American Economic Review: Papers & Proceedings 100 (May 2010): 506–510
http://www.aeaweb.org/articles.php?doi=10.1257/aer.100.2.506

Statistical Default Models and Incentives

By Uday Rajan, Amit Seru, and Vikrant Vig*

The likelihood that a bank loan will default is of interest to both regulators and investors. Under the Basel II regulatory guidelines, a bank must hold capital in proportion to the riskiness of its assets. The probability of default is a primary determinant of the riskiness of a loan. Investors, in turn, price a loan in the secondary market based on its expected cash flow, which again depends on the default probability.

How should market participants assess the default probability on a pool of bank loans? It is natural to consider historical data on loan conditions and default rates, and to estimate a statistical model that can be used to predict defaults going forward. Such statistical models have been widely used across the financial markets, to enhance market liquidity and impose capital requirements on financial institutions.

The accuracy of predictions from statistical models was especially poor in the subprime mortgage market in the period from August 2007 onwards.[1] We argue that one cause for this failure was that these models relied entirely on hard information variables and ignored changes in the incentives of lenders to collect soft information about borrowers.[2] That is, they failed to account for the change in the relationship between observable borrower characteristics and default likelihood caused by a fundamental change in lender behavior. Such a failure is in the spirit of the Lucas critique (Robert Lucas 1976): a purely statistical model ignores the idea that a change in the incentives of agents who generate the data may change the very nature of the data.

What changed the behavior of lenders in the subprime market? There was a tremendous growth in securitization in the subprime sector after 2000. Securitization increases the distance between the originator of the loan and the party that bears the default risk inherent in the loan. As Jeremy Stein (2002) points out, soft information is unverifiable to a third party. We argue that the increase in distance therefore results in lenders' choosing to not collect soft information (such as the likelihood of future income shocks) about borrowers. Consequently, among borrowers with similar hard information characteristics, the set that receives loans changes in a fundamental way as the securitization regime changes. This leads to a breakdown in the quality of predictions from default models that use parameters estimated using data from a period in which a low proportion of loans are securitized. Importantly, the breakdown is systematic, and therefore predictable: it occurs in the set of borrowers on whom, after conditioning on the hard information, soft information is potentially important.

In this piece, we outline a simple theoretical model that develops this argument, building on the intuition of Gary Gorton and George Pennacchi (1995) that a bank that makes and sells loans is subject to a moral hazard problem with respect to screening borrowers. We then comment on the empirical tests reported in a companion paper (Uday Rajan, Amit Seru, and Vikrant Vig 2009).

I. Model

There are three sets of agents in the model: borrowers, a single lender, and investors. At date 0, a borrower applies for a loan to be repaid

* Rajan: University of Michigan, Ross School, 701 Tappan St., MI 48109-1234 (e-mail: urajan@umich.edu); Seru: University of Chicago, Booth School of Business, 5807 S. Woodlawn Ave., Chicago, IL 60637 (e-mail: amit.seru@chicagobooth.edu); Vig: Institute of Finance and Accounting, London Business School, Regent's Park, London NW1 4SA UK (e-mail: vvig@london.edu). We are grateful to Anat Admati, Patrick Bolton, Joshua Coval, Douglas Diamond, Dirk Jenter, Anil Kashyap, Tobias Moskowitz, Mitchell Petersen, Raghuram Rajan, Ilya Strebulaev, Andrei Shleifer, Robert Vishny and especially Jeremy Stein for helpful discussions.

[1] For example, in November 2007, Standard & Poor's adjusted its LEVELS® default model to increase predicted defaults on no-documentation loans by approximately 60 percent (see Standard & Poor's 2007).

[2] Risk calculators used by rating agencies estimate default risk from hard information variables such as the borrower's credit score and the location of the property (see, for example, the FitchRatings report, October 2006).

at date 1. The loan size is homogeneous across types and is normalized to 1. At date 0, the lender costlessly observes a hard information signal x about the borrower. Based on the hard information signal, the lender decides whether to incur a cost c to obtain a soft information signal y. Using all available information, the lender offers the borrower an interest rate r. The borrower accepts or rejects the loan offer. Finally, a fixed proportion of loans made by the lender, $\alpha \in [0, 1]$, are securitized.

There is a continuum of borrowers, with each borrower having a type $\theta \in \{\theta_h, \theta_\ell\}$, where $\theta_h > \theta_\ell$. Types are independent and identically distributed across borrowers, with p being the probability a borrower has type θ_h. A borrower with type θ_j finds herself in a good state with respect to her personal finances at time 1 with probability θ_j. In this event, she repays her loan if the interest rate is sufficiently low (in a manner made precise below). With probability $1 - \theta_j$, she is in a bad state at time 1 and defaults, in which case the lender recovers zero. In equilibrium, the types will correspond to the likelihood of repayment on the loan.

A type θ_h borrower has a reservation interest rate r_h, which reflects her (unmodeled) outside opportunities (which could include applying to and obtaining a loan from another lender). A low-type borrower accepts any loan that is offered, so her reservation interest rate may be thought of as infinite. The lender's cost of funds is normalized to zero. There is a threshold interest rate $r_\ell > r_h$ such that the low type defaults even in the good state if $r > r_\ell$. We further assume that $\theta_h(1 + r_h) > 1 > \theta_\ell(1 + r_\ell)$, so a lender earns a positive expected profit if it makes a loan to the high type at rate r_h, but loses money on any loan made to a low type.

On each borrower, the lender obtains a hard information signal $x \in \{x_h, x_\ell\}$ at zero cost. The hard information signal incorporates verifiable data such as the borrower's FICO credit score. Let $\delta_i = \Pr(x = x_h | \theta_i)$ be the probability that the hard information signal is x_h given that the borrower has type θ_i. Given a hard information signal x_j, let μ_j denote the posterior probability the borrower has type θ_h. The hard information signal is informative: $\mu_h > p$. Hard information signals are conditionally independent across borrowers.

Having seen the hard information signal, the lender may choose to incur a cost c and obtain a soft information signal about the borrower, $y \in \{y_h, y_\ell\}$. Soft information here includes any information related to the likelihood of default that is not verifiable by a third party, such as the likelihood that the borrower's job may be terminated or she will be credit-constrained in the future. Given that the borrower's type is θ_i, let $\gamma_i = \Pr(y = y_h | \theta_i)$ be the probability the lender receives the soft information signal y_i. Conditional on borrower type, soft information signals are independent across borrowers and uncorrelated with hard information signals. The soft information signal is also informative: $\gamma_h > \gamma_\ell$. Given the signals (x_i, y_j), the posterior probability a borrower has type θ_h is denoted λ_{ij}.

Given the signals it has observed, the lender chooses to either offer the borrower a loan at a specified interest rate r or not offer a loan. A high-type borrower accepts a loan if $r \leq r_h$, and a low-type borrower accepts all loan offers. A profit-maximizing lender will charge an interest rate r_h if it chooses to offer a loan, and the borrower will accept such an offer. Let $v_i = \theta_i(1 + r_h) - 1$ be the net present value of a loan to type θ_i with interest rate r_h. Then, $v_h > 0 > v_\ell$.

A loan may be securitized; i.e., sold to investors. Any particular loan made by a lender is securitized with an exogenous probability α. With probability $(1 - \alpha)$, the lender must retain the loan. It is common in the residential mortgage market for a lender to offer a basket of loans to investors, who randomly select loans in every category. Thus, on any given loan, there is a positive probability the lender will have to retain it.

For any loan made by the lender, investors observe the interest rate on the loan, r, and the hard information associated with the borrower, x. The soft information, y, is not verifiable and therefore not contractible. Financial markets are perfectly competitive, so the price of a loan equals its expected payoff, and investors earn zero profit. Let $P(x)$ denote the price of a loan with hard information signal x. Then, if investors believe the borrower has type θ_h with probability $z(x)$, $P(x) = 1 + v_\ell + z(x)[v_h - v_\ell]$.

Since all loans are offered at the same rate in equilibrium, screening on soft information can be valuable only if there is an improvement in the pool of borrowers that receive loans. The following assumption is sufficient

to ensure that a lender retaining a loan will acquire soft information if the hard information signal is x_ℓ but not if it is x_h: (i) $\lambda_{\ell\ell} < ((-v_\ell)/(v_h - v_\ell)) < \lambda_{h\ell}$ and (ii) $c < -(1 - ((\mu_\ell\gamma_h)/(\lambda_{\ell h})))\, v_\ell - \mu_\ell(1-\gamma_h)(v_h - v_\ell)$. Note that, for brevity, we state the assumptions directly in terms of the lender's posterior beliefs. Part (i) implies that a lender retaining a loan finds it optimal to screen out borrowers who generate the signals (x_ℓ, y_ℓ), but not those with signals (x_h, y_ℓ). Part (ii) implies that a lender retaining a loan earns a higher profit from lending only to borrowers with signals (x_ℓ, y_h) rather than lending to all borrowers with signal x_ℓ.

Therefore, the lender makes a loan to all borrowers that generate the high hard information signal, but may choose to obtain soft information on those with the low hard information signal. In equilibrium, the lender acquires soft information about the borrower only if the degree of securitization is sufficiently low. Whenever the lender acquires soft information, it does not lend to borrowers with signal y_ℓ. The proof of the following proposition is in the online Appendix.

PROPOSITION 1: *There exist securitization thresholds* $\underline{\alpha}, \overline{\alpha} \in (0, 1)$*, with* $\underline{\alpha} < \overline{\alpha}$*, such that in equilibrium* (*i*) *a lender acquires soft information only if* $\alpha \leq \underline{\alpha}$ *and the hard information signal is* x_ℓ*, and* (*ii*) *a lender does not acquire soft information only if* $\alpha \geq \overline{\alpha}$.

If the degree of securitization is low, the lender collects soft information when the hard information signal is x_ℓ, and the loan is priced accordingly. However, when the degree of securitization is high, the moral hazard problem with respect to collecting soft information is too severe, and only hard information is obtained by the lender.

We ignore the possibility that the lender may choose which loans to offer for securitization. Suppose the lender had such a choice, as in the work of Christine Parlour and Guillaume Plantin (2008). Then, in equilibrium, it must be optimal for the lender to make loans to at least some borrowers with signals (x_ℓ, y_ℓ), and to offer these loans to investors. The intuition of our model therefore goes through if the lender can selectively retain loans: the average quality of loans issued in a high securitization regime is worse than the average quality in a low securitization regime.

A. *Optimal Degree of Securitization*

In our model, we treat the securitization probability α as exogenous. In practice, of course, the level of securitization will depend on the costs and benefits to a particular lender. One benefit is that securitization frees up capital that can be used to make additional investments. If a bank holds a loan on its balance sheet, it is subject to minimum capital requirements, which must be met before it can expand lending. A lender will thus find it attractive to securitize loans if it wishes to release capital to invest in new projects (see Patrick Bolton and Xavier Freixas 2001, or Andrei Shleifer and Robert Vishny 2010), or if it seeks to increase its market share or sales. On the investor side, securitization increases opportunities for risk-sharing. The cost of securitization, beyond the direct effect on loan values, can include a loss of reputation if lower quality loans are made. For example, there may be a reputational cost to the lender in other areas of business.

Since lenders are heterogeneous with respect to the costs and benefits of securitization, actual levels of securitization will vary in the cross-section. For a given lender, the benefits of freeing up capital will decline with the volume of loans, as the lender invests in less profitable projects. With decreasing marginal benefits (or increasing marginal costs), each lender optimally attains an interior degree of securitization. The optimal α for a given lender will vary across time as the costs and benefits change. In the time-series data, increasing levels of securitization over time may emerge both from individual lenders securitizing more of their loans over time and from high-securitization lenders capturing a greater share of the market.

II. Empirical Results

The theoretical model above has two key empirical predictions. First, comparing a high-securitization regime (i.e., a regime with high α) to a low-securitization one (with low α), the interest rate on newly issued loans must rely more on hard information in the high-securitization regime. Second, in moving from a low- to a high-securitization regime the composition of borrowers with weak hard information signals changes: borrowers denied credit in the low-securitization regime obtain loans

in the high-securitization one. Consequently, a statistical default model estimated using data from a low-securitization era will underestimate defaults in the high-securitization era precisely for borrowers with weak hard information signals.

These predictions are tested by Uday Rajan, Amit Seru, and Vikrant Vig (2009; henceforth RSV) on a dataset that includes all securitized subprime mortgage loans in the United States issued from 1998 to 2006. The level of securitization increased dramatically over this period, from about 30 percent of all issued loans in 1998 to over 80 percent of all issued loans by 2006. There was a concomitant increase in the volume of loans issued in the market through this period. Thus, over the years, the data exhibit an increasing level of securitization.

RSV consider a borrower's FICO credit score and the loan-to-value (LTV) ratio on a loan to be the key hard information variables. A higher FICO score and a lower LTV ratio plausibly correspond to stronger hard information signals. RSV test the first prediction on reliance on hard information variables in two ways. First, in our loan sample, the R^2 of a regression of interest rates on just two variables, the FICO score and LTV ratio, increases from three percent in 1997 to almost 50 percent in 2006. This evidence is consistent with an increased reliance on hard information in setting the interest rate as the level of securitization increases.

Second, conditioning on the FICO score, the variance of interest rates on newly issued loans shrinks over time. The latter effect occurs especially for borrowers with low FICO scores, on whom soft information is more important. The shrinkage occurs even after controlling for standardization of mortgage loan features over time. The increased level of securitization over time is therefore potentially accompanied by a loss of soft information about borrowers.[3]

To test the second prediction, RSV estimate a statistical default model from loans issued in a period with a low degree of securitization (1997–2000), using hard information variables about borrowers. A loan is considered to be in default if it is delinquent for more than 90 days at any time up to two years after issuance. We then predict default rates on loans issued in the period 2001–2006, keeping the coefficients of the statistical default model the same as in the low-securitization era. For each loan, define the prediction error as the actual default minus the predicted default. We find that the forecast errors are positive on average and greater than zero at a large majority of FICO scores in each year. More important, the degree of under-prediction progressively worsens as the securitization increases, suggesting that at the same hard information characteristics, the set of borrowers receiving loans has worsened over time.[4] Finally, we find a systematic variation in the prediction errors: they increase as the borrower's FICO score falls and the LTV ratio increases. This finding is consistent with the composition of the borrower pool being most affected for borrowers on whom soft information is valuable, with lenders no longer collecting soft information about borrowers in the high-securitization era.

Since house price movements may impact defaults and in turn the prediction errors, RSV employ several strategies to examine whether the results could be explained by house price movements. In particular, we argue that the results appear even in periods when house prices were increasing (i.e., before 2006), survive when the baseline model is estimated on a rolling window (rather than 1997–2000), and remain qualitatively similar when we forecast defaults using perfect foresight about state-level changes in house prices for two years *after* a loan has been issued. Overall, the data are consistent with the predictions of the theoretical model outlined here.

III. Discussion

Our work provides a Lucas critique on statistical models that naïvely calibrate on historical data without modeling the agent behavior that produces these data. We argue that one could observe systematic deviations from the predictions of these models if incentives of agents that produce the data change in a predictable manner

[3] This test is also useful in ruling out an alternative hypothesis that securitization results in a lower cost of capital for banks, leading to an increase in the riskiness of the marginal borrower, and hence increase in the dispersion of interest rates over time.

[4] As a placebo test, RSV estimate a default model on loans issued over a subset of the low securitization era (1997 and 1998) and show that the model performs well in another low-securitization era (1999 and 2000).

over time. This implies that regulations based on the blind use of such models may be undermined by the actions of market participants. For instance, as specified in the Basel Committee on Banking Supervision (2006) document, the Basel II guidelines suggest that a regulator may rely on a third party such as a rating agency to estimate the probability of default on bank loans. Our analysis employs a statistical model that is similar in spirit to the Standard and Poor's LEVELS® 6.1 Model. As we demonstrate, this model could produce systematic errors if it does not account for the change in the data generating process.

Going forward, incorporating the effect of incentives into models that assess the riskiness of a pool of loans remains a fruitful area of research. There are at least two possible avenues that seem promising for regulators. One approach would be a structural model of default, which incorporates a selection equation that describes how the pool of borrowers changes due to different economic forces that drive the behavior of agents. A second approach would rely on including market signals in the statistical model to improve default predictions. To the extent market participants price the risk appropriately, much like the investors in our model, including these market signals (such as the market price of loans in our model) in the statistical model would automatically account for the systematic biases that arise from a naïve approach.

REFERENCES

Basel Committee on Banking Supervision. 2006. "International Convergence of Capital Measurement and Capital Standards: A Revised Framework." Bank for International Settlements. http://www.bis.org/publ/bcbs128.pdf.

Bolton, Patrick, and Xavier Freixas. 2000. "Equity, Bonds, and Bank Debt: Capital Structure and Financial Market Equilibrium under Asymmetric Information." *Journal of Political Economy*, 108(2): 324–51.

Gorton, Gary B., and George G. Pennacchi. 1995. "Banks and Loan Sales: Marketing Nonmarketable Assets." *Journal of Monetary Economics*, 35(3): 389–411.

Lucas, Robert E. Jr. 1976. "Econometric Policy Evaluation: A Critique." In The Phillips Curve and Labor Markets, Vol. 1, *Carnegie-Rochester Conference Series on Public Policy*, ed. Karl Brunner and Allan H. Meltzer, 19–46. Amsterdam: North Holland Press.

Parlour, Christine A., and Guillaume Plantin. 2008. "Loan Sales and Relationship Banking." *Journal of Finance*, 63(3): 1291–314.

Rajan, Uday, Amit Seru and Vikrant Vig. 2009. "The Failure of Models that Predict Failure: Distance, Incentives and Defaults." Unpublished.

Shleifer, Andrei, and Robert W. Vishny. Forthcoming. "Unstable Banking." *Journal of Financial Economics.*

Standard & Poor's. 2007. "Standard & Poor's Enhances LEVELS 6.1 Model." News release, November 9, 2007. http://www2.standardandpoors.com/spf/images/products/LEVELS6.1_Commentary.pdf.

Stein, Jeremy C. 2002. "Information Production and Capital Allocation: Decentralized Versus Hierarchical Firms." *Journal of Finance*, 57(5): 1891–921.