信用リスクモデルの評価方法に関する考察と比較*

山下　智志 †　　川口　昇‡　　敦賀　智裕§

概 要

本稿は，現在用いられている様々な信用リスクモデル評価方法を列挙し，その成り立ち，特色をまとめ，モデルと評価方法の対応関係を考察した．これまでの信用リスクモデル評価方法に関する論文では，個々の評価方法について説明するだけであり，その評価方法が具体的にどのモデルに適用できるかを与える研究は少ない．また，評価方法が体系的にまとめられていないので，どういった考え方でモデルを評価するかが意識されず，予測が完全に当たるモデル，つまり，的中率が高いモデルがよいと判断することが多い．しかし，作成されるモデルは誤差を含むものであり，単純に予測的中率が高いモデルがいいモデルであると判断するのは危険である．

以上のことから，本稿では，個々の評価方法や評価指標について，その特徴をまとめ，各方法の短所・長所について個別に考察した．その後，モデルケースを想定し適用方法を考察した．考察においては，モデル評価の考え方を明確にした上で，その評価方法が適用できるかどうか考えた．その結果，各評価方法に関してどのモデルに適用できるか，その対応関係を示す表を作成することができた．

*本稿の執筆にあたり, 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, 本稿は著者の個人的な見解であり, 金融庁の公式見解ではない.

†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

‡新日鉄ソリューションズ株式会社 金融ソリューション事業部 (元 金融庁 金融研究研修センター 専門研究員)

§一橋大学大学院経済学研究科, 金融庁 金融研究研修センター 専門研究員

目 次

第1章　はじめに

1.1　信用リスクモデル評価の研究背景

信用リスクとは，**「デフォルト（貸倒れ）が発生するなど，貸付け等による投下した資本が返ってこないリスク」**のことをいう．わが国でも，バブル経済の崩壊や金利の自由化に伴い，デフォルト確率を正確に予測し，信用リスクに見合うだけの収益を確保するという考え方が重要になり，新BIS 規制の導入に向け，適切な内部格付制度の確立とその検証が大きな課題となっている．このような状況下において，本稿では，評価方法と適用するモデルの対応関係を明確にし，信用リスクモデルの評価方法を整理し体系化する．

1.2　信用リスクモデルの分類

信用リスク計量化モデルは，**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**に大きく分けられる．

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．これらの統計モデルは，データ数が多いほど説明力の高いモデルを作成することができる．したがって，企業数が多いためにデータが豊富な中小企業に対して適用する．また，社債格付けに対する要請や銀行の与信業務に用いられる内部格付け制度の実用化に伴い，格付けデータの統計モデルも重要となった．**多重判別関数**や**オーダードロジットモデル**は，財務データから格付けデータを推測する代表的なモデルである．

一方，**オプションアプローチモデル**では，株価や社債金利などの市場データを用いてデフォルト確率を推定する．例えば，マートンモデルはオプションアプローチの代表例である．株式が上場されている大企業に対して市場データを得ることにより，リアルタイムでデフォルト確率が推定できる．

以上をまとめると，信用リスクモデルは図 1.1 のようにまとめることができる．

1.3　信用リスクモデル評価方法の現状と問題点

BIS 規制に定められている信用リスクのリスク要素は，**PD**（当該企業のデフォルトのしやすさ．**デフォルト確率**），**LGD**（デフォルトが起きた場合の期待損失），**EAD**（デフォルトが起きた場合の信用エクスポージャー），**Maturity**（残存融資期間）の 4 つとされている．銀行が独自にリスク要素を計算するモデルを用いる場合，監督当局は銀行が正確にリスク要素を計測しているか，評価しなければならない．4 つのリスク要素のうち，計量化モデルの作成が進んでいるのはデフォルト確率だけであり，現状ではそれに対する評価方法が検討されている．

現在，信用リスクモデルの評価手法として多く用いられている手法は，**CAP 曲線**や **AR** といっ

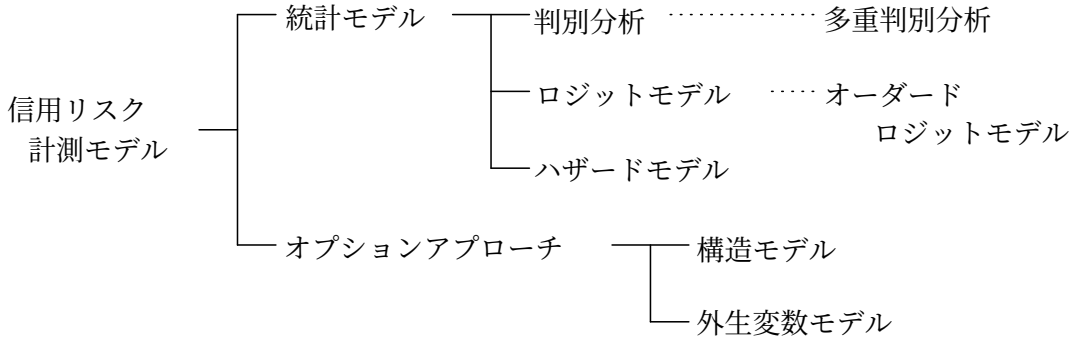


図 1.1: **信用リスク計量化モデルの分類**

た指標である．これらの指標は，デフォルト－非デフォルトの判別がどれだけ当たっているか，いわゆる的中率に重点をおいた指標である[1]．最近では，KMV によるオプションアプローチを用いて導出される EDF と，銀行の作成したモデルで与えられるデフォルト確率を比較して評価する方法も提案されている．また，格付けでは，Moody's や S＆P などで代表されるような格付け機関が作成した格付けと，銀行が独自で作成した内部格付けを比較して，同じようであれば内部格付けモデルは正しいと評価する方法も発案された．

多くの論文では，評価方法の特性について，**判別力（Discriminative Power）**と**予測と結果の合致性（Calibration Power）**の二つの側面がある，と報告している．これを言い換えると，『**どのような考え方でモデルを評価するか（Philosophy of Validation）**』という観点から，以下の二つの考え方でモデルを評価しているといえる．

(1) 予測デフォルト確率が良く当たる，**的中率の高いモデル**がいいモデルである．例えば，事前に想定されたモデルの的中率が 60%で予測結果が 60%的中になるよりも，**事前に想定されたモデルの的中率が 60%で予測結果が 90%的中になる**ほうがよいと判断する．
(2) モデルには誤差があるので，予測結果も誤差を含んでいる．つまり，**予測と予測結果が合致するモデル**がいいモデルである．例えば，事前に想定されたモデルの的中率が 60%で予測結果が 90%的中となるよりも，**事前に想定されたモデルの的中率が 60%で予測結果も 60%的中となる**ほうがよいと判断する．

運用する立場にとっては，損失を最小限に止めるために，予測結果の的中率が高いモデルを用いるはずである，つまり，的中率の高いモデルを選択するのは，さほど不自然なことではない．以上のような考え方が (1) の考え方であり，この場合，**的中率**に重点をおいてモデルを評価することになる．

一方，モデルは誤差を含むものであるという考えの下では，的中率の低いモデルが，予測結果で高い的中率をもった場合，モデルが間違っているか，たまたまいいデータが入力された，のどちらかが判断されるはずである．以上のような考え方が (2) の考え方であり，この場合，**モデル自体の正しさ**という側面から，予測と予測結果が合致しているかを評価することになる．

現状の信用リスクモデル評価方法は，前者の的中率を測る指標でモデルを評価することが多く，後者の評価方法については，あまり考えられていない．また，その評価方法も少ない．この点が，

[1]詳しい説明は，第 2 章を見よ．

現状の信用リスクモデル評価方法の問題点といえる.

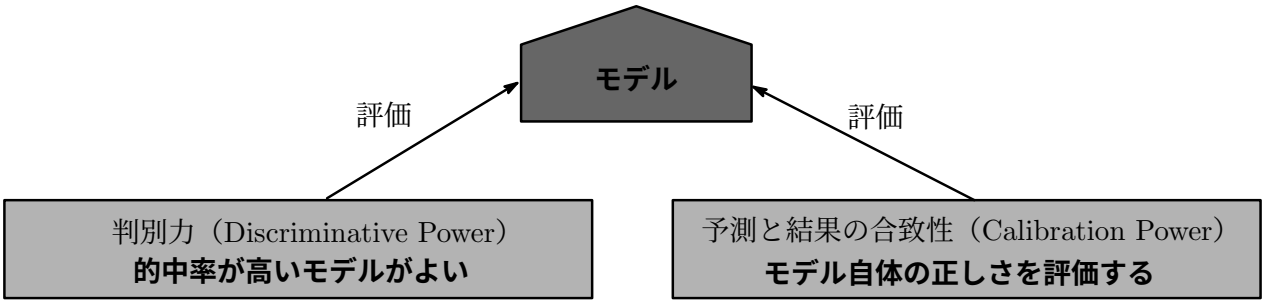


図 1.2: **信用リスクモデル評価の考え方（モデル評価のフィロソフィー）**

1.4 信用リスクモデル評価の適用にあたって

信用リスクモデルを評価するにあたって，第一に考えなければならないのは，前節で述べたように，**どのような考え方でモデルを評価するか**である．モデルの的中率を評価したいのか，モデル自体の正しさを評価したいのか，その考え方をはっきりさせた上で，モデル評価をしなければならない.

実際にモデルを評価するにあたって注意しなければならないことは，**モデルの出力結果**と**検証に用いるデータ**である．具体的には，以下の点に注意する.

(1) モデル作成時による検証（**インサンプルデータ**によるモデル検証）かモデルを用いた運用結果の**バックテスティング**（**アウトサンプルデータ**によるモデル検証）か？
(2) モデルの出力結果が**デフォルト確率 (PD)** か，**格付け (Rating)** か？
(3) モデル評価に用いるデータが **1 時点のデータ**か，**時系列データ**か？

モデル評価者は，この点に注意して適切な評価方法を適用することを心がける.（図 1.3 参照）

本稿では，第 2 章で，さまざまな評価方法を,「目的」「成り立ち」「適用方法」「長所・短所」, 利用されている文献のあるものについては「利用例」という五つの項目をたて，それぞれ詳しく説明する．その際，数学的に難しいと思われる方法については，詳細を Appendix に掲載した．第 3 章では，第 2 章で述べた評価方法を，どのようなモデル，データに対応して適用するのか，例を挙げて説明する．第 4 章では，本稿の結果と今後の課題についてまとめた.

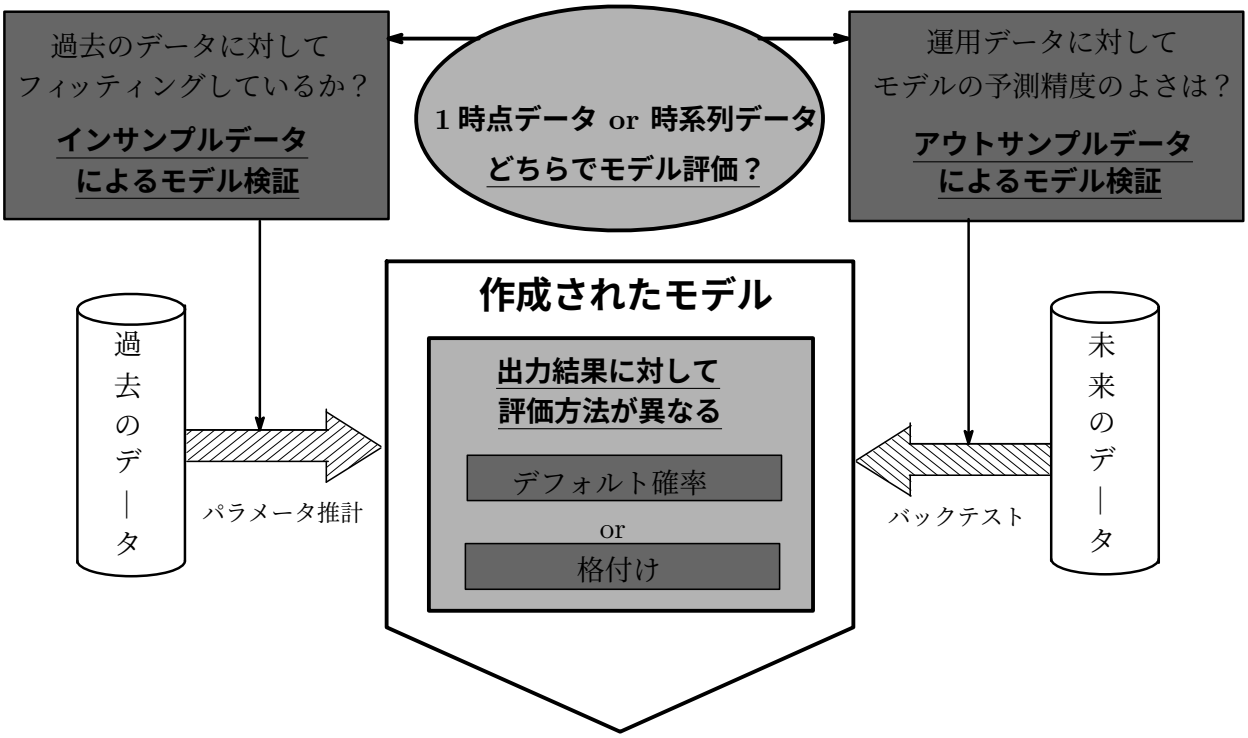


図 1.3: **信用リスクモデル評価の枠組み**

第2章　各評価方法の定義と特徴

本章では，現在用いられている評価方法および評価指標について，**目的**，**成り立ち**，**適用方法**，**長所・短所**の 4 項目に分けて説明をする．

2.1　t-値

目的

信用リスクに対する**統計モデル**では，過去の財務データから特定の財務指標を選び出し，それらを用いて，デフォルト判別やデフォルト確率を求める．**判別分析**，**一般化線形モデル**，**ハザードモデル**などでは，財務データ $(x_1, x_2, \cdots, x_m)$ の線形結合によって与えられる式，

$$z = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m \tag{2.1}$$

で**信用スコア**を求め，デフォルト予測をする．モデルは，過去データから予測が的中するように係数ベクトル $(\beta_1, \beta_2, \cdots, \beta_m)$ を**推定**することで作成される．このとき，推定の結果の係数ベクトル $(\hat{\beta_1}, \hat{\beta_2}, \cdots, \hat{\beta_m})$ の変数一つ一つが必要であるかを調べる．

$$\hat{\beta_j} = 0 \qquad (1 \leq j \leq m) \tag{2.2}$$

このときに適用する評価指標が **t-値**である．

成り立ち

t-値は母集団の分布が正規分布に従っている場合の，母平均を検定する統計量として用いられる指標である．平均が μ，分散が σ^2 の正規分布 $\mathcal{N}(\mu, \sigma^2)$ に従って得られたサンプルを $(X_1, \cdots, X_N)$ とする．**標本平均** $\overline{X}$ と**標本分散** s^2 を

$$\overline{X} = \frac{1}{N} \sum_{i=1}^{N} X_i \tag{2.3}$$

$$s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \overline{X})^2 \tag{2.4}$$

で与える．サンプル数が十分に大きい場合，ステューデントの t 統計量，

$$t = \frac{\overline{X} - \mu}{s/\sqrt{N}} \tag{2.5}$$

は，自由度 $(N-1)$ の t 分布 $t(N-1)$ に従う．

適用方法

推定に用いたデータ数を N とし，$\hat{\beta}_j$ は，平均が 0，分散が σ_j^2 の正規分布 $\mathcal{N}(0, \sigma^2)$ に従っていると仮定する．また，$\hat{\beta}_j$ の標本分散を s^2 とする．

ここで，以下の帰無仮説と対立仮説を立てる．

$$H_0 : \hat{\beta}_j = 0 \qquad H_1 : \hat{\beta}_j \neq 0$$

仮定から，t 統計量，

$$t_j = \frac{\hat{\beta}_j}{s/\sqrt{N}} \tag{2.6}$$

は t 分布 $t(N-1)$ に従うので，有意水準を α として仮説検定をおこなう．例えば，$\alpha = 95\%$ としよう．両側検定では，

$$|t_j| \geq 1.960 \tag{2.7}$$

であれば，帰無仮説が棄却される．つまり，**t-値の絶対値が 2 より大きい場合**，β_j は 0 と有意に差があると判断でき，推定された β_j を信用スコアにいれることで，モデルの説明力が高まると判断できる．

長所・短所

t-値は，モデルに取り込まれている変数（財務指標）の一つ一つが有意であるかを評価できるが，モデル全体の評価はできない．たとえば，あるパラメータが有意でないと判断されても，モデルにそのパラメータを取り込むことでモデルの説明力が高まる場合もある．これより，パラメータの t-値が有意でないという理由で，そのパラメータをモデルからはずすことは，モデルの説明力を悪化させる要因になる．そこで，t-値はモデルのバランスを整える指標として用いることを薦める．例えば，パラメータをたくさん取り入れてモデルが不安定になるのを避けるために，どのパラメータを外すかを判断する指標として適用することなどである．

利用例

[Lee, Urrutia(1996)] は 1981 年 1 月から 1991 年 6 月までにデフォルトした損害保険会社のデータを用いてロジットモデルと Cox 比例ハザードモデルの比較を行い, t-値を用いて説明変数の有意性判定を行った．その結果, ハザードモデルの方がより多くの説明変数に対して有意な結果を得ることができたと報告している．

2.2 尤度比

目的

統計モデルでは，パラメータの推定結果がよいほど，モデルの説明力が高まる.「推定結果がよい」と判断するひとつの方法は，作成したモデルを用いて，推定に使用したデータに対して予測をし，実際の結果とどれだけ適合しているかを調べることである．つまり，**推定に用いたデータとの当てはまり具合（フィッティング）の計測**である．

回帰分析の場合は，**最小二乗法**で推定し，データのフィッティングを**決定係数**を用いて計測する．しかし，信用リスクモデルのパラメータ推定では**最尤推定法**が主に用いられる．最尤推定法では，**尤度比 (Likelihood Ratio)** を用いて，データのフィッティングを計測する．

成り立ち

推定に用いるデータの企業数を N とする．i 番目の企業のデフォルト確率は p_i で与えられているとする．企業のデフォルトは互いに独立して発生すると仮定した場合，同時確率分布 L は，

$$L = \prod_{i=1}^{N} p_i^{\delta_i} (1 - p_i)^{(1-\delta_i)} \tag{2.8}$$

となる．ここで，δ_i は，

$$\delta_i = \begin{cases} 1 & (\text{i番目の企業がデフォルトしているとき}) \\ 0 & (\text{i番目の企業がデフォルトしていないとき}) \end{cases}$$

なる関数である．

デフォルト確率を信用リスクスコア $z = \beta_1 x_1 + \cdots + \beta_m x_m$ によって決定する場合，各企業のデフォルト確率をデフォルトした場合は 1 に，デフォルトしなかった場合は 0 に，近づくようにパラメータ（係数ベクトル）$\beta_1, \cdots, \beta_m$ を決定する方法を**最尤推定法**とよび，(2.8) 式を**尤度関数**と定義する．

係数ベクトルを計算する場合，(2.8) 式の対数をとった関数，

$$l = \sum_{i=1}^{N} \delta_i \log p_i + (1 - \delta_i) \log(1 - p_i) \tag{2.9}$$

にして計算する．この関数を**対数尤度関数**とよぶ．

すべての企業が平均的な同じデフォルト確率を有していると仮定した場合の対数尤度を i_{init}，財務指標を用いて作成したモデルの対数尤度を l_{opt} とする．このときモデルのフィッティングを表す**尤度比 LR** は，

$$\mathbf{LR} = \frac{l_{opt}}{l_{init}} \tag{2.10}$$

で定義される．ただし，出典によっては尤度比を，

$$\mathbf{LR} = 1 - \frac{l_{opt}}{l_{init}}$$

と定義していることもある．

適用方法

尤度比は，その成り立ちから，モデルを最尤推定法で作成した場合にのみ適用できる評価指標である．

もし，全ての変数が説明力のない変数であるならば，どの変数も選択されず，最終尤度は初期尤度のままとなり，尤度比 LR は 1 となる．また，財務データを用いたデフォルト確率推定モデルの説明力が高ければ，デフォルト企業および非デフォルト企業のデフォルト確率は，それぞれ 1 および 0 に近づく．対数尤度は，

$$l(\mathbf{b}) = \log\, L(\mathbf{b}) = \sum_{i=1}^{N} (\delta_i \log p_i + (1 - \delta_i) \log(1 - p_i)) \tag{2.11}$$

であったから，この式より推定に用いたデータの当てはまりがよいと最終的な対数尤度は 0 に近づく．したがって，尤度比がとりうる値の範囲は，

$$0 \leq LR \leq 1 \tag{2.12}$$

であり，0 に近いほどデータとよくフィットしていると判断できる．

長所・短所

尤度比は，モデル推定で用いる対数尤度関数の値をそのまま使うので，推定後，すぐに計測することができる．ただし，あくまでも**推定に用いたデータとのフィッティング**がいいだけであって，**モデルの予測的中率**のよさを示してはいないことに注意する．また，尤度比が 0 に近い場合は，**オーバーフィッティング**が起きている可能性が強い．オーバーフィッティングとは，デフォルトと非デフォルトを完全説明する状態のことを言う．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するが，そのデータに他のデータを加えて推定すると，結果が大きく異なるので，モデルの安定性が悪い．このように，オーバーフィッティングの目安としても尤度比は適用可能である．

利用例

[森平, 隅田 (2001)] は日本格付情報センター (R&I) による 1998 年 9 月から 1999 年 9 月にわたる 1 年間の格付けデータに基づき[1], 順序プロビット・モデルを用いた格付け推移確率の推定を行い, モデルの有意性を尤度比を用いて検定した．また, 同社の 1998 年 5 月から 2000 年 7 月までの製造業の債券格付けデータを用いて[2]コックス比例ハザードモデルによる格付けの取得確率の推定を行い, 同様の検定を行った．その結果, 両推定結果ともモデルが有意であったとの結果を得ている．

2.3 情報量基準

目的

財務指標を用いて信用リスクスコアを作成する場合，**どの財務指標の組み合わせがより説明力の高いモデルになるか**を判断することが重要となる．実際，多くの財務指標を用いることで，説明力の高いモデルを作成できるが，多くの財務指標を用いてモデルを作成することは，モデルの安定性を減少させる．つまり，適当な数で構成される財務指標の組み合わせで，最も説明力が高くなる組み合わせを選ぶことが必要となる．そのような場合の選択基準として**情報量基準**を適用する．

成り立ち

情報量基準には様々な種類がある．一般に用いられるのは **AIC 基準 (Akaike's Information Criterion)** である．ここでは，まず，AIC 基準について説明する．

AIC 基準は，

$$\mathbf{AIC} = -2 \cdot l_{like} + 2 \cdot m_{free} \tag{2.13}$$

で与えられる．ここで，l_{like} は**最大対数尤度**，m_{free} は**モデルに取り込まれているパラメータの数**である．

[1] この間の格付け推移を説明する説明変数は 1999 年 9 月に発表されている中間決算の財務指標の情報を採用している．
[2] 各企業が取得していた格付けが変更される直前の月の時点での最新の財務指標の情報を説明変数として採用している．

AIC 基準は **Kullback-Leibler 情報量**から導出される．導出の詳細については [Akaike(1973)] および [坂本，石黒，北川 (1983)] を参照とする．ここで重要なのは，Kullback-Leibler 情報量が**エントロピー**[3]の性質を持つことである．エントロピーとは簡単に言えば，**モデルが持つ「不確かさの量」**であるといえる．AIC 基準では「不確かさの量」を，(2.13) 式の右辺第 1 項である「最大対数尤度」にマイナスの符号をつけて表している．また，(2.13) 式の右辺第 2 項は，あるパラメータをモデルに取り込んでも，対数尤度があまり軽減しなかった場合には，そのパラメータは「モデルを不安定にさせる」として，パラメータ増加に対するペナルティと見なせる．

情報量基準のほとんどは，以上で述べたように，エントロピーとパラメータ数を調節して与えられる．もうひとつ代表的な情報量基準として，**MDL(Minimum Description Length) または BIC(Baysian Information Criterion)** がある．MDL(BIC) は，

$$\mathbf{MDL(BIC)} = -\, l_{like} + \frac{1}{2} \cdot m_{free} \cdot \log N \tag{2.14}$$

で与えられる．ここで，N はパラメータ推定に用いたデータ数である．右辺第 1 項は，AIC 基準と変わらないが，第 2 項で $\log N$ 倍されている分，パラメータの増加に対するペナルティが強いと解釈できる．

適用方法

情報量基準は，モデル選択基準として適用することが考えられる．例えば，財務指標の組み合わせを考える**変数選択**の場合である．また，AIC 基準は，その値が小さいほど，**予測誤差が小さいモデル**と解釈される．したがって，最尤推定法を用いたモデルであるならば，予測誤差を計測する指標となる．

長所・短所

尤度比でも述べたように，情報量基準も対数尤度とパラメータ数が求まれば，すぐにモデル全体の説明力を計測することができる．また，AIC 基準を用いるならば，モデルの予測誤差も計測することができる．

ただし，情報量基準では，数値そのものに意味がなく,「このぐらいの数値であれば，いいモデルである」という数値基準はない．複数のモデルに対して情報量基準を計算し，その数値を比較することで，どのモデルがよいか判断する．差が 1 以内であれば，同程度の予測誤差を持つモデルであり，1 以上ならば，より小さい数値を持つモデルが予測誤差の少ないモデルと判断する．

AIC 基準では，データ数を N とする場合，モデルに取り込めるパラメータ数は，高々$2\sqrt{N}$ 個までとなり，それ以上の場合は AIC 基準そのものの信頼性が低下する．また，AIC 基準はパラメータの増加に対するペナルティが考えられているだけで，尤度比と同じくオーバーフィッティングに関しては考慮されていない．したがって，オーバーフィッティングが発生している場合には適用しないほうが安全である．

利用例

[山下, 川口 (2003)] は CRD 運営協議会によって作成された企業数のべ 948754 件, 財務諸表項目数 93 項目という規模をもつ中小企業信用データベース[4]を用いて二項ロジットモデルを AIC が良好

[3] Appendix 参照

[4] 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

となるような変数を選択して作成した. さらに, 全件データと業種や規模といったセグメントによってセグメント分けしたデータでの推定精度を AIC を用いて比較した. その結果, セグメント分けした場合のほうが推定精度がより良好な結果を得ることができたと報告している. また, AIC を規準として最適なデータ量とセグメント数の関係について, 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.

2.4 クロスバリデーション法

目的

尤度比や情報量基準は，推定に用いたデータに対する当てはまりのよさを比較して，モデルを評価する方法であった．しかし，推定に用いたデータに対してよいモデルであると判断できても，運用結果のデータに対するデフォルト予測があたっていなければ，そのモデルがよいモデルであると判断できない．したがって，推定に用いたデータに対して当てはまりのよさだけでは絶対的な評価ができないことがいえる．そこで，データを推定用と検証用の二つに分割し，検証用データを擬似的な運用結果のデータと見なして，モデルの予測精度を比較する評価方法を考える．これを**クロスバリデーション法 (Cross Validation Method)** とよぶ.

成り立ち

クロスバリデーション法では，**擬似的な**運用結果のデータベースで検証を行うことを考える．いま，データベースを**推定用データベース**と**検証用データベース**との二つに分けても，推定に必要なデータが確保されていると仮定する．推定用データベースでモデルを作った後，検証用データベースのデフォルト予測を,「運用データベースのデフォルト確率予測」と見なし，その予測精度の比較によって，モデルを評価する.

適用方法

推定用データを用いてモデルを作成する．作成したモデルから，検証用データのデフォルト確率を計測する．検証用データの結果（デフォルトした，または，しなかったかの 2 値データ）を用いて，計測されたデフォルト確率とその結果が合致しているか，AR や AUC・ジニ係数などの評価指標を用いてモデルを評価する.

長所・短所

クロスバリデーション法を適用するには，推定用データと検証用データの二つに分割しても，推定に必要なデータが確保されなければならない．また，データの分割方法により結果はかなり異なってしまう．実は，推定用のデータと検証用のデータが同質であるかどうかを評価していることと同値である．例えば，推定用データと検証用データが同質であればモデルの予測精度や安定性が低くともよい結果を得ることができる．したがって，クロスバリデーション法では，評価の目的である予測精度の比較を十分に行えない.

利用例

[Deakin(1972)] は倒産・非倒産を判別する判別分析を行い, 1964 年から 1970 年の間に倒産した

かし, その情報を会員が利用できるようになっている.

企業のデータを推定用とし, 1963 年から 1964 年の間に倒産した企業のデータを検証用サンプルとして 5 年先までの倒産を予測した. その結果, クロスバリデーション法による精度の悪化は殆ど無かったが, 推定データ以外のデータに対するモデルの妥当性について, 継続的な観察が必要であると結論づけた.

2.5 ジャックナイフ法とブートストラップ法

目的

クロスバリデーション法では，元データから検証用データを作り出すことで，運用データに対する検証を擬似的に行えた．しかし，1 回の検証だけでは検証用データの作り方によってばらつきが生じ，高い信頼性が得られない．そこで，検証用データを複数作成することで，モデル検証者が評価したいパラメータや統計指標の分布を推定し，モデルを評価する方法を考える．これを，**統計的リサンプリング法**とよぶ．統計的リサンプリング法では，**ジャックナイフ法 (Jackknife Method)** と**ブートストラップ法 (BootStrap Method)** が代表的な方法である．

成り立ち

【ジャックナイフ法】

ジャックナイフ法は，元データベースから**非復元抽出**を行い，評価したいパラメータや統計量の分布を推定するために必要なデータベースを作成し，推測した分布を用いてモデルの予測精度や安定性を検証する方法である．

ジャックナイフ法は，以下の手順で行われる．

(1) 元のデータベースを，D 個のデータベースに分割する．
(2) $D-1$ 個のデータベースを用いてモデルを作成し，残された一つのデータを検証用データとして用いて，評価したいパラメータや統計量を計測する．
(3) D 個のデータベースに対して，順に (1)，(2) の作業を繰り返す．

以上のようにして，評価したいパラメータや統計量のサンプルが D 個得られたとする．得られた D 個のサンプルを用いて分布を推定する．

【ブートストラップ法】

ブートストラップ法は，元データベースから**復元抽出**を行い，評価したいパラメータや統計量の分布を推定するために必要なデータベースを作成し，推測した分布を用いてモデルの予測精度や安定性を検証する方法である．

ブートストラップ法は，以下の手順によって行われる．

(1) 元のデータベースから，復元抽出（同じデータを複数回抽出することを許す）を行い，データベースを作成する．
(2) 作成した検証用データベースを用いて，評価したいパラメータや統計量を計測する．
(3) この作業を繰り返し，必要としている統計量の分布を推定する．

以上のようにして，評価したいパラメータや統計量のサンプルが P 個得られたとする．得られた P 個のサンプルを用いて分布を推定する．

適用方法

例えば，推定されたパラメータは以下のようにして評価できる．全データを用いて推定されたパラメータを θ，統計的リサンプリング法で得られたパラメータの分布の平均値を $E[\hat{\theta}]$，分散を $E[\hat{V}]$ と仮定する．このとき，推定されたパラメータの値は，どのようなデータを用いて推定しても同じ値になるはずであり，あるデータに対してパラメータの値が大幅に乖離したならば，そのモデルの安定性は低いと判断できる．したがって，θ の値が $E[\hat{\theta}]$ の値に近く，$E[\hat{V}]$ が小さいならば，そのモデルは安定性の高いよいモデルであると判断できる．また，推定されるパラメータのほかにも，デフォルト確率の予測精度をはかる AR や AUC・ジニ係数などの分布を推定することも考えられる．

ジャックナイフ法およびブートストラップ法の違いは，元データからのサンプリングを**非復元抽出（ジャックナイフ法）**とするか**復元抽出（ブートストラップ法）**とするかの違いであり，**評価したいパラメータや統計量の分布を推定する**という目的においては一致している．

長所・短所

統計的リサンプリング法は，母集団分布を仮定しないノンパラメトリックな方法であり，高い適用性をもつ．また，推定された分布を用いて，評価したいパラメータや統計量の平均値や，それらの値のぶれ（分散）を計測するので，モデルの安定性を精緻に評価できる．以上のような長所を持つので，推奨する評価方法のひとつとなる．しかし，統計的リサンプリング法を適用する場合，膨大な計算時間を必要とする．したがって，計算処理速度の高いコンピュータを用いなければならない．

2.6 CAP 曲線, AR, Somers の D

目的

デフォルト確率を算出するモデルを使ってデフォルト予測をし，その予測結果を評価したい場合，特に，**運用データ全体のうち，デフォルト予測に失敗した割合**について調べたい場合は **CAP 曲線**および **AR**(**Somers の D**) を適用する．

成り立ち

【CAP 曲線】

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_D 件とする．グラフの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_D をプロットする．この曲線を **CAP 曲線**と定義する．

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 曲線は図 2.1 における C の $y = x$（45 度線）上にプロットされる．また，推計デフォルト確率が高いほうから順にデフォルトしたならば，グラフの形状は図 2.1 における A のように期待される．

【AR・Somers の D】

推計デフォルト確率が高い順にデフォルトした場合の CAP 曲線 (図 2.1 における A の CAP 曲線) と，全く説明力がないモデルが描く CAP 曲線 (図 2.1 における C の CAP 曲線) とで囲まれる面積を A_p，実際にモデルが描く CAP 曲線 (図 2.1 における B の CAP 曲線) と全く説明力がないモデル

が描く CAP 曲線 (図 2.1 における C の CAP 曲線) とで囲まれる面積を A_R とする．このとき,

$$AR = \frac{A_R}{A_p} \tag{2.15}$$

で表される値を **AR(Accuracy Ratio)** と定義する．(図 2.2 を参照)

信用スコアが低い順に高いデフォルト確率を与えるモデルの場合, AR は次の d と同じものである.

$$d = P(Z_D < Z_{ND}) - P(Z_D > Z_{ND}). \tag{2.16}$$

これは **Somers の D** と呼ばれる.

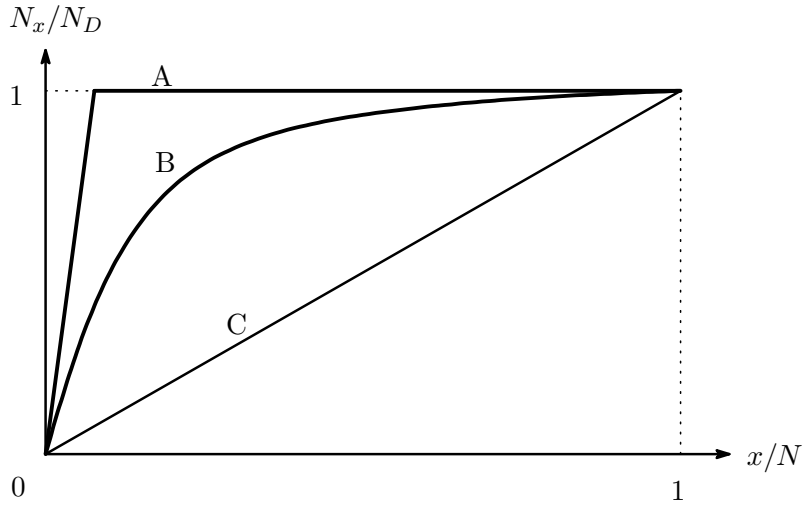


図 2.1: **CAP 曲線（予測が完全に当たった場合は** A**，モデルに説明力がなくデフォルトがランダムに起きた場合は** C **のように** CAP **曲線が描かれる．実際は，**B **のような曲線を描くことが多く，これが** A **の曲線に近いほど予測結果がよいと判断できる.）**

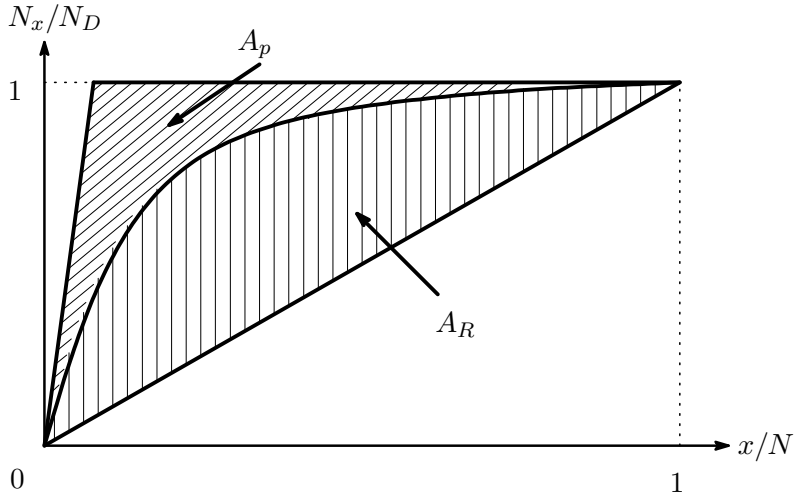


図 2.2: **AR （完全に予測した場合の** CAP **曲線と直線** $y = x$ **によって囲まれる領域の面積を** A_p**，実際の** CAP **曲線と直線** $y = x$ **によって囲まれる領域の面積を** A_R **としたとき，**A_R/A_P **で表される量．**1 **に近いほどモデルの予測結果がよいと判断できる．）**

適用方法

モデルの CAP 曲線が図 2.1 における A の曲線に近いかどうかを比較することで，予測結果が良好なものか判断できる．しかし，複数のモデルを比較する場合，それぞれの CAP 曲線だけを見て，どれが一番よいモデルであったかを判断するのは困難である．その場合は，AR を用いるとよい．

予測結果がよいモデルであるならば，CAP 曲線は A の CAP 曲線に近づくので，A_R は A_p に近づく．また，モデルの説明力が低ければ，CAP 曲線は C の CAP 曲線に近づくので，A_R は 0 に近づく．したがって，AR のとりうる値の範囲は，

$$0 \leq AR \leq 1 \tag{2.17}$$

であり，1 に近いほどモデルの予測結果がよいと判断できる．

長所・短所

CAP 曲線は，予測デフォルト確率がわかれば，それを高い順にソートすることで，比較的簡単に曲線が描ける．しかし，予測デフォルト確率の順位のみに着目し，確率の値そのものの情報が消えてしまっている．そのため，予測デフォルト確率の水準を大きく誤っていても AR は悪化しない場合があり，適用には注意が必要である．

利用例

[Moody's(2001)] は同社が保有する 1994 年から 2000 年までの 1143 社の倒産企業を含む計 41557 社の日本国内の民間企業によるデータベースを元に提供している信用リスク計量モデルを CAP・AR を用いて評価した．その結果, [Altman(1977)] などのモデルと比較して予測力がより高いことを示した.

[山下, 川口 (2003)] は中小企業信用データベースを用いて作成した二項ロジットモデルを全件データと業種や規模といったセグメントによってセグメント分けしたデータでの推定精度を AR を用い

て比較した．その結果, セグメント分けした場合のほうが推定精度がより良好な結果を得ることができたと報告している．

2.7 N/S 比

目的

デフォルト判別において，デフォルトすると予測した企業の何%がデフォルトしたか**（正しく判別できた割合）**，また，デフォルトしないと予測した企業の何%がデフォルトしたか**（間違って判別した割合）**，その判別力を計測したい場合，**N/S 比**が適用できる．N/S 比は，**配置表**と関連が強く，ここでは配置表についても説明する．

成り立ち

【配置表】

運用データに対して推計デフォルト確率を算出し，**「ある値 C を決め，C より低いデフォルト確率を持つ場合にデフォルトしない，C より高いデフォルト確率を持つときにデフォルトする」**，と予測したとする．ここで，C を**切断点 (cut-off point)** とよぶ．

以上のようにデフォルト判別予測をした場合，予測結果は，次の 4 つのケースが考えられる．

(1) デフォルトすると予測して，実際にデフォルトした．(TP:True Positive)
(2) デフォルトすると予測して，実際にデフォルトしなかった．(FP:False Positive)
(3) デフォルトしないと予測して，実際にデフォルトした．(FN:False Negative)
(4) デフォルトしないと予測して，実際にデフォルトしなった．(TN:True Negative)

予測した全企業に対して，各企業が (1) から (4) のどのケースに該当するかを調べることで，それぞれのケースに該当する企業数がわかる．その結果をわかりやすく表であらわしたのが，表 2.1 で示されている**配置表 (Classfication Table)** である．

表 2.1: **配置表** (Classfication Table)

		結果	
		デフォルトした	デフォルトしなかった
予測	デフォルトする	TP(1)	FP(2)
	デフォルトしない	FN(3)	TN(4)

【N/S 比】

デフォルト予測と，その結果が表 2.2 の配置表で得られたとする．

デフォルトすると予測した企業数を N_d，デフォルトしないと予測した企業数を N_{nd} とする．デフォルトと予測し，実際にデフォルトした企業数 TP をデフォルトすると予測した企業数 N_d で割った値を **S(Signal)**，デフォルトしないと予測し，実際にデフォルトした企業数 FN をデフォルトしないと予測した企業数 N_{nd} で割った値を **N(Noise)** とする．つまり，

$$\mathbf{N} = \frac{FN}{N_{nd}} \qquad \mathbf{S} = \frac{TP}{N_d} \tag{2.18}$$

表 2.2: **配置表**

		結果	
		デフォルトした	デフォルトしなかった
予測	デフォルトする（N_d）	TP	FP
	デフォルトしない（N_{nd}）	FN	TN

である．**NS 比**は，以上で求めた N, S を用いて，

$$\mathbf{NSR} = \frac{\mathbf{N}}{\mathbf{S}} \tag{2.19}$$

で定義される値である．

適用方法

N/S 比は，判別分析を代表とする，デフォルト判別モデルの判別力を比較したい場合に適用できる．予測結果が完全に的中したならば，TP は N_d，FN は 0 となるので，N/S 比の値は 0 になる．一方，全て外した場合は，TP が 0 となり N/S 比の値は発散する．したがって，N/S 比のとりうる値の範囲は，

$$\mathbf{0} \leq \mathbf{NSR} < \infty \tag{2.20}$$

となり，**0 に近いほど判別力が高い**と判断できる．

長所・短所

一般に，与えた判別点に対して配置表の各セルに代入される値は異なるので，N/S 比も判別点によって値が異なる．例えば，デフォルトした企業とデフォルトしなかった企業が 50%ずつ含まれるデータベースならば，判別点 C を 50%にするであろう．しかし，一般のデータベースではデフォルトした企業数が少ないので，上記の例は特殊であるといえる．このように，N/S 比ではデータベースが持つべき条件があり，適用範囲が限定されてしまう．また，判別点のとり方に基準はないので，恣意性の強い評価指標となる．N/S 比は ROC 曲線と同様，**モデルの誤判別がどのくらいであったか**を計測する評価指標であるので．恣意性の高い N/S 比を用いて評価するのではなく，ROC 曲線および AUC を用いて評価するほうがよい．

2.8 ROC 曲線および AUC・ジニ係数

目的

デフォルト確率を算出するモデルを使って，デフォルト予測をし，その予測結果を評価したい場合，特に，**デフォルトと予測してデフォルトしなかった割合，および，デフォルトしないと予測してデフォルトした割合**について調べたい場合は **ROC 曲線**および **AUC・ジニ係数**を適用する．

成り立ち

【ROC 曲線】

配置表において，TP，FN を実際にデフォルトした企業数 N_D で割った値をそれぞれ $TPR(C)$, $FNR(C)$ とし，FP，TN を実際にデフォルトしなかった企業数 N_{ND} で割った値を $FPR(C)$, $TNR(C)$ とする．これらの値は，切断点 C によって値が変化する．(表 2.3 を参照とする)

表 2.3: **配置表** (実際にデフォルトした企業数，デフォルトしなかった企業数で割った場合)

		結果	
		デフォルトした（N_D）	デフォルトしなかった（N_{ND}）
予測	デフォルトする	$TPR(C) = TP/N_D$	$FPR(C) = FP/N_{ND}$
	デフォルトしない	$FNR(C) = FN/N_D$	$TNR(C) = TN/N_{ND}$

ROC 曲線は，横軸の $[0, 1]$ 区間に，**非デフォルト企業**のなかで，推計デフォルト確率の高い j 番目までの企業の全体に占める割合 j/N_{ND} をプロットする．縦軸には，その j 番目の企業の推計デフォルト確率より，低い推計デフォルト確率を持ったデフォルト企業の数 d_j を求め，d_j をデフォルトした企業数 N_{ND} で割った値 d_j/N_D をプロットする．つまり，ROC 曲線は関数，

$$\begin{cases} x(j) = j/N_{ND} & (0 \leq j \leq N_{ND}) \\ y(j) = d_j/N_D & (0 \leq j \leq N_{ND}) \end{cases}$$

によって描かれる曲線となる．(図 2.3 参照) これは $FPR(C)$ と $TPR(C)$ に対して，C を $-\infty < C < \infty$ の範囲で動かして得られる軌跡と定義することも可能である．

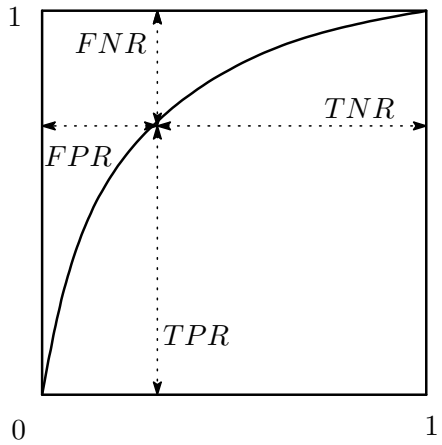


図 2.3: **ROC 曲線（x 軸はデフォルトしなかった企業に対する誤判別率** (FPR) **と成功率** (TNR)**，y 軸はデフォルトした企業に対する誤判別率** (FNR) **と成功率** (TPR) **を表している．）**

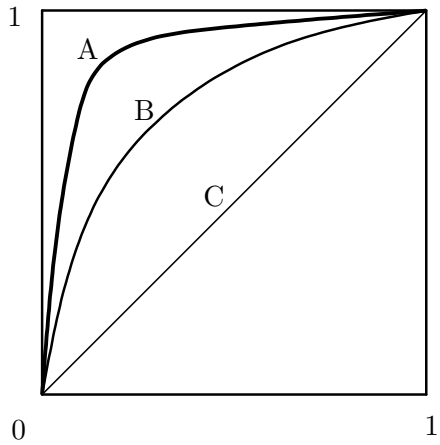


図 2.4: **ROC 曲線の判別力**（A の ROC 曲線と B の ROC 曲線では，A のほうがどの値の切断点をとっても，B より誤判別が少ない．また，C の ROC 曲線は，モデルに説明力がなく，デフォルトがランダムに起こる場合に描かれる．）

【AUC・ジニ係数】

ROC 曲線，直線 $y = 0$（x 軸）および，直線 $x = 1$ で囲まれる図形の面積を，**AUC(Area Under the Curve)**[3]と定義する．AUC に直線 $x = 1$ の下側の面積 (0.5) を除いたものは**ジニ係数**として知られている．従って両者は同じ概念である．

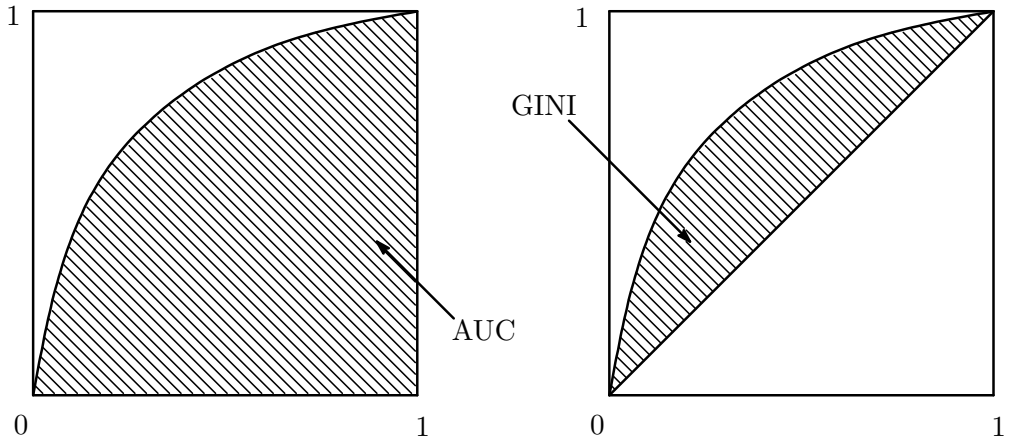


図 2.5: **AUC・ジニ係数**（ROC 曲線と x 軸，直線 $x = 1$ で囲まれる面積が AUC である．AUC のとりうる値は $0.5 \leq AUC \leq 1$ であり，1 に近いほど判別力の高かったモデルであると判断できる．ジニ係数 (GINI) はここから 0.5 を除いたものであり，$0 \leq GINI \leq 0.5$ の範囲で値をとる．）

適用方法

まず，配置表と ROC 曲線の関係について述べる．図 2.3 から ROC 曲線上の一点を決めると（つまり，切断点 C の値を決めると），x 軸はデフォルトしなかった企業に対する誤判別率 (FPR) と成功率 (TNR)，y 軸はデフォルトした企業に対する誤判別率 (FNR) と成功率 (TPR) として，配置表

[3]AUC は正式名ではない．また，AUC の正しい名称は，現在のところ付けられていない．

と関連付けられる．したがって，二つのモデルを比較する場合，モデル 1 は図 2.4 における A の曲線を描き，モデル 2 は B の曲線を描いた場合は，モデル 1 はどのような切断点 C を選んでも，モデル 2 より予測誤差が少ないモデルと判断できる．しかし，曲線が交錯した場合は，どちらがよいモデルか ROC 曲線だけでは判断できないので，その場合は AUC を用いてその数値を比較する．

モデルに説明力がなく，デフォルトがランダムに起こる場合の ROC 曲線は $y = x$ となるので，AUC は 0.5 となる．また，完全にデフォルト予測をした場合の ROC 曲線は $x = 0$ $(0 \leq y \leq 1)$ および $y = 1$ $(0 \leq x \leq 1)$ となるので，AUC は 1 となる．したがって，AUC のとりうる値は，

$$\mathbf{0.5} \leq \mathbf{AUC} \leq \mathbf{1} \tag{2.21}$$

となり，**1 に近いほど判別力の高かったモデル**であると判断できる．**ジニ係数 (GINI)** については，

$$\mathbf{0} \leq \mathbf{GINI} \leq \mathbf{0.5} \tag{2.22}$$

がとりうる値の範囲である．

長所・短所

ROC 曲線は，どんな切断点 C を設けても FP と FN の値が求まり，モデルの予測精度を**誤判別**の割合によって評価できる．したがって，デフォルト確率が算出されるモデルを用いてデフォルト予測をする場合は，ROC 曲線および AR・ジニ係数を用いて評価することを薦める．なお，AUC・ジニ係数と AR は一方が決定すれば，他方が計算できる．そのため，両方を用いてモデルを評価することにはあまり意味がない．また，AR 同様，予測デフォルト確率の順位のみに着目し，確率の値そのものの情報が消えてしまっている点にも注意する．AUC と AR との関係についていは Appendix を参照とする．

2.9 KS-値

目的

KS(Kolmogorov Smirnov) 値は, モデルのパフォーマンスや変数選択に用いるノンパラメトリックな統計量であり, デフォルトした企業とデフォルトしていない企業の信用スコアの分布がどの程度異なっているかを知りたいときに用いる．

成り立ち

KS-値はデフォルトした企業とデフォルトしなかった企業の信用スコアの分布の距離を計測する．したがって信用スコアに正規分布など特定の分布を仮定しない．

倒産企業 (D), 非倒産企業 (N) をそれぞれ N_D, N_{ND} 個ずつ含む標本に対して計算された企業の信用スコアを $Z_{Di}\,(i = 1, \cdots, N_D), Z_{NDj}\,(j = 1, \cdots, N_{ND})$ とおき, 横軸に $z\,(-\infty < z < \infty)$ をとって標本の累積分布関数 (ECDF: Empirical Cummurative Distribution Function) を描く．これはある z を固定したとき, 各標本の中で z と同じかそれ以下となる観測値の個数 $L(z)$ の, 全体の標本数に対する割合である．従って倒産企業グループの ECDF を $F_D(z)$ としたとき,

$$F_D(z) = \frac{L_D(z)}{N_D}$$

となる. これは N_D, N_{ND} が大のとき図 2.6 のようなグラフとなる. 非倒産企業グループの ECDF も同様にして求まり, これを $F_{ND}(z)$ とする.
ある z について,$|F_D(z) - F_{ND}(z)|$ の最大値を D とおく. 即ち,

$$D = \sup_{-\infty<z<\infty} |F_D(z) - F_{ND}(z)|.$$

一般に KS-値といった場合にはこの D を指す.

いま, 倒産・非倒産両方を含む全サンプルのスコア Z を小さいものから順にならびかえた $Z_D^{(i)}$ $(i = 1, 2, \cdots, N_D)$ を作り,

$$Z_D^{(1)} \leq Z_D^{(2)} \leq \cdots \leq Z_D^{(N_D)}$$

とおき, 縦軸に $F_D(z^{(i)})$ 横軸に $\frac{i}{N_D}$ をとったプロットを描く. 同様に $Z_{ND}^{(j)}$ $(j = 1, 2, \cdots, N_{ND})$ を作り, $F_{ND}(z^{(j)})$ と $\frac{j}{N_{ND}}$ のプロットを描くと, N_D, N_{ND} が大のとき, 二つのプロットが図 2.7 のようなレンズ型の曲線を描くため, 直感的に理解が容易である. 勿論この操作を行ったとしても D の値は変化しない.

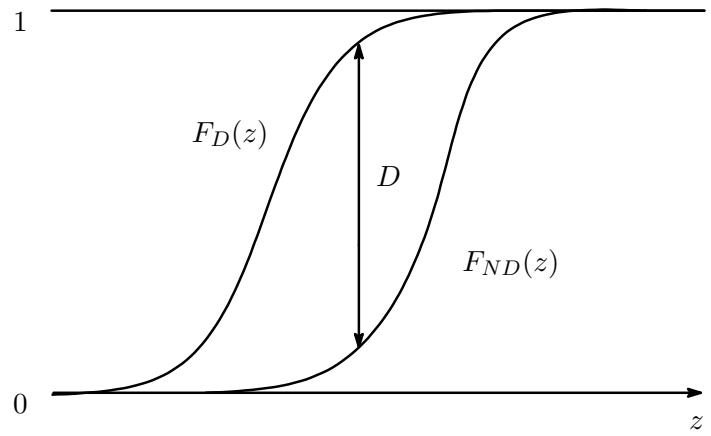


図 2.6: **累積分布関数** (**横軸は** $(-\infty < z < \infty)$ **の範囲をとり得る**.)

適用方法

D を用いて Z_D の分布と Z_{ND} の分布がどの程度異なっているかを数値的に評価し, 判別が有効に行われたかの判断に用いる.

長所・短所

統計学的には, KS-値は特定の分布に依存しない長所がある. 一方, 分布の中心付近で差異がある場合, 他の部分で差異がなくとも大きい値となる傾向がある.

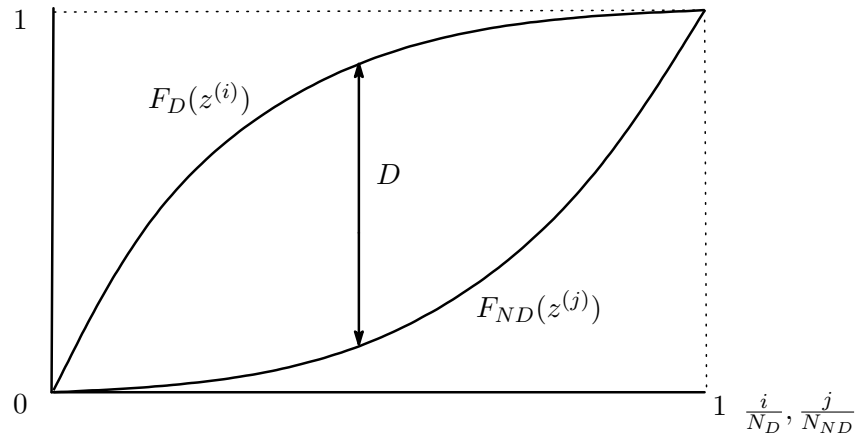


図 2.7: **レンズ型の累積分布関数** (**横軸は** $(0 \leq i/N_D, j/N_{ND} \leq 1)$ **の範囲をとり得る**. **二つのプロットがレンズ型の曲線を描くため**, **直感的に理解が容易である**.)

信用リスクモデルの評価を行う上では KS-値は ECDF を描くことで標本の直感的な把握が容易となる. しかし KS-値が大きくなったとしても, それが平均の違いによるものか, 分散か, 中央値の違いか, 密度関数のクラスが違うのか等, 原因については何も言っていない点に注意が必要である. また,KS-値は標本の数が少ないときには有効に機能しない.

利用例

データ・フォアビジョンは 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を対象としたロジットモデルに基づく信用スコアリングモデルの高度化事業において, 同社の高度化によって新たに作成された信用スコアリングモデルの KS-値が良好であることを示した. さらに, 高度化前との比較においても KS-値の改善を示した.

また, 新日本監査法人は福岡銀行の債務者格付けモデルに関する有効性の評価を行い, その際モデルの効率性の確認のために当該信用格付けモデルの KS-値の計算を実施している.

2.10 ダイバージェンス

目的

デフォルト判別において，デフォルトしたスコアの分布とデフォルトしなかったスコアの分布とがどれだけ離れているか，すなわち，それぞれの分布の乖離度を計測したい場合，**ダイバージェンス**を適用する.

成り立ち

各企業に信用リスクスコアが与えられていると仮定する．デフォルトした企業のスコア分布に対して，平均を μ_D，分散を V_D とする．同様に，デフォルトしなかった企業のスコア分布に対して，平均を μ_{ND}，分散を V_{ND} とする．このとき，**ダイバージェンス**を，

$$\mathbf{Div} = \frac{(\mu_D - \mu_{ND})^2}{V_D + V_{ND}} \tag{2.23}$$

で定義する.（図 2.8 参照）

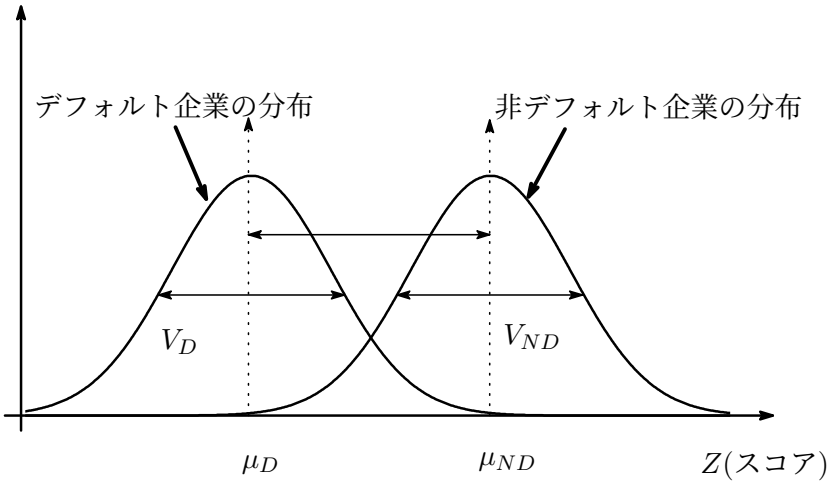


図 2.8: **デフォルト企業のスコア分布と非デフォルト企業のスコア分布（平均の差が大きく，それぞれの分布の分散が小さくなれば，ダイバージェンスの値は大きくなる）**

適用方法

判別分析や，各企業に信用リスクスコアが与えられるモデルに対して適用できる．また，デフォルト確率が算出されるモデルでは，ある確率以下ではデフォルトするという**判別点**を設けることで適用可能である．アウトサンプルデータでは，デフォルトしたか，しないかの結果がわかっているので，それに従い分布を作成すれば，ダイバージェンスを適用できる．

それぞれの分布が乖離している場合，平均の差は大きく，また，分散が小さくなる．したがって，(2.23) 式からダイバージェンスは値が**大きいほどそれぞれの分布が乖離している**と判断できる．また，$\mu_D = \mu_{ND}$ となるとき，ダイバージェンスは最小値 0 をとる．したがって，ダイバージェンスのとりうる値の範囲は，

$$0 \leq \mathbf{Div} < \infty \tag{2.24}$$

となり，値が大きいほど分布が乖離していて，デフォルト判別が良好になると判断できる．

長所・短所

ダイバージェンスは，分布の乖離度合いを表す指標なので，直感的なデフォルトと非デフォルトの判別精度を評価できる．しかし，実際に判別点を決定した場合に，どれくらい誤判別するか，その割合に関しては情報が得られない．また，格付けの場合はデフォルト確率が各ランクに与えられ，離散値をとるので精度が悪くなる．

利用例

データ・フォアビジョンは 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を対象としたロジットモデルに基づく信用スコアリングモデルの高度化事業において, 同社の高度化によって新たに作成された信用スコアリングモデルのダイバージェンスが良好であることを示した. また, 高度化前との比較においてもダイバージェンスが改善されたことを示した.

2.11 CIER

目的

デフォルトした企業のデフォルト確率が，0.6 で与えたモデルと，0.8 で与えたモデルとでは，後者のほうが前者より予測を当てたと判断できる．このように，モデルによって各企業のデフォルト確率をどれだけ 0（デフォルトする）または 1（デフォルトしない）に近づけられたかを評価する場合，**CIER(Conditional Information Entropy Ratio)** を適用する．

成り立ち

CIER は**エントロピー**に基づく指標である．ある企業のデフォルト確率が p で与えられている場合のエントロピーは，

$$E = -\Big(p\log p + (1-p)\log(1-p) \Big) \tag{2.25}$$

である．（図 2.9 参照）

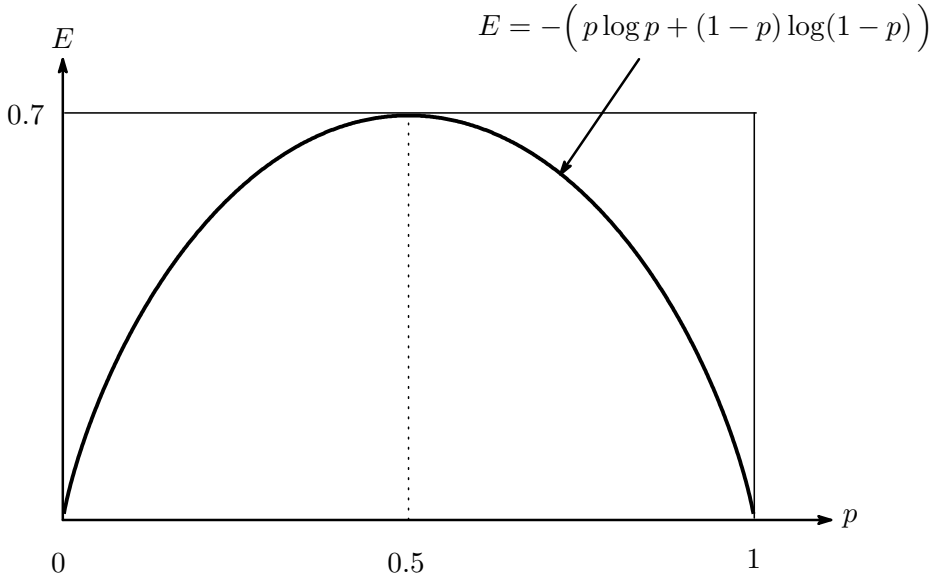


図 2.9: **エントロピーのグラフ**

いま，各格付けのデフォルト確率が $p_1, \cdots, p_K$ で表されていると仮定する．これより，各格付けに対するエントロピー $H(R_k)$ は，

$$IE(R_k) = -\Big(p_k\log p_k + (1-p_k)\log(1-p_k) \Big) \tag{2.26}$$

で与えられる．格付けされている全ての企業数を K，各ランクに属する企業数を $N_1, \cdots, N_K$ とするとき，$H(R_k)$ の加重平均 H_1 は，

$$\mathbf{IE_1} = \sum_{k=1}^{K} \frac{N_k}{N} H(R_k) \tag{2.27}$$

となる．

一方，格付けの情報がなく平均的なデフォルト率 p_d[4]のみがわかっている場合は，どの企業のエントロピーも $-\Big(\, p_d \log p_d + (1-p_d)\log(1-p_d) \,\Big)$ で与えられる．したがって，加重平均 H_0 は，

$$\begin{aligned} \mathbf{IE_0} &= \sum_{i=1}^{N} \frac{1}{N} \left(-\Big(\, p_d \log p_d + (1-p_d)\log(1-p_d) \,\Big) \right) \\ &= -\Big(\, p_d \log p_d + (1-p_d)\log(1-p_d) \,\Big) \end{aligned} \tag{2.28}$$

となる．

以上のようにして得られた H_0，H_1 を用いて，**CIER** を

$$\mathbf{CIER} = \frac{\mathbf{IE_0} - \mathbf{IE_1}}{\mathbf{IE_0}} \tag{2.29}$$

で定義する．

適用方法

CIER で重要な項は，$\mathbf{IE_1}$ である．完璧なモデルであれば，デフォルト確率を，デフォルトする企業には 1，デフォルトしない企業には 0 をそれぞれ与えるので，$\mathbf{IE_1}$ は (2.25) 式から 0 となる．また，すべての企業のデフォルト確率を 0.5 で与える場合が，もっとも情報価値のないモデルとなり，CIER は約 0.7 となる．CIER では，データの平均的なデフォルト率（全データ数を N，デフォルトした企業数を n_d とした場合，デフォルト率は n_d/N）を，デフォルト確率としてすべての企業に与えた場合にくらべ，用いたモデルがどれくらいデフォルト確率を 0 または 1 に近づけているか，比を用いて評価している．ただし，デフォルト予測が当たっているかいないに関係なく，各企業のデフォルト確率を 0 または 1 に近づけるモデルであれば，CIER は 1 に近づくことに注意する．

長所・短所

CIER は，デフォルトしたか，しなかったかに関係なく，モデルがデフォルト確率を 0 または 1 にどれだけ近づけたかを評価している．例えば，ある企業がデフォルトしたとわかっている場合に，デフォルト確率を 90%で与えたモデルと 80%で与えたモデルを比較する場合に用いる．CIER では，デフォルトした企業には 1 に近いデフォルト確率を与えるモデルがよいと判断するので，デフォルト確率を 90%で与えたモデルがよいと判断する．しかし，予測結果がわからない場合，CIER は当てにならない評価指標となる．例えば，モデルがある企業にデフォルト確率を 1%と予測したときに，デフォルトした場合と，デフォルトしなかった場合との二つのケースを考える．この企業のエントロピーは，

$$E = -\Big(\, p \log p + (1-p)\log(1-p) \,\Big)$$

であるから，予測結果にかかわらず CIER は同じ値をとる．明らかにデフォルトしなかった場合のほうが予測を当てていると判断できるが，CIER ではそれを評価できない．このように，モデルの予測結果が全て正しいと仮定できない場合，CIER を適用できない．

利用例

[4]デフォルト率 p_d は，全データ数を N，デフォルトした企業数を N_D として，$\mathbf{p_d} = \dfrac{\mathbf{N_D}}{\mathbf{N}}$ と定義する．

[Moody's(2001)] は同社が保有する 1988 年から 1999 年までの 1502 件のデフォルトを含む計 30000 社の民間企業によるデータベースを元に提供している信用リスクモデルを CIER を用いて評価した. その結果, [Altman(1977)] などのモデルを同じデータベースを用いて推定したものより同社のモデルのほうが高い CIER を示したと報告している.

2.12 ブライアスコア

目的

予測デフォルト確率が，実際の結果（デフォルトしたか，しなかったかの 2 値）をどれだけ当てているか，つまり，予測と結果の合致性について評価したい場合，**ブライアスコア**を適用する.

成り立ち

デフォルト確率が与えられた企業数を N 件，それぞれの企業に対してデフォルト確率は p_i で与えられているとする．このとき，ブライアスコアは,

$$\mathbf{BS} = \frac{1}{N}\sum_{i=1}^{N}(p_i - \theta_i)^2 \tag{2.30}$$

で表される．ここで θ_i は,

$$\theta_i = \begin{cases} 1 & (\text{i番目の企業がデフォルトしているとき}) \\ 0 & (\text{i番目の企業がデフォルトしていないとき}) \end{cases}$$

なる関数である.

次に，格付け（K ランクに分けられているとする）の場合を考える．各ランクに与えられたデフォルト確率を $P_1, \cdots, P_K$ とする．同様にして，Brier スコアは (2.30) 式で与えられるが，各 p_i は $P_1, \cdots, P_K$ のいずれかの値をとる．この場合，Brier スコアは次のように分解できる.

$$\begin{aligned} \mathbf{BS} &= \frac{1}{N}\sum_{i=1}^{N}(p_i - \theta_i)^2 && (2.31) \\ &= \frac{1}{N}\sum_{k=1}^{K} n_k(P_k - \overline{\theta_k})^2 - \frac{1}{N}\sum_{k=1}^{K} n_k(\overline{\theta_k} - \overline{\theta})^2 + \overline{\theta}(1-\overline{\theta}) && (2.32) \end{aligned}$$

ここで，$\overline{\theta_k}$ はランク内での平均デフォルト率を，$\overline{\theta}$ は格付け全体の平均デフォルト率を表す.

適用方法

デフォルト確率が各企業に与えらている場合は，デフォルト確率の誤差を 2 乗平均しているので，直感的には誤差分散を評価している考えればよい．したがって，ばらつきが小さい，つまり，ブライアスコアが小さいほうがよいと判断できる.

格付けの場合の直感的な解釈を説明する．右辺第 1 項は，格付けのランク内で予測したデフォルト確率と実際の平均的なデフォルト率の適合度，第 2 項は格付けがランクをうまく分割しているか，それを評価する指標であり，第 3 項は格付け全体としての適合度を表している．したがって，Brier スコアは以上で説明した 3 因子を足し合わせて評価する指標と見なせる．結果的には，分散を評価している指標であるから，ブライアスコアが小さいほうがよいモデルであると判断する.

長所・短所

ブライアスコアでは，予測デフォルト確率が小さいほどスコアの値が小さくなる．例えば，A ランクと格付けされた企業が 1000 件，B ランクと格付けされた企業が 1000 件あったとする．そこで，次の二通りの予測を考える．

(1) A ランクの予測デフォルト確率を 0.02，B ランクの予測デフォルト確率を 0.1.
(2) A ランクの予測デフォルト確率を 0.002，B ランクの予測デフォルト確率を 0.1.

その結果，実際にデフォルトした企業数は，A ランクでは 20 件，B ランクでは 100 件であったとしよう．このとき，それぞれのブライアスコアを計算すると，(1) では 0.1099，(2) では 0.1096 となる．(2) の A ランクは「2 件の企業がデフォルトする」と予測されており，実際の結果は 20 件で 18 件もはずしている．にもかかわらす，Brier スコアの値を比較すると正確に予測した (1) の値と，18 件はずした (2) の値とでは 0.0003 の差である．つまり，ブライアスコアは，予測デフォルト確率が 0 に近い値であればあるほど，モデル間での差がわずかになり，当てにならないことがいえる．

2.13 F 検定

目的

F-検定は次の二つの用途に用いられる．一つは説明変数の個々の説明力の有無の判定, もう一つは全体の説明力の有無の判定である．

【説明変数の説明力の有無を評価する場合】

いま, 両グループをあわせたすべての企業についてそのスコアの全平均まわりの変動（総変動）は, グループ間変動 (グループ内平均の全平均まわりでの変動）と各グループ内変動 (グループ内平均まわりでの変動）の和によって表される．

デフォルトに対して有効な指標であれば, グループ間変動はグループ内変動に比較して大きいはずである．そこで, 後者に対する前者の比率をとって, この比率の分母分子がそれぞれ自由度 1, 自由度 (標本に含まれる全企業数 − 2) の χ^2 分布に従うことを利用して, F 分布に従う統計量をつくり, F 検定を行う．

【スコア関数の説明力の大小・有無を評価する場合】

全体の説明力がどの程度であるかを評価したい場合にも用いられる．倒産・非倒産を正確に判別する上では, グループ内変動に対してグループ間変動を大きくするようにグループ分けすることが求められる．そこで, 後者に対する前者の比率をとって, スコア関数の説明力を測る指標, λ が定義される．また, スコア関数の説明力の有無を判断する際には, この比率の分母分子がそれぞれ自由度 1, 自由度 (標本に含まれる全企業数 − 2) の χ^2 分布に従うことを利用し, F 検定を行う．

成り立ち

【説明変数の説明力の有無を評価する場合】

いま, 倒産グループに含まれる企業数を N_D, 非倒産グループに含まれる企業数を N_{ND} とする．ここで, 当該変数の総変動は次のように分解される．

$$\sum_{i=1}^{N_D}\left(X_{Di}-\overline{X}\right)^2+\sum_{j=1}^{N_{ND}}\left(X_{NDi}-\overline{X}\right)^2$$

$$= \left\{N_D\left(\overline{X}_D-\overline{X}\right)^2+N_{ND}\left(\overline{X}_{ND}-\overline{X}\right)^2\right\}+\left\{\sum_{i=1}^{N_D}\left(X_{Di}-\overline{X}_D\right)^2+\sum_{j=1}^{N_{ND}}\left(X_{NDj}-\overline{X}_{ND}\right)^2\right\}$$

説明変数の判別力は, 右辺第 1 項（グループ間変動）が右辺第 2 項（グループ内変動）に対して大きいほど高まる. そこで,

$$f=\frac{N_D\left(\overline{X}_D-\overline{X}\right)^2+N_{ND}\left(\overline{X}_{ND}-\overline{X}\right)^2}{\left(\sum_{i=1}^{N_D}\left(X_{Di}-\overline{X}\right)^2+\sum_{j=1}^{N_{ND}}\left(X_{NDj}-\overline{X}\right)^2\right)/(N_D+N_{ND}-2)} \tag{2.33}$$

を定義する. 分母の $N_D+N_{ND}-2$ がグループ内変動の自由度である. 分子の自由度は (グループ数 -1) で, 2 グループのみを考える場合は式に現れない[5]. この f は自由度 $(1, N_D+N_{ND}-2)$ の F 分布に従うため, F 値と呼ばれる.

【スコア関数の説明力の大小・有無を評価する場合】

スコア関数の総変動は次のように分解される.

$$\begin{aligned}&\sum_{i=1}^{N_D}\left(Z_{Di}-\overline{Z}\right)^2+\sum_{j=1}^{N_{ND}}\left(Z_{NDi}-\overline{Z}\right)^2\\ =\ &\left\{N_D\left(\overline{Z}_D-\overline{Z}\right)^2+N_{ND}\left(\overline{Z}_{ND}-\overline{Z}\right)^2\right\}+\left\{\sum_{i=1}^{N_D}\left(Z_{Di}-\overline{Z}_D\right)^2+\sum_{j=1}^{N_{ND}}\left(Z_{NDj}-\overline{Z}_{ND}\right)^2\right\}.\end{aligned}$$

スコア関数による判別力は, 右辺第 1 項（グループ間変動）が右辺第 2 項（グループ内変動）に対して大きいほど高まる. そこでスコア関数によって推定された信用スコアの判別力を測る指標として λ を定義する.

$$\lambda=\frac{N_D\left(\overline{Z}_D-\overline{Z}\right)^2+N_{ND}\left(\overline{Z}_{ND}-\overline{Z}\right)^2}{\sum_{i=1}^{N_D}\left(Z_{Di}-\overline{Z}\right)^2+\sum_{j=1}^{N_{ND}}\left(Z_{NDj}-\overline{Z}\right)^2} \tag{2.34}$$

これを大きくするスコア関数が判別に優れている. ただし $N_D=N_{ND}$ のとき,

$$\lambda=\frac{\frac{1}{2}\left(\overline{Z}_D-\overline{Z}_{ND}\right)^2}{V_D+V_N}=\frac{1}{2}\mathbf{Div}$$

となり, λ による評価はダイバージェンスによる評価と同じになる.

また, スコア関数によって推定された信用スコアが倒産グループと非倒産グループで有意に異なるかどうかを知る際の F-値は次のように計算される. λ の分母分子はそれぞれ χ^2 分布に従うから, 自由度を考慮して第 1 項と第 2 項の比率を,

$$f=\frac{N_D\left(\overline{Z}_D-\overline{Z}\right)^2+N_{ND}\left(\overline{Z}_{ND}-\overline{Z}\right)^2}{\left(\sum_{i=1}^{N_D}\left(Z_{Di}-\overline{Z}\right)^2+\sum_{j=1}^{N_{ND}}\left(Z_{NDj}-\overline{Z}\right)^2\right)/(N_D+N_{ND}-2)} \tag{2.35}$$

と定義する. この f は自由度 $(1, N_D+N_{ND}-2)$ の F 分布に従う.

[5] 3 グループ以上を設定して判別を行う場合は, 分子の自由度を (グループ数 -1) とすることで同様に対応する.

適用方法

【説明変数の説明力の有無を評価する場合】

ある説明変数の候補変数の説明力の有無を検定するために F-値を用いる. F-値が 1 より小さい値をとるとき, 説明力がないと判断できるため, F-値が 1 より大であるか検定する. このときの検定は片側検定となり, 有意水準 α のもとでの F 分布の値を F_α と置くと,

$$F_\alpha \leq f \tag{2.36}$$

のとき帰無仮説を棄却する. 帰無仮説が棄却されれば, F-値は 1 より大と判断され当該変数は説明力があると判断され, 説明変数の候補として適切である.

【スコア関数の説明力の大小・有無を評価する場合】

推定されたいくつかのスコア関数がどの程度の説明力を持っているかを評価したい場合には λ や F-値を用いて, その数値が大きいほどそのスコア関数が優れていると言える.

また, 説明力の有無を検定したい場合には F-値を用いる. F-値が 1 より小さい値をとると説明力はないと判断できるため, F-値が 1 より大であるか検定する. 有意水準 α のもとでの F 分布の値を F_α と置くと,

$$F_\alpha \leq f \tag{2.37}$$

のとき, 帰無仮説を棄却する. 帰無仮説が棄却されれば, F-値は 1 より大と判断されスコア関数は説明力があると言える.

長所・短所

F-検定を個別の説明変数の選択に用いる場合, 個別変数の説明力によって変数を選択することができることから, F 検定を行って変数を選択することは有効である. また, スコア関数の説明力の大小・有無を評価する場合, 安定性については考慮されていない点に注意する. 通常, モデルに含める説明変数を追加することで F-値は非減少であるが安定性は損なわれる.

F-値を判別分析に利用する場合は, グループ数が 2 つ以上に増加した場合も対応することができるが, グループ間の順序性や, 複数あるグループの中の特定のペアに関しては何ら情報は得られない. また, 異なるグループ数を設定しているモデル同士を比較することはできない. そして F-値を用いる際には対象データの各グループ内での等分散性・正規性を要求する点に注意が必要である.

利用例

[Altman(1968)] は倒産・非倒産を判別する判別分析において, 個別変数の説明力をテストするために F 検定を行い, 有意水準 1% で有意でないとされた変数に関しては説明変数として採用しなかった. また, モデル全体の判別力をテストするためにスコア関数に対しても F 検定を行い, その結果モデルが有意であると結論づけた.

[Deakin(1972)] も同様の判別分析を行い, モデル全体の判別力をテストして倒産 1 年前から 5 年前までの判別スコア関数について有意水準 1% 以上で有意であるとの結論を得た.

2.14 二項検定

目的

格付けのあるランクにおけるデフォルト確率が p と予測され，実際にデフォルトした企業数が k 件であったと仮定する．このとき，予測されたデフォルト確率 p が，実際にデフォルトした企業数 k を予測できたといえるかどうか，つまり予測誤差の範囲内であるかを評価したい場合，**二項検定**を適用する．

成り立ち

格付けのあるランクに属する企業数を n とする．いま，このカテゴリーのデフォルト確率 p を p_0 と予測したが，実際にデフォルトした企業数は k 件であったとする．(ただし，デフォルトは**独立して起きる**と仮定する)

帰無仮説 H_0 と対立仮説 H_1 を，

$$H_0 : p = p_0 \qquad H_1 : p \neq p_0 \tag{2.38}$$

とおく．

有意水準を α とするとき，以下の条件を満たす最大の a および最小の b を求める．

$$P(k \leq a) \leq \frac{\alpha}{2} \qquad P(k \geq b) \leq \frac{\alpha}{2} \tag{2.39}$$

ここで，

$$P(i \leq a) = \sum_{i=0}^{a} {}_nC_i \, p_0^i (1-p_0)^{n-i} \tag{2.40}$$

$$P(i \geq b) = \sum_{i=b}^{n} {}_nC_i \, p_0^i (1-p_0)^{n-i} \tag{2.41}$$

である．計算の結果，

$k \leq a$ または $k \geq b$ ならば, H_0を棄却する
$a < k < b$ ならば H_0を採択する

と決定する．

しかし，データ数が多い場合は，標準正規分布による検定を用いることも可能である．標準正規分布による検定では，カテゴリー内のデータ数が多く，予測デフォルト確率が 0 または 1 に近くないことが必要である．この条件を満たすとき，統計量 $\mathbf{Z}$

$$\mathbf{Z} = \frac{k - np_0}{\sqrt{np_0(1-p_0)}} \tag{2.42}$$

を用いて，有意水準 α で標準正規分布による検定を行い，実際のデフォルト企業数 k が予測誤差の範囲内であるか決定できる．

適用方法

例えば，格付けのあるランクに属する企業数 n を 100 件，このカテゴリーのデフォルト確率 p を $p_0 = 0.05$ と予測し，実際にデフォルトした企業数 k は 8 件であった場合を考える．

このとき，有意水準 $\alpha = 0.05$ として計算をすると，

$$\begin{aligned} P(i \leq 0) &= P(i = 0) = 0.006 \leq 0.025 \\ P(i \leq 1) &= P(i = 0) + P(i = 1) = 0.037 \geq 0.025 \\ P(i \geq 10) &= \sum_{j=10}^{100} P(i = j) = 0.028 \geq 0.025 \\ P(i \geq 11) &= \sum_{j=11}^{100} P(i = j) = 0.012 \leq 0.025 \end{aligned}$$

となり，$0 < k < 11$ ならば H_0 を採択できる．デフォルトした企業数 k は 8 なので，有意水準 95% では有意に起こりうると判断できる．

長所・短所

格付けのカテゴリー内のデータ数が少数である場合，検定ができなくなるという短所がある．しかし，各カテゴリーにおける予測デフォルト確率が妥当であるか評価する場合，二項検定は有力な評価指標であり，推奨する評価方法である．

また，デフォルトに相関がある場合の二項検定については,「資産の相関係数から，デフォルトの相関係数を導き，その値に基づいて二項検定を行う」という方法が考えられている．結果は，二項検定と同様に，予測誤差の範囲が求まり，相関を考えない場合と比較して，誤差の範囲が大きくなることが言える．しかし，この計算結果の値は，デフォルトの相関係数が正しいとした場合であることに注意しなければならない．評価の順序としては，(1) デフォルトの相関係数の評価，(2) 誤差範囲の評価，となる．上述の方法は，(2) の方法であり，(1) の評価方法については，現在のところ考えられていない．したがって，デフォルトに相関がある場合，二項検定は有効でない．

2.15 Normal Test

目的

格付けにおける平均的なデフォルト率の時系列データが存在すると仮定する．各ランクに与えた予測デフォルト確率と予測結果の平均的なデフォルト率とが乖離していなかったかを検定したい場合，**Normal Test** を適用する．

成り立ち

あるランクの T 年間における予測デフォルト確率は $p_1, \cdots, p_T$ であったとする．T 年間での平均予測デフォルト確率 PD は，

$$PD = \frac{p_1 + \cdots + p_T}{T} \tag{2.43}$$

である．一方，時系列データから，t 年目において，そのランクに属する企業数が n_t 件，実際にデフォルトした企業数が N_{D_t} 件であったならば，実際のデフォルト率は N_{D_t}/n_t となる．したがって，時系列に対する平均デフォルト率 Z_T は，

$$Z_T = \frac{1}{T} \sum_{t=1}^{T} \frac{N_{D_t}}{n_t} \tag{2.44}$$

となる．また，N_{D_t}/n_t の分散を σ^2 とするとき，十分大きな T であれば，中心極限定理を用いて，

$$\frac{Z_T - PD}{\sigma/\sqrt{T}} \approx \mathcal{N}(0,1) \tag{2.45}$$

が成立する．ここで，$\mathcal{N}(0,1)$ は標準正規分布を表す．以上のようにして得られた Z_T を，有意水準 α を設けて検定する．

適用方法

格付けモデルにしか適用できない．アウトサンプルデータでの評価方法で，十分な時系列データが必要となる．

実際の検定では，α を 95%とする場合，$Z_T > 2.0$ で帰無仮説を棄却できる．つまり，$Z_T > 2.0$ **であるならば，予測デフォルト確率と実際のデフォルト確率には有意差がある**と判断する．

長所・短所

Normal Test は，正規母集団に対する母平均の検定方法を適用しているに過ぎない．分散が既知の場合は，以上で述べた方法で検定できるが，ここで用いた分散 σ^2 は標本分散なので，本来ならば **t 検定**を用いるのが普通である．また，検定では**十分大きな** T を確保することが必要となるので，5 年間〜10 年間のデータを用いて検定する場合，その結果の信頼性は低い．以上のことから，Normal Test は評価方法として不適であると考える．今後，十分な時系列データが蓄積されたならば，適用可能である．

2.16 多重比較法

目的

格付けにおいて，各ランクの平均的なデフォルト率に有意差があるか評価したい場合，**多重比較法**を適用する．

成り立ち

格付けは K 段階にランク分けされているとする．一般に，3 群以上の母平均を比較する場合，**分散分析**を用いて，帰無仮説 H_0 および対立仮説 H_1 を

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \cdots = \mu_K \\ H_1 &: \mu_1, \mu_2, \cdots, \mu_K \text{のうち少なくともひとつが異なる} \end{aligned}$$

とたて，F 統計量を用いて検定をする．格付けでは，母平均に平均的なデフォルト率を対応させる．帰無仮説が棄却された場合，どのランク間に有意差があるか，分散分析からではわからない．どのランクに有意差があるか，すべてのランク間で検定を繰り返すと，各ランク間に設定している有意水準を満たしていても，全体としての有意水準が大きくなる恐れがある．**多重比較法では，全体として必要とする有意水準をコントロールするために，それぞれ検定における帰無仮説の有意水準を調節し，有意差があるランクを見つけ出す方法**である．

多重比較法では，目的によって手法がことなるので，詳細については [永田，吉田 (1997)] を参照とする．ここでは，信用リスク評価として用いることが可能であると考えられる，いくつかの方法について説明する．

【ランク間の有意差検定】

格付けの各ランクを R_i と表し，R_i の平均的なデフォルト率が μ_i であったとする．このとき，任意の 2 組のペア (R_i, R_j) に対して，平均的なデフォルト率に有意差があるか，帰無仮説 H_0，対立仮説 H_1

$$H_0 : \mu_i = \mu_j \quad H_1 : \mu_i \neq \mu_j \tag{2.46}$$

として検定する．検定回数は ${}_KC_2$ 回を要する．例えば，20 段階にランク分けされていたならば，${}_{20}C_2 = 190$ 回検定を繰り返すことになる．${}_KC_2$ 回検定を繰り返すことで，全体で必要とする有意水準を確保して，具体的に有意差のあるランクを見つけ出す．

具体的な方法は，パラメトリック法である**テューキー (Tukey) の方法**またはノンパラメトリック法である**スティール・デュワス (Steel-Dwass) の方法**のいずれかを適用する[5]．

【順序性を想定した，対比較の有意差検定】

着目している格付けのランクと，着目している格付けより高い格付け（または低い格付け）の平均的なデフォルト率に対して，どのランクから有意差があるか検定する．注目しているランクを第 1 群，他のランクを第 2 群から第 K 群とし，それぞれのデフォルト率を $\mu_1, \mu_2, \cdots, \mu_K$ とする．ここで，これらのデフォルト率に対して，

$$\mu_1 \leq \mu_2 \leq \cdots \leq \mu_K \tag{2.47}$$

の関係が成り立っていると仮定する．

検定の進め方は，まず，帰無仮説

$$H_{\{1,2,\cdots,K-1,K\}} : \mu_1 = \mu_2 = \cdots = \mu_{K-1} = \mu_K \tag{2.48}$$

を検定し，$\mu_1 < \mu_K$ であるかを調べる．この帰無仮説が棄却できたならば，次に，帰無仮説

$$H_{\{1,2,\cdots,K-1\}} : \mu_1 = \mu_2 = \cdots = \mu_{K-1} \tag{2.49}$$

を検定し，$\mu_1 < \mu_{K-1}$ であるかを調べる．このように帰無仮説が棄却できたならば，一番外側の母平均を削って，順次検定を行ってゆく．また，途中で帰無仮説が棄却できない場合は，その時点で検定作業を終了する．

具体的な方法は，パラメトリック法である**ウィリアムズ (Williams) の方法**またはノンパラメトリック法である**シャーリー・ウィリアムズ (Shirley-Williams) の方法**のいずれかを適用する[6]．

適用方法

格付けモデルにおいて，アウトサンプルデータを用いた事後評価に適用する．具体的に，どの値で帰無仮説を棄却するかについては，Appendix を参照とする．また，以上で述べてきた 4 つの方法について，表 2.4 でまとめた．

長所・短所

パラメトリックな方法である，**テューキーの方法**と**ウィリアムズの方法**は，各ランクに属するデータが**正規分布に従い**かつ**等分散**が仮定できなければならない．デフォルトしたかしなかったかの 2

[5] Appendix **テューキーの方法**および**スティール・デュワスの方法**を参照
[6] Appendix **ウィリアムズの方法**および**シャーリー・ウィリアムズの方法**を参照

表 2.4: 多重比較法の適用例

	パラメトリック法	ノンパラメトリック法
格付け間の有意差を検定	テューキーの方法	スティール・デュワスの方法
格付けの順序性を検定	ウィリアムズの方法	シャーリー・ウィリアムズの方法

値データは二項分布と見れるので，データ数が多い場合は，中心極限定理を用いて正規分布に近似することができ，これらの方法を適用することができる．しかし，データ数が少ない場合は，正規分布に近似できないので，ノンパラメトリック法である**スティール・デュワスの方法**および**シャーリー・ウィリアムズの方法**を適用する．

ランク間の有意差検定であるテューキーの方法およびスティール・デュワスの方法は，2 組のランクについて有意差を検定したい場合に有効である．しかし，格付けの順序性に着目している場合には，有効な方法ではない．

一方，順序性を想定し，着目するランクがどのランクから有意差があるかを検定するウィリアムズの方法およびシャーリー・ウィリアムズの方法は，以下のことに注意しなければならない．例えば，4 つのランクが存在するとして，これらのデフォルト率には，

$$\mu_1 = \mu_2 < \mu_3 = \mu_4 \tag{2.50}$$

が実際に成り立っていると仮定する．検定者が想定している順序は，

$$\mu_1 \leq \mu_2 \leq \mu_3 \leq \mu_4 \tag{2.51}$$

を仮定し，第 1 群に着目してシャーリーウィリアムズの方法を適用すると,「第 1 群のデフォルト率は第 3 群以降のデフォルト率と有意差がある」と判断できるが，第 3 群と第 4 群には有意差が無いことを判断できない．シャーリー・ウィリアムズの方法は，**順序性を想定する**だけで，格付けの順序性を検定していないことに注意する．

以上のような短所があるが，多重比較法は複数のランクを対象としていることが長所といえる．例えば，格付けデータにおいてランク内のデータ数が少ない場合，二項検定は適用できない．その代用として多重比較法を適用できる．

利用例

新日本監査法人が行った福岡銀行の債務者格付けモデルに関する有効性の評価において, モデルの適切性の確認のために当該信用格付けモデルのスティール・デュワスの方法による t-値の計算を実施している．

2.17 田口の累積法

目的

高格付けから低格付けにゆくにしたがってデフォルト確率が高くなっているか，つまり格付けの**順序性**を評価したい場合，特に一回の検定で結論を導きたい場合，**田口の累積法**を適用する．

成り立ち

格付けは K 段階にランク分けされていて，高格付けのランクからデフォルト率が $\mu_1, \mu_2, \cdots, \mu_K$ であったとする．一般に，3 群以上の母平均を比較する場合は，多重比較法で述べたように，**分散分析**を用いて，帰無仮説 H_0 および対立仮説 H_1 を

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \cdots = \mu_K \\ H_1 &: \mu_1, \mu_2, \cdots, \mu_K \text{のうち少なくともひとつが異なる} \end{aligned}$$

とたて，F 統計量を用いて検定をする．格付けの場合，デフォルト率を母平均と見なし，デフォルト率には $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_K$ をしてランク付けしている．したがって，分散分析では，帰無仮説が棄却されてもデフォルト率の順序性が成り立っているかは判断できない．

田口の累積法では，対立仮説 H_1 を，

$$H_1 : \mu_1 \leq \mu_2 \leq \cdots \leq \mu_K \tag{2.52}$$

とし，**順序のある対立仮説に対する検定方法**とした．この場合，F 統計量では検出力が下がるため，異なった形の統計量を用いることで対応する方法である．

適用方法

格付けモデルにおいて，アウトサンプルデータを用いた事後評価方法として適用し，格付けの順序性を評価したい場合に用いる．具体的な方法は Appendix を参照とする．

長所・短所

田口の累積法は，帰無仮説が棄却されれば，1 回の検定で格付け全体に順序性があると判断できる．しかし，帰無仮説が棄却されない場合は，すべてのランクが同じデフォルト率であると判断される．したがって，実際にあるランク間において有意差があるとしても，それを確認することができない．このような場合は，多重比較法を用いて，ランク間の有意差を判断するしかない．

第3章　モデル・目的に合致した評価方法

本章では，第2章で説明してきた評価方法について，実際に想定されるモデルケースを用いてその適用方法について言及する．実際に想定するケースとしては，

(1) 信用リスクモデルが二項ロジットモデルで，出力がデフォルト確率であるとき，作成したモデルをパラメータ推定に用いたインサンプルデータによって評価する場合
(2) 信用リスクモデルがオーダードロジットモデルで，出力が格付けの各クラスごとのデフォルト確率であるとき，その運用結果をアウトサンプルデータ（バックテストデータ）を用いて評価する場合
(3) マクロ変数を組み込んだモデルを作成し，蓄積されている運用結果の時系列データを用いてモデルを評価する場合

の3つを考える．(1) はモデル作成者が作ったモデルを事前評価する場合，(2) は監督当局が運用結果を検査する場合（**バックテスト**）にそれぞれ対応する．(3) では，マクロ変数が組み込まれたモデルの事後評価（バックテスト）を考える．現状では，十分な時系列データが存在しないため，マクロ変数を用いたモデルを作成するのは困難である．しかし，将来的には時系列データも蓄積され，このようなモデルの評価方法が重要になることが予想される．

3.1　二項ロジットモデルをパラメータ推定に用いたデータ（インサンプルデータ）によって評価する方法

ここでは，パラメータ推定に用いたデータ（インサンプルデータ）を用いて二項ロジットモデルを作成する場合を想定する．出力はデフォルト確率である．一般的に，二項ロジットモデルでは，財務指標を用いてデフォルト確率を計測する．したがって，何らかの方法によってそれらの財務指標を決定しなければならない．本節では，推定に用いるデータから変数選択を行う場合と，外生的に変数を与える場合の二つのケースについて考える．次節に進む前に，インサンプルデータを用いて作成しているモデルを評価する場合，適用できない評価方法について考える．

Normal Test は，時系列データを必要とし，格付けにおいて，各ランクに与えられた予測デフォルト確率と実際のデフォルト率に有意差があるかどうかを検定する方法であった．また，**多重比較法**や**田口の累積法**は，格付けの順序性やランク間の有意差を検定する方法であった．したがって，これらの方法は企業別にデフォルト確率が与えられるモデルでは適用できない．

二項検定は，格付けのようにあるカテゴリーにデフォルト確率が与えられたとき，実際にいくつの企業がデフォルトしたか，その予測誤差を検証する方法である．したがって，各企業ごとに個別のデフォルト確率を与える二項ロジットモデルでは用いることができない．

以上のことから，**NormalTest，二項検定，多重比較法および田口の累積法は適用できない**．

3.1.1 変数選択を行う場合

二項ロジットモデルに用いる財務指標を選択する場合，**変数選択法**を用いることが多い．すべての財務指標の組み合わせを試して，モデルに用いる財務指標を決定すること（**総当り法**）も可能であるが，その場合，膨大な時間が必要となる．変数選択法では，複数の財務指標からどの指標がデフォルト確率に寄与するかシステマティックに選び出し，その結果，一番説明力のある財務指標の組み合わせを選択する方法である．したがって，変数選択法は**複数のモデルを評価**してモデルを選択していると見なせる．

どのような財務指標の組み合わせを選択するか（**どのモデルを選択するか**）は，評価指標である**尤度比**，**情報量基準**，**AR**，**AUC・ジニ係数**のいずれかを用いて，その評価指標が一番良い値をとる財務指標の組み合わせを選択することが一般的である．

尤度比は，最尤推定法における尤度関数の値を用いて与えられ，推定に用いるデータとモデルのフィッティングを表す評価指標である．変数をたくさん取り込むことで，データフィッティングが増し，説明力の高いモデルが作成できると期待される．しかし，注意しなければならないのは，変数をたくさん取り込むと，モデルが不安定になることである（**ロバストネスの低下**）．尤度比は，データフィッティングのみが考慮されている指標で，ロバストネスは考慮されていない．したがって，尤度比は，その値が少しでも良くなれば財務指標を取り入れる性質を持つので，すべての財務指標が取り入れられる可能性がある．一方，**情報量基準**では，

$$（\textbf{情報量基準}）=（\textbf{最大対数尤度}）+（\textbf{パラメータ数}）$$

を基本の式として，右辺第 2 項でロバストネスを考慮している．つまり，尤度（データフィッティング）が少し良くなったぐらいでは，パラメータ（財務指標）を取り込まないようにするように作られている．したがって，情報量基準はデータフィッティングとロバストネスとのバランスを考慮した評価指標となる．

一方，**AR** と **AUC・ジニ係数**はそれぞれ **CAP 曲線**，**ROC 曲線**を描いて得られる評価指標であり，モデルの予測的中率を評価する指標である．これらの評価指標では，デフォルトした企業のデフォルト確率は高く，デフォルトしなかった企業のデフォルト確率は低くなるモデルがよいモデルと判断する．したがって，基本的には尤度比と同じように，推定に用いたデータにフィッティングするモデルを選択する．

以上のことから，尤度比，AR，AUC・ジニ係数と比較して，**情報量基準**だけはモデルの安定性（ロバストネス）も考慮に入れた指標となっている．したがって，**変数選択基準として用いる評価指標としては，情報量基準を用いることが望ましい．**（図 3.1 参照）

変数選択におけるパラメータ推定方法に関して注意しなければならないのは，**パラメータ推定方法と評価方法の一致を考慮することである．**例えば，AR が一番良くなるモデルを作成したい場合に，情報量基準を用いて変数選択をすることは，パラメータ推定方法と評価方法が一致していない．この場合のパラメータ推定方法は，AR を変数選択基準として，AR が最もよくなるモデルを選択することである．しかし，実際に推定をする際，AR を最大にする推定方法として，最尤推定法を適用することができない[7]．評価方法に一致したパラメータ推定方法が確立されていないのは問題であり，新しい推定方法の確立が今後の課題となっている．

以上はモデル全体での評価であったが，選択された財務指標（変数）が有意であるかを評価したい場合は，**t-値**, もしくは **F 検定**を用いる．t-値が 2 以上であれば，その財務指標はモデルの説明力

[7] AR を最大にする推定方法は，[Eguchi and Copas(2002)] を参照.

を上げる有意な変数と見なせる．しかし，t-値が悪いからといって，その変数を取り除くことは，モデルの説明力を低下させてしまうことがある．したがって，t-値は，モデルの安定性を考慮し，モデルから財務指標を取り除くときの目安として活用すべき指標である．例えば，安定性を重視したモデルを作成したい場合は t-値を目安に変数を除けばよいし，説明力を重視したモデルを作成したい場合は t-値を無視して変数をたくさん含むモデルを作成すればよい．F 検定を用いる場合も同様に, 有意であると判断された財務指標はモデルの説明力を上げることができるが, 有意といえない場合であっても, 場合によっては変数を除かないほうがモデルの説明力が良いことがある．以上のように，**モデル作成の目的にあわせて評価方法を適用する**ことが大切である．

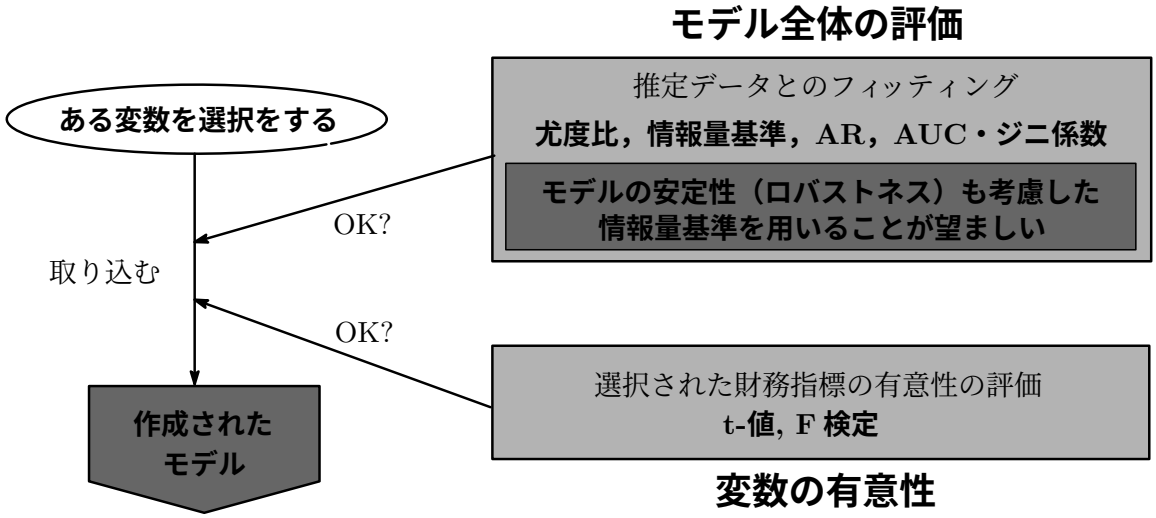


図 3.1: **変数選択法を用いる場合のモデル評価方法**

3.1.2 変数が決定している場合

ここでは，変数が決定され，モデルが一つに決まっている場合について考える．まず，**AIC 基準**であるが，AIC 基準の値そのものには意味がなく，複数のモデルを比較する場合に対して有効な指標である．したがって，モデルが一つしかない場合の評価方法としては適用できない．それに対して，**AR**，**AUC**, **KS-値**は 1 に近いほど的中率がよいと判断できる指標であり，モデルの的中率という観点からデータフィッティングを評価する方法である．ただし，AR や AUC・ジニ係数および KS-値は，モデルの的中率を推計デフォルト確率の順序性に注目して評価するので，デフォルト確率の値が水準より大きく乖離しても，悪化しないことに注意が必要である．

一方，推計デフォルト確率の水準に着目しデータフィッティングを評価したい場合は，**尤度比**，**ダイバージェンス**，**ブライアスコア**が適用できる．

尤度比は，推計デフォルト確率と推定に用いたデータとのデータフィッティングを評価する指標である．デフォルトしたかしなかったかの情報で生成されるモデルの対数尤度を i_{init}，モデルの対数尤度を l_{opt} とすることで値が得られる．

ダイバージェンスはデフォルト確率ではなく，信用リスクスコアの分布を対象に，分布の平均値と分散を用いた指標である．二項ロジットモデルでは，信用リスクスコアを用いてデフォルト確率

を計測するので適用可能である．分布の分散を用いるので，モデルが推計デフォルト確率の水準から乖離すると，ダイバージェンスの値は悪化する．

ブライアスコアは，予測デフォルト確率が実際の結果をどれだけ当てているか，予測デフォルト確率の誤差を 2 乗平均して誤差分散を評価する指標である．したがって，推計デフォルト確率が水準から乖離するとブライアスコアは悪化する．また，モデルのパラメータ推定は，デフォルトしなかった企業のデフォルト確率は 0 に，デフォルトした企業のデフォルト確率は 1 に近づくようにモデルを作成するので，ブライアスコアをインサンプルデータで用いる場合，データフィッティングを測る指標とみなせる．

したがって，データフィッティングを評価する場合，AR と AUC・ジニ係数のうちから一つ，尤度比，ダイバージェンス，ブライアスコアから一つの 2 指標で評価すればよい．

CIER は，用いた推定データのデフォルト率（全データ数を N，デフォルトした企業数を N_D とした場合，デフォルト率は N_D/N）を各企業のデフォルト確率に割り当てる場合に比べ，モデルを用いて各企業にデフォルト確率を与える場合では，どれだけデフォルト確率を 0 および 1 に近づかせることができたかを表す指標である．しかし，的中率のよいモデルであれば，デフォルトしなかった企業のデフォルト確率は 0 に，デフォルトした企業のデフォルト確率は 1 に近づくので，CIER はデータフィッティングを評価していると見なせる．また，CIER の場合，各企業のエントロピーは，

$$E = -\Big(p \log p + (1-p) \log(1-p) \Big)$$

で定義された．この場合，デフォルト確率が 0.3 でデフォルトした企業も，デフォルト確率が 0.7 でデフォルトした企業も，同じ情報量を持っているとみなす．つまり，実際の結果はともかく，モデルが与える情報量の増分しか考えていない指標であるといえる．以上のことから，CIER は評価指標として用いるべきではない．

ジャックナイフ法や**ブートストラップ法**の最大の利点は，評価したいパラメータや統計指標の分布を作成できる点であった．例えば，各企業のデフォルト確率の分布の分散を調べることで，デフォルト確率のぶれが計算できる．デフォルト確率のぶれが小さいモデルは，安定したモデルであると判断できる．したがって，どのようなデータに対しても，各企業が一定したデフォルト確率を持つので，モデルの持つ的中率も一定した値をとると推測できる．また，データをリサンプリングして評価するので，**モデルの持つ的中率や安定性を一つのデータからではなく，多くのデータから検証していると見なせる**．以上のように，ジャックナイフ法やブートストラップ法は，評価方法の中でも一番精度の高い評価方法である．高性能の計算機で計算できる場合は，これらの方法を推奨する．また，**クロスバリデーション法**も検証用データを作成し，モデルを評価する意味では，ジャックナイフ法やブートストラップ法と同じである．しかし，推定用データと検証用データはそれぞれ 1 セットしかないので，モデル検証というより，二つに分割したデータセットの類似性を評価しているとみなせる．したがって，クロスバリデーション法は有効ではない．

N/S 比はデフォルトするかしないかの 2 値を予測し，その結果がどれだけ合致しているかを評価する方法であった．二項ロジットモデルでは，企業別にデフォルト確率 p_i が与えられた．そこで，ある判別点 C を一つ決め，$p_i < C$ ならばデフォルトしない，$p_i \geq C$ ならばデフォルトすると予測して，N/S 比を適用する．したがって，N/S 比は判別点 C に依存して変化する値となる．また，判別点 C の決定方法に基準はない．例えば，推定に用いたデータにデフォルト企業，非デフォルト企業がそれぞれ 50%である場合は，判別点 C として 0.5 をとればよい．また，会計基準上の要請により判別点 C が決定されることも考えられる．しかし，一般的には，判別点 C の決定は恣意性が強く影響するので，N/S 比は評価指標として有効ではない．

F 検定はモデルの説明力の有無を統計的検定により評価する場合に用いることができる．ダイバージェンスと同様, 信用リスクスコアの分布を対象に, 分布の平均値と分散を用いた指標である．したがって二項ロジットモデルに適用可能である．F 検定を適用した結果得られる結論は有意か有意でないかのどちらかである．

以上をまとめると，図 3.2 のようになる．

3.1.1 節および 3.1.2 節の考察から，インサンプルデータでのモデル作成では，どのような目的でモデルを作成するか明確にしたうえで，それに対応した評価方法を用いなければならない．つまり，**モデル作成の目的と評価方法の一致性**が重要である．例えば，モデルの的中率を重視する場合は，AR や AUC・ジニ係数などを評価方法として用いればよいし，安定性を重視したモデルを作成したい場合は，ジャックナイフ法やブートスラップ法を用いるというように，その目的にあわせて，評価方法を対応させることである．また, 使用する評価方法に対して適切な推定方法を対応させなければならない．これは**推定方法と評価方法の一致性**といえる．この 2 点に注意して，適切な評価方法を選択しなければならない．

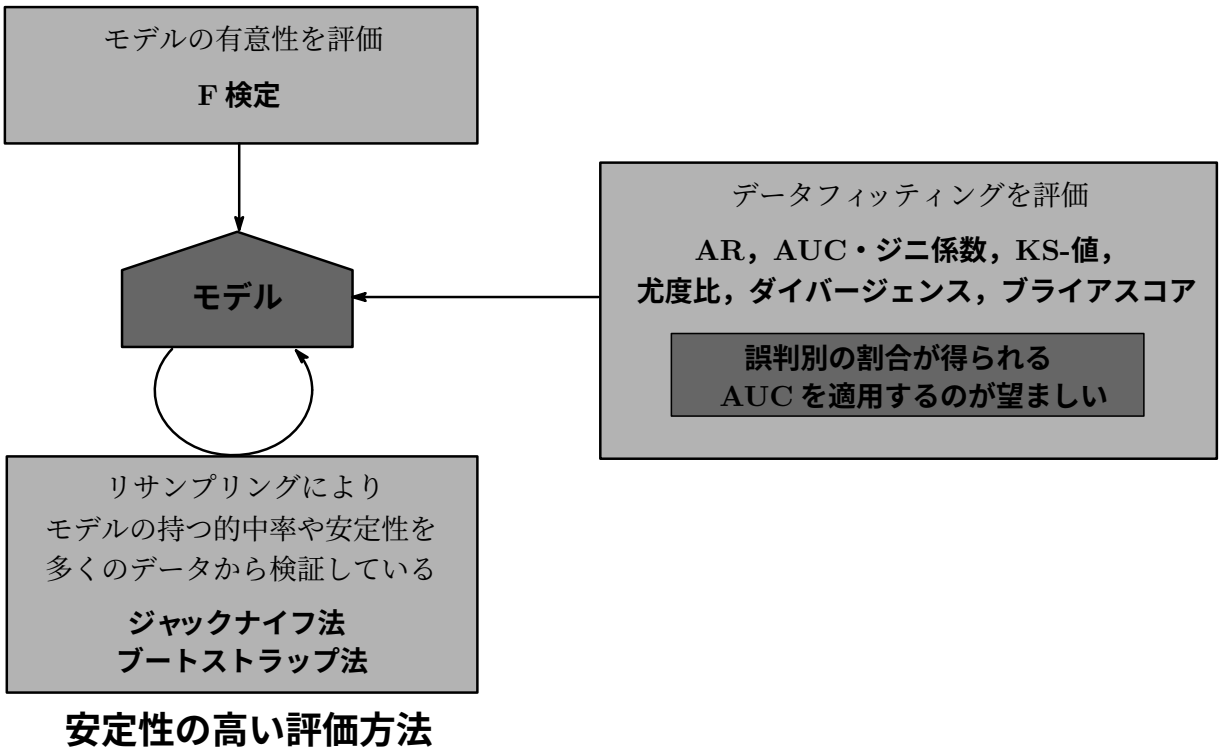


図 3.2: **変数が決定されている場合のモデル評価方法**

3.2 格付けモデルを運用結果データ（アウトサンプルデータ）によって評価する方法

本節では，オーダーロジットモデルで格付けの各ランクごとにデフォルト確率を予測し，その結果について評価したい場合を想定する．格付けモデルの予測結果の評価では，

(1) **格付けモデル全体の評価**
(2) **予測デフォルト確率を格付けの各ランクごとに評価したい場合**
(3) **各ランクに与えたデフォルト確率に有意差の評価**
(4) **格付けの順序性の評価**

の 4 つの観点から，それぞれ対応する評価方法を説明する．次節に進む前に，アウトサンプルデータを用いて，格付けモデルの予測結果を評価する場合，適用できない評価方法について考える．

尤度比，**情報量基準**は，モデル作成時（本節では，オーダーロジットモデルを仮定）において，インサンプルデータ（推定データ）とモデルのフィッティングを評価する指標であった．また，**t-値**も，モデルに用いられている財務指標が有意であるかを評価する指標であり，モデル作成時において適用される．ここでは，アウトサンプルデータによる予測結果の評価を考えているので，適用できない．

クロスバリデーション法，**ジャックナイフ法**，**ブートストラップ法**は，推定に用いるデータからモデル検証用データを作成し，そのデータを用いてアウトサンプルデータによるモデルを評価を擬似的に行う方法である．したがって，これらの方法はインサンプルデータしかない場合に適用する方法であって，本節で想定しているアウトサンプルデータによる評価では適用できない．

CIER は，デフォルトしたか，しなかったかに関係なく，モデルがデフォルト確率を 0 または 1 にどれだけ近づけたかを評価する指標であった．第 2 章で説明したように，予測結果がわからない場合，CIER は当てにならない評価指標となる．アウトサンプルデータを用いてモデルの予測結果を評価する場合，予測結果が全て正しいという仮定は成り立たたない．むしろ，その結果をもとにモデルの的中率や予測と結果の合致性を評価しなければならない．したがって，CIER は本節で想定している格付けモデルだけでなく，どのようなモデルであっても，アウトサンプルデータによる評価に適用できない．

N/S 比は，デフォルトしたか，しなかったかの判別において，正しく判別した率と誤判別率との比で与えられ，デフォルト判別に対する評価方法である．第 2 章で説明したように，N/S 比を用いる場合には，用いるデータの条件が満たされたないと，判別点 C の決定に恣意性が強く影響してしまい，評価指標として有効でなくなる．格付けデータの場合は，各ランクにデフォルト確率が与えられ，判別点 C を決定すると，C より高いデフォルト確率を与えられたランクに属する企業はデフォルトすると予測されてしまうために，単純に判別点 C を決定できない．そのため，判別点 C の決定には恣意性が強く影響するので，N/S 比を用いることは困難である．

ダイバージェンスは，信用リスクスコアの分布を対象に，分布の平均値と分散を用いた指標であった．これも，N/S 比同様，各ランクに一定のデフォルト率を与える格付けモデルでは分布が想定できないので，適用不可能であると判断できる．同じ理由で **KS-値**, **F 検定**も適用不可能である．

以上から，**尤度比，情報量基準，t-値，クロスバリデーション法，ジャックナイフ法，ブートストラップ法，CIER，N/S 比，ダイバージェンス**, **KS-値**, **F 検定の 9 つの評価方法は，格付けモデルのアウトサンプルデータを用いた予測結果の評価方法として適用は薦められない．**

3.2.1 格付けモデル全体を評価したい場合

格付けモデル全体の予測デフォルト確率とその結果の適合度については，**AR または AUC・ジニ係数**，**ブライアスコア**の 2 指標が適用可能であると考えられる．

AR や **AUC・ジニ係数**を用いる場合は，格付けモデル全体の予測的中率を評価する．この場合，各ランクでデフォルト確率が与えられている離散値なので滑らかな曲線は得られない（図 3.3 参照）．これは，個々の企業にデフォルト確率を与えるモデルと比較して，ランクごとにデフォルト確率を与えることで情報量が低下してしまい，精度が落ちてしまうことに注意する．

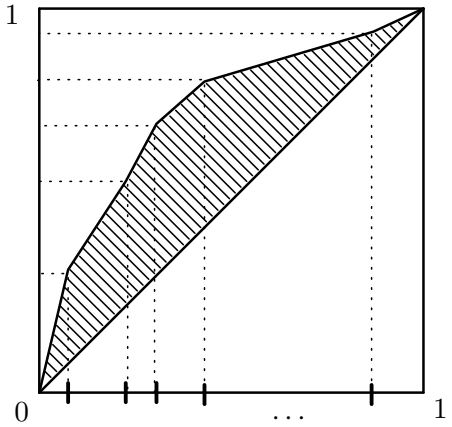


図 3.3: **格付けモデルに ROC 曲線を適用した場合（ランクごとにデフォルト確率が与えらる離散値なので，滑らかな曲線が描けない）**

ブライアスコアは，デフォルト確率の誤差を 2 乗平均して得られる，いわば，分散を評価した指標であった．格付けモデルの場合，ブライアスコアは，

$$\mathbf{BS} \quad = \quad \frac{1}{N}\sum_{k=1}^{K} n_k(P_k - \overline{\theta_k})^2 - \frac{1}{N}\sum_{k=1}^{K} n_k(\overline{\theta_k} - \overline{\theta})^2 + \overline{\theta}(1-\overline{\theta})$$

で表される（2.12 節を参照）．右辺第 1 項は，格付けのランク内で予測したデフォルト確率と実際のデフォルト率の適合度，第 2 項は格付けがランクをうまく分割しているか，それを評価する指標であり，第 3 項は格付け全体としての適合度を表している．ブライアスコアは，デフォルト確率が小さく与えられる（高格付けに与えられたデフォルト確率）と，信頼性が低くなる短所はあるが，格付け全体の適合度のみを考慮している AR や AUC に比べ，ランク内の予測確率の誤差や，分割具合も評価している指標と見れるので，格付けモデル全体の評価方法として推奨できる．

3.2.2 予測デフォルト確率を格付けの各ランクごとに評価したい場合

前節では，格付けモデル全体の予測デフォルト確率とその結果の適合度についての評価方法であったが，格付けの評価の場合，各ランクに与えられた予測デフォルト確率は，実際にそのランクでデフォルトした企業数を予測したか（予測誤差の範囲内であるか）の評価が重要となる．

二項検定は，カテゴリー内にデフォルト確率を与えたとき，実際にデフォルトが起きた件数が，与えたデフォルト確率の予測誤差内であるか，その範囲を与える方法であった．カテゴリー内に十

分多くの企業が存在する場合，非常に有効な評価方法となる．一方，高格付けのランクでは，そのランクに属する企業数が少なく，二項検定をしても良好な結果を得ることが難しい．また，例えば，10 のランクが存在し，有意水準 90%でそれぞれのランクを検定した結果，9 つは有意差がないと判断され，一つが有意差があると判断されたとしよう．この結果から，この格付けモデルがよいモデルではないと判断できない．なぜならば，有意水準 90%であるから，モデルが正しいとしても，この事象は十分起こりえるからである．このように，二項検定を繰り返すと全体で確保したい有意水準を低下させる問題点もある．

格付けの時系列データが存在する場合には，その期間のあいだ，ランクに与えたデフォルト確率と予測結果が合致していたかどうかの評価が考えれらる．この場合は，**Normal Test** が適用できる．Normal test では，中心極限定理を用いて，

$$\frac{Z_T - PD}{\sigma/\sqrt{T}} \approx N(0,1)$$

が成立することを想定する（2.13 節を参照）．つまり，Z_T（その期間における平均予測デフォルト確率）を確率変数とみなし，それが平均 PD（その期間における平均デフォルト件数），分散 σ^2 の正規分布 $\mathcal{N}(PD, \sigma^2)$ に従っていると想定して，正規検定をする方法である．現状では，5 年から 10 年の時系列データがあればこの方法を適用できると報告されている．しかし，これではサンプル数が少ない．実際は，Z_T は正規分布を仮定して，標本分散が未知のときの標本平均の検定をしているだけなので，正規検定を用いるのではなく，**t-検定**を適用するほうがよい．今後，時系列データが蓄積されたときに適用可能であると考えれるが，現状で適用するのは疑問が残る．

また，Normal Test では，**平均**デフォルト確率を考えているが，デフォルト確率のぶれ，つまり，**分散**を評価することも考えられる．信用リスクにおいては，結果が 2 値であるので，二項分布が想定できる．二項分布は，平均がわかれば分散も導出できるので，分散のモデル誤差を考えないことが一般的となっている．しかし，分散も誤差を持つものであるから，評価対象となる．市場リスクでいう **VaR** の考え方からいえば，分散がどれだけかをしっかり評価すべきである．この点は，今後の課題であると考えている．

3.2.3 各ランクに与えたデフォルト確率に有意差を評価したい場合

デフォルト確率はマクロ要因の変化に強く影響される．そのため，マクロ要因を取り込まないモデルは，マクロ変化の影響によって説明力の低いモデルになり，それらのモデルから計測されるデフォルト確率の精度が低下する事態が発生する．しかし，デフォルト確率が正確に予測できない場合でも，モデルから信用力の順序性や格付け間の有意差を判断できるならば，そのモデルを用いてある程度のリスクコントロールを行うことができる．以上のような場合を想定して，本節では格付け間の有意差を，次節において順序性の評価について説明する．

各ランクのデフォルト確率に有意差があるかを評価する場合は，**多重比較法**の**テューキーの方法**または**スティール・デュワスの方法**が適用できる．

これらの方法は，ランク間に与えた予測デフォルト確率に有意差があるかを，ランクすべての対比較を同時に検定するため，モデル全体で考えている有意水準を確保するために，個々の検定の有意水準を調節する方法であった．実際，検定をしたいランクの対を R_i，R_j，それぞれ与えられた予測デフォルト確率を d_i，d_j とするとき，帰無仮説 $H_0 : d_i = d_j$ であるかを**すべての組み合わせに対して**検定する．もし，帰無仮説 H_0 が棄却できないならば，考えているランクは分けなくてもよかったと判断でき，ランクわけの整合性について評価することができる．しかし，対になるランク

の有意差のみを検定していてるので，格付けの順序性については，この方法では評価できないことに注意する．

3.2.4　格付けの順序性を評価したい場合

格付けの順序性が保たれているかを評価する場合は，**田口の累積法**や**多重比較法**の**ウィリアムズの方法**および**シャーリー・ウィリアムズの方法**が適用可能である．

田口の累積法は，順序のある対立仮説を立てて，格付けの順序性を 1 回で検定できる方法であった．帰無仮説が棄却できた場合は，ただちに格付けに順序性があると判断できる．しかし，帰無仮説が棄却できない場合は，全ランクのデフォルト率が同じであると判断するので，この場合，どのランクに有意差があるかは確認できない．

一方，**ウィリアムズの方法**および**シャーリー・ウィリアムズの方法**は，各ランクのデフォルト率に順序性が想定できる場合，着目しているランクがどのランクから有意差があるかを検定する方法である．具体的に，高格付けから順に，$R_1, R_2, \cdots, R_K$ とし，観測されたデフォルト率を $d_1, d_2, \cdots, d_K$ とする．このとき，デフォルト率が，

$$d_1 \leq d_2 \leq \cdots \leq d_K$$

という順序性があると判断できるか検定したい．しかし，これらの方法では，あるランク（ここでは，R_1 とする）に着目して，予測結果はどのランクから有意差があるかを逐次検定していく．実際,「R_1 は R_2 から有意差がある」という結果を得られたとしても，ほかのランクがどのランクから有意差があるかは，この検定からではわからず，目的の順序性の検定ができない．したがって，R_2 に着目して検定，R_3 に着目して検定，と逐次繰り返す作業が必要となるが，検定の多重性（検定を繰り返すことによって全体で考えている有意水準を確保できない）が問題となるので注意しなければならない．

以上，3.2.1 節から 3.2.4 節にかけての考察から，アウトサンプルデータを用いて予測結果を評価する方法は図 3.4 のようにまとめられる．アウトサンプルデータで信用リスクモデルを評価する場合，モデルの的中率を測る評価方法より，**予測と結果の合致性**という考え方でモデルを評価するのが望ましい場合もある．なぜならば，モデル作成時においては，モデルの予測的中率の評価に重点が置かれ，モデルが持つ的中率がどれくらいであるかがわかる．そのモデルを用いて予測デフォルト確率を与えているのだから，当然，予測的中率もモデル作成時で得られた的中率に近づくはずである．この考え方に合致する方法は，**二項検定**，**多重比較法**および**田口の累積法**である．統計学的には，3 者とも**仮説検定法**である．**現状の信用リスクモデル評価方法では，的中率を測る指標は多く提案されているが，予測と結果の合致性を評価する方法（仮説検定法）が少ないといえる．**

3.3　マクロ変数を組み込んだモデルを時系列データによって評価する方法

3.2 節では，アウトサンプルデータを元に，将来の一時点におけるデフォルト確率を求めるモデルを想定した．我が国の現状では，十分な時系列データが存在しないため，十分な議論がなされなかった．しかし，将来は時系列データも蓄積され，このような場合の評価方法が重要になると予想される．そこで，本節では，第 2 章で取り上げてきた評価方法が，時系列データが存在する場合においても適用可能であるか考察する．

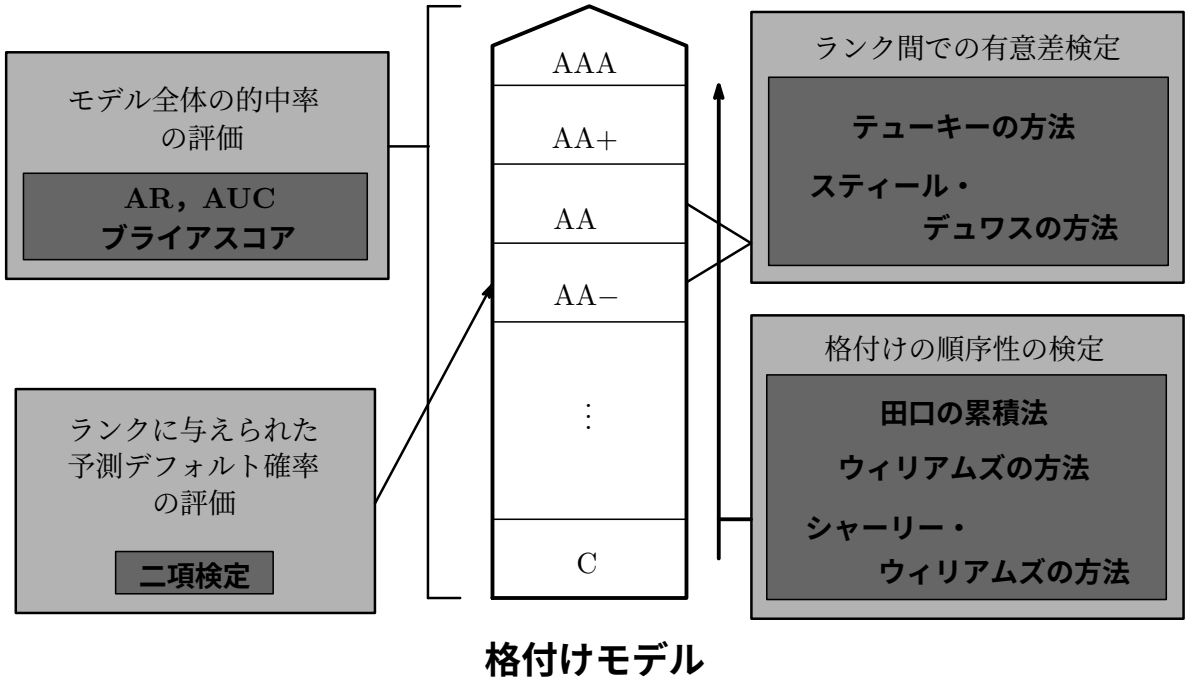


図 3.4: **アウトサンプルデータを用いて運用結果を評価する方法**

まず，二項ロジットモデルにマクロ変数を組み込み，ある一定期間（例えば，5 年〜10 年）の予測結果が得られた場合のモデル評価を想定する．第一の方法としては，時系列を無視し個別のデータとして扱う方法である．例えば，10000 件の企業の予測デフォルト確率が蓄積され，5 年間分存在すると仮定する．この場合，10000（1 年）× 5 年＝ 50000 件の個別データとして扱うことになる．

そこで，予測結果を **CAP 曲線**や **ROC 曲線**で評価することを考える．マクロ変数を取り込むことによって，不景気時の予測デフォルト確率は高く，好景気時には予測デフォルト確率が低くなる．結果としても，不景気時のデフォルト件数は高く，好景気時のデフォルト件数は低くなると予想されるので，**AR** や **AUC・ジニ係数**も予測結果がよければ 1 に近づくはずである．このような結果が得られれば，マクロ変数を取り込むことでモデルの予測結果が良好になったと見なすことができる．しかし，これは一要因であり，実際にマクロ変数を取り込んだから良好になったと単純に判断するのは危険である．それを回避するために，例えば，マクロ変数を用いないモデルとの比較によって評価することが考えられる．

もうひとつ考えられる方法は，1 年間を一つのデータと見なす方法である．上述の例を用いれば，10000 件企業データを一つのデータとしてまとめ，5 つ（5 年間）のデータでモデルを評価することである．例えば，一つのデータにまとめる方法として，平均デフォルト確率を用いる場合は，3.2 節で述べた **Normal Test** を適用することが考えられる．しかし，この方法では『個別企業にデフォルト確率が与えられる』という情報を全く無視してしまうため, この方法は用いない方がよい.

時系列データを用いる利点は，企業別の財務データでは表現できない**マクロ要因**を，信用リスクモデルに反映できることである．企業の財務指標，および，マクロ要因を表す経済指標は，1 年間に数回という単位で蓄積される．しかし，前者は企業数を考慮すると膨大なデータ量になるが，後者は一つの指標であるので 1 年間に数個しか得られない．このような理由から，マクロ要因を信用リスクモデルに反映させるためには，時系列データの蓄積が必要となるのである．

第4章　結論および今後の課題

4.1　結論

本稿では，さまざまな信用リスクモデル評価方法，および，その適用方法について説明した．モデル評価においては，**どのような考え方**で評価するかを明確にしなければならない．現状では，**モデルの的中率**および**予測と結果の合致性**と言う二つ考え方があり，その考え方によって選択する評価方法も変わってくることに，特に注意しなければならない．さらに，検証に用いるデータが**インサンプルデータであるかアウトサンプルデータ**であるか，および，モデルの出力結果が**デフォルト確率か格付け**なのか，明確にして，それに対応した方法を用いなければならない．第3章の考察を重ねた上で，本稿で紹介した評価方法について，表4.1でまとめることができる．

また，第3章の考察から，モデル作成時（インサンプルデータ）を評価する場合は，**モデルの的中率**という観点からの評価，予測結果（アウトサンプルデータ）を評価する場合，**予測と結果の合致性**という観点からの評価が重要であるといえる．

4.2　今後の課題

今後の課題については，モデル評価の考え方の確認，デフォルト相関がある信用リスクモデルの評価方法，動的モデルの評価方法，の3点について考えている．詳しい説明は以下の通りである．

4.2.1　モデル評価と考え方（フィロソフィー）の合致性

現在考えられている評価方法の傾向として，モデルの予測的中率を評価している方法に比べ，モデルと予測の合致性を評価する方法は少ない．これは第3章でも述べたが，統計学的には仮説検定法の適用を考える必要がある．モデル評価は，何度も繰り返すが，どのような考え方で評価するかが一番重要になってくる．監督当局は，予測的中率に重点が置かれた評価指標だけでモデルを評価するのか，モデルそのもの正しかったのかどうか評価するのか，モデル評価の考え方を明確にしなければならない．

4.2.2　デフォルト相関がある信用リスクモデルモデルの評価方法

本稿で想定している信用リスクモデルは，『**企業のデフォルトは独立して起こる**』と暗に仮定していた．しかし，実際にヒストリカルデータを用いてデフォルト率を分析した結果，デフォルトは相関をもつ．したがって,『企業のデフォルトには相関がある』ことを想定した，モデルおよびモデル評価方法を考える必要がある．しかし，デフォルトに相関を持つ場合の代表的なモデルは定まっていない．例えば，ペアの相関を考える場合，ある企業 j がデフォルトした場合の企業 i のデフォルト確率 $p(i|j)$ が割り当てられ，個別の企業に一つのデフォルト確率が与えられない．したがって，

表 4.1: **評価方法の分類**（◎：推奨する評価方法，○：適用可能な評価方法，△：適用できるが推奨されない評価方法，×：適用できない評価方法）

	In Sample		Out Sample	
評価方法名	**PD Model**	**Rating Model**	**PD Model**	**Rating Model**
t-値	○	○	×	×
尤度比	○	○	×	×
情報量基準	◎	◎	×	×
クロスバリデーション法	○	○	×	×
ジャックナイフ法	◎	◎	×	×
ブートストラップ法	◎	◎	×	×
CAP および AR	○	○	○	△
N/S 比	△	×	○	×
ROC および AUC・ジニ係数	○	○	◎	△
KS-値	○	×	○	×
ダイバージェンス	○	△	○	△
CIER	△	△	×	×
ブライアスコア	○	○	○	○
F 検定	○	△	○	△
二項検定	×	×	×	◎
Normal Test	×	×	×	△
多重比較法	×	×	×	○
田口の累積法	×	×	×	◎

（注：ダイバージェンス・F 検定は信用リスクスコアの分布を対象としている．したがって，信用リスクスコアを用いてデフォルト確率や格付けを計測するモデルでは，これらを評価指標として用いることができる）

本稿で説明してきた評価方法を用いることは困難である．この問題点を解決するために，データ全体での平均相関係数を用いたモデルが考えられている．しかし，このモデルを評価する場合，一つのデータセットで一つの平均相関係数しか得られないので，時系列データの蓄積が必要となる．以上のように，デフォルトに相関がある場合のモデルは，これからの研究課題といえる．

4.2.3 格付けモデルの動的変動に対する評価方法

3.3 節で述べたように，十分な時系列データが存在すれば，将来の 1 時点でなく時間とともに変動する信用リスクモデルも作成できる．例えば，時間とともに変動する格付けモデルでは，格付けのランクが変化するモデルと，ランクに与えられるデフォルト確率が変化するモデルの二つが考えられる．前者は，ランクに固定されたデフォルト確率が変化しないので，各ランクに属する企業数が増減する（図 4.1 参照）．これは，単純にデフォルト確率を推定し，どのランクに属するかを判定すればよい．したがって，マクロ変数によって予測デフォルト確率の大きさは変動するが，順位については 1 期ごとに評価しても影響はないので，前者では順序性に着目して的中率を評価する CAP 曲線や ROC 曲線などで評価が可能である．一方，後者のモデルは，ランクに属する企業数は変化しないが，与えられるデフォルト確率が変化する（図 4.2 参照）．この場合，デフォルト確率を推定してどのランクに属するかを判定するだけでは，ランクに属する企業数が一致しない．つまり，後者のモデルでは，順序性に着目して的中率を評価する評価指標では評価できない．したがって，各ランクに与えたデフォルト確率が時間とともに変化するので，時系列データを用いた評価をしなければならない．このように時系列データを用いた評価ができない場合，1 期づつに分け，本稿で説明した評価指標を用いてモデルを評価し，その指標の平均的な値で評価することも考えられる．時系列データを組み込んだモデルに対して，現状で用いられている評価方法で評価できるのか，新しい評価方法を作成しなければならないのか，その整理が必要である．

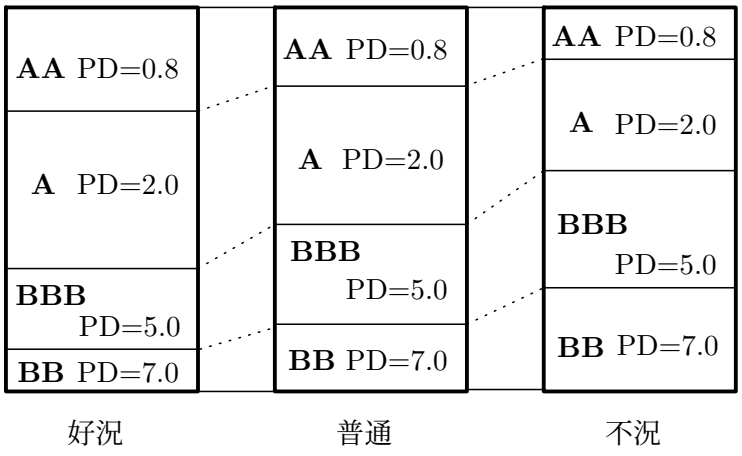


図 4.1: **ランクに与えられるデフォルト確率は変化しないが，ランクに属する企業数が推移する**

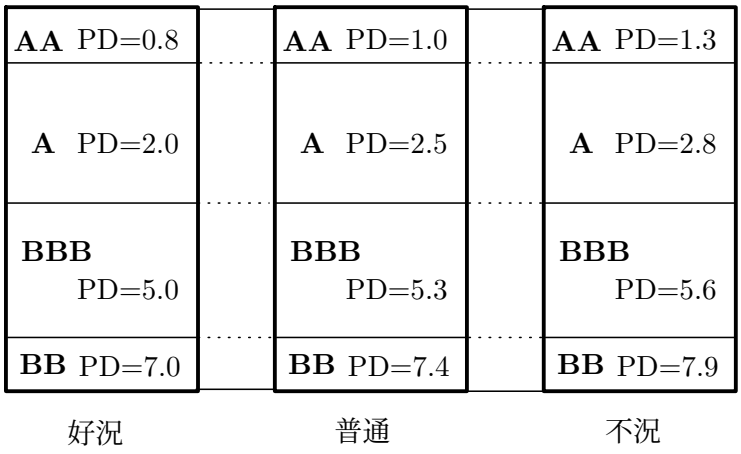


図 4.2: **ランクに属する企業数は変化しないが，ランクに与えられるデフォルト確率が推移する**

Appendix

本論では説明できなかった，数学的な事項について説明する．

エントロピー

我々はなんらかの情報を得ることにより,「何もわからない」状態から「少しだけわかった」状態へと変化する．つまり，不確かさの軽減度合いがその情報が持つ**情報量**となる．

例えば，コインの表か裏かを当てるゲームを考える．歪みのないコインであれば，コインの表のでる確率 p は $1/2$ である．情報量を,

$$I(p) = -\log_2(p) \tag{4.1}$$

と定義する．これより,「コインの表がでる」という情報の情報量は 1 であることがわかる．この情報量を **1bit(ビット)** と呼ぶ.

もうひとつの例として，明日の天気が，晴れ 50 %，曇り 30 %，雨 20 %であると予測されている場合を考える，このときの「晴れ」,「曇り」,「雨」，それぞれの情報量は,

$$\begin{aligned} I(\text{「晴れ」}) &= -\log_2 \frac{5}{10} = 1 \\ I(\text{「曇り」}) &= -\log_2 \frac{3}{10} \simeq 1.74 \\ I(\text{「雨」}) &= -\log_2 \frac{2}{10} \simeq 2.32 \end{aligned}$$

である．これより，情報の持つ確率が低いほど，情報量が大きくなっていることがわかる．ここで,これらの情報量の平均 H を考えると,

$$\begin{aligned} H &= \frac{5}{10} \cdot I(\text{「晴れ」}) + \frac{3}{10} \cdot I(\text{「曇り」}) + \frac{2}{10} \cdot I(\text{「雨」}) \\ &= 1.57 \end{aligned}$$

となり，この平均情報量を**エントロピー**と呼ぶ.

一般に，確率変数を X としたとき，X のエントロピー $H(X)$ を,

$$\begin{aligned} H(X) &= -\sum_{x \in X} p(x) \log p(x) \\ &= E\left(log \frac{1}{p(x)} \right) \end{aligned} \tag{4.2}$$

で定義する．(ここで，$E(\cdot)$ は確率変数の期待値を表す) つまり，**エントロピー**とは，確率変数 X によって得られる情報量の期待値であり，どの事象が生起するかの不確かさを表す量となる．

相対エントロピー

ふたつの確率変数が従う確率分布をそれぞれ $p(x)$，$q(x)$ とする．**相対エントロピー（Kullback-Leibler 情報量）**は，

$$\begin{aligned} D(p|q) &= \sum_{x\in X} p(x)\log\frac{p(x)}{q(x)} \\ &= \sum_{x\in X} p(x)\log p(x) - \sum_{x\in X} p(x)\log q(x) \end{aligned} \tag{4.3}$$

で定義される．いま，モデルが従う真の分布を $p(x)$，データから推測されたモデルの分布を $q(x)$ と仮定する．(4.3) 式から，データから推測された分布 $q(x)$ が真の分布 $p(x)$ に完全に一致したならば，相対エントロピーは 0 となる．したがって，モデルの分布が真の分布に近づいているかは，相対エントロピーを計測することで確認ができる．

しかし，一般に真の分布は未知である．また，真の分布がわかっていたら，(4.3) 式の第 1 項は定数になるので，重要となるのは第 2 項の $\sum p(x)\log q(x)$ である．そのため，何らかの方法によって $\sum p(x)\log q(x)$ が推定できたならば，その値を比較することで，どのモデルが真のモデルに一番近いか判断可能である．このような考え方に基づき，相対エントロピーは，モデル評価に用いられている．

AR と AUC との関係

デフォルトした企業数を N_D，デフォルトしなかった企業を N_{ND} とする．モデルが完全に予測するモデルの CAP 曲線とランダムなモデルの CAP 曲線で囲まれる面積を a_p とすると，

$$a_p = \frac{1}{2}\cdot\frac{N_{ND}}{N_D+N_{ND}} \tag{4.4}$$

とかける．

全体からランダムにある企業を選んだ場合，その選ばれた企業の信用スコアを S_T とする．同様に，デフォルトした企業の中からランダム選んだ場合は S_D，デフォルトしなかった企業の中からランダムに選んだ場合は S_{ND} とする．また，ある定数 C を用いて，S_D が C より小さかった場合の確率を $HR(C) = P(S_D < C)$，S_{ND} が C より小さかった場合の確率を $FAR(C) = P(S_{ND} < C)$ で表す．このとき，全体からランダムにある企業を選び，その企業の信用スコアが C より小さくなる確率 $P(S_T < C)$ は，

$$P(S_T < C) = \frac{N_D\,P(S_D < C) + N_{ND}\,P(S_{ND} < C)}{N_D+N_{ND}} \tag{4.5}$$

で表される．ただし，S_D，S_{ND} の分布はともに連続であると仮定する．

実際に用いたモデルの CAP 曲線とランダムなモデルの CAP 曲線とで囲まれる面積を a_R とする．a_R は，

$$\begin{aligned} a_R &= \int_0^1 P(S_D < C)dP(S_T < C) - \frac{1}{2} \\ &= \frac{N_D\int_0^1 P(S_D < C)dP(S_D < C) + N_{ND}\int_0^1 P(S_{ND} < C)dP(S_D < C)}{N_D+N_{ND}} - \frac{1}{2} \end{aligned}$$

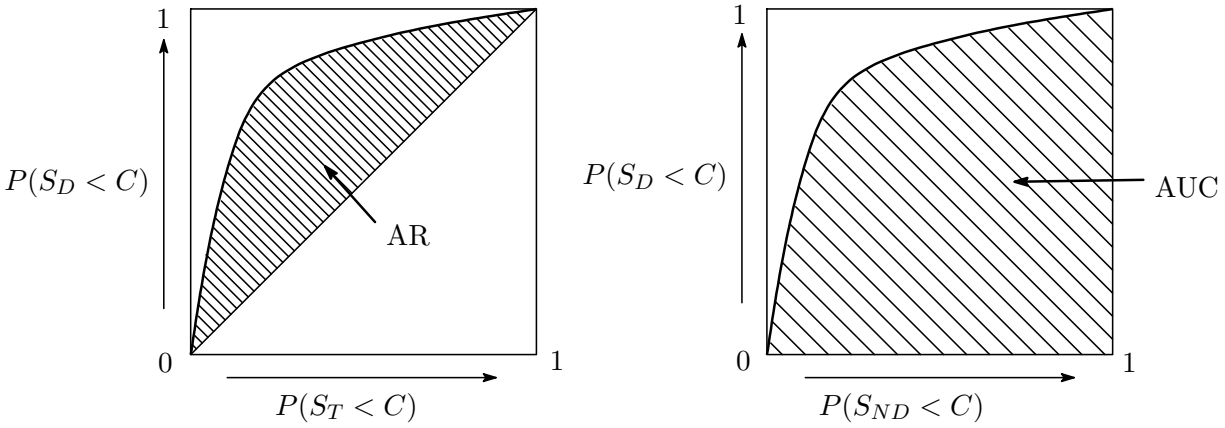
$$= \frac{\frac{1}{2} \cdot N_D + N_{ND} \cdot AUC}{N_D + N_{ND}} - \frac{1}{2}$$
$$= \frac{N_{ND}(2 \cdot AUC - 1)}{2 \cdot (N_D + N_{ND})} \tag{4.6}$$

と表される．ここで，AUC は実際に用いたモデルの ROC 曲線下の面積を表す.（図 4.1 参照）

したがって，AR は，

$$AR = \frac{a_R}{a_p} = \frac{N_{ND} \cdot (2 \cdot AUC - 1)}{N_{ND}} = 2 \cdot (AUC - \frac{1}{2}) \tag{4.7}$$

となる．



全件データ中の
信用スコアが小さい順に並ぶ

$$AR = \int_0^1 P(S_D < C) dP(S_T < C) - 0.5$$

デフォルトした企業中の
信用スコアが小さい順に並ぶ

$$AUC = \int_0^1 P(S_D < C) dP(S_{ND} < C)$$

図 4.3: **AR と AUC の関係**

多重比較法の基礎

一般に，2 群間における母平均の差の検定では **2 標本 t 検定**を用いる．3 群以上であっても，2 標本 t 検定を繰り返すことが考えられるが，検定を多数回繰り返すと各群間に設定している有意水準を満たしていても，全体としての有意水準が大きくなる．例えば，A，B，C という 3 つの群があり，それぞれの平均値が μ_A，μ_B，μ_C であったとする．この場合，考えられる帰無仮説は 3 つで，

$$H_0 : \mu_A = \mu_B \qquad H_0 : \mu_B = \mu_C \qquad H_0 : \mu_C = \mu_A \tag{4.8}$$

となる．ここで，2 標本 t 検定を繰り返し，それぞれ有意水準 5%で帰無仮説が採択されたと仮定する．その結果,「有意水準 5%で，$\mu_A = \mu_B = \mu_C$ がで成立している」と結論付けたとする．この場合，(4.8) 式の帰無仮説は，それぞれが 95%の確率で正しいとしているから，$\mu_A = \mu_B = \mu_C$ となる確率は $(0.95)^3 = 0.8574$ となる．これは，$1 - (0.95)^3 = 0.1426$，つまり，約 14%の確率で結論は誤りであるといえる．これでは，有意水準 5%で成立しているとはいえない．

このように，検定の多重性は有意水準を増加させてしまう．**多重比較法**は，**全体として必要とする有意水準をコントロールするために，それぞれの帰無仮説に対する有意水準を調節して検定する方法**である[8]．具体的な方法の説明する前に，必要となる用語についてまとめておく．

K 個の群が存在すると仮定する．第 k 群のデータ数を n_k，データを $(x_{k1}, \cdots, x_{kl}, \cdots, x_{kn_k})$ とする．また，第 k 群に属するデータの総和を T_k，群内平均を $\overline{x_k}$，群内分散を V_k とする．T_k，$\overline{x_k}$，V_k は，

$$T_k = \sum_{l=1}^{n_i} x_{kl} \qquad \overline{x_k} = \frac{T_k}{n_k}, \qquad V_k = \frac{\sum_{l=1}^{n_k} (x_{kl} - \overline{x_k})^2}{(n_k - 1)}$$

で表される．各群について具体的に表記すると，表 4.2 のようになる．

表 4.2: 1 次元配置デザインのデータ形式

群	サイズ	データ			計 T_i	平均 $\overline{x_i}$	分散 V_i
第 1 群	n_1	x_{11}	$\cdots$	x_{1n_1}	T_1	$\overline{x_1}$	V_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
第 k 群	n_k	x_{k1}	$\cdots$	x_{kn_k}	T_k	$\overline{x_k}$	V_k
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
第 K 群	n_K	x_{K1}	$\cdots$	x_{Kn_K}	T_K	$\overline{x_K}$	V_K
（全データ）$= N = n_1 + \cdots n_K$					（データ総和）$= T = \sum_{k=1}^{K} T_k$		

テューキーの方法

テューキーの方法は，**母平均について群間ですべての対比較を同時に検定するための多重比較法**である．群 G_k のデータ数を n_k，G_k に属するデータを $(x_{k1}, \cdots, x_{kn_k})$ とし，母平均を μ_k とする．ここで，各群に属するデータの平均，分散は，

$$\overline{x_k} = \frac{\sum_{l=1}^{n_k} x_{kl}}{n_k}, \qquad V_k = \frac{\sum_{l=1}^{n_k} (x_{kl} - \overline{x_k})^2}{(n_k - 1)}$$

となる．ただし，各群に属するデータは**正規分布に従い**，かつ，各分布の**分散は等しい**とする．検定をしたい群の対を第 i 群 G_i と第 j 群 G_j とし，帰無仮説と対立仮説を，

$$H_0 : \mu_i = \mu_j \qquad H_1 : \mu_i \neq \mu_j$$

とたてる．

[8] 2 群間ではなく，**全て**の群の母平均が等しいかどうかを検定する場合は，**分散分析**を適用する．

ここで，**誤差自由度** ϕ_e と**誤差分散** V_e を，

$$\phi_e = \sum_{k=1}^{K} n_k - K \tag{4.9}$$

$$V_e = \frac{\sum_{k=1}^{K} (n_k - 1)V_k}{\phi_e} \tag{4.10}$$

と定義する．有意水準を α として，検定統計量 t_{ij} を，

$$t_{ij} = \frac{\overline{x_i} - \overline{x_j}}{\sqrt{V_e\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \tag{4.11}$$

を計算し，$|t_{pq}| \geq q(K, \phi_e; \alpha)$ であるならば，帰無仮説 H_{ij} を棄却し，R_i と R_j の母平均には有意に差があると判断する．(ただし，$q(K, \phi_e; \alpha)$ は自由度 ϕ_e の**ステューデント化された分布**の上側 $100\alpha\%$ 点の値を表す)

スティール・デュワスの方法

スティール・デュワスの方法は，テューキーの方法と同様，**母平均について，群間ですべての対比較を同時に検定するための多重比較法**である．ただし，各群の従う分布は何も仮定しない，**ノンパラメトリック**な方法である．

第 1 群から第 K 群まで合わせたデータを小さいものから順位をつけ，**順位データ**に変換する．第 i 群の l 番目の順位データを r_{il} とおき，第 i 群の順位和 R_{ij} を，

$$R_{ij} = r_{i1} + \cdots + r_{in_i} \tag{4.12}$$

と定義する．(対となる群を第 j 群 G_j としているので，R_{ij} と記述している)

テューキーの方法と同様，第 i 群 G_i と第 j 群 G_j に有意差があるか，帰無仮説と対立仮説を，

$$H_0 : \mu_i = \mu_j \qquad H_1 : \mu_i \neq \mu_j$$

とたて，検定する．ただし，μ_i，μ_j は母集団分布が従う分布の位置を表すパラメータ（母平均やメディアンなど）とする．

期待値 $E(R_{ij})$ と分散 $V(R_{ij})$ を以下の式で計算する．

$$N_{ij} = n_i + n_j \tag{4.13}$$

$$E(R_{ij}) = \frac{n_i(N_{ij} + 1)}{2} \tag{4.14}$$

$$V(R_{ij}) = \frac{n_i n_j}{N_{ij}(N_{ij} - 1)} \left(\sum_{k=1}^{n_i} r_{ik}^2 + \sum_{k=1}^{n_j} r_{jk}^2 - \frac{N_{ij}(N_{ij} + 1)^2}{4} \right) \tag{4.15}$$

これらを用いて検定統計量 t_{ij}，

$$t_{ij} = \frac{R_{ij} - E(R_{ij})}{\sqrt{V(R_{ij})}} \tag{4.16}$$

を計算する．$|t_{ij}| \geq q(K, \infty; \alpha)$ であるならば帰無仮説を棄却し，有意差があると判断する．(ただし，$q(K, \infty; \alpha)$ は自由度 ∞ の**ステューデント化された分布**の上側 $100\alpha\%$ 点の値を表す)

ウィリアムズの方法

ウィリアムズの方法は，**ひとつの対照群とふたつ以上の処理群があって，各群の母平均に順序性が想定できる場合，母平均について対照群と処理群の対比較のみを検定するための多重比較法である．**

第 1 群を対照群，第 2 群から第 K 群を処理群とし，各群の母平均には，

$$\mu_1 \leq \mu_2 \leq \cdots \leq \mu_K \tag{4.17}$$

の関係が成り立つと想定できるものとする．ただし，各群に属するデータは**正規分布に従い**，かつ，各分布の**分散は等しい**であるとする．

$p = K$ とおき，帰無仮説，対立仮説を

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_p$$
$$H_1 : \mu_1 \leq \mu_2 \leq \cdots \leq \mu_p \quad \text{（ただし，少なくともひとつの不等号が成立）}$$

とたて，検定する．

テューキーの方法と同様，誤差自由度 ϕ_e および誤差分散 V_e を，

$$\phi_e = \sum_{k=1}^{K} n_k - K \tag{4.18}$$

$$V_e = \frac{\sum_{k=1}^{K} (n_k - 1)V_k}{\phi_e} \tag{4.19}$$

と定義する．ここで次の統計量，

$$\begin{aligned} y_{2p} &= \frac{T_2 + T_3 + \cdots + T_p}{n_2 + n_3 + \cdots + n_p} \\ y_{3p} &= \frac{T_3 + \cdots + T_p}{n_3 + \cdots + n_p} \\ &\cdots \\ y_{pp} &= \frac{T_p}{n_p} \end{aligned}$$

を求め，その最大値を M_p と定める．統計量 t_p を

$$t_p = \frac{M_p - \overline{x_1}}{\sqrt{V_e\left(\dfrac{1}{n_p} + \dfrac{1}{n_1}\right)}} \tag{4.20}$$

を計算して，$t_p < w(p, \phi_e; \alpha)$ ならば，帰無仮説を保留して検定を終了する．$t_p \geq w(p, \phi_e; \alpha)$ ならば，帰無仮説を棄却し，「μ_p は μ_1 より大きい」と判断する．$p = 2$ ならば，検定を終了するが，$p \geq 3$ であれば，p の値を 1 だけ減らして，再び検定を行う．（ただし，$w(p, \phi_e; \alpha)$ は，自由度 ϕ_e の**ウィリアムズの方法のための分布**上側上位 $\alpha\%$ 点の値とする．）

シャーリー・ウィリアムズの方法

シャーリー・ウィリアムズの方法は，ウィリアムズの方法と同様，**ひとつの対照群とふたつ以上の処理群があって，各群の母平均に順序性が想定できる場合，母平均について対照群と処理群の対比較のみを検定するための多重比較法である．**ただし，各群が従う分布を何も仮定しない**ノンパラメトリック**な方法である．

第 1 群を対照群，第 2 群から第 K 群を処理群とし，各群の母集団の位置を表すパラメータ（母平均やメディアンなど）には，

$$\mu_1 \leq \mu_2 \leq \cdots \leq \mu_K \tag{4.21}$$

の関係が成り立つと想定できるものとする．ただし，**各群の従う分布がわからなくてもよい**．

$p = K$ とおき，帰無仮説，対立仮説を

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_p$$
$$H_1 : \mu_1 \leq \mu_2 \leq \cdots \leq \mu_p \quad \text{（ただし，少なくともひとつの不等号が成立）}$$

とたて，検定する．

第 1 群から第 K 群まで合わせたデータを小さいものから順位をつけ，**順位データ**に変換する．第 i 群の k 番目の順位データを r_{ik} とおき，各群の順位和 R_{ip} を，

$$R_{ip} = r_{i1} + \cdots + r_{in_i} \tag{4.22}$$

で計算する．次に，以下の統計量，

$$\begin{aligned} y_{2p} &= \frac{R_{2p} + R_{3p} + \cdots + R_{pp}}{n_{2p} + n_{3p} + \cdots + n_{pp}} \\ y_{3p} &= \frac{R_{3p} + \cdots + R_{pp}}{n_{3p} + \cdots + n_{pp}} \\ &\cdots \\ y_{pp} &= \frac{R_{pp}}{n_{pp}} \end{aligned}$$

を計算し，その最大値を $M_p (= \max\{y_{2p}, \cdots, y_{pp}\})$ とする．ここで，

$$U_{1p} = \frac{R_{1p}}{n_1} \tag{4.23}$$

$$N_p = \sum_{k=1}^{p} n_k \tag{4.24}$$

$$V_p = \frac{1}{N_p - 1}\left(\sum_{k=1}^{p} \sum_{l=1}^{n_i} r_{il}^2 - \frac{N_p(N_p+1)^2}{4} \right) \tag{4.25}$$

を計算し，統計量 t_p

$$t_p = \frac{M_p - U_{1p}}{\sqrt{V_e\left(\frac{1}{n_p} + \frac{1}{n_1}\right)}} \tag{4.26}$$

を計算して，$t_p < w(p, \infty; \alpha)$ ならば，帰無仮説を保留して検定を終了する．$t_p \geq w(p, \infty; \alpha)$ ならば，帰無仮説を棄却し，「μ_p は μ_1 より大きい」と判断する．$p = 2$ ならば，検定を終了するが，$p \geq 3$ であれば，p の値を 1 だけ減らして，再び検定を行う．（ただし，$w(p, \infty; \alpha)$ は，自由度 ∞ の**ウィリアムズの方法のための分布**上側上位 $\alpha\%$ 点の値とする．）

田口の累積法

田口の累積法は，**対立仮説が順序性を持つようなモデルにおいて，通常の統計量を計算するのではなく，改良された統計量を用いて検出力を大きくする**方法である．

対象としている群の個数は K 個であると仮定する．第 k 群 G_k のデータ数を r，G_k に属するデータを $(x_{k1}, \cdots, x_{kj}, \cdots, x_{kr})$ として，各群のデータ数 r は同じとする．各群が従う分布の平均を $\mu_1, \mu_2, \cdots, \mu_K$ とするとき，帰無仮説 H_0 および対立仮説 H_1 を，

$$
\begin{aligned}
H_0 &: \mu_1 = \mu_2 = \cdots = \mu_K \\
H_1 &: \mu_1 \leq \mu_2 \leq \cdots \leq \mu_K \quad \text{（すべての不等号が成立）}
\end{aligned}
$$

とたてる．

また，各群に属するデータの平均，全データの平均，および，全データの分散を

$$
\overline{x_k} = \frac{\sum_{l=1}^{r} x_{kl}}{r} \qquad \overline{x} = \frac{\sum_{k=1}^{K}\sum_{l=1}^{r} x_{kl}}{K \cdot r} \qquad S^2 = \frac{\sum_{k=1}^{K}\sum_{l=1}^{r} (x_{kl} - \overline{x})^2}{K \cdot (r-1)}
$$

で定義する．ただし，各群に属するデータは**正規分布に従い**，かつ，各分布の**分散は等しい**とする．

ここで次の統計量 t'_i，Q を，

$$
t'_k = \sqrt{\frac{r \cdot k \cdot (K-k)}{K \cdot S^2}} \left(\frac{1}{k}(\overline{x_1} + \cdots + \overline{x_k}) - \frac{1}{K-1}(\overline{x_{k+1}} + \cdots + \overline{x_K}) \right) \tag{4.27}
$$

$$
Q = \sum_{k=1}^{K-1} {t'_k}^2 \tag{4.28}
$$

を計算する．統計量 $Q/(K-1)$ は，自由度 $(\phi, K(r-1))$ の F 分布に従うことを用いて，検定を行う．ここで，ϕ は，

$$
\frac{1}{\phi} = \frac{1}{K-1}\left(1 + \frac{2}{K-1} \sum_{k=2}^{K} \sum_{l=1}^{k-1} c_{kl}^2 \right) \tag{4.29}
$$

$$
c_{kl} = \sqrt{\frac{k \cdot (K-l)}{l \cdot (K-k)}} \tag{4.30}
$$

を用いて計算する．

参考文献

[Akaike(1973)] Akaike,H., "Information Theory and an Extention of the Maximum Likelihood Principle", *2nd International Symposium on Information Theory*, Akademiai Kiado, Budapest 267-281, 1973

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

[Altman(1977)] E.I. Altman, R. Haldeman, and P. Narayanan, "ZATA Analysis: A New Model to Identify Bunkruptcy Risk of Corporation", *Journal of Banking and Finance*, 29-55, 1977

[Deakin(1972)] Deakin, E.B., "A Discriminant Analysis of Precictors of Business Failure", *Journal of Accounting Research*,10,1, 167-179, 1972

[Efron(1982)] Efron,B., *The Jackknife, the Bootstrap and Other Resampling Plans*, SIAM, Philadelphia, 1982

[Eguchi and Copas(2002)] Eguchi,S. and Copas,J., "A class of logistic-type discriminant functions", *Biometrika*, 89, 1-22, 2002

[Hamerle, Rauhmeier, Rosch(2003)] Hamerle, A., Rauhmeier, R., Rosch, D., "Uses and Misuses of Measures for Credit Rating Accuracy", University of Regensburg, 2003

[Keenan and Sobehart(1999)] Keenan,S.C. and Sobehart,J.R., "Performance Measures for Credit Risk Models", *Moody's Techical Reports*, 1999

[Kullback and Leibler(1951)] Kullback,S. and Leibler,R.A., "On Information and Sufficiency", *Annals od Mathmatical Statistics* 22, 79-86, 1951

[Kullback(1959)] Kullback,S., *Information Theory and Statistics*, Wiley & Sons, New York, 1959

[Lee, Urrutia(1996)] Lee,S.H. and Urrutia, J.L., "Analysis and Prediction of Insolvency in the Property-Liability Insurance Industry: A Comparison of Logit and Hazard Models", *The Journal of Risk and Insurance* 63, 1, 121-130, 1996

[Moody's(2001)] Moody's Investors Service., *RISK CALC ^{TM}For Private Companies: Moody's Default Model*, 2000

[Moody's(2001)] Moody's Investors Service., *RISK CALC ^{TM}For Private Companies: Japan*, 2001

[Newson(2001)] Newson, R., "Parameters behind "non-parametric" statistics: Kendall's τ_α, Somers' D and median differences" *The Stata Journal*, 1, 1, 1-20. 2001

[Shao and Tu(1995)] Shao,J. and Tu,D., *The Jackknife and Bootstrap*, Springer-Verlag, New York, 1995

[Shumway(1999)] Shumway, Taylor., "Forcasting Bankruptcy More Accurately. *A Simple Hazard Model*", University of Michigan, Working Paper, 1999

[Stein(2002)] Stein,R.M., "Benchmarking Default Prediction Models: Pitfalls and Remedies in Model Validation", *Moody's Technical Report*, 2002

[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

[坂本，石黒，北川 (1983)] 坂本慶行 石黒真木夫 北川源四郎 著,「情報量統計学」，共立出版，1983

[東京大学教養学部統計学教室 (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

[竹内 (1980)] 竹内啓 著,「現象と行動のなかの統計数理」, 新曜社, 1980

[永田，吉田 (1997)] 永田靖，吉田道弘 著,「統計的多重比較法の基礎」，サイエンティスト社，1997

[前園 (2001)] 前園宜彦 著,「統計的推測の漸近理論」, 九州大学出版会, 2001

[武藤 (1995)] 武藤真介 著,「統計解析ハンドブック」, 朝倉書店, 1995

[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

[森平, 隅田 (2001)] 森平爽一郎, 隅田和人, "格付け推移行列のファクター・モデル",「金融研究」第 20 巻別冊第 2 号, 日本銀行金融研究所,12, 2001

[山下, 川口 (2003)] 山下智志, 川口昇, "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)", 金融庁金融研究研修センター, ディスカッションペーパー, 4, 2003