatment group to households in the control group who did not take up loans produces largely similar results - although they are weakly more positive - as comparing all treated and control households, as shown in figure 1. The results of comparing the households who took up the loans in the treatment group to households who took up in the control group for all outcomes is shown in figure 2. These effects tend to be broadly similar to the impact of mere access, in that they are zero almost everywhere, although on average the effects are estimated to lie weakly below the ITT effect. Taken together these results suggest that the bounds are themselves applicable to the microcredit studies and that the broad pattern of zero effects along most of the distribution occurs within the complier group as well as in the general population of households.

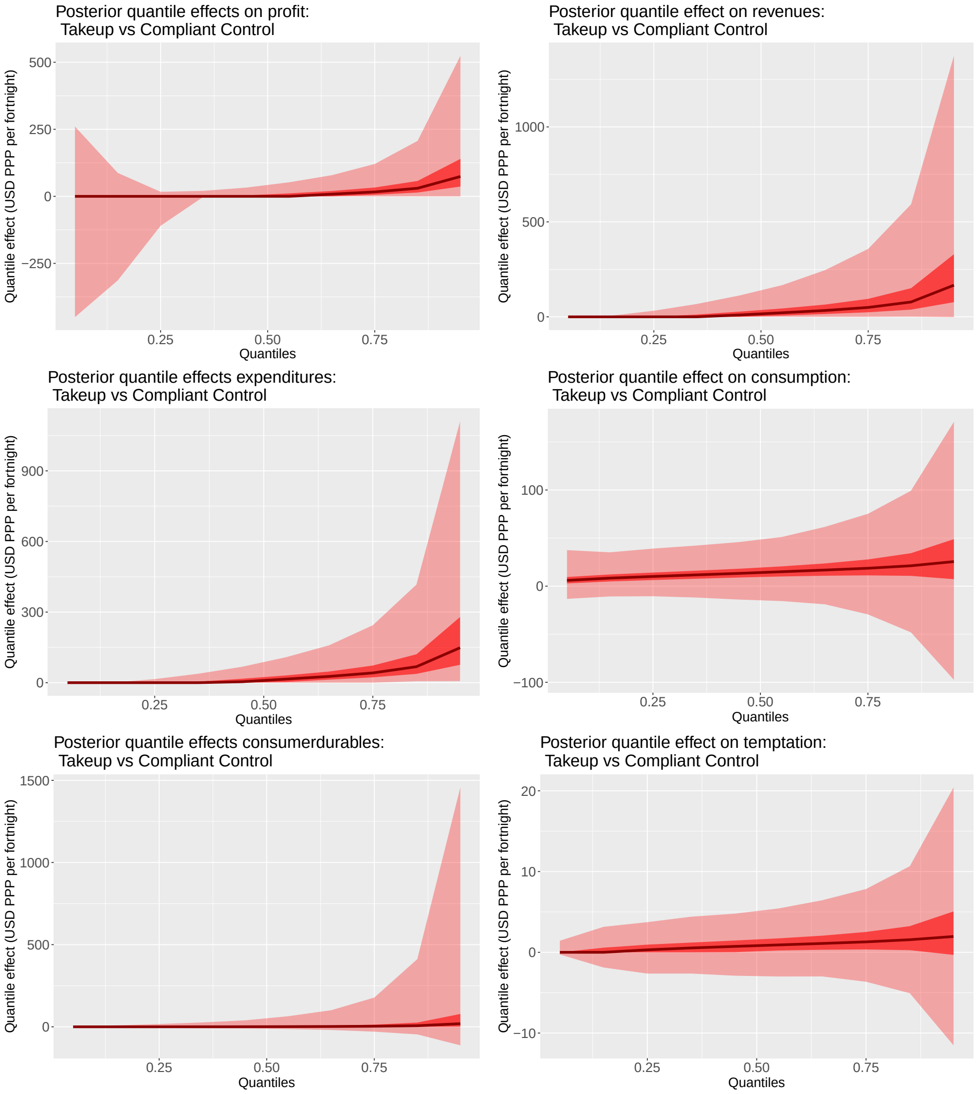


Figure 1: General Quantile Treatment Effect Curves for Business Outcomes: Treated households who took up vs Compliant control households who did not take up. This effect should overestimate the true impact of microcredit on those who take it up in a simple selection framework. The dark line is the median, the opaque bars are the central 50% interval, the translucent bands are the central 95% interval. [Back to main]

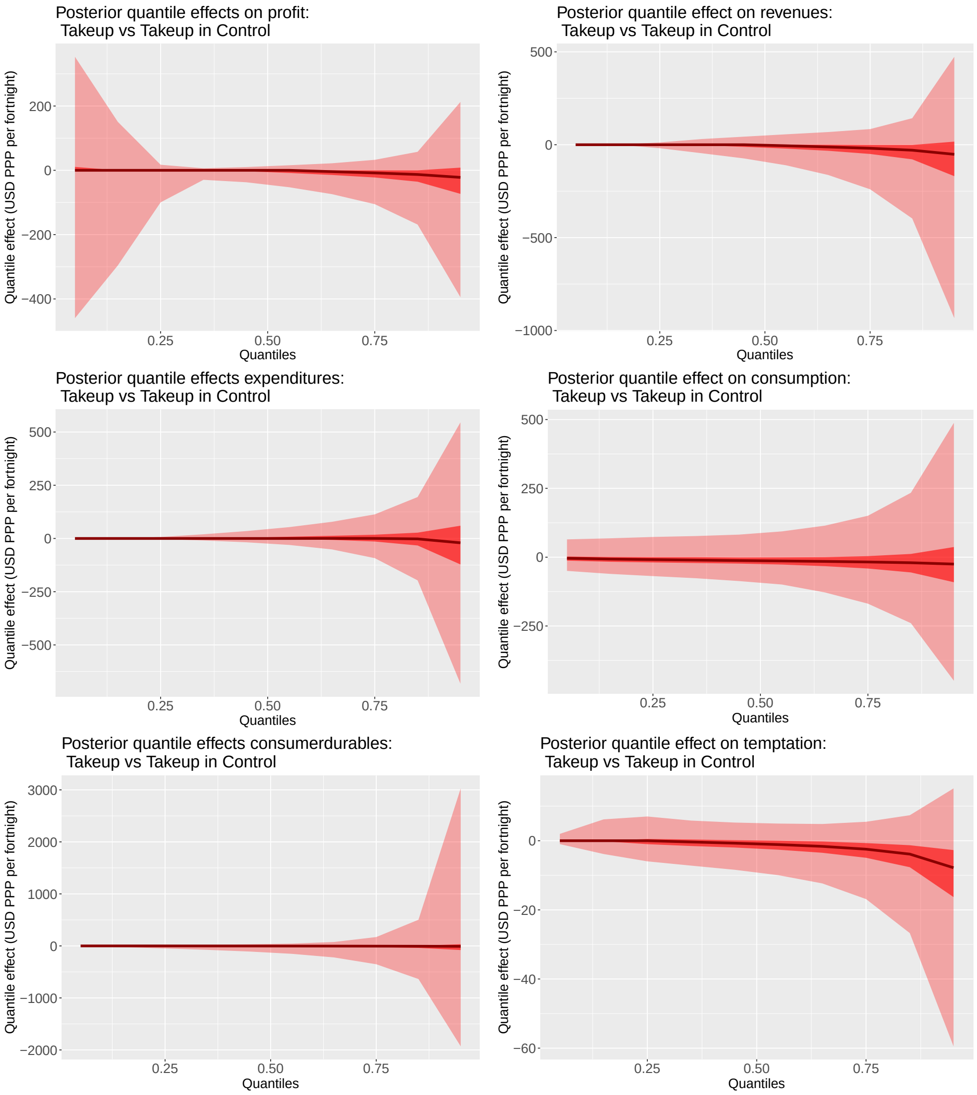


Figure 2: General Quantile Treatment Effect Curves for Business Outcomes: Treated households who took up vs Control households who took up. This effect should underestimate the true impact of microcredit on those who take it up in a simple selection framework. The dark line is the median, the opaque bars are the central 50% interval, the translucent bands are the central 95% interval. [Back to main]