Federal Deposit Insurance Corporation • Center for Financial Research

Working Paper Series

Modeling Loss Given Default

Phillip Li
Federal Deposit Insurance Corporation

Xiaofei Zhang
Office of the Comptroller of the Currency

Xinlei Zhao
Office of the Comptroller of the Currency

July 2018

FDIC CFR WP 2018-03

fdic.gov/cfr

NOTE: Staff working papers are preliminary materials circulated to stimulate discussion and critical comment. The analysis, conclusions, and opinions set forth here are those of the author(s) alone and do not necessarily reflect the views of the Federal Deposit Insurance Corporation. References in publications to this paper (other than acknowledgement) should be cleared with the author(s) to protect the tentative character of these papers.

Modeling Loss Given Default[1]

Phillip Li[2]
Federal Deposit Insurance Corporation

Xiaofei Zhang
Xinlei Zhao

Office of the Comptroller of the Currency

First version: 5/31/2017
This version: 7/13/2018

Opinions expressed in this paper are those of the authors and not necessarily those of the Office of the Comptroller of the Currency, the U.S. Department of the Treasury, or the FDIC.

Keywords: loss given default, bi-modal distribution, simulation, predicted distribution, stress testing.
JEL classifications: G21, G28.

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

[2] Corresponding author: Phillip Li, Federal Deposit Insurance Corporation, 550 17th Street, NW, Washington, DC 20429-0002, Tel: 202-898-3501, e-mail: pli@fdic.gov.

Modeling Loss Given Default

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, 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. Therefore, the findings in the literature that predictive accuracy and rank ordering cluster in a very narrow range across different parametric models are robust. We argue, however, that predicted distributions as well as the models' ability to accurately capture marginal effects are also important performance metrics for capital models and stress testing. We find that 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. Also, we find that stress testing poses a challenge to all LGD models because of limited data and relevant explanatory variable availability, and that model selection criteria based on goodness of fit may not serve the stress testing purpose well. Finally, the evidence here suggests that we do not need to use the most sophisticated parametric methods to model LGD.

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 distribution with modes close to the boundary values, as shown in Asarnow and Edwards (1995) and Qi and Zhao (2011).[3] These characteristics make standard statistical models, such as the linear regression model, theoretically inappropriate for LGD modeling. As a result, many

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

statistical models have been proposed in the literature to accommodate the unusual distribution of LGD (see for example, Qi and Zhao (2011), Li, Qi, Zhang, and Zhao (2016), and the references therein for surveys). Even with the sophistication of these models, papers such as Qi and Zhao (2011) and Li, Qi, Zhang, and Zhao (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. This noise can result from various reasons, 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 the same predictions and 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 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), Leymarie, Hurlin, and Patin (2018), and Kruger and Rosch (2017).

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) and then fit a variety of statistical models to this dataset.[4] 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 controlled simulation framework 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

[4] 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. Note that nobody knows the true LGD DGP, and there is no way for anybody to prove that any simulated data can replicate the true LGD DGP. We address this concern indirectly by trying many different DGPs to generate bi-modal data with substantial mass at the 0 and 1 values. The results from these alternative simulations are discussed in Section 3.5.

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 capital formula only requires 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.[5] 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 and Dodd-Frank Act Stress Testing (DFAST). 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

[5] 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.

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, Qi, Zhang, and Zhao (2016)),[6] fractional response regression (Papke and Wooldrige (1996))[7], 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)).[8] All models other than the standard linear regression are briefly outlined in Appendix A. The standard linear regression 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 designed to handle bi-modal distributions.[9] 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. We do not include in this study non-parametric methods (such as regression trees (Qi and Zhao (2011)) and support vector regression (Yao, Crook, and Andreeva

[6] We have investigated the inverse Gaussian regression with beta transformation and smearing estimator as well, and those results are quite similar to those from the inverse Gaussian regression with smearing estimator in this paper.

[7] The FRR is technically a semi-parametric method while the other models are parametric.

[8] 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.

[9] We mainly investigate IG smearing in this paper, because Li, Qi, Zhang, and Zhao (2016) show that the smearing estimator makes the inverse regression method more stable.

(2015)) because generating predicted distributions and marginal effects for these methods is not straightforward.

We find that, even using the full set of explanatory variables from the DGP, 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 (measured by squared error loss) and rank ordering (measured by Pearson's correlation, Spearman's Rho, and Kendall's Tau). Moreover, when we introduce noise, both predictive accuracy and rank ordering ability unsurprisingly decline across all models, and the models still perform similarly to each other. Therefore, the findings in the literature that the predictive accuracy and rank ordering ability are poor and cluster in a very narrow range are robust, and such findings are not driven by statistical model specification or too much noise in the data to be modeled. These results suggest that, in conditions where only LGD conditional means are required and performance is measured by these performance metrics, all the commonly-used models investigated in this paper perform equally well.

Furthermore, we find that the predicted conditional distributions from the sophisticated models all perform reasonably well, while the ones from the linear regression and IG smearing significantly underperform. In addition, 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.

We also find that the true model using the full set of explanatory variables is able to capture the marginal effects from a macroeconomic shock quite well. However, we find that the linear regression, IG smearing, and FRR models using the full set of explanatory variables have little macroeconomic sensitivity. Furthermore, even though the smearing estimator helps to ensure model stability when model fit is concerned, as documented in Li, Qi, Zhang, and Zhao (2016), it does not add value in terms of the marginal effect. Therefore, a method that predicts the mean better does not necessarily capture the marginal effect better. The primary challenge facing stress testing is that, when some relevant variables are unobserved or when the sample size is small, none of the models, including the true model, are able to estimate precisely the marginal effect of the macroeconomic factor.[10] This challenge suggests that we may need to rethink the design of stress testing and, in this context, stressing LGD directly instead of indirectly through macroeconomic variables might be more appropriate.

Overall, we find little difference in all performance metrics investigated in this paper among the four sophisticated parametric methods. Even though the true model has a slight advantage over the other sophisticated models when using the full set of explanatory variables and when using a large sample, the advantage disappears when some relevant explanatory variables are not included in the model or when the sample size decreases. Since unobserved explanatory variables and small sample sizes are major challenges in empirical studies, we conclude that it is not critical to use the most sophisticated parametric methods to model LGD.

Finally, 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.

[10] Our finding that the marginal effects are sensitive to omitted variables is consistent with Ramalho and Ramalho (2010). They find that omitting relevant variables biases the marginal effects for probit and binary models with loglog link functions.

For example, these results can be applied to exposure at default, which is another important risk parameter in banking practice that has a bi-modal distribution (see Jacobs (2010) and Tong et. al (2016)).

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 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 (2010a) and Li, Qi, Zhang, and Zhao (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.[11] The 400,000 observations are independent conditional on a common macroeconomic factor.

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,[12] and 9 normally distributed explanatory variables ($x_{i3}, \dots, x_{i11}$). Each explanatory

[11] 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. If most observations are from economic downturns and few observations are from normal business conditions, this data imbalance will make it more difficult to capture the macroeconomic impact. Further, the number of corporate defaults is much lower than 10,000 in any year. We use 10,000 defaults a year to enhance model fit, and we will discuss smaller samples in section 3.5.2. Our simulation design aims to investigate the ideal case to understand the challenges we would face, with the understanding that we will face more challenges in the real data.

[12] 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.

variable in $(x_{i3}, \dots, x_{i11})$ has a marginal distribution of $N(0, 0.5^2)$ 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:

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

The vector of explanatory variables, $\overrightarrow{x_\iota} = (x_{i1}, x_{i2}, \dots, x_{i11})$, is linked to the model through the following equations:

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

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

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

We set the true parameter values as $\alpha = (0.1, -5, 0.4, \dots, 0.4)$, $\beta = (-1, 6, -0.1, \dots, -0.1)$, $\gamma = (0, 0.5, -0.1, \dots, -0.1)$, and $\phi = 1.6$, and we generate 400,000 LGD_i observations according to (1)-(6).

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).

Figure 1 shows the histogram of the true LGD data. It has a strong bi-modal pattern, mimicking the distribution of real LGD data from Fig 1 of Qi and Zhao (2011). We use maximum likelihood estimation on the log likelihood implied by (1)-(6). See Li, Qi, Zhang, and Zhao (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 use two approaches to introduce additional noise in the simulation exercise. One way is to only use 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.

The second way is to 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. Due to space constraints and the fact that our main qualitative results do not change across the two definitions of noise, we only report results for the first approach.

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.

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

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 $E(LGD_i|\overrightarrow{x_\iota})$ and not the whole distribution (see Appendix A.2). 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 $\hat{E}(LGD_i|\overrightarrow{x_\iota}) = \frac{\alpha}{\alpha+\beta_i}$.[14] We then plot the 1,000 draws from this derived $beta(\alpha, \beta_i)$ distribution.

2.2.2 Predicted conditional distributions

A. 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 each of these draws, we generate 1,000 draws of LGD based on the model of interest, and calculate their Kolmogorov-Smirnov (KS) statistic and p-value relative to the

[14] 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.

DGP.[15] 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.

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.

B. 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

[15] These KS statistics and p-values correspond to a two-sample Kolmogorov-Smirnov test comparing the conditional distributions from the DGP and from the assumed model. The KS statistic measures the largest deviation between the two sets of estimated CDFs. The p-value corresponds to the null hypothesis that the two distributions are equal and the alternative hypothesis that the two distributions are not equal.

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

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 68123 among all the models. Second, although the literature has argued that the linear regression model may not be the most appropriate model to use, its SSE value of 68149 does not fall far behind that of IB and is even slightly lower than those from FRR, CG, and IG smearing. This suggests that 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, Qi, Zhang, and Zhao (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 problem 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: linear regression, IG smearing, CG, TTG, IB, and BR. Unsurprisingly, this panel shows that the predicted unconditional distribution from the linear regression has a bell shape, while all the other models show similar bi-modal patterns. 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 $\alpha = 0.0029$ leads to the most extreme bi-modal distribution with high peaks at both ends but little mass in between. 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 pictures in Panel A of Figure 4. So, we report the 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 bi-modal distributions and 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 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 quite 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 KS statistics across various models, which is to be expected for LGD distributions unconditional on the explanatory variables.

Since the 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 KS statistics in the first column of Table 2. It is clear from this panel that the predicted unconditional distribution from the linear regression is quite different from the true distribution which explains the large KS statistic. The difference between the IG smearing and true CDFs is also noticeable, particularly at the tails. These deviations at the tails are responsible for the large KS statistic for IG smearing in Table 2. Furthermore, the CDFs from the true and predicted distributions from CG, TTG, IB, and BR are quite similar, which is consistent with the small KS statistics from Table 2. Although IB is the correct distributional assumption, its advantage over CG, BR and TTG is almost undiscernible.

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, and such results indicate 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 KS tests. 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 two-sample 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 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. The p-values from Panel B support these findings as the distributions of p-values for the linear regression and IG smearing are essentially degenerate at 0 while the distributions corresponding to the sophisticated models are centered away from 0. Being centered away from 0 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. There is less dispersion for the four sophisticated models in Panel C than the linear regression and IG smearing in Panel A. Also, the 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.

3.3.2 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 quantiles stand out as they are straight lines, and they are not similar to the true quantiles. Surprisingly, the IG smearing quantile resembles the true quantile better than the linear regression quantile, despite our previous results from the conditional distribution section. However, it is obvious that the IG smearing quantiles deviate more from the true quantile than the ones for CG, TTG, BR, and IB, especially for the lower 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.

3.4 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 linear regression marginal effect curve is flat, which is not surprising because of the linearity assumption in this model. Further, both IG smearing and FRR show low macroeconomic sensitivity in the stressed scenarios. With the macroeconomic factor at 10% (corresponding to a higher level of LGD because of the positive true marginal effect in Panel A), the linear regression, FRR, and IG Smearing all under-estimate the true marginal effect, with the IG smearing marginal effect showing the largest under-estimation. The IB marginal effect curve almost entirely overlaps with 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). CG, TTG, and BR over-estimate the marginal effect, with BR showing the most over-estimation. The degree of over-estimation from the sophisticated models is not large in Panel A of Figure 8: when the macroeconomic factor is at 10%, the true marginal effect is 1.53 versus 1.58 from BR.

Panel B of Figure 8 depicts the marginal effects when four variables are dropped from the models. This panel shows that all models under-estimate the true marginal effect, with IG smearing under-estimating the most. The IB model shows more under-estimation than the other three sophisticated models. Note that given the particular set of parameters we use in the DGP, the under-estimation of all models other than IG smearing is due to our assumption of positive correlation between the macroeconomic variable and the other explanatory variables. If we assume a negative correlation while keeping the other parameters the same, we would observe drastic over-estimation in Panel B of Figure 8 for all models except IG smearing.

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.

Figure 9 contains more analyses for the IG models. 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.[16]

We can draw several conclusions from Figure 9. First, regardless of which IG method is used, there is generally a large gap between the true marginal effect and the IG marginal effects. Therefore, 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 naïve IG in terms of predicting the marginal effect. Interestingly, Li, Qi, Zhang, and Zhao (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.

3.5 Further investigations

3.5.1 Alternative DGPs

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

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 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, Qi, Zhang, and Zhao (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.

3.5.2 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.[17]

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.

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

We find that, even using the full set of explanatory variables from the data generating process and without noise, all models still provide poor model fit, and perform similarly in terms of both predicative accuracy and rank ordering. When we omit some explanatory variables from the model or add extra noise to the data generating process, both predicative accuracy and rank ordering ability decline, 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 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. FRR, on the other hand, is not a proper method when the LGD distribution is the performance metric of interest because of uncertainty in generating the predicted distributions.

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.

Appendix A: Econometric models investigated in this paper

A.1 Transformation regressions - inverse Gaussian regression with a smearing estimator (IG smearing)

Because LGDs are bounded in the unit interval [0, 1], whereas the predicted LGDs from a linear regression are not bounded, certain transformations can be applied to LGDs before running the regression and the fitted LGDs from the regression are then transformed back to (0,1). The naïve inverse Gaussian regression (Hu and Perraudin (2002) and Qi and Zhao (2011) first transforms LGDs from the unit interval (0,1) to $(-\infty, \infty)$ using an inverse Gaussian distribution function and then runs an OLS regression using the transformed LGDs. Finally, the fitted values are transformed back from $(-\infty, \infty)$ to (0,1) using the Gaussian distribution function. This method is termed naïve, because it does not account for the fact that the optimal predictions on the untransformed scale are generally not equal to the inversions of the optimal predictions on the transformed scale. This non-linear transformation is taken care of by the smearing estimator proposed by Duan (1983) and applied in Li, Qi, Zhang, and Zhao (2016).

The inverse Gaussian transformations result in infinite values when they are applied to LGD observations that equal exactly 0 or 1. Because infinite values are not useful in the statistical model, we first need to convert $LGD_i \in [0,1]$ into $L_i \in (0,1)$ before we can apply the inverse Gaussian transformations. Specifically, L_i is defined as LGD_i when $LGD_i \in (0,1)$, $LGD_i - \epsilon$ when LGD_i = 1, and $LGD_i + \epsilon$ when LGD_i = 0. In the previous expressions, ϵ is a small positive number. The inverse Gaussian transformations are applied to L_i to produce Z_i. A linear regression model is assumed and estimated for Z_i (i.e., $Z_i = X_i\beta + e_i$) and then predictions for LGD_i are obtained from predictions for Z_i and L_i.

More specifically, define $\hat{L}_i$ as the predictor for L_i, and we can invert $\hat{Z}_i = X_i\,\hat{\beta}$ to produce the retransformed predictor $\hat{L}_i = h^{-1}(\hat{Z}_i; \alpha) = h^{-1}(X_i\,\hat{\beta}; \alpha)$. The smearing estimator works as follows. First, the empirical CDF of the estimated residuals is computed as

$$\hat{F}_N(r) = \frac{1}{N}\sum_{j=1}^{N} I(\hat{e}_j \leq r) \tag{A.1}$$

where $\hat{e}_i = \hat{Z}_i - X_i\,\hat{\beta}$, N is the number of observations, and $I(A)$ denotes the indicator function of the event A. Second, using the empirical CDF, an estimate of the mean is expressed as

$$\bar{E}(L_i|X_i) = \frac{1}{N}\sum_{j=1}^{N} h^{-1}(X_i\beta + \hat{e}_j; \alpha) \tag{A.2}$$

Because β is unknown, the third step is to plug in the OLS estimator and obtain

$$\hat{E}(L_i|X_i) = \frac{1}{N}\sum_{j=1}^{N} h^{-1}(X_i\hat{\beta} + \hat{e}_j; \alpha) \tag{A.3}$$

For details of this method, please refer to Li, Qi, Zhang, and Zhao (2016).

Li, Qi, Zhang, and Zhao (2016) show that results from the IG naïve method is sensitive to the choice of ϵ, the small value to adjust values at 0 or 1. Results based on IG smearing, by contrast, is not very sensitive to the choice of ϵ. For this reason, we use the IG smearing method in this paper. We use $\epsilon = 0.000001$ in the paper.

A.2 Fractional response regression (FRR)

This simple quasi-likelihood method was proposed by Papke and Wooldridge (1996) to model a continuous variable ranging between 0 and 1 and to perform asymptotically valid inference. The model specification is as follows:

$$E(LGD|X) = G(X\beta) \tag{A.4}$$

where X is a vector of explanatory variables, β is a vector of model parameters, and we choose the logistic function for $G()$,

$$G(X\beta) = \frac{1}{1 + \exp(-X\beta)} \tag{A.5}$$

To estimate the coefficients β, Papke and Wooldridge (1996) recommend maximizing the following quasi-log-likelihood function:

$$\sum_i l_i(\hat{\beta}) = \sum\nolimits_i \{LGD_i \times \log\left[G\left(X_i\hat{\beta}\right)\right] + (1 - LGD_i) \times \log[1 - G(X_i\hat{\beta})]\} \tag{A.6}$$

A.3 Censored Gamma Regression

Sigrist and Stahel (2011) introduce gamma regression models to estimate LGD. The probability function for the ith observation is

$$P_i(LGD; \xi, \alpha, \theta_i) = \begin{cases} P(LGD = 0) = G(\xi, \alpha, \theta_i) \\ P\big(LGD \in (l, l + dl)\big) = g(l + \xi, \alpha, \theta_i)dl, \qquad i\ 0 < l < 1 \\ P(LGD = 1) = 1 - G(1 + \xi, \alpha, \theta_i) \end{cases} \tag{A.7}$$

where $g(u;\ \alpha, \theta_i) = \frac{1}{\theta_i^{\alpha}\Gamma(\alpha)}\ u^{\alpha-1}\ e^{-u/\theta_i}$ and $G(u;\ \alpha, \theta_i) = \int_0^u g(x;\ \alpha, \theta_i)dx$ are the PDF and CDF for a gamma random variable, respectively. Also, $\alpha > 0$, $\xi > 0$, and $\theta_i > 0$. Note that Sigrist and Stahel (2011) define the underlying latent variable to follow a gamma distribution shifted by $-\xi$. The use of a gamma distribution with a shifted origin, instead of a standard gamma distribution, is motived by the fact that the lower censoring occurs at zero.

The connection between explanatory variables X_i and the expected LGD for the *i*th observation is through the linear equations as follows:

$$\log(\alpha) = \alpha^*, \log(\xi) = \xi^*, \log(\theta_i) = X_i\beta \;, \tag{A.8}$$

where β is the vector of model coefficients. These coefficients and the parameters α^* and ξ^* are estimated by maximizing the log likelihood function. The resulting estimates are then used to obtain LGD predictions: $E(LGD_i) = \alpha\theta_i[G(1+\xi,\alpha+1,\theta_i) - G(\xi,\alpha+1,\theta_i)] + (1+\xi)\big(1 - G(1+\xi,\alpha,\theta_i)\big) - \xi(1 - G(\xi,\alpha,\theta_i))$. For more detail on the censored gamma regression, refer to Sigrist and Stahel (2011).

A.4 Two-Tiered Gamma Regression

Sigrist and Stahel (2011) extend the censored gamma model into the two-tiered gamma model. This extension allows for two underlying latent variables, one that governs the probability of LGD being 0 and another for LGD being in (0, 1). The extension is useful in that it allows each latent variable to have its own set of explanatory variables and parameters.

More specifically, the two-tiered gamma regression assumes that there are two latent variables: the first latent variable, L_1^*, which follows a shifted gamma distribution with density function $g(L_1^* + \xi, \alpha, \tilde{\theta}_i)$, and the second variable, L_2^*, which is a shifted gamma distribution lower truncated at zero with the density function $g(L_2^* + \xi, \alpha, \theta_i)$,. These two latent variables are then related to LGD through

$$\begin{gathered} LGD = 0 \; if \; L_1^* < 0 \\ LGD = L_2^* \; if \; 0 < L_1^*, L_2^* < 1 \\ LGD = 1 \; if \; 0 < L_1^*, 1 \le L_2^* \end{gathered} \tag{A.9}$$

The distribution of LGD can be characterized as follows:

$$P_i\left(LGD;\ \xi, \alpha, \widetilde{\theta_i}, \theta_i\right) = \begin{cases} P(LGD = 0) = G(\xi, \alpha, \widetilde{\theta_i}) \\ P\left(LGD \in (l, l + dl)\right) = g(l + \xi, \alpha, \theta_i)\frac{1-G(\xi,\alpha,\widetilde{\theta_i})}{1-G(\xi,\alpha,\theta_i)}dl, if\ 0 < l < 1 \\ \\ P(LGD = 1) = 1 - G(1 + \xi, \alpha, \theta_i)\frac{1-G(\xi,\alpha,\widetilde{\theta_i})}{1-G(\xi,\alpha,\theta_i)} \end{cases}$$

(A.10)

The connection between the explanatory variables X_i and the expected LGD is through the linear equations as follows:

$$\log(\alpha) = \alpha^*$$

$$\log(\xi) = \xi^* \qquad \text{(A.11)}$$

$$\log\left(\widetilde{\theta_i}\right) = X_i\beta \qquad \text{(A.12)}$$

$$\log(\theta_i) = X_i\gamma \qquad \text{(A.13)}$$

where β, γ are vectors of model coefficients. These coefficients and the parameters $\alpha^*$and $\xi^*$are estimated by maximizing the log likelihood.

A.5 Inflated Beta Regression

Ospina and Ferrari (2010a) propose inflated beta distributions that are mixtures between a beta distribution and a Bernoulli distribution degenerated at 0, 1, or both 0 and 1. Ospina and Ferrari (2010b) then further develop inflated beta regressions by assuming the response distribution to follow the inflated beta and by incorporating explanatory variables into the mean function. Ospina and Ferrari (2010a) propose that the probability function for the *i*th observation is

$$P_i\left(LGD;\ P_0^i, P_1^i, \mu^i, \phi\right) = \begin{cases} P_0^i & if\ LGD = 0 \\ \left(1 - P_0^i - \ P_1^i\right) f\left(LGD;\ \mu^i, \phi\right) & if\ LGD \in (0,1) \\ P_1^i & if\ LGD = 1 \end{cases} \quad \text{(A.14)}$$

where $0 < \mu^i < 1$, $\phi > 0$, and $f(\cdot)$ is a beta probability density function (PDF), i.e.,

$$f\left(LGD;\ \mu^i, \phi\right) = \frac{\Gamma(\phi)}{\Gamma(\mu^i \phi)\Gamma\left((1-\mu^i)\phi\right)} LGD^{\mu^i \phi - 1}(1 - LGD)^{(1-\mu^i)\phi - 1} \quad \text{(A.15)}$$

Note that μ^i is the mean of the beta distribution, and ϕ is interpreted as a dispersion parameter. The mean function is $E(LGD_i) = P_1^i + \ \mu^i\left(1 - P_0^i - P_1^i\right)$. The connection between explanatory variables X_i and the expected LGD is through the three equations as follows:

$$P_0^i = e^{X_i \alpha}/(1 + e^{X_i \alpha} + e^{X_i \beta}) \quad \text{(A.16)}$$

$$P_1^i = e^{X_i \beta}/(1 + e^{X_i \alpha} + e^{X_i \beta}) \quad \text{(A.17)}$$

$$\mu^i = e^{X_i \gamma}/(1 + e^{X_i \gamma}) \quad \text{(A.18)}$$

where the parameters α, β, γ are model coefficients. These coefficients along with ϕ are estimated by maximizing the log likelihood function. For details on the inflated beta regression in general, see Ospina and Ferrari (2010b), Pereira and Cribari-Neto (2010), and Yashkir and Yashkir (2013).[18]

A.6 Beta Regression

Beta regression was proposed by Duan and Hwang (2014). These authors assume that the recovery rate LGD_i is defined as

[18] Our parameterizations of the probabilities in (13) and (14) are slightly different from the ones in Yashkir and Yashkir (2013). Our parameterizations ensure that each probability is positive and that the mixture weights in (11) sum to 1, while the parameterizations in Yashkir and Yashkir (2013) do not guarantee that P_0^i+ P_1^i<1, resulting in mixture weights in (11) that may be negative for $L \in (0, 1)$.

$$LGD_i = \begin{cases} 0 & if\ Z_i \in (-C_l, 0] \\ Z_i & if\ Z_i \in (0, 1) \\ 1 & if\ Z_i \in [1, 1 + C_u) \end{cases} \tag{A.19}$$

where the two constants $C_l \geq 0$, $C_u \geq 0$, and $\frac{Z_i + C_l}{1 + C_l + C_u}$ follows a beta distribution with parameters, a_i and b_i, and $a_i = \ln\{1 + \exp(x_i\theta)\}$ and $b_i = \ln\{1 + \exp(x_i\Psi)\}$. The parameters c_l, c_u, θ, and Ψ can be estimated by maximizing the log likelihood function. For more details of this method, please see Duan and Hwang (2014).

References:

Altman, E., and E.A. Kalota, 2014, Ultimate recovery Mixtures, *Journal of Banking and Finance* 40: 116–129.

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

Bastos, J.A., 2010, Forecasting bank loans loss-given-default, *Journal of Banking and Finance* 34, 2510-2517.

Cohen, J., Cohen, P., West, S.G., and Aiken, L.S., 2013, Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, Routledge.

Duan, Jin-Chuan and Ruey-Ching Hwang, 2014, Predicting recovery rates at the time of corporate default, working paper, National University of Singapore.

Ferrari, Sivia, and Francisco. Cribari-Neto, 2004, Beta regression for modeling rates and proportions, Journal of Applied Statistics 31,799-815.

Hartmann-Wendels, T., P. Miller, and E. Tows. 2014. Loss given default for leasing: Parametric and nonparametric estimations, *Journal of Banking and Finance* 40: 364–375.

Hu,Y, and M. Perraudin, 2002, The dependence of recovery rates and defaults, Working paper, Birkbeck College.

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

Kruger, Steffen, and Daniel Rosch, 2017, Downturn LGD modeling using quantile regression, *Journal of Banking and Finance* 79, 42-56.

Leymarie, Jeremy, Christophe Hurlin, and Antoine Patin, 2018, Loss functions for LGD models comparison, *European Journal of operational Research*, 268(1), 348-360

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

Li, P., 2018, Efficient MCMC estimation of inflated beta regression models, *Computational Statistics* 33(1): 127-158.

Loterman, G., I. Brown, D. Martens, C. Mues, and B. Baesens, 2012, Benchmarking regression algorithms for loss given default modeling, *International Journal of Forecasting* 28: 161–170.

Ospina, R. and S. L. P. Ferrati, 2010b, Inflated beta regression models, working paper, Universidade Federal de Pernambuco and Universidade de Sao Paulo.

Papke, L. E., and J. M. Wooldridge, 1996, Econometric methods for fractional response variables with an application to 401(k) plan participation rates, *Journal of Applied Econometrics* 11, 619-632.

Qi, M. and X. Zhao, 2011, A comparison of methods to model loss given default, *Journal of Banking and Finance* 35, 2842-2855.

Ramalho, E and J. Ramalho, 2010, Is neglected heterogeneity really an issue in binary and fractional regression models? A simulation exercise for logit, probit and loglog models, Computational Statistics & Data Analysis 54 (4), 987-1001.

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

Tobback, E., D. Martens, T.V. Gestel, and B. Baesen,. 2014, Forecasting loss given default models: Impact of account characteristics and macroeconomic State, *Journal of the Operational Research Society* 65: 376–392.

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

Table 1: Model Fit

Panel A: Coefficient estimates of the IB, the true model

	Alpha			Beta			Gamma			Phi		
	True parameter	Estimation		True parameter	Estimation		True parameter	Estimation		True parameter	Estimation	
		Coeff	SE		Coeff	SE		Coeff	SE		Coeff	SE
X1	0.1	0.105	0.015	-1	-0.994	0.016	0	0.001	0.011	1.6	1.599	0.014
X2	-0.05	-0.051	0.002	0.06	0.060	0.002	0.005	0.004	0.002			
X3	0.4	0.400	0.008	-0.1	-0.105	0.008	-0.1	-0.110	0.006			
X4	0.4	0.406	0.008	-0.1	-0.104	0.008	-0.1	-0.111	0.006			
X5	0.4	0.398	0.008	-0.1	-0.097	0.008	-0.1	-0.090	0.006			
X6	0.4	0.403	0.008	-0.1	-0.098	0.008	-0.1	-0.104	0.006			
X7	0.4	0.391	0.008	-0.1	-0.107	0.008	-0.1	-0.114	0.006			
X8	0.4	0.404	0.008	-0.1	-0.102	0.008	-0.1	-0.100	0.006			
X9	0.4	0.396	0.008	-0.1	-0.100	0.008	-0.1	-0.094	0.006			
X10	0.4	0.418	0.008	-0.1	-0.091	0.008	-0.1	-0.107	0.006			
X11	0.4	0.406	0.008	-0.1	-0.087	0.008	-0.1	-0.106	0.006			

Panel B: Full set of explanatory variables on the RHS

	SSE	R2	Pearson	Kendall	Spearman
OLS	68148.936	0.0770	0.278	0.204	0.284
IG Smearing	68542.935	0.0717	0.277	0.204	0.284
FRR	68149.380	0.0770	0.278	0.204	0.284
CG	68175.058	0.0767	0.278	0.204	0.284
TTG	68142.629	0.0771	0.278	0.204	0.284
BR	68140.991	0.0771	0.278	0.204	0.284
IB	68122.967	0.0774	0.278	0.204	0.284

Panel C: Omission of four explanatory variables from the RHS

	SSE	R2	Pearson	Kendall	Spearman
OLS	70632	0.043	0.208	0.152	0.213
IG Smearing	70786	0.041	0.208	0.152	0.213
FRR	70631	0.043	0.208	0.152	0.213
CG	70650	0.043	0.208	0.152	0.213
TTG	70641	0.043	0.208	0.152	0.213
BR	70632	0.043	0.208	0.152	0.213
IB	70626	0.043	0.209	0.152	0.213

Table 2: KS Statistics

	Full set of explanatory variables	Omitting explanatory variables	
		4 Omitted	**8 Omitted**
OLS	0.203	0.204	0.203
IG Smearing	0.346	0.346	0.346
CG	0.023	0.023	0.023
TTG	0.021	0.021	0.020
BR	0.004	0.004	0.004
IB	0.001	0.001	0.001

Figure 1 – Original distributions of data generated from the IB distribution with the true parameters shown in Panel A of Table 1.

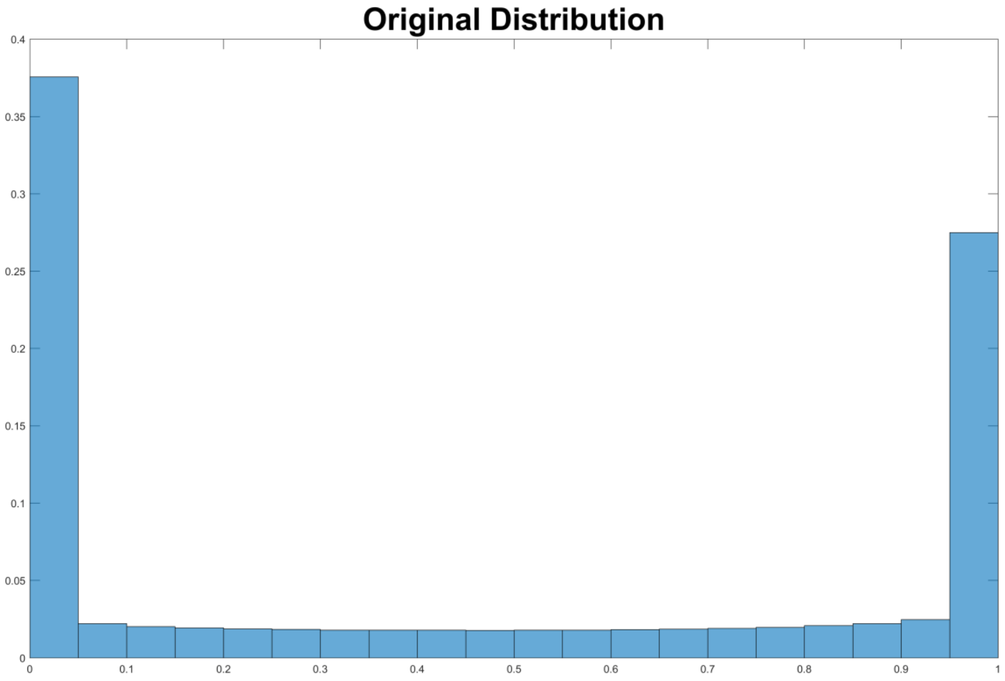


Figure 2: Predicted means from a full set of explanatory variables on the RHS

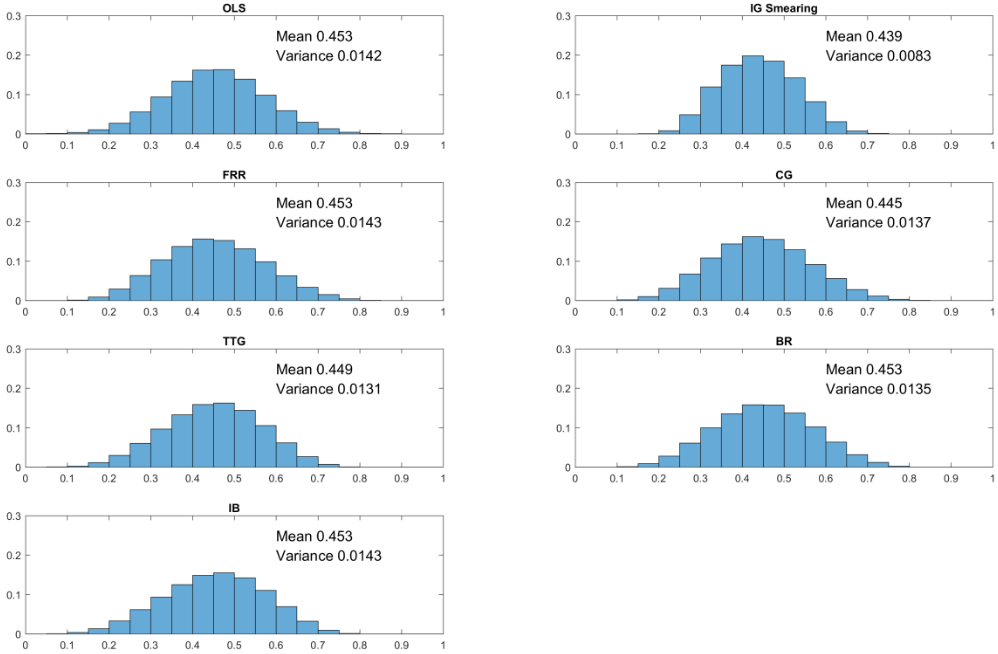


Figure 3: Kernel densities of predicted means from various model specifications

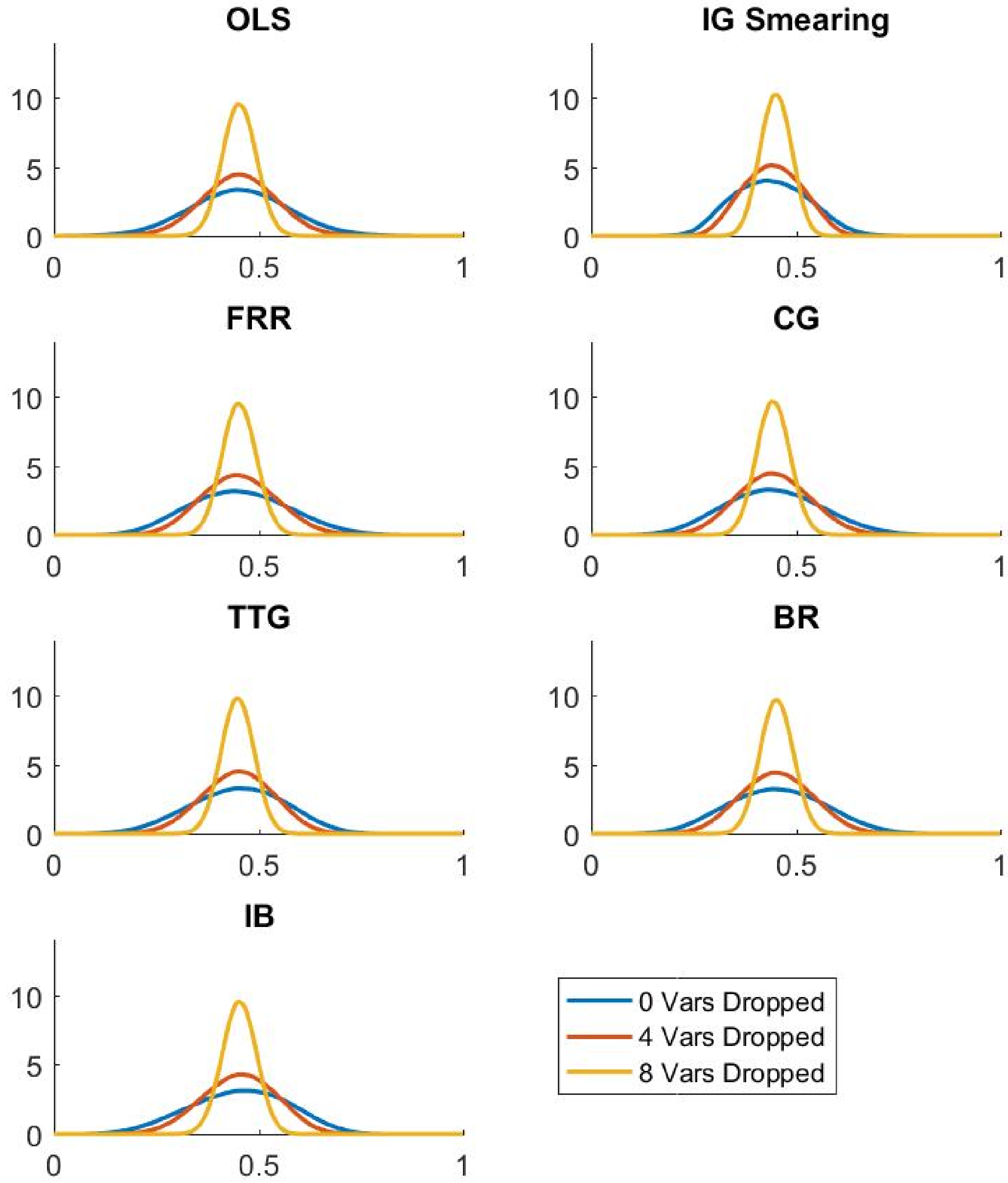
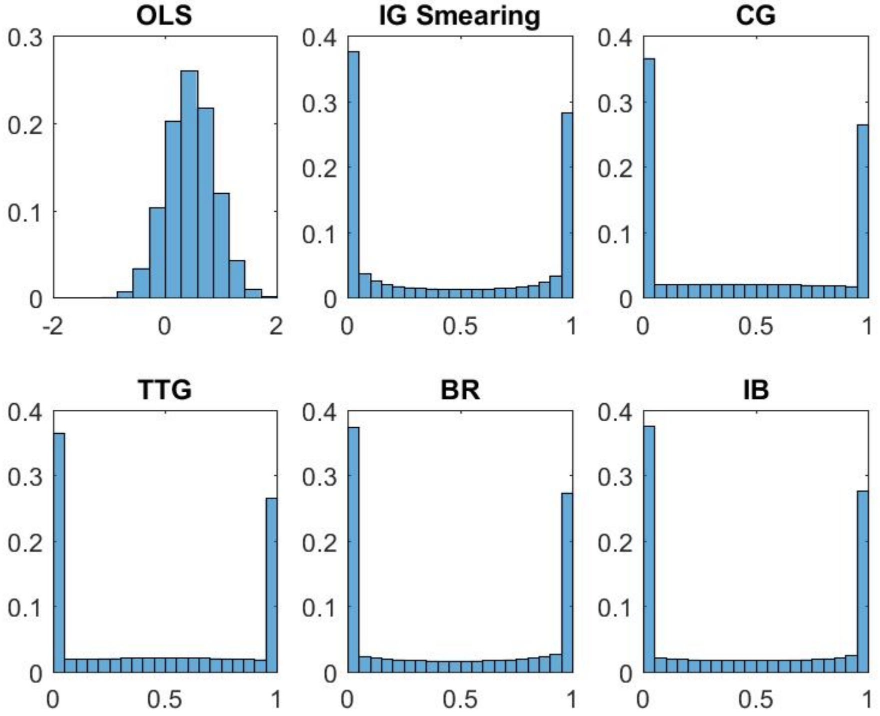


Figure 4: Predicted distributions from a full set of explanatory variables on the RHS

Panel A, Predicted distributions of OLS, IG Smearing, TTG and CG, TTG, IB, and BR



Panel B, Predicted distributions for FRR with different alphas

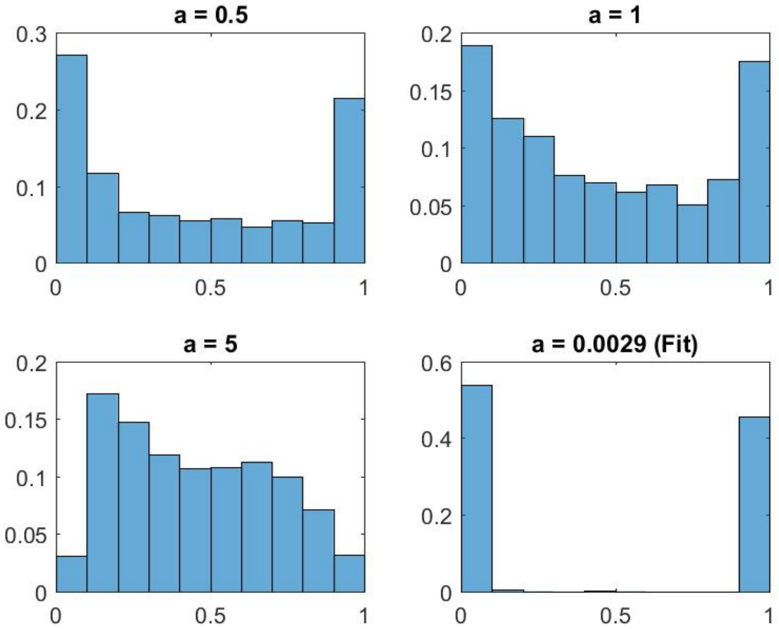
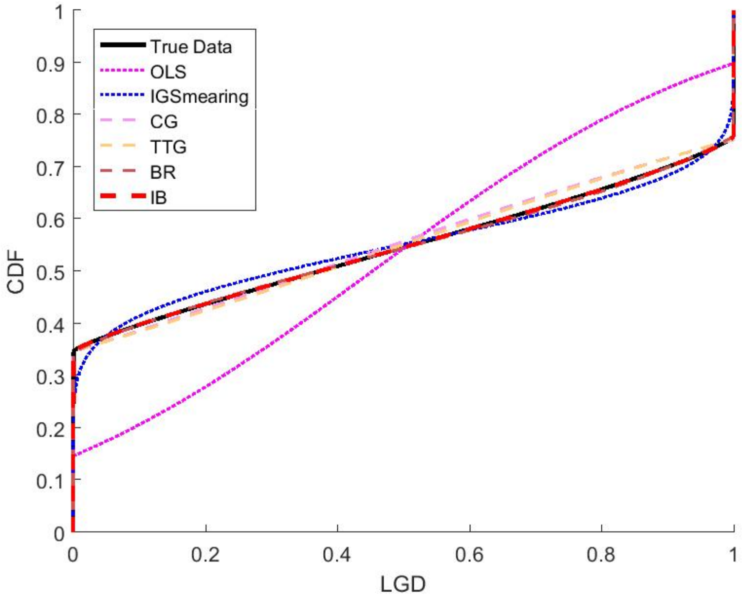


Figure 5 CDFs of the predicted unconditional distributions

Panel A: Full specification



Panel B: Omitting four explanatory variables from the RHS

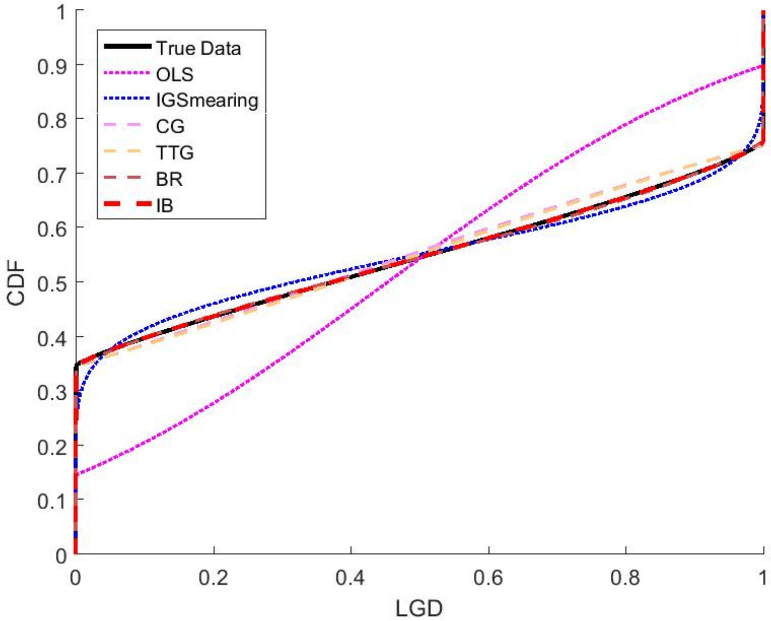
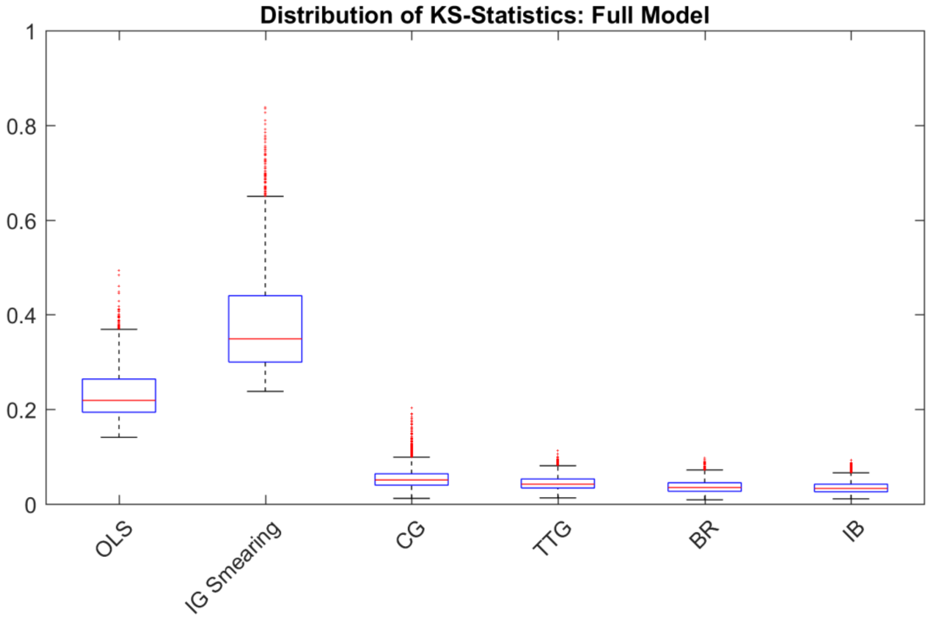


Figure 6: Predicted conditional distributions - KS statistics

Panel A:



Panel B:

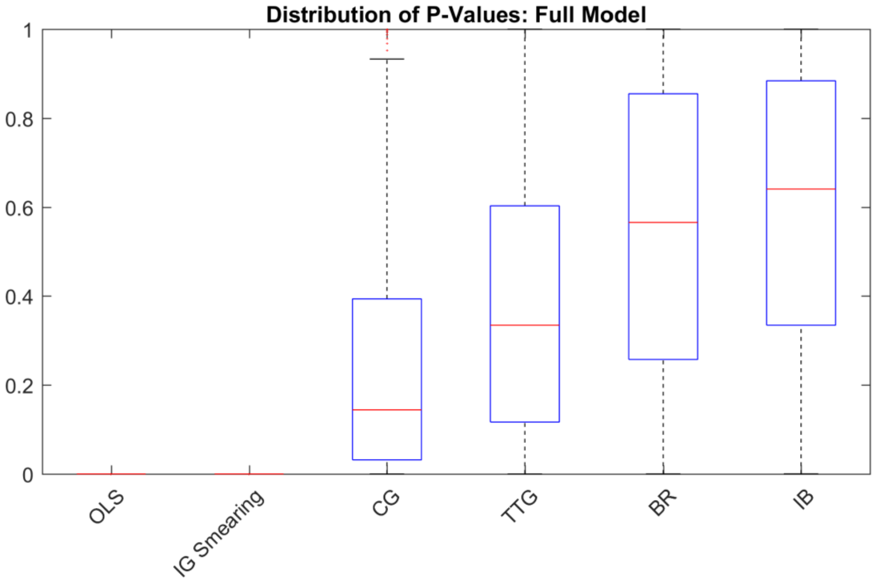
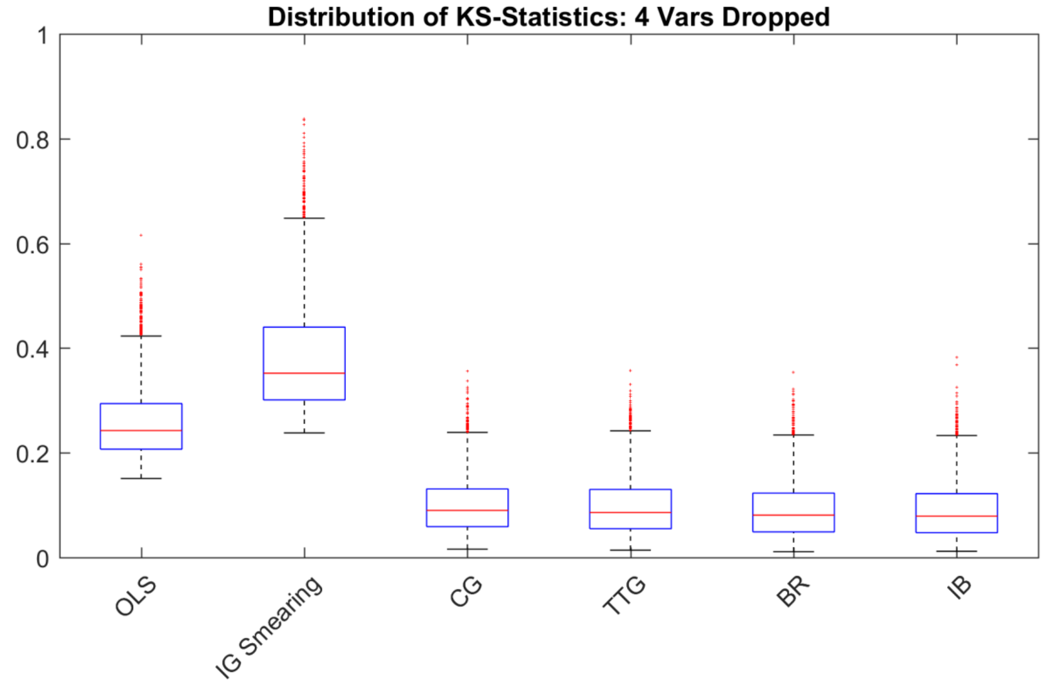


Figure 6 Continued

Panel C:



Panel D:

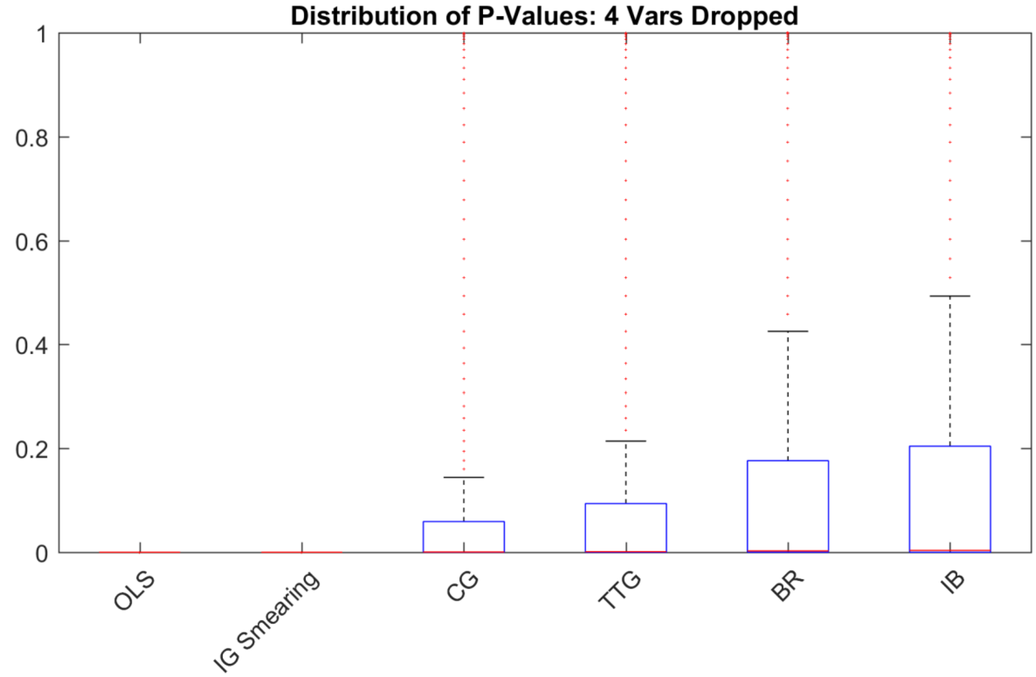
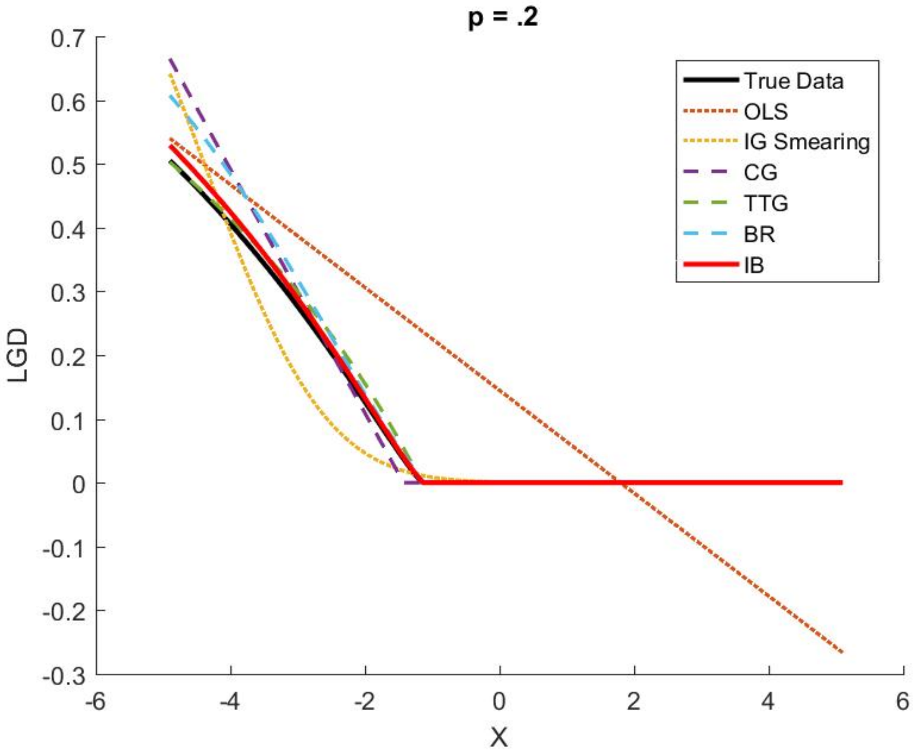


Figure 7 Predicted conditional distributions - Quantile plots – Full specification



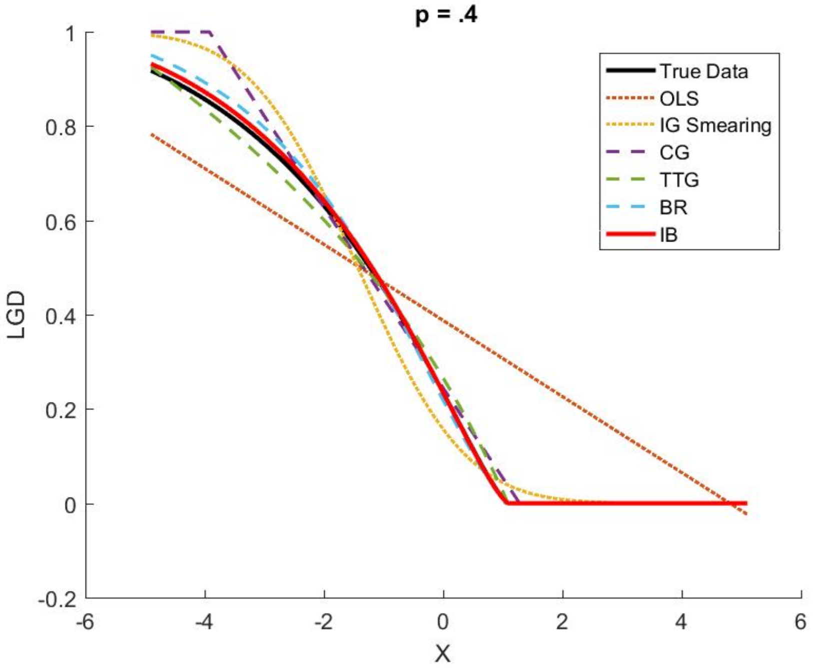
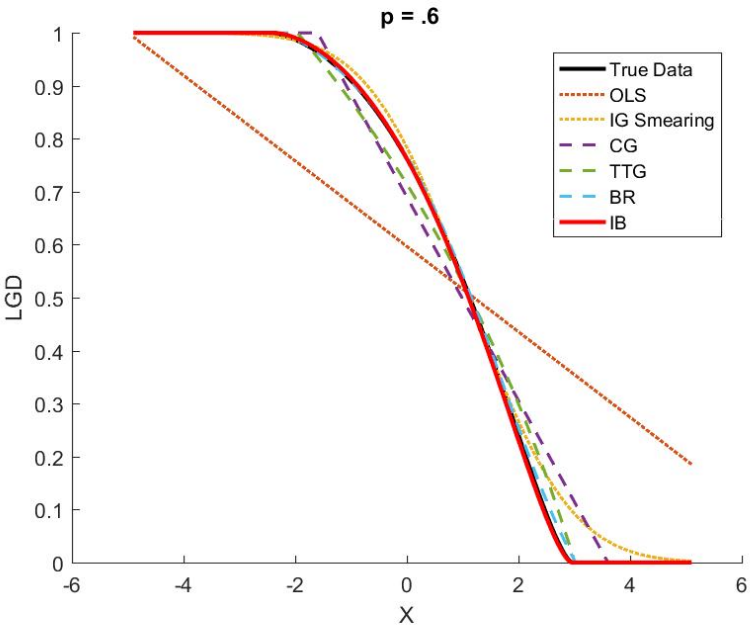


Figure 7 Continued



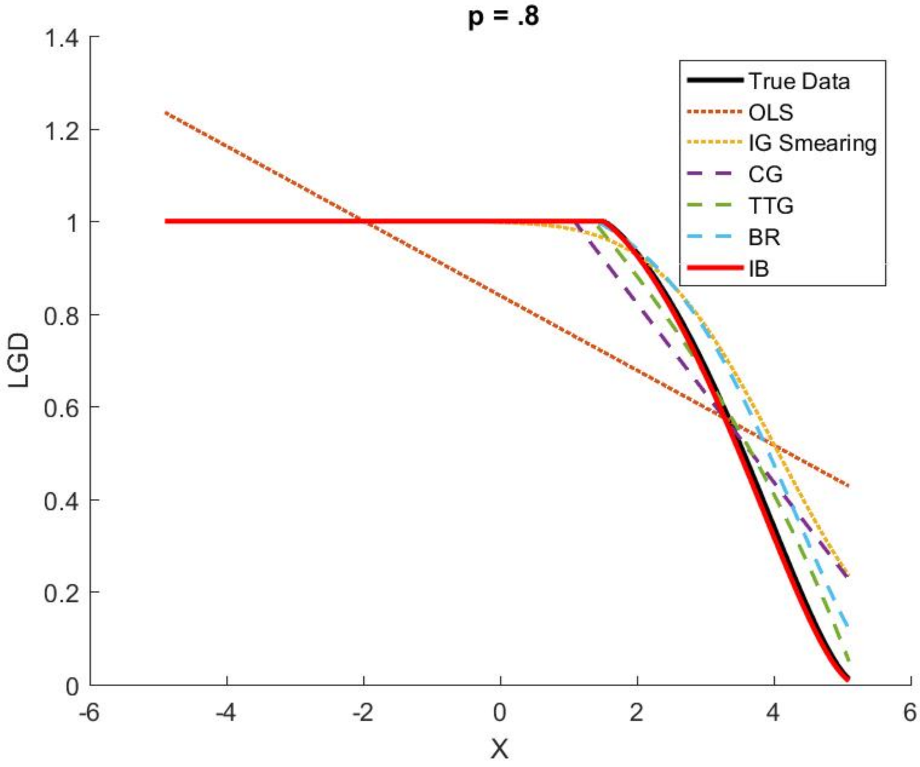


Figure 8 Marginal effects

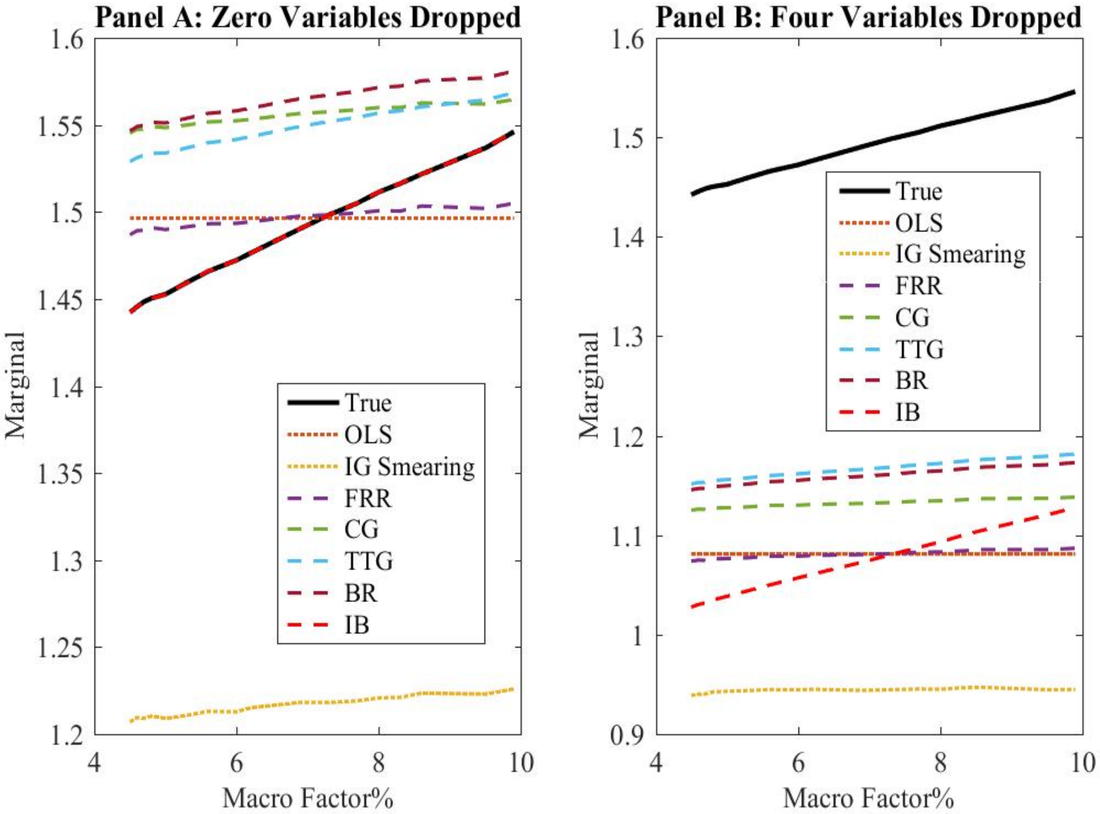
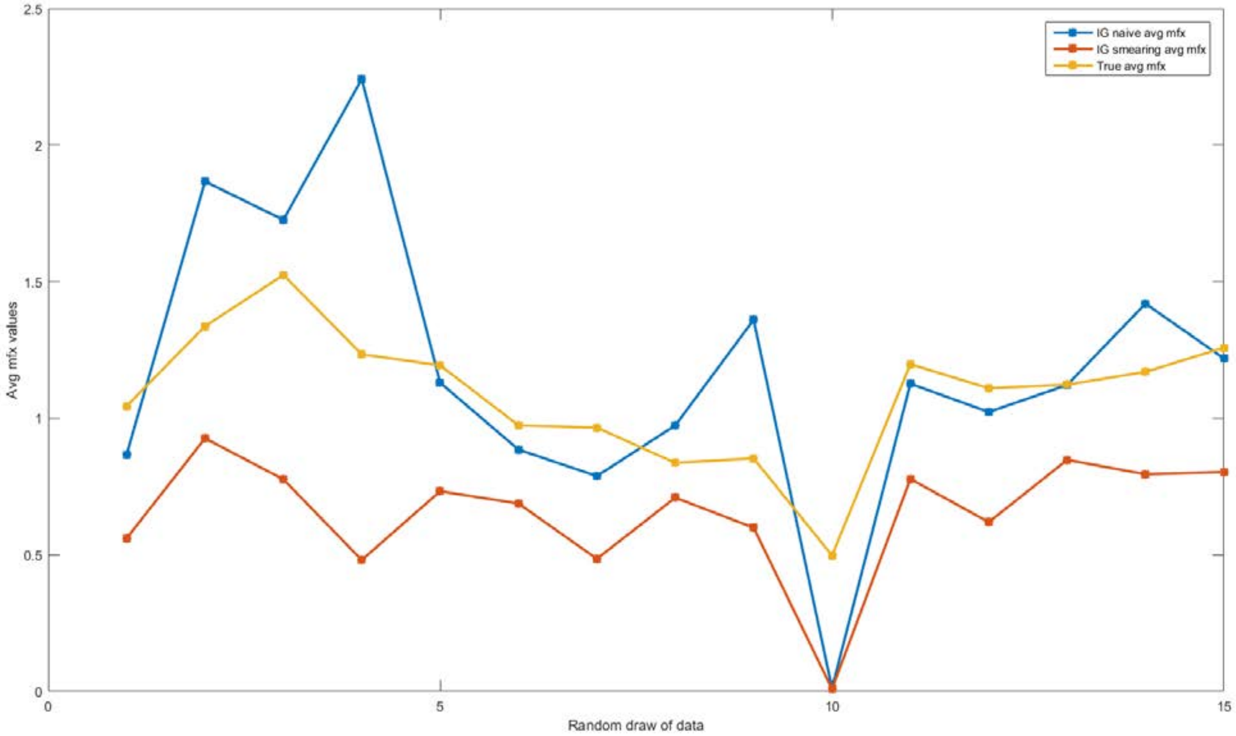


Figure 9 IG naïve versus the IG smearing method


IG naive avg mfx
IG smearing avg mfx
True avg mfx
Avg mfx values
Random draw of data