情報理論による実証分析と測定論の橋渡し

福井義高

青山学院大学大学院国際マネジメント研究科

〒150-8366　東京都渋谷区渋谷４－４－２５

fukui@gsim.aoyama.ac.jp

2012 年 12 月 28 日

2013 年 1 月 17 日改訂

要旨

実証分析と測定論の二項対立を止揚し、情報の科学たる会計研究にふさわしく、情報理論の初歩を用いて、実証と測定を統一的に理解する道を模索する。

謝辞

会計数値の情報価値と情報理論の関連について貴重な指摘をいただいた斎藤静樹教授及び本稿で引用した論文に関する質問に対して丁寧に答えてくださったProfessor Hwan-sik Choi にこの場を借りて感謝申し上げる。なお、本稿作成にあたっては科学研究費（20653026）の助成を受けた。

Causal models can never be proven to be true.
Certainty must be left to the mystic.
David Kenny

１．はじめに：実証と測定の統一的理解を目指して

日本における最近の会計研究は米国の動向に強い影響を受け、「実証にあらずんば人にあらず」といった観を呈している。一方、会計測定を中心にした伝統的規範論は、実務家にとっては有用であっても「科学」以前の研究の名に値しない存在と見る向きも、口に出していわないにせよ（過去の遺物とはいえ、大先生たちがまだご存命！）、少なくないように見受けられる。

しかし、測定論は本当に研究方法として「終わった」存在なのであろうか。本稿では、情報の科学たる会計研究にふさわしく、情報理論（の初歩）を用いて、実証と測定をむしろ統一的に理解する道を模索する。

まず、次の第 2 節で情報を量的に把握する方法を示し、第 3 節では必ずしもわかりやすいとはいえない情報量の意味を「無知の大きさ」と解釈できることを示す。さらに、第 4 節で実証分析において強調される会計数値の情報価値と密接に関連する相互情報量概念を導入する。

続く第 5～8 節では、モデル構築のカギとなる概念であるカルバック - ライブラー情報量を導入し、統計分析で広く用いられている最尤法の意味を考えるとともに、竹内情報量規準（TIC）の簡略版として、赤池情報量規準（AIC）を導出する。

AIC については、古典的統計理論の立場から様々な批判があり、とりわけその一致性の欠如が問題点として指摘される。第 9～10 節はこの AIC 批判とそれに対する反批判である。続く第 11～12 節では、しばしば対立して捉えられることの多い仮説検定と AIC によるモデル選択を融合させた考え方を紹介し、会計実証研究で多用される Vuong 検定もこの文脈のなかで検討する。

第 13～14 節では、ここまでモデル構築と選択を情報理論の立場から検討したように、測定に関しても同様の視点から理解できることを示す。続く第 15 節では、測定の背後にある実体という観点から、心理学におけるテスト理論の会計測定への適用可能を検討する。

最後の第 16 節は簡単な総括である。

２．情報とは

「情報」という言葉はいろいろな意味で使われる。しかし、ここでは標準的な情報理論[1]における情報概念を簡単に解説する。情報を測定するにはその量化が必須であることを考えれば、情報理論とは情報量の理論といってもよい。

[1] 情報理論の概要を理解するには、たとえば、甘利（2011）、豊田（1997）吉田（2009）等を参照。

情報量とは、一言で表現すれば、「伝達する情報のニュース性の大きさ」であり、「聞いてあるいは読んでビックリする驚きの程度」である[2]。

したがって、確実に起こることが分かっている事象が生じた場合、全く驚かないので情報量はゼロ。逆に、あり得ない想定外の事象が生じた場合は、驚きは非常に大きいので情報量は大である。

情報量が確率と密接に関係する概念であることはもはや自明であろう。事象 i の確率を p_i、情報量を $I(p_i)$とすれば、確率と逆相関の関係にあり、事象が必ず起こるすなわち $p_i = 1$ の場合、$I(1) = 0$ となる関数で情報量を定義すればよい。

そのような関数は無限にありそうな気がするかもしれないけれども、実はひとつしかない。独立な事象である A の情報量と B の情報量を加えると、A かつ B の情報量と等しくなければならない。つまり、A と B が同時に起こる確率は $p_A \times p_B$ なので、

$$I(p_A \times p_B) = I(p_A) + I(p_B)$$

が成り立っていなければならない。

こうした掛け算を足し算とする関係といえば対数（関数）である。したがって、$I(p_i) = -\log p_i$ とすれば、上記の情報量の定義は満たされる。しかも、それ以外には（その定数倍を除き）存在しない。対数の底は任意であるけれども、実証研究との関連を見るうえで便利なので、ここでは自然対数を用いる。

まず、個別の事象ではなく、確率的にいずれかの事象が生じるある世界そのものを、ひとつの明確な量で表すにはどうすればよいか考える。たとえば、我々が対象とする世界が歪みのないコイン投げだとすると、オモテかウラが二分の一で起こるという事象全体を数値で表すにはどうすればよいのだろうか。

先ほどの定義を用いると、もしオモテが出れば情報量は

$$I\left(\frac{1}{2}\right) = -\log\left(\frac{1}{2}\right) = \log 2$$

となり、ウラの場合も同じである。

事象が生じればそのたびに情報量を測定できるけれども、事前にはその確率分布しかわからない。それではこの確率分布を知っていることを数値化できないであろうか。個別の情報量が「あるニュースを聞いたことによる驚きの量」であることを考えれば、事後にどれくらい驚くかの（事前の）期待値 S を確率分布の（平均）情報量とすればよい。

コイン投げの場合は、オモテかウラが二分の一で起こるので、期待値 S はオモテとウラの情報量の平均、すなわち

$$S = \frac{1}{2}\log 2 + \frac{1}{2}\log 2 = \log 2 \simeq 0.693$$

となる。

事象が等確率で起こる場合以外も考え方は同じである。コインが歪んでおり、オモテが

[2] ここでの引用は豊田（1997, 1 章）に基づく。

出る確率は三分の二、ウラは三分の一の場合、平均情報量は

$$S = -\left(\frac{2}{3}\log\frac{2}{3} + \frac{1}{3}\log\frac{1}{3}\right) = \log 3 - \frac{2}{3}\log 2 \simeq 0.637$$

となる。

コイン投げに限らず、ある確率的現象を（離散）確率分布で表せる場合、平均情報量は

$$S = -\sum_{i=1}^{N} p_i \log p_i$$

となる。連続変数の場合は、確率密度関数を $p(x)$ とすれば、

$$S = -\int_{-\infty}^{\infty} p(x)\log p(x)dx$$

と定義すればよい。

３．情報量の意味

ところで情報量は多い方がよいのか、それとも少ない方がよいのだろうか。情報量をめぐる議論では、物理学におけるエントロピー概念との関連が指摘されることで、かえって混乱が生じている論点なので、コイン投げを参考に情報量の意味をもう一度考えてみる。

上記で定義した平均情報量は（情報）エントロピーと呼ばれ、少なくとも数式としては統計力学に登場するそれと同じである。

エントロピーが乱雑さの指標であるという表現はしばしば耳にする。ここまで議論してきた平均情報量もある意味、対象とする世界の乱雑さの指標と考えることができる。そもそも、最も乱雑な現象とはどのような状態を指すのであろうか。それは、何がおきるか全く見当がつかない場合、すなわちどの事象も等確率で起きる場合であろう。平均情報量は確かにこの場合最大となり、逆にある事象が確率１で起ることがわかっている場合には平均情報量は最小値ゼロとなる[3]。

情報の観点からは、「ニュースを聞く前までにもっていた情報のあいまいさ」あるいは「ニュースを聞く前までにもっていた無知の大きさ」と考えたほうがわかりやすいかもしれない。平均情報量が大きいほど、ニュースを聞くまでの我々の無知の程度は大きいということである。

しかし、我々が対象とするのが、理論や測定が不完全であるためではなく、本質的に確率的な現象[4]である場合、事前に自らの無知を完全に解消することは、定義上できない。この場合、情報のあいまいさは我々には如何ともし難い、所与の条件ともいうべきものである。かりに理論と測定が完璧であっても、事前の不確実性は残り、無知の指標としての平均情報量はゼロとはならない。

実際には、我々の理論と測定は完璧ではないので、この本質的に存在する下限としての

---

[3] $S = -\Sigma p_i \log p_i$ を $\Sigma p_i = 1$ の条件で最大化すればよい。なお、$I(0) = 0\log 0 = 0$ とする。

[4] このアインシュタインが死ぬまで認めなった立場が、現在の量子力学解釈の主流である。

理想的平均情報量に、さらに我々が理想から離れる程度に従って増加する、無知あるいは驚きの程度を加えた平均情報量の下で、現象に向かい合っている。我々の確率モデルが真のモデルから離れるにつれて、我々の無知は理想状態下のそれよりも、さらに大きくなるわけである。

４．相互情報量と情報価値

それでは、会計実証分析における中心概念である情報価値（の有無あるいは大小）は、情報理論の枠組みのもとで、どのように理解することができるのだろうか。

もう一度確認すれば、確率変数 y に関する平均情報量

$$S(y) = -\sum_{k=1}^{N} p(y_k) \log p(y_k)$$

は、変数 y の実現値という「ニュースを聞く前までにもっていた無知の大きさ」である。しかし、通常、我々が実証分析において興味があるのは、ある特定の変数の（無条件）確率分布それ自体ではなく、（ひとつあるいは複数の）説明変数を条件とする被説明変数[5]の条件付確率分布である[6]。たとえば、いわゆる価値関連性研究の目的は、株価変動そのものではなく、株価変動に会計数値がどれだけ連動しているか、すなわち株価を会計数値でどれだけ「説明」できるかにあるといえる。

そこで、変数 y とは別のそれと関連する（と予想される）現象のデータ x を既知（条件）[7]としたうえでの y の（条件付）確率を $p(y|x)$ とし、条件付エントロピー（平均情報量）を、

$$S(y|x) = -\sum_{k=1}^{N}\sum_{j=1}^{M} p(y_k \cap x_j) \log p(y_k|x_j)$$

と定義する。

説明変数 x の情報価値とは、それを知ることで、どれだけターゲットである被説明変数 y の変動をモデル化（予測）できるかに依存する。したがって、（無条件）確率分布に基づく y の平均情報量と x 既知のもとでの条件付確率分布に基づく条件付エントロピー（平均情報量）の差

$$\begin{aligned} I(y,x) &= S(y) - S(y|x) \\ &= -\sum_{k=1}^{N} p(y_k) \log p(y_k) + \sum_{k=1}^{N}\sum_{j=1}^{M} p(y_k \cap x_j) \log p(y_k|x_j) \\ &= -\sum_{k=1}^{N}\sum_{j=1}^{M} p(y_k \cap x_j) \log p(y_k) + \sum_{k=1}^{N}\sum_{j=1}^{M} p(y_k \cap x_j) \log \frac{p(y_k \cap x_j)}{p(x_j)} \end{aligned}$$

---

[5] 誤解を招く表現ではあるけれども、本稿では慣例に従って、「説明」・「被説明」変数という用語を用いる。

[6] 被説明変数の条件付期待値といってもよい。

[7] 議論の見通しを良くするため、データ（説明変数）数はひとつとする。

$$= \sum_{k=1}^{N}\sum_{j=1}^{M} p(y_k \cap x_j) \log \frac{p(y_k \cap x_j)}{p(y_k)p(x_j)}$$
$$= \sum_{k=1}^{N}\sum_{j=1}^{M} p(y_k \cap x_j)\left[\log p(y_k \cap x_j) - \log p(y_k)p(x_j)\right]$$

を使って、情報価値を測ればよい。これが情報理論において相互情報量と呼ばれる概念である。なお、最後の式から容易にわかるように、

$$I(y, x) = I(x, y)$$

である。

最も簡単な具体例であるコイン投げで相互情報量の意味を考えてみよう。

偏りのないコイン投げの場合、被説明変数 y（のとる値）はオモテ(H) かウラ(T) のどちらかが、それぞれ二分の一の確率

$$p(H) = p(T) = \frac{1}{2}$$

で生じるので、平均情報量 $S(y)$はオモテとウラの情報量の平均、すなわち、

$$S(y) = \frac{1}{2}\log 2 + \frac{1}{2}\log 2 = \log 2$$

となることは既に述べた。

さて、ある（説明）変数 x が、白(W) 及び赤(R) というふたつの値をとるとしよう。コイン投げの結果（被説明変数）を条件とするこの情報（説明変数）の条件付確率が

$$p(W|H) = 1,\ p(R|H) = 0,\ p(W|T) = 0,\ p(R|T) = 1$$

である場合、この情報を条件とするコイン投げの条件付確率と同時確率は

$$p(H|W) = 1,\ \ p(T|W) = 0$$
$$p(H|R) = 0,\ \ p(T|R) = 1$$

$$p(H \cap W) = \frac{1}{2},\ \ p(H \cap R) = 0$$

$$p(T \cap W) = 0,\ \ p(T \cap R) = \frac{1}{2}$$

となり[8]、条件付エントロピー（平均情報量）は

$$S(y|x) = \frac{1}{2}\log 1 + 0\log 0 + 0\log 0 + \frac{1}{2}\log 1 = 0$$

つまり、ゼロなので、y と x の相互情報量

[8] ベイズの定理に基づく導出は補論 4 参照。

$$I(y,x)=S(y)-S(y|x)=\log 2-0=\log 2$$

は、y の平均情報量と同じになる。x という情報があれば、平均情報量で表わされる我々の y に関する無知は完全に解消されるのである。この場合、y に関して x は最大限の情報価値を持つといってもよい。

今度は、ある別の（説明）変数 z が、やはり白(W) 及び赤(R) というふたつの値をとるとしよう。コイン投げの結果（被説明変数）を条件とするこの情報（説明変数）の条件付確率が

$$p(W|H)=\frac{1}{2},\ p(R|H)=\frac{1}{2},\ p(W|T)=\frac{1}{2},\ p(R|T)=\frac{1}{2}$$

である場合、この情報を条件とするコイン投げの条件付確率と同時確率は

$$p(H|W)=\frac{1}{2},\ \ p(T|W)=\frac{1}{2}$$
$$p(H|R)=\frac{1}{2},\ \ p(T|R)=\frac{1}{2}$$

$$p(H\cap W)=\frac{1}{4},\ \ p(H\cap R)=\frac{1}{4}$$
$$p(T\cap W)=\frac{1}{4},\ \ p(T\cap R)=\frac{1}{4}$$

となり、条件付エントロピー（平均情報量）は

$$S(y|x)=-\left(\frac{1}{4}\log\frac{1}{2}+\frac{1}{4}\log\frac{1}{2}+\frac{1}{4}\log\frac{1}{2}+\frac{1}{4}\log\frac{1}{2}\right)=\log 2$$

なので、相互情報量は

$$I(y,x)=S(y)-S(y|x)=\log 2-\log 2=0$$

つまり、ゼロとなる[9]。さきほどの例とは全く対照的に、z という「説明」変数は、y に関して何ら情報価値を持たず、y に対する我々の無知は全く改善されない。

結局、情報価値概念を焦点とする会計実証分析とは、有用な会計数値を用いたモデル化によって可能な限り対象となる被説明変数に関する我々の無知を減らすこと、条件付エントロピー（平均情報量）を最小化すなわち相互情報量を最大化することだといえる。コイン投げの例で示した通り、相互情報量がとり得る最大値は被説明変数の平均情報量であり、その場合、不確実性は完全に除去される。

５．カルバック - ライブラー情報量

現象を観察してモデルを構築する場合、我々の目標は真のモデルあるいは確率分布であ

[9] 相互情報量の最小値がゼロであることは次節を参照。

る。したがって、我々のモデルが真の確率分布とどれだけ近いかでモデルの良し悪しを判断すればよい。

しかし、その近さを如何に量として測ればよいのか。それに対する有力な提案がカルバック - ライブラー（Kullback - Leibler）情報量[10]（以下「KL 情報量」）である。前節の相互情報量は、ある（被説明）変数の（無条件）確率分布と別の（説明）変数を条件とする条件付確率分布の平均情報量の差を表わすものであった。

それに対し、ある現象（被説明変数）に対する（それと真に関連する）すべての情報を条件とした真の（条件付）確率分布を$\{p_1, p_2,\ldots, p_N\}$、我々が想定する（通常、真でない）モデルの（条件付）確率分布を$\{f_1, f_2,\ldots, f_N\}$とし、

$$D=\sum_{i=1}^{N} p_i(\log p_i-\log f_i)=\sum_{i=1}^{N} p_i\left(l\log\frac{p_i}{f_i}\right)$$

と定義されるのが KL 情報量である[11]。真のモデルと我々のモデルの「距離」を測るのが KL 情報量だといってもよい。KL 情報量は負とはならず、最小となるのは全ての確率が同じすなわち$p_i=f_i$の場合で、この時ゼロとなる[12]。もちろん、この場合、我々のモデルは真のモデルと一致しているわけである

連続変数の場合は、真の確率密度関数を$p(x)$、我々のモデルを$f(x)$とすれば、

$$D=\int_{-\infty}^{\infty} p(x)\left[\log p(x)-\log f(x)\right]dx=\int_{-\infty}^{\infty} p(x)\log\frac{p(x)}{f(x)}dx$$

となる。

現実には、我々は真の（条件付）確率分布を知らない。したがって、何らかの（真の分布とは異なっているという意味で）「間違った」モデルというレンズを通して現実を観察しているわけである。ある同じ確率現象を前にしていても、モデルの良し悪しによって、観察者の（平均）情報量は変わってくる。その意味で情報量は主観的な存在である。また、詳しくは第 3 節で述べるけれども、当然ながらモデルには測定という要素も含まれることは、会計研究者が留意しておくべき点である。

前述したように、我々が対象とするのが本質的に確率的な現象である場合、かりに理論と測定が完璧であっても不確実性は残り、無知の指標としての条件付エントロピー（平均情報量）はゼロとはならない。結局、実証研究者が目指すべきは、真の（条件付）確率分布とモデルの「距離」である KL 情報量をゼロにすることであって、条件付エントロピーすなわち不確実性をゼロにすることではない。

なお、前節の相互情報量の定義

---

[10] Kullback and Leibler (1951)。

[11] ここでは同じ被説明変数を対象とする条件付確率を前提に KL 情報量を定義している。より標準的な定義については、たとえば、吉田 (2009, 1 章)を参照。

[12] $t\log t\le t-1$ であることを利用すればよい。

$$I(y,x)=\sum_{k=1}^{N}\sum_{j=1}^{M}p(y_k\cap x_j)\left[\log p(y_k\cap x_j)-\log p(y_k)p(x_j)\right]$$

は、実は KL 情報量と同じである。したがって、相互情報量は必ず非負の値を取り、（被説明変数と説明変数の）真の同時分布と両者が独立と仮定した「モデル」の距離を測っているといってもよい。距離が最小になるのは、真の同時分布において事象と測定が独立となる、

$$p(y_k \cap x_j) = p(y_k)p(x_j)$$

つまり、「説明」変数が全く無意味な場合で、コイン投げの例で示したように、被説明変数に対する説明変数の情報価値である相互情報量はゼロとなる。

６．最尤法の考え方

まず、実証分析で広く用いられる推計手法である最尤（maximum likelihood）法を、KL 情報量との関連を念頭に、誤差が *i.i.d.*の正規分布に従うと仮定[13]された線形モデルを用いて、簡単に解説する。

具体的には、k 個の変数$\{y, x_2, x_3,\ldots, x_k\}$に関して、

$$y_i = \beta_1 + \beta_1 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2)$$

のモデルを仮定し、観測された n 個（$i = 1, 2,..., n$）のデータ$\{y_i, x_{2i}, x_{3i},\ldots, x_{ki}\}$を用いて推定する。

式の展開を見通しよくするため、定数項は「変数」1 の係数すなわち $x_{1i} = 1$ とし、モデル式を

$$y_i = \beta_1 x_{1i} + \beta_1 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

と表わす。そのうえで、定数項を含めた説明変数ベクトルと係数ベクトルを

$$x_i = (x_{1i}, x_{2i},\ldots, x_{ki})' = (1, x_{2i},\ldots, x_{ki})'$$

$$\beta = (\beta_1, \beta_2,\ldots, \beta_k)'$$

とすれば、y_i の（条件付）確率密度関数は

$$f(y_i \mid x_i;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - x_i'\beta)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\varepsilon_i^2}{2\sigma^2}}$$

となる。なお、推計を要するモデルのパラメータには係数だけでなく、分散を含まれるので、パラメータ（ベクトル）は $k+1$ の要素からなる

$$\theta = (\beta_1, \beta_2,\ldots, \beta_k, \sigma^2)'$$

である。

---

[13] 以下、この仮定が厳密には満たされていない場合、maximum likelihood は *quasi*-maximum likelihood ということになる。どのような場合に、仮定が正しくなくても、漸近的に一致性等の好ましい属性を持つかどうかは、ここでは議論しない。この点についての比較的わかりやすい解説として、Mittelhammer et al. (2000)を参照。

ここでは、すでにモデルは与えられていると仮定しているので、我々に残された課題は、n 個のデータ $\{y_i, x_{2i}, x_{3i}, \ldots, x_{ki}\}$ から $k+1$ のパラメータを推定することだけである。それでは最尤法で推定してみよう。

モデルとデータが与えられれば、

$$y = (y_1, y_2, \ldots, y_n)'$$

$$X = (x_1, x_2, \ldots, x_k)$$

と定義し、独立性の仮定から、我々が観察したデータの確率は

$$L(y|x;\theta) = f(y_1|x_1;\theta) \times f(y_2|x_2;\theta) \times \cdots \times f(y_n|x_n;\theta)$$

と表わせるはずである。これを θ の関数

$$L(\theta) = f_1(\theta) \times f_2(\theta) \times \cdots \times f_n(\theta)$$

とみた尤度関数を最大にする推定手法が最尤法である。確率が最大となるという意味で、最も尤もらしい推定値を選ぶわけである。正値をとる関数の対数変換は大小関係を維持するので、

$$l_i(\theta) = \log f_i(\theta)$$

と定義して、対数尤度関数

$$l(\theta) = \log L(\theta) = \log f_1(\theta) + \log f_2(\theta) + \cdots + \log f_n(\theta) = l_1(\theta) + l_2(\theta) + \cdots + l_n(\theta)$$

を用いても最大化問題を解いてもよい。

まず、

$$f_i(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(y_i - x_i'\beta)^2}{2\sigma^2}}$$

より、

$$l_i(\theta) = \log f_i(\theta) = -\frac{1}{2}\left[\log 2\pi + \log\sigma^2 + \frac{(y_i - x'\beta_i)^2}{\sigma^2}\right]$$

なので、対数尤度関数は

$$l(\theta) = \sum l_i(\theta) = -\frac{1}{2}\left[n\log 2\pi + n\log\sigma^2 + \sum\frac{(y_i - x'\beta_i)^2}{\sigma^2}\right]$$

$$= -\frac{1}{2}\left[n\log 2\pi + n\log\sigma^2 + \frac{1}{\sigma^2}(y - X\beta)'(y - X\beta)\right]$$

と表わすことができる。

これを θ で偏微分して、ゼロと置けば、

$$\frac{\partial l(\theta)}{\partial \beta} = -\frac{1}{2\sigma^2}\left(-2X'y + 2X'X\beta\right) = 0$$

$$\frac{\partial l(\theta)}{\partial \sigma^2} = -\frac{1}{2}\left[\frac{n}{\sigma^2} - \frac{1}{\sigma^4}(y - X\beta)'(y - X\beta)\right] = 0$$

となる。したがって、最尤推定値

$$\hat{\theta} = (b_1, b_2, \ldots, b_k, s^2)'$$

は、

$$b = (X'X)^{-1}X'y$$

$$s^2 = \frac{1}{n}(y - Xb)'(y - Xb) = \frac{\sum e_i^2}{n}$$

であることがわかる。なお、

$$e_i = y_i - x_i'b$$

である。

7．最尤法と KL 情報量

真の分布との距離を測る KL 情報量は、真の分布に基づく期待値オペレータ $E_p[\cdot]$によって表わせば、

$$D = E_p[\log p(x) - \log f(x)] = E_p[\log p(x)] - E_p[\log f(x)]$$

となる。我々の目的はなるべく真の分布に近いモデルを見つけることにあり、KL 情報量が非負であることから、結局、第二項であるモデルの平均対数尤度 $E_p[\log f(x)]$を最大化することに帰着する[14]。

とはいえ、我々は真の分布を知らないので、何らかの方法でこの平均対数尤度を推定しなければならない。まず思いつくのが、真の分布 p をデータに基づく経験（標本）分布 f で置き換えることである。つまり、（標本）対数尤度を標本数 n で割った

$$E_f\left[l_i(\hat{\theta})\right] = \frac{1}{n}\sum l_i(\hat{\theta}) = \frac{l(\hat{\theta})}{n}$$

である。要するに、最尤法による推定値は KL 情報量（第二項）推定値（の n 倍）でもある。最尤法とは、真の分布との距離を KL 情報量の意味で最小化する統計手法といえる。

しかしながら、最尤推定値は平均対数尤度の推定値としてはバイアスがある。具体的には、平均対数尤度を最大にする（真の分布ではなく）モデルのパラメータ値を θ_0 とし、

[14] 以下、本節の解説は小西・北川 (2004) による。

$$I(\theta_0) = E_p\left[\frac{\partial l_i(\theta)}{\partial \theta}\frac{\partial l_i(\theta)}{\partial \theta'}\bigg|_{\theta_0}\right]$$

$$J(\theta_0) = -E_p\left[\frac{\partial^2 l_i(\theta)}{\partial \theta \partial \theta'}\bigg|_{\theta_0}\right]$$

と定義すると、標本 x の同時分布 $\Pi p(x_i)$に関してとった期待値 $E_{p(x)}[\cdot]$の差つまりバイアスは、

$$\begin{aligned}
&E_{p(x)}\left[l(\hat{\theta}) - nE_p[l_i(\hat{\theta})]\right] \\
&= E_{p(x)}[l(\hat{\theta}) - l(\theta_0)] + E_{p(x)}\left[l(\theta_0) - nE_p[l_i(\theta_0)]\right] + E_{p(x)}\left[nE_p[l_i(\theta_0)] - nE_p[l_i(\hat{\theta})]\right] \\
&= \frac{1}{2}tr\left[I(\theta_0)J^{-1}(\theta_0)\right] + 0 + \frac{1}{2}tr\left[I(\theta_0)J^{-1}(\theta_0)\right] \\
&= tr\left[I(\theta_0)J^{-1}(\theta_0)\right]
\end{aligned}$$

となる[15]。

このバイアスの推計値を「ペナルティ」として対数尤度から引いて、(歴史的経緯から) マイナス 2 をかけた

$$TIC = -2\left[l(\hat{\theta}) - tr(\hat{I}\hat{J}^{-1})\right] = -2l(\hat{\theta}) + 2tr(\hat{I}\hat{J}^{-1})$$

が竹内情報量規準 (Takeuchi Information Criterion; TIC) である[16]。マイナス 2 をかけたため、値が小さいほど良いモデルということに注意が必要である。

ところで、最尤法が KL 情報量と密接な関連があり、最尤推定値から KL 情報量の推定値が得られるにしても、それが実証研究にどう役立つのだろうか。実は、ここまでの議論では、モデルの同定 (specification) はすでに終わり、我々に残った課題はその値を推定することであった。

しかし、実証研究における実際のモデル構築においては、どのような説明変数をモデルに採用するかを研究者が自ら決定しなければならない。したがって、説明変数の選択にあたって何らかの客観的規準があれば好都合である。

まさに TIC はその目的に適合する。研究者が理論的あるいは機械的に選んだいくつかのモデルの TIC を比較して、最も低い値となったモデルを選べばよいわけである。対数尤度は変数を追加するほど推定値が大きくなるので、変数選択規準としてはふさわしくない。ところが、TIC はバイアスを補正するペナルティ項が付いているため、変数が多いからといって低い値になるとは限らない。

---

[15] 式の導出は補論 1 を参照。

[16] 竹内 (1976)。

８．TIC 簡略版としての AIC

実際には、TIC はその理論的魅力にもかかわらず、それほど用いられていない。なぜなら、バイアス修正項を推定せねばならず、どのような手法を用いても、標本に依存する推定値の変動を避けることができない。

とはいえ、$I(\theta_0)$と $J(\theta_0)$は無関係ではなく、

$$\frac{\partial^2 l_i(\theta)}{\partial\theta\partial\theta'} = \frac{\partial^2 \log f_i(\theta)}{\partial\theta\partial\theta'} \frac{\partial\left[\dfrac{\partial \log f_i(\theta)}{\partial\theta'}\right]}{\partial\theta} = \frac{\partial\left[\dfrac{1}{f_i(\theta)}\dfrac{\partial f_i(\theta)}{\partial\theta'}\right]}{\partial\theta}$$

$$= \frac{1}{f_i(\theta)}\frac{\partial^2 f_i(\theta)}{\partial\theta\partial\theta'} - \frac{1}{f_i^2(\theta)}\frac{\partial f_i(\theta)}{\partial\theta}\frac{\partial f_i(\theta)}{\partial\theta'}$$

$$= \frac{1}{f_i(\theta)}\frac{\partial^2 f_i(\theta)}{\partial\theta\partial\theta'} - \frac{\partial \log f_i(\theta)}{\partial\theta}\frac{\partial \log f_i(\theta)}{\partial\theta'}$$

$$= \frac{1}{f_i(\theta)}\frac{\partial^2 f_i(\theta)}{\partial\theta\partial\theta'} - \frac{\partial l_i(\theta)}{\partial\theta}\frac{\partial l_i(\theta)}{\partial\theta'}$$

なので、両辺を真の分布に従って、θ_0 での期待値を取ると、

$$E_p\left[\left.\frac{\partial^2 l_i(\theta)}{\partial\theta\partial\theta'}\right|_{\theta_0}\right] = E_p\left[\left.\frac{1}{f_i(\theta)}\frac{\partial^2 f_i(\theta)}{\partial\theta\partial\theta'}\right|_{\theta_0}\right] - E_p\left[\left.\frac{\partial l_i(\theta)}{\partial\theta}\frac{\partial l_i(\theta)}{\partial\theta'}\right|_{\theta_0}\right]$$

つまり、$I(\theta_0)$と $J(\theta_0)$には、

$$-J(\theta_0) = E_p\left[\left.\frac{1}{f_i(\theta)}\frac{\partial^2 f_i(\theta)}{\partial\theta\partial\theta'}\right|_{\theta_0}\right] - I(\theta_0)$$

という関係がある。

もし、$p = f(\theta_0)$すなわちモデルの中に真の分布が含まれていれば、この式は

$$-J(\theta_0) = \int \frac{1}{f_i(\theta)}\frac{\partial^2 f_i(\theta)}{\partial\theta\partial\theta'} f_i(\theta)dx\Big|_{\theta_0} - I(\theta_0)$$

$$= \frac{\partial^2}{\partial\theta\partial\theta'}\int f_i(\theta)dx\Big|_{\theta_0} - I(\theta_0)$$

となり、$\int f_i(\theta)dx = 1$ なので、右辺第 1 項はゼロ、したがって $J(\theta_0) = I(\theta_0)$となり、$E_{k+1}$ を k+1 ×k+1 の単位行列とすると、

$$tr\left[I(\theta_0)J^{-1}(\theta_0)\right] = tr(E_{k+1}) = k+1$$

となる。

この真の分布がモデルに含まれている場合、TIC は

$$TIC = -2\left[l(\hat{\theta}) - tr(\hat{I}\hat{J}^{-1})\right] = -2l(\hat{\theta}) + 2tr(\hat{I}\hat{J}^{-1}) = -2l(\hat{\theta}) + 2(k+1)$$

となり、バイアスはパラメータの個数 k+1 となって、その推定が不要となる。実は、このモデルに真の分布が含まれているという仮定を加えて、TIC をいわば簡略化したのが、モデル選択規準として広く用いられている赤池情報規準（Akaike Information Criterion; AIC）である[17]。つまり、AIC

$$AIC = -2l(\hat{\theta}) + 2(k+1)$$

は想定したモデルが「十分よい」近似であれば、モデル推定に追加的ノイズをもたらすバイアスの推定が不要で、対数尤度の計算とパラメータの数え上げという「機械的」作業でモデル選択を可能にする、極めて使いやすい規準である[18]。AIC においては、パラメータ数がそのままペナルティ項となるので、変数を増やしたからといってモデルの当てはまりがよくなるとは言えないことが、TIC に比べ、より明瞭となっている。

９．AIC への批判

AIC の問題点としてしばしば指摘されるのが、モデル選択における一致性を持たないことである。誤差が *i.i.d.*の正規分布に従うと仮定すると、線形モデル

$$y = \beta_1 + \beta_1 x_2 + \ldots + \beta_k x_k + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$$

の最大対数尤度は

$$l(\hat{\theta}) = -\frac{1}{2}\left[n\log 2\pi + n\log s^2 + \sum \frac{e_i^2}{s^2} \right] = -\frac{1}{2}(n\log 2\pi + n\log s^2 + n)$$

となる。したがって、（分散を含む）パラメータ数が $m\,(=k+1)$のモデル M の AIC は

$$AIC = -2\left[l(\hat{\theta}) - m \right] = n\log 2\pi + n\log s^2 + n + 2m$$
$$= n\left[\log 2\pi + \log s^2 + 1 + \frac{2m}{n} \right]$$

となる。つまり、AIC は

$$\log s^2 + \frac{2m}{n}$$

の大小によってモデル選択を行うことになる。

ちなみに、モデル選択規準としてしばしば用いられる自由度修正済 R^2 は（定数項を含む）説明変数の数を k として、

---

[17] 実際は AIC の方が先に提唱された（Akaike 1973）。AIC の詳しい解説は小西・北川 (2004) を参照。赤池 (1976) は本人によるわかりやすい解説である。

[18] 実証研究において AIC が TIC を「圧倒」していることは、理論的には望ましい GLS より OLS が用いられることと似ている。

$$\bar{R}^2 = 1 - \frac{\frac{\sum e_i^2}{n-k}}{\frac{\sum (y_i - \bar{y})^2}{n-1}}$$

と表わせるので、$\bar{R}^2$ を規準に用いるということは、$\log\left(\frac{\sum e_i^2}{n-k}\right)$ の大小、つまり、

$$\begin{aligned}
\log\left(\frac{\sum e_i^2}{n-k}\right) &= \log\left(\frac{\sum e_i^2}{n}\frac{n}{n-k}\right) = \log s^2 + \log\left(\frac{n}{n-k}\right) \\
&= \log s^2 + \log\left(1 + \frac{k}{n-k}\right) \\
&\simeq \log s^2 + \frac{k}{n-k} \\
&\simeq \log s^2 + \frac{k}{n}
\end{aligned}$$

の大小でモデルを選択することを意味する。AIC に比べて、$\bar{R}^2$ のペナルティ項は半分になっている。

ところで、パラメータ数 m_0 (=k_0 + 1)のモデル M_0

$$y_i = \beta_1 + \beta_2 x_{2i} + \ldots + \beta_{k_0} x_{k_0 i} + \varepsilon_i$$

とそれを（係数に制約がかかった）部分集合として含むすなわちネストするパラメータ数 m_A (=k_A + 1)のモデル M_A

$$y_i = \beta_1 + \beta_2 x_{2i} + \ldots + \beta_{k_0} x_{k_0 i} + \ldots + \beta_{k_A} x_{k_A i} + \varepsilon_i$$

の対数尤度比検定（likelihood ratio test）は、M_0 が真のモデルであれば、対数尤度の差の二倍が漸近的に自由度 m_A－ m_0 (= k_A－ k_0)のカイ二乗分布に従う[19]ことを利用する。

それでは、標本数が多くなった場合、AIC は真のモデル M_0 を必ず選択する、すなわち、必ず $AIC_0 < AIC_A$ となるであろうか。両者の差は

$$\begin{aligned}
&AIC_0 - AIC_A \\
&= -2\left[l(\theta_0) - m_0\right] - \left[-2\left[l(\theta_A) - m_A\right]\right] = 2\left[l(\theta_A) - l(\theta_0)\right] - 2(m_A - m_0)
\end{aligned}$$

なので、

[19] 補論 3 参照。

$$P\left(AIC_0 < AIC_A \mid M_0\right) = P\left(AIC_0 - AIC_A = 2\left[l(\theta_A) - l(\theta_0)\right] - 2(m_A - m_0) < 0 \mid M_1\right)$$
$$= P\left(LR = 2\left[l(\theta_A) - l(\theta_0)\right] < 2(m_A - m_0)\right)$$
$$\to P\left(\chi^2(m_A - m_0) < 2(m_A - m_0)\right) < 1$$

となり、真のモデル M_0 を選ぶ確率は 1 に収束しない。標本数が無限大となった場合、必ず真のモデルを選択するという意味での一致性を AIC は持たないのである[20]。

一方、AIC に類似した選択基準である BIC は

$$BIC = -2l(\hat{\theta}) + m\log n$$

の大小でモデルを選択する。バイアス修正項においてパラメータ数の係数を 2 ではなく log$n$ とするわけである。

AIC と同様の代入を行い整理すると、

$$P\left(BIC_0 < BIC_A \mid M_A\right) = P\left(BIC_0 - BIC_A = 2\left[l(\theta_A) - l(\theta_0)\right] - (m_A - m_0)\log n < 0 \mid M_1\right)$$
$$= P\left(LR = 2\left[l(\theta_A) - l(\theta_0)\right] < (m_A - m_0)\log n\right)$$
$$\to P\left(\frac{\chi^2(m_A - m_0)}{\log n} < (m_A - m_0)\right) = 1$$

となる。つまり、AIC と異なり、標本数増加につれてペナルティが重くなるため、BIC は一致性を持つ。

以上の結果を見れば、AIC ではなく BIC の方が好ましいのではないか、という当然の疑問が生じる。

10. AIC 批判への反批判

モデル選択規準に一致性があるかないかという議論は、対象となるモデル集合のなかに真のモデルが含まれていることを仮定している。ところで、多くの実証研究においては、

$$y_i = \beta_1 + \beta_2 x_{2i} + \ldots + \beta_{k_0} x_{k_0 i} + \ldots + \beta_{k_A} x_{k_A i} + \varepsilon_i$$

という（研究者が「ベスト」と考える）モデル M_A に対し、

$$\beta_{k_0+1} = \beta_{k_0+2} = \ldots = \beta_{k_A} = 0$$

という $m_A - m_0$ $(= k_A - k_0)$個の制約を課した、帰無仮説であるモデル M_0

$$y_i = \beta_1 + \beta_2 x_{2i} + \ldots + \beta_{k_0} x_{k_0 i} + \varepsilon_i$$

が真のモデルだとして、仮説検定が行われる。

このように帰無仮説モデルがネストされている場合、通常の仮説検定は、帰無仮説が（五

[20] ペナルティが AIC の半分である $\overline{R}^2$ は当然ながら一致性を持たない。

パーセント等）一定の有為水準で棄却された場合にはモデル M_A が、棄却されなかった場合はモデル M_0 が選ばれるという、モデル選択問題と解釈することもできる。一方、AIC ではその閾値がパラメータの差の二倍に「機械的」に設定されるわけである。下平 (2007, 144-145 頁)が指摘するように、「仮説検定と AIC はその背景となる問題意識は異なるものの、結果として同じ統計量を異なる閾値で使っている」のである。

五パーセント等のルーティンに従うのであればともかく、有意水準をどう選ぶかというのは決め手を欠いた困難な問題である。AIC はそれに対し、KL 情報量概念に基づいて一定の答えを与えたと考えることもできる。

さらに、AIC に一致性がないという批判は、モデルというもの、あるいは観察データを利用してモデルを作るという行為への、ある意味で奇妙な考え方に基づいている。それは、データを生成した真のモデルが存在し、十分なデータがあれば、我々はそれを見つけることができるという考え方である。

それに対し、AIC は真のモデルは到達不可能な「ニルヴァナ（涅槃）」であって、モデルとはどこまで行っても近似に過ぎないという、モデルというものに対する、よりプラグマティックな発想に基づいており、実証研究という現場の実態をはるかによく捉えている。ここでは、北川 ・小西 (2004, 66 頁)の言葉を借りて、AIC に一致性がないという批判の問題点を指摘しておく。

> われわれのモデリングの目的は「よい」モデルを求めることであって、「真の」モデルを求めることではないということである。統計的モデルが複雑なシステムをある目的に沿って近似したものであることを思い起こせば、真の次数を推定するという問題設定自体が適当でないものであることは明らかである。真のモデルや真の次数が明確に定義づけられるのは、シミュレーション実験を行う場合のようにきわめて限られた状況である。モデルは複雑な現象の近似であるという立場からあえていえば、真の次数は無限大ともいえる。
>
> たとえ、有限の真の次数が存在するものと仮定しても、よいモデルの次数が真の次数と一致するとは限らないのである。データ数が少ない場合には、推定されるパラメータの不安定さを考えると、より低次のモデルを用いた方がより高い予測精度が得られる可能性があることを AIC は示している。

11. AIC と仮説検定の融合

モデルにスカラーの「点数」を与える（小さいほどよい）AIC は、極めてわかりやすいモデル選択規準である。とはいえ、当然ながら AIC は我々の得たデータに依存する確率変数（の実現値）なので、モデル間の AIC の差が「わずか」な場合、数値の小さいモデルがよいとは言い切れない。

すでにみたように、ネストされた関係にあるモデル間の AIC に基づく選択は、ネストさ

れたパラメータ数の少ないモデルを帰無仮説として行う通常の仮説検定では有意水準で決まる閾値に、パラメータの差の二倍を使っていると解釈することができた。したがって、モデル間の AIC の差がどれほど有意であるかを計算できる。

しかし、これは互いネストされた関係にあるモデル間の場合に限られる。そもそも、ネストされた関係がないすなわちノンネストの場合は、通常の仮説検定は行えない。むしろ、AIC はネストされたモデル間の比較にしか使えない仮説検定の「弱点」を持たないところにその強みがあるともいえる。

それではノンネスト・モデル間の AIC を比較した場合、その差がどれほど意味のあるものなのか、たとえば、実現値がデータに依存することを考慮しても、単なる偶然とはいえないほど「大きな」差であるのかを知る方法はないのであろうか。

こうした問題意識のもと、Linhart (1988) 、Kishino and Hasegawa (1989)及び Vuong (1989) は、ノンネスト・モデル間の対数尤度の差が漸近的に正規分布に従うことを利用した検定を提案した[21]。

ネストされた（パラメータ数の少ない）モデルを帰無仮説すなわち真のモデルと仮定して行う通常の仮説検定と異なり、ノンネスト・モデルを比較する場合には、どちらかのモデルにも帰無仮説という「優先的」地位を与えず、両方とも「間違った」（misspecified）モデルであって、「桃源郷」である真のモデルとの距離が同じかどうかという観点で検定を行う。「モデリングの目的は『よい』モデルを求めることであって、『真の』モデルを求めることではない」という AIC の基本的発想に沿った考え方だといえる。

ふたつのモデル M_a と M_b の KL 情報量で測った真のモデルからの距離が同じであれば、標本に基づいて計算した両者の（平均）対数尤度の推定値の差は平均的にゼロとなるはずである。しかし、推定値の大小が真の相違を反映したものかどうかを知るには、そのサンプリング誤差を知る必要がある。

具体的には、最尤推定値を l_a、l_b、バイアス修正項を B_a、B_b とし、（修正された）対数尤度の推定値（平均対数尤度の n 倍の推定値）の差 Δl

$$\Delta l = \left[l(\hat{\theta}_a) - B_a \right] - \left[l(\hat{\theta}_b) - B_b \right]$$

をその分散の推定値で割った

$$z = \frac{\Delta l}{\sqrt{\mathrm{var}(\Delta l)}}$$

が漸近的に分散 1 の正規分布に従うことを利用して、二つのモデルの真のモデルからの距離が同じかどうかという仮説検定を行う。z が「大きな」正値であれば M_a の方が、逆に「大きな」負値であれば M_b の方が良いモデルと考えることもできる。

---

[21] AIC に基づくモデル選択検定の詳細については、Shimodaira (1997)、下平 (1999, 2004, 2007)を参照。

z の分子の推定は自明なので、分母の推定を詳しく解説する。具体的な計算過程を見やすくするため、誤差が *i.i.d.*の正規分布に従う場合を考える[22]。

仮想上の（定数項を含む）変数 k 個のフルモデル

$$y_i = \beta_1 + \beta_1 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2)$$

にそれぞれ、

$$\beta_2 = \ldots = \beta_{k-a+1} = 0$$

$$\beta_{b+1} = \ldots = \beta_k = 0$$

の制約をかけた互いにネストされた関係にない

$$y_i = \beta_1 + \beta_{k-a+2} x_{k-a+2,i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

$$y_i = \beta_1 + \beta_2 x_{2i} + \ldots + \beta_b x_{bi} + \varepsilon_i$$

をそれぞれ、モデル M_a と M_b とする。

まず、最尤法により、

$$l_i(\hat{\theta}_a) = -\frac{1}{2}\left(\log 2\pi + \log s_a^2 + \frac{e_{ai}^2}{s_a^2} \right)$$

$$l_i(\hat{\theta}_b) = -\frac{1}{2}\left(\log 2\pi + \log s_b^2 + \frac{e_{bi}^2}{s_b^2} \right)$$

から、平均対数尤度の推定値

$$\bar{l}_a = \frac{\sum l_i(\hat{\theta}_a)}{n} = -\frac{1}{2}\left(\log 2\pi + \log s_a^2 + \frac{\sum e_{ai}^2}{s_a^2} \right) = -\frac{1}{2}(\log 2\pi + \log s_a^2 + 1)$$

$$\bar{l}_b = \frac{\sum l_i(\hat{\theta}_b)}{n} = -\frac{1}{2}\left(\log 2\pi + \log s_b^2 + \frac{\sum e_{bi}^2}{s_b^2} \right) = -\frac{1}{2}(\log 2\pi + \log s_b^2 + 1)$$

を得る。

バイアス修正は分散には影響しないことと *i.i.d.*の仮定から、

$$\begin{aligned} \mathrm{var}(\Delta l) &= \mathrm{var}(l(\theta_a) - l(\theta_b)) = \mathrm{var}\left[\sum \left(l_i(\theta_a) - l_i(\theta_b) \right) \right] \\ &= n \cdot \mathrm{var}\left[l_i(\theta_a) - l_i(\theta_b) \right] \end{aligned}$$

となるので、標本を用いて $\mathrm{var}\left[l_i(\theta_a) - l_i(\theta_b) \right]$ を推定すれば、

$$\mathrm{var}\left[l_i(\hat{\theta}_a) - l_i(\hat{\theta}_b) \right] = \frac{\sum \left[l_{ai} - l_{bi} - (\bar{l}_a - \bar{l}_b) \right]^2}{n} = \frac{1}{4n} \sum \left(\frac{e_{ai}^2}{s_a^2} - \frac{e_{bi}^2}{s_b^2} \right)^2$$

であることがわかる。したがって、

[22] この仮定が正しくない場合、quasi-maximum likelihood に基づくと考えればよい。

$$\mathrm{var}\left[l(\hat{\theta}_a) - l(\hat{\theta}_b) \right] = n \cdot \mathrm{var}\left[l_i(\hat{\theta}_a) - l_i(\hat{\theta}_b) \right] = n \cdot \frac{1}{4n} \sum \left(\frac{e_{ai}^2}{s_a^2} - \frac{e_{bi}^2}{s_b^2} \right)^2 = \frac{1}{4} \sum \left(\frac{e_{ai}^2}{s_a^2} - \frac{e_{bi}^2}{s_b^2} \right)^2$$

より、二つのモデルに差がなければ漸近的に、

$$z = \frac{\Delta l}{\mathrm{var}(\Delta l)} = \frac{\Delta l}{\sqrt{\frac{1}{4} \sum \left(\frac{e_{ai}^2}{s_a^2} - \frac{e_{bi}^2}{s_b^2} \right)^2}} = \frac{2\Delta l}{\sqrt{\sum \left(\frac{e_{ai}^2}{s_a^2} - \frac{e_{bi}^2}{s_b^2} \right)^2}} \to N(0,1)$$

となる。

バイアス修正を行わないのであれば、Δl には対数尤度差を用いればよい。いわゆる Vuong (1989) 検定である。

バイアス修正に AIC を用いる場合[23]は、パラメータ数に分散を含めてそれぞれ、

$$AIC_a = -2\left[l(\hat{\theta}_a) - (a+1) \right]$$

$$AIC_b = -2\left[l(\hat{\theta}_b) - (b+1) \right]$$

なので、平均対数尤度の n 倍（の推定値）の差 Δl は

$$\Delta l = \left[l(\hat{\theta}_a) - (a+1) \right] - \left[l(\hat{\theta}_b) - (b+1) \right] = \left[l(\hat{\theta}_a) - l(\hat{\theta}_b) \right] - (a-b)$$

となり、

$$z = \frac{\Delta l}{\mathrm{var}(\Delta l)} = \frac{2\Delta l}{\sqrt{\sum \left(\frac{e_{ai}^2}{s_a^2} - \frac{e_{bi}^2}{s_b^2} \right)^2}} = \frac{2\left[\left[l(\hat{\theta}_a) - (a+1) \right] - \left[l(\hat{\theta}_b) - (b+1) \right] \right]}{\sqrt{\sum \left(\frac{e_{ai}^2}{s_a^2} - \frac{e_{bi}^2}{s_b^2} \right)^2}}$$

$$= \frac{AIC_b - AIC_a}{\sqrt{\sum \left(\frac{e_{ai}^2}{s_a^2} - \frac{e_{bi}^2}{s_b^2} \right)^2}} \to N(0,1)$$

を得る。

バイアス修正に TIC を用いる場合、

[23] Vuong (1989, 318-319 頁)も、対数尤度比検定への修正として、AIC 及び BIC に基づく分子修正に言及している。

$$z=\frac{\Delta l}{\mathrm{var}(\Delta l)}=\frac{2\left[\left[l(\hat{\theta}_a)-tr(\hat{I}_a\hat{J}_a^{-1})\right]-\left[l(\hat{\theta}_b)-tr(\hat{I}_b\hat{J}_b^{-1})\right]\right]}{\sqrt{\sum\left(\frac{e_{ai}^2}{s_a^2}-\frac{e_{bi}^2}{s_b^2}\right)^2}}$$

$$=\frac{TIC_b-TIC_a}{\sqrt{\sum\left(\frac{e_{ai}^2}{s_a^2}-\frac{e_{bi}^2}{s_b^2}\right)^2}}=\rightarrow N(0,1)$$

であることは、もはや自明であろう。

なお、AIC 及び TIC いずれにおいても、バイアス修正を半分にした平均対数尤度の n 倍（の推定値）の差

$$\Delta l=\left[l(\hat{\theta}_a)-\frac{a+1}{2}\right]-\left[l(\hat{\theta}_b)-\frac{b+1}{2}\right]$$

$$\Delta l=\left[l(\hat{\theta}_a)-\frac{tr(\hat{I}_a\hat{J}_a^{-1})}{2}\right]-\left[l(\hat{\theta}_b)-\frac{tr(\hat{I}_b\hat{J}_b^{-1})}{2}\right]$$

を用いる方が、有意差検定においては望ましいことが Linhart (1988)及び Choi and Kiefer (2011)により指摘されている。

12. 会計実証分析における AIC と Vuong 検定

現実にはすべてのモデルは近似に過ぎないことを考えれば、真のモデルを前提としない AIC（あるいは TIC）の考え方は、モデル構築の現場にいる（会計に限らず）実証研究者にとって極めて魅力的である。しかも、ここで示したように、AIC は伝統的な仮説検定と対立するものではなく、むしろモデル選択と仮説検定を統一的に理解するうえでも有用な概念である。会計実証分析において多用される Vuong 検定もこの文脈で理解することができる。

しかしながら、下平 (2007, 145-146 頁)が指摘するように、ある（被説明）変数の変動を予測[24]することがモデル構築の目的であれば、（定数項を含む）変数 k 個のフルモデル

$$y_i=\beta_1+\beta_1x_{2i}+\ldots+\beta_kx_{ki}+\varepsilon_i \qquad \varepsilon_i\sim N(0,\sigma^2)$$

にそれぞれ、

$$\beta_2=\ldots=\beta_{k-a+1}=0$$

$$\beta_{b+1}=\ldots=\beta_k=0$$

の制約をかけた

[24] 「予測」には *pre*diction のみならず *post*diction も含むものとする。後者はむしろ「説明」と呼ぶべきかもしれない。

$$y_i = \beta_1 + \beta_{k-a+2} x_{k-a+2,i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$
$$y_i = \beta_1 + \beta_2 x_{2i} + \ldots + \beta_b x_{bi} + \varepsilon_i$$

というノン・ネストモデルの Vuong 検定型比較を行うより、「フルモデルだけを考えて、必要に応じて回帰係数が不安定にならないような工夫…をすれば十分」とも考えられる。

もちろん、それぞれのノン・ネストモデルが互いに両立しない排他的な仮説を反映している場合など、データの「当てはまり」が良いからといって、フルモデルを採用することが理論的に正当化できないこともあろう。

とはいえ、会計実証分析においては、会計数値の情報価値が論点となっており、ある一群のデータと別の一群のデータが理論的に排他的関係にあることは稀である。たとえば、純利益関連データと包括利益関連データを両方とも含むモデルを排除する確乎たる理論的根拠を筆者は寡聞にして知らない。

そもそも、純利益関連データと包括利益関連データを比べて、どちらがより多くの情報価値を持つかというような二者択一的問題設定自体が疑問である。生物学[25]における理論と実証の交差点としてのノン・ネストモデル選択の場合と異なり、フルモデルを明確に否定する理論仮説なしにデータの情報価値の大小を議論の焦点とする会計実証分析において、ノン・ネストモデル選択の有用性は限定的である。

むしろ、「モデルは複雑な現象の近似であるという立場からあえていえば、真の次数は無限大」であって、標本数が増えるとより多くの変数を含むモデルが（真ではなく）最良となる可能性が高まる AIC の枠組みのもとでは、データ生成に多大な追加コストがかかるのでない限り、確たる理論的制約がない会計数値の情報価値が論点の場合、データの種類は多ければ多いほどよいといえる。

いずれにせよ、会計数値という情報を扱う我々会計研究者の立場から見れば、AIC そのものというより、その背後にある情報理論が、次節以降に示すように、実証と測定を統一的に理解するヒントも与えてくれるのである。

13. 測定と情報量概念

ここまで、測定の対象となる現象と測定されたデータを区別せず、モデルの良し悪しを情報理論の観点から議論してきた。しかしながら、両者の区別は会計人にとって raison d'être にかかわる最重要論点である。実際、会計をめぐる言説は、如何に実体を公正忠実に測定するかに尽きるともいえる。

そこで、本節と次節では、情報理論の観点からこの実体と測定の乖離[26]（の可能性）に焦点をあてる。具体的には、第 4 節で導入した相互情報量概念を用いて、実体と測定の関係を考察する。

---

[25] たとえば、ノン・ネストモデル選択の嚆矢となった Kishino and Hasegawa (1989)を参照。

[26] 本稿では、測定が実体に影響する可能性は捨象し、如何に正確に実体を測定するかという観点で考えていく。

まず、実体 y を測定した得られたデータ[27]を x とし、x を知ったうえでの y の（条件付）確率を $\log p(y|x)$とすれば、条件付エントロピー（平均情報量）は

$$S(y|x) = -\sum_{k=1}^{N}\sum_{j=1}^{M} p(y_k \cap x_j)\log p(y_k|x_j)$$

となる。実際には存在しないものを「測定」したり、その逆に存在するものを把握できない可能性を考慮すれば、$N = M$ とは限らない。

測定によって我々の実体に関する「無知」をどれだけ解消できたかは、平均情報量と条件付エントロピー（平均情報量）の差である相互情報量

$$\begin{aligned} I(y,x) &= S(y) - S(y|x) \\ &= \sum_{k=1}^{N}\sum_{j=1}^{M} p(y_k \cap x_j)\left[\log p(y_k \cap x_j) - \log p(y_k)p(x_j)\right] \end{aligned}$$

で測ればよい。当節の文脈においては、相互情報量とは測定の情報価値だといえる。

第 4 節同様、最も簡単な具体例であるコイン投げで測定の情報価値を考えてみよう。

偏りのないコイン投げの場合、事象（実体）y はオモテ(H) かウラ(T) のどちらかで、それぞれ二分の一の確率

$$p(H) = p(T) = \frac{1}{2}$$

で生じるので、平均情報量 $S(y)$はオモテとウラの情報量の平均、すなわち、

$$S(y) = \frac{1}{2}\log 2 + \frac{1}{2}\log 2 = \log 2$$

となることは既に述べた。

さて、測定システム x からは「オモテ」(I_H)及び「ウラ」(I_T)という二通りの情報が発信されるとしよう。情報自体もそれぞれ二分の一の確率

$$p(I_H) = \frac{1}{2},\quad p(I_T) = \frac{1}{2}$$

で生じるとしよう。ただし、その測定情報が実体を反映しているとは限らない。

まず、測定情報を条件とする実体の条件付確率を

$$\begin{aligned} &p(H|I_H) = 1,\ \ p(T|I_H) = 0 \\ &p(H|I_T) = 0,\ \ p(T|I_T) = 1 \end{aligned}$$

と仮定する。この場合、情報と実体の同時確率は

[27] 現象 y そのものではなく、それと関連する（と予想される）現象の測定と考えてもよい。

$$p(H \cap I_H) = \frac{1}{2},\ p(H \cap I_T) = 0$$

$$p(T \cap I_H) = 0,\ p(T \cap I_T) = \frac{1}{2}$$

となり、条件付エントロピー（平均情報量）は

$$S(y|x) = \frac{1}{2}\log 1 + 0\log 0 + 0\log 0 + \frac{1}{2}\log 1 = 0$$

つまり、ゼロなので、実体と測定の相互情報量は実体の平均情報量と同じ、

$$I(y,x) = S(y) - S(y|x) = \log 2 - 0 = \log 2$$

となる。この測定システムのもとでは情報と実体が一致するので、実体の情報量を測定によって完全に読み取ることができる。測定ノイズがゼロといってもよい。

別の測定システムでは、測定情報を条件とする条件付確率が、

$$p(H|I_H) = \frac{1}{2},\ p(T|I_H) = \frac{1}{2}$$

$$p(H|I_T) = \frac{1}{2},\ p(T|I_T) = \frac{1}{2}$$

で与えられているとしよう。この場合、情報と実体の同時確率は

$$p(H \cap I_H) = \frac{1}{4},\ p(H \cap I_T) = \frac{1}{4}$$

$$p(T \cap I_H) = \frac{1}{4},\ p(T \cap I_T) = \frac{1}{4}$$

となり、条件付エントロピー（平均情報量）は実体の平均情報量と同じ、

$$S(y|x) = -\left(\frac{1}{4}\log\frac{1}{2} + \frac{1}{4}\log\frac{1}{2} + \frac{1}{4}\log\frac{1}{2} + \frac{1}{4}\log\frac{1}{2}\right) = \log 2$$

なので、相互情報量は

$$I(y,x) = S(y) - S(y|x) = \log 2 - \log 2 = 0$$

つまり、ゼロとなる。この測定システムから生み出される「情報」に情報価値はなく、我々の「無知」は全く改善されない。

一般には測定システムはここで挙げた二つの極端な例、完璧なシステムと全く役に立たないシステムの中間に位置するであろう。しかも、次節で解説するように、測定誤差を標本数の増加で「カバー」することはできない。いずれにせよ、測定されたデータの信頼性を考慮しない実証分析が GIGO (garbage-in garbage-out)の危険性を孕んでいることは確かである。

14. 測定誤差がもたらすバイアス

所得や利益など、我々が実証分析に使うデータは、本来あるべき理念型あるいは理想型から見れば不十分なものであり、測定というレンズを通して、ノイズやバイアスを含んだものになっている。したがって、平均情報量あるいはエントロピーを減少させるという観点から考えると、測定への考慮を払わずデータを与えられたものと考える昨今の実証分析には大きな問題がある。

会計測定の規範論から会計情報（価値）の実証という流れのなかで、どのように測定されたかよりも、株価等との相関を基準にデータとして有用かどうかが、会計情報の価値を左右するという視点が会計研究者の間で支配的となりつつある。

しかし、情報価値という視点を中心に据えるにしても、正確な測定が重要であることを前節で示した。それに対して、測定でノイズが入っても標本数が多ければ、（変数の）係数の推定値は正確になるので、バイアスさえ修正すれば大きな問題ではない。たとえば、資産簿価が平均的に 20 パーセント少なめに報告されているとすれば、報告データを 1.25 倍（＝1/0.8）したうえでモデルを推計すればよい、という反論があるかもしれない。

ところが、仮に報告データにバイアスがなくランダムなノイズが含まれているという意味でのみ測定誤差がある場合でも、係数の推計値は一致性を持たず、データがいくら多くても、真の値には収束しない。いわゆる errors-in-variables bias である。

この会計実証分析において、しばしば忘れられている測定誤差がもたらすバイアスという論点を、（すべての変数を平均からの乖離とした）真のモデルが誤差 ε の *i.i.d.*正規分布[28]に従う、定数項のない k 変数モデル

$$y_i = \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i \qquad \varepsilon_i \sim N(0, \pi^2)$$

の場合で示そう。データはそれぞれ n 個得られたとする。

測定誤差がある場合、我々は説明変数[29]

$$x_i = (x_{1i}, x_{2i}, \ldots, x_{ki})'$$

そのものではなく、正規分布に従う（説明変数とは無相関の）誤差

$$\mu_i = (\mu_{1i}, \mu_{2i}, \ldots, \mu_{ki})' \qquad \mu_i \sim N(0, \Omega)$$

を含んだ

$$z_i = x_i + \mu_i$$

しかデータとして得ることができない。なお、モデルの誤差と測定誤差は無相関

$$E(\varepsilon_i \mu_i) = 0$$

とする[30]。したがって、係数ベクトルを

---

[28] 正規分布及び分散一定の仮定は緩めることができるものの、このような「理想的」状態でも望ましい結果が得られないことを示す。

[29] 非説明変数に測定誤差があってもなくても以下の議論には関係ない。したがって、式が不必要に煩雑になるのを避けるため、被説明変数に測定誤差はないと仮定する。

[30] ここでは正規分布を仮定しているので無相関ならば独立である。

$$\beta = (\beta_1, \beta_2, \ldots, \beta_k)'$$

と置けば、我々は真のモデルである

$$y_i = \beta' x_i + \varepsilon_i$$

ではなく、

$$y_i = \beta'(z_i - \mu_i) + \varepsilon_i = \beta' z + (\varepsilon_i - \beta' \mu_i)$$

を推計することになる。被説明変数ベクトルと（観察された）説明変数行列を

$$y = (y_1, y_2, \ldots, y_n)'$$

$$Z = (z_1, z_2, \ldots, z_k)$$

と置けば、最尤（最小二乗）推定値

$$b = (b_1, b_2, \ldots, b_k)'$$

は、

$$b = (Z'Z)^{-1} Z'y = \left(\frac{Z'Z}{n} \right)^{-1} \frac{Z'y}{n}$$

となる。（観察できない）真の説明変数 x の分散共分散行列を Σ と置けば、標本数が無限大になると、

$$\left(\frac{Z'Z}{n} \right)^{-1} \to (\Sigma + \Omega)^{-1}, \; Z'y \to \Sigma\beta$$

なので、測定データに基づく係数の推定値は

$$b = (Z'Z)^{-1} Z'y \to (\Sigma + \Omega)^{-1} \Sigma\beta$$

つまり、測定誤差がある限り、真の値である β に収束しない。もちろん、容易にわかるように、測定誤差がなければ、$\Omega = 0$ なので、

$$(\Sigma + \Omega)^{-1} \Sigma\beta = \Sigma^{-1} \Sigma\beta = \beta$$

となり、推定値 b は一致性を持つ。

測定誤差間に相関がある場合はもちろん、たとえ独立であっても推計の過程で「干渉」しあうので、特定の説明変数が上下どちらにバイアスを受けるかは一概にいえない。ただし、一変数の場合は、Σ と Ω が非負のスカラーなので、

$$b = (z'z)^{-1} z'y \to \frac{\Sigma}{\Sigma + \Omega} \beta \le \beta$$

となり、標本数が多くなると（絶対値が）真の値より小さい値に収束していく。つまり、測定ノイズの存在は、下方への一方向のバイアスをもたらすのである。

実は、説明変数が複数あっても、測定誤差間だけでなく（真の）説明変数間にも相関がない、

$$\Sigma=\begin{pmatrix}\sigma_1^2 & & O\\ & \ddots & \\ O & & \sigma_k^2\end{pmatrix},\ \Omega=\begin{pmatrix}\omega_1^2 & & O\\ & \ddots & \\ O & & \omega_k^2\end{pmatrix}$$

の場合（それぞれ対角行列となる）は、

$$(\Sigma+\Omega)^{-1}=\begin{pmatrix}\sigma_1^2+\omega_1^2 & & O\\ & \ddots & \\ O & & \sigma_k^2+\omega_k^2\end{pmatrix}^{-1}=\begin{pmatrix}\dfrac{1}{\sigma_1^2+\omega_1^2} & & O\\ & \ddots & \\ O & & \dfrac{1}{\sigma_k^2+\omega_k^2}\end{pmatrix}$$

なので、

$$b\to(\Sigma+\Omega)^{-1}\Sigma\beta=\begin{pmatrix}\dfrac{1}{\sigma_1^2+\omega_1^2} & & O\\ & \ddots & \\ O & & \dfrac{1}{\sigma_k^2+\omega_k^2}\end{pmatrix}\begin{pmatrix}\sigma_1^2 & & O\\ & \ddots & \\ O & & \sigma_k^2\end{pmatrix}\beta$$

$$=\begin{pmatrix}\dfrac{\sigma_1^2}{\sigma_1^2+\omega_1^2} & & O\\ & \ddots & \\ O & & \dfrac{\sigma_k^2}{\sigma_k^2+\omega_k^2}\end{pmatrix}\beta=\begin{pmatrix}\dfrac{\sigma_1^2}{\sigma_1^2+\omega_1^2}\beta_1\\ \vdots\\ \dfrac{\sigma_k^2}{\sigma_k^2+\omega_k^2}\beta_k\end{pmatrix}$$

となり、一変数の場合と同じく、すべての説明変数の推定値が下方にバイアスがかかったものとなる。説明変数が互いに無相関の場合には、重回帰の係数推定値はそれぞれ単回帰の値と同じになることを考えれば、当然の結果である。

説明変数が一つあるいはそれぞれが無相関の場合のこの上下「対等」でない結果は、次のように考えれば、より直観的に理解できるかもしれない。

測定ノイズの大きさを表す分散 ω_i^2 が説明変数の分散 σ_i^2 より相対的に大きければ大きいほど、観測された説明変数はランダムなノイズに近づく。したがって、その係数がゼロに近づくのは自明といってもよい。不正確な測定からは説明変数の「効果」を読み取ることはできない。

実証に基づいて会計情報の有用性を確かめるという昨今の会計研究は、会計測定の規範論ではデータの有用性を明らかにすることができないという「反省」がその出発点だったといえる。しかし、測定されたデータと株価の相関を検証するだけでは、ある概念としての（純利益や包括利益など）会計数値が有用な情報かどうかを決めることはできない。測定誤差によるバイアスの存在を考えると、係数の推定値がゼロという仮説が棄却できたに

せよ、あるいはできないにせよ、その「結論」は測定の不十分さを表しているだけかもしれないのである。

学部上級レベルの標準的教科書である Stock and Watson (2012, 364 頁)が指摘するように、測定誤差がもたらすバイアスを解決する最善策は正確な測定を行うことである。推計上の様々な技法は、理論上はともかく、実際のデータに適用するに際して役に立たないのが通例である[31]。測定の不十分さは「高度な」統計的手法によってバイパスすることができない、実証研究者が正面から改善に取り組むべき課題なのだ。

とはいえ、「正確な」測定ができない限り、実証研究者は沈黙していなければならないわけではない。誤差を含んではいても、測定値が研究の基盤をなす貴重な情報であることには変わりはない。会計研究者に求められているのは、誤差を含んでいることを前提に、測定値を用いてその背後にある実体間の関係を探ることである。それでは、測定値のもつ情報価値を最大限利用するにはどうすればよいのか。それが次節の課題である。

15. テスト理論と会計測定

すでに述べたように、昨今の実証会計研究においては、会計数値がどのように測定されたかをある意味「棚上げ」し、株価等との相関が高いことを「有用」と定義したうえで、会計数値の情報価値が議論されている。しかしながら、測定の対象となる現象と測定されたデータの乖離を意識することなしに会計「測定」という考え方は意味をなさない。

ここでは、会計測定など規範論という名前の「信仰告白」に過ぎないと考える実証研究者との水掛け論に陥らないよう、心理学や教育学の分野で確固たる地位を占めているテスト理論（test theory）[32]の考え方を参考に、我々会計学徒は会計測定及び実証分析を通じて何をしようとしているのかを明確にしたい。

我々が対象とする利益や資本などの会計概念は、身長や体重のように直接見て測ることはできない。その点、心理学や教育学が対象とする学力や知力などと同じく、自然科学における以上に、実体あるいは「構成概念」（construct）そのものとその測定の区別に注意が必要である。構成概念とは、「仮説的概念（hypothetical concept）であって、人間行動を説明するために理論を展開しようと努める社会科学者による知見に基づいた科学的想像（informed scientific imagination）の産物である。」[33]

しかし、自然科学とて概念と測定の乖離すなわち測定誤差とは無縁であり得ず、ただ程度の差があるだけである。したがって、「科学の全分野における主たる目的は構成概念レベルでの関係を計測（calibration）することである。こうした関係こそ理論の基礎単位を成す。自然科学であれ社会科学であれ、あらゆる構成概念の測定は誤差を含む。誤差なき測定な

[31] たとえば、操作変数法を用いるとして、教科書的条件を満たした操作変数はまず存在しない。そんなものがあれば、実証研究者は苦労しない。

[32] その概要については、たとえば、Crocker and Algina (2006)を参照。

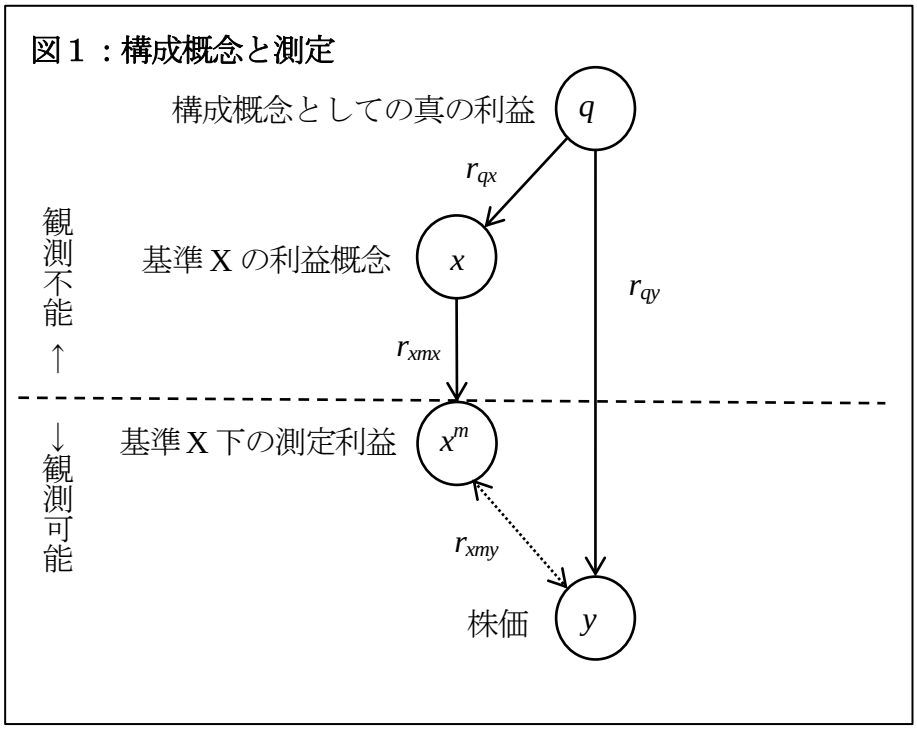
[33] Crocker and Algina (2006, 4 頁)。

どというものは存在しない。その結果、すべての観察された関係は特定の測定データ間の関係であって、構成概念そのものの関係ではなく、構成概念間の関係のバイアスをもった推計なのである。それゆえ、測定誤差という普遍的問題に対処し、測定誤差がもたらすバイアスを修正する研究が必要となる。」[34]

こうした問題意識のもと、パス解析（path analysis）[35]を利用した Hunter and Schmidt (2004, 2 章)の枠組みを参考に、実証会計研究の文脈で具体的に考えてみる。なお、ここでは測定誤差に焦点を合わせるため、サンプリング誤差の問題は捨象する。

まず、一定の理論に基づき、企業の価値と利益という構成概念が抽出される。ここでは議論を複雑にしすぎないため、観察される市場価格つまり株価 y が構成概念としての企業価値そのもの定義し、議論の対象を利益に絞る。

図１：構成概念と測定



抽象的な議論から導出された構成概念としての真の利益 q そのものは測定できないので、何らかの操作可能性を持った概念を作り出せねばならない。そこで、会計基準 X の下、測定可能な利益概念 x が構築されたとしよう。

しかしながら、操作可能な概念であっても、実際に測定するのは人間つまり企業の会計担当者や公認会計士であり、どのような理由で生じるかはともかく、測定誤差は避けられ

[34] Schmidt and Hunter (1999, 183-184 頁)。

[35] Wright (1920, 1921)。初等的解説として Kenny (1979)、測定誤差分析も含め社会科学研究におけるパス解析応用の嚆矢として Duncan (1966)を参照。

ない。したがって、観察可能なのは x そのものではなく、測定誤差を含んだ測定値 x^m となる。

実際の測定値 x^m と理論上の構成概念 q の間には二つの「関門」が存在する。まず、どれほど真の利益という構成概念を正確に捉えているかという、操作可能な会計基準 X の構成概念妥当性（construct validity）である。自然現象と異なり、社会現象の場合、理論から導出される理念型と操作可能概念には距離がある。そもそも、測定以前に何が真の利益かをめぐる議論は文字通り果てしなく続いている。とはいえ、包括利益あるいは純利益、また、FASB あるいは ASBJ の純利益のどちらが妥当かという問題設定自体は「科学的」であり得る。

そして、操作可能概念を前提として、どれだけ正確に測れているかどうかという測定の信頼性（reliability）の問題が存在する。社会現象のみならず自然現象においても、信頼性すなわち測定の正確性は重要なトピックである。実際、物理学などでは正確な測定に多大なコストがかけられ、研究分野としても確立している。

Hunter and Schmidt (2004, 44 頁)は、構成概念妥当性と信頼性を合わせて測定の操作品質（operational quality）と呼んでいる。結局、我々は操作品質に左右される測定値を観察しているのであって、構成概念である理念型としての真の利益はもちろん、捜査可能な特定の会計基準に基づく利益そのものすら観察できない。

その結果、実証研究においては、測定値と株価の相関を通常、利益の情報価値の基準としている。以上の関係を図示したのが図 1 である。

しかし、我々の関心が没理論的な観察されたデータ間の相関探しに終始するならばともかく、会計を含む社会現象のメカニズムの解明にあるのであれば、第 13 節で述べたように、測定対象に関する情報という観点から測定値を捉えることが求められる。

企業価値（株価）との関連でいえば、利益測定値と株価の相関 r_{xmy} は、真の利益と株価の相関 r_{qy} を推定するための情報であって、パス解析の公式[36]に従えば、真の利益と基準 X 利益の相関を r_{qx}、利益測定値と基準 X 利益の相関を r_{xmx} とすると、

$$r_{xmy} = r_{qy} \times r_{qx} \times r_{xmx}$$

すなわち、

$$r_{qy} = \frac{r_{xmy}}{r_{qx} \times r_{xmx}}$$

であることがわかる。

容易にわかるように、概念妥当性と信頼性[37]が満点（$r_{qx} = r_{xmx} = 1$）でないかぎり、

---

[36] 変数を平均 0、分散 1 に標準化すれば回帰係数は相関係数と一致する。誤差項が互いに無相関と仮定すれば、当節の公式は簡単に導出できる。

[37] テスト理論における信頼性の正確な定義は r_{xmx} ではなく r^2_{xmx} である。本節では前者を用いる。

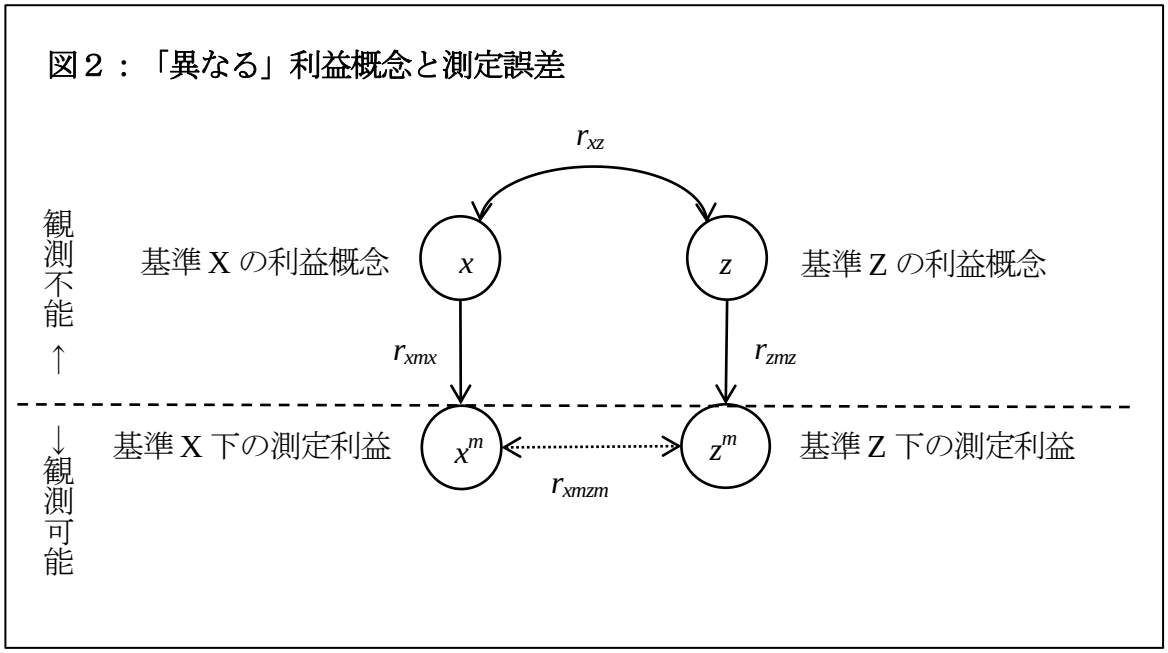
$r_{xmy} < r_{qy}$

であり、利益測定値と株価の相関 r_{xmy} が小さくても、それは必ずしも真の利益と株価の相関 r_{qy} つまり情報価値が小さいことを意味しない。真の利益と基準 X 利益の相関 r_{qx} あるいは利益測定値と基準 X 利益の相関 r_{xmx} が小さければ、「見かけ」の相関 r_{xmy} の小ささは、概念構成の拙さあるいは測定の不正確さ、それとも両方に起因する可能性がある。本質的には、前節で述べた測定誤差による回帰係数の下方バイアスの議論と同じことである。

実証研究には理論的な裏付けを持った正確な測定が不可欠であるという問題意識とともに、概念構成と測定が完全でないことを前提に、測定データに基づく相関から真の相関を推定するという視点が求められる。それに対し、多くの実証会計研究者は、特定のテストの成績という情報と所得の相関が低いことから学力は所得に関係しないという結論を出す教育「学者」のようなものだといったら、皮肉が過ぎるだろうか。

次に、概念妥当性を捨象し信頼性に焦点を当てて、操作可能な異なる利益概念の比較の問題も検討してみよう。たとえば、企業が二つの基準 X 及び Z の下での利益を両方とも報告するような状況を考えてみる。この場合、基準 X の利益概念 x に基づく測定値 x^m と、基準 Z の利益概念 z に基づく測定値 z^m が得られる。それぞれの間の相関とともに図示したのが図 2 である。

図２：「異なる」利益概念と測定誤差



通常、こうしたデータが得られると、基準 X と Z のどちらの利益概念が（株価との相関の大小という意味で）情報価値が大きいか否かが実証研究の論点となっている。しかし、測定誤差を考慮すれば、株価など他の観察値との相関を検討する前に、本当に基準 X と Z の利益概念は異なっているのかを検討する要がある。

再びパス解析の公式を用いると、観察された基準 X と Z に基づく測定値間の相関 r_{xmzm} は、両者の利益概念としての真の相関 r_{xz} を推計するための情報と理解することができ、それぞれの信頼性を r_{xmx} と r_{zmz} とすれば、パス解析の公式に従うと、

$$r_{xmzm} = r_{xz} \times r_{xmx} \times r_{zmz}$$

すなわち、

$$r_{xz} = \frac{r_{xmzm}}{r_{xmx} \times r_{zmz}}$$

となる。

利益概念間の相関 r_{xz} が 1、つまり両者は本質的に同じものを測ろうとしているにしても、測定誤差が大きければ、測定値間の相関 r_{xmzm} は小さな値となり得る。同じ基準の下、同じ企業の利益を測定する場合であっても、担当する公認会計士によって相当数値が異なることはよく知られた事実である。基準 X の下での利益測定値と基準 Z の下での利益測定値の違いをもたらしているのは、全部ではないにせよ、その一部は概念ではなく測定段階でのノイズである。会計研究者に求められるのは、どちらの利益概念が妥当かという議論以上に、正確な測定である可能性は小さくない。

一見「異なった」利益概念が実は本質的に同じ（真の相関が 1 に近い）であったとしても驚くに値しない。テスト理論を使った知力（intelligence）研究の成果の一つが、到達度（achievement）を測るとされるテストと素質（aptitude）を測るとされるテストの点数の相関から、測定誤差を考慮して推定した真の相関は極めて高く、両者は基本的に同じ何か、いわゆる *g* factor を測っているに過ぎないというものである[38]。

なお、本節ではサンプリング誤差の問題を捨象したけれども、現実に観察できるのは測定誤差を伴った標本であって母集団ではない。したがって、我々が知り（推計し）たい真（母集団）の相関と比べた観察された相関の下方へのバイアスはさらに大きくなる。また、利益同様、被説明変数である株価も背後に存在するファンダメンタル・バリューの測定値だと考えれば、やはり相関のズレは大きくなる。

Schmidt and Hunter (1996, 221-222 頁)は、心理学研究者の測定誤差無視あるいは軽視を憂慮しつつ、それでもこの点は、研究の進展と知識の累積を遅らせてきた方法論上の実践（methodological practice）としては最悪ではなく、二番目に悪いものだとしている。それでは最悪なのは？もちろん、データ解釈における統計的有意性（statistical significance）への依存である！同病相哀れむ、いや自覚が薄い分、会計研究者の病はより深いというべきか。

16. おわりに：測定と実証の間

会計数値という高度に発達した市場社会に生きる我々にとって極めて重要な情報を主たる対象とする我々会計研究者にとって、「情報とは、我々に何かを教えてくれるものであ

[38] *g* factor については、Jensen (1998)を参照。

り、それによって不確実な知識を、より確実なものにするものである」（吉田 2009）という視点に立つ情報理論の知見を無視することはできない。しかも、情報理論は会計研究における測定と実証の不毛の二項対立を止揚し、両者を統一的に考える道を指し示している。

実証が会計研究に不可欠の要素であることは認めるにしても、経済学者やファイナンス研究者に対する会計研究者の比較優位を考えると、理論的裏付けを持った正確な会計測定こそ、我々の主たる任務であるといっても過言ではない。会計測定から実証への「青い鳥」探しの旅を続けるのが本当によいのかどうか、我々は今一度立ち止まって考える局面に差し掛かっているのではなかろうか。

補論 1：バイアスの導出[39]

バイアス B を

$$B = E_{p(x)}[l(\hat{\theta}) - l(\theta_0)] + E_{p(x)}\left[l(\theta_0) - nE_p[l_i(\theta_0)]\right] + E_{p(x)}\left[nE_p[l_i(\theta_0)] - nE_p[l_i(\hat{\theta})]\right]$$
$$= B_1 + B_2 + B_3$$

と三つの部分にわけて導出する。

まず、

$$B_1 = E_{p(x)}[l(\hat{\theta}) - l(\theta_0)]$$

については、最尤推定値のまわりでテイラー展開を行い、

$$l(\theta) = l(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})' \left.\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta'}\right|_{\hat{\theta}} (\theta - \hat{\theta})$$

であることを確認する。

$$J(\theta_0) = -E_p\left[\left.\frac{\partial^2 l_i(\theta)}{\partial\theta\partial\theta'}\right|_{\theta_0}\right]$$

と置くと、上記のテイラー展開右辺の行列が

$$\frac{1}{n}\frac{\partial^2 l(\hat{\theta})}{\partial\theta\partial\theta'} \to -J(\theta_0)$$

に収束することを用いて、

$$l(\theta_0) - l(\hat{\theta}) = -\frac{n}{2}(\theta_0 - \hat{\theta})' J(\theta_0)(\theta_0 - \hat{\theta})$$

を得る。したがって、

$$B_1 = E_{p(x)}[l(\hat{\theta}) - l(\theta_0)]$$
$$= \frac{1}{2}E_{p(x)}\left[(\theta_0 - \hat{\theta})' J(\theta_0)(\theta_0 - \hat{\theta})\right]$$

となり、カッコの中はスカラーなのでトレースをとっても同じであり、$tr(AB) = tr(BA)$なので、

$$B_1 = \frac{n}{2}E_{p(x)}tr\left[(\hat{\theta} - \theta_0)' J(\theta_0)(\hat{\theta} - \theta_0)\right]$$
$$= \frac{n}{2}tr\left[J(\theta_0)E_{p(x)}\left[(\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)'\right]\right]$$

となることがわかる。さらに、

---

[39] 小西・北川 (2004, 50-53 頁)に基づく。

$$I(\theta_0) = E_p\left[\frac{\partial l_i(\theta)}{\partial\theta}\frac{\partial l_i(\theta)}{\partial\theta'}\bigg|_{\theta_0}\right]$$

と置くと、最尤推定量の漸近共分散行列は

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right)\to N\left(0, J^{-1}(\theta_0)I(\theta_0)J^{-1}(\theta_0)\right)$$

なので[40]、

$$\begin{aligned}B_1 &= \frac{n}{2}tr\left[J(\theta_0)E_{p(x)}\left[(\hat{\theta}-\theta_0)(\hat{\theta}-\theta_0)'\right]\right]\\ &= \frac{1}{2}tr\left[J(\theta_0)J^{-1}(\theta_0)I(\theta_0)J^{-1}(\theta_0)\right]\\ &= \frac{1}{2}tr\left[I(\theta_0)J^{-1}(\theta_0)\right]\end{aligned}$$

を得る。

次に、

$$B_2 = E_{p(x)}\left[l(\theta_0)-nE_p[l_i(\theta_0)]\right]$$

がゼロとなるのは自明であろう。

最後に、

$$B_3 = E_{p(x)}\left[nE_p[l_i(\theta_0)]-nE_p[l_i(\hat{\theta})]\right]$$

を計算する。モデルの真のパラメータ値 θ_0 は

$$E_p\left[\frac{\partial l_i(\theta)}{\partial\theta}\right]=0$$

の解、すなわち、

$$\frac{\partial E_p\left[l_i(\theta)\right]}{\partial\theta}\bigg|_{\theta_0}=0$$

なので、テイラー展開によって、

$$E_p[l_i(\hat{\theta})] = E_p\left[l_i(\theta_0)\right]+\frac{1}{2}(\hat{\theta}-\theta_0)'E_p\left[\frac{\partial^2 l_i(\theta)}{\partial\theta\partial\theta'}\bigg|_{\theta_0}\right](\hat{\theta}-\theta_0)$$

を得る。したがって、

[40] 補論 2 参照。

$$B_3 = E_{p(x)}\left[nE_p[l_i(\theta_0)] - nE_p[l_i(\hat{\theta})] \right]$$
$$= \frac{n}{2} E_{p(x)}\left[(\hat{\theta} - \theta_0)' J(\theta_0)(\hat{\theta} - \theta_0) \right]$$

すなわち、

$$B_3 = B_1 = \frac{1}{2} tr\left[I(\theta_0) J^{-1}(\theta_0) \right]$$

であることがわかる。

結局、バイアスは三つの部分を加えて、

$$B = B_1 + B_2 + B_3$$
$$= \frac{1}{2} tr\left[I(\theta_0) J^{-1}(\theta_0) \right] + 0 + \frac{1}{2} tr\left[I(\theta_0) J^{-1}(\theta_0) \right]$$
$$= tr\left[I(\theta_0) J^{-1}(\theta_0) \right]$$

となる。

補論 2：共分散行列の導出[41]

最大対数尤度の一階微分を θ_0 のまわりでテイラー展開すると、

$$\frac{\partial l(\hat{\theta})}{\partial \theta} = \frac{\partial l(\theta_0)}{\partial \theta} + \frac{\partial^2 l(\theta_0)}{\partial \theta \partial \theta'}(\hat{\theta} - \theta_0) = 0$$

すなわち、

$$-\frac{\partial^2 l(\theta_0)}{\partial \theta \partial \theta'}(\hat{\theta} - \theta_0) = \frac{\partial l(\theta_0)}{\partial \theta}$$

なので、大数の法則より、

$$-\frac{1}{n}\frac{\partial^2 l(\theta_0)}{\partial \theta \partial \theta'} = -\frac{1}{n}\sum \left.\frac{\partial^2 l_i(\theta)}{\partial \theta \partial \theta'}\right|_{\theta_0} \to J(\theta)$$

に収束することがわかる。さらに、

$$E_p\left[\left.\frac{\partial l_i(\theta)}{\partial \theta}\right|_{\theta_0}\right] = 0,\ E_p\left[\left.\frac{\partial l_i(\theta)}{\partial \theta}\right|_{\theta_0}\left.\frac{\partial l_i(\theta)}{\partial \theta'}\right|_{\theta_0}\right] = I(\theta)$$

より、

$$\sqrt{n}\frac{1}{n}\frac{\partial l(\theta_0)}{\partial \theta} = \sqrt{n}\frac{1}{n}\sum \left.\frac{\partial l_i(\theta)}{\partial \theta'}\right|_{\theta_0} \to N\left(0, I(\theta_0)\right)$$

なので、

$$\begin{aligned}\sqrt{n}\frac{1}{n}\frac{\partial l(\theta_0)}{\partial \theta} &= -\sqrt{n}\frac{1}{n}\frac{\partial^2 l(\theta_0)}{\partial \theta \partial \theta'}(\hat{\theta} - \theta_0) = \sqrt{n}\frac{1}{n}\frac{\partial^2 l(\theta_0)}{\partial \theta \partial \theta'}(\hat{\theta} - \theta_0) \\ &= \sqrt{n}J(\theta_0)(\hat{\theta} - \theta_0) \to N\left(0, I(\theta_0)\right)\end{aligned}$$

すなわち、

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \to N\left(0, J^{-1}(\theta_0)I(\theta_0)J^{-1}(\theta_0)\right)$$

を得る。

[41]小西・北川 (2004, 42-45 頁)に基づく。

補論 3：対数尤度比検定

誤差が *i.i.d.*の正規分布に従うと仮定し、パラメータ数 m_0 の真の（線形）モデル M_0 とそれをネストするパラメータ数 m_A のモデル M_A の対数尤度の差の二倍 LR は

$$
\begin{aligned}
LR &= 2\left[l(\theta_A) - l(\theta_0)\right] \\
&= 2\times\left[-\frac{1}{2}\left(n\log 2\pi + n\log s_A^2 + n\right) - \left\{-\frac{1}{2}\left(n\log 2\pi + n\log s_0^2 + 1\right)n\right\}\right] \\
&= n(\log s_0^2 - \log s_A^2) = n\log\frac{s_0^2}{s_A^2} = n\log\left(1 + \frac{s_0^2 - s_A^2}{s_A^2}\right) \\
&\simeq n\cdot\frac{s_0^2 - s_A^2}{s_A^2} = n\cdot\frac{\sum e_{0i}^2 - \sum e_{Ai}^2}{\sum e_{Ai}^2} = \frac{\sum e_{0i}^2 - \sum e_{Ai}^2}{s_A^2}
\end{aligned}
$$

と変形できる。

$$
n\cdot\frac{s_0^2 - s_A^2}{s_A^2} = n\cdot\frac{\sum e_{0i}^2 - \sum e_{Ai}^2}{\sum e_{Ai}^2} = \frac{\sum e_{0i}^2 - \sum e_{Ai}^2}{s_A^2}
$$

は、M_0 を帰無仮説とし、M_A のパラメータのうち M_0 に含まれない $m_A - m_0$ 個のパラメータをゼロとおいた、漸近理論に基づく仮説検定で用いる自由度 $m_A - m_0$ のカイ二乗検定量[42]なので、

$$
LR = 2\left[l(\theta_A) - l(\theta_0)\right] \to \chi^2(m_A - m_0)
$$

を得る。

[42] さらに自由度 $m_A - m_0$ で割ったものが F 検定量である。

補論４：ベイズ定理

事象と情報がそれぞれ、{*H*, *T*}、{I_H, I_T}の二通りの場合、事象を条件とする条件付確率の定義

$$p(I_H|H)=\frac{p(H\cap I_H)}{p(H)},\ \ p(I_T|H)=\frac{p(H\cap I_T)}{p(H)}$$
$$p(I_H|T)=\frac{p(T\cap I_H)}{p(T)},\ \ p(I_T|T)=\frac{p(T\cap I_T)}{p(T)}$$

より、同時確率は

$$p(H\cap I_H)=p(I_H|H)\cdot p(H),\ \ p(H\cap I_T)=p(I_T|H)\cdot p(H)$$
$$p(T\cap I_H)=p(I_H|T)\cdot p(T),\ \ p(T\cap I_T)=p(I_T|T)\cdot p(T)$$

となる。事象の確率 p(H)と p(T)が与えられていれば、ベイズ定理に従い、

$$p(H|I_H)=\frac{p(H\cap I_H)}{p(I_H)}=\frac{p(H\cap I_H)}{p(H\cap I_H)+p(T\cap I_H)}$$
$$p(T|I_H)=\frac{p(T\cap I_H)}{p(I_H)}=\frac{p(T\cap I_H)}{p(H\cap I_H)+p(T\cap I_H)}$$
$$p(H|I_T)=\frac{p(H\cap I_T)}{p(I_T)}=\frac{p(H\cap I_T)}{p(H\cap I_T)+p(T\cap I_T)}$$
$$p(T|I_T)=\frac{p(T\cap I_T)}{p(I_T)}=\frac{p(T\cap I_T)}{p(H\cap I_T)+p(T\cap I_T)}$$

を得る。したがって、相互情報量は

$$S(y|x)=-[p(H\cap I_H)\log p(H|I_H)+p(T\cap I_H)\log p(T|I_H)$$
$$+p(H\cap I_T)\log p(H|I_T)+p(T\cap I_T)\log p(T|I_T)]$$

であることがわかる。

参照文献

赤池弘次（1976）「情報量規準とは何か」『数理科学』14 巻 3 号 5-11 頁。

甘利俊一（2011[1970]）『情報理論』筑摩書房。

小西貞則・北川源四郎（2004）『情報量規準』朝倉書店。

下平英寿（1999）「モデル選択理論の新展開」『統計数理』47 巻 1 号 3-27 頁。

下平英寿（2004）『モデル選択：予測・検定・推定の交差点』岩波書店。

下平英寿（2007）「モデル選択とブートストラップ」室田一雄他編『赤池情報量規準 AIC：モデリング・予測・知識発見』共立出版所収。

竹内啓（1976）「情報統計量の分布とモデルの適切さの規準」『数理科学』14 巻 3 号 12-18 頁。

豊田正（1997）『情報の物理学』講談社。

吉田裕亮（2009）『情報理論入門：基礎から確率モデルまで』サイエンス社。

Akaike, H. 1973. Information Theory and an Extension of the Maximum Likelihood Principle. In B. N. Petrov and F. Caski (eds.), *Proceeding of the Second International Symposium on Information*. Budapest, Hungary: Akademiai Kiado.

Choi, H., and N. M. Kiefer. 2011. Geometry of the Log-Likelihood Ratio Statistic in Misspecified Models. *Journal of Statistical Planning and Inference* 141 (6): 2091-2099.

Crocker, L., and J. Algina. 2006. *Introduction to Classical and Modern Test Theory*. Mason, U.S.A.: Cengage Learning

Duncan, O. D. 1966. Path Analysis: Sociological Examples. *American Journal of Sociology* 72 (1): 1-16.

Jensen, A. R. 1998. *The g Factor: The Science of Mental Ability*. Westport, U.S.A.: Praeger.

Kenny, D. A. 1979. *Correlation and Causality*. , New York, U.S.A.: Wiley.

Kishino, H., and M. Hasegawa. 1989. Evaluation of the Maximum Likelihood Estimate of the Evolutionary Tree Topologies from DNA Sequence Data, and the Branching Order in Hominoidea. *Journal of Molecular Evolution* 29 (2): 170-179.

Kullback, S., and R. A. Leibler. On Information and Sufficiency. *Annals of Mathematical Statistics* 22 (1): 79-86.

Linhart, H. 1988. A Test Whether Two AIC's Differ Significantly. *South African Statistical Journal* 22 (2): 153-161.

Mittelhammer, R. C., G. G. Judge and D. J. Miller. 2000. *Econometric Foundations*. Cambridge, U.K.: Cambridge University Press.

Schmidt, F. L., and J. E. Hunter. 1996. Measurement Error in Psychological Research: Lessons from 26 Research Scenarios. *Psychological Methods* 1 (2): 199-223.

Schmidt, F. L., and J. E. Hunter. 1999. Theory Testing and Measurement Error. *Intelligence* 27 (3):

183-198.

Schmidt, F. L., and J. E. Hunter. 2004. *Methods of Meta-Analysis*, Second Edition. Thousand Oaks, U.S.A.: Sage.

Shimodaira, H. 1997. Assessing the Error Probability of the Model Selection Test. *Annals of the Institute of Statistical Mathematics* 49 (3): 395-410.

Stock, J. H., and M. W. Watson. 2012. *Introduction to Econometrics*, Third International Edition. Harlow, U.K.: Pearson.

Vuong, Q. H. 1989. Likelihood Ratio Tests for Model Selection and Non-Nested Hypothesis. *Econometrica* 57 (2): 307-333.

Wright, S. 1920. The Relative Importance of Heredity and Environment in Determining the Piebald Pattern of Guinea-Pigs. *Proceedings of the National Academy of Sciences* 6 (6): 320-332.

Wright, S. 1921. Correlation and Causation. *Journal of Agricultural Research* 20 (7): 557-585.