金融・資本市場とデータサイエンス

一橋大学大学院経営管理研究科准教授　横内大介

はじめに

新型コロナウイルスの世界的流行により、国内でも連日PCR検査の陽性者数の報道がなされている。PCR検査数が正確にわからなければ日々の陽性者数の議論は意味がないなどという、いわゆるデータの活用に係る問題点までもがテレビで議論されている様子をみると、一般社会におけるデータ・リテラシーの重要性は日増しに高くなっている印象を受ける。ましてや社会の血液である金融・資本市場の参加者にとってはより重要な知識であることは言うまでもないだろう。しかしながら、周りを見回すとデータに関して多くの誤解が生じていることもまた事実である。多くの企業でビッグデータがもてはやされているが、実は、データはむやみやたらに集めてよいものではなく、また、同じテーマでもデータ数が多い分析ほど価値があるというものでもない。ましてや誤解に基づいて使われた統計手法で得られた知見は我々に何ももたらさない。本稿では、データサイエンスやビッグデータ分析という名のもとに、金融市場の分析で常用されている統計手法や機械学習法の問題点について議論する。

データは現象を映す鏡

この原稿を執筆している2020年10月上旬においても、コロナはまだまだ収束する気配を見せない。ウィズコロナというスローガンとともに社会生活の在り方も大きく変わってしまった。密になることを避けつつ生活するということが当たり前になり、人間の価値観や活動様式は大きく変化した。

当然のことながら、株式市場における企業価値評価についても大きな変化が生じている。例えば同じ運輸業でも、人を運ぶ企業は苦戦する一方で、物を運ぶ企業は巣ごもり需要で大きな成長が見込まれる優良企業へと、それぞれの評価が一変したことは、市場の関係者であれば周知の事実であろう。この種のレジームの変化のたびに起こる企業評価の入れ替わりは、規模の大小の違いはあるにせよ、株式市場ではそれほど珍しいことではないだろう。それは「過去は過去、今は今」という誰もが当たり前だと思っていることを表しているに過ぎないのだから。しかしながら、いざデータを活用するとなった際には、この当たり前の原則を忘れてしまう人は少なくない。

小職の所属するプログラムの大学院生は基本的に社会人であり、金融機関に所属している学生も少なくない。彼らに実データを用いた資本資産価格モデル（CAPM）やファーマ-フレンチ3ファクター・モデル（FF3）のような線形モデルの係数推定の課題を出すと、できる限り長い期間のデータを使いたがる。彼らになぜそのようなデータの選びかたをするのかを尋ねると、決まって持ち出してくるのが大数の法則である。

大数の法則は、いわゆる実験条件が変化しないという仮定の下でデータをたくさん集めれば、それだけよい推定値が得られるということを示している。正確なさいころをたくさん振ってそのデータを記録していけば、おのおのの目が出る割合はそれぞれ6分の1に近づくという話はまさに大数の法則である。

読者の皆様はすでにお気づきかと思うが、社会情勢の大きな変化を考慮せずに長い期間のデータを取ってくるということは、大数の法則が成立するための重要な仮定である「実験条件が変化しなければ」という条件を無視してデータ分析していることにつながる。学生からは、「サンプル数が少ないと統計的な精度が落ちるから仕方がないではないか」という言い訳を聞くこともしばしばあるが、統計学のような数理科学において、仮定は最も重視すべき条件であり、それを満たさずに行うデータ分析ほど信頼のおけないものはない。

社会科学における統計モデルの濫用

仮定を満たさないという意味では、実務にお

ける資産価格モデルの使い方についても問題は多い。CAPMやFF3など多くの資産価格モデルでは、誤差項が独立同一な正規分布に従うことを仮定しており、最小二乗法で係数を推定した際に得られる残差は、この誤差項の推定値とみなすことができる。この残差自体が互いに独立で同一の正規分布に従っていなければ、資産価格モデルの大前提は崩れることになり、それ以上の議論を進めることはできない。しかしながら、実務のレポートなどで見かけるデータ分析で、残差の独立性や正規性を調べている例を見かけたことはほとんどない。仮定が成立しているか不明な状態では、係数の推定値のp値でいくら有意性を議論したとしても、その信頼性は低いと言わざるを得ない。

もう一つ実例を挙げよう。個別株式の日次の収益率データのヒストグラムを描いたことのある方ならわかると思うが、日次収益率の分布はいわゆる正規分布よりも急尖的になることが多い。もしなだらかだったとしても、値幅制限の存在を考えれば個別株式の日次収益率の分布の両裾は切断されていることは明らかであり正規分布はしていない。多くの実証分析では日次収益率のデータに対して正規性を仮定した統計手法、簡単な例でいえばt検定などを使っているケースがしばしば見受けられるが、厳密には正しい使い方とは言えないだろう。使える手法がないという意見もあるが、たとえば t検定の代わりならばウィルコクソン検定[1]のようなノンパラメトリック検定もあるので、やはり実データが示す性質に合わせて適切な手法を用いるべきである。

このような乱暴なデータ分析が氾濫した原因の一つは、社会科学系の学術論文自体のデータ分析でも全く同じことが行われているからに他ならない。分野を問わず、データを正しく分析し解釈することが要求される現代において、大学や研究機関に籍を置いている小職たち研究者の責任は決して小さいとは言えない。

データサイエンスの誤解

データサイエンスは、データの取得から解析、モデル化までをトータルに科学する学問であり、データが発生したメカニズムを、テューキー（Tukey）の提唱した探索的データ解析を通じて解明することが主な目的である。たとえば株式データの分析であれば、「なぜ」その価格（収益率）が発生したのかを調べることが一つの目的となる。一方、巷で流行している機械学習（ディープラーニングを含む）は「なぜ」を解明することに主眼はなく、ブラックボックス化を許容して予測精度の向上を目指す仕掛けである。先の例であれば、条件を入力することで精度の高い株価予測を与えるAIを作ることが機械学習の目的の１つになる。これは明らかに「なぜ」を解明するサイエンスではなく、「実用に供することを最優先としてシステムを作る」というエンジニアリングの考え方そのものである。機械学習でAI開発やビッグデータ分析を行う人をデータサイエンティストと呼称することがあるが、厳密には誤りであり、正しくはデータエンジニアと呼ぶべきであろう。

最近は金融業界でも、機械学習が席巻しており、たとえば、機械学習で作られたAIが運用を担う商品なども続々と開発された。しかしながら、「なぜ」がわからないAIを使う場合には十分な注意を払う必要があることを忘れてはならない。2019年6月に大阪で開かれたG20で採択された「人間中心のAI社会原則」では、AIサービスを提供する企業がAIの動作を保証することが明記されている。つまり、機械学習で作ったAIが運用を失敗し、顧客から訴訟を起こされたとしたら、運用側は圧倒的に不利になることは言うまでもない。

もちろん、機械学習は便利で優秀な道具であり、たとえば将棋などでは一流のプロ棋士でも考えつかない手を提案することがある。それは機械学習のアルゴリズムの下での最善手を探すためであって、それが最善手である理由を説明することは目的外である。言い換えれば、AIが指した棋譜に対してプロ棋士が解釈を与えて新たな戦術へ昇華することがAI導入の目的となっていると言える。一方で、金融機関は顧客のお金を扱うという立場上、総じて説明責任が強く求められる。将棋の例のようにAIの出す結果を解釈すれば済むという話ではない。

今、流行している機械学習も決して万能な道具ではなく、そして作られたAIは神様ではない。これらが持つ弱点の克服には、本来の意味の「なぜ」を明らかにするデータサイエンスの積極的な活用がカギを握るだろう。

---

[1] 一対の標本によるノンパラメトリック検定法。対応のあるt検定で必要とされる正規性の仮定が満たされない場合に用いる。