Bayesian Distribution Regression*

Weige Huang† Emmanuel S. Tsyawo‡

Abstract

This paper introduces a Bayesian version of distribution regression that enables inference on estimated distributions, quantiles, distributional effects, among other functionals of interest. Our estimators come in three categories: the non-asymptotic, semi-asymptotic, and asymptotic. To conduct simultaneous inference on a function of any estimator, we introduce asymmetric and symmetric Bayesian confidence bands. Inference on point estimates is conducted via posterior intervals. The Bayesian asymptotic theory we develop extends the foregoing to gains in computational time and tractability of posterior distributions. Monte Carlo simulations conducted illustrate good performance of our estimators. We apply our estimators to evaluate the impact of institutional ownership on firm innovation.

JEL classification: C11, D22

Keywords: Distribution regression, Counterfactual analysis, Bayesian inference, Simultaneous confidence bands, Institutional ownership, Innovation

*We thank Brantly Callaway, Oleg Rytchkov, Charles E. Swanson, Xiaoyong Zheng, and Mark Robinson for valuable comments. All errors are ours.

†Department of Economics, Temple University. Email: weige.huang@temple.edu

‡Corresponding Author. Department of Economics, Temple University. Email: estsyawo@temple.edu

1 Introduction

Distribution regression and quantile regression are widely used techniques in econometrics to estimate distributional parameters namely distributions, quantiles, distribution effects, quantile effects, average effects, measures of spread viz. variance, inter-quartile range among others. Though Foresi and Peracchi (1995)[1] introduced distribution regression, Chernozhukov, Fernández-Val, and Melly (2013) recently increased its popularity in applied economics. Examples of works that employ distribution regression are Doorley and Sierminska (2012), Shepherd et al. (2013), Wüthrich (2015), Han, Lutz, and Sand (2016), Richey and Rosburg (2016), Chernozhukov, Fernandez-Val, Melly, and Wüthrich (2016) Dube (2017), Chernozhukov, Fernández-Val, Han, and Kowalski (2018), and Chernozhukov, Fernandez-Val, and Weidner (2018).

In a manner akin to Yu and Moyeed (2001), Schennach (2005), and Lancaster and Jae Jun (2010) who develop Bayesian quantile regression based on working likelihoods, we develop a Bayesian distribution regression (hereafter BDR) method which leverages the likelihood in a distribution regression framework in order to arrive at a Bayesian version of distribution regression.[2] BDR can be a useful tool in the study of various questions. For instance, one can use BDR to study poverty/inequality in labor economics, Value-at-Risk (VaR) in financial economics and counterfactual and decomposition analyses in policy evaluation, and so on. Generally, our method enables one to explore questions which can be studied by distribution regression and quantile regression. Although our method can be useful in many fields, for our empirical application, we focus on counterfactual analysis by applying our method to study the impact of institutional ownership on firm innovation.

Though standard distribution regression can obtain the entire distribution of an outcome variable conditional on other covariates, one ought to rely on the bootstrap in order to conduct inference. Our approach enables direct inference on parameters of interest, such as quantiles, distributions, distribution effects, quantile effects, stochastic dominance, and so on using Bayesian techniques[3] that only require posterior and proposal distributions. More specifically, we are able to obtain not only the distribution of the outcome variable as an estimand, as obtained by the standard (i.e. frequentist) distribution regression, but the (posterior) density thereof. In this paper, we propose

[1] They estimate the conditional distribution of excess returns by estimating a sequence of conditional logit models over a grid of values of excess returns.

[2] We notice Law, Sutherland, Sejdinovic, and Flaxman (2017) developed a machine learning method termed Bayesian distribution regression. It is, however, very different from ours.

[3] We use Markov Chain Monte Carlo techniques and asymptotic Bayesian approximations.

three estimators: the non-asymptotic, semi-asymptotic and asymptotic BDR estimator. The non-asymptotic BDR uses posterior simulations based on the available data sample at fixed threshold values separately. While the semi-asymptotic uses simulated draws of parameters from the normal approximated posterior at thresholds separately, the asymptotic version simulates draws from the joint normal approximated posterior across several threshold values. For all three estimators, we provide asymmetric and symmetric Bayesian confidence bands for simultaneous inference and posterior intervals for point estimates. Other methods for correcting point-wise intervals for simultaneous coverage like the sup-t, Bonferroni, Šidák, $\theta-$projection, and $\mu-$projection bands[4] can be implemented with our estimators.

The bootstrap is usually employed for (frequentist) distribution regression in order to perform inference whereas ours provides the entire distribution over which inference is performed using Bayesian posterior intervals and confidence bands.[5] In addition, the asymptotic (normal) approximation of the posterior distribution obtains as a closed form function of the modes at different points of the outcome. This feature of the approximated posterior enables the derivation of joint distributions (and inference as a result) of the outcome, counterfactual, distribution and quantile treatment effects among other estimands of interest at arbitrarily many distinct points of the outcome. Also, where counterfactual or partial effects are of interest, we are able to obtain their distribution, thus, paving the way for hypothesis tests not only on point estimates like the mean, median, and standard deviation but also on entire processes like the distribution and quantile functions using Bayesian simultaneous inference.

Bayesian distribution regression can be viewed as an alternative to Bayesian quantile regression.[6] Beyond the advantages of doing Bayesian inference[7] on distributions estimated using distribution regression, Bayesian distribution regression that we propose enables us to carry over to the Bayesian framework, the gains in using distribution regression in general. As noted by Chernozhukov, Fernández-Val, and Melly (2013), distribution regression (DR), unlike quantile regression, does not require smooth conditional distributions; it handles discrete, continuous or mixed outcome variables fairly well.[8] DR allows heterogeneity in the impact of covariates on the outcome

[4] See Montiel Olea and Plagborg-Møller (2018, sect. 3.1) for a discussion of these.

[5] An applicable confidence band is the (bayesian) simultaneous bands of Montiel Olea and Plagborg-Møller (2018, algorithm 2). We provide two easier-to-compute alternatives; symmetric and asymmetric (bayesian) confidence bands.

[6] See Yang and Wang (2017) for a review of Bayesian quantile regression methods derived from different types of working likelihoods.

[7] Bayesian inference is conditional on the data and is exact. It allows the incorporation of prior information in a logical way that follows from Bayes' theorem. For a thorough treatment of the advantages and disadvantages of Bayesian inference, see Berger (2013, sections 4.1 and 4.12).

[8] In their simulation exercise in appendix SB of the supplemental material, Chernozhukov, Fernández-Val, and

at different points of the distribution (Chernozhukov, Fernandez-Val, Melly, and Wüthrich, 2016). Chernozhukov, Fernández-Val, and Melly (2013, Remark 3.1) shows that distribution regression involves simpler steps[9] in computing the distribution but not the quantile function. Bayesian posterior intervals and confidence bands do have a direct probabilistic interpretation, unlike frequentist methods. Also, once the posterior density of a statistic is obtained, simulating draws from it is computationally faster than, say, the bootstrap that may require the re-estimation of the model (and statistic) on each bootstrap data sample.[10] Bayesian distribution regression unifies the convenience of distribution regression and Bayesian inference.

We apply our method to study the impact of institutional ownership on innovation (as studied by Aghion, Van Reenen, and Zingales (2013)) and show that the presence of severe mass points at zero leads to an underestimation of the the effect when parametric poisson, negative binomial and hurdle models are used. Also, we find substantial heterogeneity in the distribution and quantile effects. Point-wise and simultaneous confidence bands reveal non-trivial asymmetry in the density of quantile and distribution effects.

The rest of the paper is organized as follows. In section 2, we present the Bayesian distribution regression model, define the (counterfactual) distribution, quantiles of the outcome, distribution effects, and quantile effects. In section 3, we outline the estimation algorithm and discuss the computation of parameters of interest. Tools of Bayesian inference developed in Section 4 enable simultaneous inference. Section 5 presents the asymptotic theory and joint inference at several points of the distribution of the outcome using the normal approximation of the posterior distribution. We conduct Monte Carlo simulations in section 6 and study the impact of institutional ownership on innovation using our method in section 7. Section 8 concludes.

2 The model

The focus of this paper is to develop a Bayesian approach to distribution regression. A key ingredient for this task is the likelihood function, needed in addition to the prior distribution, to compute the posterior distribution of parameters. Suppose observed data are independent and identically distributed samples $(y_i, \mathbf{x}_i)$ where the outcome variable y can be discrete, mixed or

Melly (2013) show that quantile regression is only more accurate with continuous conditional distributions and performs worse in the presence of mass points.

[9]Distribution regression involves a convenient functional form that involves neither inversion or trimming. See Chernozhukov, Fernández-Val, and Melly (2013, Remark 3.1).

[10]This assertion is at least true for the non-parametric bootstrap.

continuous and the $N \times k$ matrix $\mathbf{x}$ includes a treatment variable t (in the case of treatment effect analysis) and other covariates X. A threshold value $y_g \in \mathcal{Y}$, where $\mathcal{Y} \subset \mathbb{R}$ denotes the support of outcome y, enables us to define a binary variable $\check{y}_i^g = \mathbb{1}\{y_i \leq y_g\}$ that equals one if $y_i \leq y_g$ and zero otherwise. Distribution regression involves running a binary response model of $\check{y}^g$ on covariates $\mathbf{x}$ at a threshold or a continuum of threshold values[11] $y_g \in \mathcal{Y}$ and the conditional distribution obtains as $F_Y(y_g|\mathbf{x}) = P(y_i \leq y_g|\mathbf{x}) = \Lambda(\mathbf{x}\boldsymbol{\theta}_g)$ where $\Lambda(\cdot)$ is the link function, $\Lambda(\nu) \in (0,1)\ \forall\ \nu \in \mathbb{R}$, and $\boldsymbol{\theta}_g \in \mathbb{R}^k$ is a $k \times 1$ vector of unknown parameters.

2.1 Likelihood and posterior

For a threshold value y_g, the likelihood of an observation i is given by

$$p(\check{y}_i^g|\boldsymbol{\theta}_g) = \Lambda(\mathbf{x}_i\boldsymbol{\theta}_g)^{\check{y}_i^g}(1 - \Lambda(\mathbf{x}_i\boldsymbol{\theta}_g))^{1-\check{y}_i^g} \tag{2.1}$$

Popular choices of link functions include the logistic and normal distribution. The conditional distribution obtains as $F_Y(y_g|\mathbf{x}) = \Lambda(\mathbf{x}\boldsymbol{\theta}_g)$.[12] As Chernozhukov, Fernández-Val, and Melly (2013, section 3.1.2) notes, any link function can approximate the conditional distribution arbitrarily well by using sufficiently rich transformations of $\mathbf{x}$, for example, polynomials, b-splines and tensor products. The joint likelihood at a fixed y_g and vector of parameters $\boldsymbol{\theta}_g$ is given by

$$p(\check{y}^g|\boldsymbol{\theta}_g) = \prod_{i=1}^{N} p(\check{y}_i^g|\boldsymbol{\theta}_g) = \prod_{i=1}^{N} \Lambda(\mathbf{x}_i\boldsymbol{\theta}_g)^{\check{y}_i^g}(1 - \Lambda(\mathbf{x}_i\boldsymbol{\theta}_g))^{1-\check{y}_i^g} \tag{2.2}$$

For notational ease, we suppress covariates $\mathbf{x}$ in the likelihood. Where the entire distribution process is of interest, distribution regression proceeds by maximising (eq. (2.2)) at a continuum of thresholds y_g in $\bar{\mathcal{Y}} \in \mathcal{Y}$ that cover the support of y fairly well. The choice of the finite subset $\bar{\mathcal{Y}}$ in $\mathcal{Y}$ for a continuous or mixed y needs to satisfy the condition that the Hausdorf distance between $\bar{\mathcal{Y}}$ and $\mathcal{Y}$ is approaching zero at a rate faster than $1/\sqrt{N}$ (Chernozhukov, Fernandez-Val, and Weidner (2018, remark 2)).

Using the likelihood (eq. (2.2)) above, the following posterior distribution of $\boldsymbol{\theta}_g$ obtains using

[11] Running a binary response at a threshold or several of them depends on the application at hand. For instance, to compute the poverty rate, only a threshold value is of interest. To estimate an entire distribution process, several threshold values are required.

[12] See Koenker and Yoon (2009) for a thorough study of link functions for binary response models.

Bayes' theorem

$$p(\boldsymbol{\theta}_g|\breve{y}^g) = \frac{p(\breve{y}^g|\boldsymbol{\theta}_g)p(\boldsymbol{\theta}_g)}{p(\breve{y}^g)} \propto p(\breve{y}^g|\boldsymbol{\theta}_g)p(\boldsymbol{\theta}_g) \tag{2.3}$$

where the proportionality follows because $p(\breve{y}^g)$ does not depend on $\boldsymbol{\theta}_g$ and $p(\boldsymbol{\theta}_g)$ is a prior density.[13] Observe the dependence of $\boldsymbol{\theta}_g$ on y_g via $\breve{y}^g$ in $\breve{y}_i^g = \mathbb{1}\{y_i \leq y_g\}$. In fact, for fairly distinct values of $y_g \in \bar{\mathcal{Y}}$, $g = 1, ..., G$, the posterior distributions $\{p(\boldsymbol{\theta}_g|\breve{y}^g)\}_{g=1}^{G}$ are distinct.

2.2 Parameters of interest

Given the above posterior distribution (eq. (2.3)), it is straightforward to use Markov Chain Monte Carlo (MCMC) methods[14] to obtain draws of $\boldsymbol{\theta}_g$. All other parameters of interest are functions of $\boldsymbol{\theta}_g$ and do derive their distribution from eq. (2.3). In this paper, parameters of interest include the conditional distribution, unconditional distribution, quantiles of outcome, the distribution effect, quantile effect, (conditional) probability, and value-at-risk (VaR). The parameters are in two main categories; those that derive from the conditional distribution (or probability) and those from the quantile of the outcome.[15]

Conditional probability

Some applications are concerned with the conditional or unconditional probability of the outcome being below a threshold, for instance a poverty line. Distribution regression provides a flexible tool for such inference. Recall,

$$P(y \leq y_g|\mathbf{x}) = F_Y(y_g|\mathbf{x}) = \mathit{\Lambda}(\mathbf{x}\boldsymbol{\theta}_g)$$

measures the conditional probability that outcome $y \in \mathcal{Y}$ at most equals a threshold $y_g \in \bar{\mathcal{Y}}$ given characteristics $\mathbf{x}$. An extension over a set $\bar{\mathcal{Y}} \subset \mathcal{Y}$ obtains the conditional distribution function $F_Y(y_g|\mathbf{x}), g = 1, ..., G$.

[13]The adoption of a uniform prior on $\boldsymbol{\theta}_g$ is convenient when a theoretical basis for a proper prior density is lacking. A uniform prior in no way impairs formal posterior analyses (Berger (2013, section 4.2.3)). In addition, in high dimensional settings (see the empirical application in section 7), a proper prior density is difficult to motivate.

[14]Two options, the Independence Metropolis-Hastings and the Random Walk Metropolis-Hastings algorithms, are available in our R package `bayesdistreg`. For a thorough treatment of methods for posterior simulation and computation, see Gelman, Carlin, Stern, and Rubin (1995, ch. 11 and 13).

[15]For a comprehensive study of parameters estimable using distribution and quantile regression methods, refer to Barrett and Donald (2009), Firpo and Pinto (2015), and Callaway and Huang (2017).

Unconditional distribution

At a fixed threshold y_g, the unconditional distribution function $F_Y(y),\ y \in \mathcal{Y}$ is

$$F_Y(y) = \int F_{Y|X}(y|\mathbf{x})dF(\mathbf{x})$$

Denoting $F_Y(y_g)$ as y_θ^g at a threshold y_g for notational simplicity, the probability density function[16] of y_θ^g is

$$p(y_\theta^g) = \int p(y_\theta^g|\breve{y}^g, \boldsymbol{\theta}_g)p(\boldsymbol{\theta}_g|\breve{y}^g)p(\breve{y}^g)d\breve{y}^g d\boldsymbol{\theta}_g \tag{2.4}$$

In practice, y_θ^g is computed as $N^{-1}\sum_{i=1}^N \Lambda(\mathbf{x}_i\boldsymbol{\theta}_g)$.[17]

It is straightforward to extend our method to estimate the counterfactual distributions and compute counterfactual effects. The importance of estimating counterfactual distributions for policy analysis (Stock (1989) and Heckman and Vytlacil (2007)) lies in its ability to uncover heterogeneity in the impact of covariates on the distribution (and by extension the quantiles) of the outcome.

Counterfactual distribution

One can obtain the counterfactual distribution of the outcome by replacing $\mathbf{x}$ with $\mathbf{x}^c$ where $\mathbf{x}^c$ is a counterfactual of $\mathbf{x}$.

$$y_\theta^{g,c} = \int F_{Y|X}(y_g|\mathbf{x}^c)dF(\mathbf{x}^c) = F_Y^c(y_g), \quad y_\theta^{g,c} \in (0,1)$$

where $y_\theta^{g,c}(\cdot)$ is the counterfactual of $y_\theta^g(\cdot)$. The counterfactual expression for eq. (2.4) obtains by simply replacing y_θ^g with $y_\theta^{g,c}$.

Distribution effect

The distribution effect Δ_g^{DE} at a threshold $y_g \in \mathcal{Y}$ given by

$$\Delta_g^{DE} = y_\theta^{g,c} - y_\theta^g = F_Y^c(y_g) - F_Y(y_g) \tag{2.5}$$

[16] The distribution of other parameters derive from that of $\boldsymbol{\theta}_g$ in an analogous fashion.

[17] Since the distribution $F_Y(y_g|\boldsymbol{\theta}_g)$ may be non-monotone in y_g, we apply the monotonisation method of Chernozhukov, Fernández-Val, and Galichon (2010) based on rearrangement. In practice, we may think of rearrangement as sorting (Chernozhukov, Fernández-Val, and Galichon (2010), p. 1098).

measures the impact on the fraction of individuals with outcome less than y_g as a result creating a counterfactual of $\mathbf{x}$. It has the following probability density function

$$p(\Delta_g^{DE}) = \int p(\Delta_g^{DE}|\check{y}^g, \boldsymbol{\theta}_g)p(\boldsymbol{\theta}_g|\check{y}^g)p(\check{y}^g)d\check{y}^g d\boldsymbol{\theta}_g \tag{2.6}$$

The distribution effect conditional on $\mathbf{x}$ can also be obtained as

$$\Delta_g^{DE}(\mathbf{x}) = F_Y^c(y_g|\mathbf{x}) - F_Y(y_g|\mathbf{x})$$

In addition to the distribution effect at a threshold y_g, one may also be interested in the mean distribution effect

$$E(\Delta^{DE}) = \int_{\mathcal{Y}} (F_Y^c(y) - F_Y(y))dy$$

Quantile effect

The quantile or left inverse is important for parameters like the average (conditional) outcome, quantile effects, VaR among others that are generally easier to interpret because they are defined with respect to the outcome y. The quantile function, given a distribution function $F_Y(y_g)$, $\tau \in (0, 1)$, is defined as

$$F_Y^{-1}(\tau) = \inf\{y \in \mathcal{Y} : F_Y(y) \geq \tau\}$$

and the counterfactual $F_Y^{c-1}(\tau)$ and conditional quantile functions are defined analogously. Defining the conditional quantile in a similar fashion, average conditional outcome obtains as

$$E(Y|\mathbf{x}) = \int_0^1 F_Y^{-1}(\tau|\mathbf{x})d\tau$$

The quantile effect at the τ'th quantile of y is given by

$$\Delta_\tau^{QE} = F_Y^{c-1}(\tau) - F_Y^{-1}(\tau)$$

and the average effect obtains as

$$E(\Delta^{QE}) = \int_0^1 (F_Y^{c-1}(\tau) - F_Y^{-1}(\tau))d\tau$$

The distribution of the counterfactual quantile effect does not obtain as a direct product of distribution regression at a single index y_g but rather after inverting the entire distribution on $\mathcal{Y}$. We discuss Bayesian inference on quantiles in the next section.

Measures of spread

In addition to the above, measures of spread viz. (conditional) variance, standard deviation, quartiles, inter-quartile range, of outcome (or quantiles), distribution, distribution effect, quantile effect, etc. may be of interest. The variance of a function $\nu(\cdot)$ is defined as

$$var(\nu) = \int (\nu(\zeta) - E(\nu))^2 d\zeta$$

where $E(\nu) = \int \nu(\zeta) d\zeta$ and the index $\zeta = \tau \in (0,1)$ for quantile-based functions or $\zeta = y \in \mathcal{Y}$ for distribution-based ones.

VaR

Value-at-risk[18] is a standard quantitative measure of risk essential in risk management for both financial institutions and regulators. It measures how much of an investment can be lost with a given confidence level $1-\alpha$. Conceptually simple, VaR is a quantile of future returns y_h given current information $\mathbf{x}_{-h}$. VaR_h (VaR h periods ahead) with a $1-\alpha$ level of confidence satisfies the relationship[19] $Pr(y_h < -VaR_h|\mathbf{x}_{-h}) = \alpha$ which can be expressed as

$$F_{Y_h|\mathbf{X}_{-h}}(-VaR_h|\mathbf{x}_{-h}) = \alpha$$

The continuity and monotonicity of $F_{Y_h|\mathbf{X}_{-h}}(\cdot)$ allows us to have

$$VaR_h = -F^{-1}_{Y_h|\mathbf{X}_{-h}}(\alpha|\mathbf{x}_{-h})$$

3 Estimation

We present an algorithm for the estimation of Bayesian distribution regression. This not only makes the practical understanding of it easier but also facilitates computations. To aid computation and applicability, we provide an R package `bayesdistreg`.

[18] For further discussion on estimating VaR using quantile regression, see Engle and Manganelli (2004) and Gaglianone, Lima, Linton, and Smith (2011).

[19] The set-up of the DR model needs to use returns at a specified number of periods ahead h as the outcome variable y_h and lagged variables as the set of covariates $\mathbf{x}_{-h}$.

3.1 BDR Algorithm

All parameters (see parameters of interest in section 2.2) are derived from the parameter set $\{\boldsymbol{\theta}_g\}_{g=1}^G$. In the following algorithm, we provide steps to obtain simulated draws of $\{\boldsymbol{\theta}_g\}_{g=1}^G$ using MCMC methods. The following algorithm can be thought of as the non-asymptotic BDR for estimating and simulating draws of $\boldsymbol{\theta}_g$, $g = 1, .., G$.

Algorithm 1 (BDR algorithm)**.**

1. *Fix a grid of threshold values* $y_g \in \bar{\mathcal{Y}} \subset \mathcal{Y}$, $g = 1, ..., G$.
2. *For each* $g = 1, ..., G$
   (a) *Obtain the likelihood function* $p(\breve{y}^g|\boldsymbol{\theta}_g)$ *using eq.* (2.2) *where* $\breve{y}_i^g = 1\{y_i \leq y_g\}$ *and the posterior distribution* $p(\boldsymbol{\theta}_g|\breve{y}^g)$ *therefrom using eq.* (2.3)*.*
   (b) *For each* $m = 1, ..., M$*'th MCMC simulation*
      i. *Make a draw* $\boldsymbol{\theta}_g^{(m)}$ *from the posterior* $p(\boldsymbol{\theta}_g|\breve{y}^g)$*.*[20]
   (c) *end* m
3. *end* g

Remark 1. *Algorithm 1 does not treat the simulation from the posterior* $p(\boldsymbol{\theta}_g|\breve{y}^g)$ *for* $g = 1, ..., G$ *jointly just like in distribution regression and quantile regression. When inference over the distribution or quantile function is of interest, simultaneous inference methods that we propose can be used to correct the point-wise inference that algorithm 1 generates.*

3.2 Parameters of interest (continued)

The conditional probability can be estimated as

$$\hat{P}(y < y_g|\mathbf{x}) = \hat{F}_{Y|X}(y_g|\mathbf{x}) = \varLambda(\mathbf{x}\hat{\boldsymbol{\theta}}_g)$$

and the unconditional probability obtains by averaging across $\mathbf{x}_i$, $i = 1, ..., N$. $\hat{F}_Y(y_g) = \frac{1}{N}\sum_{i=1}^N \varLambda(\mathbf{x}_i\hat{\boldsymbol{\theta}}_g)$. In a similar vein, the counterfactual is computed as $\hat{F}_Y^c(y_g) = \frac{1}{N}\sum_{i=1}^N \varLambda(\mathbf{x}_i^c\hat{\boldsymbol{\theta}}_g)$. The distribution effect then obtains as the difference $\hat{\Delta}_g^{DE} = \hat{F}_Y^c(y_g) - \hat{F}_Y(y_g)$. The average distribution effect simply

[20] In the package `bayesdistreg`, the functions `IndepMH()` and `RWMH()` implement the independence and random-walk chain Metropolis-Hastings algorithms. For a general treatment of posterior simulation techniques, see Gelman, Carlin, Stern, and Rubin (1995, chapter 11) and Albert (2009, chapter 6).

obtains from averaging across indices $g = 1, ..., G$ i.e. $\bar{\Delta}^{DE} = \frac{1}{G}\sum_{g=1}^{G}(\hat{F}_Y^c(y_g) - \hat{F}_Y(y_g))$. Notice that the preceding point estimates are a function of the parameter vector $\boldsymbol{\theta}_g$. Algorithm 1 gives M simulations of the parameter vector at each threshold $\boldsymbol{\theta}_g,\ g = 1, ..., G$. For some point estimator $\nu(\boldsymbol{\theta}_g)$, the probability density function obtains as the vector of random draws $[\nu(\boldsymbol{\theta}_g^{(1)}), ..., \nu(\boldsymbol{\theta}_g^{(M)})]$, $m = 1, ..., M,\ g = 1, ..., G$.

The quantile function, given a distribution function, can be computed as

$$\hat{F}_Y^{-1}(\tau) = \inf\{y_g \in \bar{\mathcal{Y}} : \hat{F}_Y(y_g) \geq \tau\},\ \tau \in (0, 1)$$

and a conditional quantile obtains as $\hat{F}_Y^{-1}(\tau|\mathbf{x}) = \inf\{y_g \in \bar{\mathcal{Y}} : \hat{F}_Y(y_g|\mathbf{x}) \geq \tau\}$. The estimand of the VaR is a special case of the conditional quantile (see section 2.2) with $\alpha = \tau$. With the (conditional) quantile, it is straightforward to calculate the average conditional outcome $\hat{E}(Y|\mathbf{x}) = \frac{1}{G}\sum_{g=1}^{G}\hat{F}_Y^{-1}(\tau_g|\mathbf{x})$, the quantile effect $\hat{\Delta}_\tau^{QE} = \hat{F}_Y^{c-1}(\tau) - \hat{F}_Y^{-1}(\tau)$, and the average effect $\bar{\Delta}^{QE} = \frac{1}{G}\sum_{g=1}^{G}(\hat{F}_Y^{c-1}(\tau_g) - \hat{F}_Y^{-1}(\tau_g))$. The variance of the quantile effect, for example, can be computed as $v\hat{a}r(\hat{\Delta}^{QE}) = \frac{1}{G}\sum_{g=1}^{G}(\hat{\Delta}_{\tau_g}^{QE} - \bar{\Delta}^{QE})^2$. The variance of other estimands follows similarly.

In computing counterfactual distributions and distribution effects, note that the $G \times M$ matrices $\boldsymbol{P}^o$, $\boldsymbol{P}^c$ and $\boldsymbol{\Delta}^{DE} = \boldsymbol{P}^c - \boldsymbol{P}^o$ that obtain using MCMC simulations of $\boldsymbol{\theta}_g,\ g = 1, ..., G$ from the BDR algorithm 1 constitute draws of y_θ^g, $y_\theta^{g,c}$ and Δ_g^{DE} at thresholds $\{y_g\}_{g=1}^G \subseteq \bar{\mathcal{Y}}$. The columns of $\boldsymbol{P}^o$, $\boldsymbol{P}^c$ and $\boldsymbol{\Delta}^{DE}$ constitute random draws of the distribution, its counterfactual, and the distribution effect, respectively, across thresholds $\{y_g\}_{g=1}^G$. The corresponding matrices of quantiles $\boldsymbol{Q}^o$ and $\boldsymbol{Q}^c$ obtain by inverting the columns of the distribution matrices and the random draws of quantile effects obtain as the $G \times M$ matrix $\boldsymbol{\Delta}^{QE} = \boldsymbol{Q}^c - \boldsymbol{Q}^o$. The rows correspond to thresholds $y_g \in \bar{\mathcal{Y}}$ and corresponding quantile indices $\tau_g \in (0, 1)$ while the columns represent random draws from the posterior. For example, a row in $\boldsymbol{\Delta}^{QE}$ represents random draws from the probability density function of the quantile effect at the corresponding τ'th index.

4 Bayesian Inference

Bayesian analysis offers much flexibility in summarising posterior inference. Though summaries of location (mean, median, mode(s)) and variation (standard deviation, interquartile range, etc.) are desirable and practical, summaries on posterior uncertainty are quite important (Gelman, Carlin, Stern, and Rubin, 1995, section 2.3). In this paper, we focus on confidence bands across thresholds that comprise $100(1 - \alpha)\%$ central intervals of posterior probability (or posterior intervals) which are directly interpretable as the $\alpha/2$ and $1-\alpha/2$ quantiles of the posterior density. An alternative to posterior intervals is the $100(1-\alpha)\%$ region of highest posterior density (HPD) but this has the drawback of not being necessarily invariant to transformations (Berger, 2013, section 4.3.2, example 6). Confidence bands enable the testing of several hypotheses including those not intended by the researcher.

Definition 1 (Bayesian confidence bands of $\mathbf{F}$).
Let $\mathbb{D}$ collect non-decreasing functions that map $\bar{\mathcal{Y}}$ into $[0,1]$[21] *and let $\mathbf{F} \in \mathbb{D}$ be a target distribution.*[22] *A confidence band $I = [L, U]$ collects intervals $I(y) = [L(y), U(y)]$, $L(y) \leq F(y) \leq U(y)\ \forall\ y \in \mathcal{Y}$. If I covers $\mathbf{F}$ with probability at least $(1-\alpha)$, $I = [L, U]$ is a confidence band of $\mathbf{F}$ of level $(1-\alpha)$.*

The above definition follows definition 1 in Chernozhukov, Fernandez-Val, Melly, and Wüthrich (2016, section 2.1). Note that in our Bayesian framework, point-wise confidence bands obtain by taking posterior intervals of $F_Y(y_g)$ (or of its counterfactual $F_{Y^c}(y_g)$) at each y_g, $g = 1, ..., G$. Note that the confidence band I may not be symmetric about $\mathbf{F}$. This occurs because Bayesian inference is exact and is based on the posterior distribution which may not be symmetric about $\mathbf{F}$ in small samples.

In the following theorem, we obtain adjustments to point-wise confidence bands from Bayesian posterior intervals in algorithm 1 to obtain confidence bands with simultaneous coverage of $\mathbf{F}$.

Theorem 1 (Simultaneous Bayesian confidence bands of $\mathbf{F}$).
Let $\mathbf{F} \in \mathbb{D}$ be a target distribution function that obtains point-wise from $\{\boldsymbol{\theta}_g\}_{g=1}^G$ in the BDR algorithm 1. $\mathbf{F}$ is simultaneously covered by
(a) $\quad I^* = [L^*, U^*] = [\mathbf{F} - c_{1-\alpha},\ \mathbf{F} + c_{1-\alpha}]$ *(symmetrically) and*
(b) $\quad I^* = [L^*, U^*] = [\mathbf{F} - \underline{c}_{\alpha/2},\ \mathbf{F} + \bar{c}_{1-\alpha/2}]$ *(asymmetrically)*
with probability at least $(1-\alpha)$ where $c_{1-\alpha} = q_{1-\alpha}(\max_y |F(y) - \hat{F}(y)|)$, $\underline{c}_{\alpha/2} = -q_{\alpha/2}(\min_y \hat{F}(y) - F(y))$, $\bar{c}_{1-\alpha/2} = q_{1-\alpha/2}(\max_y \hat{F}(y) - F(y))$, and $q_\tau(\cdot)$ denotes the τ'th quantile function.

Proof. See appendix A.1. $\square$

Note that $F(y)$ is the value of the target distribution $\mathbf{F}$ at y and $\hat{F}(y)$ denotes the value of a random draw of the distribution. The simultaneous confidence bands on BDR matrices $\boldsymbol{P}^o$, $\boldsymbol{P}^c$, $\boldsymbol{\Delta}^{DE}$, $\boldsymbol{Q}^o$,

[21] In the case of $\mathcal{Y}$, we consider it as a closed interval in the extended real line $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$. See Chernozhukov, Fernandez-Val, Melly, and Wüthrich (2016, section 2.1)

[22] Target distributions collect estimands (in increasing order) of location like the mean, median or mode at $\{y_g\}_{g=1}^G \subseteq \bar{\mathcal{Y}}$.

$\boldsymbol{Q}^c$, and $\boldsymbol{\Delta}^{QE}$ are easy to construct using theorem 1. For $\boldsymbol{Q}^o$, $\boldsymbol{Q}^c$, and $\boldsymbol{\Delta}^{QE}$, replace $y_g \in \bar{\mathcal{Y}}$ with $\tau \in (0,1)$.

In the following algorithm, we demonstrate the computation of the distribution (using the mean) and the $(1-\alpha)$ confidence band using theorem 1. The algorithm is applied in the same way to the other five BDR matrices to obtain both symmetric and asymmetric BDR confidence bands.

Algorithm 2 (Constructing BDR confidence bands of $\mathbf{F}$)**.**

1. *Compute the target distribution* $\mathbf{F} = [1/M \sum_{m=1}^{M} \hat{y}_{\theta,1}^{(m)}, ..., 1/M \sum_{m=1}^{M} \hat{y}_{\theta,G}^{(m)}]$ *by taking row-wise means of* $\boldsymbol{P}^o$*.*
2. *Subtract* $\mathbf{F}$ *from each column of* $\boldsymbol{P}^o$ *denote the resulting matrix as* $\mathbf{F}_\Delta$*.*
3. *Compute the absolute value of* $\mathbf{F}_\Delta$ *element-wise and obtain a vector of length* M *corresponding to the maximum of each column.* $c_{1-\alpha}$ *obtains as the* $(1-\alpha)$*'th quantile of this vector.*
4. *Compute the confidence bands using the definitions in theorem 1.*

Asymmetric confidence bands require a modification of step 3 in algorithm 2. Compute column-wise minima of $\mathbf{F}_\Delta$ and set $\underline{c}_{\alpha/2}$ to the negative of its $\alpha/2$'th quantile and set $\bar{c}_{1-\alpha/2}$ to the $(1-\alpha/2)$'th quantile of the column-wise maxima. Observe that confidence bands on the distribution effect can be used to test for first-order stochastic dominance. Higher-order stochastic dominance can also be tested but we do not pursue this further.

4.1 Parameters of interest (continued)

In the following results, we provide Bayesian inference on some point estimates. These estimates are means of their respective distributions. The median, mode, standard deviation, skewness, kurtosis etc. are other interesting point estimates which easily lend themselves to simultaneous inference (across thresholds or quantile indices) or point inference as outlined in result 1 below.

Result 1 (Posterior inference on point estimates)**.**

1. *The average distribution effect* $\bar{\Delta}^{DE} = 1/(GM) \sum_{g=1}^{G} \sum_{m=1}^{M} \Delta_g^{DE,(m)}$ *computed from the vector of random draws* $p_{\Delta^{DE}} = [1/G \sum_{g=1}^{G} \Delta_{g,1}^{DE}, ..., 1/G \sum_{g=1}^{G} \Delta_{g,M}^{DE}]$*.*
2. *The average effect* $\bar{\Delta}^{QE} = 1/(GM) \sum_{\tau=1}^{G} \sum_{m=1}^{M} \Delta_{\tau,m}^{QE}$ *obtains from the vector of random draws* $p_{\Delta^{QE}} = [1/G \sum_{\tau=1}^{G} \Delta_{\tau,1}^{QE}, ..., 1/G \sum_{\tau=1}^{G} \Delta_{\tau,M}^{QE}]$*.*

3. *A researcher may be interested in a posterior interval on the entire quantile or distribution effect. For example, a* $100(1-\alpha)\%$ *posterior interval on vectorised* $\mathbf{\Delta}^{QE}$*,i.e.* $vec(\mathbf{\Delta}^{QE})$ *can be interpreted as one in which the quantile effect at any* $\tau \in (0,1)$ *falls with probability* $1-\alpha$.

$100(1-\alpha)\%$ *posterior intervals of the average effect estimates obtain by taking the* $\alpha/2$ *and* $(1-\alpha/2)$*'th quantiles of random draws from their posterior densities.*

5 Asymptotic Inference

Large sample theory in Bayesian analysis is often not crucial for inference since Bayesian analysis provides distributions for direct inference. However, large sample results are useful and computationally convenient approximations. Some applications have used the normal approximations of posterior distributions especially when these are relatively more tractable. See Rubin and Schenker (1987), Agresti and Coull (1998) and Clogg et al. (1991) for examples.[23] To proceed, we make the following assumptions.

Assumption 1.

(a) *For each* $y_g \in \bar{\mathcal{Y}}$, $\boldsymbol{\theta}_g$ *is defined on a compact set* $\mathbf{\Theta}_o \subset \mathbb{R}^k$.

(b) *For each* $y_g \in \bar{\mathcal{Y}}$, $\Lambda : \mathbf{\Theta}_o \to \mathbb{R}$ *is twice continuously differentiable on* $\mathbf{\Theta}_o \subset \mathbb{R}^k$.

(c) *For each* $y_g \in \bar{\mathcal{Y}}$, $\boldsymbol{\theta}_g$ *is in the interior of* $\mathbf{\Theta}_o$.

The following theorem shows the convergence of the posterior distribution to the true distribution as $N \to \infty$.

Theorem 2 (Convergence of the posterior distribution)**.** - *Gelman, Carlin, Stern, and Rubin (1995, p. 587)*
Suppose assumption 1(a) holds in a neighbourhood $\mathcal{A}_g \subset \mathbf{\Theta}_o$ *with non-zero prior probability, then* $P(\boldsymbol{\theta}_g \in \mathcal{A}_g|\check{y}^g) \to 1$ *as* $N \to \infty$ *where* $\boldsymbol{\theta}_g$ *minimises the Kullback-Leibler information* $KL(\boldsymbol{\theta}_g) = \int \log\left(\frac{p(\check{y}^g)}{p(\check{y}^g|\boldsymbol{\theta}_g)}\right) p(\check{y}^g) d\check{y}^g$.

Proof. See Gelman, Carlin, Stern, and Rubin (1995, Appendix B) for a proof. □

[23] For example, Agresti and Coull (1998) shows that tests based on the score function perform better than exact ones in terms of coverage probabilities of the confidence interval. A Bayesian inference based on the score retains the advantage in a frequentist sense.

Theorem 2 shows that the collection of posterior distributions $\{p(\boldsymbol{\theta}_g|\breve{y}^g)\}_{g=1}^G$ converge to the true posterior distributions. In the next theorem, we show asymptotic normality of the posterior distributions.

Theorem 3 (Asymptotic normality of the posterior distribution)**.** - *Gelman, Carlin, Stern, and Rubin (1995, p. 587)*
Under standard regularity assumptions[24]*, the posterior distribution eq.* (2.3) *is asymptotically normal. The asymptotic distribution is*

$$p(\boldsymbol{\theta}_g|\breve{y}^g) = \mathcal{N}(\boldsymbol{\theta}_{o,g}, \mathcal{I}(\boldsymbol{\theta}_{o,g})^{-1}) \tag{5.1}$$

where $\mathcal{I}(\boldsymbol{\theta}_{o,g}) = N\sum_{i=1}^N \left(\frac{\exp(\mathbf{x}_i\boldsymbol{\theta}_{o,g})}{(1+\exp(\mathbf{x}_i\boldsymbol{\theta}_{o,g}))^2}\right)\mathbf{x}_i\mathbf{x}_i'$, $\boldsymbol{\theta}_{o,g} = \lim_{N\to\infty} \arg\max_{\boldsymbol{\theta}_g} p(\boldsymbol{\theta}_g|\breve{y}^g)$ *and* $\breve{y}^g$ *is dependent on the sample size* N*.*

Proof. This proof is fairly standard. See appendix A.1 for proof. □

Remark 2 (Robust variance alternative to $\mathcal{I}(\boldsymbol{\theta}_{o,g})^{-1}$)**.**
Distribution regression is a semi-parametric estimation method which uses a series of binary response models to approximate the conditional distribution without making an assumption about the correct specification of the binary response model. In that case, one may want to replace $\mathcal{I}(\boldsymbol{\theta}_{o,g})^{-1}$ *with a heteroscedasticity-robust variance matrix*

$$\boldsymbol{V}_g = \Big(\sum_{i=1}^N \mathbf{H}_i(\boldsymbol{\theta}_{o,g})\Big)^{-1}\Big(\sum_{i=1}^N \mathbf{s}_i(\boldsymbol{\theta}_{o,g})\mathbf{s}_i(\boldsymbol{\theta}_{o,g})'\Big)\Big(\sum_{i=1}^N \mathbf{H}_i(\boldsymbol{\theta}_{o,g})\Big)^{-1}$$

where $\mathbf{H}_i(\boldsymbol{\theta}_{o,g}) = \frac{d^2}{d\boldsymbol{\theta}_g}L(\boldsymbol{\theta}_g|\breve{y}_i^g)\Big|_{\boldsymbol{\theta}_g=\boldsymbol{\theta}_{o,g}}$ *and* $\mathbf{s}_i(\boldsymbol{\theta}_{o,g}) = \nabla_{\boldsymbol{\theta}_g}L(\boldsymbol{\theta}_g|\breve{y}_i^g)\Big|_{\boldsymbol{\theta}_g=\boldsymbol{\theta}_{o,g}}$

For a finite set $\{y_g\}_{g=1}^G \subseteq \bar{\mathcal{Y}}$, the set of corresponding posterior distributions obtains as $\{p(\boldsymbol{\theta}_g|\breve{y}^g)\}_{g=1}^G$. The approximated posterior eq. (A.5) can be used to compute $p(y_\theta^g|\boldsymbol{\theta}_g)$ in eq. (2.4) and $p(\Delta_g^{DE}|\boldsymbol{\theta}_g)$ in eq. (2.6). This process is computationally faster than MCMC because the approximated multivariate normal eq. (A.5) density is analytical and easier to sample from.

In the following theorem and corollaries, we push the preceding asymptotic results further in order to obtain closed-form expressions for the (joint) densities of the outcome distributions and the distribution effects.

[24]See theorems 12.3 and 13.2 in Wooldridge (2010) for regularity conditions and asymptotic normality results for M- and Maximum likelihood estimators.

Theorem 4 (Asymptotic distribution of $\hat{y}_{\theta}^{g} = \hat{F}_Y(y_g|\hat{\boldsymbol{\theta}}_g)$)**.**
Let assumption 1(b) & assumption 1(c) hold, using results in theorems 2 and 3, the asymptotic distribution of $\hat{F}_Y(y_g|\hat{\boldsymbol{\theta}}_g)$ *is normal.*

$$\hat{F}_Y(y_g|\hat{\boldsymbol{\theta}}_g) \sim \mathcal{N}(F_{y_g}, \mathcal{V}_{F_{y_g}}) \tag{5.2}$$

where $F_{y_g} = \int F_{Y|X}(y_g|\mathbf{x})dF(\mathbf{x}) = \int \Lambda(\mathbf{x}\boldsymbol{\theta}_{o,g})dF(\mathbf{x})$, $\mathcal{V}_{F_{y_g}} = E[\boldsymbol{\lambda}_i^{\mathbf{x}}(\boldsymbol{\theta}_{o,g})'[I(\boldsymbol{\theta}_{o,g})]^{-1}\boldsymbol{\lambda}_i^{\mathbf{x}}(\boldsymbol{\theta}_{o,g})]$, *and* $\boldsymbol{\lambda}_i^{\mathbf{x}}(\boldsymbol{\theta}_{o,g}) = \Lambda'(\mathbf{x}_i\boldsymbol{\theta}_{o,g})\mathbf{x}_i'$.

Proof. See appendix A.1. $\square$

Theorem 4 applies to the conditional and counterfactual distributions as well, noting that $F_{y_g}^c = \int F_{Y|X}(y_g|\mathbf{x}^c)dF(\mathbf{x^c}) = \int \Lambda(\mathbf{x}\boldsymbol{\alpha}'\boldsymbol{\theta}_{o,g})dF(\mathbf{x})$ where $\boldsymbol{\alpha}$ is a $k \times k$ diagonal matrix that multiplicatively creates a counterfactual of $\mathbf{x}$.[25] In the following corollary, results in theorem 4 are extended to the joint distribution at thresholds $\{y_g\}_{g=1}^G \subseteq \bar{\mathcal{Y}}$.

Corollary 1 (Joint asymptotic distribution of $\{\hat{F}_Y(y_g|\hat{\boldsymbol{\theta}}_{o,g})\}_{g=1}^G$)**.**
Extending results from theorem 4 to the joint distribution at several indices $\{y_g\}_{g=1}^G \subseteq \bar{\mathcal{Y}}$, *the joint asymptotic distribution of* $\hat{\mathbf{F}}_Y = [\hat{F}_Y(y_1|\hat{\boldsymbol{\theta}}_1), ..., \hat{F}_Y(y_G|\hat{\boldsymbol{\theta}}_G)]'$ *is joint normally distributed with*

$$\hat{\mathbf{F}}_Y \sim \mathcal{N}(\mathbf{F}_Y, \boldsymbol{\Omega}_{F_{y_g}}) \tag{5.3}$$

where $\mathbf{F}_Y = [F_Y(y_1|\boldsymbol{\theta}_{o,1}), ..., F_Y(y_G|\boldsymbol{\theta}_{o,G})]'$,
$\boldsymbol{\Omega}_{F_{y_g}}$*'s* (g,g)*'th element* $\mathcal{V}_{F_{y_g}} = E[\boldsymbol{\lambda}_i^{\mathbf{x}}(\boldsymbol{\theta}_{o,g})'[I(\boldsymbol{\theta}_{o,g})]^{-1}\boldsymbol{\lambda}_i^{\mathbf{x}}(\boldsymbol{\theta}_{o,g})]$, *and* (g,h)*'th element* $\mathcal{V}_{F_{y_{g,h}}} = E[\boldsymbol{\lambda}_i^{\mathbf{x}}(\boldsymbol{\theta}_{o,g})'[I(\boldsymbol{\theta}_{o,g})]^{-1}I_i(\boldsymbol{\theta}_{g,h})[I(\boldsymbol{\theta}_{o,h})]^{-1}\boldsymbol{\lambda}_i^{\mathbf{x}}(\boldsymbol{\theta}_{o,h})]$, $I_i(\boldsymbol{\theta}_{g,h}) = N^{-1}\mathbf{s}_i(\boldsymbol{\theta}_{o,g})\mathbf{s}_i(\boldsymbol{\theta}_{o,h})'$.

Proof. See appendix A.1. $\square$

The asymptotic distribution of the distribution effect $\hat{\Delta}_g^{DE} = \hat{F}_Y(y_g) - \hat{F}_Y^c(y_g)$ at a threshold $y_g \in \bar{\mathcal{Y}}$ follows from theorem 4 because the theorem also applies to the counterfactual distribution. In the following corollaries, we show the asymptotic distribution of the distribution effect.

Corollary 2 (Asymptotic distribution of $\hat{\Delta}_{y_g}^{DE}$)**.**
Let assumption 1(b) and assumption 1(c) hold. Using results in theorems 2 and 3, the distribution effect at a threshold $y_g \in \mathcal{Y}$, $\hat{\Delta}_g^{DE} = \hat{F}_Y^c(y_g) - \hat{F}_Y(y_g)$ *is normally distributed,* $\hat{\Delta}_g^{DE} \sim$

[25] Suppose the second covariate is the treatment variable. A counterfactual treatment level of 5% increment replaces the second diagonal element in a $k \times k$ identity matrix with 1.05. Additive counterfactual treatment requires an observation-specific $\boldsymbol{\alpha}_i$.

$\mathcal{N}(\Delta_g^{DE}, \mathcal{V}_{\Delta_g^{DE}})$ *where* $\Delta_g^{DE} = \int(\Lambda(\mathbf{x}\boldsymbol{\alpha}'\boldsymbol{\theta}_{o,g}) - \Lambda(\mathbf{x}\boldsymbol{\theta}_{o,g}))dF(\mathbf{x})$, $\mathcal{V}_{\Delta_g^{DE}} = E[\bar{\boldsymbol{\lambda}}_i^{\mathbf{x}}(\boldsymbol{\theta}_{o,g})[I(\boldsymbol{\theta}_{o,g})]^{-1}\bar{\boldsymbol{\lambda}}_i^{\mathbf{x}}(\boldsymbol{\theta}_{o,g})']$, $\bar{\boldsymbol{\lambda}}_i^{\mathbf{x}}(\boldsymbol{\theta}_{o,g}) = (\Lambda'(\mathbf{x}_i\boldsymbol{\alpha}'\boldsymbol{\theta}_{o,g})\boldsymbol{\alpha} - \Lambda'(\mathbf{x}_i\boldsymbol{\theta}_{o,g})\mathbf{I}_k)\mathbf{x}_i'$ *and* $\boldsymbol{\alpha}$ *is a* $k \times k$ *diagonal matrix that multiplicatively creates a counterfactual of* $\mathbf{x}$ *and* $\mathbf{I}_k$ *is a* $k \times k$ *identity matrix.*

Proof. See appendix A.1. □

The asymptotic result in corollary 2 is extensible to several indices $\{y_g\}_{g=1}^G \subseteq \bar{\mathcal{Y}}$ using similar arguments in corollary 1 to establish joint asymptotic normality.

Corollary 3 (Joint asymptotic distribution of $\hat{\boldsymbol{\Delta}}^{DE}$)**.**
Extending results from corollary 2 to the joint distribution at several indices $\{y_g\}_{g=1}^G \subseteq \bar{\mathcal{Y}}$*, the joint asymptotic distribution of* $\hat{\boldsymbol{\Delta}}^{DE} = [\hat{\Delta}_1^{DE}, ..., \hat{\Delta}_G^{DE}]'$ *is joint normally distributed with*

$$\hat{\boldsymbol{\Delta}}^{DE} \xrightarrow{d} \mathcal{N}(\boldsymbol{\Delta}^{DE}, \boldsymbol{\Omega}_\Delta) \tag{5.4}$$

where

the (g, g)*'th element of* $\boldsymbol{\Omega}_\Delta$ *is* $E[\bar{\boldsymbol{\lambda}}_i^{\mathbf{x}}(\boldsymbol{\theta}_{o,g})'[I(\boldsymbol{\theta}_{o,g})]^{-1}\bar{\boldsymbol{\lambda}}_i^{\mathbf{x}}(\boldsymbol{\theta}_{o,g})]$ *and*

the (g, h)*'th element of* $\boldsymbol{\Omega}_\Delta$ *is* $E[\bar{\boldsymbol{\lambda}}_i^{\mathbf{x}}(\boldsymbol{\theta}_{o,g})'[I(\boldsymbol{\theta}_{o,g})]^{-1}I_i(\boldsymbol{\theta}_{g,h})[I(\boldsymbol{\theta}_{o,h})]^{-1}\bar{\boldsymbol{\lambda}}_i^{\mathbf{x}}(\boldsymbol{\theta}_{o,h})]$

Proof of corollary 3 .
See appendix A.1. □

Remark 3 (BDR estimators)**.**
There are three possibilities of carrying out Bayesian distribution regression, the first (non-asymptotic BDR) involves MCMC draws (see algorithm 1), a second (semi-asymptotic BDR) that samples $\boldsymbol{\theta}_g$*,* $g = 1, ..., G$ *from asymptotically approximated (normal) posterior distributions (in lieu of MCMC) from Theorem 3 at* y_g*,* $g = 1, ..., G$ *separately, and a third (asymptotic BDR) that uses a (joint) asymptotic distributions (see theorem 4, and corollaries 1 to 3).*

The first one is fairly exact even in small samples whereas the second and third are based on asymptotic approximations. While the non-asymptotic BDR can be deemed purely Bayesian, Bayesian inference on the Semi-asymptotic draws on Bayesian asymptotic theory for $\boldsymbol{\theta}_g$ at thresholds y_g, $g = 1, ..., G$ separately, and the asymptotic BDR involves a joint distribution of $\boldsymbol{\theta}_g$ across all thresholds. Inference results in section 4 on quantile and quantile effect functions hold for all three estimators by inverting draws of the distribution function (see algorithms 1 and 2 with accompanying notes). In addition, a recent paper Montiel Olea and Plagborg-Møller (2018) presents a framework (see algorithm 2 in the paper and next section, section 6) to obtain calibrated simultaneous confidence bands. We consider our the proposed confidence bands in theorem 1 as easier-to-compute alternatives to Montiel Olea and Plagborg-Møller (2018)'s simultaneous confidence bands.

6 Monte Carlo Simulation

This section enables a small sample evaluation of the three BDR estimators. The main goal is to demonstrate the performance of point and functional estimates derived from all three estimators. For this section and the empirical application, we use the logistic link function.

500 Monte Carlo simulations are carried out with three estimators, non-asymptotic, semi-asymptotic and asymptotic BDR for the distribution function $F_Y(y), y \in \mathcal{Y}$. We use the logistic link function and employ the uniform prior. The posterior distribution obtains as

$$p(\check{y}^g|\boldsymbol{\theta}_g) = \prod_{i=1}^{N} p(\check{y}_i^g|\boldsymbol{\theta}_g) = \prod_{i=1}^{N} \frac{\exp(\mathbf{x}_i\boldsymbol{\theta}_g \mathbb{1}\{y_i \leq y_g\})}{1+\exp(\mathbf{x}_i\boldsymbol{\theta}_g)} \tag{6.1}$$

For each estimator, $G = 21$ threshold values $y_g, g = 1, ..., G$ are used with 500 (600 in all, 100 as burn-in) posterior draws of $\boldsymbol{\theta}_g$. The location-shift form is used to generate the outcome with $y_i = \max\{\mathbf{x}_i\boldsymbol{\beta} + 0.05(\mathbf{x}_i\boldsymbol{\beta})\Lambda^{-1}(u_i), 0\}$, $u_i \sim U(0,1)$, $\Lambda(\cdot)$ is the logistic cumulative distribution function, $\boldsymbol{\beta} = [0.5, 0.2, 0.8]'$ (the first element is the intercept), and the last two elements of $\mathbf{x}_i$ are generated from the bi-variate normal $\mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{v}_x)$, $\boldsymbol{\mu}_x = [2, 0.5]$, and $\boldsymbol{v}_x$ is symmetric with 1 on the diagonal and 0.1 off-diagonal.

Results of the Monte Carlo exercise are summarised in the following three tables. Table 1 presents results on the distribution function in terms of bias, the second (table 2) in terms of coverage of point estimands by posterior intervals and the third (table 3) in terms of coverage by confidence bands of the distribution function across all three estimators.

Table 1: Bias of point estimates

	Mean Bias			MSE	MAE
Estimator	Mean	MAD	RMSE	Mean	Mean
non-asymptotic	-0.0003	0.0053	0.0074	0.0001	0.0175
semi-asymptotic	0.0190	0.0188	0.0205	0.0009	0.0537
asymptotic	-0.0003	0.0053	0.0074	0.0001	0.0176

Notes: Number of simulations 500, number of posterior draws: 500, sample size: 2000. The first three columns report the mean, the median absolute and the root mean square of mean biases $MeanBias = G^{-1}\sum_{g=1}^{G}(\hat{F}(y_g) - F(y_g))$ across simulations. The last two report the means of the mean squared error ($MSE = G^{-1}\sum_{g=1}^{G}(\hat{F}(y_g) - F(y_g))^2$) and maximum absolute error $MAE = \max_{y_g \in \bar{\mathcal{Y}}} |\hat{F}(y_g) - F(y_g)|$ across simulations.

Table 2: Coverage of posterior intervals (%)

	95% Posterior Interval				
Estimator	$MnCI$ (%)	$ptw_{CB}CI$ (%)	$sym_{CB}CI$ (%)	$asym_{CB}CI$ (%)	$OM_{CB}CI$ (%)
non-asymptotic	11.4	44.0	76.6	76.0	100.0
semi-asymptotic	59.2	80.4	100.0	100.0	100.0
asymptotic	56.0	99.8	100.0	100.0	100.0

Notes: Number of simulations: 500, number of posterior draws: 500, sample size: 2000. The table presents the coverage (in percentages) of the mean distribution $\bar{F} = G^{-1}\sum_{g=1}^{G} F(y_g),\ y_g \in \bar{\mathcal{Y}},\ G = 21$, by posterior intervals. The posterior intervals considered are: $MnCI$ - 95% posterior interval of $\bar{F}$, $ptw_{CB}CI$ - averaged symmetric simultaneous confidence bands, $asym_{CB}CI$ - averaged asymmetric simultaneous confidence bands, and $OM_{CB}CI$ - averaged Montiel Olea and Plagborg-Møller (2018) simultaneous confidence bands.

Table 3: Coverage of confidence bands (%)

	95% Confidence Bands			
Estimator	ptw CB (%)	sym CB (%)	asym CB (%)	OM CB (%)
non-asymptotic	0	2.8	3.4	100
semi-asymptotic	0	100	77.2	91.8
asymptotic	25.4	99.8	99.8	99

Notes: Number of simulations: 500, number of posterior draws: 500, sample size: 2000. The table presents the coverage (in percentages) of the distribution function $F(y_g),\ y_g \in \bar{\mathcal{Y}},\ g = 1, ..., G$, by confidence bands. The confidence bands considered are: ptw CB - point-wise confidence bands, sym CB - symmetric simultaneous confidence bands, asym CB - asymmetric simultaneous confidence bands, and OM CB - Montiel Olea and Plagborg-Møller (2018) simultaneous confidence bands.

Table 1 shows that the non-asymptotic BDR performs best across all bias criteria but only slightly vis-à-vis the asymptotic BDR in terms of the maximum absolute error $MAE = \max_{y_g \in \bar{\mathcal{Y}}} |\hat{F}(y_g) -$

$F(y_g)|$. Both estimators (non-asymptotic and asymptotic BDR) perform better than the semi-asymptotic in terms of all criteria. Tables 2 and 3 show that non-calibrated symmetric and asymmetric Bayesian simultaneous confidence bands used with the non-asymptotic BDR has low coverage of the distribution function. Coverage improves with the semi-asymptotic and asymptotic BDR for all simultaneous confidence bands. As generally expected, point-wise confidence bands perform poorly irrespective of the estimator.

7 Empirical Application: impact of institutional ownership on innovation

7.1 A literature review

The relevance of ownership by institutions (eg. banks, pension funds, venture funds, private equity funds, labour unions, or insurance companies) for firm innovation has gained attention in the industrial organisation and corporate finance literature (see Bushee (1998), Eng and Shackell (2001), Francis and Smith (1995), Hall and Lerner (2010), Aghion, Van Reenen, and Zingales (2013), and Berger, Stocker, and Zeileis (2017)). Aghion, Van Reenen, and Zingales (2013), a notable work, emphasise the importance of corporate governance on innovation in publicly traded firms which account for a significant proportion of private research and development (R&D) expenditure. While there abounds literature on the impact of financing constraints on research and development (R&D) and the mitigation of information asymmetry related to R&D activities of firms,[26] there is a dearth of studies on the impact of corporate governance on R&D.

A generally used measure of innovation is R&D expenditure which simply measures quantity of R&D input and may not reflect productivity and success of R&D activity adequately (Aghion, Van Reenen, and Zingales, 2013). To this end, Aghion, Van Reenen, and Zingales (2013) proxy innovation using future citation-weighted patent counts. Citation-weighted patent counts is a fairly reliable measure of innovation because it not only captures quantity (number of patents) but also relevance (citation).[27] After controlling for confounding factors and fixed effects, the Aghion, Van Reenen, and Zingales (2013) and Berger, Stocker, and Zeileis (2017) find a positive (and generally significant) impact of institutional ownership on innovation.

[26]See Aghion, Van Reenen, and Zingales (2013, p. 278) and references therein.

[27]Trajtenberg (1990), Berger, Stocker, and Zeileis (2017), Bloom, Lucking, and Van Reenen (2018), and Hirshleifer, Low, and Teoh (2012) are among works that use measures of innovation and productivity based on patent and citation counts.

There have been previous attempts to examine the impact of institutional investment on R&D and innovation. Bushee (1998) finds that institutional investors immunise managers against myopic behaviour, e.g. cutting R&D investment in order to counter a decline in earnings. On the impact of ownership on innovation, Francis and Smith (1995) finds greater innovation in firms with more concentrated ownership (mostly institutional investors). Also, Eng and Shackell (2001) finds a positive association between institutional ownership and R&D spending.

A note on the econometric modelling employed in the literature is in order. Aghion, Van Reenen, and Zingales (2013) employs ordinary least squares (on the logarithm of citation-weighted patent counts thereby excluding zeroes), the poisson and negative binomial models (on citation-weighted patent counts) to quantify the impact of institutional ownership on innovation. In a bid to retain observations with zero (citation-) patent counts, Hirshleifer, Low, and Teoh (2012) transforms outcome variables patent and citation-weighted patent counts using $log(1 + outcome)$ in a study that finds a positive effect of CEO overconfidence on firm innovation. Berger, Stocker, and Zeileis (2017) extends the poisson model in Aghion, Van Reenen, and Zingales (2013) to a two-part (zero and count) to account for differences in the determinants of innovating versus increasing the level of innovation.

In our application, we argue that Bayesian distribution regression does present a more robust alternative to econometric models in the aforementioned empirical studies among others. Firstly, the use of parametric models, viz. poisson, negative binomial, the hurdle model, tobit model, Heckman selection models, etc., to deal with the large number of zeroes in the outcome variable may be inadequate given the often restrictive assumptions that underlie their validity. Secondly, taking logarithms of non-negative outcome variables leads to the exclusion of observations with zero outcome. Logarithmic transformation does change results and even invalidate results. The poisson model, for example, supposes that determinants of the first patent-citation are identical for subsequent levels of patent-citations. We note, however, that a two-part model may fail to account for other mass points in the count part of the outcome. In fact, the hurdle model involves parametric assumptions which can lead to inconsistent results if there is misspecification of the conditional density and conditional mean in either or both parts. Also, a large number of zeroes creates a huge mass point in the outcome variable that is not adequately accounted for by the poisson or negative binomials. For example, the poisson model requires variance-mean equality which is hardly satisfied in applications (see table 4 below). Distribution regression allows flexible modelling of the conditional distribution at different points on the support. Coupled with Bayesian inference, recourse to asymptotic approximations for inference is not sine qua non and asymmetries in posterior summaries viz. confidence bands and posterior intervals can provide useful insights beyond symmetric asymptotic inference.

7.2 Data

We use data from Aghion, Van Reenen, and Zingales (2013) to study the impact of institutional ownership on innovation. The outcome variables are $Cites$ (future citation-weighted patent counts) and its logarithm $ln(Cites)$. Covariates include Share_Inst, the covariate of interest that measures the percentage of outstanding shares held by institutions, log of R&D stocks $ln(R\&DStock)$, log of capital-labour ratio $ln(K/L)$, the log of sales $ln(Sales)$, pre-sample mean scaling fixed effects of Blundell, Griffith, and Van Reenen (1999), four-digit industry dummies and year dummies.

Table 4: Summary statistics

	$Cites$	$ln(Cites)$	$Share\_Inst$	$ln(R\&DStock)$	$ln(K/L)$	$ln(Sales)$
Min.	0	0.000	0.00	-5.266	1.941	-3.963
1st Qu.	0	1.946	27.38	1.657	3.867	5.012
Median	7	3.332	48.17	3.792	4.306	6.410
Mean	176.3	3.469	45.53	3.498	4.438	6.365
3rd Qu	51	4.727	63.80	5.273	4.870	7.770
Max.	23121	10.049	99.99	10.689	8.399	12.071
Std. Dev.	923.264	1.980	23.051	2.693	0.866	2.042

Note: The table shows summary statistics for the main variables. $Cites$ and $ln(Cites)$ - the outcome variables are the citation-weighted patent counts and its logarithm, $Share\_Inst$ - percentage of outstanding shares owned by institutions (the treatment covariate of interest), $ln(R\&DStock)$ - logarithm of R&D stock, $ln(K/L)$ - logarithm of capital-labour ratio, and $ln(Sales)$ - logarithm of sales.

The variance-mean ratio of citation-weighted patent counts $\frac{var(Cites)}{mean(Cites)} = 4836.2$ illustrates substantial over-dispersion which cannot be accommodated by the poisson or negative binomial model. Such over-dispersion makes DR and QR (quantile regression) methods particularly useful since DR, for example, flexibly estimates the (conditional) distribution of the outcome variable at different points on its support without imposing strong distributional assumptions. Previous attempts in the literature include ordinary least squares[28] (see Aghion, Van Reenen, and Zingales (2013), Bloom, Schankerman, and Van Reenen (2013)) and nonlinear parametric models viz. poisson regression, negative binomial (see Aghion, Van Reenen, and Zingales (2013), Bloom, Schankerman, and Van Reenen (2013)), hurdle models (Berger, Stocker, and Zeileis (2017)), etc.

In a revisit to the problem studied by Aghion, Van Reenen, and Zingales (2013) using count

[28]OLS results in columns 1 and 2 of Aghion, Van Reenen, and Zingales (2013) obtains after taking the log of $Cites$ and dropping observations with zero cites.

data hurdle models, Berger, Stocker, and Zeileis (2017) notes that unlike the poisson model, the two hurdle parts (zero and count parts) do not coincide. This is empirical evidence against the suitability of the poisson model. On the hurdle model of Berger, Stocker, and Zeileis (2017), we note that while allowing flexibility at the zero-count threshold, modelling zero and count parts using a truncated (from the left) binomial model and a censored negative binomial respectively, disallows flexibility at other (possibly mass) points on the count part of the outcome.

7.3 Results

Aghion, Van Reenen, and Zingales (2013) considers ordinary least squares, poisson and negative binomial models (see table 1 in that paper). Table 5 shows the (successful) replication of results. For the purpose of illustrating the BDR empirically, we employ the same set of variables in order to make our results comparable to the OLS (column 2), poisson (column 4), and negative binomial (column 7) of table 1 in Aghion, Van Reenen, and Zingales (2013) and hurdle negative binomial models (see Hurdle Negbin (1) and Hurdle Negbin(2) of table 2 in Berger, Stocker, and Zeileis (2017)).

Table 5: Replicated Regression results

Method	OLS	Poisson	Neg. Bin
Outcome	$ln(Cites)$	$Cites$	$Cites$
variable	(1)	(2)	(3)
Share_Inst	0.0055***	0.007***	0.0058***
ln(R&D Stock)	0.3371***	0.009***	0.1776***
ln(K/L)	0.2611***	0.4401***	0.2644***
ln(Sales)	0.3099***	0.1839***	0.1271***
Observations	4025	6208	6208

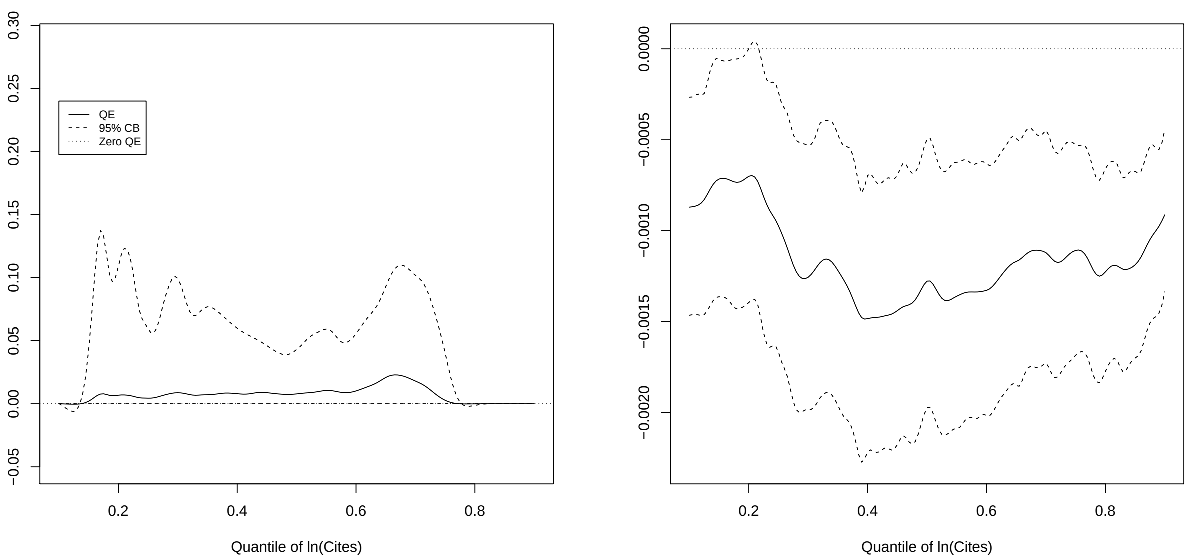
Note: The table above presents a replication of the linear, poisson, and negative binomial regression results of Aghion, Van Reenen, and Zingales (2013, Table I, columns 2, 5 & 8). Number of firms: 803, estimation period: 1991-1999. All regressions include four-digit industry dummies, time dummies and fixed effects à la Blundell, Griffith, and Van Reenen (1999). *** denotes significance at the 1% level.

From table 5, an increase in institutional ownership by one percentage point increases innovation by 0.55 percent. The OLS, however, involves dropping observations with zero $Cites$. The poisson and negative binomial models do allow for observations with zero $Cites$. Coefficients on Share_Inst from the hurdle negative binomial approach in Berger, Stocker, and Zeileis (2017, table 2, last two columns) are 0.003 for the count part and 0.009 for the zero part.[29]

[29]The coefficients are significant at 10% and 1% levels of significance respectively.

The difference in magnitude and significance of the coefficients on institutional ownership in both parts of the hurdle model leads Berger, Stocker, and Zeileis (2017) to conclude that a single equation specification like the poisson or negative binomial may not suffice. In our empirical application, we introduce further flexibility by using distribution regression that models the outcomes ($Cites$ and $ln(Cites)$) at different threshold values on their respective support. This approach obviates restrictive distributional assumptions which can lead to misleading conclusions.

In order to make our results comparable to those in Aghion, Van Reenen, and Zingales (2013) and Berger, Stocker, and Zeileis (2017), we generate the counterfactual level of treatment[30] by increasing institutional ownership by 1%. Details of the implementation of Bayesian Distribution Regression and other results are relegated to appendix C.



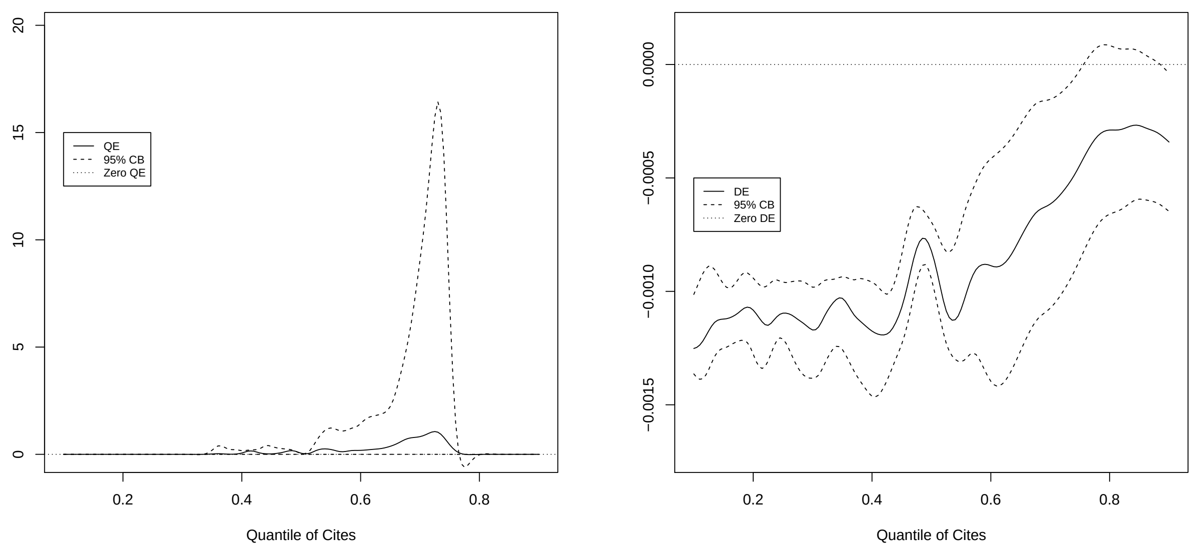
(a) Quantile effects (b) Distribution effects

Figure 1: Quantile and distribution treatment effects, $ln(Cites)$, point-wise CB

Notes: The figures above plot quantile and distribution effects with 95% point-wise confidence bands. The point-wise confidence bands comprise posterior intervals at different values of $ln(Cites)$, i.e $\alpha/2$ and $1-\alpha/2$ quantiles of quantile and distribution effects, $\alpha = 0.05$. To obtain point-wise confidence bands of quantile effects, the Bayesian distribution matrix and its counterfactual from algorithm 1 are inverted column-wise and a quantile effect matrix obtains as the difference. The row-wise $\alpha/2$ and $1-\alpha/2$ quantiles of the quantile effect matrix constitute the point-wise confidence bands in Figure 1a. The mean distribution effect is $-0.00116 \in [-0.00121, -0.00111]$ with probability 0.95 and the average effect is $0.00723 \in [0.00426, 0.01067]$ with probability 0.95. Distribution effect (at any threshold $y_g \in \bar{\mathcal{Y}}$) lies in the posterior interval $[-0.00196, -0.00034]$ while the quantile effect (at any index $\tau \in [0.1, 0.9]$ of $ln(Cites)$) falls in the posterior interval $[0, 0.06899]$ with probability 0.95.

[30]Chernozhukov, Fernandez-Val, and Weidner (2018, section 2.2) states ways of generating counterfactual treatments based on the scale of treatment.

Distributions and quantiles are negatively related. Not surprising, we find significantly positive quantile effects on $ln(Cites)$ (see Figure 1a) on most of the support of $ln(Cites)$ and zero quantile effect in the tails. On average, an increase in institutional ownership by 1% increases citation-weighted patents by 0.7%. The average quantile effect $0.007 \in [0.00426, 0.01067]$ does not[31] exclude Aghion, Van Reenen, and Zingales (2013)'s estimate of 0.0055 (see table 5, column (1)). At its maximum (the 66'th quantile of $ln(Cites)$), a 1% increase in institutional ownership leads to a 2.54% rise in innovation. The entirety of the quantile effect on $ln(Cites)$ is contained in the interval $[0, 0.06899]$ with probability 95%.



(a) Quantile effect

(b) Distribution effect

Figure 2: Quantile and Distribution effects, $Cites$, point-wise CB

Notes: The distribution effect and quantile effect across thresholds are row-wise means of $\mathbf{\Delta}^{DE}$ and $\mathbf{\Delta}^{QE}$ (see section 3.2) with (discrete) $Cites$ as outcome variable. The confidence bands comprise point-wise 95% posterior intervals. The average effect (mean quantile effect) is $0.1482 \in [0.03727, 0.33970]$ with probability 0.95 while that of the distribution effect is $-0.00086 \in [-0.00089, -0.00082]$ with probability 0.95. The distribution effect (at any threshold $y_g \in \bar{\mathcal{Y}}$ of $Cites$) is contained in the interval $[-0.00151, -0.00009]$ with probability 0.95 while the quantile effect (at any quantile index $\tau \in [0.1, 0.9]$ of $Cites$) is contained in the posterior interval $[0, 1]$ with probability 0.95.

On average, increasing institutional ownership by 1% increases citation-weighted patent counts $Cites$ by $0.1482 \in [0.03727, 0.33970]$ (see Figure 2). Note that the 95% posterior interval of the average effect excludes values (ranging from 0.003 to 0.009) obtained by Aghion, Van Reenen, and Zingales (2013) and Berger, Stocker, and Zeileis (2017). The immediate implication is that

[31] The notation $x \in [\underline{x}, \bar{x}]$ means estimate x falls in the (closed) posterior interval with the indicated probability.

parametric assumptions on the distribution of $Cites$ are unable to uncover the highly significant impact (economically and statistically) of institutional ownership that we obtain. At its maximum (the 73rd quantile of $Cites$), an increment in institutional ownership by 1% leads to an increase in citation-weighted patent counts $Cites$ by 1.23.

The quantile effect (at any threshold) is contained in the interval $[0, 1]$ while the distribution effect is contained in the posterior interval $[-0.0015, -0.0001]$ with a 0.95 probability. The interpretation of both distribution and quantile effects rules out a negative impact of institutional ownership at any threshold on the support of $Cites$. The quantile and average effects on $Cites$ represent a more economically significant impact (see table 5 columns (2) and (3)) relative to 0.007 and 0.0058 (for poisson and negative binomial models respectively) of Aghion, Van Reenen, and Zingales (2013). It is clear that the huge mass point at zero in $Cites$ is not sufficiently accommodated by the poisson and negative binomial models.

Though our confidence bands do not exclude results obtained by Aghion, Van Reenen, and Zingales (2013) and Berger, Stocker, and Zeileis (2017), we note that the estimated effect of institutional ownership changes drastically by including observations with zero $Cites$. In addition, confidence bands constructed from posterior intervals show substantial skewness in quantile and distribution effects across thresholds on the support of $Cites$ and $ln(Cites)$. Also, we note that in the case of $ln(Cites)$ and $Cites$, the impact of institutional ownership is relatively low for firms with zero or fairly low citation-weighted patent counts. Institutional ownership has maximum (positive) impact on innovation for greater-than-median level patent-count firms but sharply declines to zero in the upper tail.

8 Conclusion

In sum, we introduce a Bayesian approach to distribution regression by leveraging the likelihood function of binary response models at a grid of points on the support of the outcome. Bayesian distribution regression can be useful to study various research questions related to quantiles, distributions and counterfactual effects. We provide three BDR estimators, with posterior intervals, asymmetric and symmetric simultaneous confidence bands for conducting inference. Our empirical application, which studies the impact of institutional ownership on innovation (as studied by Aghion, Van Reenen, and Zingales (2013)), shows that the presence of huge mass points at zero leads to an underestimation of the effect when parametric models like the poisson, negative binomial and hurdle models are used. Besides, we uncover substantial heterogeneity in distribution and quantile effects. Bayesian confidence bands on quantile effects reveal non-trivial asymmetry.

References

[1] Aghion, Philippe, John Van Reenen, and Luigi Zingales. “Innovation and institutional ownership”. *American economic review* 103.1 (2013), pp. 277–304.

[2] Agresti, Alan and Brent A Coull. “Approximate is better than “exact”? for interval estimation of binomial proportions”. *The American Statistician* 52.2 (1998), pp. 119–126.

[3] Albert, Jim. *Bayesian computation with R*. Springer Science & Business Media, 2009.

[4] Barrett, Garry F and Stephen G Donald. “Statistical inference with generalized Gini indices of inequality, poverty, and welfare”. *Journal of Business & Economic Statistics* 27.1 (2009), pp. 1–17.

[5] Berger, James O. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

[6] Berger, Susanne, Herbert Stocker, and Achim Zeileis. “Innovation and institutional ownership revisited: an empirical investigation with count data models”. *Empirical Economics* 52.4 (2017), pp. 1675–1688.

[7] Bloom, Nicholas, Mark Schankerman, and John Van Reenen. “Identifying technology spillovers and product market rivalry”. *Econometrica* 81.4 (2013), pp. 1347–1393.

[8] Bloom, Nick, Brian Lucking, and John Van Reenen. “Have R&D spillovers changed?” *Working paper* (2018).

[9] Blundell, Richard, Rachel Griffith, and John Van Reenen. “Market share, market value and innovation in a panel of British manufacturing firms”. *The Review of Economic Studies* 66.3 (1999), pp. 529–554.

[10] Bushee, Brian J. “The influence of institutional investors on myopic R&D investment behavior”. *Accounting review* (1998), pp. 305–333.

[11] Callaway, Brantly and Weige Huang. “Counterfactual Analysis of Intergenerational Income Mobility”. Working paper. 2017.

[12] Chernozhukov, Victor, Iván Fernández-Val, and Alfred Galichon. “Quantile and probability curves without crossing”. *Econometrica* 78.3 (2010), pp. 1093–1125.

[13] Chernozhukov, Victor, Iván Fernández-Val, Sukjin Han, and Amanda Kowalski. *Censored quantile instrumental variable estimation with Stata*. Tech. rep. National Bureau of Economic Research, 2018.

[14] Chernozhukov, Victor, Iván Fernández-Val, and Blaise Melly. "Inference on counterfactual distributions". *Econometrica* 81.6 (2013), pp. 2205–2268.

[15] Chernozhukov, Victor, Ivan Fernandez-Val, Blaise Melly, and Kaspar Wüthrich. "Generic inference on quantile and quantile effect functions for discrete outcomes". *arXiv preprint arXiv:1608.05142* (2016).

[16] Chernozhukov, Victor, Ivan Fernandez-Val, and Martin Weidner. "Network and panel quantile effects via distribution regression". *arXiv preprint arXiv:1803.08154* (2018).

[17] Clogg, Clifford C, Donald B Rubin, Nathaniel Schenker, Bradley Schultz, and Lynn Weidman. "Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression". *Journal of the American Statistical Association* 86.413 (1991), pp. 68–78.

[18] Doorley, Karina and Eva Sierminska. "Myth or fact? The beauty premium across the wage distribution" (2012).

[19] Dube, Arindrajit. "Minimum wages and the distribution of family incomes" (2017).

[20] Eng, Li Li and Margaret Shackell. "The implications of long-term performance plans and institutional ownership for firms' research and development (R&D) investments". *Journal of Accounting, Auditing & Finance* 16.2 (2001), pp. 117–139.

[21] Engle, Robert F and Simone Manganelli. "CAViaR: Conditional autoregressive value at risk by regression quantiles". *Journal of Business & Economic Statistics* 22.4 (2004), pp. 367–381.

[22] Firpo, Sergio and Cristine Pinto. "Identification and Estimation of Distributional Impacts of Interventions Using Changes in Inequality Measures". *Journal of Applied Econometrics* (2015).

[23] Foresi, Silverio and Franco Peracchi. "The conditional distribution of excess returns: An empirical analysis". *Journal of the American Statistical Association* 90.430 (1995), pp. 451–466.

[24] Francis, Jennifer and Abbie Smith. "Agency costs and innovation some empirical evidence". *Journal of Accounting and Economics* 19.2-3 (1995), pp. 383–409.

[25] Gaglianone, Wagner Piazza, Luiz Renato Lima, Oliver Linton, and Daniel R Smith. "Evaluating value-at-risk models via quantile regression". *Journal of Business & Economic Statistics* 29.1 (2011), pp. 150–160.

[26] Gelman, Andrew, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

[27] Hall, Bronwyn H and Josh Lerner. "The financing of R&D and innovation". *Handbook of the Economics of Innovation*. Vol. 1. Elsevier, 2010, pp. 609–639.

[28] Han, Lu, Chandler Lutz, and B Sand. *The effects of macroprudential mortgage insurance regulation during a housing boom: Evidence from Canada*. Tech. rep. Citeseer, 2016.

[29] Heckman, James J and Edward J Vytlacil. "Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation". *Handbook of econometrics* 6 (2007), pp. 4779–4874.

[30] Hirshleifer, David, Angie Low, and Siew Hong Teoh. "Are overconfident CEOs better innovators?" *The Journal of Finance* 67.4 (2012), pp. 1457–1498.

[31] Koenker, Roger and Jungmo Yoon. "Parametric links for binary choice models: A Fisherian–Bayesian colloquy". *Journal of Econometrics* 152.2 (2009), pp. 120–130.

[32] Lancaster, Tony and Sung Jae Jun. "Bayesian quantile regression methods". *Journal of Applied Econometrics* 25.2 (2010), pp. 287–307.

[33] Law, Ho Chung Leon, Dougal J Sutherland, Dino Sejdinovic, and Seth Flaxman. "Bayesian Distribution Regression". *stat* 1050 (2017), p. 11.

[34] Montiel Olea, José Luis and Mikkel Plagborg-Møller. "Simultaneous Confidence Bands: Theory, Implementation, and an Application to SVARs" (2018).

[35] Richey, Jeremiah and Alicia Rosburg. "Understanding intergenerational economic mobility by decomposing joint distributions" (2016).

[36] Rubin, Donald B and Nathaniel Schenker. "Logit-based interval estimation for binomial data using the Jeffreys prior". *Sociological methodology* (1987), pp. 131–144.

[37] Schennach, Susanne M. "Bayesian exponentially tilted empirical likelihood". *Biometrika* 92.1 (2005), pp. 31–46.

[38] Shepherd, Debra et al. "A question of efficiency: decomposing South African reading test scores using PIRLS 2006". *Stellenbosch University, Department of Economics Working Papers (20* (2013).

[39] Stock, James H. "Nonparametric policy analysis". *Journal of the American Statistical Association* 84.406 (1989), pp. 567–575.

[40] Trajtenberg, Manuel. "A penny for your quotes: patent citations and the value of innovations". *The Rand Journal of Economics* (1990), pp. 172–187.

[41] Van der Vaart, Aad W. *Asymptotic statistics*. Vol. 3. Cambridge university press, 1998.

[42] Wooldridge, Jeffrey M. *Econometric analysis of cross section and panel data*. MIT press, 2010.

[43] Wüthrich, Kaspar. *Semiparametric estimation of quantile treatment effects with endogeneity*. Tech. rep. Discussion Papers, Universität Bern, Department of Economics, 2015.

[44] Yang, Yunwen and Huixia Judy Wang. "Bayesian Quantile Regression". *Handbook of Quantile Regression*. Chapman and Hall/CRC, 2017, pp. 61–74.

[45] Yu, Keming and Rana A Moyeed. "Bayesian quantile regression". *Statistics & Probability Letters* 54.4 (2001), pp. 437–447.

A Proofs

A.1 Theorems

Proof of Theorem 1.

We need to show that $P(L^*(y) \leq \hat{F}(y) \leq U^*(y)) \geq 1 - \alpha \; \forall y \in \bar{\mathcal{Y}}$.

(a) symmetric case

$$
\begin{aligned}
& P(F(y) - c_{1-\alpha} \leq \hat{F}(y) \leq F(y) + c_{1-\alpha}) \\
=& P(-c_{1-\alpha} \leq \hat{F}(y) - F(y) \leq c_{1-\alpha}) \\
=& P(-q_{1-\alpha}(\max_{y \in \bar{\mathcal{Y}}} |\hat{F}(y) - F(y)|) \leq \hat{F}(y) - F(y) \leq q_{1-\alpha}(\max_{y \in \bar{\mathcal{Y}}} |\hat{F}(y) - F(y)|)) \\
=& P(|\hat{F}(y) - F(y)| \leq q_{1-\alpha}(\max_{y \in \bar{\mathcal{Y}}} |\hat{F}(y) - F(y)|)) \\
\geq & 1 - \alpha \quad \forall \, y \in \bar{\mathcal{Y}}
\end{aligned}
$$

The first line uses the definition in theorem 1, the second line rearranges the terms and the third uses the fact that $c_{1-\alpha} \equiv q_{1-\alpha}(\max_{y \in \bar{\mathcal{Y}}} |F(y) - \hat{F}(y)|)$. The last line follows from the definition of the quantile function $q_{1-\alpha}(\cdot)$. Notice that the argument to the quantile function is the Kolmogorov maximal statistic.

(b) asymmetric case

$$
\begin{aligned}
& P(F(y) - \underline{c}_{\alpha/2} \leq \hat{F}(y) \leq F(y) + \bar{c}_{1-\alpha/2}) \\
&= P(-\underline{c}_{\alpha/2} \leq \hat{F}(y) - F(y) \leq \bar{c}_{1-\alpha/2}) \\
&= P(\hat{F}(y) - F(y) \leq \bar{c}_{1-\alpha/2}) - P(\hat{F}(y) - F(y) < -\underline{c}_{\alpha/2}) \\
&= P(\hat{F}(y) - F(y) \leq q_{1-\alpha/2}(\max_{y} \hat{F}(y) - F(y))) - P(\hat{F}(y) - F(y) < q_{\alpha/2}(\min_{y} \hat{F}(y) - F(y))) \\
&\geq 1 - \alpha \quad \forall \, y \in \bar{\mathcal{Y}}
\end{aligned}
$$

The first probability in the fourth line $= 1 - \alpha/2$ while the second is $< \alpha/2$ hence their difference is $\geq 1 - \alpha$. □

Proof of Theorem 3.

This proof follows Gelman, Carlin, Stern, and Rubin (1995, Appendix B). We use the logit link function and (non-informative) uniform prior for demonstration. Results using other suitable link functions follow similarly. The log of the posterior obtains as

$$L(\boldsymbol{\theta}_g|\breve{y}^g) = \log p(\boldsymbol{\theta}_g|\breve{y}^g) = \sum_{i=1}^{N} \mathbf{x}_i\boldsymbol{\theta}_g \mathbb{1}\{y_i \leq y_g\} - \sum_{i=1}^{N} \log(1+\exp(\mathbf{x}_i\boldsymbol{\theta}_g)) \tag{A.1}$$

such that $y_g \in \bar{\mathcal{Y}}$. The score function of $L(\boldsymbol{\theta}_g|\breve{y}^g)$ is given by

$$\mathbf{s}(\boldsymbol{\theta}_g) = \nabla_{\boldsymbol{\theta}_g} L(\boldsymbol{\theta}_g|\breve{y}^g) = \sum_{i=1}^{N} \Big(\mathbb{1}\{y_i \leq y_g\} - \frac{\exp(\mathbf{x}_i\boldsymbol{\theta}_g)}{1+\exp(\mathbf{x}_i\boldsymbol{\theta}_g)}\Big)\mathbf{x}_i' \tag{A.2}$$

Taking the second derivative with respect to $\boldsymbol{\theta}_g$,

$$\frac{d^2}{d\boldsymbol{\theta}_g} L(\boldsymbol{\theta}_g|\breve{y}^g) = -\sum_{i=1}^{N} \Big(\frac{\exp(\mathbf{x}_i\boldsymbol{\theta}_g)}{(1+\exp(\mathbf{x}_i\boldsymbol{\theta}_g))^2}\Big)\mathbf{x}_i\mathbf{x}_i' \tag{A.3}$$

obtains as the hessian matrix. Notice that the expression in eq. (A.3) above is only dependent on y_g via $\boldsymbol{\theta}_g$. Taking the Taylor expansion of $L(\boldsymbol{\theta}_g|\breve{y}^g)$ around the mode $\boldsymbol{\theta}_{o,g}$ gives

$$L(\boldsymbol{\theta}_g|\breve{y}^g) = L(\boldsymbol{\theta}_{o,g}|\breve{y}^g) + \frac{1}{2}(\boldsymbol{\theta}_g - \boldsymbol{\theta}_{o,g})'\Big[\frac{d^2}{d\boldsymbol{\theta}_g} L(\boldsymbol{\theta}_g|\breve{y}^g)\Big]\Big|_{\boldsymbol{\theta}_g=\boldsymbol{\theta}_{o,g}}(\boldsymbol{\theta}_g - \boldsymbol{\theta}_{o,g}) + (s.o.) \tag{A.4}$$

where $(s.o.)$ are negligible smaller order terms. The first term is constant and the second is proportional to the logarithm of the multivariate normal density of $\boldsymbol{\theta}_g$ with

$$p(\boldsymbol{\theta}_g|\breve{y}^g) \approx \mathcal{N}(\boldsymbol{\theta}_{o,g}, \mathcal{I}(\boldsymbol{\theta}_{o,g})^{-1}) \tag{A.5}$$

where $\mathcal{I}(\boldsymbol{\theta}_g) = -E[\frac{d^2}{d\boldsymbol{\theta}_g} L(\boldsymbol{\theta}_g|\breve{y}^g)]$ is the information matrix. □

Proof of Theorem 4.

The results follow from the delta method[32] and noting the exchangeability of the derivative and the integral which holds under general regularity conditions.

$$\begin{aligned} &\sqrt{N}(\hat{F}_Y(y_g|\hat{\boldsymbol{\theta}}_g) - F_Y(y_g|\boldsymbol{\theta}_{o,g})) = \sqrt{N}(\hat{F}_Y(y_g|\hat{\boldsymbol{\theta}}_g) - \hat{F}_Y(y_g|\boldsymbol{\theta}_{o,g}) + \hat{F}_Y(y_g|\boldsymbol{\theta}_{o,g}) - F_Y(y_g|\boldsymbol{\theta}_{o,g})) \\ &= \sqrt{N}(N^{-1}\sum_{i=1}^{N} \Lambda(\mathbf{x}_i\hat{\boldsymbol{\theta}}_g) - N^{-1}\sum_{i=1}^{N} \Lambda(\mathbf{x}_i, \boldsymbol{\theta}_{o,g})) + \sqrt{N}(N^{-1}\sum_{i=1}^{N} \Lambda(\mathbf{x}_i, \boldsymbol{\theta}_{o,g}) - F_Y(y_g|\boldsymbol{\theta}_{o,g})) \end{aligned} \tag{A.6}$$

The second term converges to zero in probability. Applying the delta method (see Van der Vaart

[32] See (Van der Vaart (1998), chapter 3).

(1998, Chapter 3)) to the first term, we have

$$\sqrt{N}(\hat{F}_Y(y_g|\hat{\boldsymbol{\theta}}_g) - F_Y(y_g|\boldsymbol{\theta}_{o,g})) = N^{-1}\sum_{i=1}^{N} \Lambda'(\mathbf{x}_i\boldsymbol{\theta}_{o,g})\mathbf{x}_i'\sqrt{N}(\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_{o,g}) + o_p(1) \tag{A.7}$$

Applying the central limit theorem,

$$\sqrt{N}(\hat{F}_Y(y_g|\hat{\boldsymbol{\theta}}_g) - F_Y(y_g|\boldsymbol{\theta}_{o,g})) \xrightarrow{d} \mathcal{N}(\mathbf{0}, N\mathcal{V}_{F_{y_g}}) \tag{A.8}$$

where $\mathcal{V}_{F_{y_g}} = E[\boldsymbol{\lambda}_i^{\mathbf{x}}(\boldsymbol{\theta}_{o,g})'[I(\boldsymbol{\theta}_{o,g})]^{-1}\boldsymbol{\lambda}_i^{\mathbf{x}}(\boldsymbol{\theta}_{o,g})]$ and $\boldsymbol{\lambda}_i^{\mathbf{x}}(\boldsymbol{\theta}_{o,g}) = \Lambda'(\mathbf{x}_i\boldsymbol{\theta}_{o,g})\mathbf{x}_i'$. □

Proof of corollary 1 .

From the score function of the posterior by $\mathbf{s}(\boldsymbol{\theta}_g)$ eq. (A.2), the influence function representation of $\hat{\boldsymbol{\theta}}_g$ (see Wooldridge (2010, equations 12.15 - 12.17)) obtains as

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_{o,g}) = [I(\boldsymbol{\theta}_{o,g})]^{-1}N^{-1/2}\sum_{i=1}^{N}\mathbf{s}_i(\boldsymbol{\theta}_{o,g}) + o_p(1) \tag{A.9}$$

Expanding eq. (A.7) using eq. (A.9) obtains

$$\sqrt{N}(\hat{F}_Y(y_g|\hat{\boldsymbol{\theta}}_g) - F_Y(y_g|\boldsymbol{\theta}_{o,g})) = N^{-1}\sum_{i=1}^{N} \Lambda'(\mathbf{x}_i\boldsymbol{\theta}_{o,g})\mathbf{x}_i'[I(\boldsymbol{\theta}_{o,g})]^{-1}N^{-1/2}\mathbf{s}_i(\boldsymbol{\theta}_{o,g}) + o_p(1) \tag{A.10}$$

Applying the multivariate central limit theorem (see Van der Vaart (1998, Section 2.18)) to $\sqrt{N}[\hat{\mathbf{F}}_Y - \mathbf{F}_Y]' = \sqrt{N}[(\hat{F}_Y(y_1|\hat{\boldsymbol{\theta}}_1) - F_Y(y_1|\boldsymbol{\theta}_{o,1})), ..., (\hat{F}_Y(y_G|\hat{\boldsymbol{\theta}}_G) - F_Y(y_G|\boldsymbol{\theta}_{o,G}))]'$ using the representation in eq. (A.10),

$$\sqrt{N}[\hat{\mathbf{F}}_Y - \mathbf{F}_Y]' \xrightarrow{d} \mathcal{N}(\mathbf{0}, N\boldsymbol{\Omega}_{F_y}) \tag{A.11}$$

where $\boldsymbol{\Omega}_{F_{y_g}}$ comprises the following elements: (g,g)'th element $\mathcal{V}_{F_{y_g}} = E[\boldsymbol{\lambda}_i^{\mathbf{x}}(\boldsymbol{\theta}_{o,g})'[I(\boldsymbol{\theta}_{o,g})]^{-1}\boldsymbol{\lambda}_i^{\mathbf{x}}(\boldsymbol{\theta}_{o,g})]$, (g,h)'th element $\mathcal{V}_{F_{y_{g,h}}} = E[\boldsymbol{\lambda}_i^{\mathbf{x}}(\boldsymbol{\theta}_{o,g})'[I(\boldsymbol{\theta}_{o,g})]^{-1}I_i(\boldsymbol{\theta}_{g,h})[I(\boldsymbol{\theta}_{o,h})]^{-1}\boldsymbol{\lambda}_i^{\mathbf{x}}(\boldsymbol{\theta}_{o,h})]$, and $I_i(\boldsymbol{\theta}_{g,h}) = N^{-1}\mathbf{s}_i(\boldsymbol{\theta}_{o,g})\mathbf{s}_i(\boldsymbol{\theta}_{o,h})'$.

□

Proof of corollary 2.

$$
\begin{aligned}
\sqrt{N}(\hat{\Delta}_g^{DE} - \Delta_g^{DE}) &= \sqrt{N}((\hat{F}_Y^c(y_g) - \hat{F}_Y(y_g)) - (F_Y^c(y_g) - F_Y(y_g))) \\
&= \sqrt{N}(\hat{F}_Y^c(y_g) - F_Y^c(y_g)) - \sqrt{N}(\hat{F}_Y(y_g) - F_Y(y_g)) \\
&= N^{-1}\sum_{i=1}^{N}(\Lambda'(\mathbf{x}_i\boldsymbol{\alpha}'\boldsymbol{\theta}_{o,g})\boldsymbol{\alpha} - \Lambda'(\mathbf{x}_i\boldsymbol{\theta}_{o,g})\mathbf{I}_k)\mathbf{x}_i'[I(\boldsymbol{\theta}_{o,g})]^{-1}N^{-1/2}\mathbf{s}_i(\boldsymbol{\theta}_{o,g}) + o_p(1) \\
&= N^{-1}\sum_{i=1}^{N}\bar{\boldsymbol{\lambda}}_i^{\mathbf{x}}(\boldsymbol{\theta}_{o,g})[I(\boldsymbol{\theta}_{o,g})]^{-1}N^{-1/2}\mathbf{s}_i(\boldsymbol{\theta}_{o,g}) + o_p(1)
\end{aligned}
\tag{A.12}
$$

where $\bar{\boldsymbol{\lambda}}_i^{\mathbf{x}}(\boldsymbol{\theta}_{o,g}) = (\Lambda'(\mathbf{x}_i\boldsymbol{\alpha}'\boldsymbol{\theta}_{o,g})\boldsymbol{\alpha} - \Lambda'(\mathbf{x}_i\boldsymbol{\theta}_{o,g})\mathbf{I}_k)$. From the above influence function representation, it follows from the CLT that

$$
\sqrt{N}(\hat{\Delta}_g^{DE} - \Delta_g^{DE}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, N\mathcal{V}_{\Delta_g^{DE}}) \tag{A.13}
$$

where $\mathcal{V}_{\Delta_g^{DE}} = E[\bar{\boldsymbol{\lambda}}_i^{\mathbf{x}}(\boldsymbol{\theta}_{o,g})[I(\boldsymbol{\theta}_{o,g})]^{-1}\bar{\boldsymbol{\lambda}}_i^{\mathbf{x}}(\boldsymbol{\theta}_{o,g})']$ □

Proof of corollary 3.

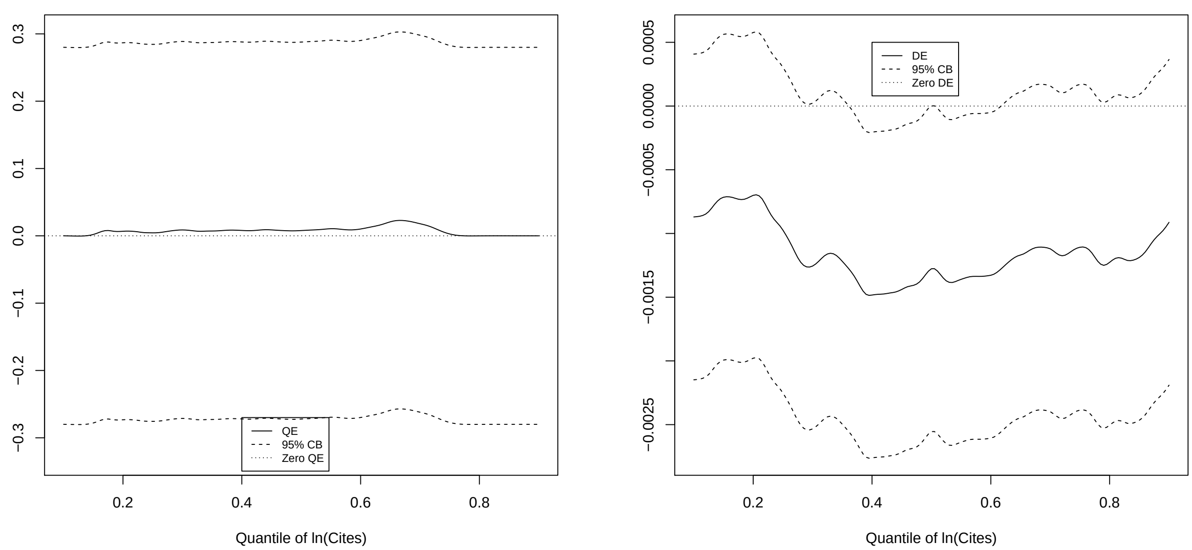
This result obtains by applying the multivariate central limit theorem (see Van der Vaart (1998, Section 2.18)) to $\sqrt{N}(\hat{\boldsymbol{\Delta}}^{DE} - \boldsymbol{\Delta}^{DE}) = \sqrt{N}[(\hat{\Delta}_1^{DE} - \Delta_1^{DE}), ..., (\hat{\Delta}_G^{DE} - \Delta_G^{DE})]'$. Using the representation in eq. (A.12),

$$
\sqrt{N}(\hat{\boldsymbol{\Delta}}^{DE} - \boldsymbol{\Delta}^{DE}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, N\boldsymbol{\Omega}_{\Delta}) \tag{A.14}
$$

where the (g,h)'th element of $\boldsymbol{\Omega}_{\Delta}$ is $E[\bar{\boldsymbol{\lambda}}_i^{\mathbf{x}}(\boldsymbol{\theta}_{o,g})'[I(\boldsymbol{\theta}_{o,g})]^{-1}I_i(\boldsymbol{\theta}_{g,h})[I(\boldsymbol{\theta}_{o,h})]^{-1}\bar{\boldsymbol{\lambda}}_i^{\mathbf{x}}(\boldsymbol{\theta}_{o,h})]$ and the (g,g)'th element is $E[\bar{\boldsymbol{\lambda}}_i^{\mathbf{x}}(\boldsymbol{\theta}_{o,g})'[I(\boldsymbol{\theta}_{o,g})]^{-1}\bar{\boldsymbol{\lambda}}_i^{\mathbf{x}}(\boldsymbol{\theta}_{o,g})]$. □

B Figures

The left panels in Figure 5 show the quantile and distributions of $ln(Cites)$ with their 95% confidence bands while those of $Cites$ are in Figure 7. Corresponding counterfactuals are on the right. From both sets of graphs, one observes confidence bands widening in upper quantiles of the outcome variables $Cites$ and $ln(Cites)$. In spite of the less precision in the upper quantiles of distributions, using both the point-wise and simultaneous results of the distribution effect on $ln(Cites)$ (see Figure 1 and appendix B) leads one to conclude that increasing institutional ownership by 1% has a stochastically dominant effect at 5% significance level. A similar conclusion on the stochastic dominance for the discrete outcome $Cites$ can also be drawn (see Figure 2).



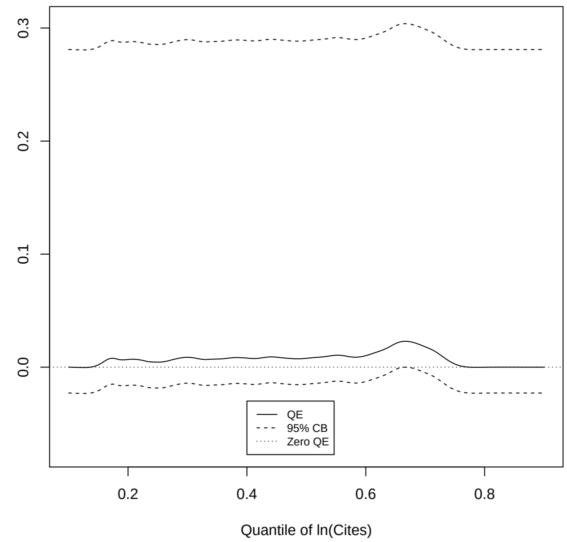
(a) Quantile effects (b) Distribution effects

Figure 3: Quantile and distribution treatment effects, symmetric simultaneous CB

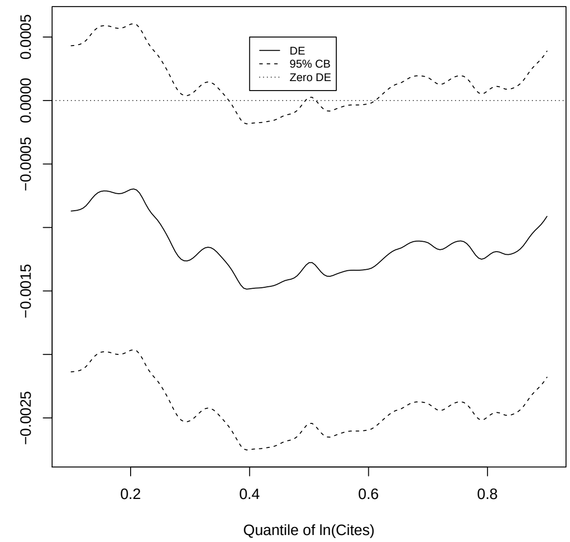
Notes: The figures above plot quantile and distribution effects with symmetric simultaneous 95% confidence bands using theorem 1.

C Empirical Model Estimation

In this section, we provide details of the empirical estimation. Relevant codes in `R` for the empirical estimation and Monte Carlo simulations are available on request from the authors. Functions used in the codes are available in the R package `bayesdistreg`. The OLS, Poisson, and



(a) Quantile effects



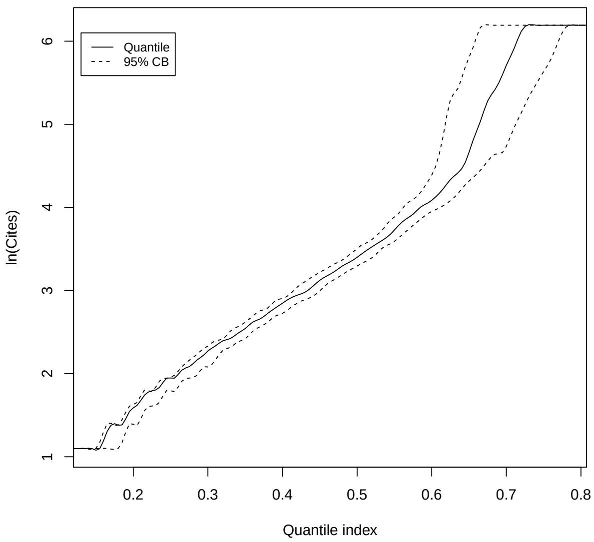
(b) Distribution effects

Figure 4: Quantile and distribution treatment effects, simultaneous CB

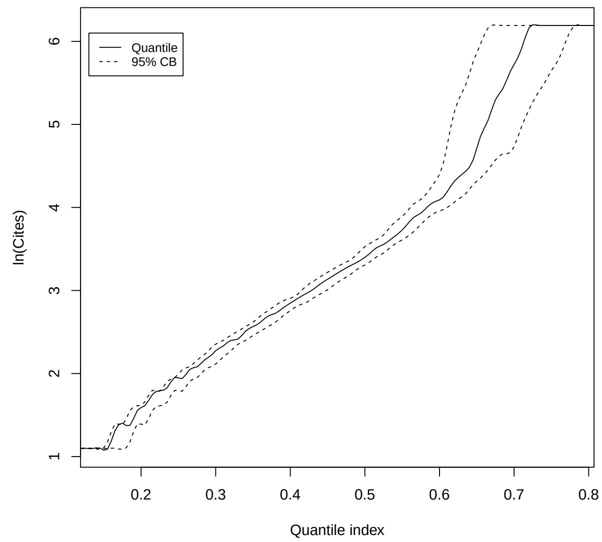
Notes: The figures above plot quantile and distribution effects with 95% asymmetric simultaneous confidence bands using theorem 1.

Negative binomial replications in table 5 use the same variables as in Aghion, Van Reenen, and Zingales (2013). $ln(Cites)$ is the outcome in the OLS while $Cites$ is the outcome variable in the Poisson and Negative Binomial models. The same set of variables are used in the two Bayes BDR models we run. The first BDR model has $ln(Cites)$ as the outcome while the second has $Cites$ as the outcome variable. 161 threshold values on $ln(Cites)$ and $Cites$ (corresponding to the 10-90'th evenly spaced quantile indices) are used in the DR estimation. In all, there are 150 covariates (including all dummies but excluding the intercept) in each model. In each regression, the counterfactual treatment level is generated by increasing institutional ownership by 1% with the intention of making our results directly comparable to those of Aghion, Van Reenen, and Zingales (2013) and Berger, Stocker, and Zeileis (2017). For MCMC implementations, we use the Independence Metropolis-Hastings algorithm. The proposal density of parameters is formed using the normal approximation of the likelihood[33] with the variance-covariance scaled up by a factor of 1.5 to ensure that the proposal covers the posterior adequately. Due to the large number of covariates, memory and computational time constraints, we make 5000 draws of which the first 1000 are discarded as burn-in.

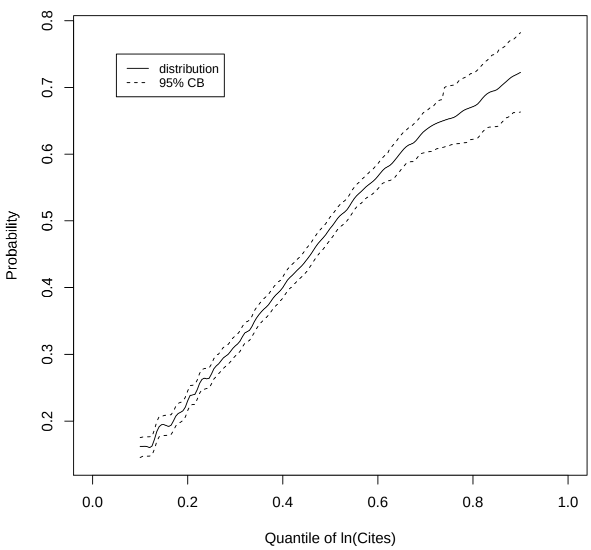
[33] Recall we use the uniform prior on parameters.



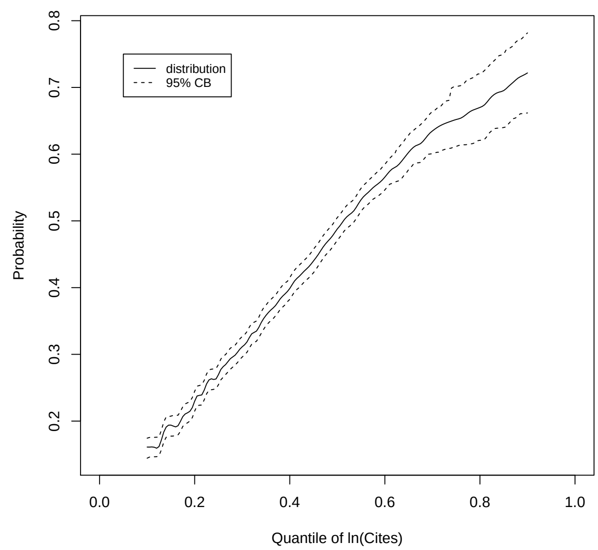
(a) Quantiles of ln(Cites)



(b) Counterfactual quantiles of ln(Cites)



(c) Distribution of ln(Cites)



(d) Counterfactual distribution of ln(Cites)

Figure 5: (Counterfactual) quantiles and distribution

Notes: The top left panel plots the quantiles of ln(Cites) with 95% simultaneous confidence bands while the top right panel plots the counterfactual quantiles with 95% simultaneous confidence bands. The bottom left and right panels plot the distribution and counterfactual distributions, respectively with 95% simultaneous confidence bands. Counterfactuals are generated from increasing institutional ownership Share_Inst by 10%. All the above figures are based on bayesian distribution algorithm 1. Simultaneous confidence bands on the distributions are constructed using theorem 1.

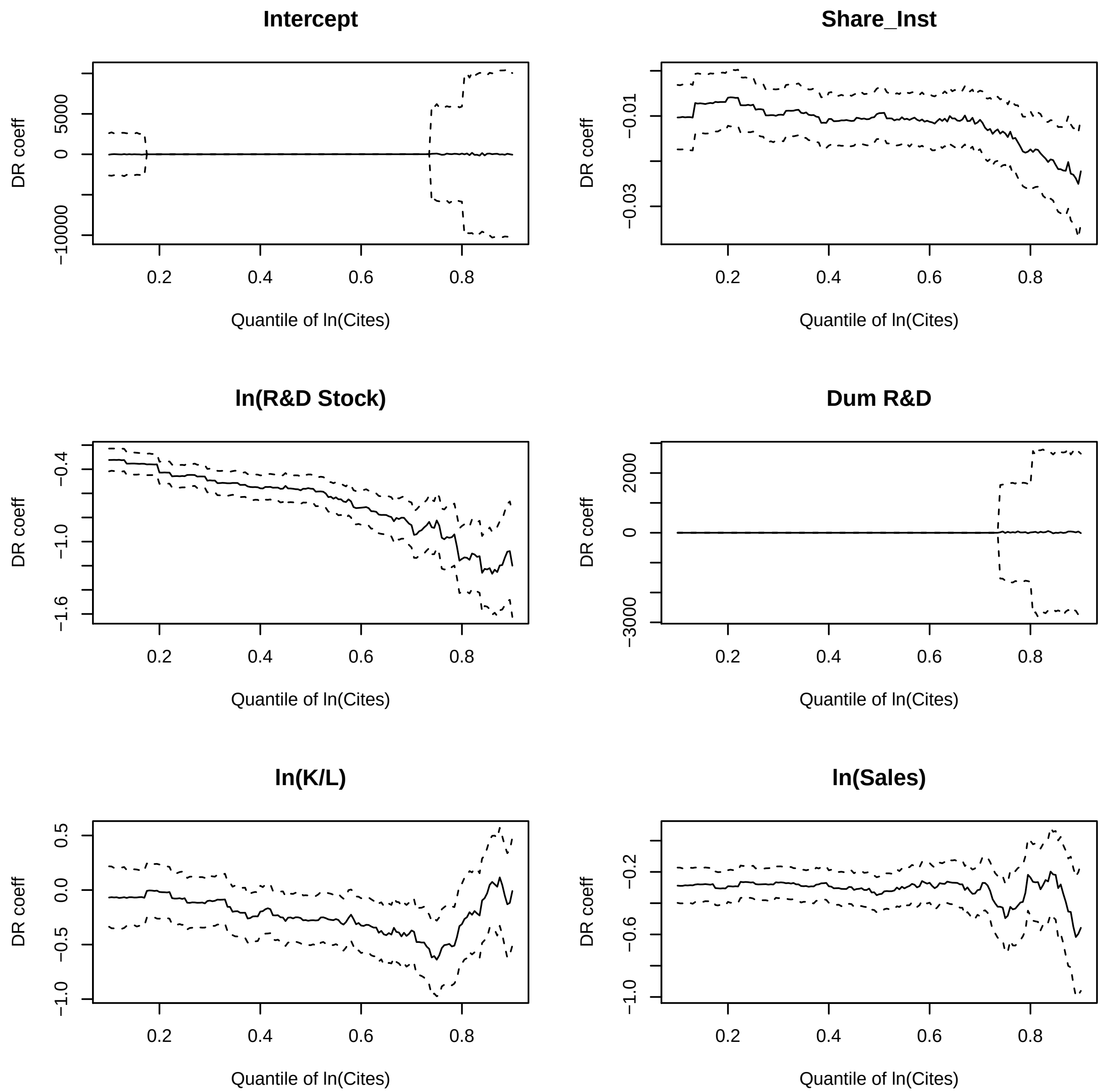
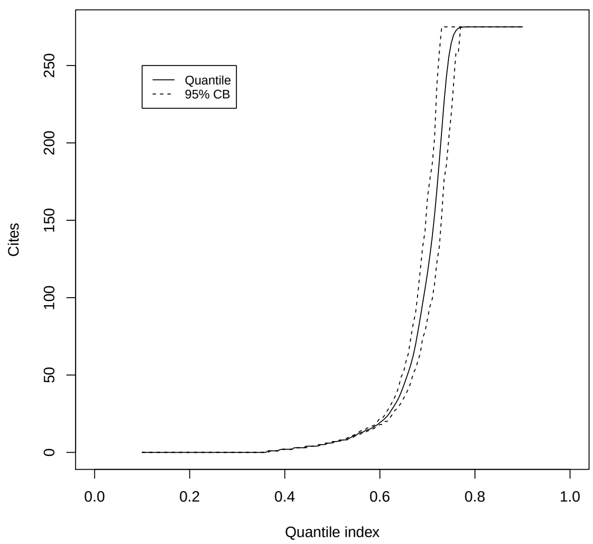
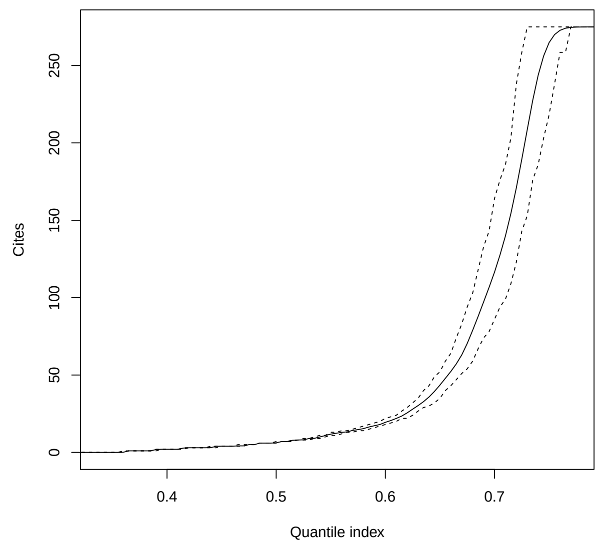


Figure 6: Parameter estimates with 95% point-wise confidence bands, outcome: $ln(Cites)$

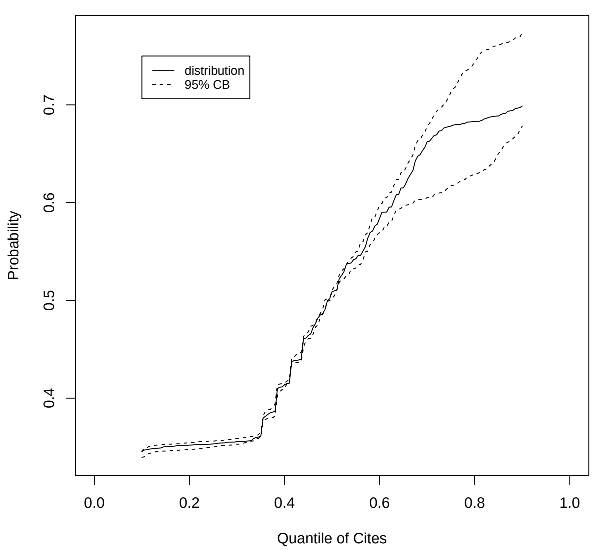
Notes: The above figures correspond to Bayesian distribution regression coefficients on covariates in tables 4 and 5. Note that distribution coefficients are negatively proportional to the respective quantile effect.



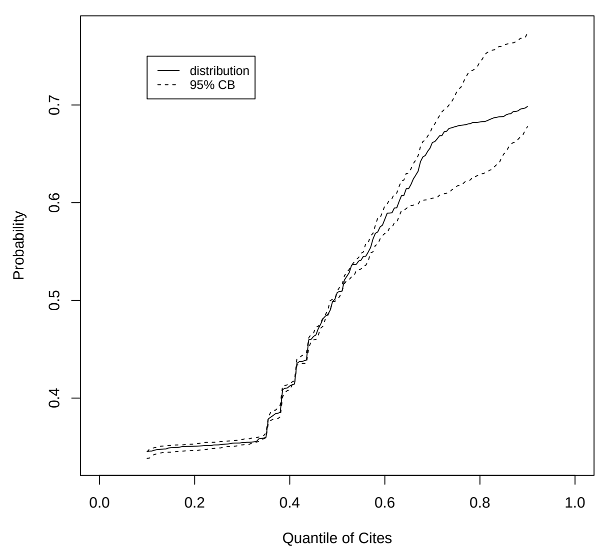
(a) Quantiles of Cites



(b) Counterfactual quantiles of Cites



(c) Distribution of Cites



(d) Counterfactual distribution of Cites

Figure 7: (Counterfactual) quantiles and distribution of *Cites*

Notes: The top left panel plots the quantiles of Cites with 95% point-wise confidence bands while the top right panel plots the counterfactual quantiles with 95% point-wise confidence bands. The bottom left and right panels plot the distribution and counterfactual distributions, respectively with 95% point-wise confidence bands. Counterfactuals are generated from increasing institutional ownership Share_Inst by 1%. All figures above are based on Bayesian distribution algorithm 1. Point-wise confidence bands on the distributions and quantiles are constructed 95% posterior intervals. The distribution of quantiles obtain by taking the left inverse column-wise of the respective distribution matrix ($\boldsymbol{P}^o$ and $\boldsymbol{P}^c$). See algorithm 1 and section 3.2. All target distributions and quantiles above are point-wise mean values.

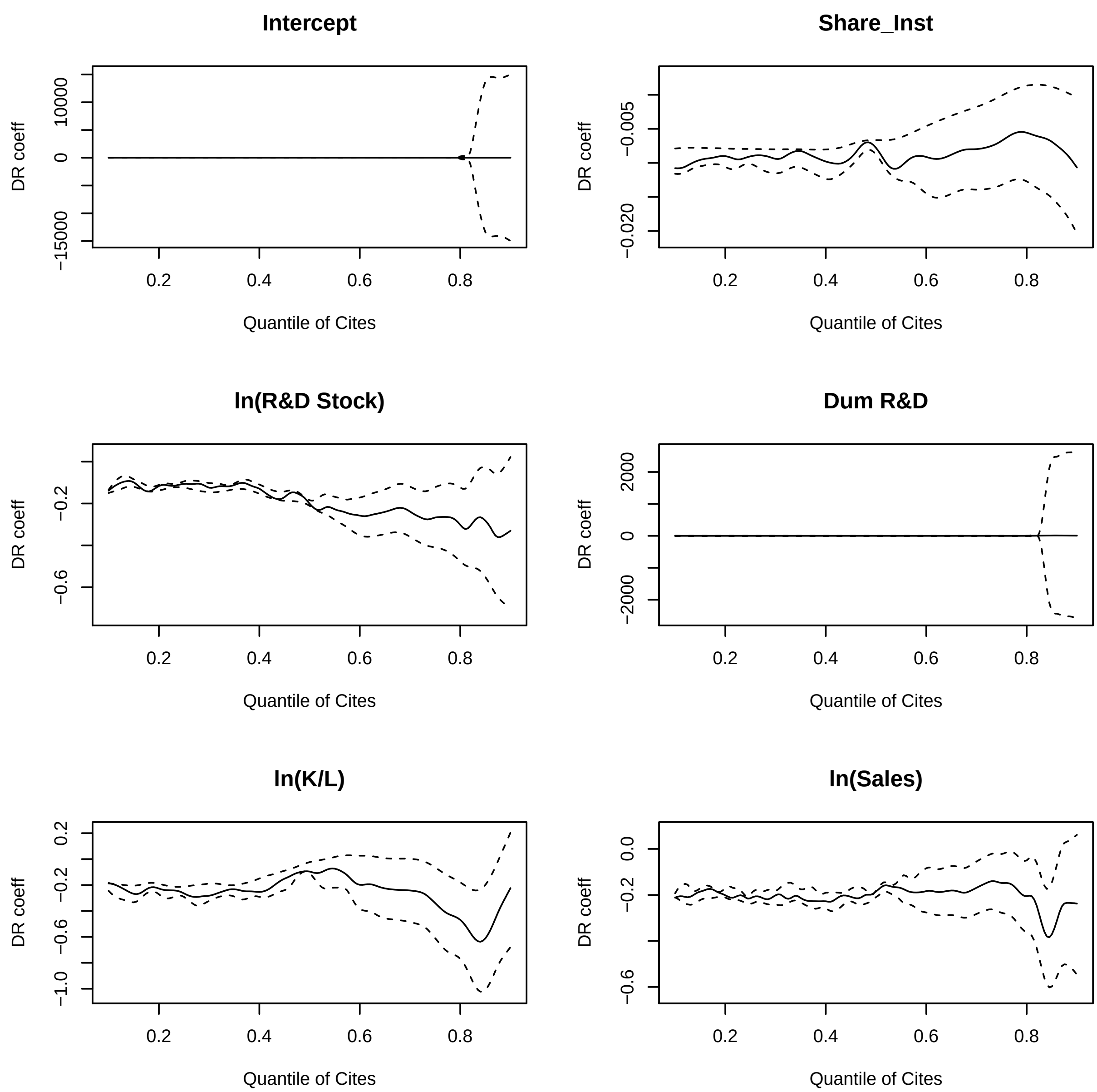


Figure 8: Parameter estimates with 95% point-wise confidence bands, outcome: *Cites*

Notes: The figures above correspond to Bayesian distribution regression coefficients on covariates in tables 4 and 5. Note that distribution coefficients are negatively proportional to the respective quantile effect.