Models, Forecasts and Validation[1]

Nicholas M. Kiefer
Cornell University Departments of Economics and Statistical Sciences and Office of the Comptroller of the Currency (OCC) Risk Analysis Division

June 8, 2009

[1] Wharton Financial Institutions Center Conference: Modeling Retail Credit Risk After the Sub-Prime Crisis– Looking Ahead: Techniques, Validation, Testing?

Important Note

Disclaimer: The statements made and views expressed herein are solely those of the author, and do not necessarily represent official policies, statements or views of the Office of the Comptroller of the Currency or its staff.

nicholas.kiefer@cornell.edu

Issues in Inference and Validation

Prediction of the default probability θ for a portfolio of loans.

Modeling of uncertainty through probabilities: forecast uncertainty, parameter uncertainty, etc.

Combination of expert and data information

Result: Posterior distribution for next-period's default rate θ_{T+1} (in a specified bucket)

Validation: How well can default rates be predicted?

Problem: Overrejection of "good" models using standard tests

Validation: Sources of Forecast Uncertainty

Typically, a model is used to forecast θ_{T+1}, that forecast, θ^F_{T+1}, is regarded as a fixed number, and the sampling distribution of the period $T+1$ default rate $\widehat{\theta_{T+1}}$ is used to test $H_0 : \theta_{T+1} = \theta^F_{T+1}$. It is widely thought that this procedure leads to too many rejections.

There are two different sources of uncertainty about θ_{T+1},
1) that generated by the economic model given parameter values,
2) that generated by parameter uncertainty.

Model 1: Binomial

The value of the ith asset in time t is

$$v_{it} = \epsilon_{it}$$

Default occurs if $v_{it} < T$
$d_i = 1$ for default, otherwise $d_i = 0$.
The distribution of $d_i$is $p(d_i|\theta) = \theta^{d_i}(1-\theta)^{1-d_i}$, where $\theta = \Phi(T^*)$
$D = \{d_i, i = 1, ..., n\}$ is the whole data set; $r = r(D) = \sum_i d_i$
Then

$$p(D) = p(r) = \theta^r(1-\theta)^{n-r}$$

Extensions: Models 2 & 3

Model 2: Single-Factor Model (B2)

$$v_{it} = \rho^{1/2} x_t + (1-\rho)^{1/2} \epsilon_{it}$$

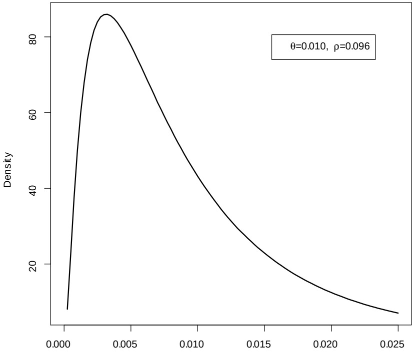
x_t is a common shock, inducing correlation ρ across asset values
For $\rho \neq 0$ there is random variation in the default probability over time and this induces correlation in defaults across assets within a period.
Model 3 adds temporal correlation in x_t so this period's realized default rate helps predict next period's.

Data and Estimation

Sample: mid-portfolio corporate bonds from S&P-rated firms in the KMV North American Non-Financial Dataset.
Cohorts of firms starting in September 1993 and running through September 2004.
2197 asset/years of data and 20 defaults: empirical rate 0.00913.
The binomial model is straightforward using direct calculations involving numerical integration
Model 2,and Model 3 are analyzed using MCMC.

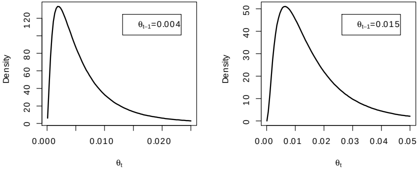
Distributions of Forecasts Given Parameters

In Model 1, the default rate is constant over time so the $T+1$ forecast default rate θ^F_{T+1} is simply $E(\theta|R)$ =0.0096.
In Model 2, the density of the default rate θ_{T+1} conditional on $\eta = (\theta, \tau)$ has $E(\theta_{T+1}|\theta, \rho, R) = 0.010$ and $\sigma_{\theta_{T+1}} = 0.009$

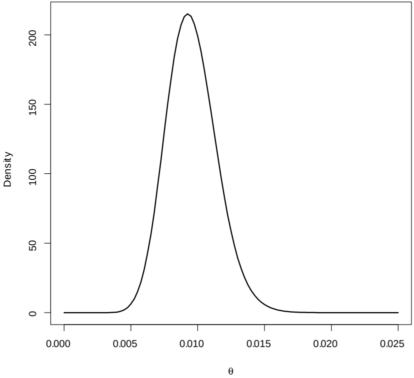


Distributions of Forecasts Given Parameters

In Model 3 The period T default rate is useful in predicting θ_{T+1} because of the dynamics of x_t.$E(\theta_{T+1}|\theta_T = 0.004, \theta, \rho, R) =$ 0.007 with $\sigma_{\theta_{T+1}} = 0.006$, and $E(\theta_{T+1}|\theta_T = 0.015, \theta, \rho, R) =$ 0.016 with $\sigma_{\theta_{T+1}} = 0.013$.

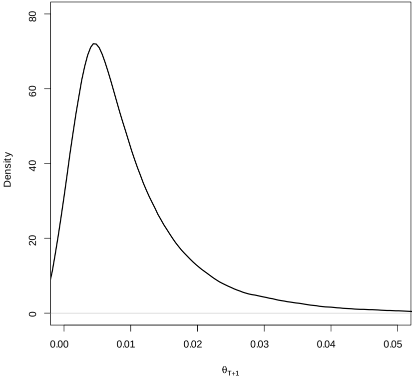


Unconditional Forecast Distribution: Binomial model



This density has $E(\theta|R) =$0.0096 and $\sigma_\theta = 0.0012$..

Unconditional Forecast Distribution: Model 2



$E(\theta_{T+1}|R) =$0.010 and $\sigma_{\theta_{T+1}} =$0.010. The variation predicted by the one-factor model increases the prediction se by a factor of 8. Unconditioning only adds about 11%

Unconditional Forecast Distribution: Model 3

$E(\theta_{T+1}|\theta_T = 0.004, \theta, \rho, R) = 0.006$ with $\sigma_{\theta_{T+1}} = 0.004$, and
$E(\theta_{T+1}|\theta_T = 0.015, \theta, \rho, R) = 0.014$ with $\sigma_{\theta_{T+1}} = 0.010$

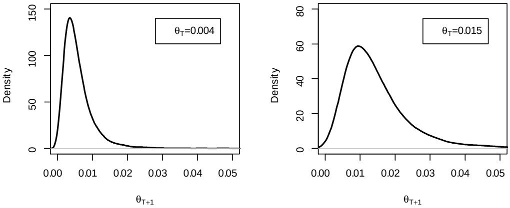


Table 1: Highest Posterior Density Intervals

Model	$p = .9$	$p = .75$	$p = .5$
Binomial	0.0096	0.0096	0.0096
$2 - parameter$	0.0008, 0.0241	0.0008, 0.01506	0.0012, 0.0089
$3 - p, \theta_T = 0.004$	0.0002, 0.0143	0.0007, 0.0093	0.0006, 0.0053
$3 - p, \theta_T = 0.015$	0.0009, 0.0346	0.0016, 0.0218	0.0031, 0.0138
$Binomial(m)$	0.0063, 0.0121	0.0077, 0.0118	0.0089, 0.0111
$2 - parameter(m)$	0.0002, 0.0217	0.0005, 0.0133	0.0011, 0.0078
$3 - p, \theta_T = 0.004(m)$	0.0006, 0.0109	0.0010, 0.0075	0.0017, 0.0053
$3 - p, \theta_T = 0.015(m)$	0.0022, 0.0267	0.0038, 0.0194	0.0056, 0.0143

Note: (m) denotes marginal with respect to parameters.

Forecast default rates are uncertain

1) because the generating model is stochastic and
2) because of uncertainty about the parameters.
1) is a much larger source of uncertainty in the application and dataset considered in this paper.
The stochastic model generates too much uncertainty: default rates are not as difficult to predict as the model implies.
Another and more difficult source of forecast uncertainty: model uncertainty. Solutions?