大規模データベースを用いた信用リスク計測の問題点と対策 (変数選択とデータ量の関係)

山下 智志* 川口 昇†

概 要

本研究は，大規模データベースを用いた信用リスク計量に伴う問題点と，その対策方法についてまとめた．これまでの信用リスク計量モデルは，データの蓄積が進んでいないため，大規模データベースで計測することができなかった．そこで，CRD 運営協議会によって作られた，450,000 件，86 財務変数を持つデータベースを用いて，2 項ロジットモデルによりデフォルト確率の推定を行った．大規模データの推定では，計算時間が膨大になる．我々は，数値計算プログラムを改良して，計算時間を短縮した．そして，推定結果から選択される財務指標の傾向を分析することができた．

また，本研究では，データ量に対する最適なセグメント数についても検討した．一般に，業種や規模が信用リスクに与える影響を考慮する場合，データセグメント法を用いることが多い．データセグメント法では，セグメントにおけるデータが減少し，推定精度が悪化する場合がある．また，推定精度がよくなる場合には，オーバーフィッティングがおきて不安定な推定結果を得ることが多い．そこで，データ数とそれに含まれるデフォルト数を変化させて，そのデータ数がセグメントに分けられるほど十分なデータ数であるか分析した．その結果，データ数，それに含まれるデフォルト数，および変数選択候補数に関して，セグメントに分けるかどうか決定する表を得た．

*文部科学省統計数理研究所 助教授, CRD 運営協議会 顧問, 金融庁金融研究研修センター 特別研究員

†早稲田大学大学院 理工学研究科, 金融庁金融研究研修センター 専門研究員

本稿の執筆にあたり，データの使用を認めていただいた CRD 運営協議会に深謝したい．また，中央青山プライスウォーターハウスクーパース・フィナンシャル・アンド・リスクマネジメント 安川武彦氏をはじめ，金融庁金融研究研修センターにおけるワークショップの参加者各位から多くの有益なコメントを頂いた．なお，本稿は筆者の個人的な見解であり，金融庁の公式見解ではない．

目 次

第1章　はじめに

企業の倒産 (デフォルト) 判別，倒産 (デフォルト) 確率を分析する研究は，金融の分野において重要なテーマである．とくに，近年の銀行をはじめとする金融機関のリスク許容の低下に伴い，貸し出し先企業の信用力を正確に把握することが，金融機関の健全性に直接関係するようになった．そのため，貸し出し先企業の信用力を計測する統計モデルの開発が重要な課題となっている．

これまでの信用リスク計量モデルの研究では，大企業を対象としたモデルと，中小企業を対象としたモデルとで大きく分けられる．前者は，財務指標に関する情報開示が進み，徐々に企業の正確な情報が蓄積しつつある．また，株式公開企業や社債を発行・流通している企業に対しては，オプションアプローチなどの計量化モデルが開発され，統計モデル以外の方法でも信用力を計算することが可能である．一方，後者である中小企業に対しては，信用リスク計量モデルに必要なデータベースの蓄積が進んでいないため，財務データ情報の入手は一般的に困難であった．

このような背景から，中小企業に対する信用リスクデータを整備，蓄積しようという動きがある．中小企業信用リスク情報データベース運営協議会 (CRD 運営協議会) はこのような目的のために設立された団体の一つである．本研究では CRD 運営協議会によって作成された中小企業信用データベース[1]を用いて，以下のテーマについて研究を行った．

1. 中小企業信用データベースを用いて，2 項ロジット分析を行い，信用リスク計測に必要な経営指標の組み合わせを推測する．

2. 業種や規模が信用リスクに与える影響についてデータセグメント法で取り扱うことが一般的であるが，その有効性について，データ量とセグメント数の関係から検討する．

3. 1，2 の過程において判明する，大規模データを用いた信用リスクモデル作成における問題点を整理し，その対策について検討する．

第 2 章において，これまでの信用リスクの計測方法について言及する．第 3 章では，2 項ロジットモデルの推定方法とデータベースの加工方法および推定結果を示し，その分析をまとめた．第 4 章では，データ量とセグメント数の関係について，分析方法を示しその結果についてまとめた．第 5 章では，本研究の結果と意義についてまとめた．

[1]CRD 運営協議会で作成されたデータベースは，信用保証協会，政府系中小企業金融機関，および民間金融機関の与信データを，統一したフォームで収集・蓄積し作成された．このようにして作成されたデータを信用リスク分析の研究に活かし，その情報を会員が利用できるようになっている．

第2章　信用リスク計測方法についての既存研究

信用リスク計量化モデルは，統計学を基本とした統計モデルとオプション理論を用いたオプションアプローチモデルとに大きく分けられる．それらの代表的なモデルと方法論について説明する．

2.1　統計モデル

統計モデルでは，貸し出し先企業の財務データをもとにデフォルトの判定やデフォルト確率の推定を行う．代表的なモデルとして，判別分析，ロジットモデル，ハザードモデルがあげられる．

判別分析は，貸し出し先企業の財務データをもとに，デフォルトグループの群に属するか，非デフォルトグループの群に属するか判別する方法である．[Altman(1968)] による研究以来，信用リスク分析でよく使われる手法である．それぞれの群が多変量正規分布に従い，かつ，分散共分散行列の構造が等しいときには，財務データ $(x_1, x_2, \cdots, x_m)$ に重み付け $(b_1, b_2, \cdots, b_m)$ し，その線形結合によって与えられる式 (線形判別式) $z = b_1x_1 + b_2x_2 + \cdots + b_mx_m$ を用いてどちらの群に属するかを判別する．

判別分析ではデフォルトの判定を予測するが，その後の発展に伴い，デフォルトの判定だけでなく，デフォルト確率を正確に予測し信用リスクに見合うリターンを確保するという考え方が重要になってきた．ロジットモデルは財務データの線形結合 $(z = b_1x_1 + b_2x_2 + \cdots + b_mx_m)$ をロジスティック分布関数,

$$p(z) = \frac{\exp(z)}{1 + \exp(z)} \tag{2.1}$$

にあてはめて，デフォルト確率を推定するモデルである．ロジットモデルは，[森平，小松，湯山 (1996)] において信用組合のデフォルト予測が試みられ，良好な結果を得ており，デフォルト確率を推定する一般的方法の 1 つである．本研究では，ロジットモデルを用いてデフォルト確率の推定を行った．詳しい推定方法については第 3 章で改めて説明する．

また今日では，BIS 基準などの社債格付けに対する実務界の要請の変化や社債の複雑化に伴い，格付けデータの統計モデルも重要である．多重判別関数やオーダードロジットモデル (ordered logit model) は，財務データから格付けデータを推測するモデルである．格付けデータは順序をもった名義変数なので単純な非線形回帰モデルにより定式化できない．そこで，$AAA = 1, AA = 2$ とナンバリングし順序変数として扱う．多重判別関数とオーダードロジットモデルは，それぞれ判別分析・ロジットモデルを応用したモデルである．その詳細については，[Kaplan,Urwitz(1979)], [中山，森平 (1998)]，[安川，椿 (1999)] を参照.

判別分析，ロジットモデルはともに将来の一時点におけるデフォルト確率を求める手法である．しかし，デフォルトの可能性のある債券の現在価値を求める場合，将来のクーポンや元本支払いが

生じる「すべての」時点でのデフォルト確率を知る必要がある．ハザードモデルは，ハザード関数を用いてデフォルト確率の期間構造を推測するモデルである．ハザード関数とは，ある時点までデフォルトしないという条件の下で，次の瞬間にデフォルトが発生する率を時間 t に依存した関数 $h(t)$ として定義される関数である．[Lane, Looney,and Wansley(1986)] では，医療や信頼性工学で用いられる Cox の比例ハザードモデルを応用してデフォルト確率の期間構造を推測した．この研究以来，Cox の比例ハザードモデルはハザードモデルの代表的なモデルとなっている．Cox の比例ハザードモデルでは，ハザード関数 h を，時間を変数にもつある関数 $f(t)$ とデフォルトに影響を与える要因 $(x_1, x_2, \cdots, x_m)$ で構成される関数 $g(x_1, x_2, \cdots, x_m)$ を用いて，

$$h(t, x_1, x_2, \cdots, x_m) = f(t) \times g(t, x_1, x_2, \cdots, x_m) \tag{2.2}$$

として定義することが特徴である．

2.2 オプションアプローチモデル

オプションアプローチモデルは，市場データを用いてデフォルト確率を推定するという考え方である．[Merton(1974)] は，企業の資産は資本と負債で構成されていると仮定し，資本が負債を下回っている状況をデフォルトと規定した．企業資産価値を株式で代用し，これがある確率過程で変動した結果，ある閾値を下回った時点をデフォルト時点として考える．ここでオプション理論を用いて，リスク中立化法に基づくブラック・ショールズ式よりデフォルト確率の推定ができる．このモデルは構造モデルと呼ばれる．株式が上場されている企業には，市場データを得ることでリアルタイムにデフォルト確率が推定できる．また,「デフォルトは，対象となる企業の財務状況とは無関係に発生する」という考え方で，市場データから上述したハザード関数の時間変化 (デフォルト過程) を推測してデフォルト確率を推定するモデルがある．このモデルを外生変数モデルと呼ぶ．代表的なモデルとして，デフォルトリスクのない債券とデフォルトリスクのある債券の利回りの差 (スプレッド) と回収率によってデフォルト過程を推定するモデル [Duffie,Singleton(1999)]，吸収マルコフ連鎖を用いてデフォルト過程を推定するモデル [Jarrow,Lando,and Turnbull(1997)] がある．

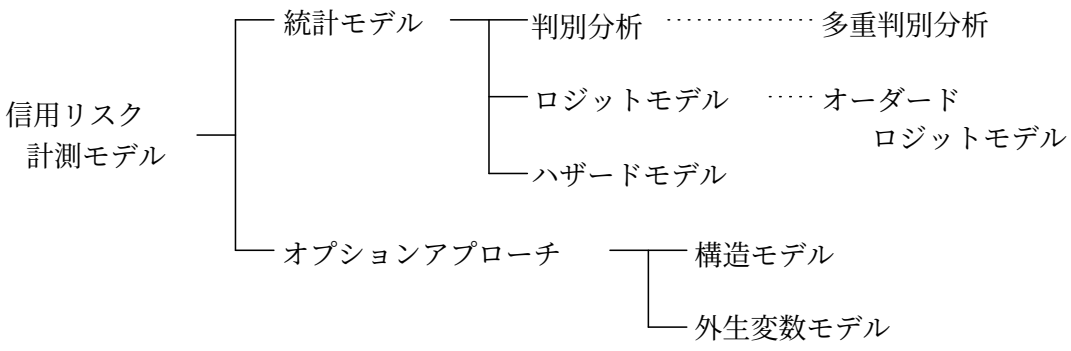

図 2.1: 信用リスク計量化モデルの分類

第3章　2項ロジットモデルを用いたデフォルト確率の推定

CRD 運営協議会によって作成された大量のデータを用いて，2 項ロジットモデルを用いたデフォルト確率の推定を行う．大量データを用いて推定を行う際の，データベースの加工方法や注意点を説明し，デフォルト確率に寄与する財務変数の組み合わせを見つける．また，データセグメント方法を用いて，全体データで推定を行うのとセグメントに分けて推定を行うのでは，どちらのほうが予測誤差の少ない推定ができるか比較分析を行う．

3.1　2項ロジットモデル

分析するデータの企業数を n 件，考慮している財務指標の数を m 個とする．各レコードの財務データ $(x_{i1}, x_{i2}, \cdots, x_{im})$ を用いて，線形結合

$$z = b_1 x_1 + b_2 x_2 + \cdots + b_m x_m \tag{3.1}$$

でスコア化する．z は信用スコアといい，デフォルト確率がこれによって決定されていると仮定する．デフォルト企業のデフォルト確率は，デフォルトしたという事後的な事実から考えて確率 1 である．また，非デフォルト企業はデフォルトしていない事実から，確率は 0 である．その結果をグラフにプロットしたのが図 3.1 である．

2 項ロジットモデルは，このスコアをロジスティック分布関数,

$$p(z) = \frac{\exp z}{1 + \exp z} \tag{3.2}$$

に代入してデフォルト確率を求める手法である．(図 3.2 参照)

係数ベクトル $\mathbf{b} = (b_1, b_2, \cdots, b_m)$ は，最尤法を用いて推定する．実際,「企業のデフォルトは独立して発生する」と仮定するならば，i 番目の企業のデフォルト確率を p_i として，尤度関数 $L(\mathbf{b})$ は,

$$L(\mathbf{b}) = \prod_{i=1}^{n} p_i^{\delta_i} (1 - p_i)^{1-\delta_i} \tag{3.3}$$

で与えられる．ここで，δ_i は,

$$\delta_i = \begin{cases} 1 & \text{(i番目の企業がデフォルトの場合)} \\ 0 & \text{(i番目の企業が非デフォルトの場合)} \end{cases}$$

なる関数である．これより対数尤度関数 $l(\mathbf{b})$ は,

$$l(\mathbf{b}) = \log\ L(\mathbf{b}) = \sum_{i=1}^{n} \delta_i \log p_i + (1-\delta_i)\log(1-p_i) \tag{3.4}$$

となり，最尤法では $l(\mathbf{b})$ が最大となるように係数ベクトル $\mathbf{b}$ を決定する．係数ベクトル $\mathbf{b}$ は，連立方程式

$$\frac{\partial l(\mathbf{b})}{\partial b_j} = \sum_{i=1}^{n} (p_i - \delta_i) x_{ij} = 0, \qquad j = 0, 1, 2, \cdots, m \tag{3.5}$$

を用いて求められるが，この連立方程式は一般に非線形方程式となるので，解析的に解くのは困難である．そのため数値計算法を用いて解くことが一般的である．しかし，大規模データを用いる場合，計算時間が膨大となり，計算方法の工夫が必要となる．[2]

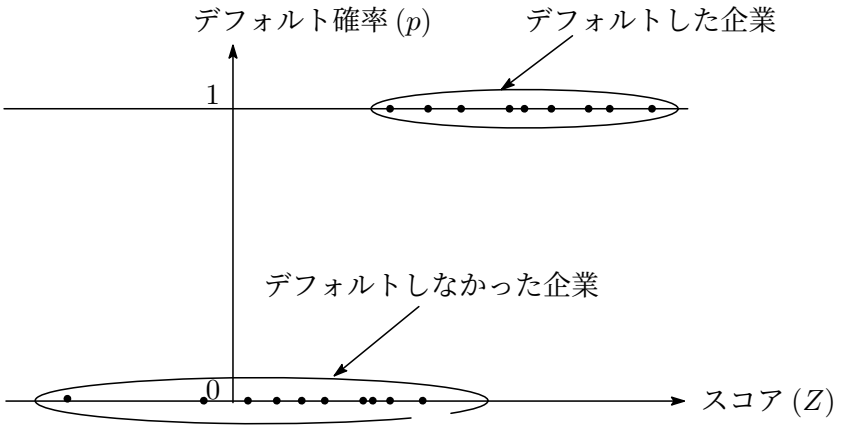

図 3.1: データのプロット

[2]一般に，市販されている統計用パッケージソフトの最適化ツールを用いれば，計算プログラムを作成する必要がない．本研究においても，統計用パッケージソフトを用いて分析を試みたが，大規模データのため，計算時間が膨大になった．そこで，計算時間短縮のため，C 言語を用いて計算用プログラムを作成した．この場合，最適化ルーチンに対する信頼性を失いかねないので，最適化ルーチンは参考文献 [Press,Teukolsky,Vetterling, and Flannery(1993)] をベースに，計算が高速化するようにプログラムを作成した．最適化ルーチンとしては，準ニュートン法を採用している．

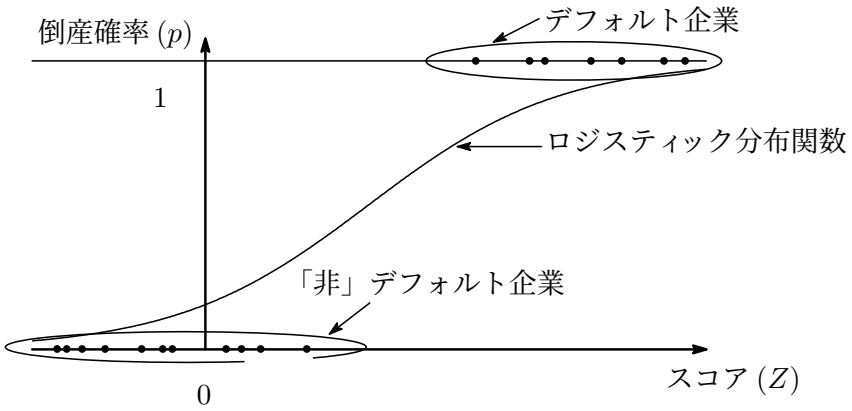

図 3.2: ロジットモデル

3.2 データベース加工方法について

CRD 運営協議会によって作成されたデータベースの企業数はのべ 948754 件，財務諸表項目数は 93 項目であった．以降，個々の企業のデータについては**レコード**，財務諸表項目を**フィールド**と呼ぶ．

3.2.1 デフォルトの定義

デフォルトは決算年月から向こう 1 年以内に「初回延滞発生」,「直近延滞発生」,「実質破綻発生」,「破綻発生」,「代弁発生」が生じた場合と定義した．

3.2.2 多くの欠損値を持つフィールドの削除

財務諸表項目のうち細かい内訳に関しては入力されていない企業が多数存在したので 1 割以上のデータが欠損しているフィールドについては削除した．

3.2.3 伸び率および財務比率変数の作成

財務諸表項目を用いて，伸び率 (異なる年度の，同一財務諸表項目の伸び率) と財務比率 (同一年度の，異なる財務諸表項目の演算から得られる指標) を作成する．これらの指標を求める際に 0 割りが生じた場合は，欠損値として扱った．また，財務諸表項目と同様に，1 割以上のデータが欠損している項目について削除を行った．

3.2.4　計算不能レコードの削除

伸び率を求めるには前年度データが必要なので，前年度データがないレコードは削除した．また，デフォルトの定義より，決算年月から向こう 1 年以内にデフォルトしたかどうか分からないレコードについても削除した．

3.2.5　外れ値の処理とデータ変換

パラメータを推定する際，元データの数値をそのまま使うと外れ値がパラメータ推定に影響し，推定結果をゆがめてしまう可能性がある．この場合，外れ値のみを一定の基準で削除する方法が考えられるが，基準値の与え方によって推定結果が大きく変わってしまうことが考えられる．そこで本研究では，元データの変換を施した．また，Box-Cox 変換などの非線形変換によって，データを平均に寄せる方法もあるが，本研究では採用しなかった．

変換方法としては，正規分布に当てはめる変換と，一様分布に当てはめる変換の 2 種類を考えた．それぞれの性質として，前者は外れ値の影響が比較的強く残る変換であり，後者は外れ値の影響が弱く，平均値まわりのデータが強く影響する変換方法であることがいえる．本研究では，後者の一様変換を採用した．

一様分布に当てはめる変換方法は，10 分位点を求め，小さい方から順に $1, 2, \cdots, 10$ とランク付けをした．また，レコードのあるフィールドに欠損値がある場合，そのフィールドについてはランク付けができないので，ランクの平均 5.5 を割り当てた．

3.2.6　多重共線性を持つフィールドの削除

相関が高い項目が含まれていると多重共線性の問題が生じる．そこで，相関係数を計算し 0.9 以上，または -0.9 以下の相関を持つものについてはどちらかの項目を削除した．どちらの項目を削除するかについては，複数の項目と相関を持つ項目を優先に削ること，より代表的な項目を残すことの 2 点を考慮し総合的な観点から選択した．

3.2.7　年度別の定数項 (フラグ) の作成

会計年度による影響を考慮して，定数項フラグを 1996 年度〜1999 年度のそれぞれにたてる．年度フラグは，会計年度に 1，それ以外には 0 をたてた．

3.2.8　欠損値フィールドの作成

欠損値による影響を考慮して，欠損値フィールドを設けた．欠損値フィールドはレコードにおける欠損値の個数をあてた．

以上の手順でデータベースの加工を行った結果，レコード数 442,686 件，フィールド数 86 項目として，データを整理した．

表 3.1: フィールド一覧

1	1996 年度	44	その他流動負債合計伸び率
2	1997 年度	45	固定負債合計伸び率
3	1998 年度	46	資本合計伸び率
4	1999 年度	47	売上高営業収益伸び率
5	欠損値に対するペナルティ	48	売上総利益伸び率
6	流動資産合計	49	営業利益伸び率
7	現金預金	50	支払利息割引料伸び率
8	受取手形	51	経常利益伸び率
9	売掛金	52	当期利益伸び率
10	棚卸資産合計	53	期末従業員数人伸び率
11	その他流動資産合計	54	総資本当期利益率
12	固定資産合計	55	（受取利息－支払利息）対総資本
13	土地	56	総資本経常利益率
14	その他固定資産	57	自己資本当期利益率
15	繰延資産	58	（受取利息－支払利息）対自己資本
16	資産合計	59	売上高総利益率
17	流動負債合計	60	総資本回転率
18	支払手形	61	固定資産回転率
19	買掛金	62	流動資産回転日数
20	短期借入金	63	売掛金回転日数
21	その他流動負債合計	64	買掛金回転日数
22	固定負債合計	65	棚卸資産回転日数
23	その他固定負債	66	売上高減価償却費
24	資本合計	67	売上高支払利息・割引料率
25	資本金	68	営業利益支払い率（逆数）
26	売上高営業収益	69	一人当たり売上高
27	売上総利益	70	一人当たり総資本
28	営業利益	71	キャッシュフロー
29	受取利息割引料配当金	72	固定負債キャッシュフロー倍率（逆数）
30	支払利息割引料	73	現預金比率
31	経常利益	74	支払準備率
32	当期利益	75	預借率
33	受取手形割引高	76	当座比率
34	受取手形裏書譲渡高	77	流動比率
35	減価償却実施額	78	正味運転資本額
36	期末従業員数	79	正味運転資本比率
37	流動資産合計伸び率	80	固定負債対有形固定資産比率
38	現金預金伸び率	81	固定比率（逆数）
39	その他流動資産合計伸び率	82	固定長期適合率（逆数）
40	固定資産合計伸び率	83	自己資本比率
41	その他固定資産伸び率	84	借入金依存度
42	資産合計伸び率	85	有利子負債利子率
43	流動負債合計伸び率	86	インタレストカバレッジ

3.3 変数選択方法

上述したように加工したデータベースには 86 フィールドの変数があり，どの変数がデフォルト確率に寄与するのかシステマティックに選び出さなければならない．この場合，多重共線性を避けるように注意しながら，有用な少数個の独立変数を精選することが重要である．変数選択を探索的に行う方法として,「変数増加法」,「変数減少法」,「ステップワイズ法（逐次選択法）」などがある．

本研究では,「ステップワイズ法」を試みた．変数選択の基本的な方法としては，適当な基準により説明変数を１つずつ加えていく「変数増加法」と，逆にすべての説明変数を用いた分析から１つずつ変数を減らしていく「変数減少法」がある．変数増加法では，一度取り込んだ変数は，新たな変数の追加によって，デフォルト確率への寄与がほとんど無くなっても除外されることがない．同様に，変数減少法において，一度除外された変数であっても，その後他の変数との関係でデフォルト確率に寄与することがある場合でも追加することができない．ステップワイズ法は，一度取り込んだ変数であっても，ある基準を満たさなくなった場合，その変数を除いたり，また，基準を満たした場合，もう一度採用することを繰り返し行うことで最適な変数選択をする方法である．変数を取り入れる，または取り外す基準としては AIC 基準を用いた．AIC 基準は

$$AIC = -2\,(\text{最大対数尤度}) + 2\,(\text{パラメータ数}) \tag{3.6}$$

出来上がった式の予測誤差の大きさを評価する基準であり，この値を小さくするようにモデルを決定する．

具体的には，まず変数増加法で AIC 基準を満たす変数を取り込み，すべての変数に対してこの操作を終えたら，次に変数減少法で AIC 基準を満たさなくなった変数を取り除く．この操作を繰り返し，取り入れる変数，または，取り除く変数が存在しなければ，その時点で取り入れた変数が最適であると決定する．[3]

3.4 全件データ推定とセグメントデータ推定との比較方法

全件データで推定した場合とセグメントデータで推定した場合とを AIC 基準を用いて比較できる．本節では，その方法について説明する．まず，セグメントに分けて推定したモデルが 1 つのモデルとして扱えることを説明する．その例として用いるセグメント数は n セグメントを考える．

いま，推定に用いる企業の全体データ数を N 件とし，k 番目のセグメント（以降，セグメント k と呼ぶ）のデータ数を N_k 件とする．セグメント k で変数選択を行った結果，得られた最大対数尤度と選択された変数の数をそれぞれ l_k，m_k，推定された係数を $\mathbf{b}_k = (b_{k1}, \cdots, b_{km_k})$，また，セグメント k に属するデータは $\mathbf{x}_{ik} = (x_{ik1}, \cdots, x_{ikm_k})$ とする．

[3]市販の統計用パッケージソフトを用いれば変数選択のプログラムが用意されており，具体的な方法を考える必要はない．注記 2 で示したように，本研究では C 言語で計算用プログラムを作成したので，変数選択の方法を考え，そのプログラムを作成した．

ここで，以下のような変数 $\mathbf{b}_{seg}$ と $\mathbf{x}_{seg}$ を考える．

$$
\begin{aligned}
\mathbf{b}_{seg} &= (b_{11},\cdots,b_{1m_1},b_{21},\cdots,b_{2m_2},\cdots\cdots,b_{n1},\cdots,b_{nm_n}) \\
\mathbf{x}_{seg} &= \begin{cases}
(\underbrace{x_{i11},\cdots,x_{i1m_1}}_{\text{セグメント 1 で選択}},0,\cdots,0,\cdots\cdots,0,\cdots,0) & \text{(セグメント 1 に属する場合)} \\
(0,\cdots,0,\underbrace{x_{i21},\cdots,x_{i2m_2}}_{\text{セグメント 2 で選択}},\cdots\cdots,0,\cdots,0) & \text{(セグメント 2 に属する場合)} \\
\qquad\qquad\qquad\vdots & \\
(0,\cdots,0,0,\cdots,0,\cdots\cdots,\underbrace{x_{in1},\cdots,x_{inm_n}}_{\text{セグメント n で選択}}) & \text{(セグメント } n \text{ に属する場合)}
\end{cases}
\end{aligned}
$$

このとき，k 番目のデータに対して，

$$
\begin{aligned}
\mathbf{b}_k\cdot\mathbf{x}_{ik}^T &= (b_{k1},\cdots,b_{km_k})\cdot(x_{ik1},x_{ik2},\cdots,x_{ikm_k})^T \\
&= (b_{11},\cdots,b_{1m_1},\cdots\cdots,b_{n1},\cdots,b_{nm_n})\cdot(0,\cdots,0,\underbrace{x_{ik1},\cdots,x_{ikm_n}}_{\text{セグメント } k \text{ で選択}},0,\cdots,0)^T \\
&= \mathbf{b}_{seg}\cdot\mathbf{x}_{seg}^T
\end{aligned}
$$

であることに注意すれば，全セグメントとの和 l_{seg} は，

$$
\begin{aligned}
l_{seg} &= \sum_{k=1}^{n} l_k \\
&= \sum_{k=1}^{n}\sum_{i=1}^{N_k}\left(\delta_i\log\frac{\exp(\mathbf{b}_k\mathbf{x}_{ik}^T)}{1+\exp(\mathbf{b}_k\mathbf{x}_{ik}^T)}(1-\delta_i)\log\frac{1}{1+\exp(\mathbf{b}_k\mathbf{x}_{ik}^T)}\right) \\
&= \sum_{i=1}^{N}\left(\delta_i\log\frac{\exp(\mathbf{b}_{seg}\mathbf{x}_{seg}^T)}{1+\mathbf{b}_{seg}\mathbf{x}_{seg}^T}+(1-\delta_i)\log\frac{1}{1+\mathbf{b}_{seg}\mathbf{x}_{seg}^T}\right)
\end{aligned}
$$

と変形することができ，ひとつのモデルとみなせる．（ここで，添え字 T はベクトルの転置を表す）

この場合の AIC 基準 AIC_{seg} は，

$$
\begin{aligned}
AIC_{seg} &= -2\cdot\sum_{k=1}^{n} l_k + 2\cdot\sum_{k=1}^{n} m_k \\
&= \sum_{k=1}^{n}\left(-2\cdot l_k + 2\cdot m_k\right)
\end{aligned}
$$

となるので，セグメントごとに算出される AIC 基準の和に等しくなる．したがって，全体データで算出された AIC 基準と，以上のように算出したセグメントデータでの AIC 基準とを比較して，どちらが予測誤差の小さいモデルか判断できる．

3.5 全件データでの推定結果

全てのデータをもとに推定した結果を分析する．

3.5.1 デフォルト確率に寄与するパラメータ

係数の推定結果から分析する．表 3.2 は選択されたフィールドの表である．この表で係数の値が正の場合，デフォルト確率を低めるように働き，係数の値が負の場合は，デフォルト確率を高めるように働く．係数の大きさからデフォルト確率に寄与する変数として，自己資本比率，預借率，借入金依存度，期末従業員数，支払準備率があげられる．逆に，欠損値に対するペナルティ，営業利益伸び率，総資産回転率は係数が小さいと同時に，t-値を見ると有意でないことがわかる．これは，変数選択において AIC を用いたために，将来の予測誤差は低くなるとして変数を取り込んだが，変数そのものの有意性について考慮しなかったためである．

3.5.2 パラメータの符号条件

一般にロジットモデルでは，用いる説明変数が高い相関を持つ場合，多重共線性の問題が発生する．そのため，本研究では推定を行う前に 2 変数間での相関が高い変数を削除した．しかし，財務指標では，2 変数間の相関がなくても複数の変数の組み合わせによって相関が高くなり，多重共線性が発生する場合がある．そのような変数の推定結果や符号条件は信用できず，また，その結果を用いて考察することも有意であるとはいえない．しかし，用いる財務データが，そのような相関を持つデータである限り，多重共線性が推定結果に影響しない．例えば，推定結果では営業利益の係数が負の値で推定されているが，これは一般的な認識と異なる．これは次のように説明ができる．

営業利益は概ね次のように分解できる；

$$(\text{営業利益}) = (\text{経常利益}) - (\text{受取利息}) + (\text{支払利息})$$

これを移項して，

$$(\text{支払利息}) = (\text{営業利益}) + (\text{受取利息}) - (\text{経常利益}) \tag{3.7}$$

が得られる. ここで，支払利息，営業利益，受取利息，経常利益をそれぞれ，x_1, x_2, x_3, x_4 としてスコア z を説明しようとすると，

$$z = \beta_1 x_1 + (\text{その他の変数}) \tag{3.8}$$

$$z = \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + (\text{その他の変数}) \tag{3.9}$$

となる．ここで，$\beta_1, \beta_2, \beta_3, \beta_4$ はそれぞれの係数とする．推定結果より，支払利息の係数 β_1 は負である．一方，(3.7) 式を考慮すると，(3.8) 式の $\beta_1 x_1$ と (3.9) 式の $\beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ は同じ符号を持つという条件が成り立つ．経常利益，受取利息はともに推定結果では正の係数であった．従って，この条件を満たすためには，営業利益の係数が負になると推測できる．この影響を受け，営業利益を使って求められる営業利益支払率やインタレストカバレッジなどの係数の符号が逆転してしまう可能性も考えられる．推定結果では，インタレストカバレッジも負の係数として推定されたが，この影響を受けていると考えられる．

また，営業利益の符号が負で推定されていても，経常利益は正に推定されており，デフォルト確率に寄与する大きさを比較しても，経常利益のほうが効いていることがわかる．

以上のことから，符号条件に関して直感と整合しない変数が存在したとしても，それが矛盾であるとはいえない．むしろ，個々の変数の符号条件が一般的な認識と異なっていても，AIC 基準を用いて推定したこの結果は，予測誤差が最小のモデルとして選ばれた最良のモデルであると考えるべきである．[4]

3.5.3 非線形に効く指標

ロジットモデルでは，スコア z を財務データ $(x_1, x_2, \cdots, x_m)$ の線形結合 $z = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$ で与えた．従って，スコア z は単調に増加 (または減少) するので非線形効果を表現できない．そのため，値が一定の範囲である場合に健全であると経験的に考えられる財務指標について，2 項ロジットモデルではその指標のデフォルト傾向を十分に反映できない．

推定結果では，借入金依存度の係数が正に推定されている．これは「借入金の増加はキャッシュを増加させる」と考えることで直感と整合性が合うが,「限度を超えるとデフォルトに陥る可能性が高まる」ことも考えられるので，単純に多いから安全であるとはいえない．

現在，このような問題を解消するためにスコア z に 2 次項を含んだロジットモデルなどの新しいモデルが考えられている．ここで詳細を述べるのは本研究の範囲を超えるので，参考文献 [今野，武 (2001)] を参照とする．

3.5.4 推定精度の検討

推定精度について，図 3.3 の CAP(Cumulative Accuracy Profile) 曲線および AR(Accuracy Ratio) を用いて検討した．いま，推定に用いた企業のデータ数を N 件，そのうち実際にデフォルトした企業を n 件とする．CAP 曲線は，横軸に推計デフォルト確率の高い上位 x 件の全体に占める割合 x/N を，縦軸に推計デフォルト確率の高い上位 x 件のうち実際にデフォルトした件数 n_x の割合 n_x/n をプロットする．例えば，表 3.3 を見ると推計デフォルト確率が高い上位 10% の企業には，デフォルト企業の約 60% が含まれていることがわかる．

ロジットモデルに全く説明力がなく，推計デフォルト確率と実際のデフォルトに関係がない場合，どのようなレベルの推計デフォルト確率であろうと，同じ割合でデフォルト企業が含まれているため，CAP 曲線は 45 度線上にプロットされる．また，モデルの予測が完全に的中した場合，推計デフォルト確率が高いほうから順にデフォルトするので，グラフの形状は図 3.4 のように期待される．従って，図 3.4 のような曲線に近い曲線が描けるほうがよい．

AR は，モデルの予測が完全に的中した場合の CAP 曲線と 45 度線とで囲まれる面積 (これを S_1 とする) と CAP 曲線と 45 度線とで囲まれる面積 (これを S_2 とする) の比 S_2/S_1 で定義される．全

[4]大規模データでは，モデルが直感と整合性がつかない符号をもつパラメータを含んでしまうことが多い．実際，例で挙げた営業利益だけをパラメータとするモデルを作れば，営業利益は大きいほどデフォルト確率が低くなるという結果を得て，符号条件が合致した．また，用いるデータによっては変数選択を行っても符号条件が合うこともあるので，直感と整合性がつかない符号をもつパラメータは，推定に用いるデータに依存して符号条件がぶれる傾向を持つことがわかった．以上述べてきた問題を解決する方法として，ブートストラップ法を適用することが考えられる．ブートストラップ法は，着目しているパラメータの分布を推定し，パラメータの安定性を確認する方法である．しかし，ブートストラップ法を適用すると計算量が膨大となるため，計算プログラムに何らかの工夫を施すか，スーパーコンピュータなどの高性能なマシンを用いなければ適用できないと考えられる．

件データにおける AR は 0.70678 となった．AR は 1 に近いほど推定精度が良好であると判断でき，AR を比較することでモデルの推定精度の比較ができる．そこで，CAP 曲線や AR は次節のセグメントに分けた場合の推定と本節の全体データによる推定の推定精度を比較する指標として用いる．

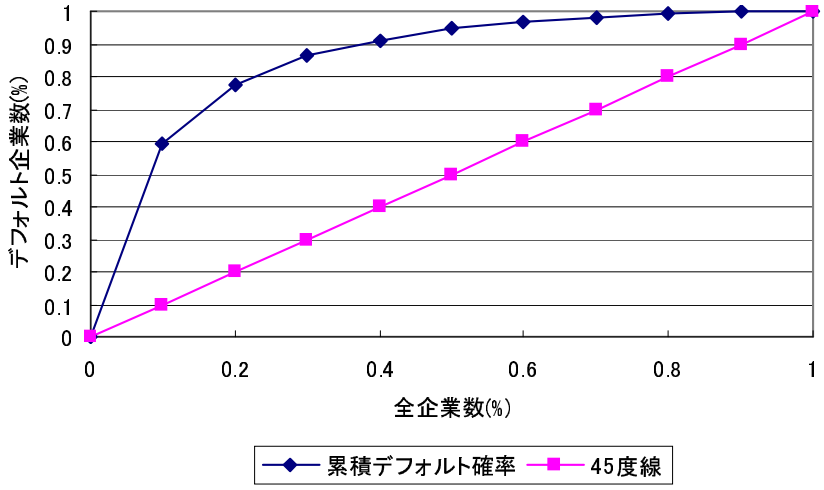

図 3.3: CAP 曲線 (全件データ)

表 3.2: 推定結果 (全件データ，AIC: 28230.78，対数尤度: -14074.4)

フィールド名	係数	t-値
1996 年度	2.696	19.500
1997 年度	2.912	21.629
1998 年度	2.824	21.441
1999 年度	2.318	18.196
欠損値に対するペナルティ	-0.015	-0.964
受取手形	0.053	6.297
その他流動資産合計	-0.070	-8.566
その他固定資産合計	-0.033	-3.529
支払手形	-0.047	-5.791
その他固定負債	0.036	5.611
資本金	-0.067	-8.627
売上総利益	0.069	7.497
営業利益	-0.030	-4.160
受取利息割引料配当金	0.037	4.531
支払利息割引料	-0.085	-6.224
経常利益	0.055	7.914
受取手形割引高	-0.045	-8.975
受取手形裏書譲渡高	0.065	10.447
期末従業員数	-0.105	-16.769
流動資産合計伸び率	0.012	2.166
現金預金伸び率	0.022	3.869
その他流動資産合計伸び率	0.024	3.132
その他固定資産合計伸び率	0.050	8.867
売上総利益伸び率	0.046	6.583
営業利益伸び率	-0.008	-1.131
期末従業員数伸び率	0.078	12.765
（受取利息ー支払利息) 対総資本	0.091	11.493
売上総利益率	0.032	4.839
総資産回転率	0.024	1.758
流動資産回転日数	-0.087	-12.238
買掛金回転日数	0.026	5.295
売上高減価償却	0.064	12.041
一人あたりの資産	-0.089	-12.790
固定負債キャッシュフロー倍率 (逆数)	0.077	8.140
支払準備金	0.100	9.878
預借率	0.193	14.617
固定負債対有形固定資産比率	-0.051	-7.410
固定比率 (逆数)	-0.050	-5.902
自己資本比率	0.221	24.076
借入金依存度	0.115	15.963
インタレストカバレッジ	-0.056	-7.009

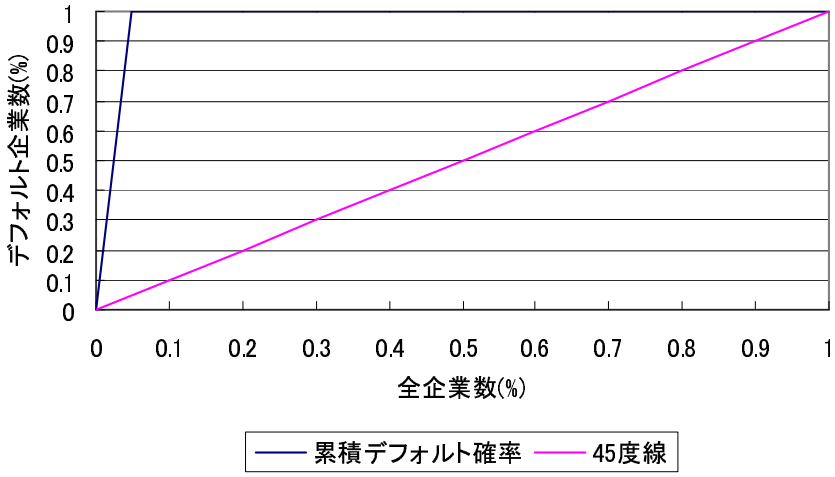

図 3.4: 期待される CAP 曲線

表 3.3: CAP 曲線の 10 分位点における数値 (全件データ)

全件数の割合	デフォルト企業の割合
10 %	0.596
20 %	0.771
30 %	0.864
40 %	0.913
50 %	0.948
60 %	0.971
70 %	0.981
80 %	0.992
90 %	0.998

3.6　セグメントしたデータでの推定結果

全件データで推定した場合，業種や規模が信用リスクにあたえる影響を推定できない．そこで，そのような影響を考慮する方法としてデータセグメント法が一般的に用いられる．本節では，全件データ推定とセグメントでの推定との比較分析を行う．

3.6.1　全件データとセグメントデータとの推定精度の比較

全件データをセグメントに分けて，全データの推定とセグメントデータの推定との推定精度について比較分析を行う．セグメント方法であるが，(1) 業種のみに分類した場合と，(2) 業種・規模セグメントの 2 通りを考えた．

まず，セグメントの分け方が良好なものであるか，尤度比を用いて検討する．尤度比が小さいほど，最尤推定量が良好だといえる．全件データでロジットモデルを適用する場合と業種セグメントで行う場合を比較すると，尤度比は業種セグメントのほうが全て小さくなっている．また，業種・規模セグメントと比較してもいくつかを除いては，同様のことが言える．一方，業種セグメントと業種・規模別セグメントの尤度比を比較すると，建設業，サービス業に関しては規模ごとに切ることで，各セグメントの尤度比が業種セグメントのそれより小さくなっている．これは，製造業の大，卸売業の中・大，小売業の中・大にも言える．しかし，製造業の小，卸売業の小，小売業の小については逆に尤度比が大きくなっている．本研究では，規模の分け方を表 3.7 のように，資産規模で分け，かつ，小・中・大の企業件数が等分されるように切り分けた．どのような基準で規模を分けるかは工夫が必要である．[5]

次に，AIC 基準を比較する．全件データでは 28230.78，業種セグメントでは 27511.28，業種・規模セグメントでは 27015.56 となった．全件データと比較すれば，業種セグメントも業種・規模セグメントも AIC 基準は小さくなり，セグメントに分けたことでモデルに対する予測誤差も小さくなった．

最後に，AR を用いて推定精度を比較する．全件データと同様に AR を求めると，業種セグメントでは 0.73775 ，業種・規模セグメントでは 0.76725 となった．この結果から，細かくセグメントに分けることで 1 に近づいていることがわかる．この値は 1 に近いほどよい指標であることを考えれば，セグメントに分けたほうが良好な結果を得たと判断できる．

3.6.2　全体データとセグメントデータとの選択されたフィールドの比較

全件データと比較して，セグメントに分けた場合に変数として選択されるフィールドの差異について分析を行った．全件データの分析で有効であった，自己資本比率，預借率，借入金依存度は，セグメントに分けた場合でもデフォルト確率に大きく寄与する．しかし，飲食店業と建設業・小においてはこれらの変数はいずれも選択されていない．これは，セグメントに分けたことでデータ数が減少し，これら 3 変数の説明力が低下したため選択されなかったと考えられる．

[5]セグメントの分け方であるが，資産規模以外にも試したが，結果はあまり変わらなかった．むしろ，セグメントに分ける数が尤度比に影響すると感じられた．

セグメントデータでは，受取利息割引料配当金，支払利息割引料が多く選択された．前節より，

$$(\text{営業利益}) = (\text{経常利益}) - (\text{受取利息}) + (\text{支払利息})$$

であった．従って，営業利益のようなよく知られた財務指標を変数として選択するより，それを構成する受取利息割引料配当金，支払利息割引料を変数として選択するほうがよい場合もあると考えられる．このような変数を見つけるためには，変数選択候補を多くしなければならない．これは大規模データを用いて推定する利点となる．

一方，全件データでは選択されなかったがセグメントに分けたことで選択された変数として，現金預金，当座比率，現預金比率，固定負債対キャッシュフロー倍率，を選択するセグメントが多く存在した．このような結果から，預借率，支払準備率，当座比率，現金預金，現預金比率，固定負債対キャッシュフロー倍率，借入金依存度など，キャッシュに関する変数がデフォルト確率に寄与することがわかった．このような選択される変数の傾向も，変数選択候補数が多いことから発見されるものであり，大規模データを用いた推定の利点となっている．

3.6.3 各種セグメントにおける推定結果の特徴

セグメントに分けたそれぞれの推定結果に，特徴的な財務指標が存在するかどうか分析を行った．製造業では，売上高減価償却比率がデフォルト確率を低める要因として寄与するという結果を得た．売上高が低くとも減価償却をすることで資金が留保されたり，または，減価償却をすることで新しい技術や設備を整え生産性の向上ができる，と考えることで説明がつく．特に製造業であるから，新しい技術や設備の導入が可能な企業はデフォルトしにくいと考えられるので，この指標がデフォルト確率に寄与するのは直感と整合する．

サービス業では，流動資産回転日数が大きいとデフォルト確率を高める要因として寄与するという結果を得た．流動資産回転日数は分子が流動資産，分母が売上高営業収益であり，これに 1 年の日数（365）を掛けてまとまる財務指標である．分子の流動資産において，回収不能売掛債権や仮払金といった不良性資産の割合が多くなることで，デフォルト確率が高まることが考えられる．

飲食店業では，売上高営業収益がデフォルト確率を低める要因として大きく寄与するという結果を得た．売上高営業収益が高ければデフォルトしにくいことは，単純に考えても直感と整合性がつく．そのほかのセグメントではこのような特徴的な財務指標は見受けられず，全件データと同様，自己資本比率，借入金依存度，預借率などがデフォルト確率に寄与するということが言える．

また，現預金比率または預借率は，どの業種においてもどちらか一方が必ず選択され，寄与率も高かった．現預金比率を選択するセグメントは，小売業，サービス業，運輸通信業，飲食店業であり，預借率を選択するセグメントは，建設業，製造業，卸売業，金融・保険・不動産業であった．現預金比率および預借率は，

$$\text{現預金比率} = \frac{\text{現金預金}}{\text{売上高営業収益}} \qquad \text{預借率} = \frac{\text{現金預金}}{\text{有利子負債額}}$$

で与えらる．すなわち，前者は売上げにおける現金の割合を示し，後者は有利子負債に対して返済できるだけの十分なキャッシュをどれだけ持っているかを示している．小売業，サービス業，飲食店業はその日に売上げた現金を用いて，材料などを即座に調達しなければならない．そのため，売

上げにおける現金の割合が多くなければ調達が不可能となりデフォルトしやすくなってしまうと考えられる．一方，建設業，製造業などは比較的現金の出入りが穏やかなので，来たるべき負債の返済に対して十分な現金を持つことがデフォルトしないために必要であると考えられる．以上のようなことから，業種の特性を考えて，これら 2 指標のうちどちらかを変数として選択すべきであり，また 2 つ同時に必要ではないと考えられる．

表 3.4: 業種セグメント

業種	件数
建設業	104005
製造業	110933
卸売業	63308
小売業	58030
サービス業	55133
運輸・通信業	21731
飲食店業	10197
金融・保険・不動産業	16036
その他	2313

表 3.5: 業種・規模セグメント

業種	件数	資産規模 (単位：千円)
建設業 (小)	34167	～50000 未満
建設業 (中)	38802	50000 以上～200000 未満
建設業 (大)	31036	200000 以上～
製造業 (小)	38307	～100000 未満
製造業 (中)	49622	100000 以上～1000000 未満
製造業 (大)	23004	1000000 以上～
卸売業 (小)	18749	～100000 未満
卸売業 (中)	29164	100000 以上～1000000 未満
卸売業 (大)	15395	1000000 以上～
小売業 (小)	20073	～50000 未満
小売業 (中)	21661	50000 以上～200000 未満
小売業 (大)	16296	200000 以上～
サービス業 (小)	20752	～50000 未満
サービス業 (中)	20258	50000 以上～300000 未満
サービス業 (大)	14123	300000 以上～
運輸・通信業	21731	
飲食店業	10197	
金融・保険不動産業	16036	
その他	2313	

表 3.6: 尤度比

業種	初期尤度	最終尤度	尤度比
全件	-17137.800	-14074.400	0.821
建設業 (全体)	-3708.510	-3023.080	0.815
建設業 (小)	-818.503	-655.757	0.801
建設業 (中)	-1419.840	-914.932	0.644
建設業 (大)	-1682.680	-1323.710	0.787
製造業 (全体)	-4674.500	-3776.040	0.808
製造業 (小)	-838.507	-699.200	0.834
製造業 (中)	-2293.650	-1880.640	0.820
製造業 (大)	-1460.200	-1120.020	0.767
卸売業 (全体)	-3181.880	-2548.720	0.801
卸売業 (小)	-480.090	-406.697	0.847
卸売業 (中)	-1462.200	-1151.850	0.788
卸売業 (大)	-1177.880	-906.087	0.769
小売業（全体）	-1737.580	-1415.990	0.815
小売業（小）	-441.322	-370.509	0.840
小売業（中）	-549.691	-426.897	0.777
小売業（大）	-718.319	-546.531	0.761
サービス業（全体）	-1686.260	-1274.400	0.756
サービス業（小）	-280.643	-196.324	0.700
サービス業（中）	-539.678	-391.417	0.725
サービス業（大）	-795.580	-580.806	0.730
運輸・通信業	-648.586	-509.968	0.786
飲食店業	-300.769	-213.311	0.709
金融・保険・不動産業	-1015.330	-701.480	0.691
その他	-185.000	-0.649	0.004

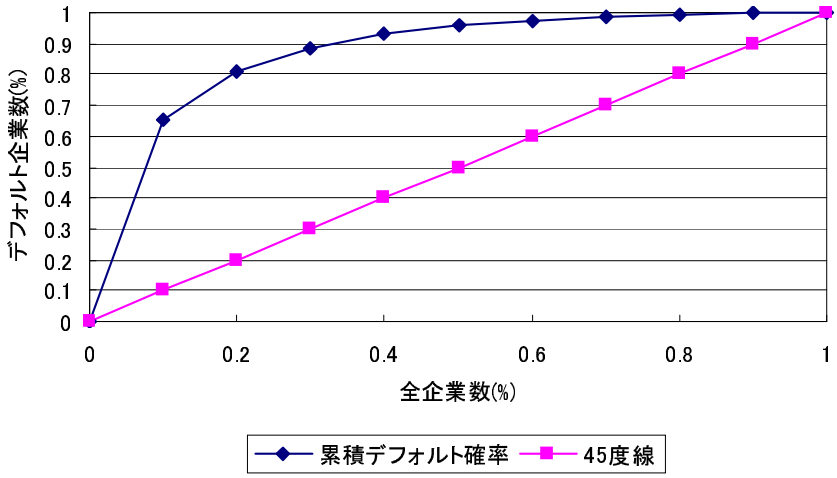

図 3.5: CAP 曲線 (業種セグメント)

表 3.7: CAP 曲線の 10 分位点における値 (業種セグメント)

全件数の割合	デフォルト企業の割合
10 %	0.653
20 %	0.806
30 %	0.886
40 %	0.932
50 %	0.957
60 %	0.975
70 %	0.984
80 %	0.994
90 %	0.999

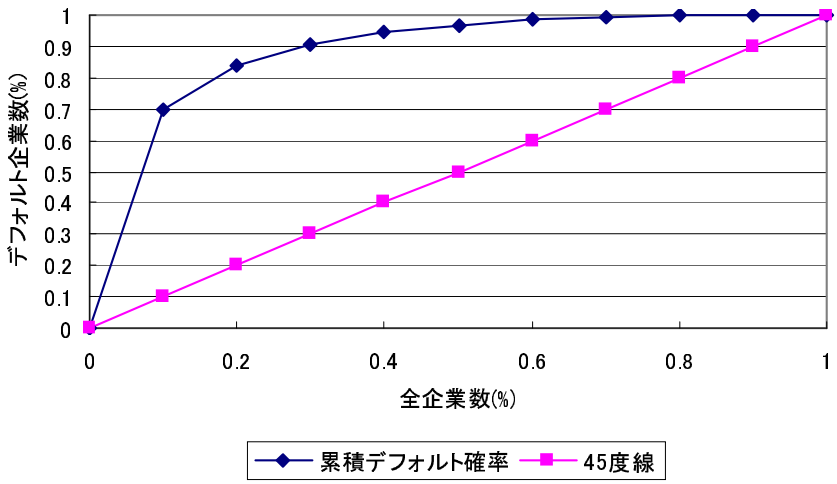

図 3.6: CAP 曲線 (業種・規模セグメント)

表 3.8: CAP 曲線の 10 分位点における値 (業種・規模セグメント)

全件数の割合	デフォルト企業の割合
10 %	0.697
20 %	0.842
30 %	0.908
40 %	0.944
50 %	0.967
60 %	0.985
70 %	0.994
80 %	0.999
90 %	1.000

表 3.9: 推定結果 (飲食店業・全件)

フィールド名	係数	t-値
1996 年度	5.859	5.209
1997 年度	5.815	5.611
1998 年度	6.464	6.248
1999 年度	5.424	5.577
欠損値に対するペナルティ	-0.238	-2.755
現金預金	-0.375	-2.064
その他流動資産合計	-0.154	-2.317
固定資産合計	0.259	1.918
その他固定資産	-0.141	-1.781
流動負債合計	-0.221	-2.077
支払手形	-0.186	-2.433
買掛金	-0.369	-3.665
固定負債合計	-0.580	-4.694
売上高営業収益	0.657	3.912
経常利益	0.198	2.279
期末従業員数	0.168	1.893
その他流動資産合計伸び率	0.111	1.472
固定資産合計伸び率	0.225	2.834
資産合計伸び率	-0.163	-1.953
売上高営業収益伸び率	0.170	2.325
経常利益伸び率	-0.101	-1.636
一人当たり売上高	0.290	2.989
キャッシュフロー	0.145	1.628
現預金比率	0.377	2.898
当座比率	0.224	1.899
AIC	478.62	
対数尤度	-213.31	

表 3.10: 推定結果 (建設業・全件)

フィールド名	係数	t-値
1996 年度	2.735	5.935
1997 年度	2.879	6.284
1998 年度	2.715	5.975
1999 年度	2.426	5.356
欠損値に対するペナルティ	0.053	0.992
受取手形	0.042	2.476
棚卸資産合計	0.130	2.853
その他流動資産合計	-0.100	-5.090
固定資産合計	-0.087	-1.885
土地	0.038	1.912
その他固定資産	-0.054	-2.534
支払手形	-0.032	-1.568
買掛金	0.058	2.657
資本合計	-0.053	-2.266
資本金	-0.096	-4.882
売上総利益	0.141	6.880
営業利益	-0.052	-2.281
受取利息割引料配当金	0.041	2.329
当期利益	0.044	1.360
受取手形割引高	-0.040	-2.616
期末従業員数	-0.200	-7.355
現金預金伸び率	0.069	4.562
その他流動資産合計伸び率	0.027	1.776
その他固定資産伸び率	0.060	4.419
資産合計伸び率	-0.033	-1.753
その他流動負債合計伸び率	-0.025	-1.802
売上高営業収益伸び率	0.030	1.660
売上総利益伸び率	0.033	1.967
支払利息割引料伸び率	0.040	2.383
期末従業員数伸び率	0.112	3.524
総資本当期利益率	-0.067	-1.686
（受取利息－支払利息）対総資本	0.162	6.871
総資本経常利益率	0.224	6.080
総資本回転率	0.141	3.780
固定資産回転率	-0.069	-1.766
流動資産回転日数	-0.034	-1.363
棚卸資産回転日数	-0.095	-2.601
売上高減価償却費	0.159	5.402
営業利益支払い率（逆数）	-0.122	-2.999
一人当たり売上高	-0.114	-4.875
キャッシュフロー	0.036	1.435
預借率	0.182	7.0778
当座比率	0.106	4.546
固定比率（逆数）	-0.066	-1.487
固定長期適合率（逆数）	-0.058	-2.590
自己資本比率	0.265	5.520
借入金依存度	0.100	3.999
AIC	7650.08	
対数尤度	-3776.04	

表 3.11: 推定結果 (サービス業・全件)

フィールド名	係数	t-値
1996 年度	5.952	6.652
1997 年度	5.922	6.784
1998 年度	5.588	6.521
1999 年度	5.014	5.876
欠損値に対するペナルティ	-0.112	-2.834
現金預金	0.131	2.406
受取手形	0.068	2.161
その他流動資産合計	-0.065	-2.101
固定資産合計	-0.151	-1.988
その他固定資産	-0.089	-2.460
支払手形	-0.116	-4.452
資本金	-0.056	-1.935
売上総利益	0.131	3.684
支払利息割引料	-0.192	-2.866
経常利益	0.079	1.775
受取手形裏書譲渡高	0.305	2.665
期末従業員数	-0.082	-2.429
その他固定資産伸び率	0.073	3.020
流動負債合計伸び率	0.043	1.824
売上総利益伸び率	0.084	3.152
期末従業員数人伸び率	0.169	3.569
総資本当期利益率	-0.160	-3.406
総資本経常利益率	0.107	1.604
固定資産回転率	-0.141	-2.786
流動資産回転日数	-0.154	-5.820
売上高支払利息・割引料率	-0.060	-1.308
一人当たり総資本	-0.095	-2.762
キャッシュフロー	0.054	1.482
固定負債キャッシュフロー倍率（逆数）	0.114	2.158
現預金比率	0.137	3.300
正味運転資本額	0.103	3.922
固定負債対有形固定資産比率	-0.114	-3.219
自己資本比率	0.164	4.534
インタレストガバレッジ	-0.130	-2.742
AIC	2616.80	
対数尤度	-1274.40	

第4章　データ量とセグメント数の関係

前章で述べたように，業種や規模が信用リスクに与える影響についてはデータセグメント法を用いるのが一般的である．しかし，データセグメント法を行うと，各セグメントのデータ量が減少し推定結果が悪化するおそれがある．本章では，データ量に対して適切なセグメント数がいくつであるかを示す．

4.1　分析方法

分析方法は前章と同様にロジットモデルを用いて，AIC 基準を比較し最適モデルを決定する．本研究では，　全件データと業種セグメントの比較，　業種セグメントと業種・規模セグメントとの比較，を行った．

4.1.1　データベースの作成およびセグメント方法

前章で作成したデータベースおよびデータセグメント方法を用いて分析をはじめたが，データベース全体のデフォルト件数が 2,861 件であったため，セグメントに分けた結果，デフォルト企業を含まないセグメントが現れる状態がしばしば発生した．この状態を解消するためにデフォルト件数をより多く含むデータベースに更新した．本章で用いたデータベースは，CRD 運営協議会が作成したもので，前章と同様のデータ加工を行った．その結果，レコード数 879,430 件 (うちデフォルト件数 7126 件) となった．

また，前章の業種セグメント方法では，極端にデフォルト件数が少なくなるセグメントが存在したため，セグメント方法についても改めた．業種セグメントについては，　建設業，　製造業，　卸売業，　小売業および飲食店業 (以下，小売業と呼ぶ)，　サービス業，運輸通信業および金融・保険・不動産業 (以下, サービス業と呼ぶ)，とした．業種・規模セグメントについては，表 4.1 を参照とする．

4.1.2　最適モデルの比較基準

本研究の目的は，データ量に対して適切なセグメント数がいくつであるかを分析することである．例えば,「全データ件数 10000 件，うちデフォル件数 100」というデータベースであれば，セグメントをいくつに分ければよいモデルになるかを求めることである．これを決定するために，全体のデータベースからサンプリングを行い，目的にあわせたデータベースを作成した．また，全データ推定モデルとセグメントデータ推定モデルを AIC 基準を用いて比較し，セグメントに分割するか否か

表 4.1: 業種・規模セグメント

業種	資産規模 (単位：千円)
建設業 (小)	～50000 未満
建設業 (中)	50000 以上～200000 未満
建設業 (大)	200000 以上～
製造業 (小)	～70000 未満
製造業 (中)	70000 以上～300000 未満
製造業 (大)	300000 以上～
卸売業 (小)	～100000 未満
卸売業 (中)	100000 以上～500000 未満
卸売業 (大)	500000 以上～
小売業 (小)	～50000 未満
小売業 (中)	50000 以上～200000 未満
小売業 (大)	200000 以上～
サービス業 (小)	～50000 未満
サービス業 (中)	50000 以上～300000 未満
サービス業 (大)	300000 以上～

を決定した.

データサンプリング法を用いる場合，モデルのよさを判定する方法として，分析用データと検証用データを作成し，分析用データで推定されたパラメータを用いて，検証用データのデフォルト予測の推定精度を比較する方法，すなわち，クロスバリデーション法も考えられる．その場合，全データ推定モデルとセグメントデータ推定モデルのよさを比較する指標として，AR などの指標を用いる．しかし，1 回の検証だけでは検証用データのサンプリングの仕方により結果がばらつくので，安定的な結果を得るために，何度も検証を行い，AR の平均や分散を求め，それを比較することが多い.

しかし，AR は統計学的に確立された指標でもなく，現在のところ AR に変わる評価基準が確立されていない．そのような基準を用いた分析では信頼性が低下してしまう．また，クロスバリデーション法と AIC 基準を用いた比較方法では，まったく異なる結果が得られるのではなく，ほぼ同様の結果が得られる．特に，本研究のように，微妙な差異を判断する場合は，統計量である AIC 基準を用いたほうが良いと判断し，クロスバリデーション法を用いた AR 基準による比較方法を選択せず，統計量である AIC 基準を用いた比較方法を採用した.

4.1.3 データ数によるオーバーフィッティングの可能性

データ数が少ない場合，直感的に考えれば，データに含まれる情報が少ないため説明力が低下し尤度が悪くなると推測できる．しかし，データ数が少なくて変数選択に用いる説明変数が多いとき

は，図 4.1 のようにデフォルト企業と非デフォルト企業を完全に説明できる係数を推定してしまうことがある．この現象を**オーバーフィッティング**と呼ぶ．オーバーフィッティングは推定に用いたデータに対して完全説明ができる．しかし，新たなデータを加えて再び推定を行うと得られた推定結果とはかけ離れた結果を得ることが多く，安定した推定結果が得られない．従って，セグメントに分けて推定をした際，オーバーフィッティングが起きた場合はセグメントに分けないほうがよいと判断する．

この点に注意して推定を行うが，2 項ロジットモデルでは以上で述べたオーバーフィッティングの影響を受けずに尤度がよくなる場合がある．いま，図 4.2 のようにデータが分布していると仮定する．もし，図 4.3 のように丸で囲まれたデータ (以下，データ X と呼ぶ) がないならば，グラフの右側に分布するデフォルト企業のデフォルト確率を 1 と推測できる．一方，データ X がある場合はそれらのデフォルト企業のデフォルト確率を 1 と推測できない．従って，データ X の有無により推定結果は大きく変化するので，このような場合も安定した推定結果を得られたとはいえないと考える．これはデータ数の少なさと，デフォルトしたかしないかの 2 値の推定において起こる特殊な現象である．一般に，統計学では前者をオーバーフィッティングとして定義しているが，2 項ロジットモデルの場合にはこの現象も含めてオーバーフィッティングと定義する．

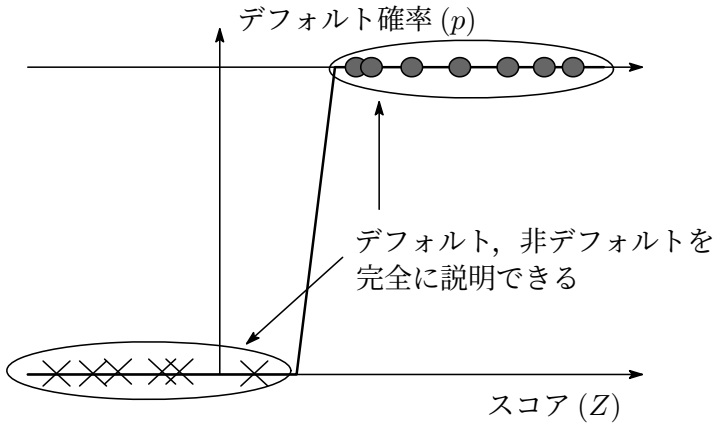

図 4.1: オーバーフィッティング

4.1.4 変数数によるオーバーフィッティングの可能性

オーバーフィッティングが起こる要因をデータ数の減少に対して解説したが，変数選択に用いる説明変数の数にも関係がある．変数選択に用いる説明変数の数が多くなれば，説明力が増す変数を多く選択することが可能になり，オーバーフィッティングが起こりやすくなる．

デフォルト確率を財務データから推定する以上，説明変数の候補の組み合わせは多く存在する．しかし，実際に数値計算を行うためには説明変数の数を固定する必要がある．その際，変数選択に用いる適切な変数の数の基準はなく，その数はモデル使用者の目的によって決定されるものである．本研究では，分析を通して「AIC 基準の比較により，全体データで推定するほうがセグメントに分けて推定するよりモデルが良くなる場合があるか」という命題を解明したい．この目的を達成する

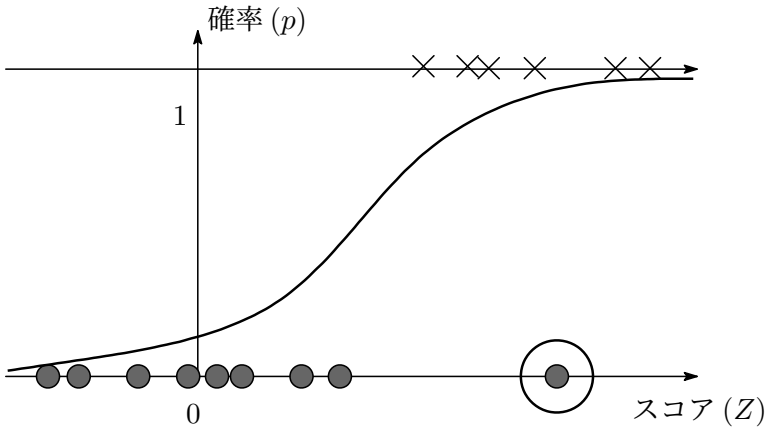

図 4.2: 2 項ロジットモデルの特殊なオーバーフィッティングの例 (データがある場合)

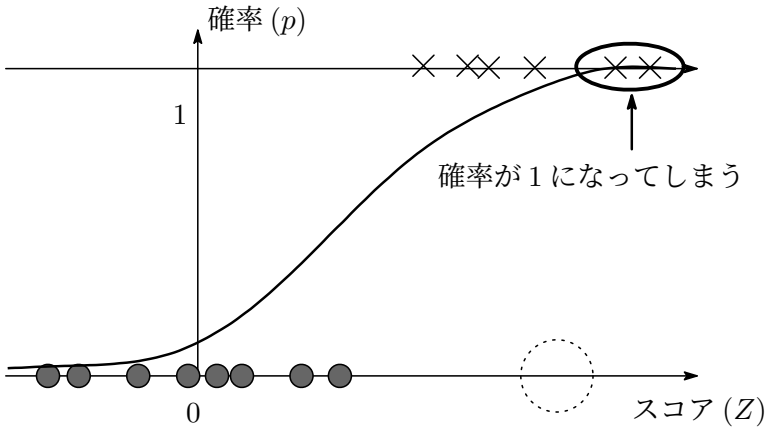

図 4.3: 2 項ロジットモデルの特殊なオーバーフィッティングの例 (データがない場合)

ため，本研究では表 4.2 で示す 33 変数を用いて分析を行った．

表 4.2: 変数一覧

1	1998 年度フラグ	18	資産合計伸び率
2	1999 年度フラグ	19	流動負債合計伸び率
3	2000 年度フラグ	20	固定負債合計伸び率
4	欠損値に対するペナルティ	21	資本合計伸び率
5	流動資産合計	22	売上総利益伸び率
6	固定資産合計	23	支払利息割引料伸び率
7	資産合計	24	経常利益伸び率
8	流動負債合計	25	期末従業員人数伸び率
9	固定負債合計	26	総資本経常利益率
10	資本合計	27	売上高総利益率
11	売上総利益	28	流動資産回転日数
12	受取利息割引料配当金	29	一人当たり総資産
13	支払利息割引料	30	キャッシュフロー
14	経常利益	31	現預金比率
15	期末従業員人数	32	預借率
16	流動資産合計伸び率	33	自己資本比率
17	固定資産合計伸び率		

4.2 変数選択を行う場合の分析結果および考察

データ数とそれに含まれるデフォルト件数を変化させて分析した．その結果，表 4.3 および表 4.4 が得られた．表中の OF はセグメントに分けた結果オーバーフィッティングが発生した，non-cut はセグメントに分けたほうが AIC 基準が悪くなった，cut はセグメントに分けたほうが AIC 基準がよくなった，をそれぞれ表している．以降，OF が分布している領域を OF 領域，non-cut が分布している領域を non-cut 領域，cut が分布している領域を cut 領域と呼ぶ．これらの領域がどのように変化するか， 全体件数の変化， データに含まれるデフォルト数の変化， セグメント数の変化， 変数選択に用いる変数数の変化，により考察する．その準備として，AIC 基準の特徴とセグメント数説明変数の関係について説明する．

4.2.1 AIC 基準の特徴

AIC 基準は以下で定義される；

$$AIC = -2\,(\text{最大対数尤度}) + 2\,(\text{パラメータ数}) \tag{4.1}$$

一般に，対数尤度はモデルのあてはまりの良さを表し，その値が負値で 0 に近いほどよい．また，説明変数を多くすることで説明力の高いモデルを構築できることを考えれば，説明変数を多く用いて対数尤度を 0 に近づけることが可能となる．しかし，ひとつの現象を説明するのに多くの説明変数を用いてモデルを構築するとモデルが不安定となる．そのため，AIC 基準では式 (4.1) で示しているように，右辺第 2 項でモデルに説明変数を加えることに対するペナルティを与えている．また，一般にデータ量に比例して対数尤度は大きくなる．一つのパラメータを加えることでペナルティが 2 増加するが，対数尤度が大きくなればこのペナルティの効果が減少する．従って，AIC 基準はデータ数が多いほど，説明変数を多く取り込む性質を持つことがわかる．

4.2.2 セグメント数と説明変数の関係

セグメントに分けることと説明変数を増やすことが同値であることを説明する．例えば，5 セグメントに分けることを考える．変数選択を行った結果，選択された変数の数を m とする．対数尤度 $l(\mathbf{b})$ とすると AIC 基準は，

$$AIC_{all} = -2 \cdot l(\mathbf{b}) + 2 \cdot m \tag{4.2}$$

となる．同様に，各セグメントで選択された変数数をそれぞれ $m_1, m_2, \cdots, m_5$ とし，対数尤度を $l(\mathbf{b}_1), l(\mathbf{b}_2), \cdots, l(\mathbf{b}_5)$ とするとセグメントデータにおける AIC 基準は

$$AIC_{seg} = -2 \cdot \sum_{i=1}^{5} l(\mathbf{b}_i) + 2 \cdot \sum_{i=1}^{5} m_i \tag{4.3}$$

となる．一般に，m と $\sum_{i=1}^{5} m_i$ の大きさを比較すると，後者のほうが大きくなる．つまり，セグメントに分けたほうが説明変数を多く用いたモデルとなる．以上のことから，セグメントに分けることは説明変数を増やすことで，説明変数の増加があてはまりのよさ，つまり，対数尤度の減少を導くこととなる．

4.2.3 全体件数の差異による領域の変化

データに含まれるデフォルト数を固定して全体データを変化させると，全体件数が多くなるにつれ，OF から non-cut, そして non-cut から cut へと変化することがわかる．全体データが少ない場合は前節で説明したオーバーフィッティングが発生している．データ数が多くなるにつれ non-cut が表れる．ここではセグメントに分けることで説明変数が増加し，そのペナルティが効いてセグメントに分けないほうがよいと決定している．さらに全体件数を増加すると，対数尤度が大きくなり，パラメータを加えるペナルティの効果が減少し，セグメントに分けたほうがよいと決定している．

4.2.4 データに含まれるデフォルト数の差異による領域の変化

全体件数を固定してデータに含まれるデフォルト件数を変化させると，同データ数でもそのデータに含まれるデフォルト件数が多くなると対数尤度が増加することを考慮することで，4.2.3 節と同

様のことが言える．

4.2.3 節および本節では，データ数の観点から考察してきたが，データ数が少ない場合にセグメントに分けるとオーバーフィッティングしやすいことがわかった．また，セグメントに分けるか分けないかは対数尤度の大きさに関係していることがわかった．

4.2.5 セグメント数の差異による領域の変化

表 4.3 および 4.4 は，データ数とそれに含まれるデフォルト件数を変化させる分析を，全データと業種セグメント (1 セグメントと 5 セグメント) および業種セグメントと業種・規模セグメント (5 セグメントと 15 セグメント) で行った結果の表である．それぞれの表を比較すると，セグメント数を多くすることで OF 領域が拡大することがわかる．また，cut 領域はあまり変わらないが non-cut 領域は縮小していることがわかる．以上のことから，セグメント数の増加は OF 領域に強く影響し，AIC 領域を縮小させる作用があると考えられる．(図 4.4 参照)

4.2.6 変数選択候補数の差異による領域の変化

表 4.5 は，データ数とそれに含まれるデフォルト件数を変化させる分析を，全データと業種セグメント (1 セグメントと 5 セグメント) を変数選択の変数の数を 50 変数で行った結果の表である．表 4.5 では，non-cut 領域がなくなり，OF 領域と cut 領域だけで構成されていることがわかる．従って，変数選択に用いる変数数の増加は cut 領域に強く影響し，AIC 領域を縮小させる作用があると考えられる．(図 4.5 参照)

表 4.3: 全データ対業種セグメント (1 セグメント対 5 セグメント)

		データに含まれるデフォルト件数				
		100	500	1000	2000	5000
全体件数	1000	OF	OF	-	-	-
	5000	OF	non-cut	non-cut	non-cut	-
	10000	OF	non-cut	non-cut	non-cut	cut
	20000	OF	non-cut	non-cut	cut	cut
	50000	non-cut	non-cut	cut	cut	cut
	100000	non-cut	cut	cut	cut	cut
	200000	non-cut	cut	cut	cut	cut

表 4.4: 業種セグメント対業種・規模セグメント (5 セグメント対 15 セグメント)

		データに含まれるデフォルト件数				
		100	500	1000	2000	5000
全体件数	1000	OF	OF	-	-	-
	5000	OF	OF	OF	OF	-
	10000	OF	OF	OF	OF	cut
	20000	OF	OF	OF	cut	cut
	50000	OF	OF	non-cut	cut	cut
	100000	OF	non-cut	non-cut	cut	cut
	200000	OF	non-cut	cut	cut	cut

表 4.5: 全データ対業種セグメント (50 変数の場合)

		データに含まれるデフォルト件数				
		100	500	1000	2000	5000
全体件数	1000	OF	OF	-	-	-
	5000	OF	cut	cut	cut	-
	10000	OF	cut	cut	cut	cut
	20000	OF	cut	cut	cut	cut
	50000	cut	cut	cut	cut	cut
	100000	cut	cut	cut	cut	cut
	200000	cut	cut	cut	cut	cut

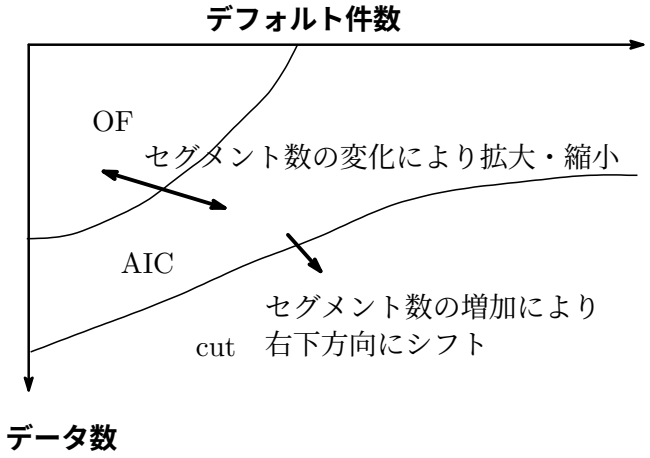

図 4.4: セグメント数の変化による領域変化

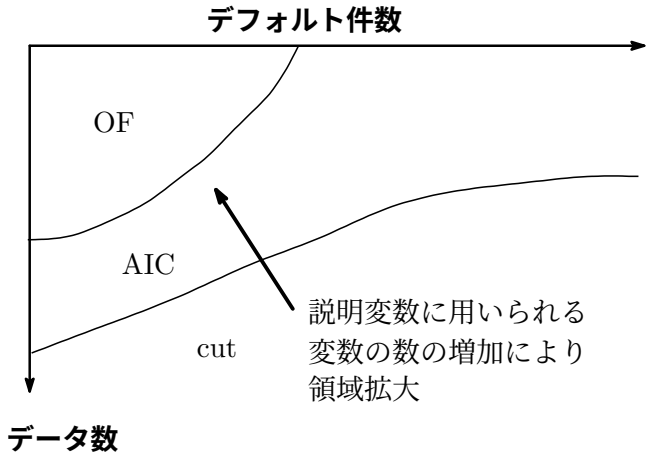

図 4.5: 変数選択に用いる変数数の変化による領域変化

4.3 固定パラメータにおける分析結果

4.2 節では，セグメントにおいても全体データと同様にして変数選択を行った．しかし，この方法でデータ数を増加すると計算時間が膨大となる．[6]この作業時間は変数選択に用いる説明変数数，データ量，およびセグメント数の増加に伴い，劇的に増加する．そこであらかじめデフォルト確率に寄与するパラメータを選別して，説明変数を固定して推定を行ったり，全体データからサンプリングをしてデータ数を減らして計算するのが一般的となっている．しかし，そのような場合のデータセグメント法における適切なセグメント数は今まで言及されていないので，本節ではそれについて分析し，4.2 節と同様の観点から考察を与える．

4.3.1 分析方法

分析方法は 4.1 節と同様に行うが，用いる説明変数は [日本銀行 (2002)] にならい，総資本経常利益率，買入債務回転日数，売上高支払利息割引料率，現預金比率，自己資本比率，有利子負債利子率，の 6 変数に，1998 年度フラグ，1999 年度フラグ，2000 年度フラグを加え計 9 変数を用いた．分析結果は表 4.6 および表 4.7 に示した．

4.3.2 セグメント数の差異による領域の変化

表 4.6 および表 4.7 を比較すると，4.2.5 節で考察したとおり，セグメント数を増加させることで OF 領域が拡大することがわかった．

[6]筆者の作成したプログラムで 50000 件のデータで 32 変数を用いて変数選択を行う場合，1 回計算に要する時間は約 30 分であった．また，統計データであるので安定した結果を得るために何度も計算を繰り返すことになる．表 4.3 および表 4.4 を得るためにはデータの加工なども含めて約 3ヶ月を要している．

4.3.3 変数選択候補数の差異による領域の変化

表 4.3 と表 4.6 を比べると，OF 領域がなくなり non-cut 領域が拡大したが，cut 領域に関しては変化はほとんど見られないことがわかる．4.2.6 節では，変数選択に用いる変数数の増加は cut 領域に強く影響し，AIC 領域を縮小させる作用があると考えたが，ここではその様子が見られない．

変数選択に用いる変数数が少ないときは，変数数によるオーバーフィッティングが生じず，データ数の変化が強く効き，non-cut 領域と cut 領域にわける．徐々に増加していくと，変数数によるオーバーフィッティングが生じ OF 領域が現れ，non-cut 領域を縮小させるが，cut 領域はまだ強く影響しない．さらに増加させると，OF 領域は変わらないが，cut 領域の影響が大きくなり non-cut 領域をつぶしてしまうと考えられる．

表 4.6: 全データ対業種セグメント (1 セグメント対 5 セグメント，9 変数)

		データに含まれるデフォルト件数				
		100	500	1000	2000	5000
全体件数	1000	non-cut	non-cut	-	-	-
	5000	non-cut	non-cut	non-cut	non-cut	-
	10000	non-cut	non-cut	non-cut	non-cut	cut
	20000	non-cut	non-cut	non-cut	cut	cut
	50000	non-cut	non-cut	cut	cut	cut
	100000	non-cut	cut	cut	cut	cut
	200000	non-cut	cut	cut	cut	cut

表 4.7: 業種セグメント対業種・規模セグメント (5 セグメント対 15 セグメント，9 変数)

		データに含まれるデフォルト件数				
		100	500	1000	2000	5000
全体件数	1000	OF	OF	-	-	-
	5000	OF	OF	OF	OF	-
	10000	OF	OF	OF	OF	cut
	20000	OF	non-cut	non-cut	cut	cut
	50000	OF	non-cut	non-cut	cut	cut
	100000	OF	non-cut	non-cut	cut	cut
	200000	OF	non-cut	cut	cut	cut

第5章　結論および今後の課題

5.1　結論

本研究では，大量データによる2項ロジット分析を行い，信用リスク計測に必要な経営指標の組み合わせを推定した．また，データセグメント法におけるデータ量とセグメントの関係からその有意性について検討してきた．全件データによる2項ロジット分析では，自己資本比率がデフォルト確率に大きく寄与することがわかり，預借率，支払準備率などキャッシュに関する財務指標がデフォルト確率に寄与するという結果が得られた．また，営業利益のような代表される指標より，その指標を構成する受取利息割引料配当金や支払利息割引料がデフォルト確率に寄与することもわかった．そして，現預金比率と預借率は，業種の特性を考えて，2指標のうちどちらかを変数として選択すべきであり，また2つ同時に必要ではない指標であることがわかった．

次に，データセグメント法におけるデータ量とセグメントの関係であるが，全体のデータ数，それに含まれるデフォルト数，変数選択に用いる説明変数数と分けるセグメント数には次のような関係があった．

1. 全体件数の変化は，AIC基準の対数尤度と加えるペナルティの関係により，全体件数が多くなるにつれ，OFからnon-cut, そしてnon-cutからcutへと変化する．

2. データに含まれるデフォルト件数の変化は，AIC基準の対数尤度と加えるペナルティの関係により，データに含まれるデフォルト件数が多くなるにつれ，OFからnon-cut, そしてnon-cutからcutへと変化する．

3. セグメント数の増加はOF領域に強く影響し，non-cut領域を縮小させる作用がある．

4. 変数選択に用いる変数数の増加は，cut領域に強く影響しnon-cut領域を縮小させる作用がある．

AIC基準を用いた2項ロジットモデルでのデフォルト確率を推定する場合，データセグメント方法は以上の点に留意して行わなければならない．最後に，データ数が多いと計算負荷がかかることは第3章でも述べてきたが，計算プログラムを改良することでかなりの時間短縮が行える．実際，筆者はCPU 2.0GHz，メモリー1.0Gbiteのパソコンを用いて40万件，86変数の変数選択を1日で処理している．統計モデルではデータ数があればそれだけデフォルト確率に寄与する財務指標を見つけることができる．多くの情報を含む大量データではサンプリング手法を用いて，その情報を落としてしまいかねない．また，データ数の多さは先に述べたオーバーフィッティングの問題を回避できる．計算プログラムを改良し多くのデータで推測することが可能でありまた重要であることを確

認してもらいたい.

5.2 今後の課題

統計モデルでは，推定精度を高めることでオーバーフィッティングの可能性を高めることになる．そのために，オーバーフィッティングの評価と解釈，そして，その問題を回避するための新しい方法が必要である．その目的を達成するために以下の 3 点については今後の課題にしたい.

5.2.1 オーバーフィッティングの評価

本研究では，適切セグメント数を決定する際，デフォルト企業の推定デフォルト確率が 1 になることもオーバーフィッティングであるとし．オーバーフィッティングの定義を拡張した．しかし，非デフォルト企業についても推定デフォルト確率が限りなく 0 に近づくことがある．これもオーバーフィッティングとして定義可能である．また，オーバーフィッティングはそのデータについては完全説明ができるのでよい推定を行えたとも考えられる．従って，オーバーフィッティングをどのように解釈するか，また，その定義について厳密に定義しなければならない.

5.2.2 最適化の方法と変数選択基準について

パラメータを推定する際，統計モデルでは最尤法を用いるのが一般的である．最尤法は，与えられたデータに対して実際との当てはまりのよさを最大にするようにパラメータを推定していく方法である．それに伴う変数選択基準を本研究では変数選択基準として AIC 基準を用いた．その他に自由度調整済み決定係数や Cp 統計量などが存在する．いずれもモデルの当てはまりと説明変数を調節するものであり，そのような基準を用いて最尤法を行えば，先に述べたオーバーフィッティングの問題は付きまとうことになる.

この問題を解決するために，推定精度を評価した CAP 曲線やベイズ理論に基づく ROC 曲線などを用いた推定方法を考えている．また，実際デフォルトが発生すると損失が発生するので，データの当てはまりより，損失が最小になるようにパラメータを推定することが重要なのではないか．最尤法に変わるパラメータ推定方法が必要である.

5.2.3 潜在変数モデル

統計モデルでは，オーバーフィッティングの問題を避け，推定精度の高いモデルを作ることが大切である．5.2.2 節でも，今後の課題としてオーバーフィッティングを解決するような方法を考えた．現在，このような問題に対しては潜在変数モデルを用いることで回避することが出来る．第 3 章の全体データを用いた推定結果では，説明変数を 41 変数も選択している．財務指標を用いたロジットモデルのパラメータ推定では，個々の変数がそれほど情報を持っていないので，変数を多く選択する傾向がある．そこで，情報の少ない多くの変数を情報が縮約された変数 (潜在変数) に変換してパラメータを推定をすることが考えられる．潜在変数を作成する方法としては，主成分分析や因子分

析がある．財務データを主成分分析 (または，因子分析) して，そこで得られる主成分得点 (因子得点) を用いてロジットモデルで推定する方法である．主成分分析，因子分析は方法論は異なるにせよ，得られる結果はあまり差異がない．モデルに対する自由度という面で，因子分析のほうが自由度が高い．自由度が高いとよりいいモデルを作成することが出来る．また，さらに自由度が高いモデルとして，LISREL という方法がある．主成分分析，因子分析では，全ての変数を用いて潜在変数を作成するが，LISREL では，一部の変数の線形結合によって潜在変数を作ることが出来る．これは，柔軟性が高く，過去のノウハウを直接利用できるモデルである．

参考文献

[Altman(1968)] Altman, E.I. “Financial Ratios, Discriminant Analysis and The Prediction of Corporate Bankruptcy”. *Journal of Finance*, 1968, 23(4), 589-609

[森平，小松，湯山 (1996)] 森平爽一郎，小松幹生，湯山智教 “倒産確率と考査モデルー信用組合の事例をめぐって” 1996 年度日本金融・証券計量・工学学会 (JAFEE) 夏季大会予稿集，10-24

[Kaplan,Urwitz(1979)] Kaplan,R.S. and Urwitz,G. “Statistical Models of Bond Rating: A Methodological Inquiry”. *The Journal of Bisiness*, 1979, 52, 231-261

[中山，森平 (1998)] 中山めぐみ，森平爽一郎 “格付け選択確率の推定結果と信用リスク量” 1998 年度日本金融・証券計量・工学学会 (JAFEE) 夏季大会予稿集，210-225

[安川，椿 (1999)] 安川武彦，椿広計 “社債格付けの決定要因に関する分析” 第 67 回日本統計学会講演報告集，238-239

[Lane, Looney,and Wansley(1986)] Lane, W.R., Looney,S.W. and Wansley,J.W. “An application of the Cox proportinal hazard model to bank failure”. *Journal of Banking and Finance*. 10. 511-532

[Merton(1974)] Merton,R.C. “On The Pricing of Corporate Debt: The Risk Structure of Interest Rates”. *Journal of Finance*, 1974, 29(2), 449-470

[Duffie,Singleton(1999)] Duffie, D. and Singleton,K. “Modeling term structures of defaultable bonds”, *Review of Financial Studies*. 1999. 12. 687-720

[Jarrow,Lando,and Turnbull(1997)] Jarrow, R.A., Lando,D. and Turnbull,S.M. “A Markov model for the term structure of credit risk spread”. *Review of Financial Studies*. 1997. 10. 481-523

[今野，武 (2001)] 今野浩，武黛，“半定値計画法による倒産確率推計”東京工業大学理財工学研究センター WP01-5,2001

[日本銀行 (2002)] 日本銀行金融市場局金融市場課 市場企画グループ “中小企業売掛債権の証券化に関する勉強会報告書”, 金融市場局ワーキングペーパーシリーズ 2002-J-6

[Press,Teukolsky,Vetterling, and Flannery(1993)] Press, W.H., Teukolsky,S.A., Vetterling, W.T. and Flannery,B.P. 著 丹慶勝市，奥村晴彦，佐藤俊郎，小林誠 訳，NUMERICAL RECIPES in C[日本語版]，技術評論社, 1993

[木島，子守林 (1999)] 木島正明 子守林克哉 著，信用リスク評価の数理モデル，朝倉書店, 1999

[森平 (1999)] 森平爽一郎，“信用リスク測定と管理－第二回：定性的従属変数回帰分析による倒産確率の推定－”，証券アナリストジャーナル, 11, 81-101, 1999

[東京大学出版会 (1992)] 東京大学教養学部統計学教室 編，自然科学の統計学，東京大学出版会, 1992

The problems and solutions of credit risk measurement using large-scale data base (The relation between the variables selection and the amount of data)

Abstruct

This paper provides an advice for the credit risk measurement using a large-scale database. In previous researches, it was impossible to estimate credit risk using a large-scale database, because there were no sufficient accumulations of data. To solve this problem, CRD (Credit Risk Database Association) have been developing a large-scale database, which has 450,000 company records with 86 management indices.

In this paper, we estimate binominal logit models using this database. However if we estimate the parameters of the model naively, it takes significant amount of time for the calculation. To cope with this problem, we improve the optimization algorithm to reduce the estimation time and then select the explanatory variables affecting the probabilities of default.

This paper also shows the optimal number of segments depend on the amount of data. We divide the database into the several segments and estimated the probabilities of default. We evaluate the goodness of fit and the robustness of the estimation by AIC. It becomes clear that the estimation using divided data was often worse than the one with all data. Then, we discuss the optimal number of segments depending on the amount of data. As a result, we obtained a table that determines the optimal number of segments depending on the number of data, defaulted companies in it and candidate explanatory variables.

YAMASHITA, Satoshi
The Institute of Statistical Mathematics Assistant professor
Credit Risk Database Association Adviser
Financial Services Agency Special research fellow

KAWAGUCHI, Sho
Mathematical principle Science, Science and Engineering, Waseda Univercity
Financial Services Agency Technical research fellow