Modeling Loss Given Default Regressions

Phillip Li
Federal Deposit Insurance Corporation,
550 17th Street, NW
Washington, DC 20429
Phone: (202) 898-3501
E-mail: pli@fdic.gov

Xiaofei Zhang
Office of the Comptroller of the Currency
United States Department of the Treasury
400 7th Street SW, Mail Stop 6E-3
Washington, DC 20219
Phone: (202) 649-5556
E-mail: xiaofei.zhang@occ.treas.gov

Xinlei Zhao
Office of the Comptroller of the Currency
United States Department of the Treasury
400 7th Street SW, Mail Stop 6E-3
Washington, DC 20219
Phone: (202) 649-5544
E-mail: xinlei.zhao@occ.treas.gov

First version: 5/31/2017
This version: 5/11/2020

The authors would like to thank the excellent research support by Hugh Zhao, Peter Deffebach, Matthew Reardon, and Benjamin R. Rodriguez. The authors also thank Jon Frye, Jeremiah Richey, Yan Yu, Emily Johnston-Ross, Lynn Shibut, and the seminar participants at the OCC, The Federal Reserve Bank of Chicago, and the Federal Deposit Insurance Corporation for helpful comments. The authors take responsibility for any errors.

* The views expressed in this paper do not necessarily reflect the views of the Office of the Comptroller of the Currency, the U.S. Department of the Treasury, or any federal agency and do not establish supervisory policy, requirements, or expectations.

Modeling Loss Given Default Regressions

Abstract

We investigate the puzzle in the literature that various parametric loss given default (LGD) statistical models perform similarly by comparing their performance in a simulation framework. We find that, even using the full set of explanatory variables from the assumed data generating process where noise is minimized, these models still show similar poor performance in terms of predictive accuracy and rank ordering when mean predictions and squared error loss functions are used. However, the sophisticated parametric models that are specifically designed to address the bi-modal distributions of LGD outperform the less sophisticated models by a large margin in terms of predicted distributions. Our results also suggest that stress testing may pose a challenge to all LGD models because of limited loss data and limited availability of relevant explanatory variables, and that model selection criteria based on goodness of fit may not serve the stress testing purpose well.

1. INTRODUCTION

Loss given default (LGD) is one of the key determinants of the premium on risky bonds, credit default swap spreads, and credit risks of loans and other credit exposures, as well as a key parameter in calculating regulatory capital requirements. Despite its importance, statistical modeling of LGD has been challenging in the academic literature and in banking practice, because LGDs for commercial loans or bonds have unusual distributional characteristics. Specifically, LGD values are often bounded to the closed interval [0,1] and tend to have a bi-modal or multi-modal distribution with modes close to the boundary values, as shown in Asarnow and Edwards (1995) and Qi and Zhao (2011).[1] These characteristics make standard statistical models, such as the linear regression model, theoretically inappropriate for LGD modeling. As a result, many statistical models have been proposed in the literature to accommodate the unusual distribution of LGD.[2] Even with the sophistication of these proposed models, papers such as Qi and Zhao (2011) and Li et al. (2016) find that these models do not necessarily provide better model fit than the simpler models, such as linear regressions, when applied to real LGD data. This finding is quite puzzling, and there may be several explanations for it.

One explanation could be that the studies in the literature are based on real but noisy LGD data. Noise can come from various sources, such as omitted variables in the LGD model specification, or measurement error in LGD or the explanatory variables. As a result, the predictable portion of LGD can be overwhelmed by the noise in the unpredictable portion, regardless of the sophistication of the statistical model, which leads to similar levels of performance across the models.

Another possible explanation is that the previous studies only based their findings on a specific type of LGD prediction and model performance metric while it may be possible for their conclusions to change using different types of predictions and performance metrics. Specifically, the previous papers have mostly used estimates of the conditional mean LGD (i.e., estimates of E(LGD|X) for the corresponding parametric model of LGD|X, where X is a vector of LGD risk drivers) as predictions and assessed model performance with squared-error loss functions, e.g. sum of squared errors or mean squared error. Given that E(LGD|X) is the minimum mean squared error (MMSE) predictor of LGD|X, it is

[1] Even in each industry or debt seniority segments, LGDs still have bi-modal distribution patterns.

[2] For example, see Altman and Kalotay (2014), Bagnato and Punzo (2013), Bastos (2010, 2014), Calabrese and Zenga (2010), Hartmann-Wendels et al. (2014), Li et al. (2016), Loterman et al. (2012), Nazemi et al. (2017), Qi and Zhao (2011), Tobback et al. (2014), Tomarchio and Punzo (2019), and Yao et al. (2015).

unsurprising that these studies did not find much differentiation in model performance across models. Moreover, since $X\beta$ from a linear regression is the best linear approximation to E(LGD|X), it is foreseeable that the linear regression model from the previous studies performed well relative to the sophisticated models, even though E(LGD|X) could be nonlinear. See Angrist and Pischke (2009) for detailed expositions. This argument that alternative performance metrics can also be important in LGD modeling has been explored in the literature. For example, see Duan and Hwang (2014), Kruger and Rosch (2017), and Leymarie et al. (2018).

This paper uses a simulation framework to shed fresh light into the puzzling finding in the literature that various parametric LGD statistical models tend to perform similarly. We first generate the explanatory variables and the "true" LGD data from a zero-and-one inflated beta regression data generating process (DGP) without error terms (i.e., noise is minimized) and then fit a variety of statistical models to this dataset.[3] Estimates of the conditional means, the predicted distribution functions, and the marginal effects implied by each one of the models are then produced. Next, we introduce additional "noise" to the exercise by omitting some explanatory variables from the DGP and then we recalculate the estimates. Finally, results across different statistical models, noise levels, and various performance metrics are compared. Unlike the previous literature, our simulation framework is more comprehensive in terms of the number of models, performance metrics, and noise scenarios. Most importantly, our findings are based on a simulation framework, where we can control the level of noise, as opposed to real but potentially noisy data.

This simulation framework allows us to answer a few important questions: 1) Using conditional mean predictions and squared error loss functions, do the various parametric models perform similarly if they use the full set of explanatory variables from the DGP? 2) Do the various parametric models perform similarly using the full set of explanatory variables from the DGP based on other predictor types and performance metrics, e.g. predicted distributions and marginal effects? and 3) What is the impact of noise and sample size on the conclusions from 1) and 2)?

The predicted distributions are key quantities to study because they are of practical importance. For example, although the Basel Advanced Internal Rating-Based (AIRB) capital formula only requires

[3] Even though our simulated data can mimic the distribution of the true LGD data, one can still criticize our simulation exercise because there is no guarantee that we can replicate the true DGP for the real LGDs. We address this concern indirectly by trying many different DGPs to generate bi-modal or multimodal data with substantial mass at the 0 and 1 values; the results from these alternative simulations are discussed in Section 3.4.

the means as the input for LGDs, conservative adjustments of the parameter inputs to the capital formula are usually required by the regulators when there is uncertainty in mean estimation due to data limitations.[4] In practice, these conservative adjustments are typically based on a certain percentile or quantile of the estimated LGD distribution, so the predicted LGD distribution is useful in this situation. As another example, the LGD distribution is a major component of expected loss distributions, and estimation of loss distributions is the focus of the Basel market rules, such as the incremental risk capital and comprehensive risk measures, as well as Comprehensive Capital Analysis and Review (CCAR) and Dodd-Frank Act Stress Testing (DFAST). Percentiles of the LGD distributions, instead of the means, are typically used in such calculations. Therefore, accurately predicting the LGD distribution is of critical importance in multiple aspects of bank risk management practice.

In addition, marginal effects are crucial in the context of stress testing, because the success of stress testing depends on a model's ability to accurately estimate the impact of a large macroeconomic shock on the risk parameters, including LGD. In other words, a model that cannot measure the impact of a macroeconomic shock well but has decent performance in all other dimensions, e.g., sum of squared errors across the whole sample, may be unfit for stress testing purposes. Ex ante, it is impossible to predict which models perform better for stress testing.

In the simulation exercise we investigate seven commonly-used models: linear regression, inverse Gaussian regression with the smearing and naïve estimators (Li et al., 2016), fractional response regression (Papke and Wooldrige, 1996)[5], censored gamma regression (Sigrist and Stahel, 2011), two-tiered gamma regression (Sigrist and Stahel, 2011), inflated beta regression (Ospina and Ferrari, 2010), and beta regression (Duan and Hwang, 2014).[6] A schema for all the models that we considered along with references are in Appendix A. The standard linear regression (OLS) is the only model that cannot restrict the predicted mean values within the [0,1] range or address the bi-modal distribution pattern. The inverse Gaussian regression with a smearing estimator (IG smearing) and fractional response regression (FRR) ensure that the mean predictions will fall in the interval [0,1], but these models are not specifically

[4] In AIRB implementation, bucket-level mean LGDs are used in some cases, while conditional mean LGDs from loan-level models are used in other cases. The results of our paper are useful for the latter cases.

[5] The FRR is technically a semi-parametric method while the other models are parametric. We chose to include FRR because it is a commonly used regression in practice.

[6] Note that the "beta regression" from Duan and Hwang (2014) is not the same as the one from Ferrari and Cribari- Neto (2004). See Duan and Hwang (2014) for the details and motivation.

designed to handle bi-modal distributions.[7] The remaining four models, censored gamma regression (CG), two-tiered gamma regression (TTG), inflated beta (IB), and beta regression (BR) are all sophisticated and designed specifically to address the bi-modal distribution of LGD; their mean predictions are also inside the [0,1] interval. Among the four, TTG and IB are more complicated as they involve more parameters and structure, and TTG is particularly challenging to fit. Because of space limitations, we do not include in this study other statistical methods that have been used in LGD modeling in the literature, such as regression trees (Qi and Zhao, 2011) and support vector regression (Yao et al., 2015).

The results in this paper are not restricted to LGD modeling and can be generally applied to situations where either the outcome or explanatory variable has a bi-modal pattern. For example, exposure at default, another important risk parameter in banking practice, also has a bi-modal distribution (see Jacobs, 2010 and Tong et al., 2016). The results in this paper can be applied to exposure at default as well.

This paper proceeds as follows. In Section 2, we describe the simulation framework and the predicted quantities of interest. We present simulation results in Section 3 and conclude in Section 4.

2. SIMULATION DESIGN, PREDICTED DISTRIBUTIONS, AND PREDICTED MARGINAL EFFECT

2.1 Data generating process (DGP)

Our LGD data are generated using a zero-and-one inflated beta regression model (see Ospina and Ferrari, 2010 and Li et al., 2016). We simulate a total of 400,000 observations, with 40 time periods and 10,000 observations in each period. The economic interpretation is that we observe 10,000 defaults in each period over 40 periods.[8] The 400,000 observations are independent conditional on a common macroeconomic factor. We assume conditionally independent LGD observations, because LGDs are correlated largely due to the impact from macroeconomic factors. Note that our dataset does not have a traditional panel or longitudinal structure as we typically do not observe LGDs for a set of firms over time. A firm only appears in the dataset when it defaults. Also, even though real LGD datasets in financial institutions have at most a few thousand observations, we use a larger sample size in the simulation to illustrate an ideal

[7] We mainly investigate IG smearing in this paper, because Li et al. (2016) show that the smearing estimator makes the inverse regression method more stable.

[8] In reality, there are more defaults and thus more LGD observations during economic downturns relative to benign periods. We simulate 10,000 observations every period, regardless of whether it is a normal period or economic downturn period, to increase model fit and to understand the challenges we would face under this ideal situation.

case where LGD data are abundant. We reduce the sample sizes to investigate the impact of sample size on model estimates in later subsections.

We have 11 explanatory variables in the DGP, including a constant ($x_{i1} = 1$), a macroeconomic factor (x_{i2}) set to the actual quarterly national unemployment rates from 2006 to 2015,[9] and 9 normally distributed explanatory variables ($x_{i3}, \dots, x_{i11}$). Each explanatory variable in ($x_{i3}, \dots, x_{i11}$) has a marginal $N(0,0.5^2)$ distribution and is generated to have a correlation of 0.05 with the macroeconomic factor. We introduce the correlation by assuming that each element in ($x_{i3}, \dots, x_{i11}$) has a correlated bivariate normal distribution with the macroeconomic factor, and we generate each element from its implied conditional distribution, e.g., $x_{i3}|x_{i2}$, using a copula function with sample averages and sample variances of the unemployment rate as the means and variances for x_{i2}. This correlation is introduced to make the stress testing and noise impact exercises more realistic.

The zero-and-one inflated beta regression model for the i-th LGD observation is

$$\Pr(LGD_i = 0) = P_0^i, \tag{1}$$

$$\Pr\big(LGD_i \in (l, l + dl)\big) = \big(1 - P_0^i - P_1^i\big) f\big(l; \mu^i, \phi\big) dl, \tag{2}$$

$$\Pr(LGD_i = 1) = P_1^i \tag{3}$$

for $l \in (0,1)$, where $0 < \mu^i < 1$, $\phi > 0$, and $f(\cdot)$ is the probability density function (PDF) of a beta random variable with two parameters:

$$f\big(l; \mu^i, \phi\big) = \frac{\Gamma(\phi)}{\Gamma(\mu^i\phi)\Gamma\big((1-\mu^i)\phi\big)} l^{(\mu^i\phi-1)}(1-l)^{(1-\mu^i)\phi-1}.$$

This parameterization of the beta distribution follows Ferrari and Cribari-Neto (2004). The vector of explanatory variables, $\overrightarrow{x_i} = (x_{i1}, x_{i2}, \dots, x_{i11})$, is linked to the model through the following equations:

$$P_0^i = \frac{e^{\overrightarrow{x_i}\alpha}}{1 + e^{\overrightarrow{x_i}\alpha} + e^{\overrightarrow{x_i}\beta}} \tag{4}$$

$$P_1^i = \frac{e^{\overrightarrow{x_i}\beta}}{1 + e^{\overrightarrow{x_i}\alpha} + e^{\overrightarrow{x_i}\beta}} \tag{5}$$

[9]Although the x_{i2} notation suggests that there could be a different macroeconomic factor for each observation i, this is not the case. Since there are 40 quarters of unemployment data from 2006 to 2015, we set the first 10,000 values (x_{i2} for $i = 1, \dots, 10000$) equal to the 2006 Q1 unemployment rate, the next 10,000 values to the 2006 Q2 unemployment rate, and so on until the last 10,000 observations is the unemployment rate for 2015 Q4; in other words, there is a common macroeconomic factor for every 10,000 observations.

$$\mu^i = \frac{e^{\overrightarrow{x_i}\gamma}}{1 + e^{\overrightarrow{x_i}\gamma}} \tag{6}$$

We set the true parameter values as $\alpha = (-0.54, -5, 0.4, \dots, 0.4)$, $\beta = (-1.46, 6, -0.1, \dots, -0.1)$, $\gamma = (0, 0.5, -0.1, \dots, -0.1)$, and $\phi = 5$, and we generate 400,000 LGD_i observations according to (1)-(6). The parameterization for the inflated beta regression in (1)-(6) follows Li et al. (2016) and Yashkir and Yashkir (2013).

Note that there are no error terms in Equations (4) and (5), and the only noise in our DGP is the difference between the latent variable from (4) and (5) and the outcome variable, which has values 0 and 1. As a result, the noise level is minimized in our DGP.

We refer to this set of observed LGD and explanatory variables as the true LGD and explanatory variables from the DGP, and we refer to the inflated beta distribution implied by (1)-(6) as the true distribution or model. Also, the true quantile functions and marginal effects are the ones derived from (1)-(6).

A histogram of the true LGD data is in Figure 1. It has a tri-modal pattern with substantial observations in the two extreme ends and a third mode close to the center. A third mode near the center is included because there is some degree of tri-modality in certain empirical LGD distributions (see for example, Altman and Kalotay, 2014); our conclusions do not change qualitatively whether a bi-modal or tri-modal distribution is used. We use maximum likelihood estimation on the log likelihood implied by (1)-(6). See Li et al. (2016) for the explicit form of this likelihood function and estimation details. Denote the estimates as $\hat{\phi}$, $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\gamma}$. Mean estimates are generated by plugging in the maximum likelihood estimates into the analytic mean functions.

We can add additional noise at the modeling step by using a subset of the true explanatory variables from the DGP when fitting the various LGD models. Using the full set of explanatory variables is necessary to be consistent with the DGP, so dropping at least one of them would result in omitted variables or noise. We drop either four or eight of the non-macroeconomic explanatory variables in this paper.

2.2 Predicted distributions

2.2.1 Unconditional distributions

We estimate the distribution of LGD unconditional on the explanatory variables using simulation output. In general, we first sample the explanatory variables and then simulate LGD conditional on the

explanatory variables according to the model of interest. We repeat this procedure many times and retain only the LGD draws. These retained draws are from the marginal distribution of LGD, unconditional on the explanatory variables, with which we use to estimate or test the desired distribution.

Specifically, we take 1,000 independent draws of the vector of explanatory variables and then, for each draw, we generate 1) a realization of LGD based on the true distributional assumption and 2) a realization from the model-predicted distributions. For example, for linear regressions, for each draw of the vector of explanatory variables, $\overrightarrow{x_i}$, we simulate 1) a realization of LGD from the zero-and-one inflated beta regression using the true parameter values and then 2) a realization from $N(x_i\,\hat{\beta}, \widehat{\sigma^2})$, where $\hat{\beta}$ and $\widehat{\sigma^2}$ are estimates from the linear regression. Using these draws, we plot and compare their histograms and estimated cumulative distribution functions (CDF).

This process can be easily implemented for IG smearing, CG, TTG, IB, and BR, based on their model structure and estimated parameters. However, estimating the predicted distribution for FRR is a challenge as it only specifies up to the mean function $E(LGD_i|\overrightarrow{x_i})$ and not the whole distribution. In order to generate a distribution consistent with FRR, we assume a beta distribution for LGD. That is, for each observation, we draw a realization from an assumed beta distribution, $beta(\alpha, \beta_i)$. Since there is no theoretical value for the beta distribution parameters, we try three different values for α: 0.5, 1, and 5. In addition to these α values, we empirically estimate α by fitting a beta distribution to the simulated dataset and matching the first two moments. Once α is known, we solve for β_i for each observation, based on the property that the i-th fitted value from FRR should be equal to the mean of the beta distribution for observation i. That is, we solve for β_i in the equation $\hat{E}(LGD_i|\overrightarrow{x_i}) = \frac{\alpha}{\alpha+\beta_i}$.[10] We then plot the 1,000 draws from this derived $beta(\alpha, \beta_i)$ distribution.

2.2.2 Predicted conditional distributions

2.2.2.1 Kolmogorov-Smirnov (KS) box plots

In this section, we compare the model-predicted and the true conditional distributions for each observation. We first take 5,000 independent draws of the vector of explanatory variables, and then for

[10] This method of assuming a beta distribution for LGD and estimating the parameters of the mean function using FRR is mentioned in page 620 of Papke and Wooldridge (1996). See Li (2018) for a Bayesian application of this method.

each of these draws, we generate 1,000 draws of LGD based on the model of interest, and calculate their modified Kolmogorov-Smirnov (KS) statistic and p-value relative to the DGP. We then plot the distributions of the 5,000 KS statistics and p-values using boxplots. A model with a predicted conditional distribution close to the conditional distribution from the DGP will have KS statistics closest to 0, and a wilder dispersion in the KS statistics would point towards inferior model performance.

These modified KS statistics and p-values correspond to the mixed Kolmogorov-Smirnov test from Dimitrova et al. (2017). The commonly-used Kolmogorov-Smirnov tests (Serfling, 1980) assume that the underlying distributions being tested are continuous, and thus they would be inappropriate for our mixed discrete-continuous LGD application. In contrast, the tests from Dimitrova et al. (2017) allow for the underlying distributions to be continuous, purely discrete, or mixed discrete and continuous. These are one-sample tests that compare the analytic conditional distributions from the DGP against the draws from the assumed model. In this paper, we only use the statistics and p-values from Dimitrova et al. (2017) and refer to them as either the modified KS statistics or simply the KS statistics.

This process can be easily repeated for IG smearing, CG, TTG, IB, and BR. Constructing the conditional distribution for FRR is again subject to the assumption of α in the beta distribution.

2.2.2.2 Quantile plots

Similar in spirit to the previous subsection, we compare the quantile functions between the estimated models and the DGP. This exercise was suggested in Sigrist and Stahel (2011). To calculate the quantiles, we essentially invert the analytic CDFs implied by the PDFs for each model, and then we plot the quantile functions for a specific quantile as a function of a single explanatory variable, with the remaining explanatory variables set at their sample means. We choose the 0.2, 0.4, 0.6, and 0.8 quantiles, and vary them as a function of x_{i3} from -6 to 6. A model with a predicted quantile function close to the quantile function from the DGP would indicate a good model.

2.3 Marginal effect

We numerically compute the average marginal effect of each model with respect to the macroeconomic factor, x_{i2}. Specifically, the average marginal effect without any omitted variables at each point in time is

$$\frac{1}{10000}\sum_{i=1}^{10000}\frac{\hat{E}(LGD_i|(x_{i1},x_{i2}+h,\dots,x_{i11}))-\hat{E}(LGD_i|(x_{i1},x_{i2},\dots,x_{i11}))}{h}$$

where $h = 0.0001$. The marginal effects with omitted variables are calculated the same way except that the set of explanatory variables in the conditioning set is appropriately reduced. We calculate these marginal effects when the macroeconomic factor varies between 4% to 10% in Figure 8.

3. RESULTS

3.1 Mean predictions

We first discuss the mean prediction results using the full set of explanatory variables from the DGP, followed by the mean prediction results when four and eight variables are omitted from the models.

To demonstrate that we are able to recover consistent maximum likelihood estimates, we present the estimates of the IB model in Panel A of Table 1. As expected, the coefficient estimates are quite close to the true values of the parameters used in the DGP, and the standard errors are small. This shows that we can recover the true values from the DGP quite well using IB and the full set of explanatory variables.

Panel B of Table 1 reports various performance metrics for the mean predictions using the full set of explanatory variables: sum of squared errors (SSE), R-squared (R^2), Pearson's correlation, Spearman's Rho, and Kendall's Tau. First, as expected, IB has the lowest SSE value of 50131 among all the models. Second, the SSE of the linear regression is 50161, which is lower than all the models except for IB. Therefore, the linear regression is not necessarily the most inappropriate model to use for LGD, assuming that the researcher is only interested in obtaining mean predictions and evaluates model performance using SSE and R^2. Third, even though all the models, including IB, use the full set of explanatory variables from the DGP, the R^2 metrics are very low at around 8%. Therefore, contrary to some of the literature and banking practice, we cannot gauge LGD model fit based on the magnitude of R^2 alone. For instance, an R^2 of 8% for a particular model might be interpreted as very low, but from this exercise, even the IB model using the full set of explanatory variables from the DGP cannot get an R^2 of above 8%. All in all, Panel B of Table 1 indicates that the various models perform very similarly under these two types of performance metrics. This finding is consistent with the existing literature and suggests that the mean predictions across these models perform similarly when assessed with squared error and rank ordering loss functions.

Furthermore, Figure 2 depicts a histogram of the predicted means used in Panel B of Table 1. It is clear that all histograms in this figure are bell-shaped and do not have mass at the boundaries. This finding is consistent with Qi and Zhao (2011) and Li et al. (2016).

The aforementioned findings from Panel B of Table 1 and Figure 2 may be initially unintuitive to some because the fitted IB model using the full set of explanatory variables does not appear to "fit" the data very well despite being consistent with the DGP. For example, one might ask: why is the SSE not closer to zero and why are the predictions not closer to the realized values when the parameter estimates from Table 1 Panel A are so close to the true values? We explain this finding in the context of the simple logistic regression. Assume the DGP is $LGD_i|X_i \sim Bernoulli(\exp(X_i\beta)/(1+\exp(X_i\beta))$, where $X_i = 0.5$ and $\beta = 1$. Plugging these values in, we know that theoretically $LGD_i|X_i = 0.5 \sim Bernoulli(0.6225)$. An observed value of $LGD_i|X_i = 0.5$ would be a realization from the distribution $Bernoulli(0.6225)$, for example 0. Now, even in the best case scenario in which we could perfectly estimate the unknown parameters β to be $\hat{\beta} = 1$, our conditional mean estimate for this LGD value would be $\exp(X_i\hat{\beta})/(1+\exp(X_i\hat{\beta})) = 0.6225$, which is obviously not equal to the realized LGD value of 0. From this example, it should be obvious that even though if we could perfectly estimate the unknown parameters and knew the true parametric form of the model (i.e., the Bernoulli distribution), the conditional mean estimates do not need to be very "close" to the realized values due to randomness in the realizations. Also, because means are measures of central tendency, they are typically closer to the "center" of the distribution and thus away from the LGD boundary value, which explains why there aren't values close to 0 or 1 in Figure 2.

We introduce additional noise into the model specification by dropping four explanatory variables and refitting the various models. The results are reported in Panel C of Table 1. Unsurprisingly, the performance metrics decline in every dimension from Panel B to Panel C of Table 1, when some relevant explanatory variables are omitted. The decline occurs at about the same rate across various models, and again, the performance metrics do not differ much across the various models in Panel C. We do not report the performance metrics when fewer or more explanatory variables are omitted from the specifications due to space constraints, but the qualitative conclusions from Panel B to C of Table 1 remain the same. This finding provides additional evidence for the findings in the literature that the mean predictions from these models perform very similarly in terms of the squared error and rank ordering loss functions.

As a different visualization of the results from our noise exercises, Figure 3 depicts, for each method, the kernel density plots of the predicted means when zero, four, and eight explanatory variables are omitted from the model specification. It is clear from Figure 3 that the distributions are again bell-shaped for each case of omitted variables and for each model. Furthermore, the empirical distributions of the predicted means become more concentrated towards the unconditional empirical average of LGD, when more explanatory variables are omitted. Because the percentiles of the distributions of predicted mean LGDs are often used as inputs to capital formulas, our results suggest that these percentiles are likely to be underestimated when there are omitted explanatory variables. This is a common challenge in empirical work and banking practice.

3.2 Predicted distributions

3.2.1 Predicted unconditional distributions

This section compares the predicted unconditional distributions using histograms, estimated CDFs, and KS statistics.

Figure 4 illustrates the predicted unconditional distributions from the various models. In Panel A, we depict the distributions from the six models: OLS, IG smearing, CG, TTG, IB, and BR. Unsurprisingly, this panel shows that the predicted unconditional distribution from the linear regression has a bell shape, the IG smearing method produces a bi-modal pattern, and all the other models show tri-modal patterns. IB, the true model, produces the predicted unconditional distribution that is closest to the true distribution, as depicted in Figure 1. Panel B shows various distributions for FRR resulting from using different values of α. The shapes of the distributions are clearly quite different for different values of α. Smaller values of α lead to distributions that are more bi-modal. The estimated value of $\alpha = 0.0051$ leads to the most extreme bi-modal distribution with high peaks at both ends but little mass in between. None of the choices for α yield the tri-modal distribution as illustrated in Figure 1. The results from Panel B suggest that the FRR is very sensitive to the choice of α, and because there is no theoretical basis to determine this parameter, FRR has a clear disadvantage over the other models when predicted distributions are needed.

It is difficult to assess the similarity or the differences between the distribution from the true data and the predicted unconditional distributions from various models based on the figures in Panel A of Figure 4. So, we report the modified KS statistics of various models in Table 2. We do not include FRR here, as it is very sensitive to the α parameter, and any choice of α might be difficult to justify.

Table 2 contains the KS statistics from the comparisons of the predicted unconditional distributions against the true unconditional distribution from the DGP. Using the full set of explanatory variables, the results from the first column suggest that the predicted unconditional distributions of LGD generated by the IG smearing and linear regression models differ the most from the true unconditional distribution. This is unsurprising as these two models do not accommodate multi-modal distributions, especially the positive probability masses at LGD values of 0 and 1. Furthermore, from the rest of the results in the first column, the IB model has the lowest modified KS statistic, and the CG, TTG, and BR models all have very small KS statistics. This suggests that the "sophisticated" CG, TTG, and BR models are able to capture the unconditional distributions reasonably well, despite not having the correct distributional assumption as the DGP. The other columns show that, when some variables are dropped from the full set of explanatory variables, there is little change in the modified KS statistics across various models, which is to be expected for LGD distributions unconditional on the explanatory variables.

Since the modified KS statistics cannot fully capture differences across the entire distribution, we plot the CDFs in Figure 5. The models in Panel A use the full set of explanatory variables, and this figure corresponds to the modified KS statistics in the first column of Table 2. It is clear from this panel that the predicted unconditional distributions from the linear regression and IG smearing are quite different from the true distribution which explains the large KS statistics. Furthermore, the CDFs from the true and predicted distributions from CG, TTG, IB, and BR are quite similar, which is consistent with the small modified KS statistics from Table 2. Although IB is the correct distributional assumption, its advantage over CG, BR, and TTG is rather minor in this figure.

Panel B of Figure 5 compares the true unconditional distribution and the predicted unconditional distributions from the various models when four explanatory variables are omitted; this figure corresponds to the KS statistics in the second column of Table 2. We can see that the predicted unconditional distributions from the linear regression and IG smearing are again quite different from the true distribution. The CDFs for CG, TTG, IB, and BR are once again quite close in this panel, suggesting that these four models generate predicted unconditional distributions that are quite similar. Again, this is the expected result as these are LGD distributions that are unconditional on the explanatory variables. The similarity of Panels A and B of Figure 5 indicates that we cannot rely on unconditional predicted distributions to assess variable selection.

3.2.2 Predicted conditional distributions

3.2.2.1 KS boxplots

This section compares the predicted conditional distributions of LGD against the true conditional distribution using results from the modified KS tests (Dimitrova, 2017). We do not include the FRR model.

Panels A and B of Figure 6 show the distributions of the KS statistics and p-values from our 5,000 KS tests using the full set of explanatory variables. From Panel A, it is clear that the distribution of KS statistics for the linear regression and IG smearing models are centered quite far away from 0, which suggests that the predicted conditional distributions generated by the linear regression and IG smearing models are dissimilar to the true conditional distribution. The analogous KS statistic distributions for the sophisticated models (i.e., CG, TTG, IB, and BR) are centered much closer 0, which suggests that the predicted conditional distributions for the sophisticated models are similar to the true conditional distribution. Panel B shows that the distributions of p-values for the linear regression and IG smearing are essentially degenerate at 0. By contrast, the distributions corresponding to the sophisticated models are centered away from 0, which means that there is not enough evidence in the data to suggest that the conditional distributions for the sophisticated models are statistically different from the true conditional distribution.

Panels C and D of Figure 6 show the distributions of the KS statistics and p-values from our KS tests when four explanatory variables are dropped from the full set. Inhe IB model does not seem more similar to the true distribution relative to CG, TTG, and BR. We tried dropping other numbers of explanatory variables, but the results are qualitatively similar to those in Panel C. In summary, we find that the sophisticated models all behave similarly when some of the explanatory variables are dropped from the model. Furthermore, we find that, in terms of the similarity with the true conditional distribution, even with the majority of explanatory variables dropped and the sophisticated models not performing well, these sophisticated models still outperform the linear regression and IG smearing models with the full set of explanatory variables.

3.2.3 Quantile plots

We next depict the quantile plots for the various estimated models and compare them against the true model. To save space, we only plot the 0.2, 0.4, 0.6, and 0.8 quantiles without dropping any explanatory variables in Figure 7. We do not include FRR for the same reason as in the previous sections.

In all four panels of Figure 7, the linear regression and IG smearing quantiles stand out as they are not similar to the true quantiles. The quantiles for the sophisticated models (i.e., CG, TTG, IB and BR) are rather close to each other, and the quantile function for IB is almost entirely on top of the true quantile. The latter result is not surprising, as we have shown earlier that the IB model using the full set of explanatory variables can recover the true DGP quite well.

We do not report the quantile results when some explanatory variables are dropped from the models due to space limitations, so we briefly discuss the results here. With some explanatory variables dropped, the IB quantile plots show the most shift, and as a result, the IB model no longer substantially outperforms the other three sophisticated methods (CG, TTG, and BR). Also, similar to the previous results, the linear regression and IG smearing quantiles always deviate much from the true quantile.

In summary, we find convincing evidence that the sophisticated models produce conditional distributions that are much more similar to the true conditional distribution than both the linear regression and IG smearing models. Also, the performance difference between the four sophisticated models is not large, especially when there are missing explanatory variables from the model specification.

3.3 Marginal effect

The marginal effect results using the full set of explanatory variables are depicted in Panel A of Figure 8. This figure shows that the true marginal effect is slightly upward sloping. The marginal effect curves for both the linear regression and FRR are flat, and both methods show low macroeconomic sensitivity of LGD even at the 10% level of the macroeconomic factor which we define as the stressed scenario. Furthermore, the IB marginal effect curve is very close to the true marginal effect curve, which is not surprising, as the IB model with the full set of explanatory variables can recover the true coefficients quite well with 400,000 observations (see Table 1). All remaining methods over-estimate the marginal effect. IG smearing shows the most over-estimation followed by BR. When the macroeconomic factor is at 10%, the marginal effect is about 1.55 for IG smearing and 1.35 for BR, while the true marginal effect is roughly 1.25.

Panel B of Figure 8 depicts the marginal effects when four variables are dropped from the models. This panel shows that all models, except for IG smearing, under-estimate the true marginal effect. In addition, the IB model does not outperform CG, TTG, and BR, while the linear regression and FRR still yield flat marginal effect curves. We do not report results when more or fewer explanatory variables are

dropped from the model due to space constraints, but we briefly describe the results here. We find that, the more explanatory variables we drop, the larger the gap between the true marginal effect and the estimated marginal effect. This finding poses a serious challenge to stress testing in practice, because it is likely that only a subset of the key risk drivers that are important for stress testing are observed by practitioners.

We conduct more analysis for IG models in Figure 9 to see if IG smearing always over-estimates the marginal effect as Figure 8 suggests. In this figure, we also plot the marginal effects for the inverse Gaussian with naïve estimator (IG naïve) from Li et al. (2016), which is a commonly-used estimator that is very similar to IG smearing. We plot the IG smearing and IG naïve average marginal effects across the whole range of the macroeconomic variable for 15 new sets of random data, where for each set of data we randomly select the values for the true parameters ϕ, α, β, and γ, and randomly draw a new set of full explanatory variables like in Section 2.1. The correlation between the macroeconomic factor and the other explanatory variables is 0.05, and we use 40,000 observations in this exercise.[11]

We can draw several conclusions from Figure 9. First, there is typically a large gap between the true marginal effect and the IG marginal effects, and the IG smearing method under-estimates the marginal effect in each one of these 15 sets of results. Therefore, the over-estimation we observe from Figure 8 is not robust, and the IG methods generally cannot capture the true marginal effect well. Second, the smearing estimator does not seem to add value in terms of capturing the true marginal effect. For most of the new sets of random data, the IG smearing is worse than the IG naïve in terms of predicting the marginal effect. Interestingly, Li et al. (2016) find that IG smearing improves IG model fit in terms of SSE and conditional mean LGD predictions. However, the results in Figure 9 suggest that a method predicting the mean better, in this case IG smearing, is not necessarily better at capturing the marginal effects. Therefore, using SSE with conditional mean LGD predictions may not be the best model evaluation strategy for stress testing models.

3.4 Further investigations

3.4.1 Alternative DGPs

The exercises in this subsection aim to address the concerns that the DGP in the main results may not mimic the complexity in real data and that our findings may be restricted to the specific set of parameters

[11] This can be interpreted as observing 1000 defaults in each period over 40 periods.

we choose. We tried several other DGPs, including the IB model with different parameter values and the TTG model. In order to increase the similarity with real data, we also generate explanatory variables with moments matching the explanatory variables from Moody's URD data used in Li et al. (2016), and use their estimated IB and TTG parameters. We account for the correlations between the loan–level variables and the macroeconomic variables by resampling with replacement from the real dataset. In all these exercises, we use a total of 400,000 observations with noise in the DGP.

Results from the alternative DGPs are very similar to those reported in the previous sections. That is, there is little variation in predictive accuracy and rank ordering ability across all models investigated in this paper. Therefore, if the main focus is model fit in terms of SSE or rank ordering using conditional mean predictions, all models show similar performance. In addition, using the full set of explanatory variables, the four sophisticated models are able to generate predicted conditional distributions more similar to the true conditional distributions and show higher sensitivity to the macroeconomic factor, relative to the simpler models.

Recovering the true parameters is more difficult for some DGPs. Among all the performance metrics we have investigated, the marginal effects show the most sensitivity to parameter estimates. For some DGPs, even using the full set of explanatory variables with the true distributional assumptions, we cannot recover the true parameters very precisely, in which case even the true models do not appear to be better than the other sophisticated models in terms of accurately capturing the marginal effects. Also, we find similar problems with accurately capturing the marginal effects as in previous sections when key explanatory variables are missing, which suggests that this conclusion is robust for different DGP assumptions.

3.4.2 Different number of observations in the DGP

We also tried simulating data with a different number of observations. The purpose of such an exercise is to investigate whether the sophisticated models perform similarly when the sample size is small. Banks typically only have between 1,000 to 4,000 internal LGD data points, which is a lot less than the number of observations we used in previous sections.

We find that the four sophisticated models still out-perform the less sophisticated models by a large margin even for small samples. However, it is more difficult to recover the true parameters within a small sample, even with the full set of explanatory variables and the true distributional assumptions. As a result, similar to our findings from the previous section, the marginal effects show the most sensitivity to

small sample sizes. When the sample size is a few thousand, we do not observe any advantage in using IB or TTG over CG or BR, even though IB and TTG are the true models with a full set of explanatory variables. This problem is particularly severe for TTG, because parameter estimation is exceptionally challenging.[12]

3.4.3 Alternative approach to add additional noise

We also tried a second way to add noise. That is, we add "error terms" or random quantities to equations (4)-(6). For example, instead of generating the data according to (4)-(6), we generate the data according to $P_0^{i*} = \frac{e^{\overrightarrow{x_i}\alpha}}{1+e^{\overrightarrow{x_i}\alpha}+e^{\overrightarrow{x_i}\beta}} + z_{0i}$, $P_1^{i*} = \frac{e^{\overrightarrow{x_i}\beta}}{1+e^{\overrightarrow{x_i}\alpha}+e^{\overrightarrow{x_i}\beta}} + z_{1i}$, and $\mu^{i*} = \frac{e^{\overrightarrow{x_i}\gamma}}{1+e^{\overrightarrow{x_i}\gamma}} + z_{01i}$, where z_{0i}, z_{1i}, and z_{01i} are unobserved random terms that are unaccounted for during the fitting of our models.[13] Because we do not account for these terms during estimation, this approach can be thought of as a different type of misspecification or noise. We find the same qualitative conclusions as before and do not report the results due to space constraints.

4. CONCLUSIONS

We compare via a simulation exercise seven parametric models to estimate LGDs: linear regression, inverse gamma regression with a smearing estimator (IG smearing), fractional response regression (FRR), censored gamma regression (CG), two-tiered gamma regression (TTG), inflated beta regression (IB), and beta regression (BR). The last four of these models are designed specifically to address the bi-modal distribution unique to the LGD data.

We find that, even using the full set of explanatory variables from the DGP without error terms (i.e., noise is minimized), the mean predictions from the various models, including the true model from the DGP, perform very similarly and poorly in terms of both predictive accuracy and rank ordering. Moreover, when we introduce additional noise, both predictive accuracy and rank ordering ability unsurprisingly decline across all models, but various models still perform similarly in these two dimensions. Therefore, the finding in the literature that model fit across different LGD models cluster in

[12] Even when it is the true model, TTG often underperforms the other sophisticated models when the sample size is small, e.g., in thousands.

[13] In our simulation exercises, each one of these unobserved random terms is generated from a standard normal distribution.

very narrow and poor ranges is robust and not driven by omitted explanatory variables or noise in the data. If the only focus of LGD modeling is in producing mean predictions, then all models investigated in this paper can serve that purpose reasonably well.

However, we argue that, in addition to predicative accuracy and rank ordering, we should also investigate predicted LGD distributions from various models, because the LGD distribution is important in various aspects of risk management in the banking industry. Based on predicted conditional distributions, the four sophisticated models, CG, TTG, IB, and BR, show similar levels of performance, outperforming the linear regression and IG smearing by a large margin. On the other hand, because FRR is only focused on estimating the mean LGD but does not have other assumptions about the underlying parametric structure, generating the predicted distributions under FRR involves much uncertainty, and it is difficult to assess the performance of FRR based on predicted distributions. Because of this uncertainty, in circumstances when knowledge about LGD distributions becomes critical, we conclude that FRR is not the most appropriate model to use.

Further, we assess the marginal effects generated from various models given their critical importance in stress testing. We find that, with missing explanatory variables or when the sample size is small, none of the models, including the true model, can accurately capture the marginal effect from the macroeconomic factor. This latter finding poses a challenge in practice as we always face the problem of unobserved risk factors and limited data on LGD. Our results further indicate that model fit in terms of SSE and mean predictions may not be a good criterion to evaluate stress testing models. Evidence on the challenges of capturing the marginal effect from the macroeconomic variables suggests that we might need to rethink the design of stress testing. Instead of indirectly stressing LGD via a macroeconomic variable translation, it might be more appropriate to stress the LGDs directly.

Finally, we do not observe a clear advantage for the true model, especially if there are missing explanatory variables or if the sample size is small. Under such conditions, the less computationally challenging models, i.e., CG and BR, can perform as well as the more complicated ones, i.e., TTG and IB. As a result, in real practice, we may not need the most sophisticated statistical models for LGD.

REFERENCES:

Altman, E., and Kalotay, E.A. (2014). Ultimate recovery Mixtures. *Journal of Banking and Finance* 40,116–129. (https://doi.org/10.1016/j.jbankfin.2013.11.021)

Angrist, J., and Pischke, J. (2008). Mostly harmless econometrics. Princeton University Press.

Asarnow, E., and Edwards, D. (1995). Measuring loss on defaulted bank loans: A 24-year study. *Journal of Commercial Lending* March 1995, 11-23.

Bagnato, L., and Punzo, A. (2013). Finite mixtures of unimodal beta and gamma densities and the k-bumps algorithm. *Computational Statistics* 28, 1571-1597. (https://doi.org/10.1007/s00180-012-0367-4)

Bastos, J. (2010). Forecasting bank loans loss-given-default. *Journal of Banking and Finance* 34, 2510-2517. (https://doi.org/10.1016/j.jbankfin.2010.04.011)

Bastos, J. (2014). Ensemble predictions of recovery rates. *Journal of Financial Services Research* 46, 177-193. (https://doi.org/10.1007/s10693-013-0165-3)

Calabrese, R., and Zenga, M. (2010). Bank loan recovery rates: measuring and nonparameteric density estimation. *Journal of Banking and Finance* 34, 903-911. (https://doi.org/10.1016/j.jbankfin.2009.10.001)

Dimitrova, D., Kaishev, V., and Tan, S, (2017). "Computing the Kolmogorov–Smirnov Distribution when the Underlying cdf is Purely Discrete, Mixed or Continuous". *Journal of Statistical Software*. Forthcoming.

Duan, J.C., and Hwang, R.C. (2014). Predicting recovery rates at the time of corporate default. Working paper, National University of Singapore.

Ferrari, S., and. Cribari-Neto, F. (2004). Beta regression for modeling rates and proportions. *Journal of Applied Statistics* 31,799-815. (https://doi.org/10.1080/0266476042000214501)

Hartmann-Wendels, T., Miller, P., and Tows, E. (2014) Loss given default for leasing: Parametric and nonparametric estimations. *Journal of Banking and Finance* 40, 364–375. (https://doi.org/10.1016/j.jbankfin.2013.12.006)

Jacobs, M. (2010). An empirical study of exposure at default. *Journal of Advanced Studies in Finance* 1(1), 31-59.

Kruger, S., and Rosch, D. (2017). Downturn LGD modeling using quantile regression. *Journal of Banking and Finance* 79, 42-56. (https://doi.org/10.1016/j.jbankfin.2017.03.001)

Leymarie, J., Hurlin, C., and Patin, A. (2018). Loss functions for LGD models comparison. *European Journal of operational Research* 268(1), 348-360. (https://doi.org/10.1016/j.ejor.2018.01.020)

Li, P., Qi., M., Zhang, X., and Zhao, X. (2016). Further investigation of parametric loss given default modeling. *Journal of Credit Risk 12(4), 17-47*. (https://doi.org/10.21314/JCR.2016.215)

Li, P. (2018). Efficient MCMC estimation of inflated beta regression models. *Computational Statistics* 33(1), 127-158. (https://doi.org/10.1007/s00180-017-0747-x)

Loterman, G., Brown, I., Martens, D., Mues, C., and Baesens, B. (2012). Benchmarking regression algorithms for loss given default modeling. *International Journal of Forecasting* 28, 161–170. (https://doi.org/10.1016/j.ijforecast.2011.01.006)

Nazemi, A., Fatemi, F, Konstantin, H., and Fabozzi, F. (2017), Fuzzy decision fusion approach for loss-given default modeling, *European Journal of Operational Research* 262, 780-791. (https://doi.org/10.1016/j.ejor.2017.04.008)

Ospina, R., and Ferrari, S. L. P. (2010). Inflated beta distributions, *Statistical Papers 51,* Article 111 (https://doi.org/10.1007/s00362-008-0125-4)

Papke, L. E., and Wooldridge, J. M. (1996). Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics* 11, 619-632. (https://doi.org/10.1002/(SICI)1099-1255(199611)11:6%3C619::AID-JAE418%3E3.0.CO;2-1)

Qi, M., and Zhao, X. (2011) A comparison of modeling methods for loss given default. *Journal of Banking and Finance* 35, 2842-2855. (https://doi.org/10.1016/j.jbankfin.2011.03.011)

Serfling, R. (1980). Approximation theorems of mathematical statistics. New York: Wiley

Sigrist F., and Stahel, W.A. (2011). Using the censored Gamma distribution for modeling fractional response variables with an application to loss given default. ASTIN Bulletin 41, 673-710 (https://doi.org/10.2143/AST.41.2.2136992)

Tobback, E., Martens, D., Gestel, T.V., and Baesen, B. (2014). Forecasting loss given default models: Impact of account characteristics and macroeconomic State. *Journal of the Operational Research Society* 65, 376–392. (https://doi.org/10.1057/jors.2013.158)

Tomarchio S., and Punzo A. (2019). Modelling the loss given default distribution via a family of zero-and-one inflated mixture models. *Journal of the Royal Statistical Society: Series A* 182(4), 1247-1266. (https://doi.org/10.1111/rssa.12466)

Tong, E.N., Mues, C., Brown, I., and Thomas, L.C. (2016). Exposure at default models with and without the credit conversion factor. *European Journal of Operational Research 252*(3), 910-920. (https://doi.org/10.1016/j.ejor.2016.01.054)

Yao, X., Crook, J., and Andreeva, G. (2015) Support vector regression for loss given default modeling. *European Journal of Operational Research* 240, 52-538. (https://doi.org/10.1016/j.ejor.2014.06.043)

Yashkir, O., and Yashkir, Y. (2013). Loss given default modeling: a comparative analysis. *Journal of Risk Model Validation* 7(1), 25-59. (https://doi.org/10.21314/JRMV.2013.101

TABLE 1 PANEL A: Parameter estimates for the inflated beta model with a full set of explanatory variables

	$\boldsymbol{\alpha}$			$\boldsymbol{\beta}$			$\boldsymbol{\gamma}$			$\boldsymbol{\phi}$		
	True	Est	SE	True	Est	SE	True	Est	SE	True	Est	SE
x_1	-0.54	-0.537	0.015	-1.46	-1.457	0.017	0	0.003	0.007	5	5.010	0.414
x_2	-5.0	-5.084	0.214	6.0	5.949	0.225	0.5	0.438	0.094			
x_3	0.4	0.405	0.008	-0.1	-0.115	0.008	-0.1	-0.100	0.003			
x_4	0.4	0.404	0.008	-0.1	-0.107	0.008	-0.1	-0.102	0.003			
x_5	0.4	0.407	0.008	-0.1	-0.084	0.008	-0.1	-0.097	0.003			
x_6	0.4	0.412	0.008	-0.1	-0.098	0.008	-0.1	-0.096	0.003			
x_7	0.4	0.398	0.008	-0.1	-0.108	0.008	-0.1	-0.090	0.003			
x_8	0.4	0.399	0.008	-0.1	-0.090	0.008	-0.1	-0.103	0.003			
x_9	0.4	0.404	0.008	-0.1	-0.100	0.008	-0.1	-0.100	0.003			
x_{10}	0.4	0.404	0.008	-0.1	-0.102	0.008	-0.1	-0.097	0.003			
x_{11}	0.4	0.404	0.008	-0.1	-0.103	0.008	-0.1	-0.098	0.003			

This table has the true parameter values (True) from the data generating process, maximum likelihood estimates (Est), and standard errors (SE).

TABLE 1 PANEL B: Performance metrics across models with the full set of explanatory variables

	SSE	R^2	Pearson	Kendall	Spearman
OLS	50161.491	0.076	0.275	0.189	0.270
IG smearing	50368.541	0.072	0.274	0.189	0.270
FRR	50161.985	0.076	0.275	0.189	0.270
CG	50182.672	0.075	0.275	0.189	0.270
TTG	50162.113	0.076	0.276	0.189	0.270
BR	50161.568	0.076	0.275	0.189	0.270
IB	50130.834	0.076	0.276	0.189	0.270

This table has the results for the linear regression (OLS), Inverse Gaussian with smearing estimator (IG smearing), Fractional Response Regression (FRR), Censored Gamma regression (CG), Two-tiered Gamma regression (TTG), Beta Regression (BR), and Inflated Beta regression (IB).

TABLE 1 PANEL C: Performance metrics across models with four explanatory variables dropped

	SSE	R^2	Pearson	Kendall	Spearman
OLS	51974.985	0.042	0.205	0.141	0.202
IG Smearing	52196.308	0.038	0.205	0.141	0.201
FRR	51976.005	0.042	0.205	0.141	0.202
CG	51998.007	0.042	0.205	0.141	0.201
TTG	51993.207	0.042	0.206	0.141	0.201
BR	51986.691	0.042	0.205	0.141	0.201
IB	51965.951	0.042	0.206	0.141	0.202

This table has the results for the linear regression (OLS), Inverse Gaussian with smearing estimator (IG smearing), Fractional Response Regression (FRR), Censored Gamma regression (CG), Two-tiered Gamma regression (TTG), Beta Regression (BR), and Inflated Beta regression (IB).

TABLE 2: KS Statistics comparing the predicted and true unconditional distributions across models

	Full set of explanatory variables	Omitted explanatory variables	
		4 Omitted	8 Omitted
OLS	0.167	0.164	0.164
IG Smearing	0.259	0.259	0.259
CG	0.061	0.063	0.063
TTG	0.032	0.032	0.032
BR	0.056	0.055	0.057
IB	0.013	0.012	0.01

This table has the results for the linear regression (OLS), Inverse Gaussian with smearing estimator (IG smearing), Fractional Response Regression (FRR), Censored Gamma regression (CG), Two-tiered Gamma regression (TTG), Beta Regression (BR), and Inflated Beta regression (IB).

FIGURE 1: Histogram of true LGD data from the inflated beta model

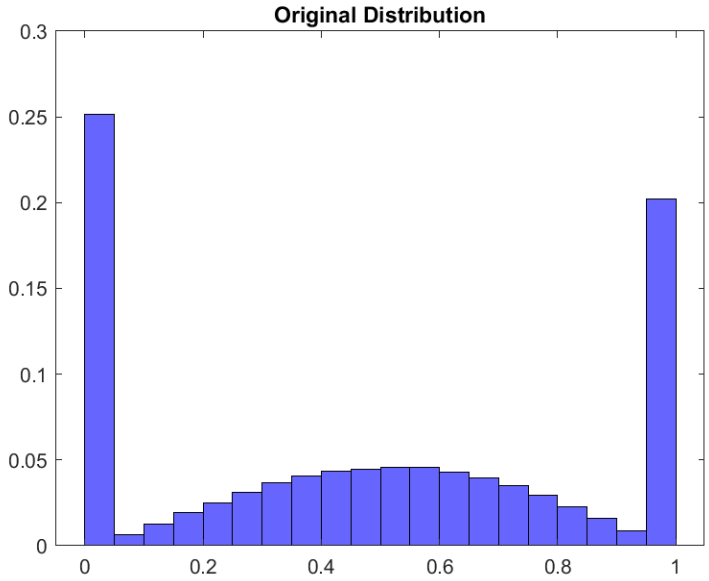
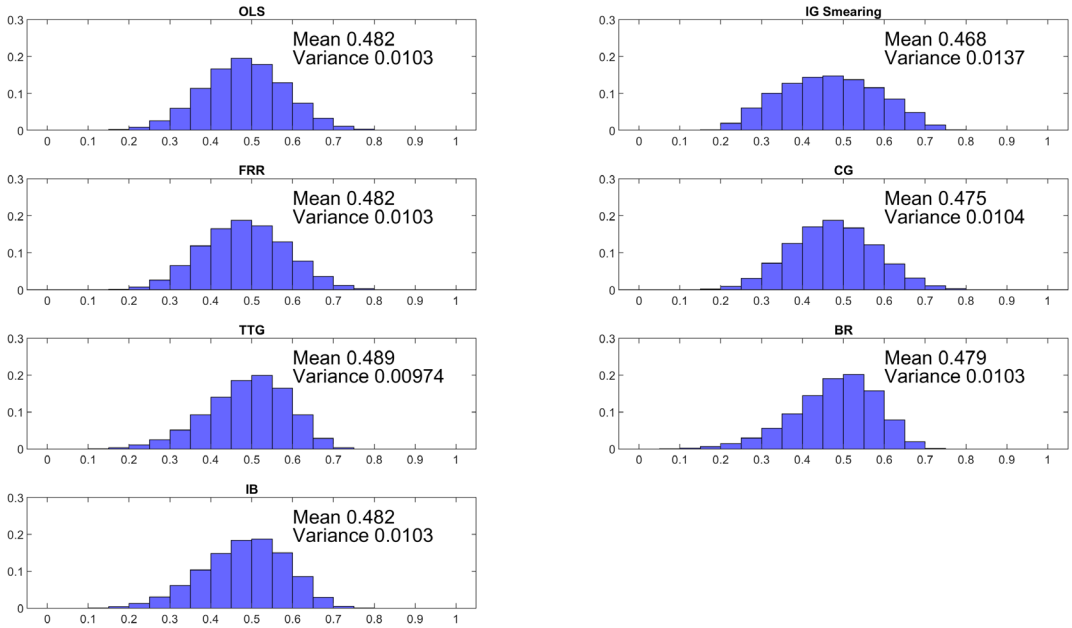
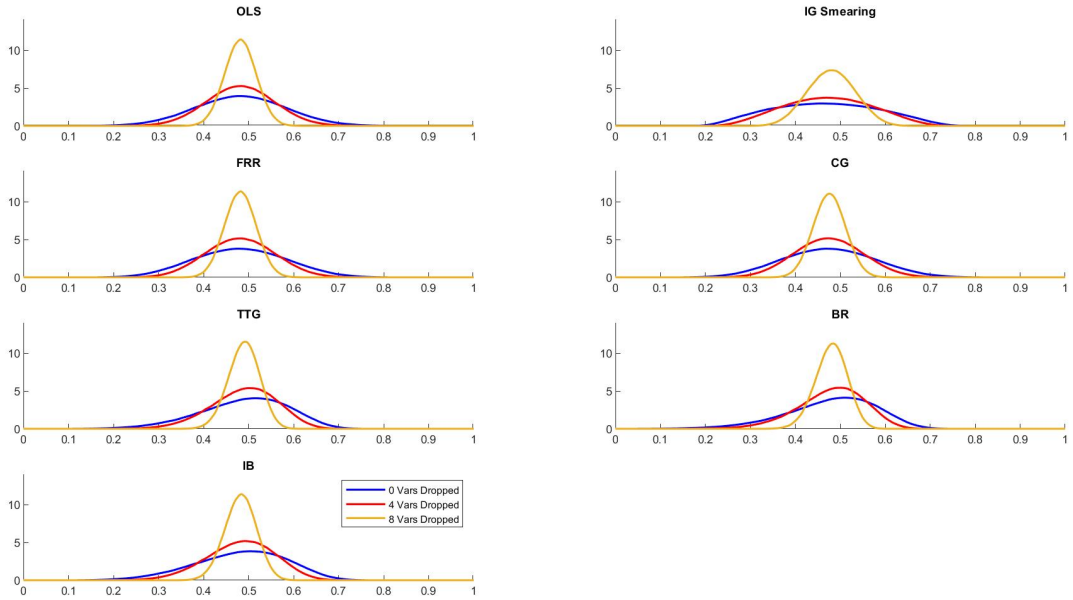


FIGURE 2: Histograms of predicted conditional means using a full set of explanatory variables



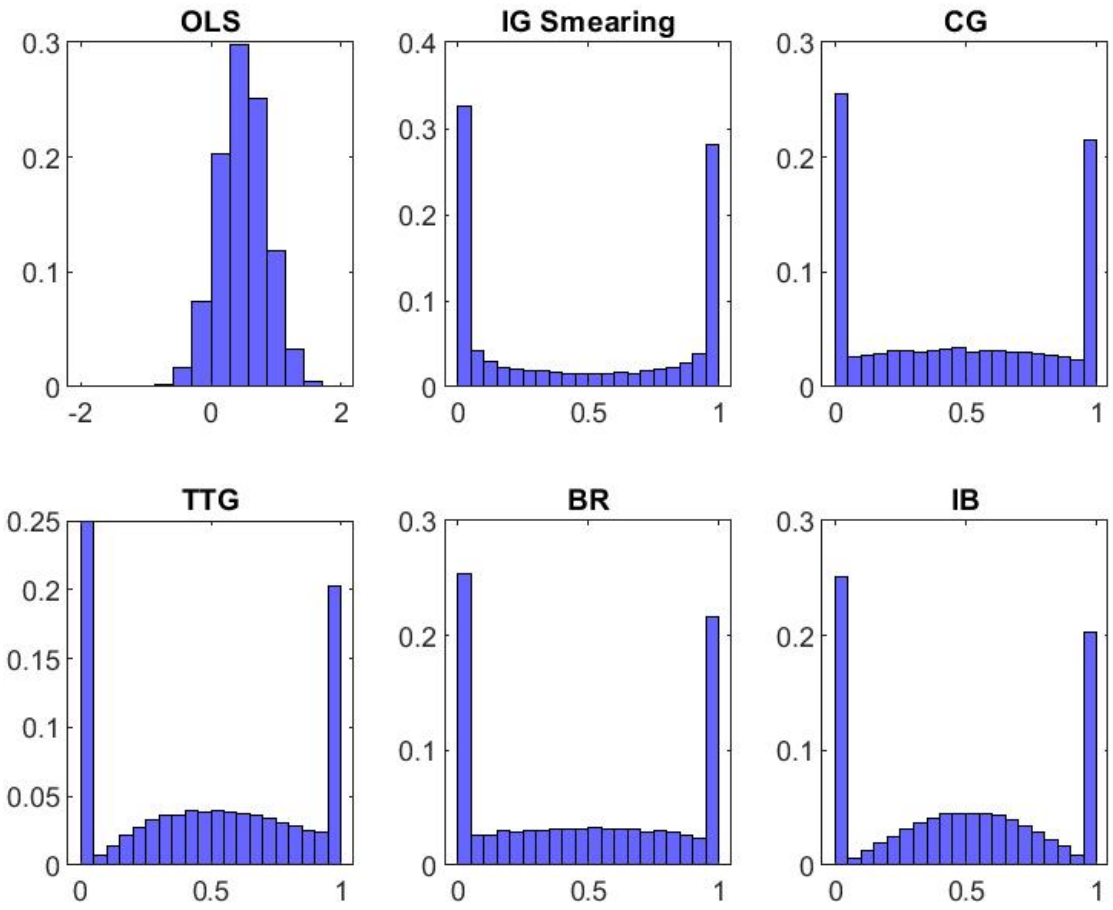
The models are the linear regression (OLS), Inverse Gaussian with smearing estimator (IG smearing), Fractional Response Regression (FRR), Censored Gamma regression (CG), Two-tiered Gamma regression (TTG), Beta Regression (BR), and Inflated Beta regression (IB).

FIGURE 3: Kernel densities of predicted conditional means across models and noise



The models are the linear regression (OLS), Inverse Gaussian with smearing estimator (IG smearing), Fractional Response Regression (FRR), Censored Gamma regression (CG), Two-tiered Gamma regression (TTG), Beta Regression (BR), and Inflated Beta regression (IB).

FIGURE 4 PANEL A: Predicted unconditional distributions



The models are the linear regression (OLS), Inverse Gaussian with smearing estimator (IG smearing), Censored Gamma regression (CG), Two-tiered Gamma regression (TTG), Beta Regression (BR), and Inflated Beta regression (IB).

FIGURE 4 PANEL B: Predicted unconditional distributions for FRR with different α values

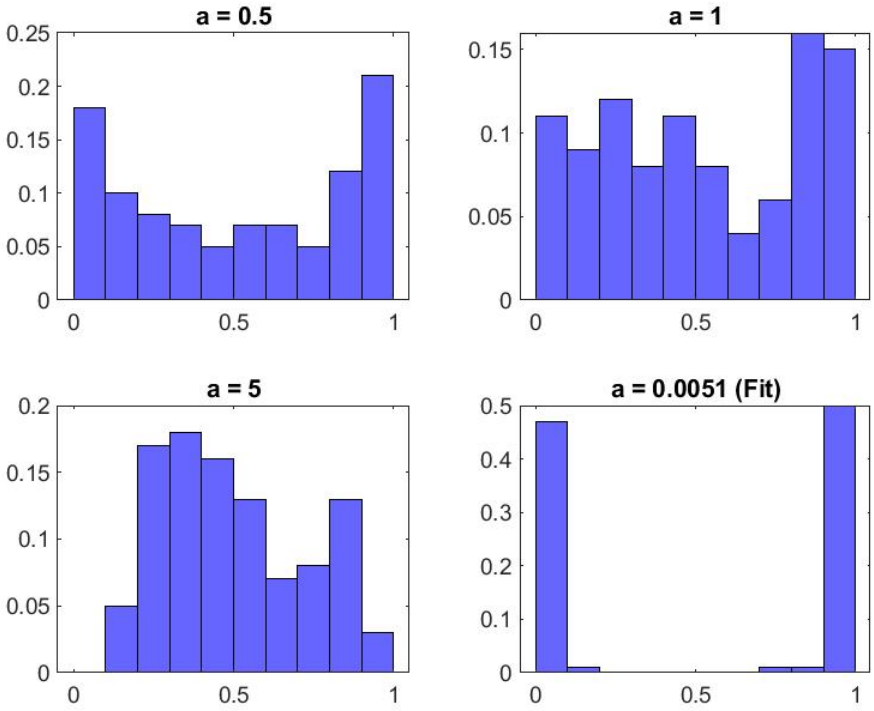
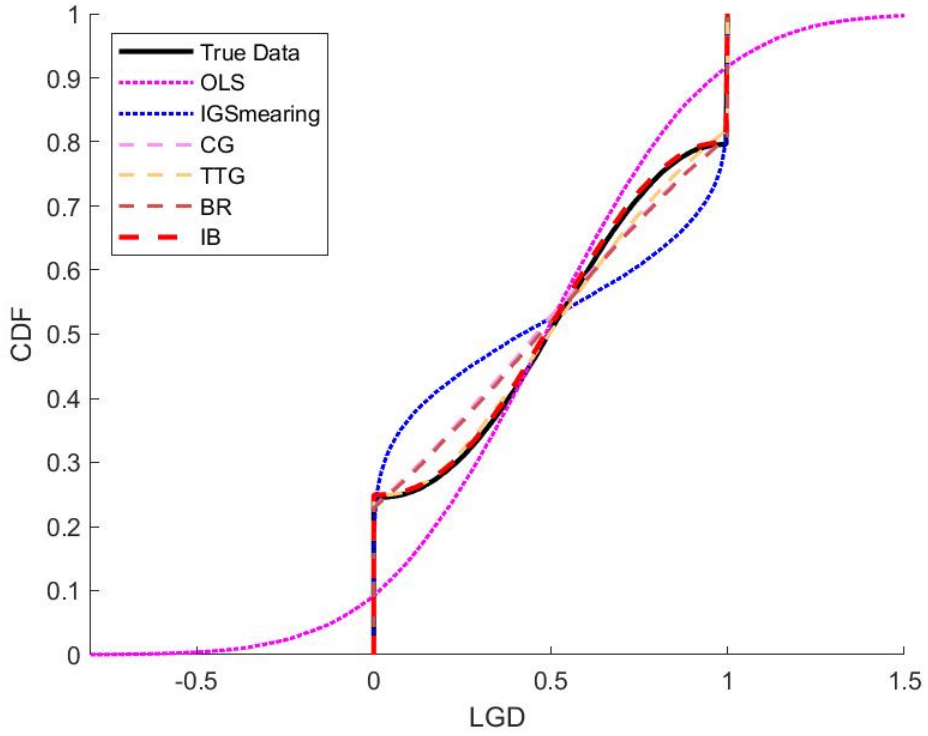
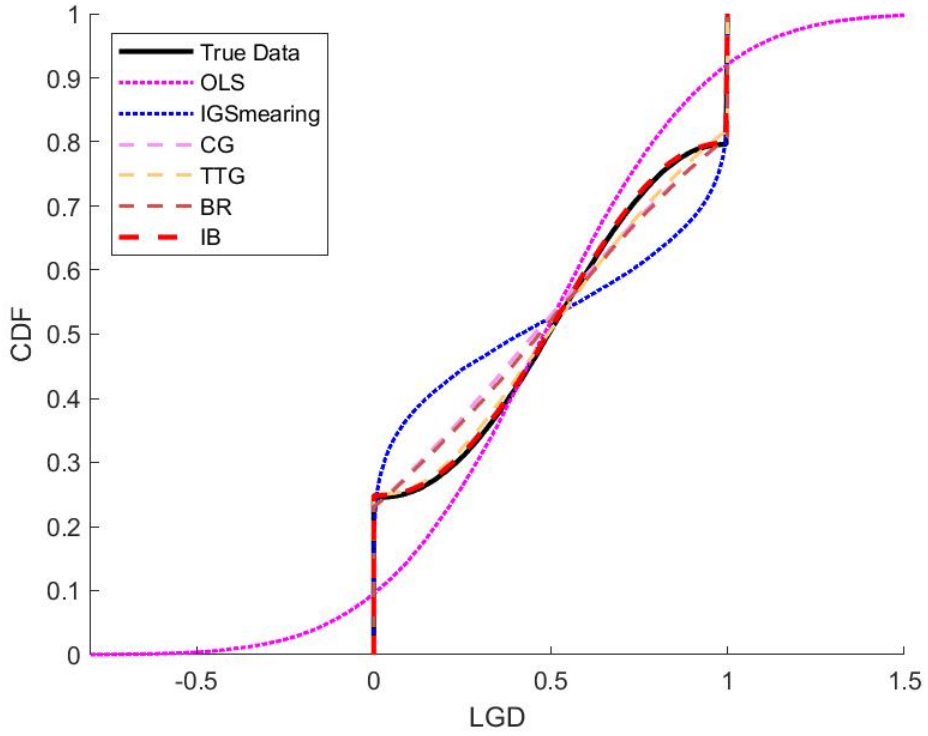


FIGURE 5 PANEL A: CDFs based on predicted unconditional distributions using full explanatory variables



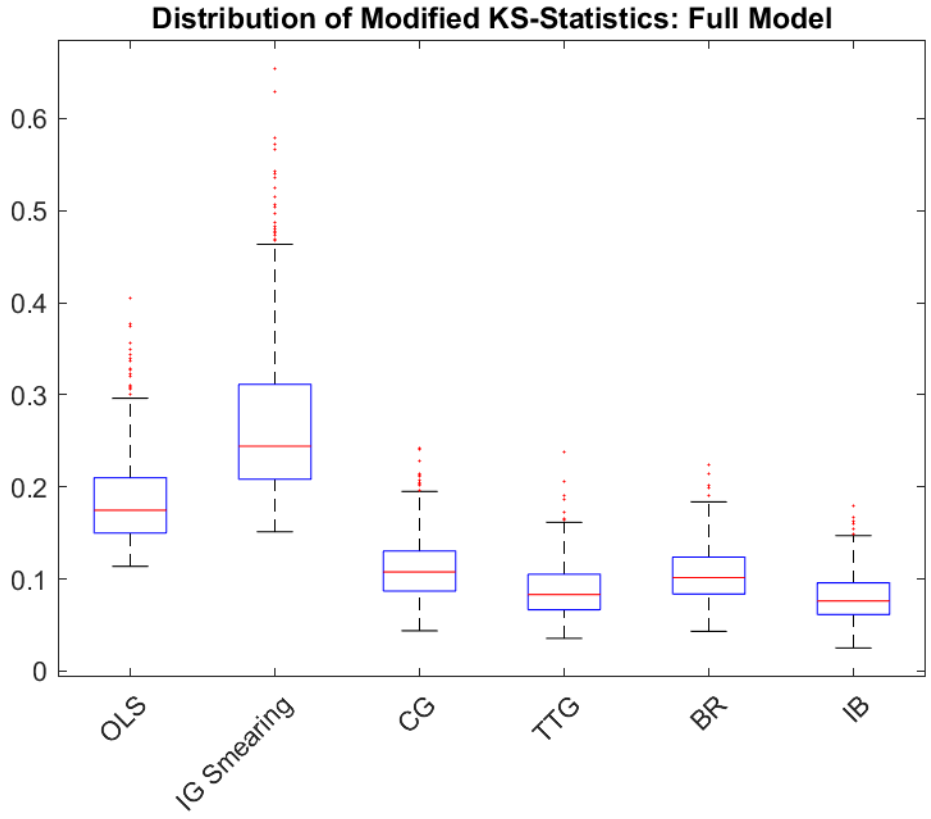
The models are the linear regression (OLS), Inverse Gaussian with smearing estimator (IG smearing), Censored Gamma regression (CG), Two-tiered Gamma regression (TTG), Beta Regression (BR), and Inflated Beta regression (IB).

FIGURE 5 PANEL B: CDFs based on predicted unconditional distributions with four explanatory variables omitted



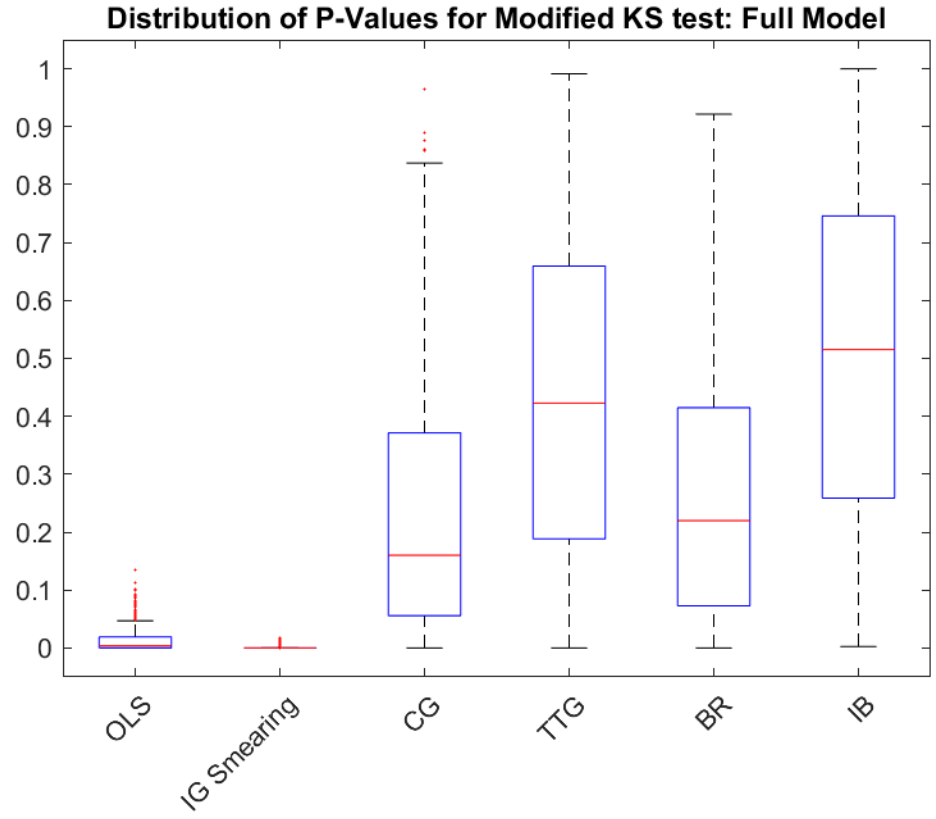
The models are the linear regression (OLS), Inverse Gaussian with smearing estimator (IG smearing), Censored Gamma regression (CG), Two-tiered Gamma regression (TTG), Beta Regression (BR), and Inflated Beta regression (IB).

FIGURE 6 PANEL A: Distribution of KS statistics comparing predicted conditional distributions



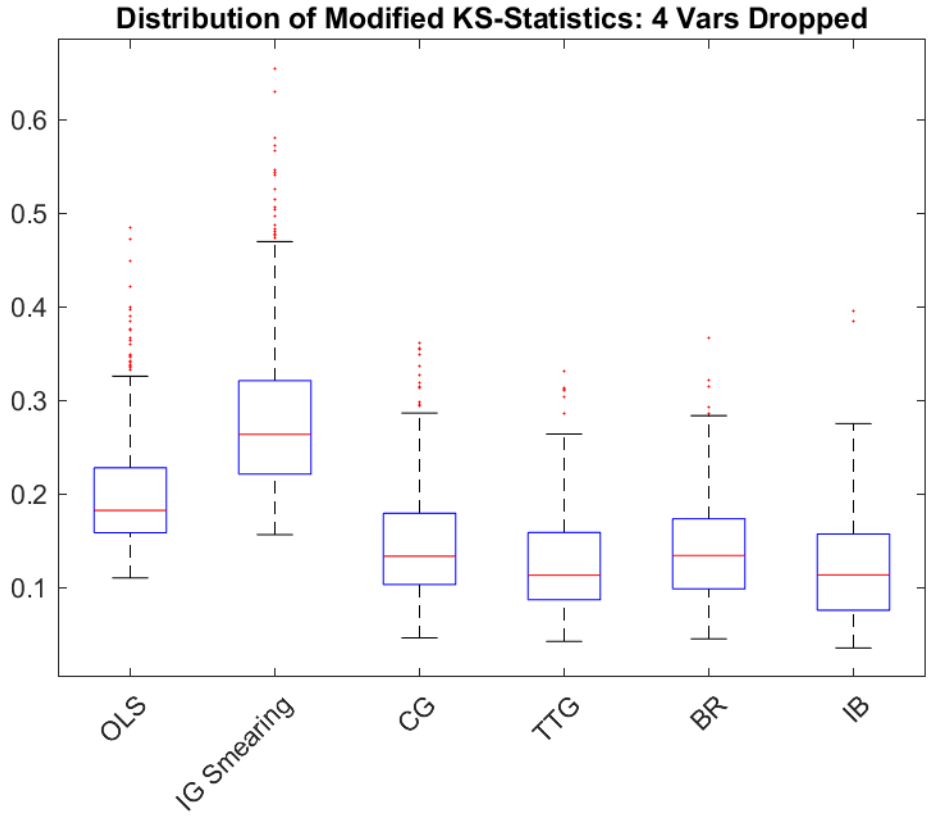
The models are the linear regression (OLS), Inverse Gaussian with smearing estimator (IG smearing), Censored Gamma regression (CG), Two-tiered Gamma regression (TTG), Beta Regression (BR), and Inflated Beta regression (IB).

FIGURE 6 PANEL B: Distribution of p-values comparing predicted conditional distributions



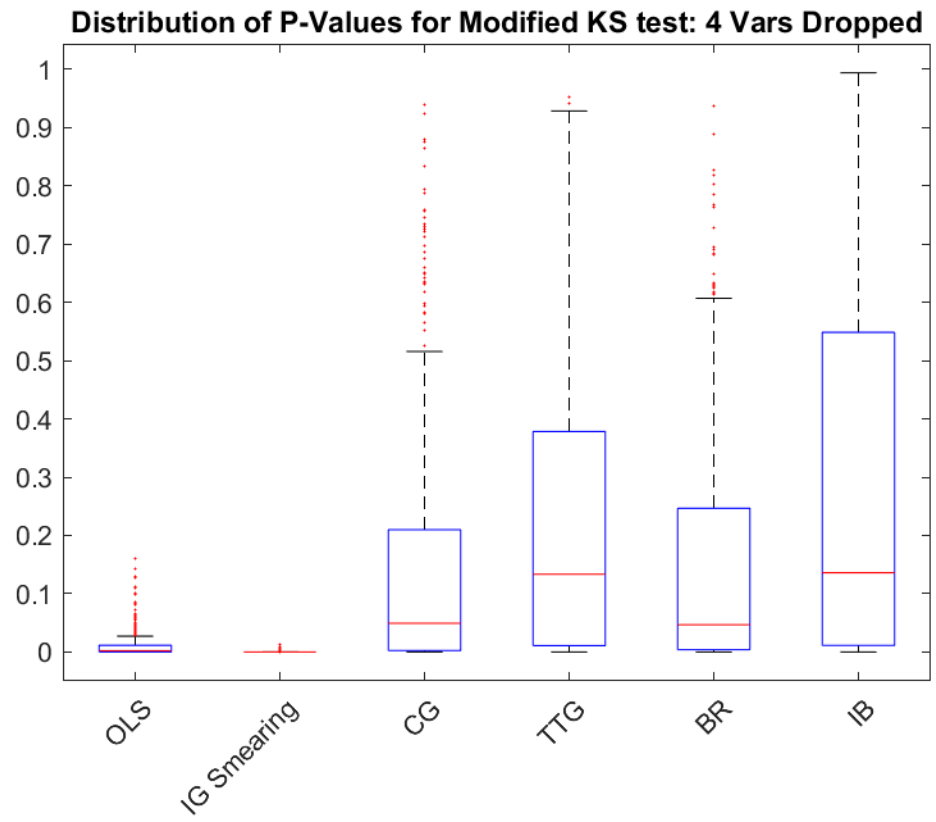
The models are the linear regression (OLS), Inverse Gaussian with smearing estimator (IG smearing), Censored Gamma regression (CG), Two-tiered Gamma regression (TTG), Beta Regression (BR), and Inflated Beta regression (IB).

FIGURE 6 PANEL C: Distribution of KS statistics comparing predicted conditional distributions



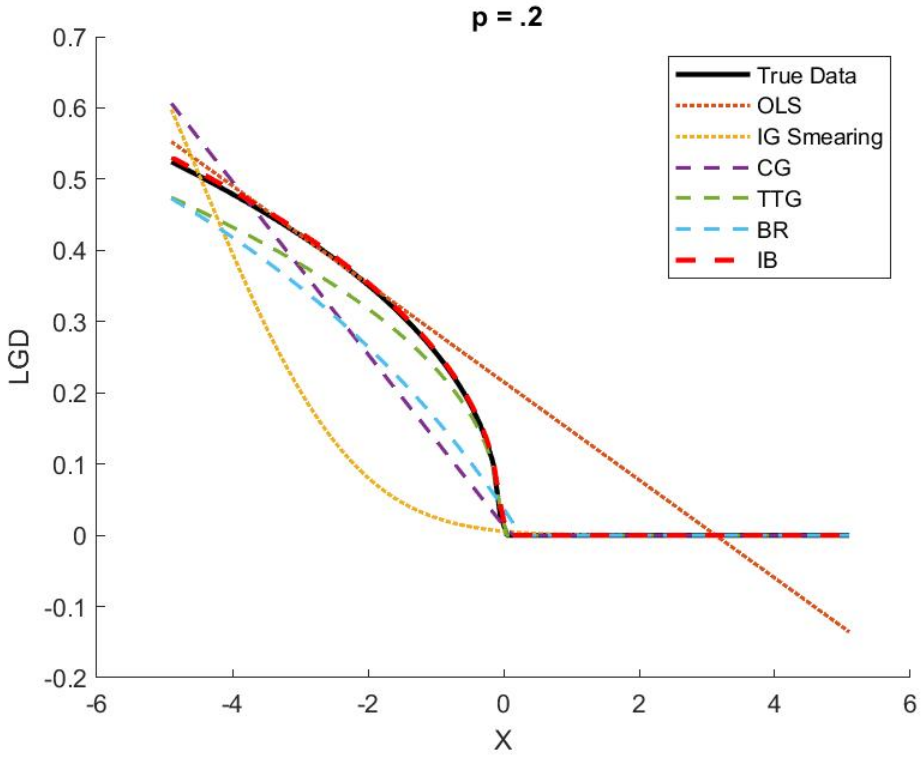
The models are the linear regression (OLS), Inverse Gaussian with smearing estimator (IG smearing), Censored Gamma regression (CG), Two-tiered Gamma regression (TTG), Beta Regression (BR), and Inflated Beta regression (IB).

FIGURE 6 PANEL D: Distribution of p-values comparing predicted conditional distributions



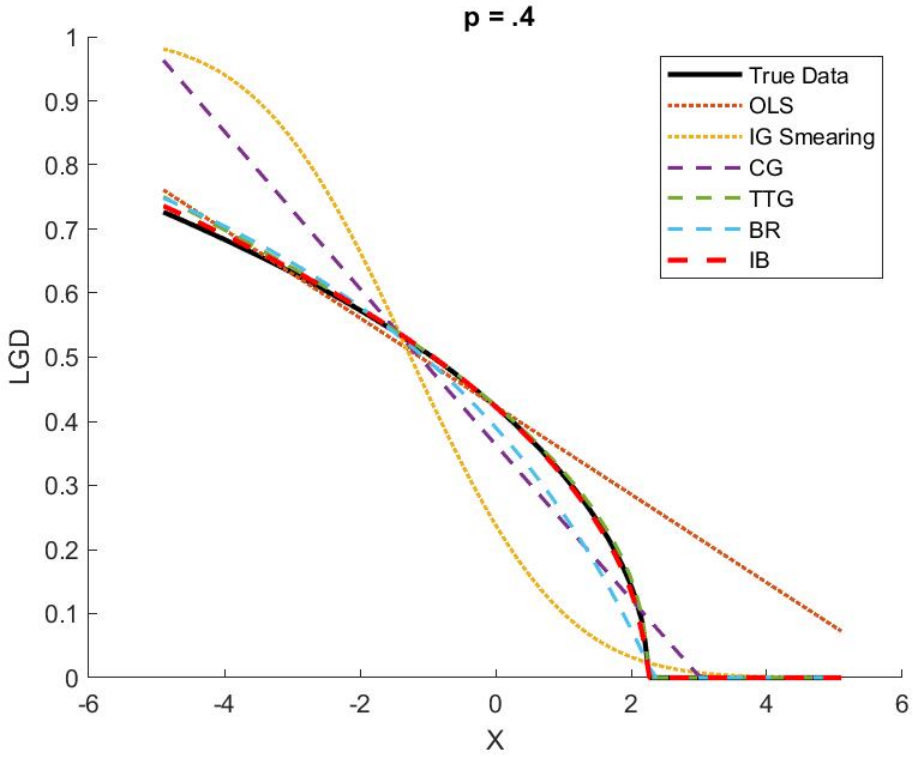
The models are the linear regression (OLS), Inverse Gaussian with smearing estimator (IG smearing), Censored Gamma regression (CG), Two-tiered Gamma regression (TTG), Beta Regression (BR), and Inflated Beta regression (IB).

FIGURE 7 PANEL A: 20th quantile based on predicted conditional distributions with a full set of explanatory variables



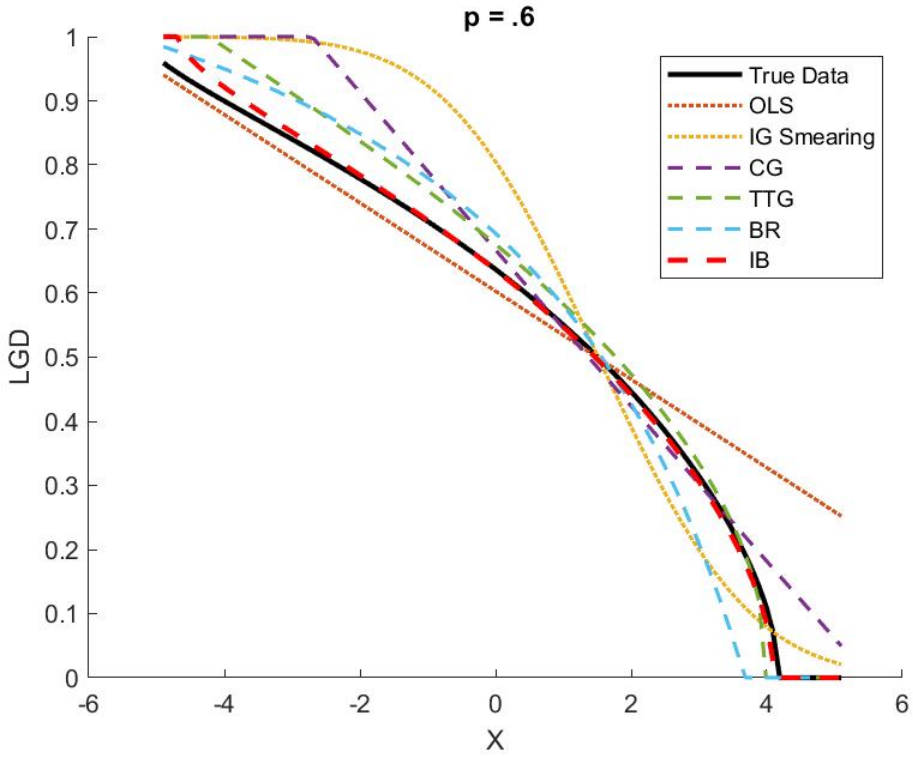
The models are the linear regression (OLS), Inverse Gaussian with smearing estimator (IG smearing), Censored Gamma regression (CG), Two-tiered Gamma regression (TTG), Beta Regression (BR), and Inflated Beta regression (IB).

FIGURE 7 PANEL B: 40th quantile based on predicted conditional distributions with a full set of explanatory variables



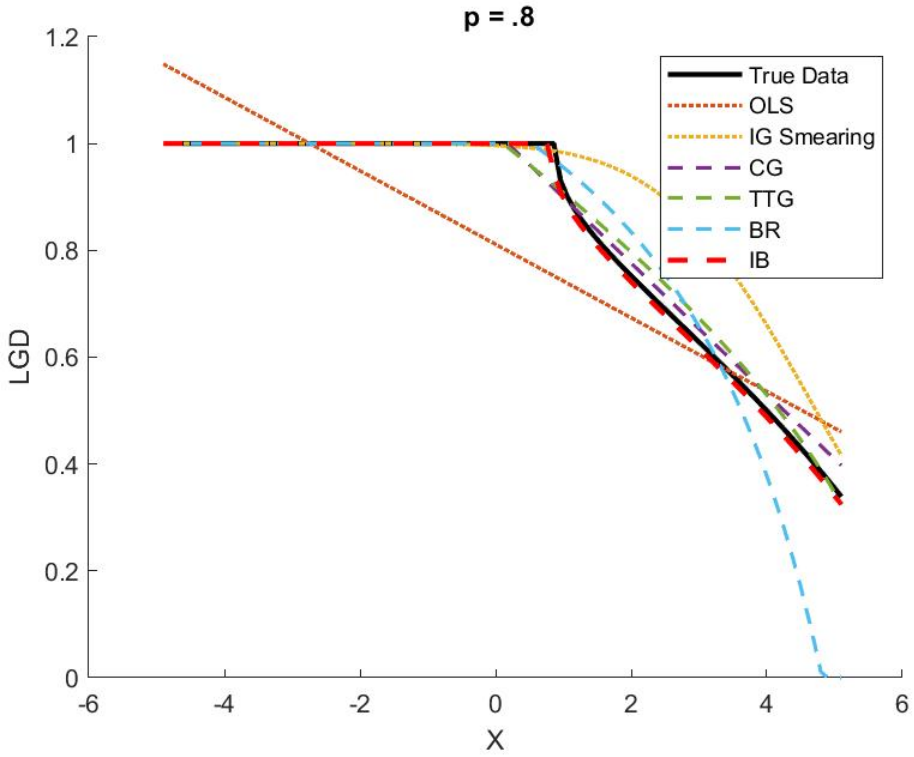
The models are the linear regression (OLS), Inverse Gaussian with smearing estimator (IG smearing), Censored Gamma regression (CG), Two-tiered Gamma regression (TTG), Beta Regression (BR), and Inflated Beta regression (IB).

FIGURE 7 PANEL C: 60th quantile based on predicted conditional distributions with a full set of explanatory variables



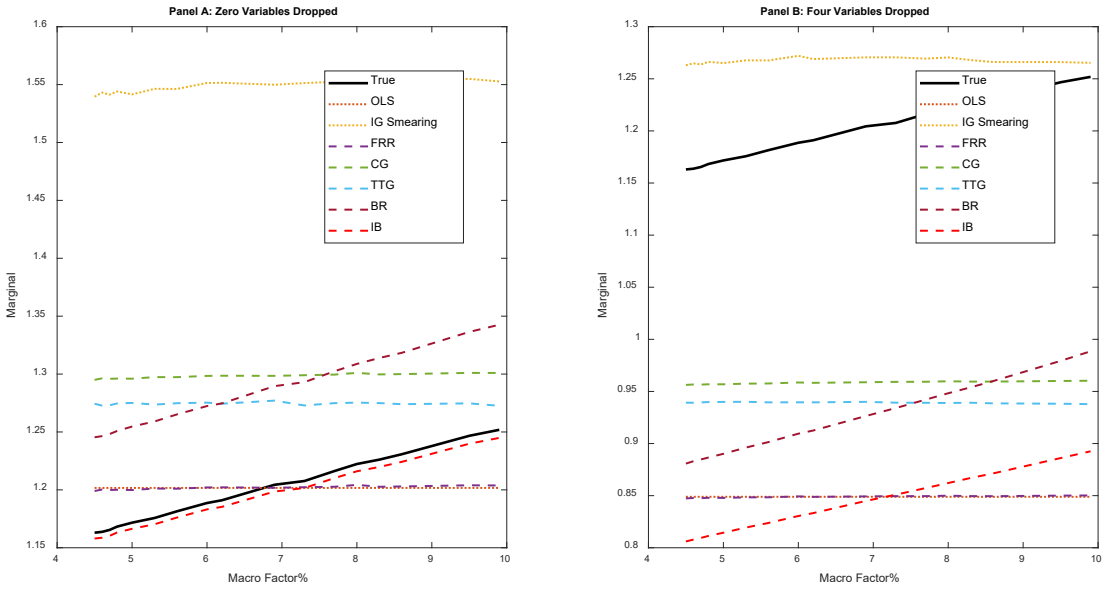
The models are the linear regression (OLS), Inverse Gaussian with smearing estimator (IG smearing), Censored Gamma regression (CG), Two-tiered Gamma regression (TTG), Beta Regression (BR), and Inflated Beta regression (IB).

FIGURE 7 PANEL D: 80th quantile based on predicted conditional distributions with a full set of explanatory variables



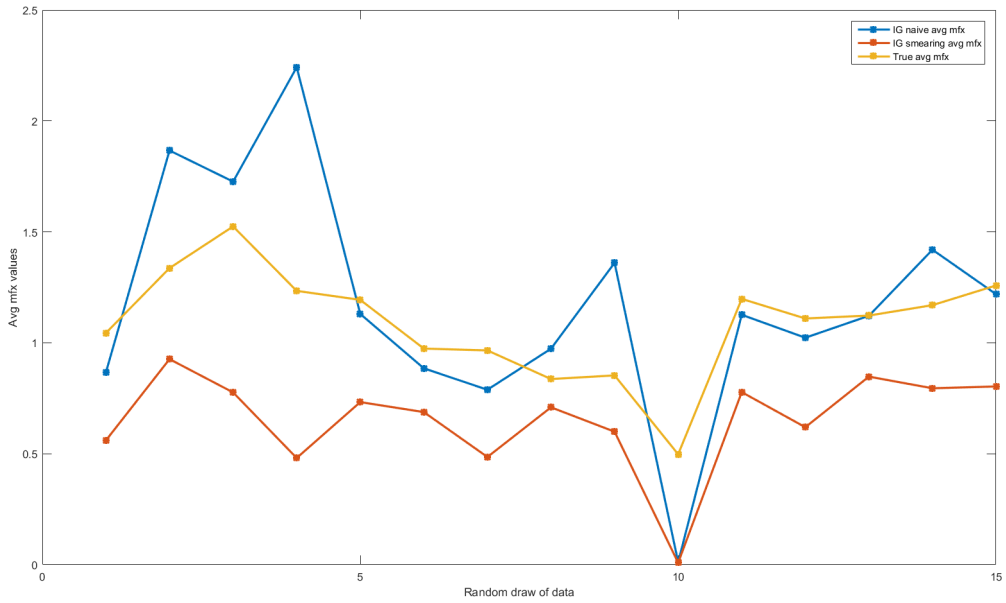
The models are the linear regression (OLS), Inverse Gaussian with smearing estimator (IG smearing), Censored Gamma regression (CG), Two-tiered Gamma regression (TTG), Beta Regression (BR), and Inflated Beta regression (IB).

FIGURE 8: Average marginal effects across models and noise levels



The models are the linear regression (OLS), Inverse Gaussian with smearing estimator (IG smearing), Fractional Response Regression (FRR), Censored Gamma regression (CG), Two-tiered Gamma regression (TTG), Beta Regression (BR), and Inflated Beta regression (IB).

FIGURE 9: Average marginal effects of IG naïve and the IG smearing models



The models are Inverse Gaussian with naïve estimator (IG naïve) and Inverse Gaussian with smearing estimator (IG smearing) from Li et al. (2016).

Appendix A: Models investigated in this paper

Methods	Full Name of the method	Reference(s)
OLS	Linear regression	
IG smearing	Inverse Gaussian regression with the smearing and naïve estimators	Li et al. (2016)
FRR	Fractional response regression	Papke and Wooldrige (1996)
CG	Censored gamma regression	Sigrist and Stahel (2011)
TTG	Two-tiered gamma regression	Sigrist and Stahel (2011)
BR	Beta regression	Duan and Hwang (2014)
IB	Inflated beta regression	Li et al. (2016), Ospina and Ferrari (2010), Yashkir and Yashkir (2013)