# Discrete Choice Data with Unobserved Heterogeneity: A Conditional Binary Quantile Model

*Online appendix for publication in Political Analysis*

Xiao Lu

# A The conditional mean-based binary choice model and its limitations

## A.1 The conditional mean-based binary model

It is called the conditional mean-based model because the expected probability of the success (choosing 1) is modeled as equivalent to the mean of the response conditioning on a set of explanatory variables. Following convention, I take a random utility view of the observed choices. That is, actors make a choice based on their latent utilities from the alternatives and choose the one providing them with the highest utility.

I start from the simplest case. Suppose actor $i$ faces two choices, $Y \in \{0, 1\}$, and her utility from the two choices is

$$\begin{array}{ll} U(y_i = 1) & = \boldsymbol{x_i'\beta_1} + \varepsilon_{i1}, \\ U(y_i = 0) & = \boldsymbol{x_i'\beta_0} + \varepsilon_{i0}, \end{array} \tag{1}$$

where $x_i$ is a set of individual characteristics (independent variables) of actor $i$, $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ are $k \times 1$ vectors of coefficients with respect to choosing 0 and 1, and $\varepsilon$ is the error term. Therefore, the probability of choosing $Y = 1$ is simply the probability of $U_i(Y = 1) > U_i(Y = 0)$, which is

$$\begin{array}{rl} Pr(y_i = 1|\boldsymbol{x_i}) & = Pr(U(y_i = 1) > U(y_i = 0)) \\ & = Pr(\boldsymbol{x_i}'(\boldsymbol{\beta_1} - \boldsymbol{\beta_0}) + \varepsilon_{i1} - \varepsilon_{i0}) \\ & = Pr(\boldsymbol{x_i'\beta} + \varepsilon_i), \end{array} \tag{2}$$

where $\boldsymbol{\beta} = \boldsymbol{\beta_1} - \boldsymbol{\beta_0}$ and $\varepsilon_i = \varepsilon_{i1} - \varepsilon_{i0}$. Since the constant term in the latent utility specification is unchanged by the choice of switching thresholds, we can normalize

the utility of the second choice to 0. Therefore, we have the following latent utility specification:

$$y_i^* = \boldsymbol{x_i'}\boldsymbol{\beta} + \varepsilon_i. \tag{3}$$

By adding a probability measure over (3), it follows naturally that

$$Pr(y_i = 1|\boldsymbol{x_i}) = Pr(y_i^* > 0) = Pr(\boldsymbol{x_i'}\boldsymbol{\beta} + \varepsilon_i > 0) = Pr(\varepsilon_i > -x_i'\beta), \tag{4}$$

which is the upper-tail distribution of the error term conditional on $\boldsymbol{x_i}$. By assuming a symmetric error distribution, we obtain

$$Pr(y_i = 1) = Pr(y_i^* > 0) = Pr(\varepsilon_i < \boldsymbol{x_i'}\boldsymbol{\beta}) = F(\boldsymbol{x_i'}\boldsymbol{\beta}), \tag{5}$$

where $F(.)$ is a cumulative density function. In general, the conditional mean-based regression for binary outcomes assumes the following form

$$Pr(y_i = 1|\boldsymbol{x_i}) = E(y_i|\boldsymbol{x_i}) = F(\boldsymbol{x_i'}\boldsymbol{\beta}). \tag{6}$$

According to the formulation of the conditional mean-based estimator, the summary statistics of the conditional behavior of the response variable depend solely on its conditional mean, which ignores heterogeneity across units. Both theoretically and empirically, such an assumption of a homogeneous dataset is rather strong.

## A.2 Unobserved heterogeneity

There are many types of unobserved heterogeneity. I focus here on the one corresponding to the varying behavior of individuals when faced with a similar set of choices. In other words, the effects of each covariate on the response vary over units. Conditional mean-based models typically assume homogeneity among individuals or subgroups of a population. It regards decision makers as possessing the same preference over alternatives and making rather similar decisions if exposed to similar choices. However, counterexamples are heterogeneous decision makers who are also influenced by unobserved contextual factors and therefore have different preferences in facing the same choice. Formally,

$$U_i = \boldsymbol{x'}\boldsymbol{\beta_i} \neq U_j = \boldsymbol{x'}\boldsymbol{\beta_j}, \text{for} i, j \in I \tag{7}$$

where $U_i$ and $U_j$ are utilities of individual $i$ and $j$, $\boldsymbol{x'}$ is the choice-specific features, $\boldsymbol{\beta_i}$ and $\boldsymbol{\beta_j}$ are individual components of the respective utility functions, and $I$ is a population set. When $\boldsymbol{\beta}$ differs across individuals, the assumption of homogeneity

will not hold. A typical solution to this unobserved heterogeneity problem is by differencing with each unit, using a fixed effect estimator in time-series-cross-section data so that unobserved heterogeneity across units cancels out. However, this kind of data structure is not always available from the dataset we can obtain. More importantly, we may be interested in unobserved heterogeneity itself, and therefore, differencing heterogeneity within units is undesirable.

Take government formation as an example: the policy-office trade-offs of potential governments, the personalities of party leaders, and contextual election features that are difficult for researchers to observe may differentiate the coalition choice and formation likelihood of potential governments, even though they are all impacted by the same observed factors.

## A.3 Independence of irrelevant alternatives

If we assume that the error terms of each observation are uncorrelated with each other, we must also assume that IIA holds. The IIA assumption implies that for any given pair of alternatives, adding additional alternatives will not change the probability ratio between that pair of alternatives. For example,

$$\frac{Pr(y_i = 1|\{0,1\})}{Pr(y_i = 0|\{0,1\})} = \frac{Pr(y_i = 1|\{0,1,2\})}{Pr(y_i = 0|\{0,1,2\})} \tag{8}$$

means that the probability ratio (or odds ratio) of choosing 1 over 0 given the set of alternatives $\{0,1\}$ is the same by adding an additional alternative 2 to the original set of alternatives $\{0,1\}$.

This is a very restrictive assumption, particularly when alternatives 0 and 2 share similar features but alternative 1 is distinct. In such cases, the probability of choosing 0 decreases due to the presence of a similar alternative 2, but the probability of choosing 1 remains the same. As a result, the probability ratio between 1 and 0 increases due to the presence of alternative 2, and the IIA assumption is violated.

## A.4 Distributional misspecification

The most commonly used binary models, such as the logit and probit, are also limited by their restrictive distributional assumptions. Due to the need to transfer a linear combination of covariates with complete support over the real line to a probability space ranging from zero to one, the traditional binary models use a certain type of cumulative density function as the link function. The common choices of the cumulative distribution function include the normal cumulative distribution, which

gives rise to the probit model:

$$Pr(y_i = 1|\boldsymbol{x_i}) = \int_{-\infty}^{\boldsymbol{x_i'\beta}} \phi(t)dt = \Phi(\boldsymbol{x_i'\beta}), \tag{9}$$

and the logistic cumulative distribution, which gives rise to the logit model:

$$Pr(y_i = 1|\boldsymbol{x_i}) = \frac{exp(\boldsymbol{x_i'\beta})}{1 + exp(\boldsymbol{x_i'\beta})}. \tag{10}$$

However, these distributional assumptions are rather strong. As illustrated by Figure 1, by simulating data from different distributions and estimating them with the probit model, we find that except for the correctly specified model in the top-left panel, the models produce the incorrect predicted probabilities. This incorrectness indicates that, in the discrete choice settings, the specification of the error terms also matters. Once the underlying distribution deviates from the assumption, the estimator produces biased estimates.

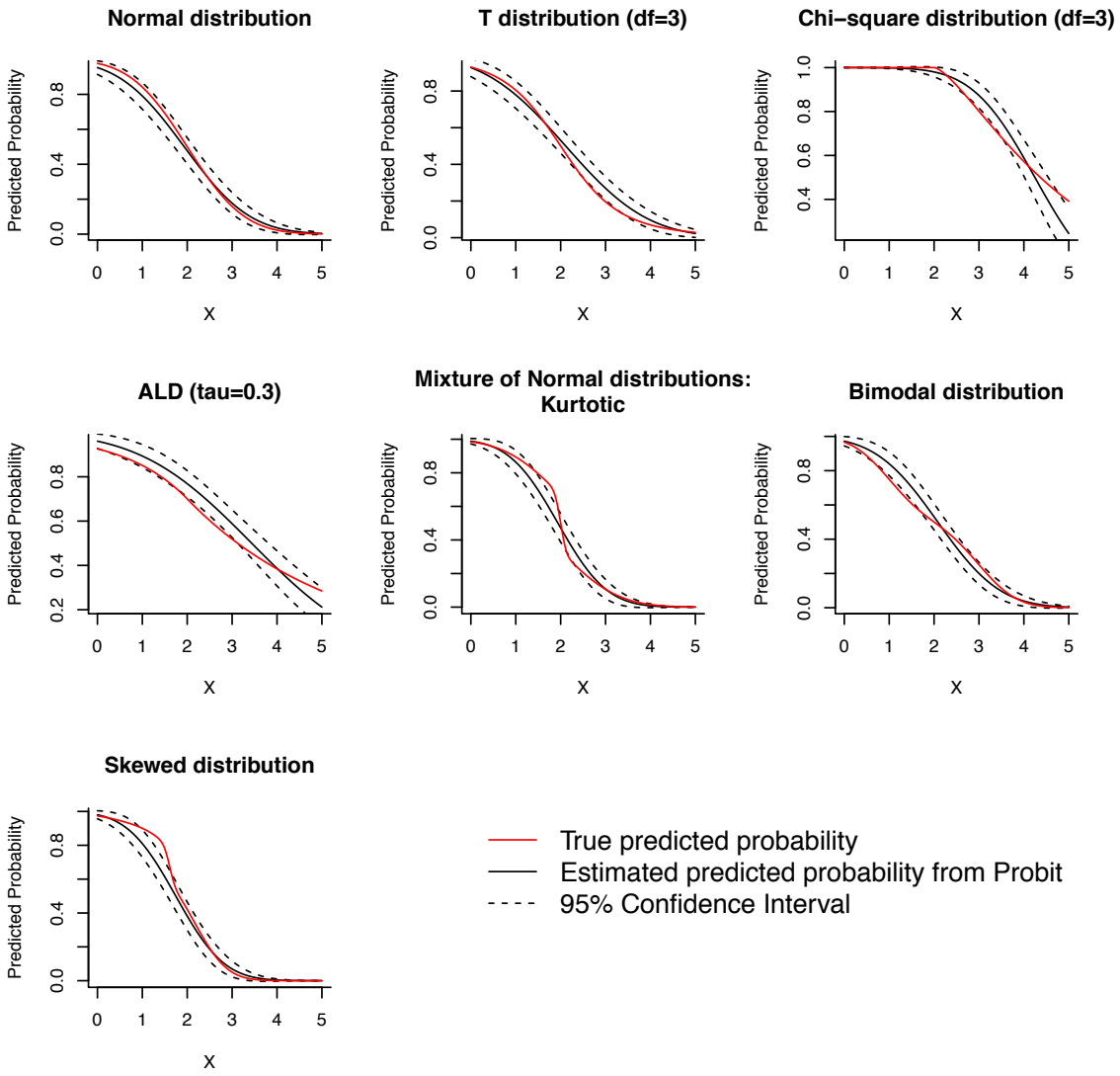


Figure 1: Comparison between the predicted probabilities and the true probabilities under different error distributions

## A.5 Specifications of error distributions of Figure 1

For the above distributions, I simulate data from the form:

$$\begin{aligned} &y_i* = \beta_0 + x_i'\beta_1 + \varepsilon_i \\ &Pr(y_i = 1|x_i) = 1 - F(-\beta_0 - x_i'\beta_1) \end{aligned} \tag{11}$$

where $\beta_0 = 2$, $\beta_1 = -1$, $x \sim$ uniform$(0, 5)$, and $F(.)$ is the cumulative densities by assuming the following seven distributions:

- **Distribution 1**: Normal distribution: $N(0, 1)$
- **Distribution 2**: T distribution with 3 degrees of freedom
- **Distribution 3**: $\chi^2$ distribution with 3 degrees of freedom
- **Distribution 4**: Asymmetric Laplace distribution with location 0 and scale 1
- **Distribution 5**: Kurtotic distribution from a weighted mixture of normal distributions $\frac{2}{3}N(0, 1) + \frac{1}{3}N(0, 1/100)$
- **Distribution 6**: Bimodal distribution from a symmetric mixture of normal distributions $(N(-1, 4/9) + N(1, 4/9))/2$
- **Distribution 7**: Skewed distribution with a mixture of three normal distributions $(N(-22/25, 1) + N(-49/125, 9/4) + 3N(29/250, 25/81))/5$

The error terms are distributed as follows:

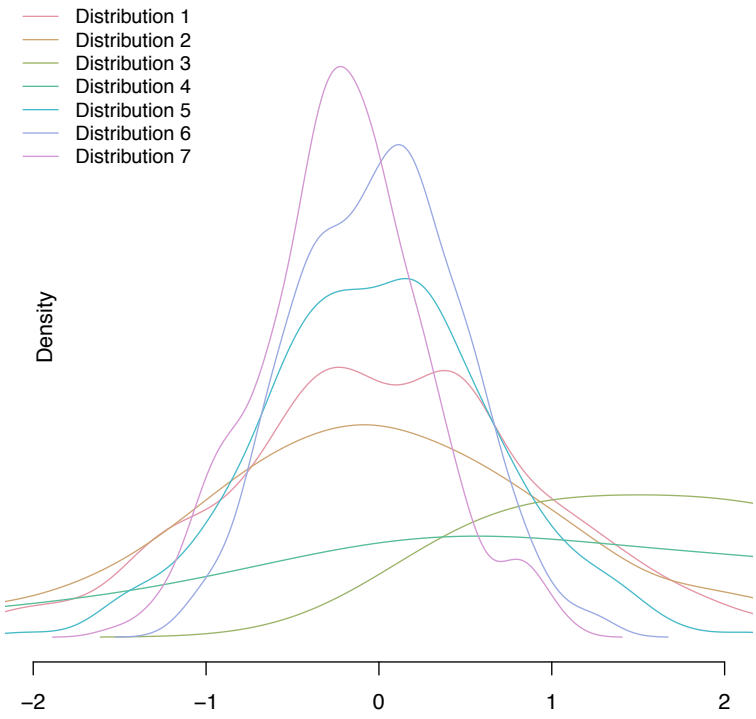


Figure 2: Empirical distributions of the error terms

# B Derivation of equations in Section 2.2

I show how to derive Equation (8) in accordance with the random utility framework, and the calculation of Equation (9) follows similarly. Following the latent linear utility specification $y^*_{ij} = \boldsymbol{x_{ij}}'\boldsymbol{\beta} + \varepsilon_{ij}$, for three alternatives $j \in \{1, 2, 3\}$ in choice set $i$, we have

$$\begin{aligned} Pr(y_{i1} = 1|\boldsymbol{x_i}) &= Pr(y^*_{i1} > y^*_{i2}, y^*_{i1} > y^*_{i3}) \\ &= Pr(\boldsymbol{x_{i1}}'\boldsymbol{\beta} + \varepsilon_{i1} > \boldsymbol{x_{i2}}'\boldsymbol{\beta} + \varepsilon_{i2}, \boldsymbol{x_{i1}}'\boldsymbol{\beta} + \varepsilon_{i1} > \boldsymbol{x_{i3}}'\boldsymbol{\beta} + \varepsilon_{i3}) \\ &= Pr(\varepsilon_{i2} - \varepsilon_{i1} < \boldsymbol{x_{i1}}'\boldsymbol{\beta} - \boldsymbol{x_{i2}}'\boldsymbol{\beta}, \varepsilon_{i3} - \varepsilon_{i1} < \boldsymbol{x_{i1}}'\boldsymbol{\beta} - \boldsymbol{x_{i3}}'\boldsymbol{\beta}) \\ &= \int_{-\infty}^{(\boldsymbol{x_{i1}} - \boldsymbol{x_{i2}})'\boldsymbol{\beta}} \int_{-\infty}^{(\boldsymbol{x_{i1}} - \boldsymbol{x_{i3}})'\boldsymbol{\beta}} f(\varepsilon_{i3} - \varepsilon_{i1}, \varepsilon_{i2} - \varepsilon_{i1}) d(\varepsilon_{i3} - \varepsilon_{i1}) d(\varepsilon_{i2} - \varepsilon_{i1}) \\ &= \int_{-\infty}^{(\boldsymbol{x_{i1}} - \boldsymbol{x_{i2}})'\boldsymbol{\beta}} \int_{-\infty}^{(\boldsymbol{x_{i1}} - \boldsymbol{x_{i3}})'\boldsymbol{\beta}} f(\varepsilon_{i3} - \varepsilon_{i1}) f(\varepsilon_{i2} - \varepsilon_{i1}) d(\varepsilon_{i3} - \varepsilon_{i1}) d(\varepsilon_{i2} - \varepsilon_{i1}) \\ &\text{(by assuming } \varepsilon_{i3} - \varepsilon_{i1} \text{ and } \varepsilon_{i2} - \varepsilon_{i1} \text{are independent).} \end{aligned} \tag{12}$$

The above calculation is also similar to what was adopted by McFadden (1974, 115), and the calculation for more alternatives is similar and thus omitted. When $\varepsilon_{ij}$ are assumed to have an independent standard type-1 extreme value distribution, the differences between $\varepsilon_{ij}$ follow the independent logistic distribution. In the CBQ model, the differences between error terms are assumed to have independent ALDs, which lead to quantile estimations.

While there are other ways to express the multinomial probabilities, the above way reduces the multinomial problem into simpler binary problems. In addition, recall the multinomial distribution with $k$ alternatives out of a total of $n$ items

$$Pr(x_1, \ldots, x_k|p_1, \ldots, p_k) = \frac{K!}{\prod_{i \in \{1,\ldots,k\}} s_i!} p_1^{s_1} \times \cdots \times p_k^{s_k}, \tag{13}$$

where $\sum s_i = K$, and $p_i$ is the probability of alternative $i$ with $\sum p_i = 1$. This is adapted to reweight the individual probabilities within each choice set so as to balance the differing numbers of alternatives, and establishes the rationale for the (weighted) multiplication of individual binary choice probabilities in the joint likelihood function in Equation (10) and the joint posterior distribution in Equation (21).

# C Plots of the CBQ estimates

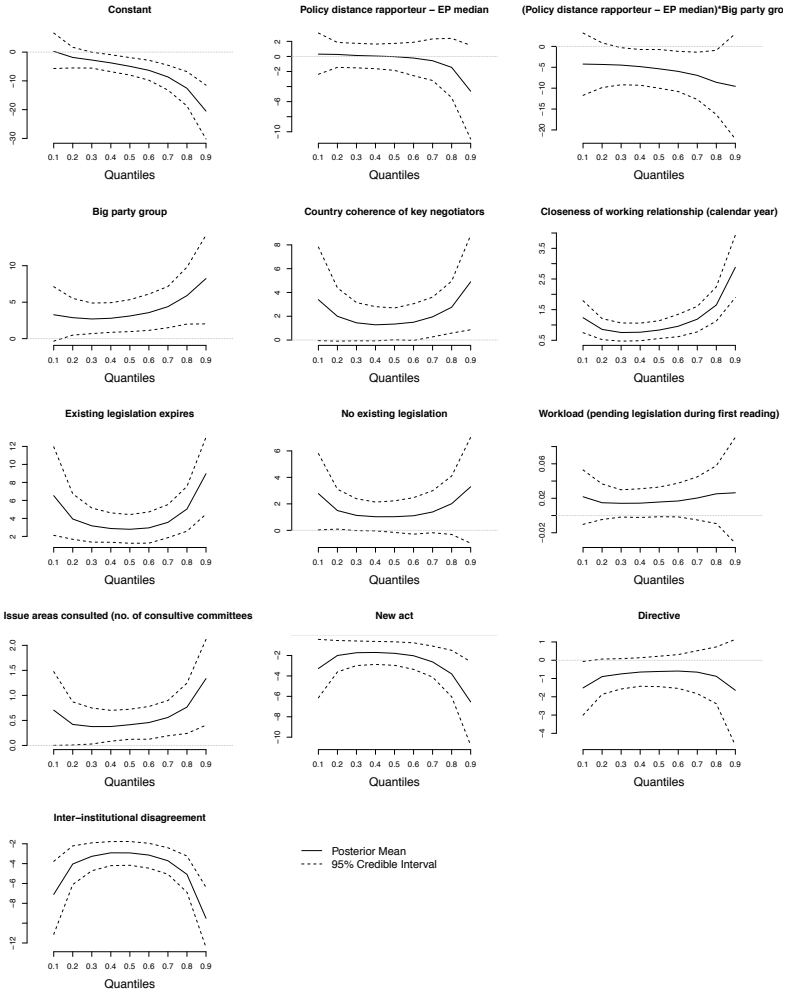


Figure 3: Estimated coefficients at the quantiles ranging from 0.1 to 0.9 for the EU legislature dataset

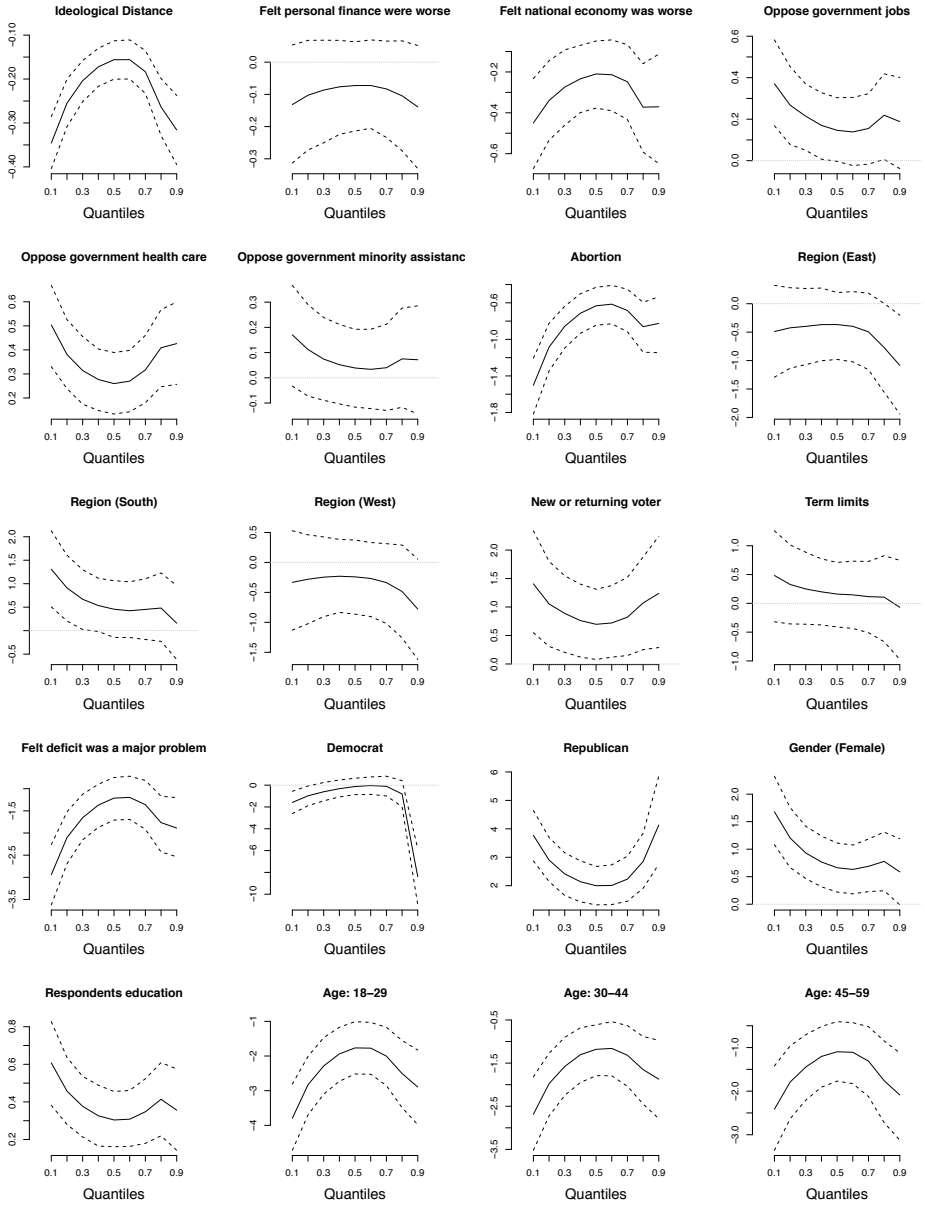


Figure 4: Estimated coefficients at the quantiles ranging from 0.1 to 0.9 for the US presidential election dataset with choice alternative Bush (Note: the estimates of variable *Ideological Distance* are the same between Bush and Clinton).

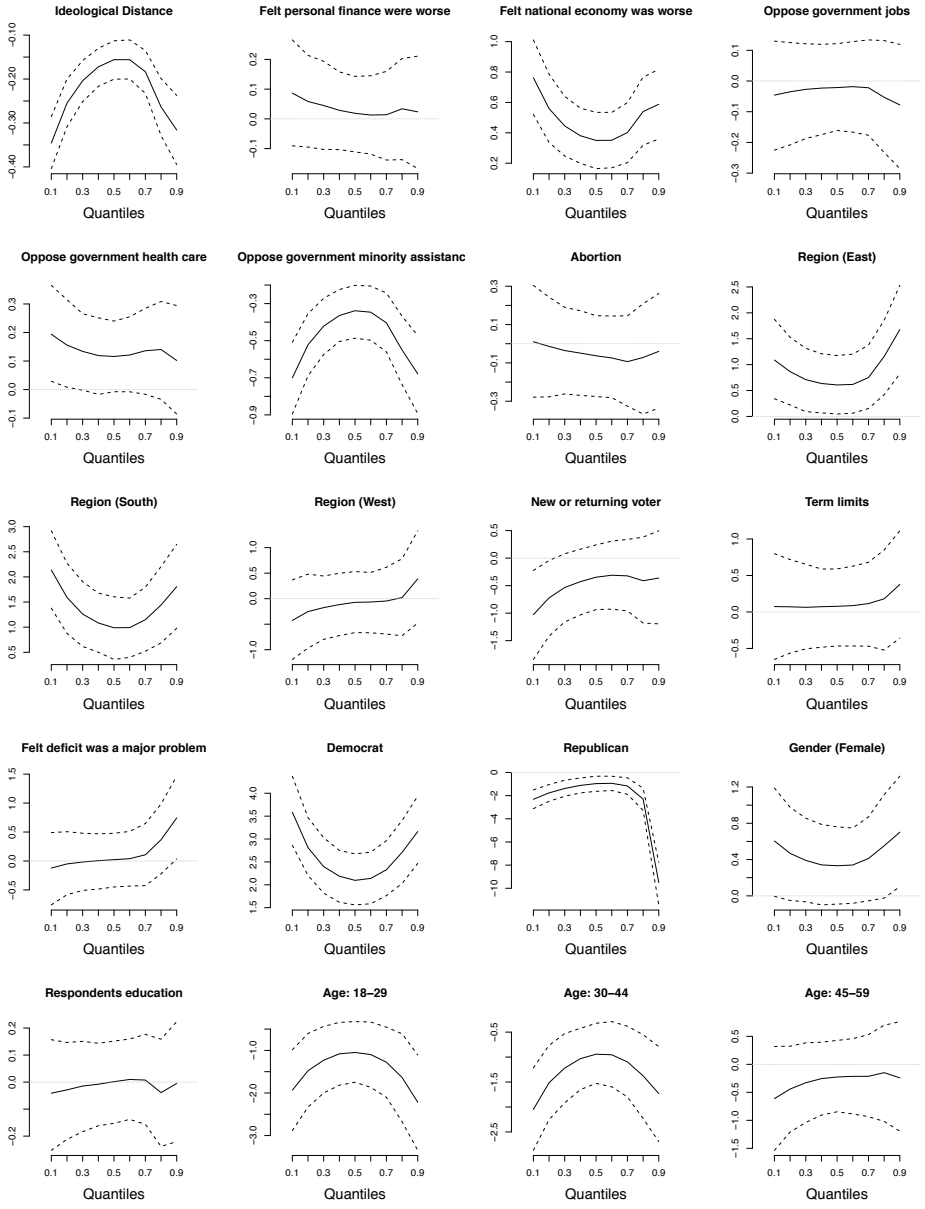


Figure 5: Estimated coefficients at the quantiles ranging from 0.1 to 0.9 for the US presidential election dataset with choice alternative Clinton (Note: the estimates of variable *Ideological Distance* are the same between Bush and Clinton).

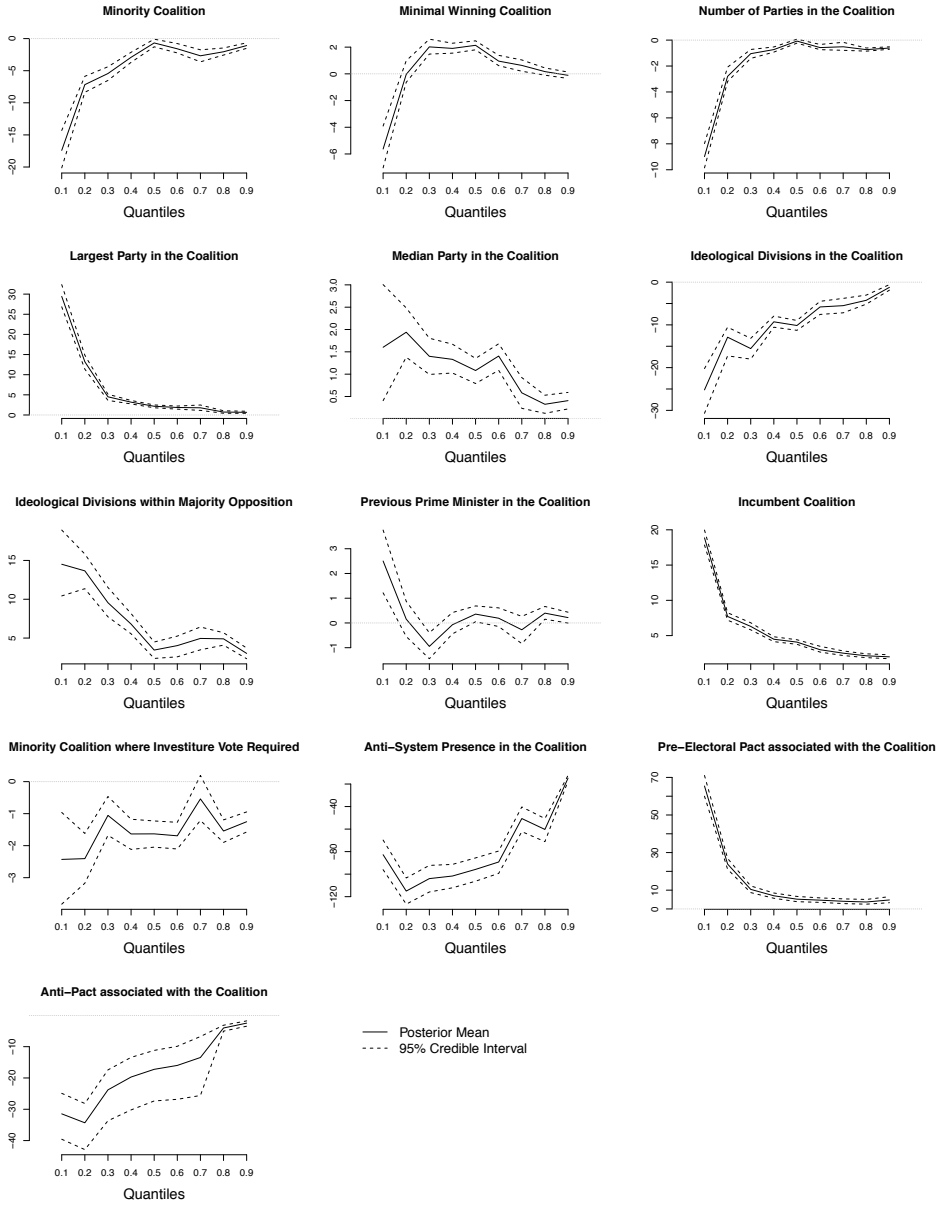


Figure 6: Estimated coefficients at the quantiles ranging from 0.1 to 0.9 for the government formation dataset