仮説成長型データ解析の教育と S-PLUS　ー社会人大学院での実践ー

筑波大学大学院ビジネス科学研究科
経営システム科学専攻
椿　広計

社会人データサイエンス教育は、知識を実質有るものとして固定しようという強いモチベーションのために、教育導入期における簡便なソフトウェア、合理的かつ有用なデータ解析戦略、実務と直結した解析体験を与えれば、容易に自己研鑽へ誘導することが可能であり、一般大学教育よりは明らかに効率が高い。しかし、一方で職業統計家レベルに近い社会人層の知的欲求を満たすような教育体系を与えるためには、担当教員が常に現時点で最高水準のデータ解析戦略体系とは何かを問い続ける必要がある。これらの特徴を明らかにするために、筆者の個人的意見や体験に過ぎないが、筑波大学東京地区夜間社会人大学院における Splus などを用いたデータサイエンス教育の理念、実態などを紹介する。本稿は、椿(1999)をベースにしているが、その後の進展も若干含まれている。

１．社会人データ・サイエンス教育への挑戦

本小論の目的は、社会人に対するデータ・サイエインス教育について、筆者の経験を提供し、今後の議論を喚起することである。筆者が、高度専門社会人教育に責任を感じるようになったのは、平成９年４月に筑波大学東京地区夜間社会人大学院に移動し、その統計教育を任されるようになってからである。

社会人教育と通常の大学教育は、何が同じで何が違うのだろうか？先ず、社会人大学院教育の特殊性は、学生に統計実務家がいることである。実際、筆者の講義の聴講生には日常的に実務で統計解析を行っている院生が２割程度存在する。しかも、その一部は社内のデータ解析指導層とも言える人達である。実務で日常使う手法の講義は彼らが行った方が、余程上手く行うと考えられる。従って、講義をマネジメントする上で必要なのは、一流統計家としての彼らが納得する知識体系を提供し、彼らの理解を梃子にして、彼らを助手として一般的水準の院生に最新のデータ・サイエンス技術の有効性を認識してもらうことである。

もう一つの困難は、古典的統計手法の黄昏を彼らが認知していることである。実際、人工知能技術に基づく推論技術の進展は、既存統計的モデリング技術を過去のものとしかねない勢いである。そもそも、この種の方法論とその優位性も専攻の中で別途講義されており、受講生が認識していることを前提にしなくてはならない。

以上の状況の下、理屈の難易度や手法の簡易性、手法の普及度は一切問わず、筆者が現在考案し得る最良のデータ解析戦略の幾つかを学生にぶつけることが重要であると考えた。着任後のデータ・サイエンス系の科目「多変量解析」でのオリエンテーションで、「聴講されている皆さんが、私の戦略の有効性を実証して下さることを期待する」という態度で臨んだのである。データ・サイエンスには、個別的の方法論の学理を追求する面白さもあることは、勿論である。しかし、高度社会人教育においては、その実践的側面、データから情報を獲得する総合技術の講義担当者による体系化を提示し、受講者集団に存在する体系と対決させることが、理論的説得以上に必要だというのが、筆者の信念である。

２．社会人大学院でのデータサイエンス教育の実際

ここでは、筑波大学社会人大学院でどのようなデータ・サイエンス教育を行っているかについて具体的に紹介する。

２．１　筑波大学東京地区夜間社会人大学院の概要

筆者の勤務する筑波大学大学院ビジネス科学研究科企業科学専攻（博士）、経営システム科学専攻（修士）は、高度職業人専門教育を使命とした社会人のみを対象とした夜間大学院（講義時間帯：火曜日ー金曜日 18 時 20 分-21 時、土曜日 13 時 45 分ー21 時）である。修士課程は平成元年に、博士課程は平成 8 年に東京都文京区大塚（茗荷谷）の旧東京教育大学跡地に設置された。学生は、東京都心に勤務する社会人を中心に、メーカー、金融サービス産業、民間研究機関、公務員、大学研究者などが万遍なく集まっている。所属企業の承認を入学資格要件として課していないので、企業派遣院生は、少数派であり、勤務先が夜間大学院に通学していることを認知していない学生が３割近くいるとされている。現在、社会人夜間大学院への志願者は非常に多く、例えば修士課程倍率は、5 倍から 10 倍を推移している。

スタッフは、平成 13 年 11 月現在 20 名であり、その内訳は、経営理論分野（組織論、戦略論、会計学、マーケティング、人材開発）7 名、ネットワーク・情報分野（ソフトウェア工学、人工知能など）6 名、数

理(Quantitative Analysis)分野（品質マネジメント、オペレーションズ・リサーチ、ゲーム理論、応用確率論、統計学）７名となっている。筆者は、量的分析分野に属している。

スタッフは、原則として上記 3 分野から１名ずつ合計３名で１名の大学院生（修士課程定員１学年 30 名、博士課程１学年 11 名）を指導する。特に経営分野の研究指導方針として、データに基づく仮説検証の重視ということがあり、数理分野の応用統計担当教員は、経営関連修士・博士論文のデータアプローチ（調査計画、データ分析）支援・指導という役割を果たす必要があり、そのための個別指導を行っている。しかし、個別指導では不効率な面も多いので、データ・サイエンスに関しては、社会人の問題解決能力向上に必要な高度教養科目としての位置づけで、かなり積極的なカリキュラム体系を構築している。

２．２　データ・サイエンス科目

筑波大学大学院の講義は、１回１時間 15 分で 10 回で１単位という構成になっている。筑波大学は、３学期制であるが、学生側の強い要望もあり、夏休み、秋休み、春休みに集中講義や単位とは関係ない準備コース（補修）を配している。経営システム科学専攻でデータサイエンス関連講義として「データ解析」、「多変量解析」、「社会調査法」、「調査計画」、「統計モデル」、「統計的管理」、「計量経済学」が配置されている。この中で「多変量解析」と「統計的管理」の前半では、Splus を用いた実践的教育を行っている。以下では、この中で「多変量解析」について、その狙いなどを紹介する。

修士１年の２学期は、専門科目として「多変量解析」が配置されており、平成 13 年度の担当は筆者である。テーマは「統計的予測」のための戦略であり、回帰分析の発展形態を習得する。ある目的変数を予測するために一般線形モデル（重回帰モデル、対数線形モデル、線形ロジスティックモデルなど）を大規模データベースを用いて当てはめるには、そのために、交互作用や非線型性といった障害要素を十分考慮しなくてはならない。筆者が「仮説成長型データ解析」と呼んでいる手続きを Splus を用いて実際に経験するのが、講義の目的である。

「多変量解析」シラバス
１時間目：オリエンテーション：データからの知識獲得
２時間目：演習データの読み込み及び観察
３時限目：樹形モデル(tree model)当てはめ：データから論理構造を発見する
４時限目：樹形モデル実習
５時限目：一般加法モデル(Generalized Additive Model)当てはめ：論理構造から非線形要因パターン効果（ノンパラメトリック回帰モデル）に知識を進化させる
６時限目：一般加法モデル実習
７時限目：一般線形モデル(Generalized Linear Model):パターンの数式化
8 時限目：一般線形モデル実習
9,10 時限目：班別成果発表会

講義計画のスケジュール面での特徴は、講義の大半を実習や報告会が占めていることである。実習のために、１班４名程度で８班を組織するとともに、上場企業（約 2000 社）の財務データ（約 100 変量）を年度を変えて各班に配布している。班構成に当たっては、計算機やデータ解析に強い人間が、各班に一人は入るように指示している。

実習では、各班は予測の目的変数を選んで（経常利益率や、最近の安全性に関する「格付け」、従業員の平均給与など）、それがどのような変数でどのように説明されるかを４週間かけて分析し、９，１０時限目に１５分ずつ発表し、５分ほどの討論を行う。実習は、通常の講義時間の枠組みを超えて深夜まで演習している班もある。また、発表会直前の週末には徹夜で分析を行っている班も複数生じ、発表会が終わると打ち上げ会が企画される位である。同一班内に財務分析専門家とデータ解析専門家とが存在し、その両者が協力ではなく、競争し２つの予測モデルを別に作り発表会に臨むことなども社会人大学院ならではのことである。「多変量解析」の講義が入門的講義である「データ解析」に比して不利な点として、「現時点で、最高の方法論とデータ解析戦略を提供する」という筆者の理想のために、解析に用いる Splus の UNIX 版が、学生が自宅で簡単に使えるものでなくなってしまったことにある。本年度からは、学生に R も配布して自身の PC でも演習できるようにしたが、樹形モデルや一般化加法モデルでは若干互換性がないことが気になる。

この種の困難にも関わらず、筆者がこのスタイルの講義を継続しているのは、

１）教育した手法が修士論文で活用されることが多くなっている。
２）樹形モデルなど実務（信用分析）でも活用したが分からないところがあるといった質問を受ける
３）理工系出身の院生から、手法の理論的側面を知りたいので、輪講科目を開いてくれと要求される。

といった所にある。
この種の動機付け成功の原因は、次のようなことだと考えている：

1）配布している財務データが、社会人には、親密感があり実務的興味を刺激する性質であり、最終回の発表会が大変盛り上がって良い雰囲気である。

2）配布データの持つ情報が膨大であるために、アドバンスな手法を使って初めて分かるという事が頻発する。

3）班別演習の中で、リーダー格の院生のマネジメント下で、様々なディスカッションがなされる。

具体的に、どのような講義が行われているかを想像していただくために、本年度の第 1 回の講義（オリエンテーション）のパワーポイントファイルの一部を付録に示す。

3．終わりに：どれ位ニーズに応えられているのか？

データサイエンス系科目を積極的に配備してはいるものの、専任スタッフの講義だけで社会人が必要な統計的方法の全てを網羅することは全く困難である。特に、金融・経済データの分析を必要とする院生や、マーケティングを業務としている院生への必要な知識を提供しているとは言えない。例えば、次のような方法は、一部の院生の学位論文作成にとって欠くことのできないものである。

1）時系列データの分析：とくに多変量非定常時系列データの解析（因果性分析、共和分関係の抽出など）と GARCH など、分散(Volatility)変動を記述するモデル。

2）質的選択のモデル：多項ロジットモデルや累積ロジットモデルなど多項分布に対するモデリング

3）生存時間モデル：最近は、人間や部品の寿命ばかりではなく、会社の寿命が問題となっている。

これらを補う講義として、「計量経済学」、「経済学演習」、「数理ファイナンス」といった非常勤の先生方にお願いしている科目の中で、可能な限りテーマとして取り上げていただくことをお願いしている。

また、特に経済時系列解析、計量経済学に関しては、自主ゼミで補うようにしている。実際、筆者の主催する輪講は実質的に「計量経済学」と「経済時系列」であり、研究室横断的なメンバーで運営されている。

一般的な統計家の観点からカリキュラム体系で欠けていると指摘されるであろうのは、「記述多変量データ解析」、すなわち、正準分析、数量化Ⅲ類、多次元尺度構成、クラスター分析といった項目である。本来は、調査データの分析手法として国内では広く活用されている方法なので、何とかしなくてはならないと考えている。記述的方法は、データを利用するという立場よりは、単にデータを虚心にかつ上手く眺めるという方法で、筆者にとっては個人的に動機付けが難しいと悩んでいる方法論である。記述多変量データ解析が、データ解析の初動段階で必要な方法論であることは疑いないが、これを行った事のメリット、すなわち情報予測ではなく「要約」のメリットを具体的に表現できる能力が教員側に欠如すると、講義でも専門家層の説得に失敗するのではないかと危惧している。

参考文献

椿　広計(1999), データサイエンスの社会人教育, KEIO SFC REVIEW, 4巻 38-43, 慶應湘南藤沢学会.