



Research article

Validation of corporate probability of default models considering alternative use cases and the quantification of model risk

Michael Jacobs, Jr.*

CFA, Lead Quantitative Analytics & Modeling Expert, Head – C&I 1st Line Model Development Validation & Quality Assurance, PNC Financial Services Group – Balance Sheet Analytics & Modeling / Model Development, 340 Madison Avenue, New York, N.Y. 10173, U.S.A.

* **Correspondence:** Email: michael.jacobsjr@pnc.com; Tel: 917-324-2098.

Abstract: In this study we consider the construction of *through-the-cycle* (“TTC”) *probability-of-default* (“PD”) models designed for credit underwriting uses and *point-in-time* (“PIT”) PD models suitable for early warning uses, considering which validation elements should be emphasized in each case. We build PD models using a long history of large corporate firms sourced from Moody’s, with a large number of financial, equity market and macroeconomic candidate explanatory variables. We construct a Merton model-style *distance-to-default* (“DTD”) measure and build hybrid structural and reduced-form models to compare with the financial ratio and macroeconomic variable-only models. In the hybrid models, the financial and macroeconomic explanatory variables still enter significantly and improve the predictive accuracy of the TTC models, which generally lag behind the PIT models in that performance measure. While all classes of models have high discriminatory power by most measures, on an out-of-sample basis the TTC models perform better than the PIT models. We measure the model risk attributable to various model assumptions according to the principle of *relative entropy* and observe that omitted variable bias with respect to the DTD risk factor, neglect of interaction effects and incorrect link function specification has the greatest, intermediate and least impacts, respectively. We conclude that care must be taken to judiciously choose how we validate TTC vs. PIT models, as criteria may be rather different and apart from standards such as discriminatory power. This study contributes to the literature by providing expert guidance to credit risk modeling, model validation and supervisory practitioners in managing model risk.

Keywords: probability of default; point-in-time; through-the-cycle; credit risk; model validation; credit underwriting; early warning systems; regulatory capital; model risk

JEL Codes: G21, G28, M40, E47.

1. Introduction and summary

It is expected that financial market participants have accurate measures of a counterparty's capacity to fulfill future debt obligations, conventionally measured by a credit rating or a score, and typically associated with a *probability of default* ("PD"). Most extant risk rating methodologies distinguish model outputs considered *point-in-time* ("PIT") vs. *through-the-cycle* ("TTC"). Although these terminologies are widely used in the credit risk modeling community, there is some confusion about what these terms precisely mean. In our view, based upon first-hand experience in this domain and a comprehensive literature review, at present a generally accepted definition for these concepts remains elusive, apart from two points of common understanding. First, PIT PD models should leverage all available information, borrower-specific and macroeconomic, which most accurately reflect default risk at any point of time. Second, TTC PD models abstract from cyclical effects and measure credit risk over a longer time period encompassing a mix of economic conditions, exhibiting "stability" of ratings wherein dramatic changes are related mainly to fundamental and not transient economic fluctuations. However, in reality this distinction is not so well defined, as idiosyncratic factors can influence systematic conditions (e.g., credit contagion), and macroeconomic conditions can influence obligors' fundamental creditworthiness.

There is an understanding in the industry of what distinguishes PIT and TTC constructs, typically defined by how PD estimates behave with respect to the business cycle. However, how this degree of "TTC-ness" vs. "PIT-ness" is defined varies considerably across institutions and applications, and there is no consensus around what thresholds should be established for certain metrics, such as measures of ratings volatility. As a result, most institutions characterize their rating systems as "Hybrid". While this may be a reasonable description, as arguably the TTC and PIT constructs are ideals, this argument fails to justify the use cases of a PD model where there may be expectations that the model is closer to either one of these poles.

In this study, we develop empirical models that avoid formal definitions of PIT and TTC PDs, rather deriving constructs based upon common sense criteria prevalent in the industry, and illustrating which validation techniques are applicable to these approaches. Based upon this empirical approach, we characterize PIT and TTC credit risk measures and discuss the key differences between both rating philosophies. In the process, we address the validation of PD models under both rating philosophies, highlighting that the validation of either system exhibits a particular set of challenges. In the case of the TTC PD models, in addition to flexibility in determining measurement of the cycle, there are unsettled questions around the rating stability metric thresholds. In the case of PIT PD models, there is the additional question of demonstrating the accuracy of PD estimates at the borrower level, which may not be obvious from observing average PD estimates versus default rates over time. Finally, considering both types of models, there is the question of whether the relative contributions of risk factors are conceptually intuitive, as we would expect that certain variables would dominate in either of these constructs.

There are some additional comments in order to motivate this research. First, there is a misguided perception in the literature and industry that PIT models contain only macroeconomic factors, and that TTC models contain only financial ratios, whereas from a modeling perspective there are other dimensions that define this distinction that we elaborate upon in this research. Furthermore, it may be argued that the validation of a TTC or PIT PD model involves assessing the validity of the cyclical factor, which if not available to the validator, may be accounted for only implicitly. One possibility is for the underlying cycle to be estimated from historical data based upon some theoretical framework, but in this study, we prefer commonly used macroeconomic factors in conjunction with obligor level default data, in line with industry practice. Related to this point, we do not explicitly address how TTC PD models can be transformed into PIT PD rating models, or vice versa. While the advantage of such alternative constructs is that they can be validated based upon an assumption regarding the systematic factor using the methodologies applicable to each type of PD model, we prefer to validate each as specifically appropriate. The rationale for our approach is that the alternative runs the risk of introducing significant model risk, thereby leading to the validity of such validation being rendered questionable as compared to testing a pure PIT or TTC PD model.

We employ a long history of borrower level data sourced from Moody's, around 200,000 quarterly observations from a large population of rated larger corporate borrowers (at least USD 1 billion in sales and domiciled in the U.S. or Canada), spanning the period from 1990 to 2015. The dataset is comprised of an extensive set of financial ratios, macroeconomic and equity market variables as candidate explanatory variables. We build a set of PIT models with a 1-year default horizon and macroeconomic variables, and a set of TTC models with a 3-year default horizon and only financial risk factors.

The position of this research in the academic literature is at the intersection of two streams of inquiry. First, there are a series of empirical studies that focus on the factors that determine corporate default and the forecasting of this phenomenon, which include Altman (1968), Jarrow and Turnbull (1995) and Duffie and Singleton (1999b). At the other end of the spectrum, there are mainly theoretical studies that focus on modeling frameworks for either understanding corporate default e.g., Merton, (1974), or else for perspectives on the TTC vs. PIT dichotomy e.g., Kiff et al., 2004; Aguais et al., 2008; Cesaroni, 2015. In this paper, we blend these considerations of theory and empirics, while also addressing the prediction of default and TTC/PIT construct.

We would like to emphasize what we believe to be the principal contributions of this paper. First, in terms of methodology, we assert this to be mainly in the domain of practical application rather than methodological innovation. Many practitioners, especially in the wholesale credit and banking book space, still use the techniques employed in this paper. We see our contribution as proposing a structured approach to constructing a suite of TTC and PIT models, while combining reduced form and structural modeling aspects, and then by further proposing a framework for model validation. We would note that many financial institutions in this space do not have such a framework. For example, a lot of banks are still using TTC Basel models that are modified for PIT uses, such as stress testing or portfolio management. Furthermore, a preponderance of banks in this space do not employ hybrid financial and Merton-style models for credit underwriting. In sum, our contribution transcends the academic literature to address issues relevant to financial institution practitioners in the credit risk modeling space, which we believe uniquely positions this research. Second, we would further like to emphasize our contribution in terms of modeling data, which we believe to be more extensive and richer than most of the prior literature, in that we combine a variety of datasets over a rather long historical period and have a very large number of candidate explanatory variables. Finally, we believe that we have made a contribution in terms of our

conclusions, which are rather multifaceted and nuanced in contrast with the majority of the prior literature. The implications of our study span the domains of prudential supervision guidance, financial theory, as well as tools for bank risk modelers and validators.

The summary of our empirical results are as follows. We present the leading two models each in the classes of PIT and TTC design, all having favorable rank ordering power, intuitive relative weights on explanatory variables and rating mobility metrics. We also perform predictive accuracy analysis and specification testing, where we observe that the TTC designs are more challenged than the PIT designs in performance, and that unfortunately all designs show some signs of model misspecification. This observation argues for the consideration of alternative risk factors, such as equity market information. In view of this, from the market value of equity and accounting measures of debt for these firms, we are able to construct a Merton model-style *distance-to-default* (“DTD”) measure and build hybrid structural reduced-form models, which we compare with the financial ratio and macroeconomic variable-only models. We show that adding DTD measures to our leading models does not invalidate the other variables chosen, significantly augments model performance and in particular increases the obligor-level predictive accuracy of the TTC models. We also find that while all classes of models have high discriminatory power by all measures, there are some conflicting results regarding predictive accuracy depending upon the measure employed, and that on an out-of-sample basis the TTC models actually perform better than the PIT models. Finally, we perform an exercise in which we measure the model risk attributable to violating various model assumptions according to the *principle of relative entropy*. In the latter experiment we observe that omitted variable bias (with respect to the DTD) has the greatest impact, the incorrect specification of the link function has the least impact, and the neglect of interaction effects amongst risk factors has an intermediate impact to measured model risk.

Finally, let us introduce the remainder of this paper, which will proceed as follows. In Section 2, we review the relevant literature, where we address a survey of PD modeling in general, as well as issues around rating philosophy in particular. In Section 3, we address modeling methodology, which we partition into the domains of econometric modeling and statistical assumptions. Section 4 encompasses the empirical analysis of this study, a description of the modeling data, estimation, validation results and the quantification of model risk. In Section 5, we conclude and summarize the study, discuss policy implications and provide thoughts on avenues for future research.

2. Review of the literature

Traditional credit risk models focus on estimating the PD, rather than on the magnitude of potential losses in the event of default (or *loss-given-default* – “LGD”), and typically specify “failure” to be bankruptcy filing, default, or liquidation, thereby ignoring consideration of the downgrades and upgrades in credit quality that are measured in *mark-to-market* (“MTM”) credit models. Such *default mode* (“DM”) models estimate credit losses resulting from default events only, whereas MTM models classify any change in credit quality as a credit event. There are three broad categories of traditional models used to estimate PD: expert systems, including artificial neural networks; rating systems; and credit scoring models.

The most commonly used traditional credit risk measurement methodology is the PD scoring model. The seminal model in this domain is the *multiple discriminant analysis* (“MDA”) of Altman (1968). While MDA is computationally convenient as it relies on the assumption of normal error terms and a linear model equation, with the increase in computational power of computers recently this is a very marginal benefit.

Mester (1997) documents the widespread use of credit scoring models amongst banks in the U.S., with 97% and 70% of them using them to approve credit card and small business loan applications, respectively. We are not surprised by this rapid spread that she documents, as credit scoring models are relatively inexpensive to implement and do not suffer from the subjectivity and inconsistency of expert systems. The spread of these models throughout the world was first surveyed by Altman and Narayanan (1997). Departing slightly from the conclusions of Mester (1997), the authors find that it is not so much the models' differences across countries of diverse sizes and in various stages of development that stand out, but rather their similarities. An example of popularly used vended PD scoring model in the industry is the private firm model of *Moody's Analytics* ("MA"; Dwyer et al, 2004), the flexibility and economy of which explains why so many banks use it.

In a departure from the credit scoring approach, Merton (1974) models equity in a levered firm as a call option on the firm's assets with a strike price equal to the debt repayment amount, basing the framework on financial contingent claims theory rather than a purely empirical construct. The PD is determined by valuing the call option using an iterative method to estimate the unobserved variables that determine this, the market value of assets and the volatility of assets, combined with the amount of debt liabilities that have to be repaid at a given credit horizon in order to calculate the firm's *distance-to-default* ("DTD"). DTD is the number of standard deviations between the current asset values and the debt repayment amount, so the higher it is, the lower the PD. While this is indeed an elegant construct, there are some very restrictive assumptions in play, which have been explored in both subsequent academic literatures, as well as in the practical realm in ways to implement this construct. In an important example of this, in the *CreditEdge*TM ("CE") public firm model of MA, historical default experience is used to estimate an empirical measure of the PD, denoted the *expected default frequency* ("EDF"). As CE EDF scores are obtained from equity prices, they are more sensitive to changing financial circumstances than external credit ratings that rely predominately on credit underwriting data.

Similar to the previous construct in drawing from financial contingent claims theory, which is sometime called the option-theoretic structural approach, other modern methods of credit risk measurement can be traced to alternative branches in the asset pricing literature of academic finance. In contrast to the structural approach, the *reduced form* approach utilizing intensity-based models to estimate stochastic hazard rates follows a study pioneered by Jarrow and Turnbull (1995) and Duffie and Singleton (1999b). This school of thought offers a differing methodology to accomplish the task of estimating of PDs. While the structural approach models the economic process of default, the reduced form models decompose risky debt prices in order to estimate the random intensity process underlying default. While this has the advantage of not relying on a description of the economy that is bound by strong assumptions that introduce model risk, the reliance on risky debt prices has been criticized as not purely measuring credit risk, as there are elements of market liquidity that confound the relationship. The proprietary model *Kamakura Risk Manager*TM, where the econometric approach (the so-called *Jarrow-Chava Model*) is a reduced-form model based upon the research of Chava and Jarrow (2004), attempts to explicitly adjust for such liquidity effects. However, noise from embedded options and other structural anomalies in the default risk-free market further distorts risky debt prices, thereby impacting the results of such intensity-based models.

One of the key motivations behind the new generation of PD models being developed in the industry, as well as in this research, is to provide a suite of models that can accommodate multiple uses, such as TTC models for credit underwriting or *risk weighted assets* ("RWA"), as well as PIT models for credit portfolio management or early warning. One point to highlight is that despite the growing literature on

TTC credit ratings, there is still no consensus on the precise definition of this concept, except the general agreement that TTC ratings do not reflect cyclical effects. The Basel guidelines (BIS 2006) describe a PIT rating system as a construct that uses all currently available obligor-specific and aggregate information to estimate an obligor's PD, in contrast to a TTC rating system that, while using obligor-specific information, tends not to adjust ratings in response to changes in macroeconomic conditions. However, the types of such cyclical effects and how they are measured differ considerably in the literature as well as in practice.

First, a number of studies have proposed formal definitions of PIT and TTC PD estimates and rating systems. These include Loeffler (2004), who explores the TTC methodology in a structural credit risk model based on Merton (1974), in which a firm's asset value is separated into a permanent and a cyclical component. In this model, TTC credit ratings are based on forecasting the future asset value of a firm under a stress scenario for the cyclical component. While a downside of this approach is that it relies on a stress scenario which is a subjective construct, this has the benefit of making a model robust to a downturn. Kiff et al. (2004) investigate the TTC approach also in a structural framework in which the definition of TTC ratings follows the one applied by Hamilton et al. (2011), emphasizing that while anecdotal evidence from credit rating agencies confirm their use of this TTC approach, it turns out that there is no single and simple definition of what a TTC rating actually means. In contrast to studies such as these that define PIT and TTC credit measures on the basis of a decomposition of credit risk into idiosyncratic and systematic risk factors, Aguais et al. (2008) follow a frequency decomposition view in which a firm's credit measure is split up into a long-term credit quality trend and a cyclical component which are filtered from the firm's original credit measure by using a smoothing technique based on the filter in Hodrick and Prescott (1997). Furthermore, the authors argue that in the existing literature, there has been little discussion about whether the C in TTC refers to the business cycle or the credit cycle and highlight that these cycles differ considerably from each other regarding their length. They describe a practical framework for banks to compute PIT and TTC PDs through converting PIT PDs into TTC PDs based on sector-specific credit cycle adjustments to the DTD credit measures of the Merton (1974) model derived from a credit rating agency's rating or MA's CE model. Furthermore, they qualitatively discuss key components of PIT-TTC default rating systems and how these systems can be implemented in banks. This approach offers practitioners much flexibility in adaptation to multiple uses, but is computationally intense and has many sub-dependencies in the model components which make it susceptible to model risk. In contrast, Cesaroni (2015) analyzes PIT and TTC default probabilities of large credit portfolios in a Merton single-factor model, where the author defines the TTC PD as the expected PIT PD, and where the expectation is taken over all possible states of a systematic risk factor. This more stylized construct has the benefit of being more parsimonious and easier to implement than some of the literature just described, but gives rise to model risk through relying on more restrictive assumptions. Finally, along this line of research, Repullo et al. (2010) propose translating PIT PDs into TTC PDs by *ex post* smoothing the estimated PIT PDs with countercyclical scaling factors, which is similar to the previously described papers in relying on some kind of translation between PIT and TTC designs based upon some model. This is connected with the industry's next-generation PD model redevelopment efforts and this research, as it aligns with the objective of supporting TTC vs. PIT ratings while not having formal definitions of what TTC or PIT means.

Second, several studies analyze the ratings of major rating agencies regarding their PIT vs. TTC orientation. These include the Altman and Rijken (2004) who find, based on credit scoring models, that major credit rating agencies pursue a long-term view when assigning ratings, putting less weight on short-

term default indicators which indicate a TTC orientation. In relation to this argument, Loeffler (2013) shows for Standard and Poor's and Moody's rating data that these agencies have a policy of changing a rating only if it is unlikely to be reversed in the future and argues that this can explain the empirical finding that rating changes lag changes of an obligor's default risk, consistent with the general view of TTC ratings. While Altman and Rijken (2006) also analyze the TTC methodology of rating agencies, they take an investor's PIT perspective and quantify the effects of this methodology on the objectives of rating stability, rating timeliness, and performance in predicting defaults. Among other results, they find that TTC rating procedures delay migration in agency ratings on average by $\frac{1}{2}$ a year on the downgrade side and $\frac{3}{4}$ of a year on the upgrade side, as well as that from the perspective of an investor's one-year horizon, TTC ratings significantly reduce the short-term predictive power for defaults. Several papers, such as Amato and Furfine (2004) and Topp and Perl (2010), take this line of inquiry a step further by analyzing actual rating data, showing that these ratings vary with the business cycle, even though these ratings are supposed to be TTC according to the policies of the credit rating agencies. Going back to Loeffler (2013), in relation to this thesis he estimates long-run trends in market-based measures of one-year PDs using different filtering techniques. He shows that agency ratings contribute to the identification of these long-run trends, thus providing evidence that credit rating agencies follow to some extent a TTC rating philosophy. In summary of this stream of research, many studies find that the ratings of major rating agencies show both PIT as well as TTC characteristics, which is consistent with the notion of hybrid rating systems. In connection with this research and industry redevelopment efforts, with the objective of supporting TTC vs. PIT ratings, these results support not having "hard" mobility metric thresholds in evaluating the model output.

Third, rating philosophy is important from a regulatory and supervisory perspective, as well as from a credit underwriting perspective, not least because capital requirements for banks and insurance firms depend upon credit risk measures. Studies that discuss TTC PDs in the context of Basel II (Bank for International Settlements, 2006 – "BIS"), or as a remedy for the potential procyclical nature of Basel II, include Repullo et al. (2010) who compare smoothing the input of the Basel II formula by using TTC PDs or smoothing its output with a multiplier based on GDP growth. They prefer the GDP growth multiplier because TTC PDs are worse in terms of simplicity, transparency, cost of implementation, and consistency with banks' risk pricing and risk management systems. Cyclicity of credit risk measures also plays an important role in the context of Basel III (BIS 2011), which states that institutions should have sound internal standards for situations where realized default rates deviate significantly from estimated PDs, and that these standards should take account of business cycles and similar systematic variability in default experience. In two separate consultation papers issued in 2016, the European Banking Authority (2016) proposes to explicitly leave the selection of the rating philosophy to the banks, whereas the Basel Committee for Banking Supervision (BIS, 2016; "BCBS") proposes requiring banks to follow a TTC approach to reduce the variability in PDs and thus RWAs across banks.

Finally, while it is widely accepted that the rating philosophy should influence the validation of rating systems, the challenges to validate TTC models have been largely ignored in the academic or practitioner literature. The BCBS (BIS 2005) further stresses that in order to evaluate the accuracy of PDs reported by banks supervisors there is a need to adapt their PD validation techniques to the specific types of banks' credit rating systems, in particular with respect to their PIT vs. TTC orientation. However, methods to validate rating systems have paid very little attention to the rating philosophy or focused on PIT models. For example, Cesaroni (2015) observes that predicted default rates are PIT, and thus the validation of a rating system "should" operate on PIT PDs from a theoretical perspective. In relation to this argument,

Petrov and Rubtsov (2016) explicitly mention that they have not yet developed a validation framework consistent with their PIT/TTC methodology.

3. Model methodology and conceptual framework

In this section, we outline our econometric technique and statistical PD modeling methodology. In principle, for classification tasks including default prediction, while one could use the same loss functions as those used for regression (i.e., the least squares criterion) in order to optimize the design of the classifier, this would not be the most reasonable way to approach such problems. This is because in classification the target variable is discrete in nature, hence alternative measures than employed in regression are more appropriate for quantifying the quality of model fit. This discussion could be motivated the classification problem for default prediction through *Bayesian decision theory*, which has the benefits of conceptual simplicity and alignment with common sense, as well as possesses a strong optimality flavor with respect to the probability of an error in classification. However, given that the focus and contribution of this paper does not lie in the domain of econometric technique, we will defer such discussion and focus on the *logistic regression modeling* (“LRM”) technique, as it is widely understood in the literature and applied by practitioners.

Considering the 2 class $\{\omega_i\}_{i=1}^2$ case for the LRM that is relevant to PD modeling, the first step is to express the *log-odds* (or the *logit function*) of the posterior probabilities as a linear function of the risk factors:

$$\ln \left(\frac{P(\omega_1|\mathbf{x})}{P(\omega_2|\mathbf{x})} \right) = \boldsymbol{\theta}^T \mathbf{x}, \quad (1)$$

Where $\mathbf{x} = (x_1, \dots, x_k) \in R^k$ is a k dimensional feature vector and $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k) \in R^k$ is a vector of coefficients and we define $x_1 = 1$ so that the intercept is subsumed into $\boldsymbol{\theta}$. In that $P(\omega_1|\mathbf{x}) + P(\omega_2|\mathbf{x}) = 1$:

$$P(\omega_1|\mathbf{x}) = \frac{1}{1 + \exp(-\boldsymbol{\theta}^T \mathbf{x})} = \sigma(-\boldsymbol{\theta}^T \mathbf{x}), \quad (2)$$

Where the function $\sigma(-\boldsymbol{\theta}^T \mathbf{x})$ is known as the *logistic sigmoid* (or *sigmoid link*) and has the mathematical properties of a cumulative distribution function that ranges between 0 and 1, with a domain on the real line. Intuitively, this can be viewed as the conditional PD of a score $\boldsymbol{\theta}^T \mathbf{x}$ where higher values indicate greater default risk.

We may estimate the parameter vector $\boldsymbol{\theta}$ by the method of *maximum likelihood estimation* (“MLE”) given a set of training samples, with observations of explanatory variables $\{\mathbf{x}_n\}_{n=1}^N$ and binary dependent variables $\{y_n\}_{n=1}^N$, where $y_n \in \{0,1\}$. The likelihood function is given by:

$$P(y_1, \dots, y_N | \boldsymbol{\theta}) = \prod_{n=1}^N (\sigma(-\boldsymbol{\theta}^T \mathbf{x}_n))^{y_n} (1 - \sigma(-\boldsymbol{\theta}^T \mathbf{x}_n))^{1-y_n}. \quad (3)$$

The practice is to consider the negative *log-likelihood function* (or the *cross-entropy error*), a monotonically increasing transformation of (3), for the purposes of computational convenience:

$$L(\boldsymbol{\theta}) = - \sum_{n=1}^N y_n \ln(\sigma(-\boldsymbol{\theta}^T \mathbf{x}_n)) + (1 - y_n) \ln(1 - \sigma(-\boldsymbol{\theta}^T \mathbf{x}_n)). \quad (4)$$

The expression in Eq4 is minimized with respect to θ using iterative methods such as steepest descent or Newton's scheme.

We note an important property of this model that is computationally convenient and leads to stable estimation under most circumstances. Since $\sigma(-\theta^T \mathbf{x}_n) \in (0,1)$ according to the properties of the sigmoid link function, it follows that the covariance matrix \mathbf{R} is positive definite, which implies that the Hessian matrix $\nabla^2 L(\theta)$ is positive definite. In turn this implies that the negative log-likelihood function $L(\theta)$ is *convex*, and as such this guarantees the existence of a unique minimum to this optimization. However, maximizing the likelihood function may be problematic in the case where the development dataset is *linearly separable*. In such a case, any point on the hyperplane $\hat{\theta}_{MLE}^T \mathbf{x} = 0$ (out of an infinite number of such hyperplanes) that solves the classification task and separates the training samples in each class does so *perfectly*, which means that every training point is assigned a posterior probability of class membership equal to one (or $\sigma(\hat{\theta}_{MLE}^T \mathbf{x}) = \frac{1}{2}$). In this case, the MLE procedure forces the parameter estimate to be infinite ($\hat{\theta}_{MLE}^T \rightarrow \infty$), which means geometrically that the sigmoid link function approaches a step function and not an s-curve as a function of the score. This basically is a case of overfitting the development sample, which can be controlled by techniques such as k-fold cross-validation, or including a regularization term inside a corresponding cost function that controls the magnitudes of the parameter estimates (e.g., LASSO techniques for a linear penalty function $C(\theta|\lambda) = \lambda|\theta|$ with a cost parameter λ).

We conclude this section by discussing the statistical assumptions underlying the LRM model. Logistic regression does not make many of the key assumptions of *ordinary least squares* ("OLS") regression regarding linearity, normality of error terms, homoscedasticity of the error variance and the measurement level. Firstly, LRM does not assume linearity relationship between the dependent variable and estimator¹, which implies that we can accommodate non-linear relationships between the independent and dependent variables without non-linear transformations of the former (although we may choose to do so for other reasons, such as treating outliers), which yields more parsimonious and more intuitive models. Another way to look at this is since we are applying the log-odds transformation to the posterior probabilities in (1), by construction we have a linear relationship in the risk drivers and do not require additional transformations. Secondly, the independent variables do not need to be multivariate normal, which equivalently means that the error terms need not be multivariate normal either. While there is an argument that if the error terms are actually multivariate normal (which is probably not true in practice), then imposing this assumption leads to efficiency gains and possibly a more stable solution, at the same time there are many more parameters to be estimated. That is because in the normal case we not only have to estimate the k regression coefficients $\theta = (\theta_1, \dots, \theta_k) \in R^k$, but we also have to estimate the entire covariance matrix (i.e., the covariance matrix in the LRM is a function of θ), which is $O\left(\frac{k^2}{2}\right)$ additional operations and could lead to a more unstable model depending upon data availability as well as more computational overhead. Thirdly, since the covariance matrix also depends on \mathbf{x} by construction through the sigmoid link function, variances need not be homoscedastic for each level of the independent variables

¹ Note that linearity *does not* mean that the dependent variable has a linear relationship with the explanatory variables (i.e., we can have non-linear transformations of the latter), but rather that the estimator is a linear function (or weighted average) of the dependent variable, which implies that we can obtain our estimator analytically using linear algebra operations as opposed to iterative techniques such as in the LRM.

(while if we imposed a normal assumption that we would require this assumption to hold as well). Lastly, the LRM can handle ordinal and nominal independent variables as they need not be metric (i.e., interval or ratio scaled), which leads to more flexibility in model construction and again avoids counterintuitive transformations and more parameters to be estimated.

However, some other assumptions still apply in the LRM setting. First, the LRM requires the dependent variable to be binary, while other approaches (e.g., *ordinal logistic regression* – “OLR” or the *multinomial regression model* – “MRM”) allow the dependent variable to be polytomous, which implies more granularity in modeling. This is because reducing an ordinal or even metric variable to a dichotomous level loses a lot of information, which makes this methodology inferior compared to OLR or MRM in these cases. In the case of PD modeling, if credit states other than default is relevant (e.g., significant downgrade short of default, or prepayment), then this could result in biased estimates and mismeasurement of default risk. However, we note in this regard that for many portfolios data limitations (especially for large corporate or commercial & industrial portfolios) prevent application of OLR for more states than default (e.g., prepayment events may not be identifiable in data), and conceptually we may argue that observations of ratings have elements of expert judgment and are not “true” events (although in wholesale credit, the definition of default is partly subjective). An assumption related to this is the *independence of irrelevant alternatives*, which states that relative odds of a binary outcome should not depend on other possible outcomes under consideration. In the statistics and econometrics literature, there is debate not only about how critical this assumption is, but also on ways to test this assumption and the value of such tests (Cheng and Long, 2006; Fry and Harris, 1996; Hausman and McFadden, 1984; Small and Hsiao, 1985).

Another important assumption is that the LRM requires the observations be independent, which means that the data-points should not be from any *dependent samples design* (e.g., matched pairings or panel data.) While obviously that is not the completely case in PD modeling in that we have dependent observations, in practice this may not be a very material violation, since if we are capturing most or all of the relevant factors influencing default, then anything else is likely to be idiosyncratic (especially if we are including macroeconomic factors).

While we are not in this implementation assuming a parametric distribution for the error terms in the LRM, there are still certain properties that the errors should exhibit, in order that we have some assurance that the model is not grossly misspecified (e.g., symmetry around zero, lack of outliers.) However, there is some debate in the literature on the criticality of this assumption, as well as the best way to evaluate LRM residuals (Li and Shepherd, 2012; Liu and Zhang, 2017).

Finally, we conclude this section by a discussion of the model methodology within the empirical context. The modeling approach as outlined in this section, and the model selection process as elaborated upon in subsequent sections, is common to both PIT and TTC constructs. However, we impose the constraint that only financial factors are considered in the TTC construct, while macroeconomic variables are additionally considered for the PIT models. This is in addition to the difference in default horizon and other model selection criteria, which results in a differentiation in the TTC and PIT outcomes, in terms of rating mobility and relative factor weights considered intuitive in each construct – i.e., high (lower) rating mobility, and greater (lower) weight on shorter (longer) term financial factors for the PIT (TTC) models.

4. Empirical analysis

4.1. Description of modeling data

The following data is used for the development of the models in this study:

- **CompustatTM** Standardized fundamental and market data for publicly-traded companies including financial statement line items and industry classifications (*Global Industry Classification Standards* - “GICS” and *North American Industry Classification System* - “NAICS”) over multiple economic cycles from 1979 and onward. This data includes default types such as bankruptcy, liquidation, and rating agency’s default rating, all of which are part of the industry standard default definitions.
- **Moody’s Default Risk ServiceTM (“DRS”) Rating History** An extensive database of rating migrations, default and recovery rates across geographies, regions, industries and sectors.
- **Bankruptcydata.com** A service provided by New Generation Research, Inc. providing information on corporate bankruptcies.
- **The Center for Research in Security PricesTM (“CRSP”) U.S. Stock Databases** This product is comprised of a database of historical daily and monthly market and corporate action data for over 32,000 active and inactive securities with primary listings on the NYSE, NYSE American, NASDAQ, NYSE Arca and Bats exchanges and includes CRSP broad market indexes.

A series of filters are applied to this Moody’s population to construct a population that is closely aligned with the North American large corporate segment of companies that are publicly rated and have publicly traded equity. In order to achieve this using Moody’s data, the following combination of NAICS and GICS industry codes, regional and historical yearly Net Sales restrictions are applied:

1. Non-C&I obligors defined by the following NAICS codes below (see Table 1 below), are not included in the population:
 - Financials
 - Commercial Real Estate and Real Estate Investment Trusts
 - Government
 - Dealer Finance
 - Not-for-Profit
2. A similar filter is performed according to GICS (see Table 2 below) classification:
 - Education
 - Financials
 - Real Estate
3. Only obligors based in the U.S. and Canada are included.
4. Only obligors with maximum historical yearly Net Sales at least \$1B are included.
5. There are exclusions for obligors with missing GICS or NAICS codes, and for modeling purposes obligors are categorized into different industry segments on this basis.
6. Records prior to 1Q91 are excluded, the rationale being that capital markets and accounting rules were different before the 1990’s, and the macroeconomic data used in the model development is only available beginning in 1990. As one-year change transformations are amongst those applied to the macroeconomic variables, this cutoff is advanced a year from 1990 to 1991.

Table 1. Large corporate modeling data - GICS industry segment composition for all Moody's obligors vs. defaulted Moody's obligors (1991–2015).

GICS Industry Segment	All Moody's Obligor	Defaulted Moody's Obligor
Consumer Discretionary	19.6%	30.9%
Consumer Staples	8.4%	6.4%
Energy	7.6%	5.9%
Healthcare Equipment & Services	2.9%	2.9%
Industrials	31.6%	15.1%
Materials	10.5%	11.3%
Pharmaceuticals & Biotechnology	2.7%	0.2%
Software & IT Services	2.5%	1.8%
Technology Hardware & Communications	4.3%	11.3%
Utilities	7.6%	5.6%

Table 2. Large corporate modeling data - NAICS industry segment composition for all Moody's obligors vs. defaulted Moody's obligors (1991–2015).

NAICS Industry Segment	All Moody's Obligor	Defaulted Moody's Obligor
Agriculture, Forestry, Hunting & Fishing	0.2%	0.4%
Accommodation & Food Services	2.3%	2.9%
Waste Management % Remediation Services	2.4%	2.1%
Arts, Entertainment & Recreation	0.7%	1.0%
Construction	1.7%	2.5%
Educational Services	0.1%	0.2%
Healthcare & Social Assistance	1.6%	1.6%
Information Services	11.5%	12.1%
Management Compensation Enterprises	0.1%	0.1%
Manufacturing	37.7%	34.4%
Mining, Oil & Gas	6.8%	8.6%
Other Services (ex Public Administration)	0.4%	0.6%
Professional, Scientific & Technological Services	2.3%	2.5%
Real Estate, Rentals & Leasing	0.9%	1.6%
Retail Trade	9.6%	12.4%
Transportation & Warehousing	5.4%	7.0%
Utilities	8.3%	5.4
Wholesale Trade	7.0%	2.7

- Records that are too close to a default event are not included in the development dataset, which is an industry standard approach, the rationale being that the records of an obligor in this time window do not provide information about future defaults of the obligor, but more likely the obligor's existing problems. Furthermore, a more effective practice is to base this on data that are 6–18 (rather than 1–12) months prior to default date, as this typically reflects the range of timing between when statements are issued and when ratings are updated (i.e., usually it takes

up to six months, depending on time to complete financials, receive them, input, and complete / finalize the ratings).

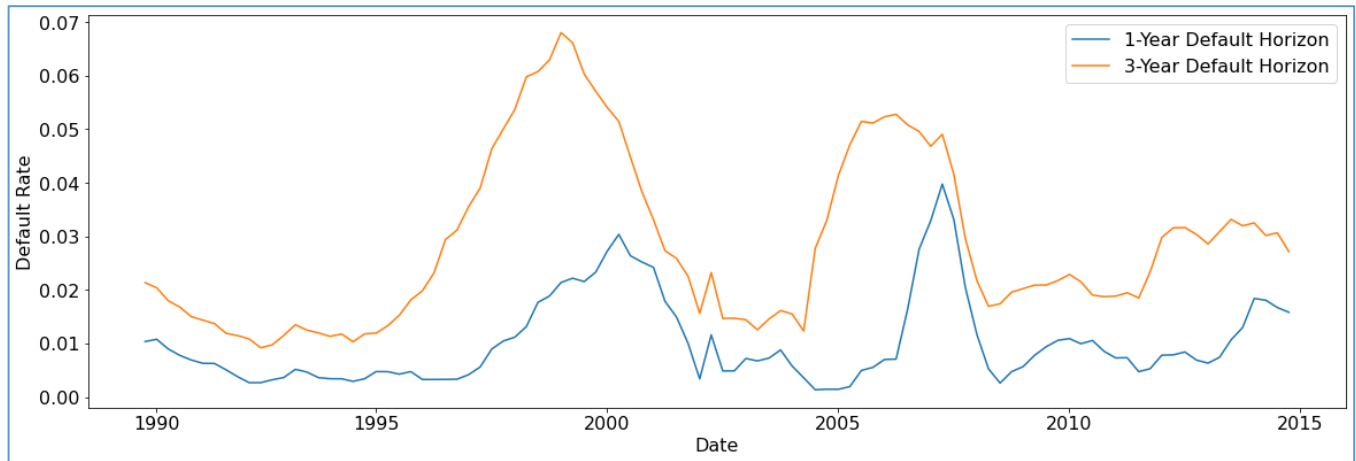


Figure 1. PD model large corporate modeling data – Moody’s obligors one and three year horizon default rates over time (1991–2015).

8. In general, the defaulted obligors’ financial statements after default date are not included in the modeling dataset. However, in some cases obligors may exit a default state or “cure” (e.g., emerge from bankruptcy), in which cases only the statements between default date and cured date are not included.

In our opinion, these data exclusions are reasonable and in line with industry standards, sufficiently documented and supported and do not compromise the integrity of the modeling dataset.

The time periods considered for the Moody’s data is the development period 1Q91–4Q15. Shown in Table 1 above is the comparison of the modeling population by GICS industry sectors, where for each sector defaulted obligors columns represent the percent of defaulted obligors in the sector out of entire population. The data are concentrated in Consumer Discretionary (20%), Industrials (17%), Tech Hardware and Communications (12%), and Energy except E&P (11%). A similar industry composition is shown below in Table 2 according to the NAICS classification system.

The model development dataset contains financial ratios and default information that are based upon the most recent data available from DRSTM, CompustatTM and bankruptcydata.com, so that the data is timely and a priori should be give the benefit of the doubt with respect to favorable quality. Furthermore, the model development time period of 1Q91-4Q15 spans two economic downturn periods and a complete business cycle, the length of which being another factor supporting a verdict of good quality. Related to this point, we plot the yearly one and three year default rates in the model development dataset, as shown above in Figure 1. As the goal of model development is to establish for each risk driver that the preliminary trends observed match that of our expectations, there is sufficient variation in this data to support quantitative methods of parameter estimation, further supporting the suitability of the data from a quality perspective.

In the subsequent tables we present the summary statistics for the variables that appear in our final models. These final models were chosen based upon an exhaustive search algorithm in conjunction with 5-fold cross-validation, and we have chosen the leading two models in either the PIT and TTC designs, as

well as incorporating the DTD risk factor or not². The counts and statistics vary slightly across models, as the Python libraries that we utilize do not accommodate missing values, but nonetheless the differences in these statistics across models are minimal. The counts of observations vary narrowly from about 150K to 165K. The default rate is consistently about 1% (3%) for the PIT (TTC) models. The following are the categories and names of the explanatory variables appearing in the final candidate models³:

- **Leverage** Total Liabilities to Total Assets Ratio (“TLTAR”)
- **Size** Change in Total Assets (“CTA”), Total Liabilities (“TL”)
- **Leverage** Total Liabilities to Total Assets Ratio (“TLTAR”)
- **Coverage** Cash Use Ratio (“CUR”), Debt Service Coverage Ratio (“DSCR”)
- **Efficiency** Net Accounts Receivables Days Ratio (“NARDR”)
- **Liquidity** Net Quick Ratio (“NQR”), Net Working Capital to Tangible Assets Ratio (“NWCTAR”)
- **Profitability** Before Tax Profit Margin (“BTPM”)
- **Macroeconomic** Moody’s 500 Equity Price Index Quarterly Average Annual Change (“SP500EPIQAAC”), Consumer Confidence Index Annual Change (“CCIAC”)
- **Merton Structural** Distance-to-Default (“DTD”)

The *Area under the Receiver Operating Characteristic Curve* (“AUC”) statistics and missing rates for the explanatory variables are summarized in Table 7 at the end of this section⁴. The univariate AUCs range in 0.6–0.8 across risk factors, with some expected deterioration when going from the 1- to 3-year default horizon, which is indicative of strong default rank ordering capability amongst these explanatory variables. The missing rates are generally between 5 and 10%, which is indicative of favorable data quality to support model development.

Table 3. Summary statistics – Moody’s large corporate financial and macroeconomic explanatory variables and default indicators: 1-year PIT model 1.

Variable	Count	Mean	Standard Deviation	Minimum	25 th Percentile	Median	75 th Percentile	Maximum
Default Indicator	157,353	0.01	0.10	0.00	0.00	0.00	0.00	1.00
Change in Total Assets		0.14	0.35	−0.40	−0.01	0.06	0.17	3.21

Continued on next page

² Clarifying our model selection criteria and process, we balance multiple criteria, both in terms of statistical performance as well as some qualitative considerations. Firstly, all models have to exhibit the stability of factor selection (where the signs on coefficient estimates are constrained to be economically intuitive) and statistical significance in k-fold cross validation subsample estimation. However, this is constrained by the requirement that we have only a single financial factor chosen from each category. Then the models that meet these criteria are evaluated according to statistical performance metrics such as AIC and AUC, as well as other considerations such as rating mobility and relative factor weights.

³ All candidate explanatory variables are Winsorized at either the 10th, 5th or 1st percentile levels at either tails of the sample distribution, in order to mitigate the influence of outliers or contamination in data, according to a customized algorithm that analyzes the gaps between these percentiles and caps / floors where these are maximal.

⁴ The plots are omitted for the sake of brevity and are available upon request.

Variable	Count	Mean	Standard Deviation	Minimum	25 th Percentile	Median	75 th Percentile	Maximum
Total Liabilities to Total Assets		0.60	0.23	0.12	0.45	0.59	0.71	1.53
Cash Use Ratio		1.90	2.84	-22.43	1.41	2.06	2.65	19.00
Net Accounts Receivables Days		130.25	101.44	11.26	68.98	106.74	159.43	754.09
Net Quick Ratio		0.34	1.07	-0.85	-0.28	0.06	0.59	6.11
Before Tax Profit Margin		5.94	21.00	-146.67	1.85	7.09	12.85	48.70
Moody's Equity Price Index		1.91	6.09	-27.33	-0.19	2.19	5.68	12.81
Consumer Confidence Index		2.34	21.58	-60.97	-7.02	4.89	15.35	73.21

Table 4. Summary statistics – Moody's large corporate financial, macroeconomic, Merton / structural model distance-to-default proxy measure explanatory variables and default indicators: 1-year PIT model 2.

Variable	Count	Mean	Standard Deviation	Minimum	25 th Percentile	Median	75 th Percentile	Maximum
Default Indicator	160,002	0.01	0.10	0.00	0.00	0.00	0.00	1.00
Change in Total Assets		0.14	0.35	-0.40	-0.01	0.06	0.17	3.21
Total Liabilities to Total Assets		0.60	0.23	0.12	0.45	0.60	0.71	1.53
Cash Use Ratio		1.90	2.83	-22.43	1.40	2.06	2.64	19.00
Net Quick Ratio		0.34	1.06	-0.85	-0.28	0.06	0.59	6.11
Before Tax Profit Margin		5.98	20.93	-146.67	1.86	7.10	12.88	48.70
Moody's Equity Price Index		1.93	6.08	-27.33	-0.19	2.19	5.68	12.81
Consumer Confidence Index		2.37	21.56	-60.97	-7.02	4.89	15.35	73.21
Distance-to-Default		0.20	0.43	-1.32	0.02	0.07	0.18	5.26

Table 5. Summary statistics – Moody’s large corporate financial and explanatory variables and default indicators: 3-year TTC model 1.

Variable	Count	Mean	Standard Deviation	Minimum	25 th Percentile	Median	75 th Percentile	Maximum
Default Indicator	150,064	0.03	0.17	0.0	0.0	0.0	0.0	1.0
Total Liabilities		3,640.65	6,741.93	8.86	422.60	1,170.45	3,374.12	41,852.00
Total Liabilities to Total Assets		0.62	0.22	0.12	0.49	0.61	0.72	1.53
Debt Service Ratio		16.44	52.82	-25.07	1.74	4.09	9.80	409.64
Net Quick Ratio		0.24	0.93	-0.85	-0.30	0.02	0.47	6.11
Before Tax Profit Margin		5.50	21.08	-146.67	1.57	6.72	12.40	48.70

Table 6. Summary statistics – Moody’s large corporate financial and Merton / structural model distance-to-default proxy measure explanatory variables and default indicators: 3-year TTC model 2.

Variable	Count	Mean	Standard Deviation	Minimum	25 th Percentile	Median	75 th Percentile	Maximum
Default Indicator	150,064	0.03	0.17	0.0	0.0	0.0	0.0	1.0
Total Liabilities		3,640.65	6,741.93	8.86	422.60	1,170.45	3,374.12	41,852.00
Total Liabilities to Total Assets		0.62	0.22	0.12	0.49	0.61	0.72	1.53
Debt Service Ratio		16.44	52.82	-25.07	1.74	4.09	9.80	409.64
Net Quick Ratio		0.24	0.93	-0.85	-0.30	0.02	0.47	6.11
Before Tax Profit Margin		5.50	21.08	-146.67	1.57	6.72	12.40	48.70
Distance-to-Default		0.20	0.42	-1.32	0.02	0.07	0.28	5.26

Table 7. Moody’s large corporate financial and macroeconomic explanatory variables areas under the receiver operating characteristic curve (AUC) and missing rates for 1-year default horizon PIT and 3-year default horizon TTC default indicators.

Category	Explanatory Variables	PIT 1-Year Default Horizon		TTC 3-Year Default Horizon	
		AUC	Missing Rate	AUC	Missing Rate
Size	Change in Total Assets Total Liabilities	0.726	8.52%	0.582	4.64%
Leverage	Total Liabilities to Total Assets Ratio	0.843	4.65%	0.783	4.65%
Coverage	Cash Use Ratio Debt Service Coverage Ratio	0.788	7.94%	0.796	17.0%
Efficiency	Net Accounts Receivables Days Ratio	0.615	8.17%		
Liquidity	Net Quick Ratio	0.653	7.71%	0.617	7.17%
Profitability	Before Tax Profit Margin	0.827	2.40%	0.768	2.40%
Macroeconomic	Moody’s 500 Equity Price Index Quarterly Average Annual Change	0.603	0.00%		
	Consumer Confidence Index Annual Change	0.607	0.00%		
Merton Structural	Distance-to-Default	0.730	4.65%	0.669	4.65%

4.2 Econometric specifications and model validation

In the subsequent tables we present the estimation results and in-sample performance statistics for our final models.

We shall first discuss general features of the model estimation results. Across models, signs of coefficient estimates are in line with economic intuition, and significance levels are indicative of very precisely estimated parameters. AUC statistics indicate that the models have a strong ability to rank order default risk, and while as expected this level of discriminatory power declines somewhat at the longer default horizon, in all cases the levels are in line with favorable performance by industry standards.

Regarding measures of predictive accuracy the *Hosmer-Lemeshow* (“HL”) tests show that the PIT models fit the data well while the TTC models fail to do so. However, we observe that when we introduce DTD into the TTC models, predictive accuracy increases markedly, as the p-values of the HL statistics increase significantly to the point where there is marginal evidence of adequate fit (i.e., the p-values indicate that the TTC models fail at only significance levels greater than 5%). AIC measures are also much higher in the TTC vs. the PIT models, but do decline materially when the DTD risk factors are introduced, consistent with the HL statistics.

We next discuss general features of the estimation that speak to the TTC or PIT qualities of the models. As expected, the TTC models have much lower *Singular Value Decomposition* (“SVD”) rating mobility metrics as compared to the PIT models, in the range of about 30-35% in the former as compared about a 70-80% neighborhood in the latter. The relative magnitude of the *factor contribution* (“FC”) measures, which quantify the proportion of the total score that is accounted for by an explanatory variable, also support that the models are exhibiting TTC and PIT characteristics. This is because intuitively, we observe that in the TTC models there are greater weights on categories considered more important in credit underwriting (i.e., Size, Leverage and Coverage), whereas in the PIT models this trend is reversed and there is greater emphasis on factors considered more critical to early warning or credit portfolio management (i.e., Liquidity, Profitability and Efficiency).

In Table 8 below we show the estimation results and in-sample performance measures for PIT Model 1 having both financial and macroeconomic explanatory variables for a 1-year default horizon. FCs are higher on more PIT relevant factors as contrasted to factors considered more salient to TTC constructs. Financial risk factors carry a super-majority of the FC than do the macroeconomic factors, about 90% in the former as compared to about 10% in the latter, which is a common observation in the industry for PD scorecard models. The model estimation results provide evidence of high discriminatory power, as the AUC is 0.8894. The AIC is 7,231.9, which relative to the TTC models is indicative of favorable predictive accuracy, which is corroborated by the very high the HL p-value of 0.5945. Finally, the SVD mobility metric of 0.7184 supports that this model exhibits PD rating volatility consistent with a PIT model.

Table 8. Logistic regression estimation results – Moody’s large corporate financial and macroeconomic explanatory variables 1-year default horizon PIT reduced form model 1.

Explanatory Variable	Parameter Estimate	P-Value	Factor Weight	AIC	AUC	HL Value	P-	Mobility Index
Change in Total Assets	-0.4837	0.0000	0.0455					
Total Liabilities to Total Assets	2.6170	0.0104	0.1091					
Cash Use Ratio	-0.0428	0.0000	0.1545					
Net Accounts Receivables Days Ratio	0.0005	0.0000	0.2273					
Net Quick Ratio	-0.4673	0.0000	0.0909					
Before Tax Profit Margin	-0.0161	0.0000	0.2736					
Moody’s Equity Index Price Index Quarterly Average	-0.0189	0.0000	0.0759					
Consumer Confidence Index Year-on-Year Change	-0.0099	0.0000	0.0232	7,231.00	0.8894	0.5945		0.7184

Table 9. Logistic regression estimation results – Moody’s large corporate financial, macroeconomic and distance-to-default explanatory variables 1-year default horizon PIT hybrid reduced form / structural-Merton model 2.

Explanatory Variable	Parameter Estimate	P-Value	Factor Weight	AIC	AUC	HL Value	P-	Mobility Index
Change in Total Assets	-0.4664	0.0000	0.0485					
Total Liabilities to Total Assets	2.5385	0.0000	0.1165					
Cash Use Ratio	-0.0428	0.0000	0.1650					
Net Quick Ratio	-0.0169	0.0000	0.0971					
Before Tax Profit Margin	-0.0169	0.0000	0.2913					
Moody’s Equity Index Price Index Quarterly Average	-0.0186	0.0000	0.0801					
Consumer Confidence Index Year-on-Year Change	-0.0100	0.0000	0.0267					
Distance to Default	-0.1913	0.0052	0.1748	7,290.00	0.8895	0.5782		0.7617

In Table 9 below we show the estimation results and in-sample performance measures for PIT Model 2 having the financial, macroeconomic and structural-Merton DTD explanatory variables for a 1-year default horizon. Results are similar to PIT Model 1 in terms of signs of coefficient estimates, statistical significance and relative FCs of financial and macroeconomic variables. DTD enters the model without any deleterious effect on the statistical significance of the other variables, although the relative

contribution of 0.17 absorbs a fair amount of the other variables' FCs and eclipses that of the macroeconomic variables. That said, we observe that collectively, financial and Merton DTD risk factors carry a super-majority of the FC than do the macroeconomic factors, about 89% in the former as compared to about 11% in the latter, which is a common observation in the industry for PD scorecard models. The model estimation results provide evidence of high discriminatory power, as the AUC is 0.8895, which is immaterially lower than then the Model 1 version not having DTD. The AIC is 7,290.0, which relative to the TTC models is indicative of favorable predictive accuracy and also indicates an improvement in fit as compared to the Model 1 version not having the structural model DTD variable, which is corroborated by the very high the HL p-value of 0.5782. Finally, the SVD mobility metric of 0.7616 supports that this model exhibits PD rating volatility consistent with a PIT model, and moreover the addition of the DTD variable improves the PIT aspect of this model relative to its Model 1 counterpart not having this feature.

Table 10. Logistic regression estimation results – Moody's large corporate financial and macroeconomic explanatory variables 3-year default horizon TTC reduced form model 1.

Explanatory Variable	Parameter Estimate	P-Value	Factor Weight	AIC	AUC	HL Value	P-Value	Mobility Index
Value of Total Liabilities	-6.97E-06	0.0000	0.1773					
Total Liabilities to Total Assets	2.0239	0.0030	0.3133					
Debt Service Coverage Ratio	-0.0431	0.0000	0.2332					
Net Quick Ratio	-0.2412	0.0000	0.1372					
Before Tax Profit Margin	-0.0129	0.0000	0.1390	17,751.00	0.8232	0.0039		0.3295

In Table 10 above we show the estimation results and in-sample performance measures for TTC Model 1 having financial explanatory variables for a 3-year default horizon. The signs of coefficient estimates are all intuitive and are all highly statistically significant. FCs are higher on more TTC relevant factors as contrasted to the factors considered more salient to PIT constructs. The model estimation results provide evidence of high discriminatory power, as the AUC is 0.8232, but which as expected is somewhat lower than in the comparable PIT models not containing DTD where they are in the range of 0.88-0.89. The AIC is 17,751.6, which relative to the comparable PIT models is indicative of rather worse predictive power, which is corroborated by the very low the HL P-Value of 0.0039, which rejects the null hypothesis that the model is properly specified with respect to a "saturated model" that perfectly fits the data. Finally, the SVD mobility metric of 0.3295 supports that this model exhibits PD rating volatility consistent with a TTC model.

In Table 11 below we show the estimation results and in-sample performance measures for TTC Model 2 having the financial and structural-Merton DTD explanatory variables for a 3-year default horizon. The signs of coefficient estimates are all intuitive and are all highly statistically significant. FCs are higher on more TTC relevant factors as contrasted to the factors considered more salient to PIT constructs. Note that in this model, adding the DTD explanatory variable results in TL not being statistically significant, and we drop it from this specification; also, the FC of DTD is 0.17, so that the financial factors still carry most of the relative weight. The model estimation results provide evidence of high discriminatory power, as the AUC is 0.8226, but which as expected is somewhat lower than in the comparable PIT models containing DTD where they are in the range of 0.88-0.89. The AIC is 11,834.6,

which relative to the comparable PIT models containing DTD is indicative of rather worse predictive power, which is corroborated by the somewhat low the HL P-Value of 0.0973, which rejects the null hypothesis that the model is properly specified with respect to a “saturated model” that perfectly fits the data at the 5% significance level, where we would note that this marginal rejection is an improvement over the comparable TTC version of this model not having the DTD variable. Finally, the SVD mobility metric of 0.3539 supports that this model exhibits PD rating volatility consistent with a TTC model, but we note that the rating volatility measure is somewhat higher than in the comparable TTC model not containing the DTD variable.

Table 11. Logistic regression estimation results – Moody’s large corporate financial, macroeconomic and distance-to-default explanatory variables 3-year default horizon TTC hybrid reduced form / structural-Merton model 2.

Explanatory Variable	Parameter Estimate	P-Value	Factor Weight	AIC	AUC	HL P-Value	Deviance / Degrees of Freedom	Pseudo R-Squared	Mobility Index
Total Liabilities to Total Assets	2.9580	0.0000	0.3707	11,834.00	0.8226	0.0973	0.2365	0.1491	0.3539
Debt Service Coverage Ratio	-0.0428	0.0000	0.2917						
Net Quick Ratio	-0.2403	0.0000	0.0808						
Before Tax Profit Margin	-0.0129	0.0000	0.0902						
Distance to Default	-0.1541	0.0000	0.1666						

In the subsequent figures we present additional in-sample and out-of-sample performance statistics and diagnostic plots for our final models. We observe that in-sample that these optical diagnostics (time series and calibration plots; fit histograms) and additional GOF statistics (e.g., Binomial and Jeffrey’s P-Values; OLS R-Squared) confirm the previously discussed results, in that the PIT models are showing much better fit to the data than the TTC models. However, the improvement in predictive accuracy in the TTC from including the DTD risk factor is not so evident from these measures, as it was from the HL statistics previously discussed. Furthermore, other optical residual diagnostic measures (e.g., Residual vs. Fitted Values, Quantile-Quantile Plot, Residual Histograms and Leverage Plots) show that both PIT and TTC models are almost equally showing issues in predictive accuracy or model specification. Finally, the out-of-sample analysis shows that the PIT models do not perform well, while the TTC model perform much better, and that the latter outperformance is augmented when including the structural DTD explanatory variable.

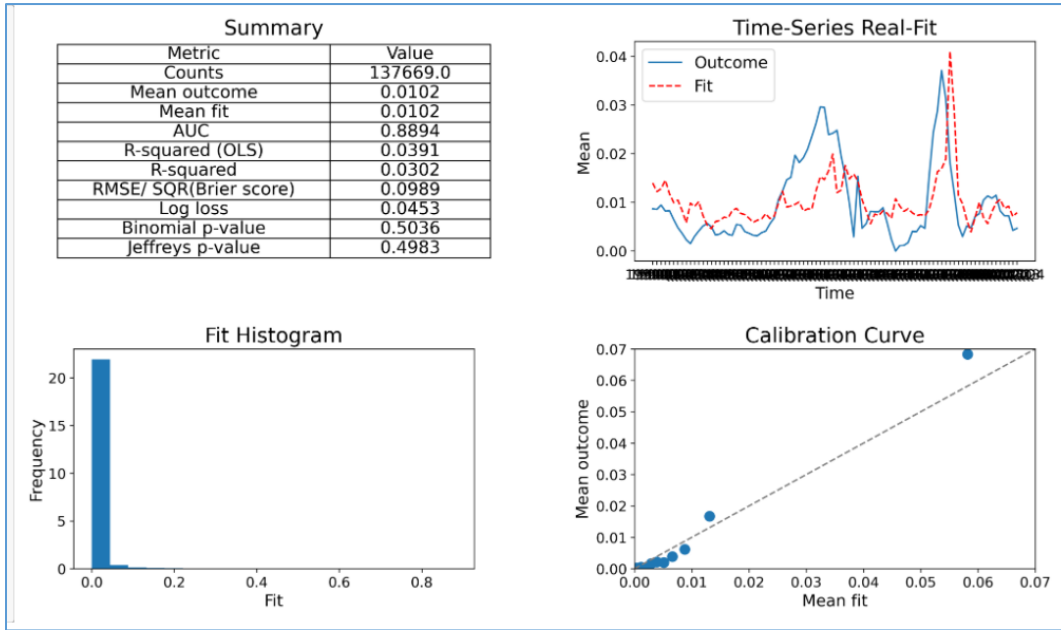


Figure 2. Logistic regression estimation in-sample performance measures, time series accuracy plot, fit histogram and calibration curve – Moody’s large corporate financial and macroeconomic explanatory variables 1-year default horizon PIT reduced form model 1.

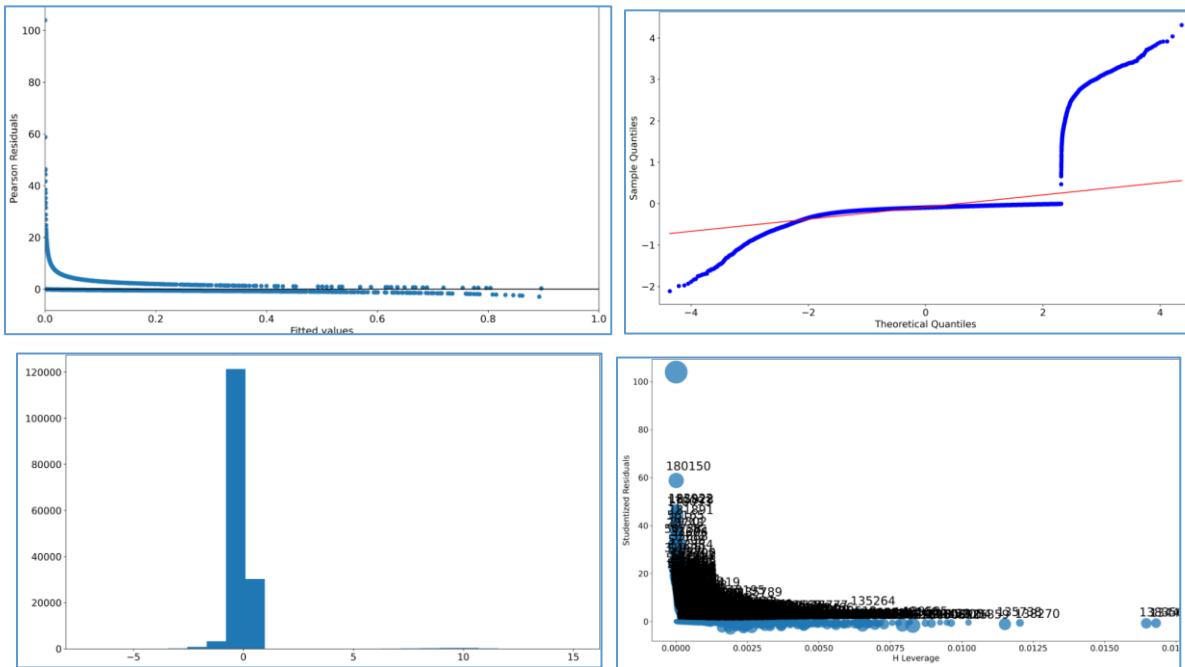


Figure 3. Logistic regression estimation in-sample residual diagnostics (residual vs. fitted values, quantile-quantile plot, residual histogram and leverage plots) – Moody’s large corporate financial and macroeconomic explanatory variables 1-year default horizon PIT reduced form model 1.

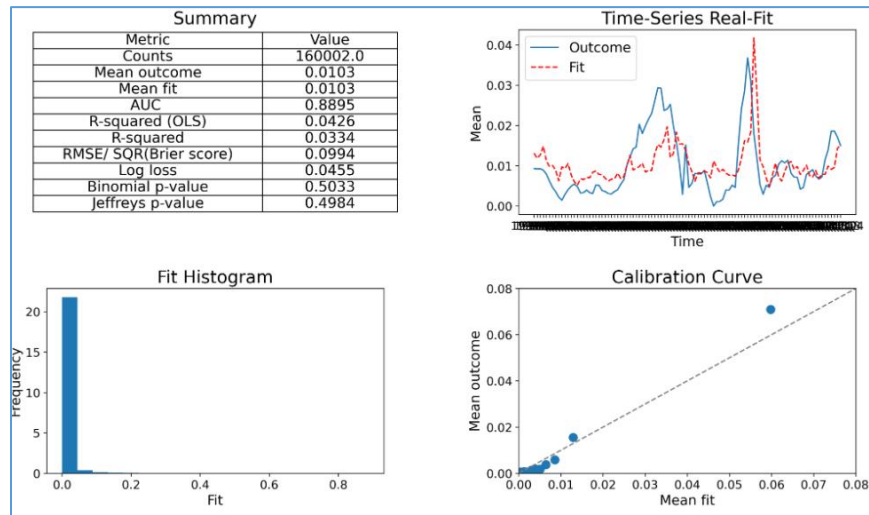


Figure 4. Logistic regression estimation in-sample performance measures, time series accuracy plot, fit histogram and calibration curve – Moody’s large corporate financial, macroeconomic and distance-to-default explanatory variables 1-year default horizon PIT hybrid reduced form / structural-Merton model 2.

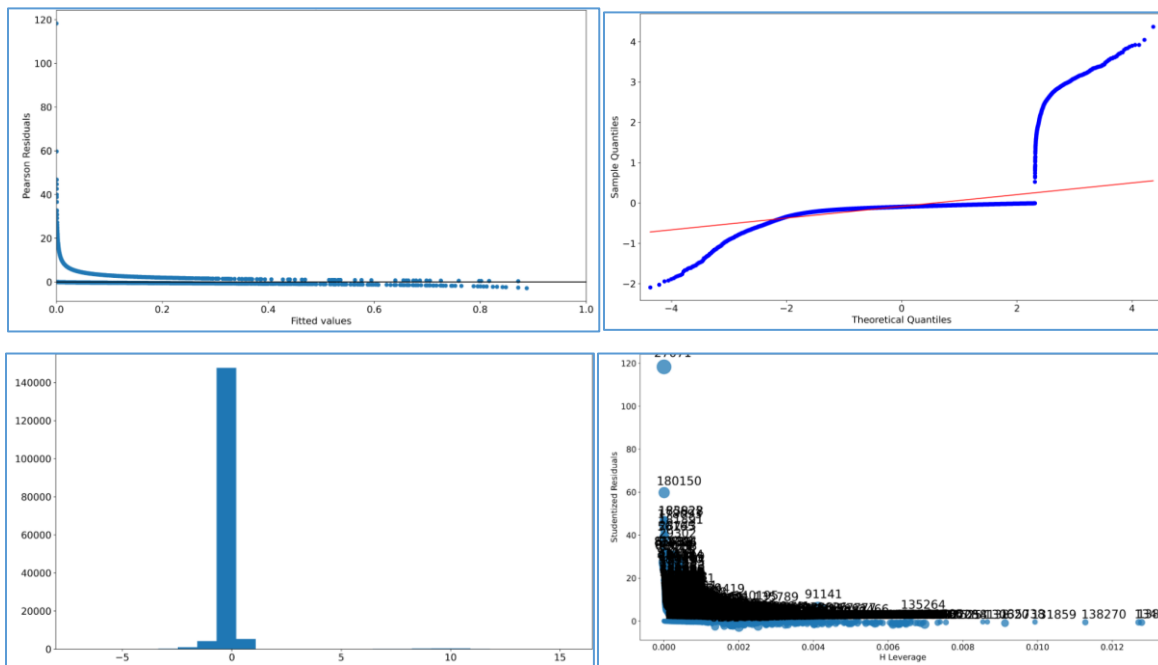


Figure 5. Logistic regression estimation in-sample residual diagnostics (residual vs. fitted values, quantile-quantile plot, residual histogram and leverage plots) – Moody’s large corporate financial, macroeconomic and distance-to-default explanatory variables 1-year default horizon PIT hybrid reduced form / structural-Merton model 2.

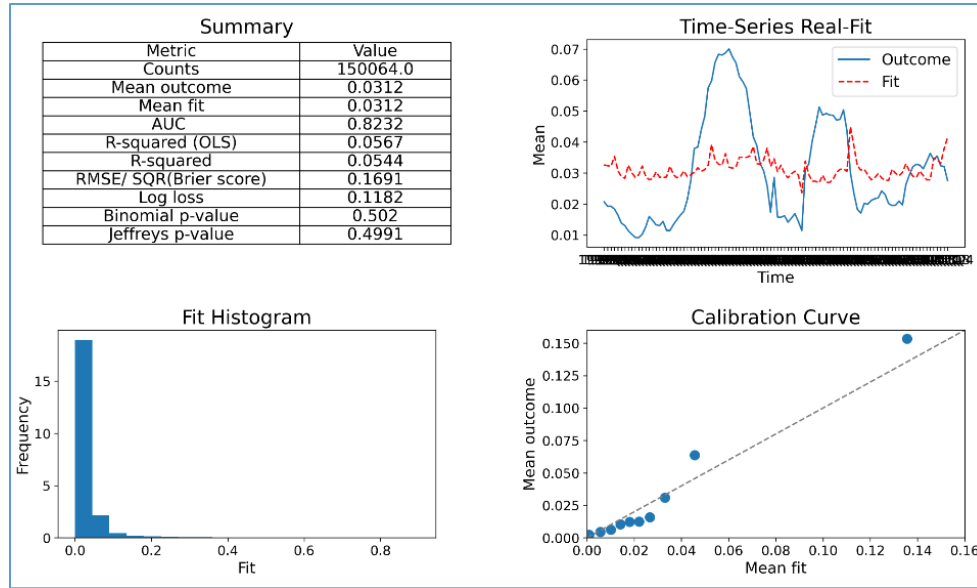


Figure 6. Logistic regression estimation in-sample performance measures, time series accuracy plot, fit histogram and calibration curve – Moody’s large corporate financial explanatory variables 3-year default horizon TTC reduced form model 1.

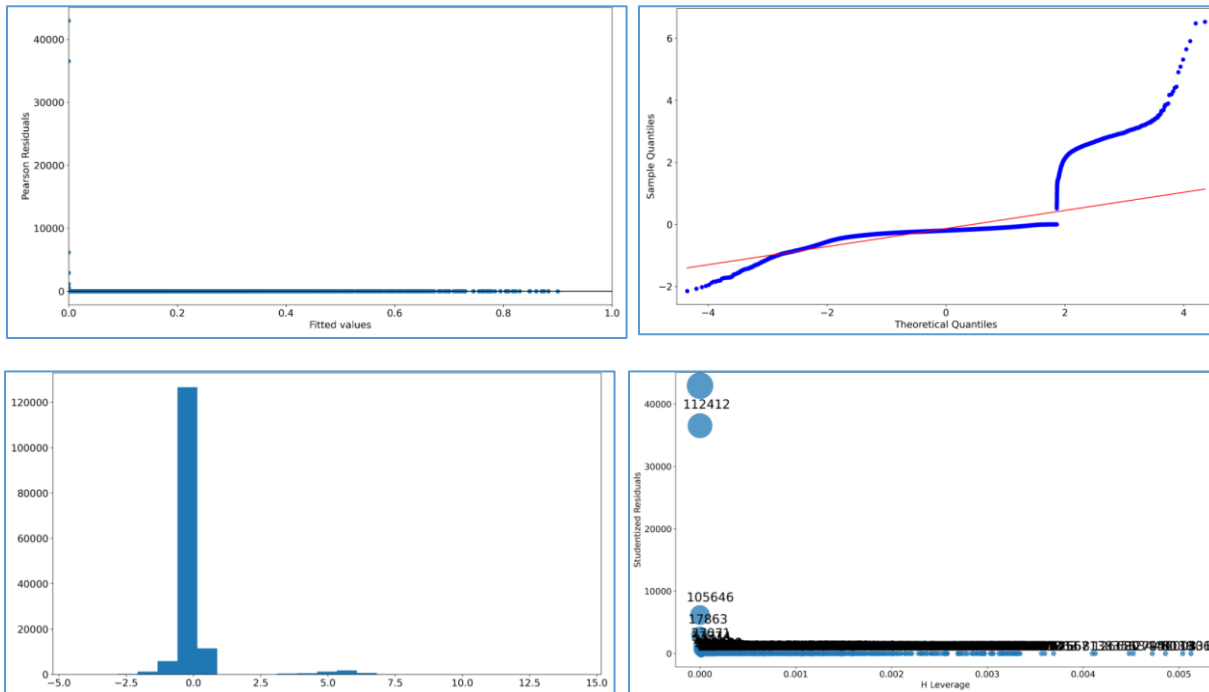


Figure 7. Logistic regression estimation in-sample residual diagnostics (residual vs. fitted values, quantile-quantile plot, residual histogram and leverage plots) – Moody’s large corporate financial explanatory variables 3-year default horizon TTC reduced form model 1.

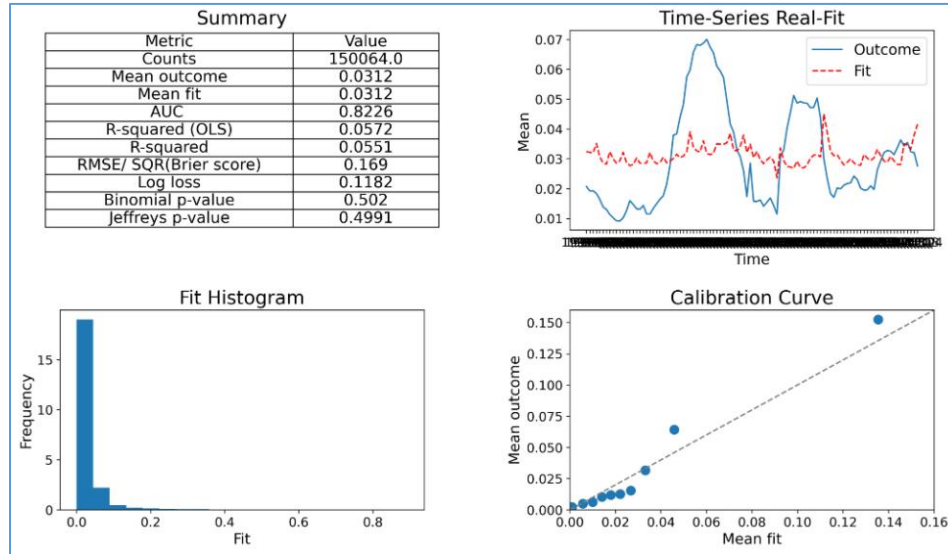


Figure 8. Logistic regression estimation in-sample performance measures, time series accuracy plot, fit histogram and calibration curve – Moody’s large corporate financial and distance-to-default explanatory variables 3-year default horizon TTC hybrid reduced form / structural-Merton model 2.

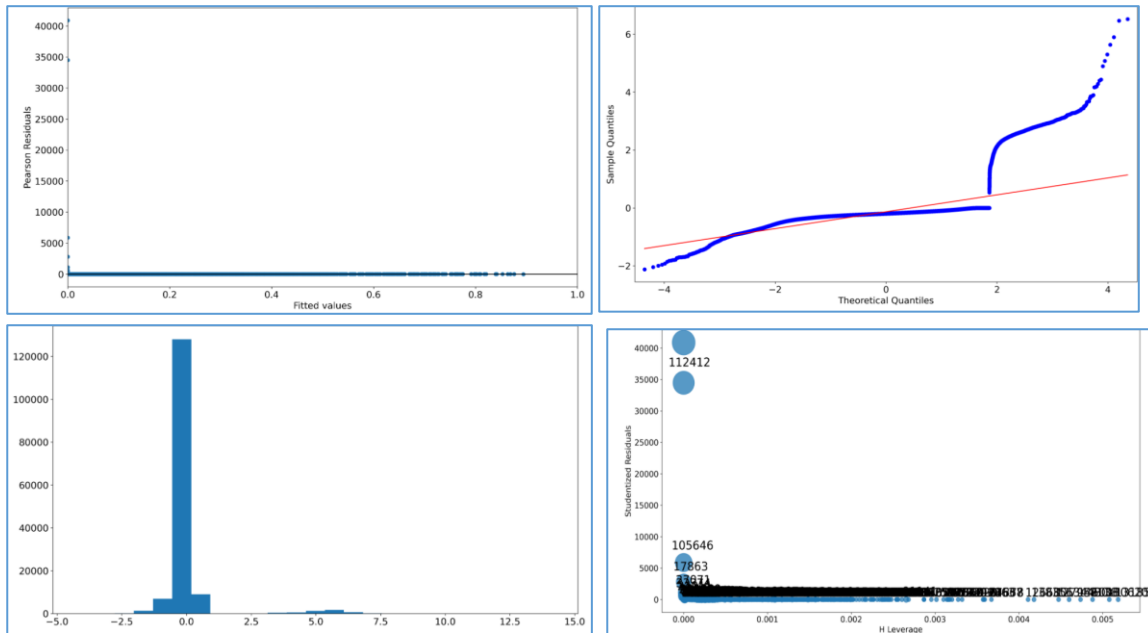


Figure 9. Logistic regression estimation in-sample residual diagnostics (residual vs. fitted values, quantile-quantile plot, residual histogram and leverage plots) – Moody’s large corporate financial and distance-to-default explanatory variables 3-year default horizon TTC hybrid reduced form / structural-Merton model 2.

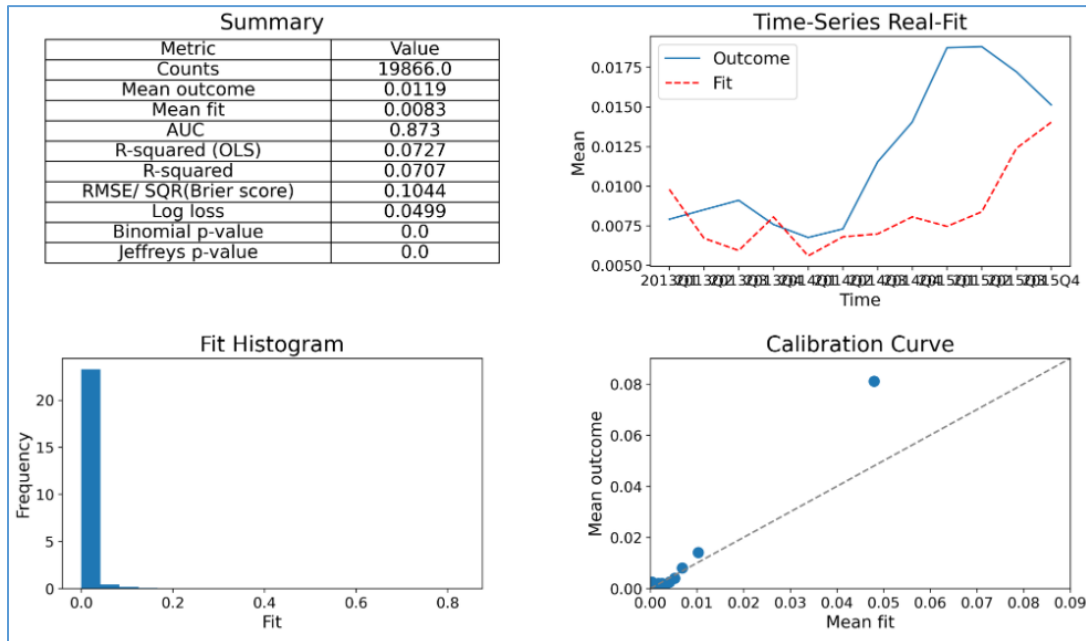


Figure 10. Logistic regression estimation out-of-sample performance measures, time series accuracy plot, fit histogram and calibration curve – Moody's large corporate financial and macroeconomic explanatory variables 1-year default horizon PIT reduced form model 1.

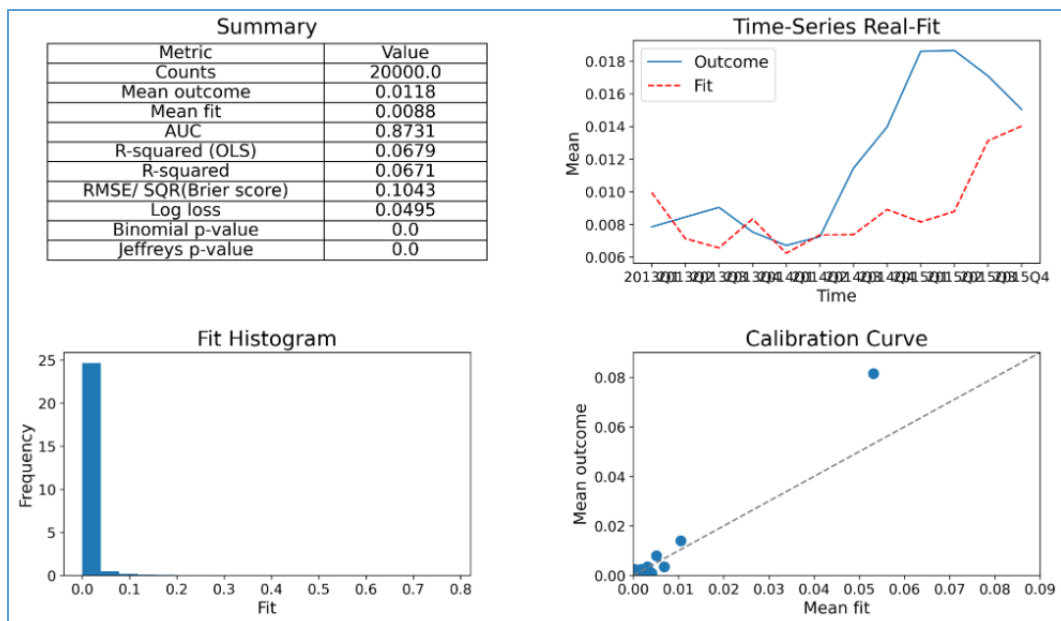


Figure 11. Logistic regression estimation out-of-sample performance measures, time series accuracy plot, fit histogram and calibration curve – Moody's large corporate financial, macroeconomic and distance-to-default explanatory variables 1-year default horizon PIT hybrid reduced form / structural-Merton model 2.

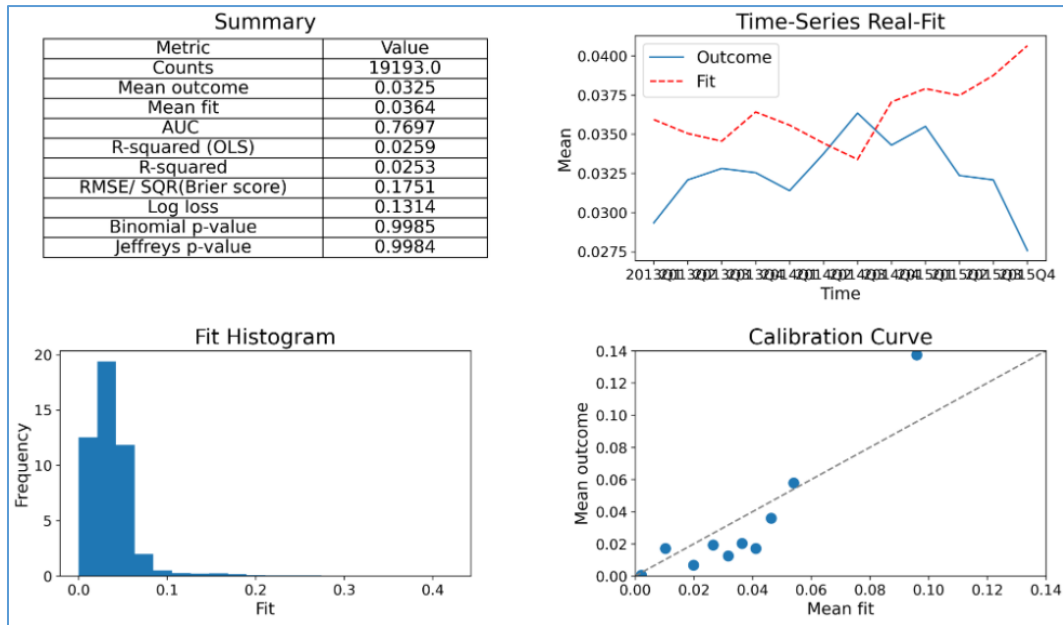


Figure 12. Logistic regression estimation out-of-sample performance measures, time series accuracy plot, fit histogram and calibration curve – Moody's large corporate financial explanatory variables 3-year default horizon TTC reduced form model 1.

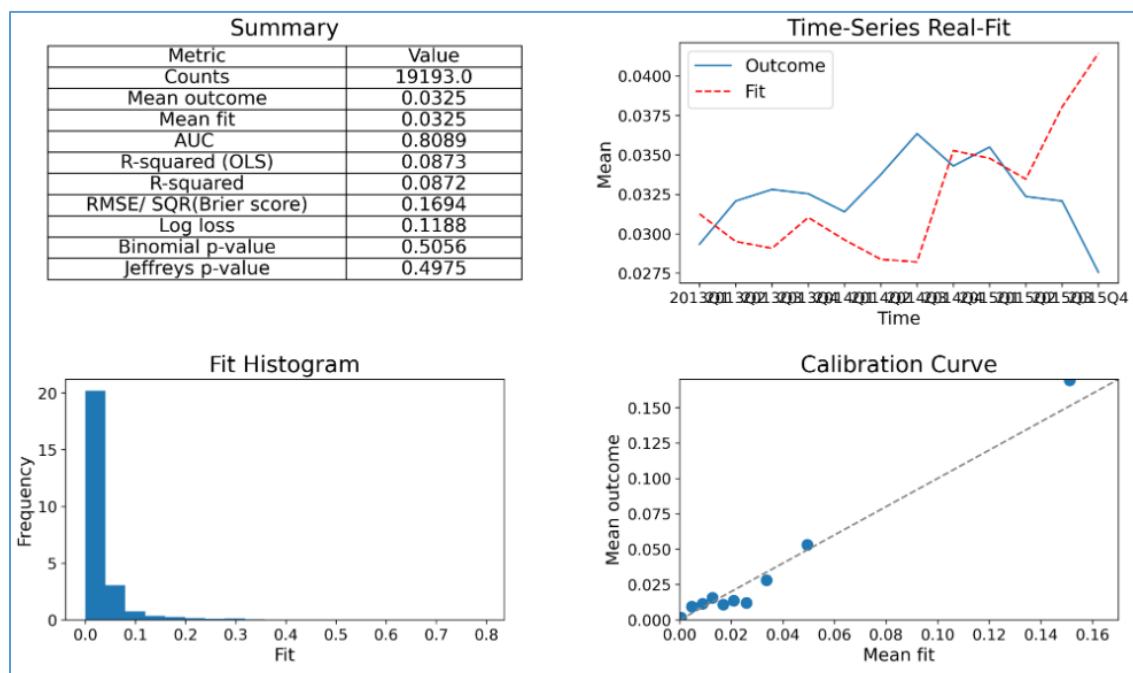


Figure 13. Logistic regression estimation out-of-sample performance measures, time series accuracy plot, fit histogram and calibration curve – Moody's large corporate financial and distance-to-default explanatory variables 3-year default horizon TTC hybrid reduced form / structural-Merton model 2.

4.3 The quantification of model risk according to the principle of relative entropy

In building of risk models we are subject to errors from model risk, one source of which being the violation of modeling assumptions. In this section we apply a methodology for the quantification of model risk that is a tool in building models robust to such errors. A key objective of model risk management is to assess the likelihood, exposure and severity of model error in that all models rely upon simplifying assumptions. It follows that a critical component of an effective model risk framework is the development of bounds upon model error resulting from the violation of modeling assumptions. This measurement is based upon a reference a nominal risk model and is capable of rank ordering the various model risks as well as indicating which perturbation of the model has a maximal effect upon some risk measure.

In line with the objective of managing model risk in the context of obligor-level PD modeling, we calculate confidence bounds around forecasted PDs spanning model errors in the vicinity of a nominal or reference model defined by a set of alternative models. These bounds can be likened to confidence intervals that quantify sampling error in parameter estimation. However, these bounds are a measure of model robustness that instead measures model error due to the violation of modeling assumptions. In contrast, a standard error estimate conventionally employed in credit risk modeling does not achieve this objective, as this construct relies on an assumed joint distribution of the asset returns or correlation in defaults.

We meet our objective referenced previously in the context of PD modeling through bounding a measure of loss, in this case the AIC, which can reflect a level of model error within reason. We have observed that while amongst practitioners one alternative means of measuring model risk is to consider challenger models, an assessment of estimation error or sensitivity in perturbing parameters is in fact a more prevalent means of accomplishing this objective, which captures only a very narrow dimension of model risk. In contrast, our methodology transcends the latter aspect to quantify potential model errors such as incorrect specification of the probability law governing the model (e.g., the distribution of error terms; or the specification of a link function in generalized linear regression, of which logistic regression is a sub-class), variables belonging in the model (e.g., omitted variable bias) or the functional form of the model equations (e.g., neglected transformations or interaction terms).

As the commonality of these types of model errors under consideration all relate to the likelihood of such error, which in turn is connected to perturbation of probability laws governing the entire modeling construct, we apply the principle of *relative entropy* (Hansen and Sargent, 2007; Glasserman and Xu, 2013). Relative entropy between a posterior and a prior distribution is a measure of information gain when incorporating incremental data in Bayesian statistical inference. In the context of quantifying model error, relative entropy has the interpretation of a measure of the additional information requisite for a perturbed model to be considered superior to a champion or null model. Said differently relative entropy may be interpreted as measuring the credibility of a challenger model. Another useful feature of this construct is that within a relative entropy constraint the so-called worst-case alternative (e.g., in our case the upper bounds on an AIC measure due to ignoring some feature of the alternative model) can be expressed as an *exponential change of measure*.

Model risk with respect to a champion model $y = f(x)$ is quantified by the *Kullback–Leibler relative entropy* divergence measure to a challenger model $y = g(x)$ and is expressed as follows:

$$D(f, g) = \int \frac{g(x)}{f(x)} \log \left(\frac{g(x)}{f(x)} \right) f(x) dx. \quad (5)$$

In this construct, the mapping $g(x)$ is an alternative PD model, and the mapping $f(x)$ is some kind of baseline, the latter being the base PD models which we have estimated in this paper that may be violating some model assumption. In a model validation context this is a critical construct as the implication of these relations is a robustness to model misspecification with respect to the alternative model – i.e., we do not have to assume that either the reference or alternative models are correct and we need only quantify its distance of the alternative from the reference model to assess the impact of the modeling assumption at play.

Define the *likelihood ratio* $m(f, g)$, which characterizes our modeling choice, which is expressed as follows:

$$m(f, g) = \frac{g(x)}{f(x)}. \quad (6)$$

As is the standard in the literature, Equation 6 may be expressed as an equivalent expectation of a *relative deviation in likelihood*:

$$E_f[m \log(m)] = D(f, g) < \delta, \quad (7)$$

Where δ is an upper bound to deviations in model risk (which should be small on a relative basis), which may be determined by the model risk tolerance of an institution for a certain model type, interpretable as a threshold for model performance.

A property of relative entropy dictates that $D(f, g) \geq 0$ and $D(f, g) = 0$ only if $f(x) = g(x)$. Given a relative distance measure $D(f, g) < \delta$ and a set of alternative models $g(x)$, model error can be quantified by the following *change of numeraire*:

$$m_\theta(f, g) = \frac{\exp(\theta f(x))}{E_f[\exp(\theta f(x))]} \quad (8)$$

Where the solution (or inner supremum) to Eq 14 is formulated in the following optimization:

$$m_\theta(f, g) = \inf_{\theta > 0} \sup_{m(x)} E_f \left[m(x) f(x) - \frac{1}{\theta} (m(x) \log(m(x)) - \delta) \right]. \quad (9)$$

Equation 9 features the parameterization of model risk by $\theta \in [0, 1]$, where $\theta = 0$ is the best case of no model risk and $\theta = 1$ the worst case of model risk in extremis.

The change in measure of Eq 8 has important property of being *model-free*, or not dependent upon the specification of the challenger mode $lg(x)$. As mentioned previously, this reflects the robustness to misspecification of the alternative model that is a key feature of this construct, which from a model validation perspective is a desirable property – i.e., we do not have to assume that either the champion or alternative models are correct and only have to quantify the distance of the alternative from the base model to assess the impact of violating modeling assumptions.

We study the quantification of model risk with respect to the following modeling assumptions:

- Misspecification According to Omitted Variable Bias
- Misspecification According to Neglected Interaction Effects
- Misspecification According to an Incorrect Link Function

Table 12. Quantification of model risk according the principle of relative entropy resampled distribution of the relative deviation of the in-sample AIC performance measure – Moody’s large corporate financial, macroeconomic and distance-to-default explanatory variables 1- and 3-year default horizon TTC and PIT models.

Type of Model	Model Specification	Model Assumption	Min.	25 th Prctl.	Median	Mean	75 th Prctl.	Max.	Std. Dev.
Through-the-Cycle	Model 1	Omitted Variable Bias	0.0093	0.1137	0.2009	0.2290	0.3208	0.8328	0.1461
		Neglected Interaction Effects	0.0221	0.1116	0.1626	0.1759	0.2267	0.5262	0.0861
		Incorrectly Specified Link Function	0.0134	0.0721	0.0960	0.1005	0.1233	0.2714	0.0380
	Model 2	Omitted Variable Bias	0.0079	0.1010	0.1746	0.1962	0.2687	0.7362	0.1251
		Neglected Interaction Effects	0.0081	0.0830	0.1203	0.13389	0.1719	0.5239	0.0699
		Incorrectly Specified Link Function	0.0158	0.0606	0.0821	0.0866	0.1077	0.24061	0.03541
Point-in-Time	Model 1	Omitted Variable Bias	0.0044	0.0816	0.1306	0.1759	0.2149	0.5528	0.0995
		Neglected Interaction Effects	0.0123	0.0572	0.0876	0.0978	0.1266	0.4128	0.0543
		Incorrectly Specified Link Function	0.0062	0.0352	0.0486	0.0635	0.0685	0.1783	0.0256
	Model 2	Omitted Variable Bias	0.0113	0.0873	0.1414	0.1587	0.2118	0.5911	0.0945
		Neglected Interaction Effects	0.0033	0.0500	0.0765	0.0869	0.1131	0.3436	0.0505
		Incorrectly Specified Link Function	0.0077	0.0304	0.0414	0.0461	0.0580	0.1621	0.0222

Omitted variable bias is analyzed by consideration of the DTD risk factor as discussed in the main estimation results in this paper, where we saw that including this variable in the model specification did not result in other financial or macroeconomic variables falling out of the model, and improved model performance. The second assumption is based upon estimation of alternative specifications that include interaction effects amongst the explanatory variables. Finally, we analyze the third assumption above through estimation of these specifications with the *complimentary log-log* (“CLL”) as opposed to the Logit link function⁵.

⁵ The Logit links function used commonly in logistic regression is symmetric whereas CLL is asymmetric. When the probability of the binary or binomial response approaches to 0 at a different rate than it approaches to 1 (as a function of covariate), symmetric link functions cannot be appropriate and do not always provide the best fit for the given dataset in

The loss metric that we consider is AIC, and we develop a distribution of the *relative proportional deviation in AIC* (“RPD-AIC”; where we take the negative of the values as lower AICs are associated with a better fitting model specification) from the base specifications through a simulation exercise as follows. In each iteration, we resample the data with replacement (stratified in order that the history of each obligor is preserved), re-estimate the models considered in the main body of the paper, as well as three variants that either include or exclude DTD, interaction effects or a CLL link function. In the case of the DTD risk factor, we will be comparing the variants as considered in the main results which have already been estimated except that in each run, the results will be perturbed according to the different bootstraps of dataset, and in the other two cases there will be alternative estimations⁶.

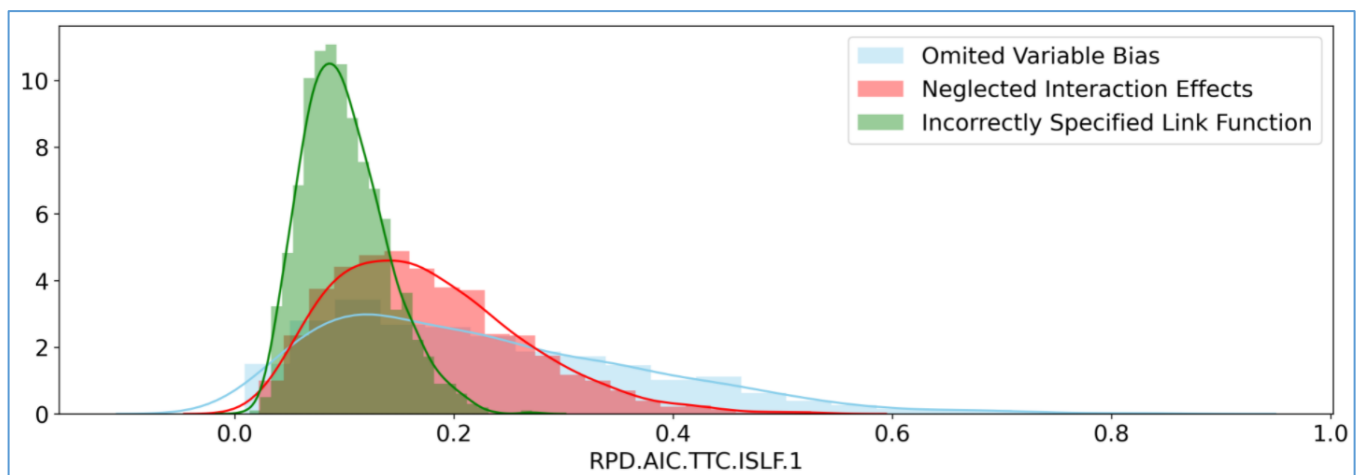


Figure 14. Quantification of model risk according the principle of relative entropy resampled distribution of the relative deviation of the in-sample AIC performance measure – Moody’s large corporate financial and distance-to-default explanatory variables 3-year default horizon TTC model 1.

The results of the model risk quantification exercise are shown in Table 12 above where we tabulate the sample moments of the bootstrapped RPD-AIC, as well as in the figures below where we plot the histograms of these. Note that we show the results of a second challenger next-best model for the models described previously in this paper, to demonstrate the robustness of our results to an alternative model specification. It is observed that omitted variable bias with respect to DTD results in the highest model risk (mean RPD-AICs ranging 0.20-0.23 and 0.15-0.17 for the TTC and PIT models, respectively), and an incorrectly specified link function has the lowest measured model risk (mean RPD-AICs ranging in 0.09-0.10 and 0.05-0.06 for the TTC and PIT models, respectively), while neglected interaction effects is intermediate in the quantity of measured model risk (mean RPD-AICs ranging in 0.13-0.18 and 0.11-0.13 for the TTC and PIT models, respectively). The other conclusion that we reach is that across violations of model

logistic regression. This is the case for the unbalanced data that we have, as defaults are very rare events, so that asymmetric link functions such as the CLL are sometimes good alternatives.

⁶ We do not include those results for the sake of brevity, but they are available upon request. Across 100,000 iterations results are stable and robust across the base as well as alternative specifications.

assumptions, the PIT models are more robust than the TTC models in terms of lower measured model risk, which is at variance with the observation that the PIT models showed worse out-of-sample model accuracy performance, and illustrates that in validating these constructs we should be looking at diverse dimensions of model performance. We further note that the distribution of the RPD-AIC is rather volatile relative to the mean and highly skewed to the right, where values in the tails of the distributions are orders of magnitude greater than measures of central tendency. This exercise shows that we should exercise caution in over-reliance on measures of model fit derived from a single historical dataset, even if out-of-sample performance is favorable, as we could be unpleasantly surprised when adding to our reference datasets when re-estimating our models.

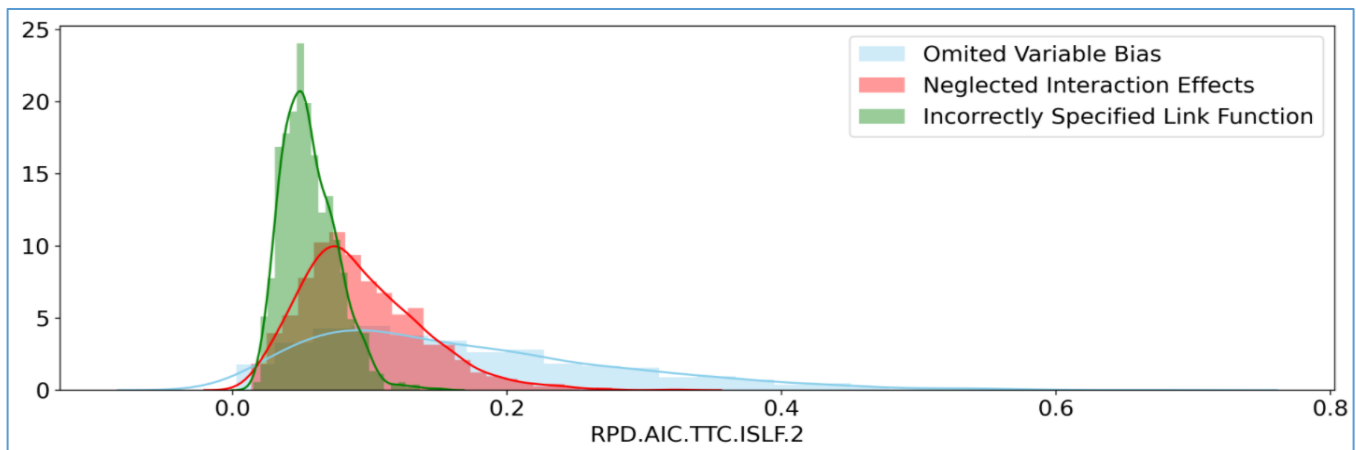


Figure 15. Quantification of model risk according the principle of relative entropy resampled distribution of the relative deviation of the in-sample AIC performance measure – Moody’s large corporate financial and distance-to-default explanatory variables 3-year default horizon TTC model 2.

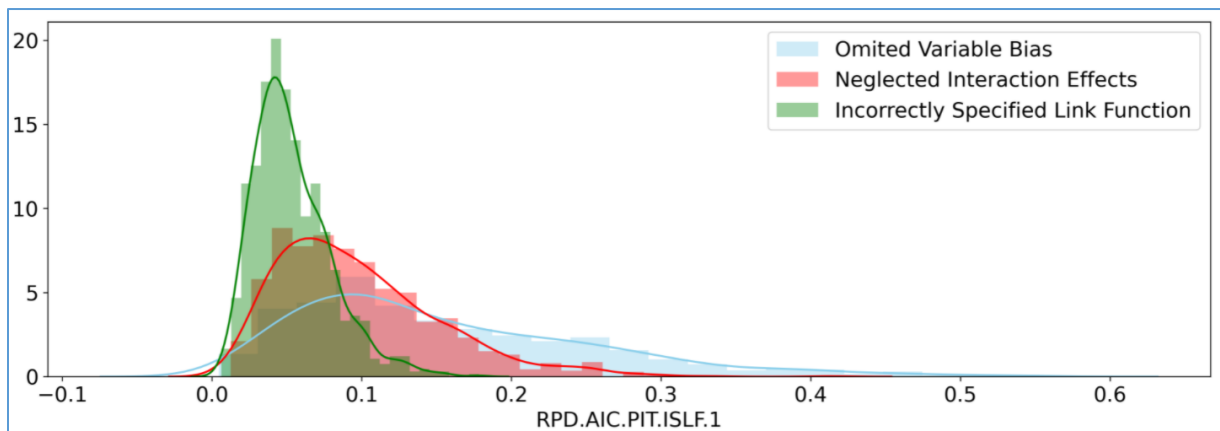


Figure 16. Quantification of model risk according the principle of relative entropy resampled distribution of the relative deviation of the in-sample AIC performance measure – Moody’s large corporate financial, macroeconomic and distance-to-default explanatory variables 1-year default horizon PIT model 1.

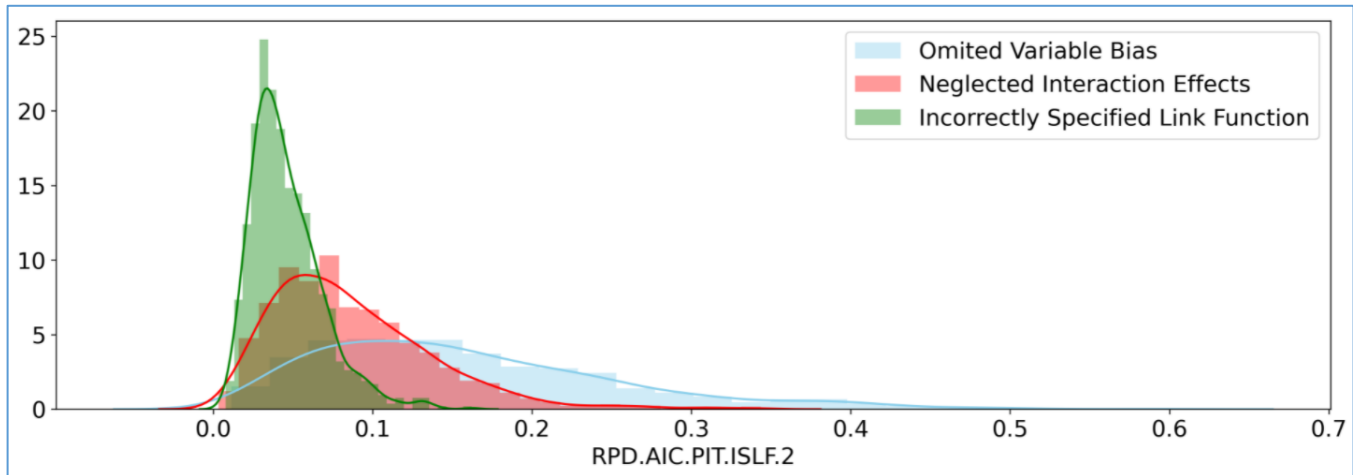


Figure 17. Quantification of model risk according the principle of relative entropy resampled distribution of the relative deviation of the in-sample AIC performance measure – Moody’s large corporate financial, macroeconomic and distance-to-default explanatory variables 1-year default horizon PIT model 2.

4. Conclusions and directions for future research

In this study, we have developed alternative simple and general econometrically estimated PD models of both TTC and PIT designs. We have avoided formal definitions of PIT vs. TTC PDs, and rather derived constructs based upon common sense criteria prevalent in the industry, and in the process have illustrated which validation techniques are applicable to these different approaches. Based upon this empirical approach to modeling, we have characterized PIT and TTC credit risk measures and have discussed the key differences between both rating philosophies. In the process, we have addressed the validation of PD models under both rating philosophies, highlighting that the validation of either rating system exhibits particular challenges. In the case of the TTC PD rating models, we have answered questions around the thresholds for rating stability metrics and level PD accuracy performance that are not settled. In the case of PIT PD rating models, we have spoken to questions around the rigorous demonstration that PD estimates are accurately estimated at the borrower level, which may not be obvious from optically observing the degree to which average PD estimates track default rates over time. Finally, this study has looked into the debate around the challenges in demonstrating the robust performance of PD models on an out-of-sample basis, which may differ between the PIT and TTC frameworks.

We have observed that the validation of a TTC or PIT PD model involves assessing the economic validity of the cyclical factor, which depending upon modeling methodology may not be available to the validator, or else may be accounted for only implicitly. One possibility is for the underlying cycle of the PD rating model to be estimated from historical data based upon some theoretical framework. However, in this study we have chosen to propose commonly used macroeconomic factors in conjunction with obligor-level default data, in line with the industry practice in building such models.

We have highlighted features of PIT vs. TTC model design in our empirical experiment, yet have not explicitly addressed how TTC PD models can be transformed into corresponding PIT PD models, or vice versa. While the advantage of such a construct is that it can be validated based upon an assumption

regarding the systematic factor and validated using a common methodology applicable to both types of PD models, we have chosen to validate each as specifically appropriate. The rationale for our approach is that the alternative runs the risk of introducing significant additional model risk (i.e., if the theoretical model is mis-specified).

We have employed a long history of borrower-level data sourced from Moody's, around 200,000 quarterly observations from a population of rated larger corporate borrowers (at least USD 1 billion in sales and domiciled in North America), spanning the period from 1990 to 2015. The dataset is comprised of an extensive set of financial, equity market and macroeconomic variables to form the basis of candidate explanatory variables. We built a set of PIT models with a 1-year default horizon and macroeconomic variables, and a set of TTC models with a 3-year default horizon and only financial ratio risk factors. We presented the leading two models in each class of PIT and TTC designs, both having favorable rank ordering power, which were chosen based upon the relative weights on explanatory variables (i.e., certain variables are expected to have different relative contributions in TTC vs. PIT constructs), as well as rating mobility metrics (e.g., PIT models are expected to show more responsive ratings and TTC models more stable ratings.) We also performed specification testing, where we observed that the TTC designs were more challenged than the PIT designs in predictive performance. The latter observation argues for the consideration of alternative risk factors, such as equity market information. In view of this, from the market value of equity and accounting measures of debt, we constructed a Merton model-style DTD measure and built hybrid structural-reduced form models, which we compared with the models containing only financial and macroeconomic variables.

We showed that adding DTD measures to our leading models did not invalidate the other variables chosen, significantly augments model performance, and in particular improved predictive accuracy of the TTC models. We also found that while all classes of models have high discriminatory power by all measures, there are some conflicting results regarding predictive accuracy depending upon the measure employed, and also that on an out-of-sample basis the TTC models actually perform better than the PIT models in this regard.

Finally, we measured the model risk attributable to various modeling assumptions according to the *principle of relative entropy*. We observed that for both TTC and PIT designs that omitted variable bias (with respect to the DTD risk factor) had the greatest, the incorrect specification of the link function had the least, and the neglect of interaction effects amongst risk factors an intermediate impact upon model risk, respectively. Another key finding in this exercise is that the PIT models had on the whole lower measured model risk across these assumptions, which is at variance with the observation that the PIT models had worse out-of-sample performance that may be evidence of over-fitting. The implication of these findings is that we should be cautious in drawing strong conclusions from analysis based on limited out-of-time data, and that we are advised to view model performance through alternative lens, such as this approach to quantifying model risk.

There are various implications for model development and validation practice, as well as supervisory policy, which can be gleaned from this study. First, it is a better practice to take into consideration the use case for a PD model in establishing the model design the model, from a fitness for purpose perspective. That said, we believe that a balance must be struck, since it would be infeasible to have separate PD models for every single use, and what we are arguing for is a parsimonious number of separate designs for major classes of use that satisfy a set of common requirements. Second, in validating PD models that are designed according to TTC or PIT constructs, we should have different emphases on which model

performance metrics are scrutinized. In light of these observations and contributions to the literature, we believe that this study provides valuable guidance to model development, model validation and supervisory practitioners. Additionally, we believe that our discourse has contributed to resolving the debates around which class of PD models is best fit for purpose in large corporate credit risk applications, providing evidence that reduced form and Merton structural models can be combined in hybrid frameworks in order to achieve superior performance. This better performance is manifest in a broad sense, both better fit to the data as well as lower measured model risk due to model mis-specification.

We would like to emphasize that we believe the principal contribution of this paper to be mainly in the domain of practical application rather than methodological innovation. Many practitioners, especially in the wholesale credit and banking book space, still use the techniques employed in this paper. We see our main contribution as proposing a structured approach to constructing a suite of TTC and PIT models, while combining reduced form and structural modeling aspects, and then by further proposing a framework for model validation. We understand that many financial institutions in this space do not have such a framework. For example, a lot of banks are still using TTC Basel models that are modified for PIT uses, such as stress testing or portfolio management. Furthermore, a preponderance of banks in this space do not employ hybrid financial and Merton-style models for credit underwriting. We believe that our contribution transcends both the academic and practitioners streams of the literature to address issues relevant to financial institution practitioners, which is also informed by the leading thought about credit risk in financial economics, which uniquely positions this research.

Given the wide relevance and scope of the topics addressed in this study, there is no shortage of fruitful avenues along which we could extend this research. Some proposals include but are not limited to:

- Alternative econometric techniques, such as various classes of machine learning models, including non-parametric alternatives;
- asset classes beyond the large corporate segments, such as small business, real estate or even retail;
- applications to stress testing of credit risk portfolios⁷;
- the consideration of industry specificity in model specification;
- investigation of machine learning or non-linear techniques;
- different modeling methodologies such as, ratings migration or hazard rate models; and,
- datasets in jurisdictions apart from the U.S., else pooled data encompassing different countries with a consideration of geographical effects.

References

- Aguais S (2008) Designing and implementing a Basel II compliant PIT-TTC ratings framework.
 Altman EI (1968) Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *J Finance* 23: 589–609. <https://doi.org/10.2307/2978933>

⁷ Refer to Jacobs Jr. et al (2015), Jacobs Jr. (2020a), Jacobs Jr. (2020b) and Jacobs Jr. (2022) for studies that address model validation and model risk quantification methodologies. These studies include supervisory applications such as *comprehensive capital analysis and review* (“CCAR”) and *current expected credit loss* (“CECL”); and further feature alternative credit risk model specifications (including machine learning model), macroeconomic scenario generation techniques, as well as the quantification and aggregation of model risk (including the principle of relative entropy as studied in this paper.)

- Altman EI, Narayanan P (1997) An international survey of business failure classification models. *Financ Mark Inst Instrum* 6: 1–57. <https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0416.00010>
- Altman EI, Rijken HA (2004) How rating agencies achieve rating stability. *J Bank Financ* 28: 2679–2714. <https://doi.org/10.1016/j.jbankfin.2004.06.006>
- Altman EI, Rijken HA (2006) A point-in-time perspective on through-the-cycle ratings. *Financ Anal J* 62: 54–70. <https://doi.org/10.2469/faj.v62.n1.4058>
- Amato JD, Furfine CH (2004) Are credit ratings procyclical? *J Bank Financ* 28: 2641–2677. <https://doi.org/10.1016/j.jbankfin.2004.06.005>
- Cesaroni T (2015) Procyclicality of credit rating systems: how to manage it. *J Econ and Bus* 82: 62–83. <https://doi.org/10.1016/j.jeconbus.2015.09.001>
- Chava S, Jarrow RA (2004) Bankruptcy prediction with industry effects. *Rev Financ* 8: 537–569. <https://doi.org/10.1093/rof/8.4.537>
- Cheng S, Long JS (2007) Testing for IIA in the multinomial logit model. *Sociol Methods Res* 35: 583–600. <https://journals.sagepub.com/doi/abs/10.1177/0049124106292361>
- Duffie D, Singleton K (1998) Simulating correlated defaults. *Paper presented at the Bank of England Conference on Credit Risk Modeling and Regulatory Implications Working Paper*, Stanford University. <https://kenneths.people.stanford.edu/sites/g/files/sbiybj3396/f/duffiesingleton1999.pdf>
- Duffie D, Singleton KJ (1999) Modeling term structures of defaultable bonds. *Rev Financ Stud* 12: 687–720. <https://doi.org/10.1093/rfs/12.4.687>
- Dwyer, Douglass, Ahmet E. Kogacil, Roger M. Stein (2004) Moody's KMV RiskCalc™ v2.1 Model. Moody's Analytics. Available from: <https://www.moody's.com/sites/products/productattachments/riskcalc%202.1%20whitepaper.pdf>
- Fry TRL, Harris MN (1996) A Monte Carlo study of tests for the independence of irrelevant alternatives property. *Transp Res Part B: Methodol* 31: 19–32. <https://journals.sagepub.com/doi/abs/10.1177/0049124106292361>
- Glasserman P, Xu X (2014) Robust risk measurement and model risk. *Quant Financ* 14: 29–58. <https://doi.org/10.1080/14697688.2013.822989>
- Hamilton DT, Sun Z, Ding M (2011) Through-the-cycle EDF credit measures. *Moody Anal.* <https://ssrn.com/abstract=1921419>
- Hausman J, McFadden D (1984) Specification tests for the multinomial logit model. *Econometrica J Econometric Soc* 1219–1240. <https://doi.org/10.2307/1910997>
- Hansen LP, Sargent TJ (2007) *Robustness*. Princeton: Princeton University Press, Available from: <http://www.library.fu.ru/files/Robustness.pdf>
- Hodrick RJ, Prescott EC (1997) Postwar US business cycles: an empirical investigation. *J Money, Credit Bank* 1–16. <https://doi.org/10.2307/2953682>
- Jacobs Jr M (2020) The accuracy of alternative supervisory methodologies for the stress testing of credit risk. *Int J Financ Eng Risk Manage* 3: 254–296. http://michaeljacobsjr.com/files/Jacobs_2020_AccAltSupMdlStrTstCrRisk_IJE_RM_vol3no3_pp254-296.pdf
- Jacobs Jr. M (2020) A holistic model validation framework for current expected credit loss (CECL) model development and implementation. *Int J Financ Stud* 8: 1–36. <https://doi.org/10.3390/ijfs8020027>
- Jacobs Jr. M (2022) Borrower level models for stress testing corporate probability of default and the quantification of model risk. *Int J Econ Finance* 14: 75–99. <https://doi.org/10.5539/ijef.v14n4p75>

- Jacobs Jr. M, Karagozoglu AK, Sensenbrenner F (2015) Stress testing and model validation: application of the Bayesian approach to a credit risk portfolio. *J Risk Model Validation* 9: 41–70. <https://ssrn.com/abstract=2684227>
- Jarrow RA, Turnbull SM (1995) Pricing derivatives on financial securities subject to credit risk. *J Financ* 50: 53–85. <https://doi.org/10.1111/j.1540-6261.1995.tb05167.x>
- Kiff MJ, Kisser M, Schumacher ML (2013) *Rating through-the-cycle: what does the concept imply for rating stability and accuracy?* International Monetary Fund, 2013. <https://www.imf.org/external/pubs/ft/wp/2013/wp1364.pdf>
- Li C, Shepherd BE (2012) A new residual for ordinal outcomes. *Biometrika* 99: 473–480. <http://dx.doi.org/10.1093/biomet/asr073>
- Liu D, Zhang H (2018) Residuals and diagnostics for ordinal regression models: a surrogate approach. *J Am Stat Assoc* 113: 845–854. <https://doi.org/10.1080/01621459.2017.1292915>
- Löffler G (2004) An anatomy of rating through the cycle. *J Bank Financ* 28: 695–720. [https://doi.org/10.1016/S0378-4266\(03\)00041-4](https://doi.org/10.1016/S0378-4266(03)00041-4)
- Löffler G (2013) Can rating agencies look through the cycle? *Rev Quant Financ Account* 40: 623–646. <https://doi.org/10.1007/s11156-012-0289-9>
- Merton RC (1974) On the pricing of corporate debt: The risk structure of interest rates. *Journal Financ* 29: 449–470. <https://doi.org/10.2307/2978814>
- Mester LJ (1997) What’s the point of credit scoring? *Federal Reserve Bank of Philadelphia Business Review*, 3–16. https://fraser.stlouisfed.org/files/docs/historical/frbphi/businessreview/frbphil_rev_199709.pdf
- Rubtsov M, Petrov A (2016) A point-in-time-through-the-cycle approach to rating assignment and probability of default calibration. *J Risk Model Validation* 10: 83–112.
- Repullo R, Saurina J, Trucharte C (2010) Mitigating the Pro-cyclicality of Basel II. *Econ Policy* 25: 659–702. <https://doi.org/10.1111/j.1468-0327.2010.00252.x>
- Small K A, Hsiao C (1985) Multinomial logit specification tests[J]. *Int Econ Rev* 26: 619–627. <https://doi.org/10.2307/2526707>
- The Bank for International Settlements—Basel Committee on Banking Supervision (BIS) (2005) *Studies on the Validation of Internal Rating Systems*. Working Paper 14. Basel: The Bank for International Settlements—Basel Committee on Banking Supervision. Available from: https://www.bis.org/publ/bcbs_wp14.htm.
- The Bank for International Settlements—Basel Committee on Banking Supervision (BIS) (2006) *International Convergence of Capital Measurement and Capital Standards: A Revised Framework*. Basel: The Bank for International Settlements—Basel Committee on Banking Supervision. Available from: <https://www.bis.org/publ/bcbsca.htm>.
- The Bank for International Settlements—Basel Committee on Banking Supervision (BIS). 2011. *Basel III: A Global Regulatory Framework for More Resilient Banks and Banking Systems*. Basel: The Bank for International Settlements—Basel Committee on Banking Supervision. Available from: <https://www.bis.org/publ/bcbs189.htm>.
- The Bank for International Settlements—Basel Committee on Banking Supervision (BIS) (2016) *Reducing Variation in Credit Risk-Weighted Assets—Constraints on the Use of Internal Model Approaches*. Consultative Document. Basel: The Bank for International Settlements—Basel Committee on Banking Supervision. Available from: <http://www.bis.org/bcbs/publ/d362.htm>.

The European Banking Authority (2016) The European Banking Authority. Guidelines on PD Estimation, LGD Estimation and the Treatment of Defaulted Exposures. Consultation Paper. Paris: The European Banking Authority. Available from: <https://www.eba.europa.eu/regulation-and-policy/model-validation/guidelines-on-pd-lgd-estimation-and-treatment-of-defaulted-assets>.

Topp R, Perl R (2010) Through-the-Cycle Ratings Versus Point-in-Time Ratings and Implications of the Mapping Between Both Rating Types. *Financ Mark Inst Instrum* 19: 47–61. <https://doi.org/10.1111/j.1468-0416.2009.00154.x>



AIMS Press

© 2022 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>)