ISSN 0956-8549-614

Forecasting Bankruptcy and Physical Default Intensity

Ping Zhou

DISCUSSION PAPER NO 614

DISCUSSION PAPER SERIES

September 2007

Ping Zhou received her BSc degree in Economics from Renmin University of China in 1996, the MA degree in Economics from Université catholique de Louvain in 2003, and the MSc degree in Statistics from Katholieke Universiteit Leuven in 2004. She is currently a PhD student in Finance in the Department of Finance at London School of Economics and the University of Lugano. Her research interests include credit risk modelling and financial econometrics.

Forecasting Bankruptcy and Physical Default Intensity

Ping Zhou *

Abstract

This report presents two of our investigations: one is to obtain an accurate forecast for the corporate bankruptcy; the other is to obtain a physical default intensity. Both investigations were based on the hazard model, using only firm-specific accounting variables as predictors. Different methods, such as the list-wise deleting, closest-value imputation and multiple imputation, were applied to tackling the problem of missing values. Our empirical studies showed that the multiple imputation performed the best amongst these methods and led to a forecasting model with economically reasonable predictors and corresponding estimates.

1 Introduction

The purpose of this report is, considering both the firm-specific accounting and market information and the macro-economics information, to investigate the following two questions.

Which factors are the determinants to predict the firm bankruptcy? And, in order to examine the relationship between the physical bankruptcy risk premium and risk-neutral bankruptcy risk premium, how to obtain the physical default intensity?

*London School of Economics and University of Lugano. I would like to thank Professor Ron Anderson for his constructive comments. This research has been supported by the EPSRC Grant No. EPRC522958/1, "Integrating Historical Data and Market Expectations in Risk Assessment for Financial Institutions".

Although many investigations have been performed, these two questions remain open in the empirical research. The contribution of this report is that: our empirical studies showed that, compared with the list-wise deleting and closest-value imputation to tackle the problem of missing values of the predictor variables, the multiple imputation performed the best and led to a forecasting model with economically reasonable predictors and corresponding estimates, which reflected firm-specific features of profitability, leverage and stock market information and their impact on the bankruptcy.

The problem of missing values often hinders the statistical inference for panel data, such as those collected in clinical trials, biostatistics and credit risk management. In the context of credit risk management, the data of a financially-distressed firm are more likely to have missing values than those of a healthy firm; this leads to a self-selection bias of the data. In general a distressed firm is more reluctant to provide the accounting information such as its net income, because, for example, auditors may be unwilling to sign off on financial reports, or the investors' expectation about the firm's performance may be hurt. Consequently, methods to cope with the missing values and thus correct the self-selection bias may play a vital role in forecasting bankruptcy. As observed from our empirical studies, the results of parameter estimation are indeed sensitive to the method chosen to deal with the missing values, at least in terms of bias and efficiency of the estimates.

The simplest method is to list-wisely delete the missing values, i.e., to delete all the observations with any missing values. However, this method is not reasonable if the missing values count nontrivial portion of the data set or play an important role in the analysis, because the important information, which is implicitly conveyed by the pattern of the missing values themselves, is lost. Also the inference based on this method may suffer from selection bias due to the drop of observations.

Another simple method is to simply impute the missing values by the closest non-missing values; however, it is still not able to sufficiently recover the information of the missing values, e.g., changes in values at crucial times are missed.

Alternatively, we can use the method of multiple imputation to impute the missing values where the uncertainty about the right values to impute are taken into account.

Our empirical studies with these three methods are detailed in Section 4. In the literature, missing values might be either substituted from past observations (e.g., Shumway (2001)), or list-wisely deleted or substituted by

cross-sectional means or medians (e.g., Campbell et al. (2005)).

After the processing of missing values, a bankruptcy forecast can be performed within either a framework of statistical models or a framework of credit risk models.

Within the framework of credit risk models, structure models and reduced-form models were widely used. Merton (1974) pioneered in using the structure models for forecasting default: a default occurs when the firm's value falls below the face value of the firm's bond at maturity. Black and Cox (1976) extended the models in Merton (1974) to first-passage models, which allow the occurrence of a default at any time. Leland (1994), Anderson and Sundaresan (1996) and Longstaff and Schwartz (1995), among others, were subsequent extensions. Reduced-form models, as used by Jarrow and Turnbull (1995) and Duffie and Singleton (1999), define a default as the first arrival time of a Poisson process at a mean arrival rate.

Within the framework of statistical models, Shumway (2001) developed a hazard model to forecast bankruptcy by using yearly frequency data. Altman (1968) pioneered in using classification models for forecasting bankruptcy, which were referred to as static models in Shumway (2001). Shumway (2001) compared the empirical estimates obtained from the hazard model with those obtained from the static models, and concluded that the hazard model was more appropriate than the static models for forecasting bankruptcy. Chava and Jarrow (2004) confirmed the superior forecasting performance of the hazard model of Shumway (2001) to that of the models of Altman (1968) and Zmijewski (1984), using both yearly and monthly frequency data. Campbell et al. (2005) used a similar model to predict the firm bankruptcy and failure at short and long time periods, and claimed that their best model had greater explanatory power than those of Shumway (2001) and Chava and Jarrow (2004). Duffie et al. (2005) incorporated the time dynamics of the predictor variables into their model. Our report can be located within this framework.

We used the hazard model for a sample between 1995 and 2005; our empirical studies showed that, by using list-wise deleting or closest-value imputation for the missing values, the results were not fully in lines with the literature (e.g., Shumway (2001) and Campbell et al. (2005)) in terms of statistical significance of the estimates of the predictor variables. However, by using the multiple imputation, the estimation results conformed to those in the literature, in terms of not only statistically significance but also expected signs. Using the estimated coefficients of the predictor variables, we were

able to obtain the physical default intensity.

2 The model

The hazard model is used to describe the physical default intensity with a merit of no assuming a joint distribution for the predictor variables. Shumway (2001) shows that a multi-period logit model is equivalent to a discrete-time hazard model with a hazard function $\phi(\tau, X; \alpha, \beta)$. The hazard function is defined as

$$\phi(\tau, X; \alpha, \beta) = \frac{f(\tau, X; \alpha, \beta)}{1 - \sum_{j<\tau} f(j, X; \alpha, \beta)} \ , \tag{1}$$

where $f(\tau, X; \alpha, \beta)$ is the probability mass function of failure and provides the conditional probability of failure at time τ conditional on survival to τ. That is, if we assume that the failure time is the time when the firm filed for bankruptcy, then the conditional probability of the firm i filing for bankruptcy at time t, given the information to time $t-1$, is given by

$$Pr(y_{i,t} = 1 | X_{i,t-1}, y_{i,t-1} = 0) = \frac{1}{1 + e^{-\alpha - X'_{i,t-1}\beta}} \ , \tag{2}$$

where $y_{i,t}$ is the indicator, which equals one when the firm i filed for bankruptcy at time t, X is the vector of predictor variables, α is the scalar constant term and β is the vector of the parameters for the predictor variables. The α and β can be estimated by maximum likelihood estimation.

The likelihood function is written as

$$\mathcal{L} = \prod_{i=1}^{n} \left(\phi(t_i, X_i; \alpha, \beta)^{y_{i,t_i}} \prod_{k_i < t_i} [1 - \phi(k_i, X_i; \alpha, \beta)]^{y_{i,k_i}} \right) \ . \tag{3}$$

If the data was collected quarter by quarter, then, in order to forecast the bankruptcy in one quarter ($j = 1$), six months ($j = 2$) or one year ($j = 4$), a logit specification can be rewritten, for the probability of the firm filing for bankruptcy in j quarters, as (Campbell et al., 2005)

$$Pr(y_{i,t-1+j} = 1 | X_{i,t-1}, y_{i,t-2+j} = 0) = \frac{1}{1 + e^{-\alpha_j - X'_{i,t-1}\beta_j}} \ . \tag{4}$$

If we assume that the probability of the firm filing for bankruptcy does not change with the prediction horizon, i.e., $\alpha_j = \alpha$ and $\beta_j = \beta$, then the cumulative probability of the firm filing for bankruptcy over j quarters is

$$1 - \prod_{l=1}^{j} Pr(y_{i,t-1+l} = 0 | X_{i,t-1}, y_{i,t-2+l} = 0) = 1 - \left(\frac{e^{-\alpha - X'_{i,t-1}\beta}}{1 + e^{-\alpha - X'_{i,t-1}\beta}} \right)^j . \quad (5)$$

The physical default intensity, λ_j^P, over j quarters at time t, can then be estimated as

$$\lambda_t^P(j) = j e^{\hat{\alpha} + X'_{t-1}\hat{\beta}} , \quad (6)$$

by using the estimated parameters.

3 The data

3.1 Raw variables

Our sample period is from the beginning of 1995 to the end of 2005; our raw variables consist of 10 firm-specific and 9 macro-economic variables for the US market.

The 10 firm-specific variables include the indicator of the timing of firms filing for bankruptcy, the accounting variables, and the quarterly and daily stock prices for non-financial firms, which are publicly listed in the US market. The timing of firms filing for bankruptcy is collected from FISD (Fixed Investment Securities Database). The accounting variables and the quarterly stock price are collected from Compustat North America. The daily stock price is collected from CRSP.

The 9 macro-economic variables include the VIX (Volatility Index), the 3-month, 1-year and 10-year Treasury bill/note rates, three Fama-French factors, the level and market capitalisation of S&P 500. Daily observations of the VIX are obtained from the website of Chicago Board Options Exchange, monthly observations of the Treasury bill/note rates are obtained from the website of the Federal Reserve Board, the monthly Fama-French factors are obtained from Ken French's website, and the monthly data on S&P 500 are obtained from CRSP.

The firm-specific variables are first matched into the quarterly frequency data set by using the common identifier CUSIP (Committee on Uniform Securities Identification Procedures) code amongst Compustat, FISD and

CRSP data resources. Then the macro-economic variables are added into the data set by matching the year and the quarter with the firm-specific variables.

In more detail, for each firm, we have 44 quarterly observations (rows); for each observations (rows), we have 16 variables (columns). In this quarterly-frequency data set, there are in total $89,276$ observations representing $2,029$ firms, where 79 firms filed for bankruptcy.

Variable	Label	N^*	N^* Missing	Mean	Std. Dev.
DATA14	Price Close 3rd Month of Quarter ($&c)	60151	29125	28.43	32.49
DATA36	Cash and Short Term Investments (MM)	66809	22467	348.66	1549.34
DATA44	Assets Total (MM)	67077	22199	5439.02	19641.68
DATA49	Current Liabilities Total (MM)	63974	25302	1088.52	3173.61
DATA51	Long Term Debt Total (MM)	66615	22661	1492.96	6860.17
DATA54	Liabilities Total (MM)	67068	22208	3711.53	16312.75
DATA59	Common Equity Total (MM)	66522	22754	1693.67	5184.17
DATA61	Common Shares Outstanding (MM)	63963	25313	154.58	466.41
DATA69	Net Income (Loss) (MM)	68863	20413	46.30	451.32

Table 1: The simple statistics of the raw firm-specific data (N^*: the number of observations).

The simple statistics of the raw firm-specific variables are shown as Table 1; the data set is illustrated in Table 2. It is observed that the data set has a sever problem of missing values and some possible occurrence of extreme values of the variables.

3.2 The dependent variable

The dependent variable is the binary indicator y_t such that y_t equals one if the timing of the firm filing for bankruptcy falls at t, and otherwise zero. The timing is described in the FISD as the date on which the bankruptcy petition was filed under the Chapter 7 (Liquidation) or Chapter 11 (Reorganisation) of the US bankruptcy laws.

3.3 Predictor variables

From the raw data and the matched quarterly-frequency data set, we constructed 15 predictor variables: 9 firm-specific predictor variables to capture

Obs	CNUM	y	DATA14	DATA36	DATA44	DATA49	..	VIX	SRRATE
1	000361	0	16.625	28.557	411.362	59.484	..	13.58	0.0591
2	000361	0	18.375	22.960	421.450	67.828	..	12.88	0.0564
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
44	000361	0	24.080	NaN	NaN	NaN	..	11.77	0.0397
45	00081T	0	NaN	NaN	NaN	NaN	..	13.58	0.0591
46	00081T	0	NaN	NaN	NaN	NaN	..	12.88	0.0564
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
80	00081T	0	NaN	60.500	886.70	265.800	..	17.51	0.0091
81	00081T	0	NaN	NaN	NaN	NaN	..	16.73	0.0095
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
84	00081T	0	NaN	79.800	984.50	324.8	..	13.58	0.0222
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
88	00081T	0	24.500	91.100	1929.50	453.000	..	11.77	0.0397
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 2: Illustration of a sample in the raw quarterly data set.

the firm's profitability, leverage, liquidity and stock price variation, and 6 macro-economic predictor variables to capture the macroeconomic status.

The description for the raw firm-specific and the macro-economic predictor variables is shown in Table 3.

Amongst the 9 firm-specific predictor variables, the net income over total asset (NITA), the total liability over total asset (TLTA), the cash to total asset (CASHTA), the market over book ratio (MB) were calculated directly from the raw data. The PRICE, an indicator of financial distress as reverse stock splits are relatively rare, was calculated by the natural logarithm of the minimum between firm's share price and \$15. The distance to default (DtD) was constructed based on the existing literature (see Section 3.4 for its construction). The firm's relative size (RSIZE) and excess return (EXRET) were based on the firm's market capitalisation and stock price, and on the market capitalisation and the level of S&P500. An annualised three-month sample standard deviation of the firm's daily return was calculated as a proxy of the firm's equity volatility (SIGMA), i.e.,

$$SIGMA_t = \left(252 \times \frac{1}{N-1} \sum_{j \in t} r_j^2 \right)^{\frac{1}{2}} ,$$

Predictor variables	Description
Firm-specific predictor variables	
NITA	net income / book value of total asset
TLTA	liability / book value of total asset
CASHTA	cash / book value of total asset
MB	market value / book value of total asset
PRICE	log(minimum of firm's equity price or \$15)
SIGMA	volatility of equity return of the firm
RSIZE	log(market capitalisation of the firm / that of S&P 500 index)
EXRET	excess log-return
DtD	distance to default
Macro-economic predictor variables	
VIX	implied volatility option index
SRRATE	three-month T-bill rate
LRRATE	ten-year T-note rate
MKTRF	excess return on the market, Fama-French factor
SMB	small minus big, Fama-French factor
HML	high minus low, Fama-French factor

Table 3: The description of the predictor variables

where r is the firm's daily stock return, j is the daily time index, t is the quarterly time index, and N is the daily observation numbers within the quarter t.

To avoid the effect of extreme values and thus obtain an accurate and robust estimations, we winsorized all the firm-specific predictor variables at the 5-th and 95-th percentiles after processing the missing values.

3.4 Construction of distance to default

To construct the distance to default, we need the firm's market asset value and asset volatility. As both the market asset value and the asset volatility are not observable, we use a call option formula to work out them.

According to the Black-Scholes and Merton model, the market asset value A_t follows the Geometric Brownian motion, $\frac{dA_t}{A_t} = \mu_A dt + \sigma_A dW_t$, and the equity value of the firm, E_t, can be viewed as a call option on A_t with the strike price as the face value of debt L_t. The face value of debt is conventionally obtained by a proxy of the short-term debt plus half of the long-term debt. Hence, the call option formula is

$$E_t = A_t N(d_1) - L_t e^{-rT} N(d_2) \ , \tag{7}$$

where $N(.)$ is the cumulative distribution function of the standard normal distribution, r is the risk-free return, T is the time to maturity which is assumed to be 1 year, and

$$d_1 = \frac{\ln(\frac{A_t}{L_t}) + (r + \frac{1}{2}\sigma_A^2)T}{\sigma_A\sqrt{T}} \ , \tag{8}$$

$$d_2 = d_1 - \sigma_A\sqrt{T} \ . \tag{9}$$

Using Eq.(7)-(9), we can back out the market asset value A_t and asset volatility σ_A from the market equity and accounting information in two ways.

One way (denoted by Method-1 hereafter) to back out A_t and σ_A is through an iterative algorithm including the following five steps (Vassalou and Xing, 2004).

1. Set the initial value of A_t to be the sum of equity E_t and the firm's short-term liability and the long-term liability; set the initial value of σ_A to be the standard deviation of daily initial asset value from past 12 months; and use the one-year Treasury bill rate as the risk free return r.

2. For each trading day of past 12 months, use Eq.(7)-(9) to get the daily value of A_t; compute the standard deviation of A_t over past 12 months; take this standard deviation as σ_A for the next iteration.

3. Continue the procedure until the values of σ_A from two consecutive iterations converge at a tolerance level, say, 10^{-4}. Once the converged value of σ_A is obtained, Eq.(7)-(9) is used to back out A_t.

4. μ_A is obtained by taking the mean of the daily value of log return, $\ln A_t - \ln A_{t-1}$.

5. If in the Step 1-3 the quarterly data are processed and the size of the time window is kept as 4 quarters, then we can obtain the estimate of the quarterly value of σ_A and back out the quarterly asset value of A_t.

It follows that the distance to default can be obtained as

$$DtD_t = \frac{\ln(\frac{A_t}{L_t}) + (\mu_A - \frac{1}{2}\sigma_A^2)T}{\sigma_A\sqrt{T}} \ . \tag{10}$$

An alternative way (denoted by Method-2 hereafter) to back out A_t and σ_A is through simultaneously solving two equations for these two unknowns. Campbell et al. (2005) take the same equations as Eq.(7)-(9), and use the optimal hedge equation as another equation

$$SIGMA_t = \sigma_A N(d_1) \frac{A_t}{E_t} \ , \tag{11}$$

where $SIGMA_t$ is the firm's equity volatility.

The distance to default can be then obtained as

$$DtD_t = \frac{\ln(\frac{A_t}{L_t}) + (0.06 + r - \frac{1}{2}\sigma_A^2)T}{\sigma_A \sqrt{T}} \ , \tag{12}$$

where the equity premium directly takes the value of 0.06 instead of being estimated by the average firms' daily returns as with the Method-1, which might be a noisy estimate.

The Method-2 avoids keeping a rolling window of the previous observations, and thus it works for incomplete data sets. Moreover, the Method-2 does not require the daily stock price, and thus it facilitates the preparation of the data. In this report, we use the Method-2 to calculate the distance to default.

For convenience, we hereafter refer to a row in the quarterly-frequency data set as a firm-quarter, a data column as a variable, a cross intersect of the row and the column as an entry, and the quarter in which the firm is filed for bankruptcy as an event-quarter.

Before data processing for the missing values, we first take the following steps to help clean data.

1. Take a one-quarter lag for all the predictor variables to ensure that the predictor information is available before the quarter over which the probability of bankruptcy is to be estimated; and thus the firms with only the first firm-quarter data are removed, giving rise to a decrease in the total number of firms to 1,713 and the number of firms filing for bankruptcy to 65.

2. When any accounting variable at the 4th quarter of a year Y for the firm i is missing, fill in the value with its corresponding annual value of the year Y if the firm's annual data is not missing.

3. Replace any occurrence of zero values in the firm-specific accounting variables as missing, because the zero values were apparently misrepresented for our accounting and stock price variables, and the data resources did not provide explanations for the occurrence of such zero values.

4 Empirical studies

In this section, we shall apply three methods, the list-wise deleting, the closest-value imputation and the multiple imputation, to our sample for the missing values, and investigate the impact of these methods on the parameter-estimation results, with or without variable selection.

4.1 Empirical studies (ES-1) with list-wise deleting

The simplest method to process the missing values is to list-wisely delete the firm-quarters which have missing entries. For our sample, the list-wise deleting is performed through the following steps.

1. We delete any firm-quarters with missing entries.

2. For a firm filing for bankruptcy, if its event-quarter has missing entries and thus has been deleted in the last step, we remove such a firm from our sample.

3. For a firm filing for bankruptcy, we delete any of its firm-quarters after the even-quarter.

We observed that, in our data set, some of the empty firm-quarters were generated from automatically spanning the data to cover the whole sample period while being downloaded from the data resources, so that, for a firm-quarter, even if its entries were all missing, it still appeared in the sample. In addition, for some firms filing for bankruptcy, non-missing entries may reappear several quarters after their event-quarters, as the firms were re-listed in the market. For such firms, we only remove these reappearing observations. The intuition is that the firm is not expected to have any information in our sample after its event-quarter and we are to forecast the bankruptcy from data before the event, rather than back out the bankruptcy from the data after the event.

In a nutshell, after the above processing, we have in total 45,460 firm-quarters representing 1,637 firms.

Parameter	Estimate	Std Error	Wald χ^2	$Pr > \chi^2$
Intercept	-16.2758	4.3434	14.0418	0.0002
NITA	-8.7861	8.2405	1.1368	0.2863
TLTA	8.1357	2.4747	10.8082	0.0010
CASHTA	-1.7476	1.8793	0.8648	0.3524
PRICE	-1.1990	0.9888	1.4704	0.2253
MB	-0.3045	0.2038	2.2328	0.1351
RSIZE	-0.4215	0.3226	1.7068	0.1914
EXRET	-0.4195	0.2350	3.1878	0.0742
SIGMA	5.2340	1.6358	10.2377	0.0014
DtD	0.5294	0.1883	7.9037	0.0049
VIX	-0.0165	0.0360	0.2103	0.6465
SRRATE	-10.6599	21.3551	0.2492	0.6177
LRRATE	8.2882	43.3919	0.0365	0.8485
MKTRF	-0.0402	0.0683	0.3461	0.5563
SMB	0.0733	0.0724	1.0255	0.3112
HML	-0.0297	0.0881	0.1134	0.7363

Table 4: The parameter estimates for the full model for the ES-1.

The estimation results for the full model with all the predictor variables are shown in Table 4. Three predictor variables, TLTA, SIGMA and DtD are statistically significant at the 5% significance level. Reflecting the firm's leverage and stock price volatility, TLTA and SIGMA enter the model with expected signs. However, DtD, the volatility-adjusted measure of leverage, has an unexpected positive sign.

Parameter	Estimate	Std Error	Wald χ^2	$Pr > \chi^2$
Intercept	-14.3934	3.4603	17.3017	$< .0001$
TLTA	9.0198	2.4608	13.4348	0.0002
PRICE	-2.4598	0.8487	8.4000	0.0038
EXRET	-0.4769	0.2293	4.3246	0.0376
SIGMA	5.7652	1.5954	13.0586	0.0003
DtD	0.5894	0.1965	8.9985	0.0027

Table 5: The parameter estimates for the ES-1 with the predictor variables by stepwise model selection.

Furthermore, from the 15 predictor variables, a subset of 5 predictor variables, TLTA, PRICE, EXRET, SIGMA and DtD, were selected by a stepwise model selection. The estimation results for the new model are shown in Table 5. All estimates of the predictor variables have expected signs, expect for that of DtD.

4.2 Empirical studies (ES-2) with closest-value imputation

Another simple way to process the missing values is to do a simple imputation of the missing entries with the closest non-missing entries. For our sample, such a closest-value imputation is performed through the following steps.

1. Code all the missing entries as NaN.

2. For each firm, if a missing entry is between any two non-missing entries with regard to an accounting variable, then this missing entry (NaN) is replaced with -99999.

3. Remove the firm-quarters whose missing entries are still shown as NaN. In fact, these missing entries are either before the first non-missing entries or later than the last non-missing entries.

4. Replace -99999 as NaN.

5. For each firm, replace NaN with the closest non-missing entries later than them, then replace the remaining NaN with the closest non-missing entries before them to ensure that all NaN are imputed.

In a nutshell, after the above processing, we have in total 59,378 firm-quarters representing 1,667 firms.

Using all predictor variables, we estimated the full model for the ES-2. The estimation results are shown in Table 6. Compared with the estimates of the full model for the ES-1 (in Table 4), we observe that, for the ES-2, NITA and EXRET become statistically significant and with the expected signs. The changes in statistical significance is in line with the economical significance. However, DtD becomes nonsignificant. Meanwhile, all estimates of the predictor variables remain the same signs as those for the ES-1, except for that of DtD, which changes to the expected negative sign.

Parameter	Estimate	Std Error	Wald χ^2	$Pr > \chi^2$
Intercept	-11.2265	2.6090	18.5153	$< .0001$
NITA	-10.3210	4.9988	4.2631	0.0389
TLTA	5.8879	1.4365	16.7998	$< .0001$
CASHTA	-2.5708	1.4465	3.1586	0.0755
PRICE	-1.0319	0.5456	3.5776	0.0586
MB	-0.1412	0.0926	2.3249	0.1273
RSIZE	-0.1708	0.2053	0.6919	0.4055
EXRET	-0.4129	0.1508	7.5026	0.0062
SIGMA	3.3053	0.8421	15.4063	$< .0001$
DtD	-0.0312	0.0871	0.1282	0.7203
VIX	-0.0182	0.0262	0.4840	0.4866
SRRATE	-11.4882	15.4431	0.5534	0.4569
LRRATE	14.1017	30.5004	0.2138	0.6438
MKTRF	-0.0233	0.0479	0.2354	0.6275
SMB	0.0375	0.0483	0.6036	0.4372
HML	-0.0265	0.0615	0.1850	0.6671

Table 6: The parameter estimates for the full model for the ES-2

Parameter	Estimate	Std Error	Wald χ^2	$Pr > \chi^2$
Intercept	-11.1471	1.8742	35.3761	$< .0001$
TLTA	7.1606	1.4145	25.6274	$< .0001$
PRICE	-1.6014	0.4281	13.9899	0.0002
EXRET	-0.4692	0.1511	9.6376	0.0019
SIGMA	3.6498	0.7593	23.1079	$< .0001$

Table 7: The parameter estimates for the ES-2 with the predictor variables by stepwise model selection

Furthermore, from the 15 predictor variables, a subset of 4 predictor variables, TLTA, PRICE, EXRET and SIGMA, were selected by a stepwise model selection. The estimation results for the new model are shown in Table 7.

Compared with the corresponding estimates in Table 5 for the ES-1, DtD is not selected while other four predictor variables remain the same. In addition, TLTA and SIGMA are more statistically significant and the magnitudes of their estimates are increased. The possible reason of the removal of DtD could be because the information about the firm's volatility and leverage has already been reflected by TLTA and SIGMA in our model, and TLTA and SIGMA could be better measures than DtD for the firm's leverage and volatility.

4.3 Empirical studies (ES-3) with multiple imputation – our best model

The third way to process the missing values is to impute the missing entries by the multiple imputation. The multiple imputation has been widely used for incomplete data analysis in biostatistics. The basic idea is first to obtain m complete data sets through imputing the missing entries m times, then to obtain m estimation results for the m complete data sets, and finally to obtain the estimation results by combining the m estimates. The merit of the multiple imputation is that it considers the uncertainty about the right value to impute. Effectively incorporating uncertainty caused by the missing entries, the statistical inference is valid for the final estimation results (Rubin, 1987).

The approach to obtaining the m complete data sets depends on the pattern of missing data, which could be either monotonic or arbitrary, with the assumption of missing at random.

Given a data set with variables $X_1, X_2, \cdots, X_j, \cdots, X_p$ (arranged in this order), if, for an observation, the fact that X_j is missing means the values of the following variables from X_{j+1} to X_p are all missing, then this data set has a monotonic missing pattern. For a monotonic missing pattern, either a regression or a nonparametric method can be used (Rubin, 1987) to impute the missing entries.

An arbitrary missing pattern is all other missing patterns rather than the monotonic missing pattern. For a data set with the arbitrary missing pattern,

a Markov Chain Monte Carlo method with an assumption of multivariate normality can be used (Schafer, 1997) to impute the missing entries.

The approach to combining the m estimates of the m complete data sets is as the following (SAS Institute Inc., 1999). Given the point estimate $\hat{Q}_i$ and its variance estimate $\hat{U}_i$ for a parameter Q from the i-th imputed data set, $i = 1, \cdots, m$, the combined point estimate $\bar{Q}$ is the average of the $\hat{Q}_1, \cdots, \hat{Q}_m$ such that

$$\bar{Q} = \frac{1}{m} \sum_{i=1}^{m} \hat{Q}_i \ . \tag{13}$$

Its total variance estimate T is calculated as the weighted sum of the so-called within-imputation variance $\bar{U}$ and between-imputation variance B as

$$T = \bar{U} + (1 + \frac{1}{m})B \ , \tag{14}$$

where the within-imputation variance $\bar{U}$ is the average of the $\hat{U_1}, \cdots, \hat{U_m}$ such that

$$\bar{U} = \frac{1}{m} \sum_{i=1}^{m} \hat{U_i} \ , \tag{15}$$

and the between-imputation variance B is given by

$$B = \frac{1}{m-1} \sum_{i=1}^{m} (\hat{Q}_i - \bar{Q})^2 \ . \tag{16}$$

If only doing a simple imputation such as with the ES-2, the inference of estimates for the variables are based solely on the $\hat{U_i}$; while doing multiple imputation, besides the within-imputation variance $\bar{U}$, we are able to exploit the between-imputation variance B; with multiple imputation, the confidence intervals of the estimates are narrowed.

For our sample, the multiple imputation is performed through the following six steps, with the first four steps the same as those for the ES-2.

1. Code all the missing entries as NaN.

2. For each firm, if a missing entry is between any two non-missing entries with regard to an accounting variable, then this missing entry (NaN) is replaced with -99999.

3. Remove the firm-quarters whose missing entries are still shown as NaN. In fact, these missing entries are either before the first non-missing entries or later than the last non-missing entries.

4. Replace -99999 as NaN.

5. Test for the normality of each predictor variable and take logarithmic transform for the non-normality predictor variables.

6. Obtain $m = 10$ data sets using the MI procedure of SAS for the missing entries coded as NaN and then inverse the log-transformed predictor variables.

In a nutshell, after the above processing, we have in total 59,716 firm-quarters representing 1,673 firms.

Parameter	Estimate	Std Error	LCL	UCL	t	$Pr > \|t\|$
Intercept	-10.2310	2.3315	-14.8070	-5.6550	-4.39	$< .0001$
NITA	-11.8802	4.4913	-20.7504	-3.0100	-2.65	0.0090
TLTA	4.7511	1.2512	2.2688	7.2335	3.80	0.0003
CASHTA	-1.3481	1.1466	-3.6022	0.9059	-1.18	0.2404
PRICE	-0.9110	0.4180	-1.7357	-0.0863	-2.18	0.0306
MB	-0.0418	0.0392	-0.1207	0.0370	-1.07	0.2908
RSIZE	-0.2616	0.1787	-0.6128	0.090	-1.46	0.1441
EXRET	-0.2394	0.1186	-0.4721	-0.0068	-2.02	0.0437
SIGMA	1.6400	0.5890	0.4756	2.8043	2.78	0.0061
DtD	-0.0208	0.2014	-0.4163	0.3748	-0.10	0.9178
VIX	-0.0041	0.0248	-0.0527	0.0445	-0.17	0.8687
SRRATE	-12.4174	15.2311	-42.2705	17.4358	-0.82	0.4149
LRRATE	1.7469	29.7873	-56.6352	60.1289	0.06	0.9532
MKTRF	-0.0363	0.0460	-0.1265	0.0538	-0.79	0.4290
SMB	0.0716	0.0484	-0.0233	0.1665	1.48	0.1392
HML	-0.0026	0.0600	-0.1201	0.1149	-0.04	0.9649

Table 8: The parameter estimates for the full model for the ES-3.

The estimation results of the full model for the ES-3 using all the predictor variables are shown in Table 8. Compared with the estimates of the full model for the ES-2 (in Table 6), we observe the following. First, there are five predictor variables, NITA, TLTA, EXRET, SIGMA and PRICE, statistically significant at the 5% level for the ES-3. Nonsignificant for the ES-2, the

variable PRICE becomes significant and keeps the expected negative sign for the ES-3. Secondly, DtD is again nonsignificant.

The intervals of 95% confidence limit of the full model for the ES-3 are calculated as the lower confidence limit (LCL) and the upper confidence limit (UCL) and shown in Table 8. We observe that all the corresponding estimates in Table 6 for the ES-2 fall into that confident intervals, except for those of MB and SIGMA. Meanwhile, all estimates of the predictor variables remain the same signs as those in the ES-2. In addition, the macroeconomic variables are all nonsignificant for the ES-3, the same as with the ES-1 and the ES-2.

Variables	NITA	TLTA	PRICE	MB	EXRET	SIGMA	SRRATE
Frequency	10	10	10	4	7	10	1

Table 9: The frequencies of the most-frequent significant predictor variables (MFSPV) from the stepwise model selections for the ES-3 and that for the variable SRRATE.

Parameter	Estimate	Std Error	LCL	UCL	t	$Pr > \|t\|$
Intercept	-9.3022	1.3226	-11.8993	-6.7051	-7.03	$< .0001$
NITA	-10.3148	4.1458	-18.4963	-2.1332	-2.49	0.0138
TLTA	4.8065	1.0734	2.6910	6.9220	4.48	$< .0001$
PRICE	-1.3812	0.3448	-2.0588	-0.7036	-4.01	$< .0001$
EXRET	-0.2514	0.1150	-0.4770	-0.0258	-2.18	0.0290
SIGMA	1.8190	0.4387	0.9545	2.6835	4.15	$< .00014$

Table 10: The parameter estimates for the ES-3 with the predictor variables by stepwise model selection.

We perform stepwise model selection for the 10 imputed data sets, respectively. As the 10 stepwise model selections give us distinct subsets of the predictor variables, we choose NITA, TLTA, PRICE, EXRET and SIGMA as the most-frequent significant predictor variables (MFSPV) to analyse the 10 imputed data sets. The frequencies of the MFSPV obtained from these 10 model selections are shown in Table 9. The estimation results are reported in Table 10.

Compared with the corresponding estimates by stepwise model selection in Table 5 for the ES-1 and Table 7 for the ES-2, the predictor variable

NITA enters into the model for the first time with statistical significance at the 5% level, while the other four variables TLTA, PRICE, EXRET and SIGMA remain in the model. With the expected negative sign and the high magnitude of the estimate, NITA, as a profitability measure, becomes the most influence predictor variable in forecasting firms filing for bankruptcy instead of the leverage measure of TLTA. In this sense, ES-3 reestablishes the role of NITA in forecasting firms filing for bankruptcy. We argue that the model with these five predictor variables, NITA, TLTA, PRICE, EXRET and SIGMA is the best model for our data set, based on the above empirical studies.

5 Empirical comparison based on a model by Campbell et al. (2005)

In the section, we use a model (denoted by Campbell-M hereafter), proposed in the column (1) of Table 3 of Campbell et al. (2005), as a benchmark to compare the performance of the three methods of processing the missing values. The Campbell-M, using NITA, TLTA, RSIZE, EXRET and SIGMA as predictor variables, has a similar specification to our models, as shown in Tables 5, 7 and 10, obtained from stepwise model selections, respectively. We applied the Campbell-M to our data sets generated for ES-1, ES-2 and ES-3, respectively. The results are shown as follows.

5.1 ES-1

Compared with those in Campbell et al. (2005), our results, as listed in Table 11 for the ES-1, show the same signs but different magnitudes of estimates, and NITA is nonsignificant here. Note that when constructing the predictor variables, Campbell et al. (2005) adjusted the book value of total assets by adding 10% of the difference between market and book equity to them, whereas we do not make such an adjustment.

5.2 ES-2

Compared with those in Table 11 for the ES-1, our results, as listed in Table 12 for the ES-2, show that, although still nonsignificant at the 5% level, NITA becomes significant at the 10% level.

Parameter	Estimate	Std Error	Wald χ^2	$Pr > \chi^2$
Intercept	-18.5870	2.5997	51.1198	$< .0001$
	-15.214	–	*39.45**	**
NITA	-8.2006	7.6499	1.1491	0.2837
	-14.05	–	*16.03**	**
TLTA	9.2075	2.5011	13.5521	0.0002
	5.378	–	*25.91**	**
RSIZE	-0.7467	0.2915	6.5607	0.0104
	-0.188	–	*5.56**	**
EXRET	-0.4665	0.2315	4.0606	0.0439
	-3.297	–	*12.12**	**
SIGMA	3.6198	1.1284	10.2907	0.0013
	2.148	–	*16.40**	**

Table 11: The parameter estimates for the ES-1 with the predictor variables in Campbell et al. (2005). Contents in italic are the results in the column (1) of Table 3 of Campbell et al. (2005), where the value with * is the absolute value of z statistics; ** represents statistical significance at 1% level.

Parameter	Estimate	Std Error	Wald χ^2	$Pr > \chi^2$
Intercept	-16.0864	1.5358	109.7056	$< .0001$
	-15.214	–	*39.45**	**
NITA	-8.3294	4.6490	3.2100	0.0732
	-14.05	–	*16.03**	**
TLTA	6.8877	1.4001	24.2014	$< .0001$
	5.378	–	*25.91**	**
RSIZE	-0.5233	0.1691	9.5782	0.0020
	-0.188	–	*5.56**	**
EXRET	-0.4715	0.1500	9.8793	0.0017
	-3.297	–	*12.12**	**
SIGMA	3.8771	0.7471	126.9305	$< .00014$
	2.148	–	*16.40**	**

Table 12: The parameter estimates for the ES-2 with the predictor variables in Campbell et al. (2005). Contents in italic are the results in the column (1) of Table 3 of Campbell et al. (2005), where the value with * is the absolute value of z statistics; ** represents statistical significance at 1% level.

5.3 ES-3

Parameter	Estimate	Std Error	LCL	UCL	t	$Pr > \|t\|$
Intercept	-13.7896	1.1363	-16.0202	-11.5591	-12.14	$< .0001$
	-15.214	–	–	–	*39.45**	**
NITA	-10.9615	4.0802	-19.0098	-2.9132	-2.69	0.0079
	-14.05	–	–	–	*16.03**	**
TLTA	4.7678	1.1099	2.5743	6.9613	4.30	$< .0001$
	5.378	–	–	–	*25.91**	**
RSIZE	-0.5462	0.1506	-0.8415	-0.2508	-3.63	0.0003
	-0.188	–	–	–	*5.56**	**
EXRET	-0.2717	0.1155	-0.4982	-0.0453	-2.35	0.0187
	-3.297	–	–	–	*12.12**	**
SIGMA	1.9172	0.4046	1.1224	2.7120	4.74	$< .0001$
	2.148	–	–	–	*16.40**	**

Table 13: The parameter estimates for the ES-3 with the predictor variables in Campbell et al. (2005). Contents in italic are the results in the column (1) of Table 3 of Campbell et al. (2005), where the value with * is the absolute value of z statistics; ** represents statistical significance at 1% level.

In contrast to those in Tables 11 for the ES-1 and 12 for the ES-2, our results, as listed in Table 13 for the ES-3, show that all the five predictor variables are statistically significant at the 5% level, which is in lines with with that of Campbell et al. (2005).

6 Conclusions and Discussion

In this report, we used three different methods, list-wise deleting (ES-1), closest-value imputation (ES-2) and multiple imputation (ES-3), to cope with the severe problem of missing values in our raw data set. Using the data sets obtained from the ES-1 and ES-2, we estimated the hazard model, and found that estimation results are not fully in lines with the literature in terms of statistical significance of the estimates of the predictor variables. However, when we used the data sets obtained from multiple imputation for the ES-3, our estimation results are conforming to the literature. Moreover, using stepwise model selection, we obtained models and parameter estimations for the ES-1, ES-2 and ES-3, respectively, and chose NITA, TLTA, PRICE, EXRET, and SIGMA as the predictor variables of our best model.

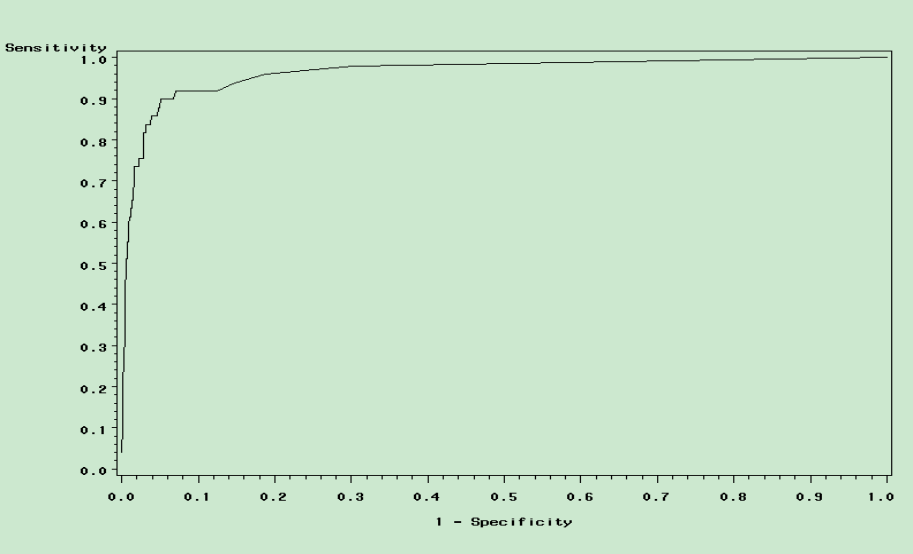

Figure 1: ROC plot for the best model.

In order to visualise the prediction performance of the model, we plot the receiver operating characteristic (ROC) curve in Figure 1. To its extreme, if the model predicts perfectly, the ROC curve passes through the point $(0, 1)$ and the area below the curve will be one; if the model has no predicting ability, the curve is a line at diagonal through the points $(0, 0)$ and $(1, 1)$. Figure 1, based on the average of the 10 multiple-imputed data sets obtained for the ES-3, shows that our model has a good prediction performance.

From our investigation, we draw the following conclusion. Amongst the three methods we used to process the missing values, we empirically confirmed that the multiple imputation helped to correct the self-selection bias and outperformed the list-wise deleting and the closest-value imputation in the sense that the obtained results are consistent to those in the literature. Subsequently, in terms of the determinants of forecasting the probability of bankruptcy, we empirically found that the predictor variables NITA, TLTA, PRICE, EXRET and SIGMA were the most promising ones over our sample period of 1995–2005.

Furthermore, we suggest that three issues merit further investigation as follows.

The first issue is: although following the way of assuming and using a

hazard model to forecast the probability of bankruptcy as with the existing literature (Shumway, 2001; Campbell et al., 2005), we think that there are still some aspects worth discussion. These aspects include: each firm-quarter is treated as an independent observation although the data are panel data; the proportion of the firms filing for bankruptcy is very small so that the two classes for logistic regression is very unbalanced, which makes the prediction less reliable; and the random effect resulted from the discrepancy between individual firms are not considered explicitly in the model.

The second issue is: after we obtained the physical default intensity (λ^P), in order to further explore our second question as asked in the beginning of Section 1, we need to "back out" the risk-neutral default intensity (λ^Q). After we obtain the risk neutral default intensity, we are able to explore the relationship between risk-neutral default intensity and physical default intensity. The simplest way is to regress the physical default intensity with other variables on the risk-neutral default intensity, so that a rough magnitude of the default premium defined as λ^Q/λ^P can be obtained.

Currently, credit default swap (CDS) is trading in a huge volume and a high liquidity, so that it can be regarded as a purer security traded for credit risk than corporate bonds. Researchers have shown strong interest in seeking the credit risk premium through the CDS rate. Duffie et al. (2005) and Berndt et al. (2005) are the papers mostly related with this topic using the CDS rate, while Driessen (2005) uses corporate bond ratings.

The third issue is: although none of the macroeconomics predictor variables is significant in our model, we intend to explore the effect of the macroeconomic variables on the default risk premium, because the variables reflecting business cycles have been well documented affecting on the probability of bankruptcy (Duffie and Kenneth, 2003). In addition, we would like to incorporate industry effect into the determinants of default risk premium.

References

Altman, E., 1968. Financial ratios, discriminant analysis, and the prediction of corporate bankruptcy. Journal of Finance 23, 589–609.

Anderson, R., Sundaresan, S., 1996. Design and valuation of debt contracts. Reviews of Financial Studies 9, 37–68.

Berndt, A., Douglas, R., Duffie, D., Ferguson, M., Schranz, D., 2005. Mea-

suring default risk premia from default swap rates and EDFs. BIS working paper .

Black, F., Cox, J., 1976. Valuing corporate securities: Some effects of bond indenture provisions. Journal of Finance 31, 351–367.

Campbell, J., Hilscher, J., Szilagyi, J., 2005. In search of distress risk. Working paper .

Chava, S., Jarrow, R. A., 2004. Bankruptcy prediction with industry effects. Review of Finance 8, 537–569.

Driessen, J., 2005. Is default event risk priced in corporate bonds? Reviews of Financial Studies 18, 165–195.

Duffie, D., Kenneth, J. S., 2003. Credit Risk: Pricing, Measurement, and Management. Princeton University Press, Princeton and Oxford.

Duffie, D., Saita, L., Wang, K., 2005. Multi-period corporate default prediction with stochastic covariates. Working paper .

Duffie, D., Singleton, K. J., 1999. Modelling term structures of defaultable bonds. Reviews of Financial Studies 12, 687–720.

Jarrow, R. A., Turnbull, S. M., 1995. Pricing derivatives on financial securities subject to credit risk. Journal of Finance 50 (1), 53–86.

Leland, H., 1994. Corporate debt value, bond covenants, and optimal capital structure. Journal of Finance 49, 1213–1252.

Longstaff, F., Schwartz, E., 1995. A simple approach to valuing risky fixed and floating rate debt. Journal of Finance 50 (3), 789–821.

Merton, R., 1974. On the pricing of corporate debt: The risk structure of interest rates. Journal of Finance 29 (2), 449–470.

Rubin, D. B., 1987. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, Inc, New York.

SAS Institute Inc., 1999. SAS OnlineDoc^{TM} Version 8. Cary, NC, USA.

Schafer, J. L., 1997. Analysis of Incomplete Multivariate Data. Chapman and Hall, New York.

Shumway, T., 2001. Forecasting bankruptcy more accurately: A simple hazard model. Journal of Business 74 (1), 101–124.

Vassalou, M., Xing, Y., 2004. Default risk in equity returns. Journal of Finance 59, 831–868.

Zmijewski, M. E., 1984. Methodological issues related to the estimation of financial distress prediction models. Journal of Accounting Research 22, 59–82.