医療介護とデータサイエンスの新たな視点

慶應義塾大学　山内 恒人　君
スイス再保険　藤澤 陽介　君
慶應義塾大学　小林 凌雅　氏
慶應義塾大学　小暮 厚之　氏

山内

それでは、今日はこれから「医療介護とデータサイエンスの新たな視点」ということで、１時間半にわたるプレゼンテーションを行いたいと考えております。データサイエンスに関しましては、この場で何もいろいろと申し上げるまでもなく、今後のアクチュアリーを支える、一つの大きな柱になるだろうと思います。その中で、やはりコーディングスキルや、あるいは、その統計に関する最近の話題というものにキャッチアップするというのは、なかなか機会がないと難しいということもございまして、最先端の話だけではなくて、もちろん今日はかなりその断片も見えるわけですけれども、われわれが学んできたことも含めて、どのようにデータサイエンスという形で新たな進化が得られているのかということについて、３名の方からお話を伺うという予定にしております。詳細はそれぞれのお話の中で語られますので、この場では申し上げることはいたしません。

早速でございますけれども、その３名の方々をこれからご紹介したいと思います。最初はスイス再保険会社日本支店シニアヘルスソリューションズマネージャー、バイスプレジデントでいらっしゃいます藤澤陽介さんです。よろしくお願いします。お次でございますが、慶應義塾大学政策メディア研究科後期博士課程に在学中でいらっしゃいます、小林凌雅さんです。よろしくお願いします。最後でございますけれども、慶應義塾大学総合政策学部の教授でいらっしゃいます、小暮厚之先生です。それでは、早速、藤澤さんから、よろしくお願いします。

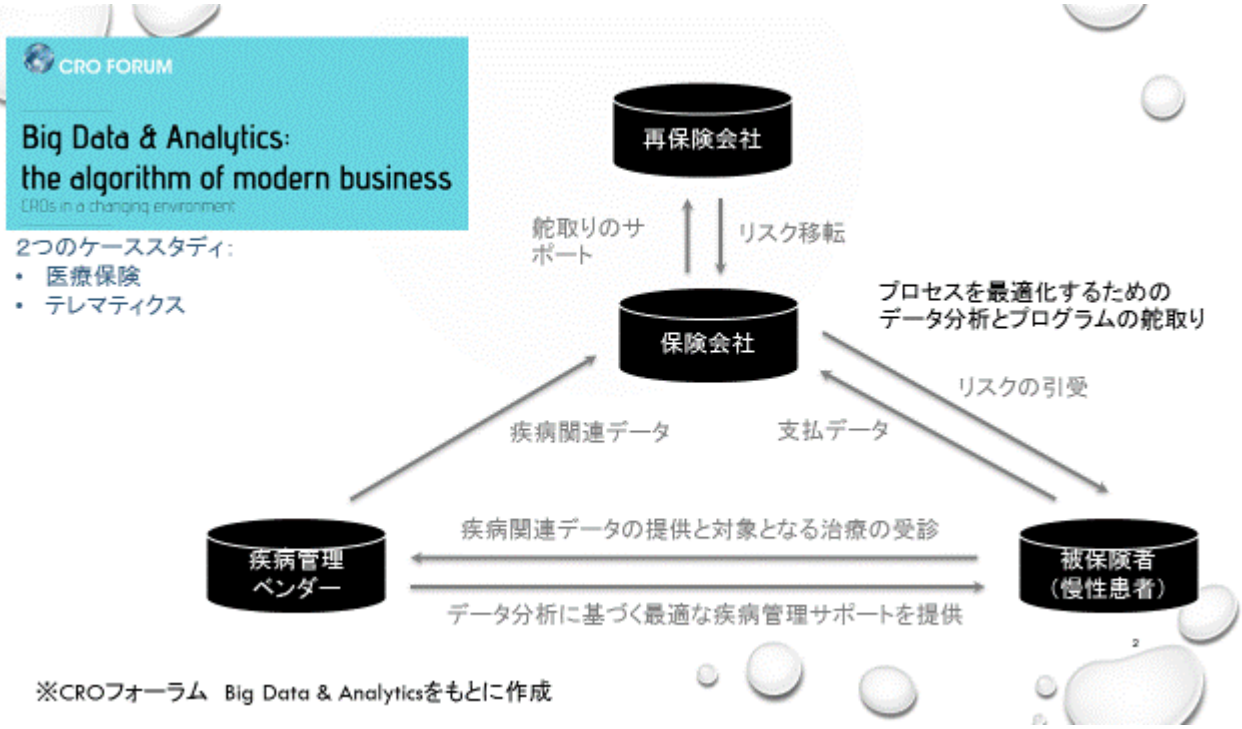
藤澤　ご紹介ありがとうございます。スイス再保険会社で働いています、藤澤と申します。本日はよろしくお願いします。

平成29年度　日本アクチュアリー会　年次大会

医療介護とデータサイエンスの新たな視点

スイス再保険
藤澤陽介

私のパートは導入部分です。これまでも何度か小暮先生や小林さんと意見交換させていただいたのですが、私も一人のオーディエンスの気持ちで、早く先生方のプレゼンを聞きたいと思っています。私の役目はこのパートをなるべく早く終わらせて、次のプレゼンを、オーディエンスになった気持ちで楽しみたいと思っています。

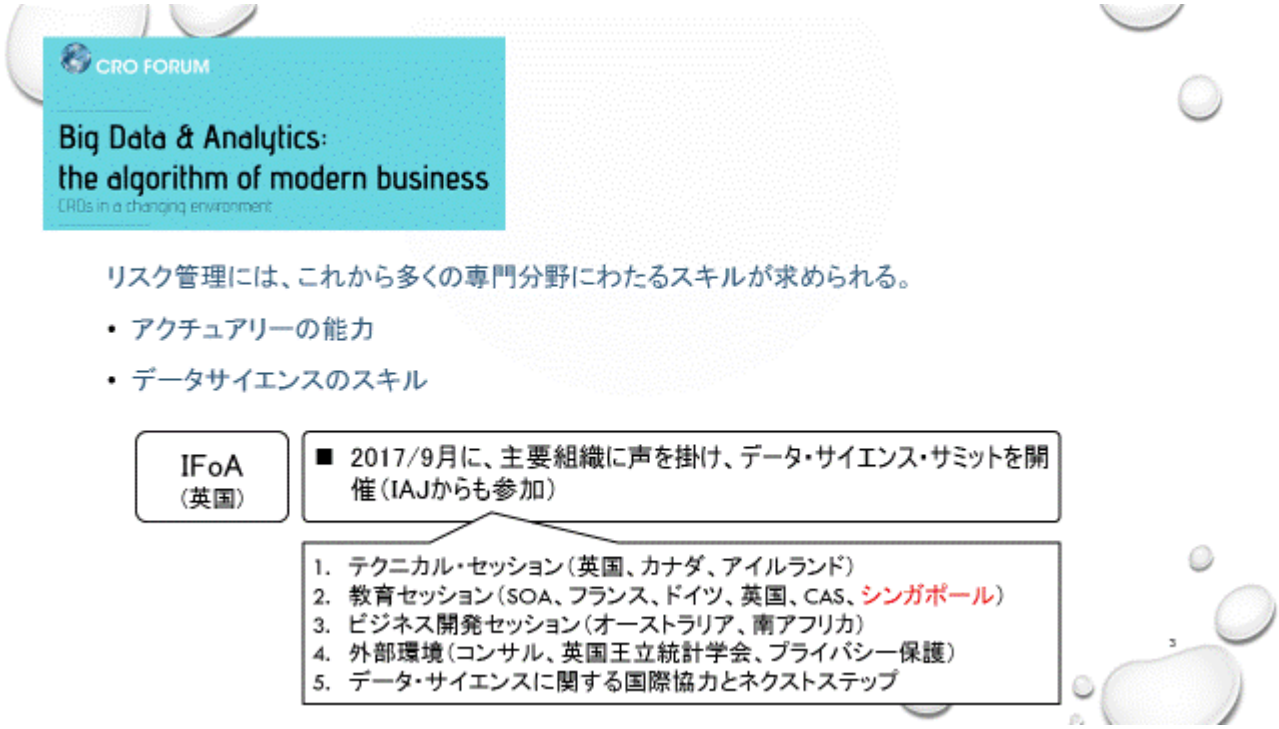


※CROフォーラム　Big Data & Analyticsをもとに作成

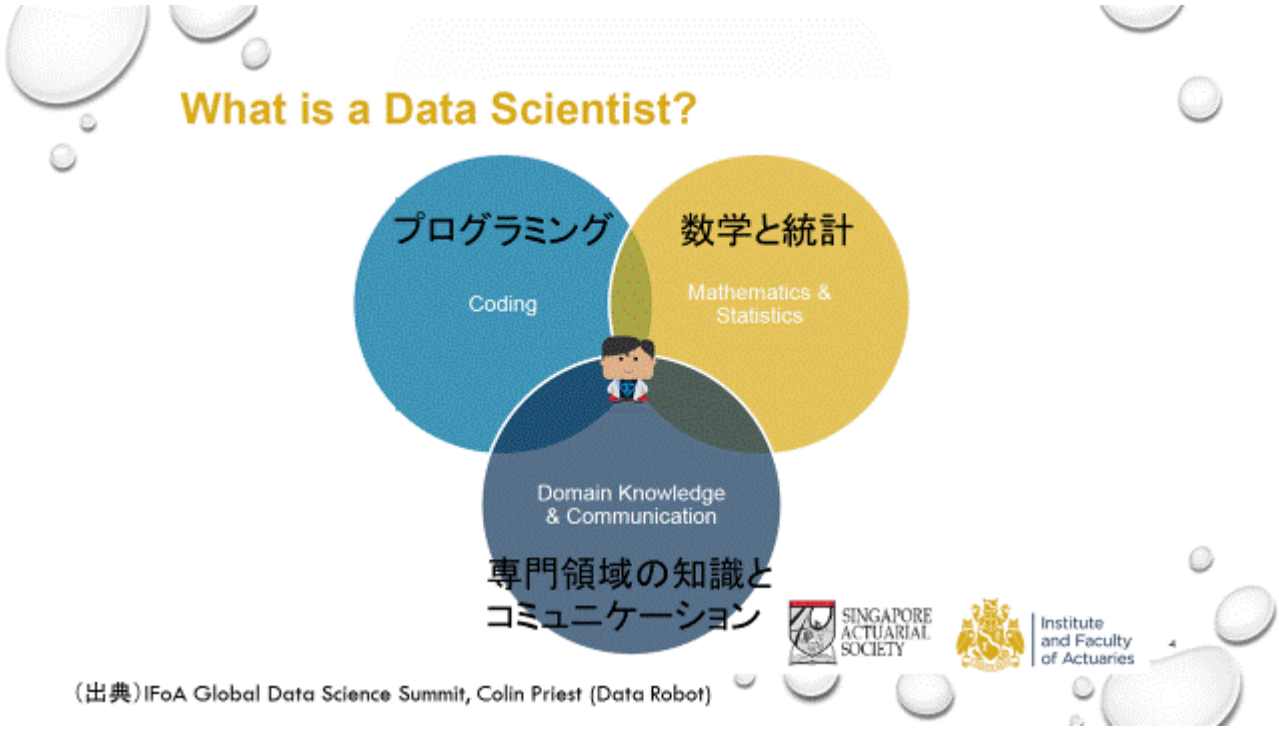
最初は導入ということで、実務的な視点の話をしたいと思って作ったスライドです。ＣＲＯフォーラムというところがございます。ここが幾つか、外部環境の変化を踏まえて、今後どのようなリスク管理が必要になるのかということをレポートにまとめています。その中の一つがこのビッグデータとアナリティクスというテーマでございます。このペーパーの中に二つのケーススタディが載っていまして、損保で言うとテレマティクスが有名だと思うのですけれども、生保、もしくは医療の世界だと、医療保険のところで

ちょっと実務的なその動向が変わりつつあるということが紹介されています。

一般的に保険会社は被保険者、リスクを引き受けて、そのデータを入手するという形のビジネスをしているケースが多いと思います。ここに疾病管理のベンダーが入ってきて、例えば被保険者の中でも典型的なものは糖尿病の患者さんですが、糖尿病の患者さんが合併症を予防するためにどのようなことができるのかという形で疾病管理のベンダーが入ってきて、どちらかと言うとパーソナライズした疾病管理のサポートをする。そこでまた新たなデータが集まってくる。これを保険会社が、ここでそのままつないで良いのかという法的な問題もあろうかと思いますし、これが例えば遺伝子になると、倫理的に本当につないで良いのかという問題も当然出てくるかと思いますが、そのようなデータを仮につなぐことができた場合に、これまでいわゆる保険会社でやっていた経験値分析に新たな変数が入ってくるという状況になり、使える統計的なスキルは少しずつ変わってくるのではないかという流れになっています。それを再保険会社がバックでサポートしているという事例が、幾つか海外の方で出てきています。

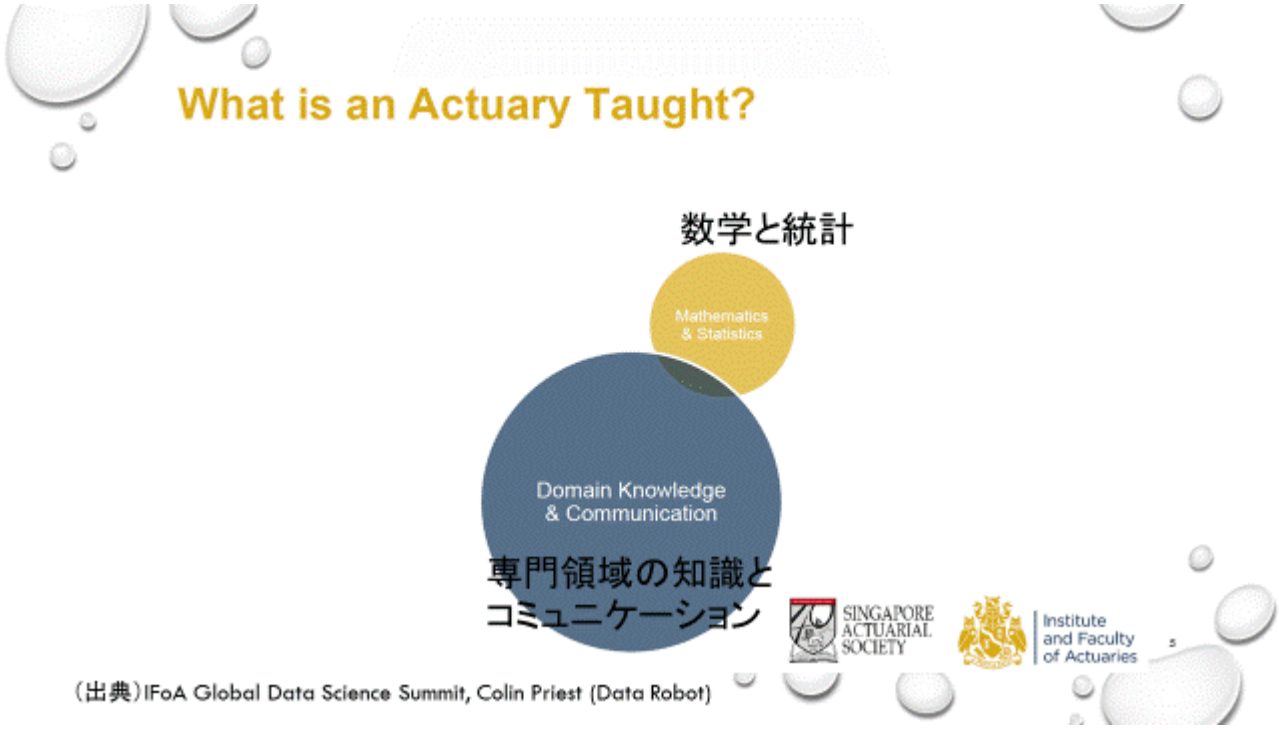


このペーパーの中に書いているのですけれども、リスク管理という観点で、アクチュアリーのスキルはもちろん必要だと思いますが、データサイエンスのスキルを、今後磨いていかないといけないということも書かれています。このようなことを背景に、これは９月に開催されたデータサイエンスの集中セミナーでも少しお話ししたのですが、イギリスのアクチュアリー会が９月にデータサイエンスのサミットを開催しています。その導入ということで、一部分、ここは教育のセッションの、シンガポールの方の話を紹介させていただこうと思っています。



（出典）IFoA Global Data Science Summit, Colin Priest (Data Robot)

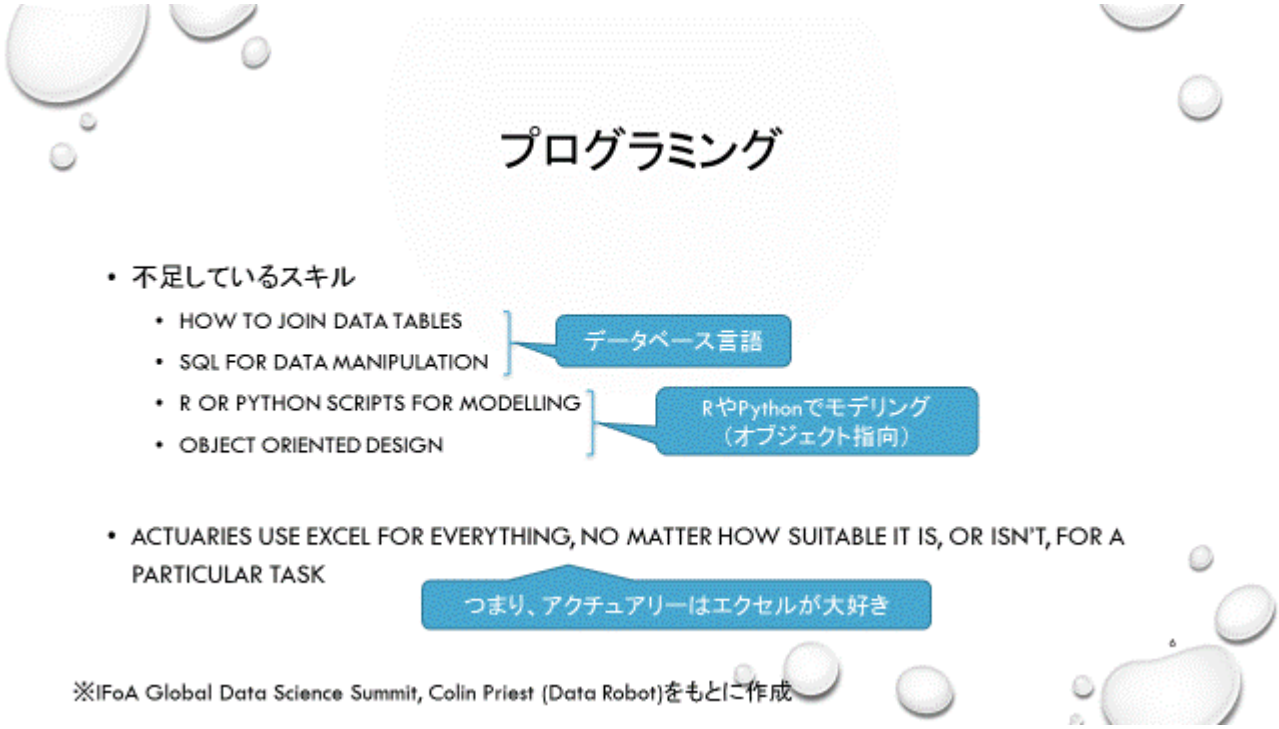
これは、データサイエンスはどのようなスキルが必要になるのか。これもよく出てくる図なのですけれども、プログラミングと、あと数学と統計、そして専門領域の知識とコミュニケーションが必要になる。これはシンガポール・アクチュアリー会の方のプレゼンを抜粋させていただいたものです。一方、アクチュアリーはどのような領域の教育を受けてきたのかという図が、こちらです。これも多分国によって違っていて、その場でもディスカッションがあったのですけれども、「やはり国によって、もっと数学と統計が強い国もあるのではないか」というような意見なども、サミットの中では出ていました。



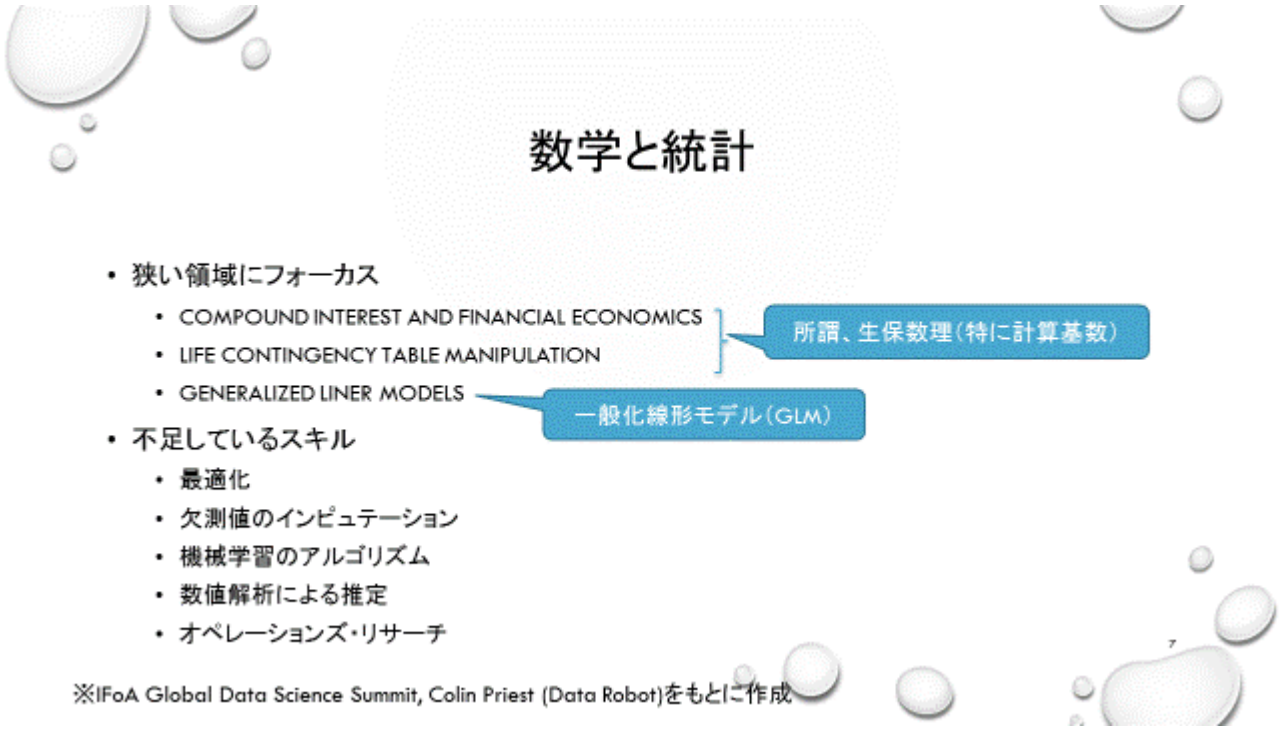
（出典）IFoA Global Data Science Summit, Colin Priest (Data Robot)

例えば、日本で言うと、多分数学的なスキルは海外のアクチュアリーに比べると、もうちょっと高いと思っています。このバブルがもう少し広いのかもしれませんが、統計の場合、なかなか日本で統計教育を受けるような機会はなかったり、現代的な統計技術が増える中で、そのような技術にキャッチアップして

いく教育がちょっと足らないのではないかという話や、あと、プログラミングを学ぶ機会がやはりなくて、これは各国のアクチュアリー会として、そのような教材であったり、勉強の機会を提供していく必要があるという話がありました。

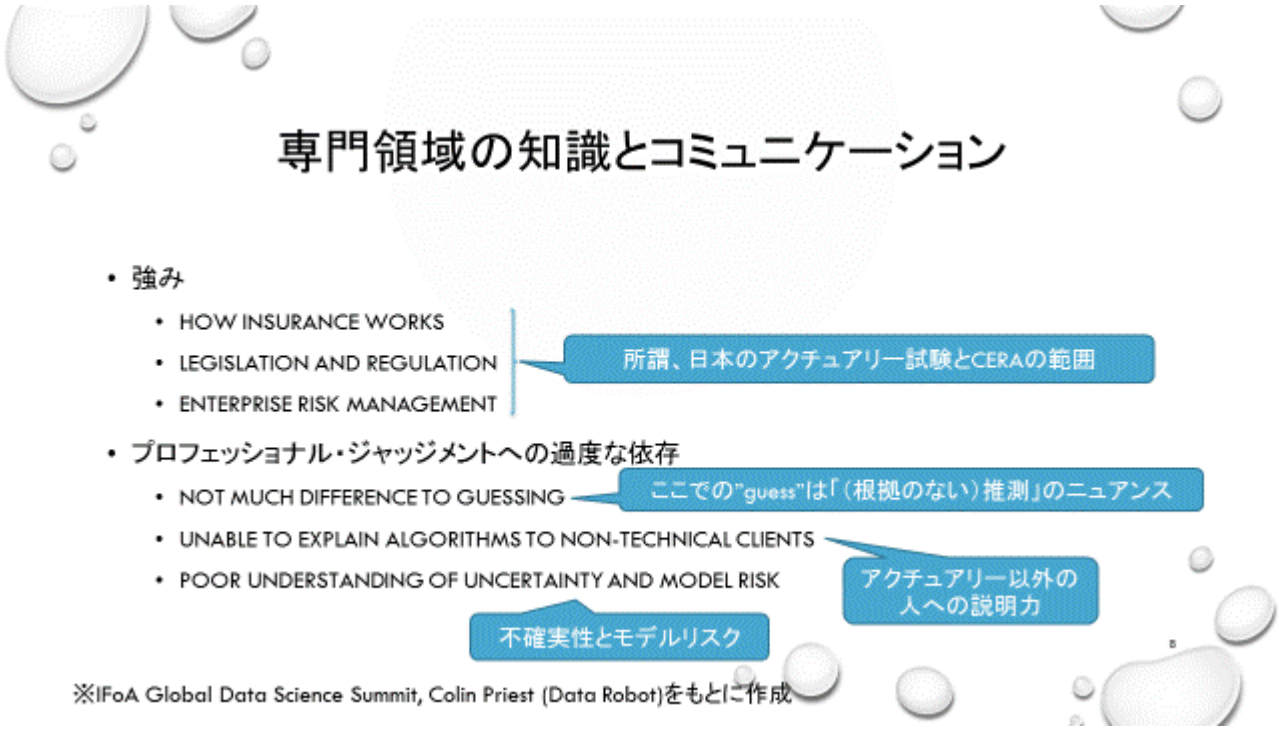


例えば、プログラミングという観点で、不足しているスキルとして、ちょっと英語のままで恐縮なのですけれども、例えばＳＱＬやデータベースの取り扱いなどです。ＲやPython でのプログラミングのスキルというのは、やはり絶対的に不足しているという話がありました。下に書いているのが、日本語に訳すとインパクトがありすぎると思って、あえて訳していないのですけれども、やはりアクチュアリーはエクセルで何でも実装してしまう。マクロを書いて、ＶＢＡを書いて、シミュレーションをするようなところが、特徴としてあるのではないかという話がございました。



次に、数学と統計の部分ですけれども、これもその場で議論になっていたのは、やはり生保数理の計算基数は、狭い領域にフォーカスしすぎているのではないかという話や、一般化線形モデルについても、これも９月の集中セミナーの時にも少し話があったのですが、やはり真のモデルというものが分からない中で、このＧＬＭは説明がとてもしやすいということもあって、過度に使いすぎているのではないかというような意見が出ていました。

不足しているスキルとしましては、最適化や、欠損値があったときにどのように補完するのか、インピュテーションするのか。機械学習、数値解析、オペレーションズ・リサーチなど、このような分野を学ぶ機会はなかなかないのですけれども、データサイエンスの世界では、このようなスキルも必要になってくるというような話がありました。



あと、専門領域の知識とコミュニケーションのところですけれども、ここは強い部分ということで、保険の仕組みや法律や規制などはアクチュアリー試験にも入っていますし、その下のエンタープライズ・リスク・マネジメントも、ＣＥＲＡの範囲に含まれていたり、アクチュアリーの人でもリスク管理をやっている人が増えてきているので、ここは一つの強みなのではないかと言われています。

一方で、ここも国によって違うと思いますが、プロフェッショナル・ジャッジメントに過度に依存しているのではないかと。日本の場合は、個人的にはそこまでジャッジせずに、なるべくデータを探してきて、プライシングするというカルチャーが多分あると思っていますが、海外のアクチュアリーのプライシングを見ていると、結構ジャッジに頼っているところもあります。guess はいろいろな意味がありますが、根拠のない推測に近いようなニュアンスで、ジャッジをしている人がいるのではないかという意見も出ていました。

あとは、説明力のコミュニケーションのところ、これもよく言われるところですけれども。あとは不確実性やモデルリスクなど、そのような部分に対する理解が少し不足しているかもしれないというような話でした。

以下の分野の知識拡充

- プログラミング
  - RとPYTHON
- 数学と統計
  - 罰則付回帰とスパースモデリング（LASSO）
  - ベイズ統計
- 専門領域の知識とコミュニケーション
  - モデルリスクとパラメータリスク

そのようなことを背景に、今回小暮先生と小林さんにお願いして、主にそのような話をしていただきたいと、私自身がお願いしたのですけれども、その以下の部分の知識を拡充するような、一種の継続教育のような場を提供したいということで、このセッションを企画した次第です。

ここに書いているような技術の中には、もしかしたら、実務で使えるような、面白い手法があるかもしれないと思いますので、そのような部分を今後もっと深掘りして、実務で使ったり、論文を書いたり、そのような方向につなげていっていただけるといいなと思っている次第です。

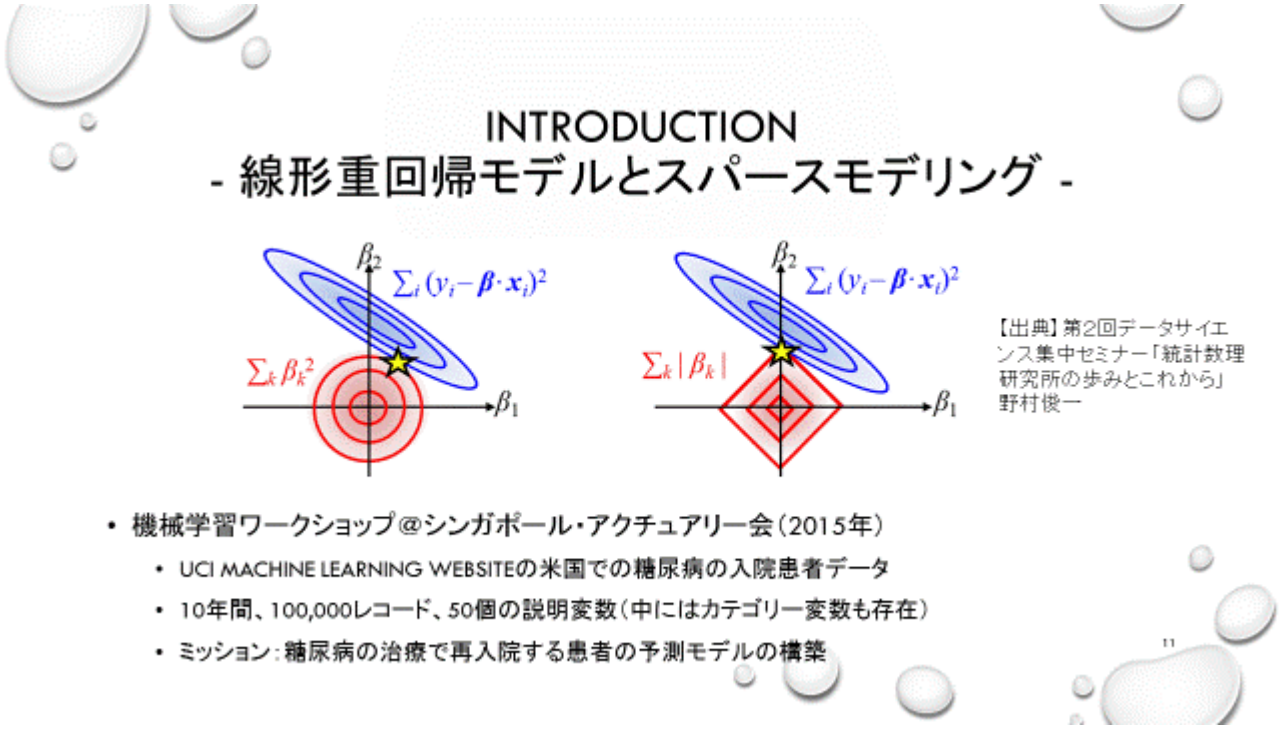
INTRODUCTION
- 線形重回帰モデルとスパースモデリング -

- $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$
- どうやって β_j を決めるのか？
- P値？ それとも AIC？
- 最小二乗法 $RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$ を最小化
- RIDGE回帰　$RSS + \Lambda \sum_{j-1}^{p} \beta_j^2$ を最小化　⇔ RSSを最小化 SUBJECT TO $\sum_{j-1}^{p} \beta_j^2 \leq s$
- LASSO回帰 $RSS + \Lambda \sum_{j=1}^{p} |\beta_j|$ を最小化　⇔ RSSを最小化 SUBJECT TO $\sum_{j=1}^{p} |\beta_j| \leq s$
  - L1ノルムの罰則項を加えることで、パラメータ推定と同時に、変数選択を自動的に行う手法

10

導入部分ということで、一般化線形モデルや線形重回帰モデルですね。これはアクチュアリー試験にも出てきていますので、皆さんご存じだと思います。どのようにパラメーターを選ぶのかというときに、例えばＰ値や、ＡＩＣというのも一つの手法にはなるかと思います。

一般的にこれまで教えられていた統計学に基づくと、最小二乗法を使って誤差が最小になるようにパラメーターを選ぶというのが常套手段だと思うのですけれども、例えばそこに罰則化項を入れて、このリッジ回帰やLassoの部分ですけれども、この λ で、λ がついている罰則化項を入れて、その上で最小化する。これをLagrangeを使って表現すると、右側の表現になって、制約条件付きの最小化問題になります。このような手法を用いて、パラメータの推定と変数選択を自動的にやるという概念が、Lassoです。もう少し詳しい概念は小林さんの方から、この後説明があると思います。



９月のデータサイエンスの集中セミナーで、統数研の野村先生がお話しされていた中にも、この罰則付き回帰の話がございました。これはイメージなのですけれども、左側がリッジ回帰、右側がLassoですね。その前のスライドで、制約条件付きの最適化問題の所ですけれども、ＲＳＳを最小化するというときに、制約条件が二乗の場合がリッジで、Lassoは絶対値でということですので、その制約条件が二乗の場合はこの円のような形で決まって、その中でＲＳＳが最小化できるようにパラメーターを決めると、そのβ1 とβ2というパラメーター空間の中では、星で書いた所が最適化の解になるということです。

一方で、右側のLassoの所で見ると、このβ1 とβ2の所で、絶対値ですので、とがっているような形になっています。ここの例で言うと、そのβ1が０になります。このような形で変数を自動的に選択して、β1は０になって、β2を使えば良いというように、自動的に変数選択してくれるような特徴があるので、スパースモデルと言っています。

Lassoに私が初めて出会ったのは、2015年にシンガポール・アクチュアリー会でして、マシンラーニングのワークショップに参加した時に、シンガポールのアクチュアリー会の人がLassoを用いていました。それが糖尿病のデータでして、実際に匿名化した生データを使って、糖尿病の患者が、次に再入院する確率を予測するモデルを作るということを、ワークショップでパソコンを持ち込んでやりました。その中でも、説明変数が50を超えるような中で、どれが効くのかというところを、ＡＩＣやＰ値でやってももちろん良いのですけれども、Lassoを使ってやると、数行のコードで一気にその変数が選択できるという説明を受けて、結構、私自身は衝撃的に思ったのが、そのLassoと初めて会った時の印象です。

INTRODUCTION
- 要介護者の死亡率 -

- 実務的な課題：推定に用いるデータが不足
- "長寿化を考慮した要介護状態別死亡率の予測－混合リー・カーター・モデルによるアプローチ"（小暮厚之、神谷信一、伏見考弘、2016年度JARIP 研究発表大会）
  - この論文では、長寿化を考慮して要介護状態における死亡率を予測するための統計的手法について考察する。このような手法を開発する際に困難となる点は、要介護状態の人口集団に対する死亡データが十分に整備されていない点である。本稿では、全人口集団が健康状態別の部分人口集団に区分されるような状況で、各健康状態別の死亡数データは欠如しているが、対応する人口データは利用可能という状況下で、健康状態別死亡率を予測する新たな手法を提案する。この手法を我が国の介護年金制度のデータに適用し、要介護状態を反映した死亡率の予測を議論する。

12

あと、介護の観点で、これは介護のプライシングをされた方は皆さん、共通の苦労をされていると思うのですけれども、やはりデータがとても不足していて、どのように死亡率や要介護の発生率を推定するのかというところが、かなり難しいのですね。このような部分を補完するときに、小暮先生からお話しいただくベイズ的なアプローチを使うと、データがない所をうまく事前分布を置きながら、要介護者の死亡率の推定が可能となります。

本セッションの目的

- 一種の継続教育
- データ・サイエンスに関する集中セミナー（第3回：岩沢宏和）
  - 以下のたぐいの（もっとずっと詳細な）整理を早急に行う必要がある．
  - 機械学習の手法や道具立てと称されるもので，アクチュアリー分野での予測に直接的ないし間接的に役立ちそうなもののうち，
    - I) 統計学の手法や道具立てとして理解できるもの（主成分分析，劣則付回帰，EMアルゴリズムなど）→ 統計学の枠組みの中で取り入れ
    - II) I)以外で汎用的な手法 → 統計解析の基本手法として取り入れるべきもの（クラスタリングの諸手法など）と（将来に備えて）その他の機械学習の基本手法として習得しておくべきもの（NN，SVMなど）とに分類して取り入れ

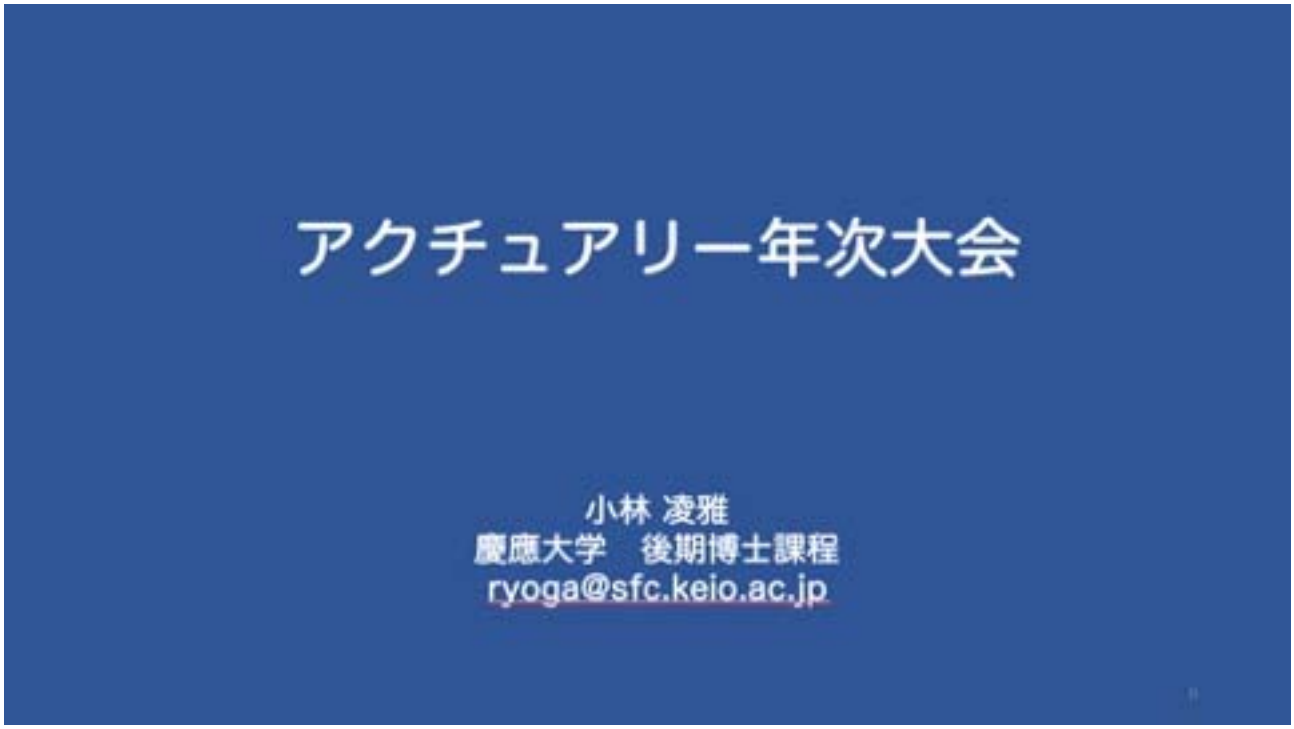
13

これは最後のスライドです。データサイエンスの集中セミナーの時に、岩沢先生がおっしゃっていたように、データサイエンスと言うと、何かふわっとしていて、定義がなくて、どこまでがデータサイエンスなのか、よく分からない印象があるかと思うのですけれども、そのような中でも、粒々のその手法の中に、アクチュアリーが使える手法は多分幾つかあるはずで、私が個人的に思っているのは、この Lasso やベイ

ズ的アプローチは、比較的アクチュアリーに近いデータサイエンスの範ちゅうの手法なのではないのかと思っています。ですので、これは繰り返しになるのですけれども、今日この後に続く話を聞いていただいた上で、面白いなと思えば、それを深掘りして実務に使ってみたり、論文を書いてみるなど、そのような方向に向かうといいなと思っております。私の発表は以上です。

山内　藤澤さん、どうもありがとうございました。それでは、引き続きまして、小林さんの方からプレゼンテーションをお願いしたいと思います。

小林　ご紹介ありがとうございます。



慶應大学の小林です。今日はどのような話をするかというと、まず統計学と機械学習についていろいろ思うことを書いてみたので、皆さん、聞いてくれたら嬉しいです。その後に、先ほど藤澤さんからありましたスパース推定について、話していこうと思っています。

まずビッグデータについてです。これは発表する前に藤澤さんや小暮先生、山内先生と話していて、他にもアクチュアリーの方はいたのですけれども、そのアクチュアリーの中の１人が、「私はビッグデータという言葉はなるべく使わないようにしている」と言っていました。そこで私はひねくれているので、あえてビッグデータという言葉を使ってみました。ここでビッグデータという言葉を使うために「ビッグデータとは何か」と改めて考えてみました。どのようなものとして捉えるのがいいかと思ったのですが、一般的にはこのガートナー社の３Ｖ、Volume、Variety、Velocity のような話によって定義されることが多いと思います。ただ、今までの自分の周りの現状などを見て、ビッグデータをどう定義すればいいかと考えたときに、計画されずに集められたデータのことをビッグデータとして定義するのが、とても自分の中で直感的であると思いました。なぜかと言うと、昔のコンピュータ、その左の図のような、紙に穴を開けてプログラミングするというコンピュータがあったらしいのです。自分は知らないのですが、そうですよね、小暮先生。

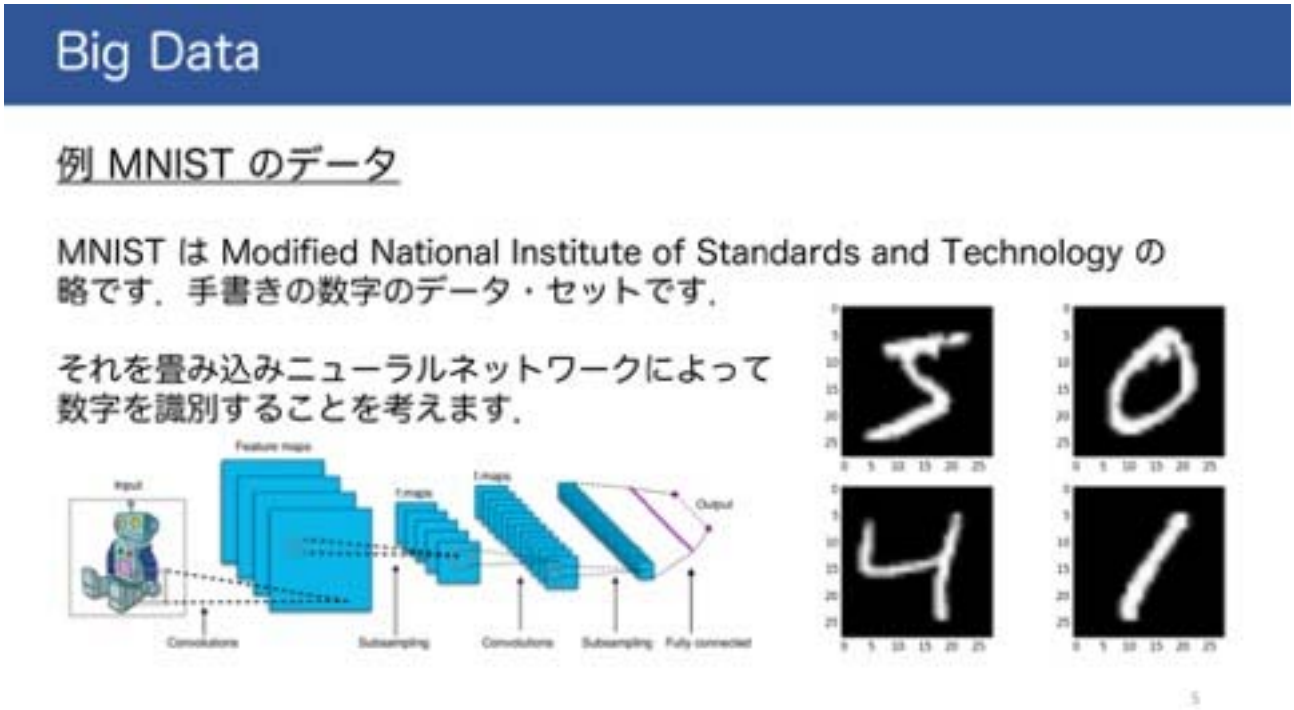
小暮　そうなのです。

小林　ただ、今のコンピュータは右側のコンピュータのように、例えば回帰分析、逆行列の演算は大きな処理が必要です。逆行列を計算することも、あの左側のパソコンだったらどれくらいかかるのかと考えたら、到底その何も考えられずに集められたデータなどは分析できるような時代ではなかったと思われます。ただ、その逆行列の演算などもぱっと計算できるので、やはり計画されずに集められたデータを分析するような余裕が出てきたと、最近感じています。

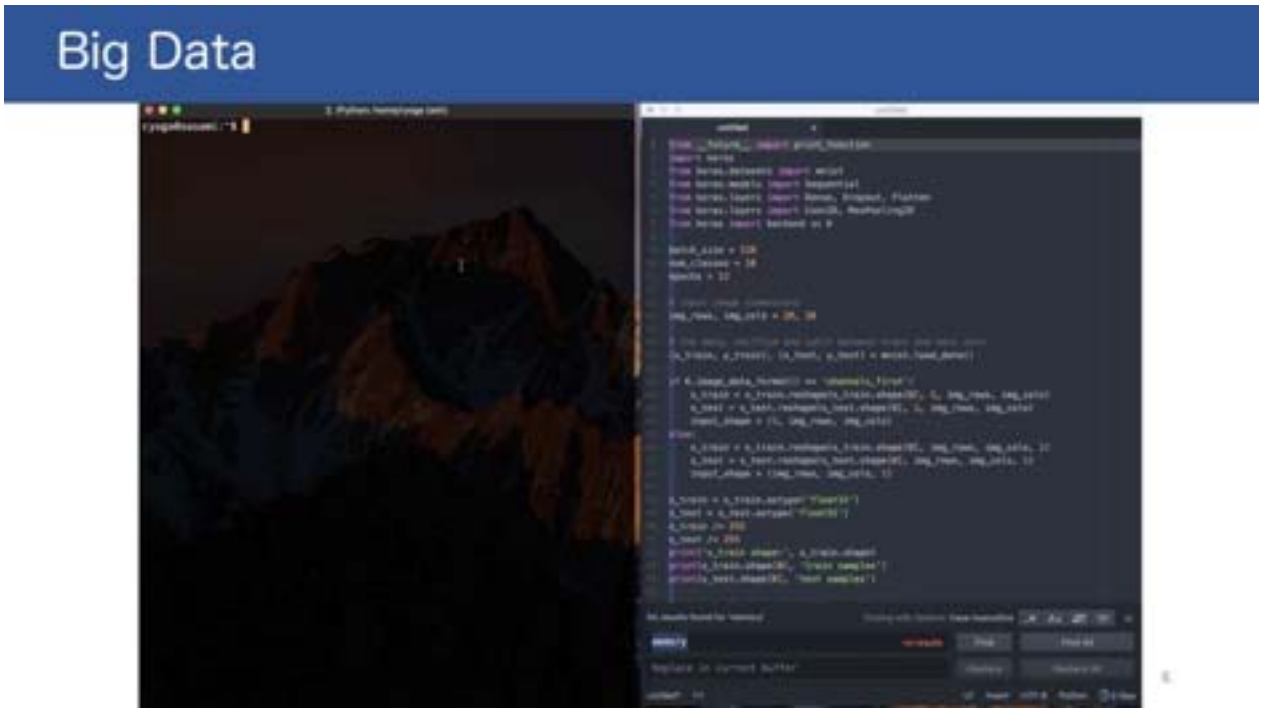
これは、あるコンサル業界の友人から聞いた話なのですけれども、コンサル業界に集まってくるデータも、やはり計画されずに、「とりあえず何とかしてくれ、ドーン」というように渡されるようなデータが多いらしいのです。これは業界用語で「ドーン案件」と言うらしいのです。やはり計画されずに集められたデータというのが重要なのかなと、このスライドを作った時はとても思っていました。ただ、その後にあ

る企業さんに行かせていただいて、画像のデータがあるのですけれども、この画像のデータをパートの人が一つ一つ「どこが手で、どこが頭だ」とプロットしているような会社がありまして、この定義もちょっと古くなりつつあるのかなと、今、この場で話していて思っています。

例えばですけれども、画像解析などは結構、最たる例だと思います。画像解析の分野では機械学習の方法を用いて、盛んに研究がされてきています。ここで扱われる画像データは、典型的には何か明確な目的があって、このような分析をしようなどと考えずに集められたデータではなく、ただウェブ上にある取得可能なデータを適当に引っ張ってきて、それから何ができるのかと考えていることが、とても多いように感じます。このように扱われるデータを、ここではビッグデータとして話していきます。



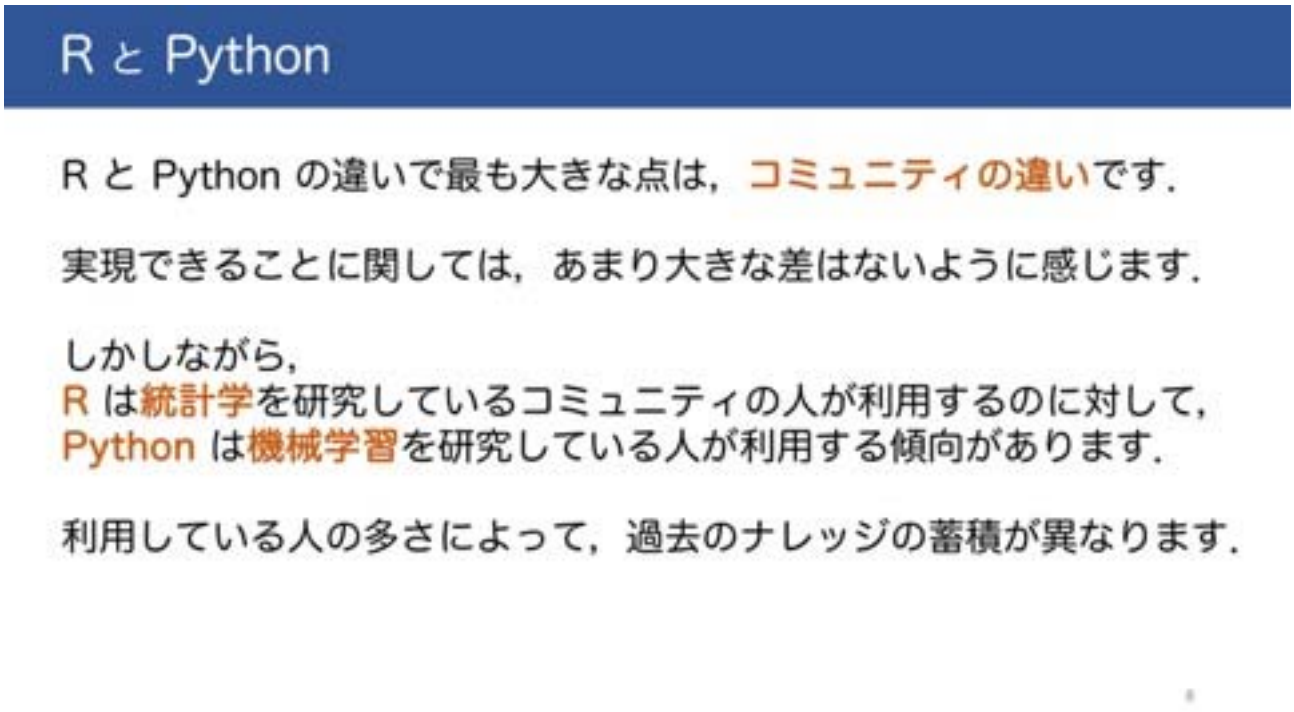
一番有名なのは、この MNIST のデータなどなのですが、手書きの数字のデータを、はやっているので、ニューラルネットワークのような方法を使って分析していました。



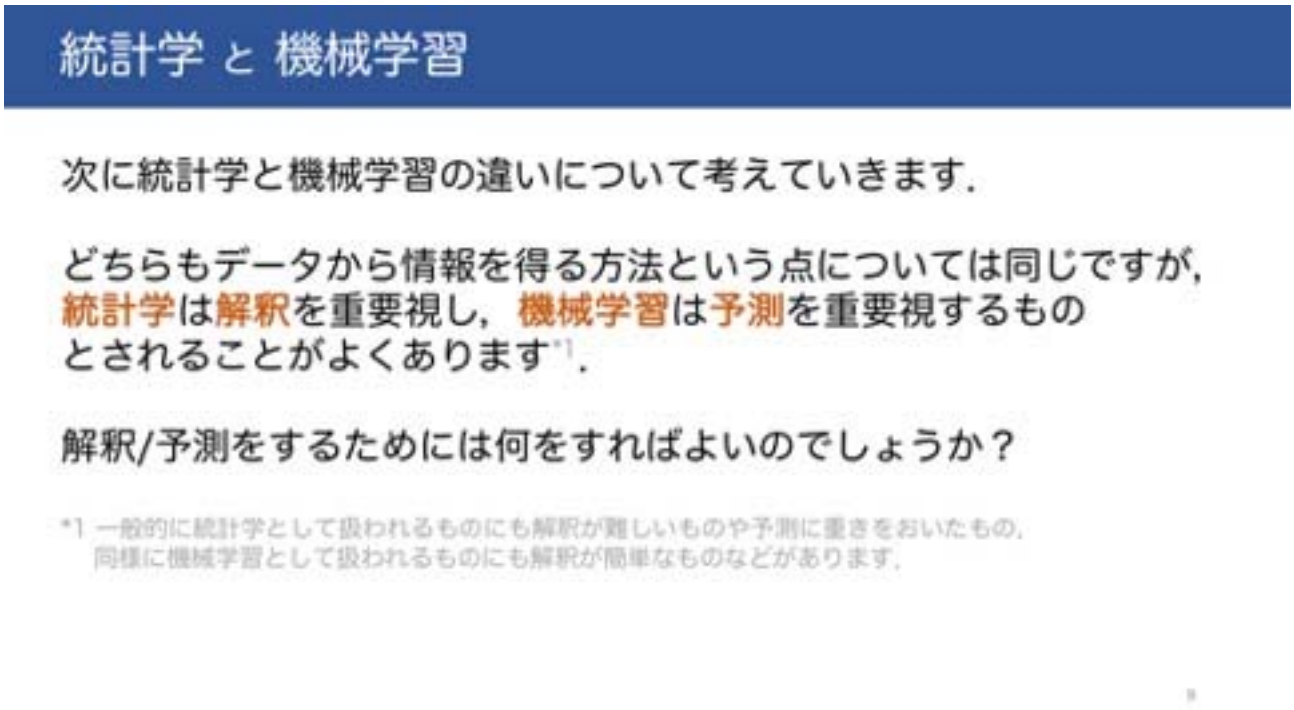
話ばかり聞いてしまうと、多分飽きてしまうので、実装をお見せしようと思います。小さくなってしまうのですが、このような感じで、さっきの手書きの画像のデータを分析していて。ただの手書きの画像、この文字が幾つなのかと分類のモデルを回していて。これは Python ですね。あえて iPython で、ちょっと派手にやっています。これがこう分析していくと、このように計算ができて……。ここですね。たしか 10 分強あるのですけれども、これがたった 10 分、この手元のパソコンで回すだけで。そろそろですね。「正答率が 99%も出るって、結構すごい時代になったな」と、個人的に思います。先ほどのこのパソコンでは考えられなかったような時代になってきたのかなと思っています。

これも最近よく聞かれるので、あえて入れたのですけれども、「R と Python、どちらを勉強した方がよいで

すか」という質問を結構されます。結構、Rと Python の違いについて、真剣に考えてみました。いろいろ考えてみたのですが、このRと Python の違いで最も大きいことは、やはりコミュニティの違いかなと思います。

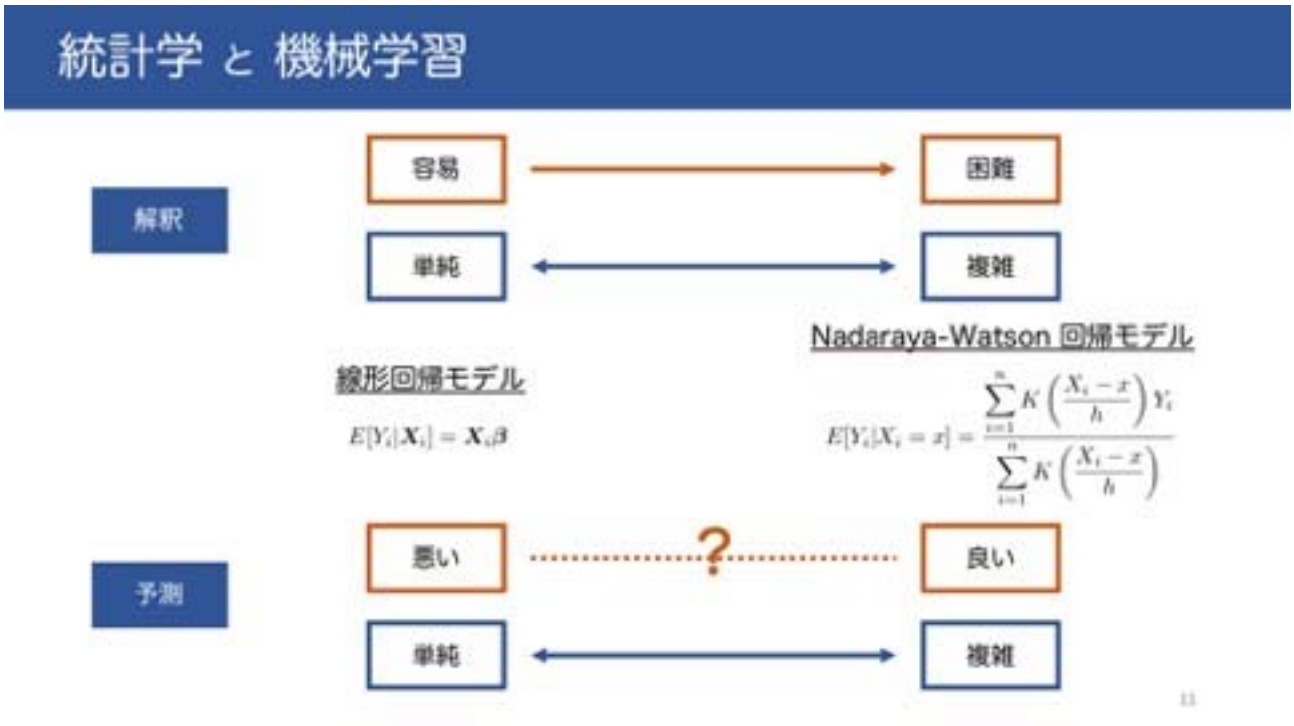


言語でできることは、問題がきちんと定義できて、関数さえ書ければ、全部同じような処理ができるので、多分あまり違いはないと思います。ただ、扱っている人のコミュニティは大きく違うなと感じています。Rは統計学を研究しているコミュニティの人がとても利用していると思います。統計学会などへ行くと、ほとんどRでコードを書いていて、結構自分はびっくりしました。逆に言うと、この Python、では、機械学習のコミュニティの研究発表に行くと、ほとんどの人が Python で書いています。利用している人の多さによって、過去のナレッジの蓄積が異なるので、それは「自分がやりたいことが、どっちが多いのかな」ということをきちんと考えないと、判断できないものなのかなと思いました。あまり答えになっていないですがこれがどちらを勉強したらよいかの答えです。

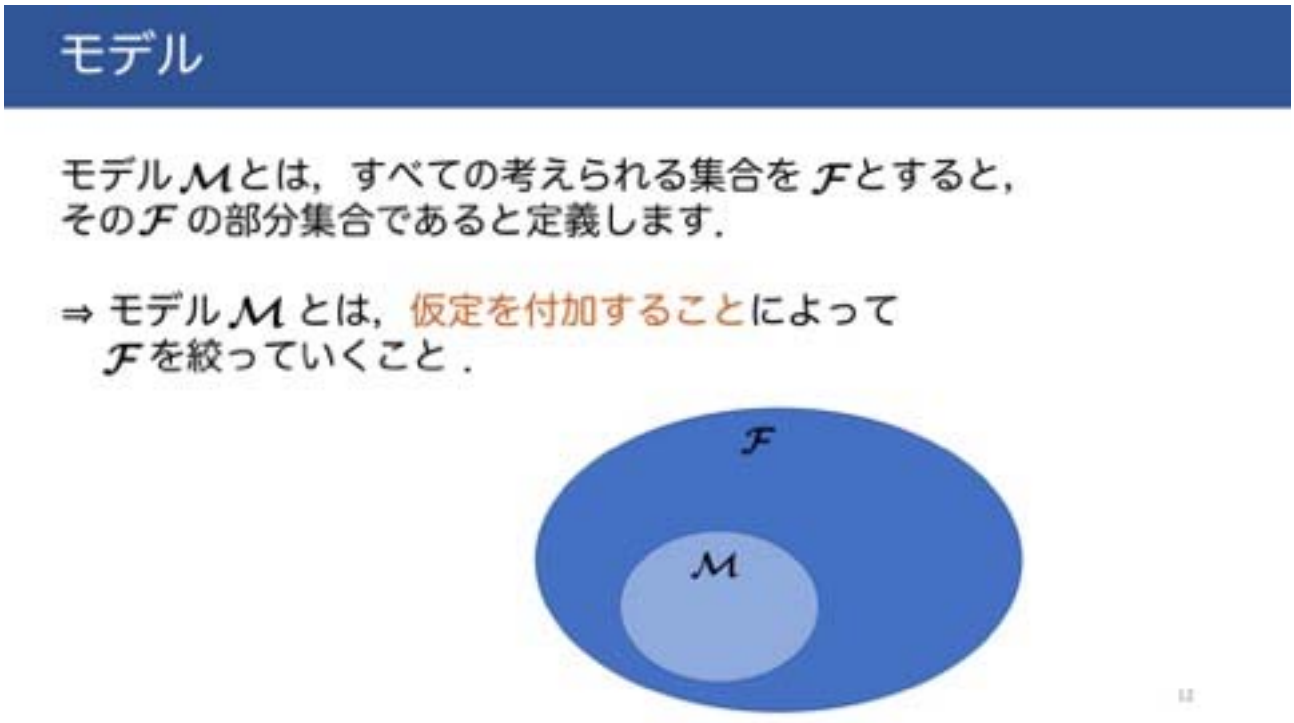


次に統計学と機械学習の違いについて、考えていきたいと思います。統計学と機械学習は何が違うのかということも改めて考えてみたのです。これは人の回答を借りたのですが、どちらもデータから情報を得る方法という点については同じです。ただし、統計学は解釈を重要視していて、機械学習は予測を重要視しているものとされることがよくあります。ただ一般的に統計学として扱われるようなものにも、解釈が難しいものや、予測に重きを置いたものもあります。同様に機械学習でも解釈を簡単にするためのものもありますが、こちらは、とりあえずこれで話を進めていきたいと思います。

その解釈、予測をするためには何が必要なのでしょうか。恐らくですけれども、どちらもモデル、モデル化によって実現することが典型的な方法だろうと考えました。ただし、異なる点として、解釈することは、より単純なモデルを使った方が実現できるのに対して、予測することはモデルの複雑さはあまり関係ないです。



そもそも複雑さというものは何なのかと考えていくと、結構複雑さというのは人間が定義するものなので、解釈できるかどうかによって定義されると考える方が自然なのかもしれません。もちろん何か数学的にアプローチする方法もあるのでしょうが、この解釈に関しては、直感的に容易なものと困難なものがあります。ただ予測に関しては、「単純なモデルだから悪い」「良い」というのは、正直よく分からないですよね。単純なモデルの方が、予測がよくなる場合もありますし、悪くない場合もあります。



そもそも、そのモデル化によって実現できるという話をしましたが、そもそもモデルとか何かという話もしておきたいなと思いました。モデルとは何かと言うと、すべての考えられることの中に、幾つか仮定

を置くことだと。仮定を置いて、集合を小さくしていくことをモデルと言うこととします。

ちょっと分かりやすい例かどうか分からないですけれども、友人が車に乗っていて海まで行ったのですけれど、鍵をなくしてしまいました。海で鍵をなくしてしまって、鍵を探そうとなったのですけれど、皆さんは海でかぎをなくしたら、どのような所を探しますか。いわゆる機械学習の人は、海全部を掘れば鍵は見つかるだろうと考える人達が先ほどの定義の機械学習だと思います。ただ、統計学の場合は、この辺りにかぎがあるだろうということを、モデルを置いて探していきます。これはとても一長一短があって、自分がいた場所から誰かが持っていってしまったと。そうなってしまうと、そのかぎは一生見つけることはできないです、その自分のいた所だけを探していては。ただし、この全部を掘り返してしまえば、いずれかは見つかると思います。

前半のまとめです。これまで統計学と機械学習の違いを比較してきました。しかしながら、最近ではやはり統計学と機械学習は別に反発し合うのではなくて、昔の確率論と統計学のように対立し合うのではなくて、統計学と機械学習は互いにいいところを認めて、取り込んでいく方向に動いているのではないのかと個人的には思っています。また、統計学と機械学習の間のような、統計的機械学習と言われる分野も、非常に近年注目を集めたりしています。

スパース推定

次の分析例を通して考えていきます．

下記のようなデータ・セットが得られたとします[2]．

被説明変数: 中性脂肪（diabetes patients）

説明変数：年齢（age），性別（sex），BMI（body mass index），
血圧（ average blood pressure）など残り 6 変数．

データ数は 442 個で，1 年前をベースラインとします．

[2] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R., 2004, "LEAST ANGLE REGRESSION," *The Annals of Statistics*, Vol. 32, No. 2, 407-499.

ここからは、スパース推定を話していきたいと思います。今回は特に次の小暮先生の発表に合わせて、ベイズを意識して、このスパース推定について説明しています。

まず例を通して考えていきましょう。次のようなデータセットが得られたとします。例えば、説明するものとしては、中性脂肪を説明したいとします。ただ被説明変数としては年齢、性別、ＢＭＩ、血圧などを説明変数となるデータを考えましょう。これは結構有名な論文なのですけれども、Efron などが書いている、Lars という最適化方法の具体例で扱われているデータをそのまま持ってきました。

スパース推定

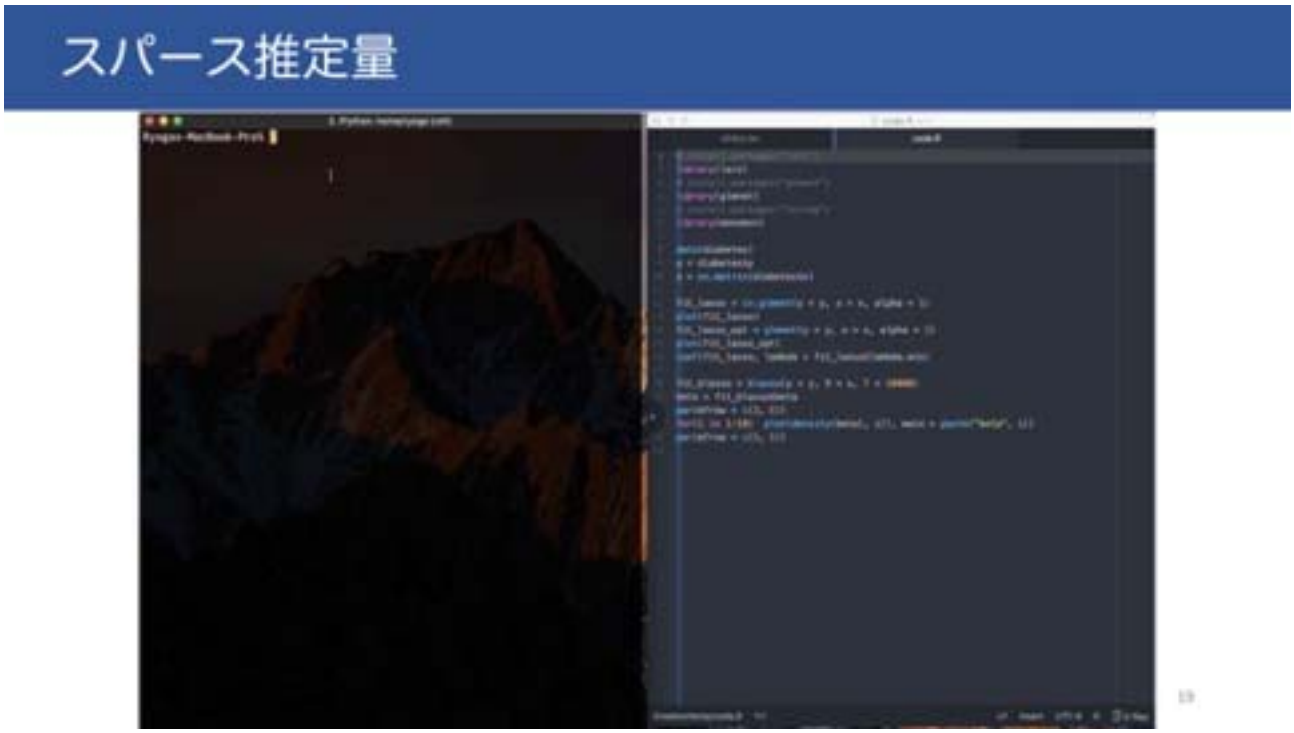
	AGE	SEX	BMI	BP	Serum measurements						Response
Patient	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	y
1	59	2	32.1	101	157	93.2	38	4	4.9	87	151
2	48	1	21.6	87	183	103.2	70	3	3.9	69	75
3	72	2	30.5	93	156	93.6	41	4	4.7	85	141
4	24	1	25.3	84	198	131.4	40	5	4.9	89	206
5	50	1	23.0	101	192	125.4	52	4	4.3	80	135
6	23	1	22.6	89	139	64.8	61	2	4.2	68	97
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
441	36	1	30.0	95	201	125.2	42	5	5.1	85	220
442	36	1	19.6	71	250	133.2	97	3	4.6	92	57

Efron et al. (2004) から引用．

このようなデータセットですね。

ここで、データ・サイエンティストという言葉を最近よく聞きます。そこで、データ・サイエンティストという言葉も自分なりに定義してみようと思いました。このデータ・サイエンティストとは、ビッグデータを分析する人です。つまり、この計画されずに集められたデータを分析する人と定義します。

データ・サイエンティストが、さっきのデータセットを分析することになったら、どうすればよいでしょうか。データ・サイエンティストは、これも自分の周りの話を聞いた状況を考えて、まだとても新しい職種なので、専業の人がデータ・サイエンティストになるというよりは、従来の仕事をしつつ、データ・サイエンティストの仕事もしている人も多いように感じました。つまり、何が言いたいかというと、多分時間がない人が多いと、皆さん、言っていました。そのようなときに、スパース推定が使えるのではないのかなと思っています。



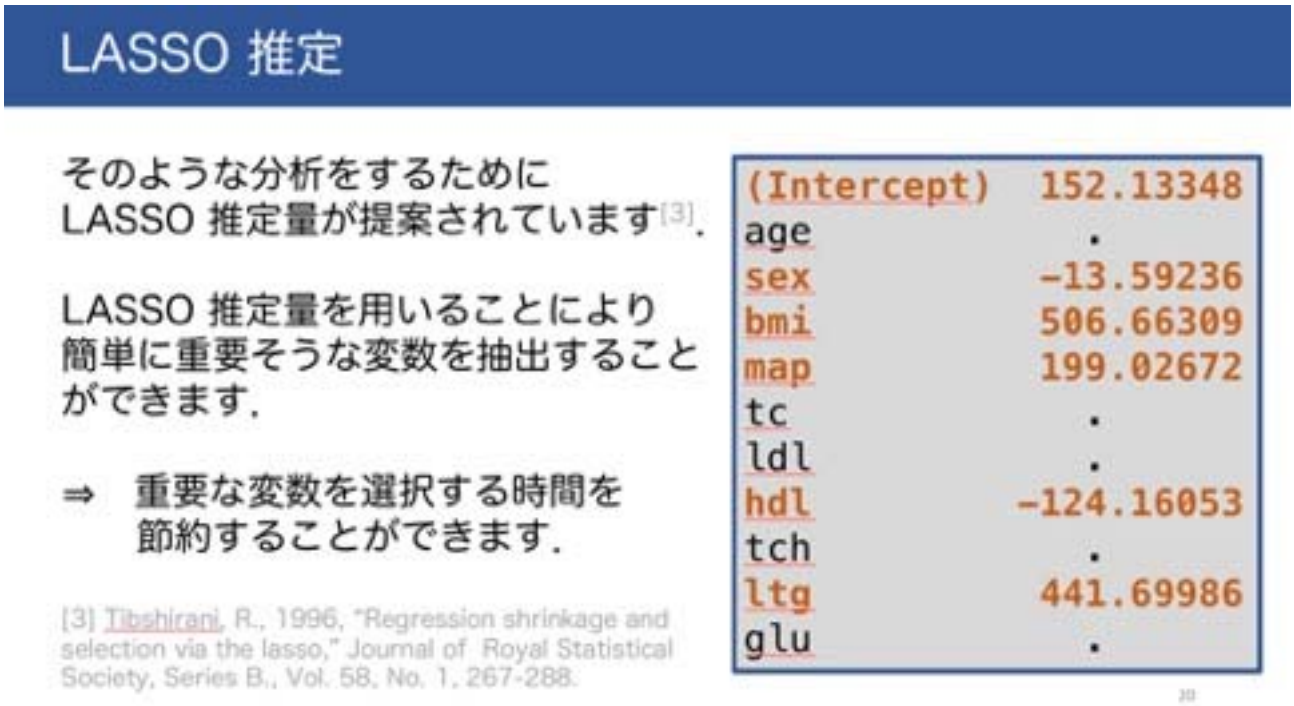
これですね。このように、今度はＲで書いているのですけれども。もし興味があれば、参考としてコードもつけましたので、ぜひ見てみてください。

山内　小林さん、これは具体的に何をしようとしているものなのですか。

小林　それは、これから説明していきます。すみません。多分、先に見せた方が後に聞くモチベーションになるかなと思ったので。

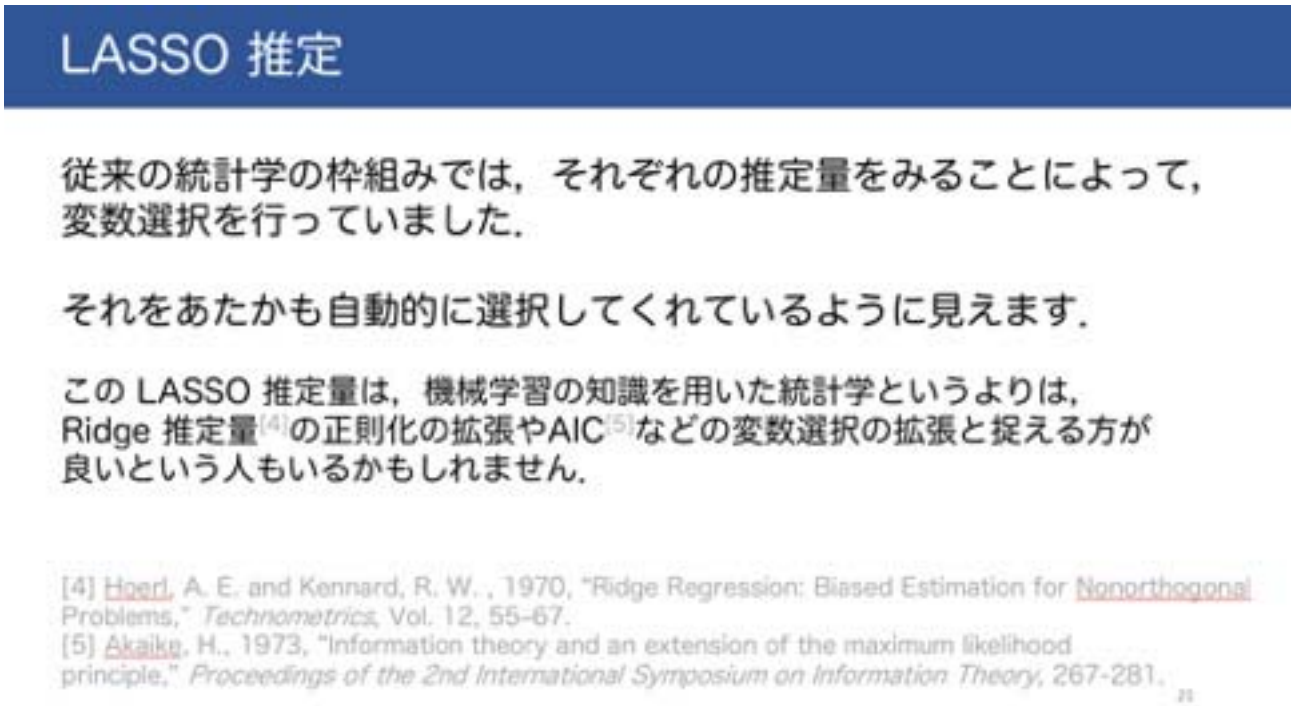
山内　失礼しました。

小林　ここで推定していますね、何かしら。ＭＣＭＣをして、かっこいいなと思って、このような感じで説明できたらなと思っています。



今まで何をしていたかということを説明しますと、Lasso 推定量というものをやっていました。Lasso 推定量とはどのようなものかというと、Lasso 推定量を用いることによって、簡単に重要そうな変数を抽出することができます。重要な変数を選択する時間を節約することができます。

そこに出ているように、先ほどのデータセットの線形モデルを考えたのですけれども。こちらのデータセットですね。そうすると、例えばこのデータセットにおいては、年齢などは関係ありませんよ。その代わり、性別やＢＭＩなどがとても重要な要素となっていますよと。直感的にも、中性脂肪ならＢＭＩなどは何となく合っているのかなという感じがします。



従来の統計学の枠組みでは、それぞれの推定量を見ることによって、先ほど藤澤さんもおっしゃっていましたが、Ｐ値やｔ値、もしくはＡＩＣを見ることによって、変数選択を行ってきました。それをあたかも自動的に選択してくれているように見えます。

この Lasso 推定量は、機械学習の延長として書いているのですけれども、本当はリッジ推定量の正則化の拡張や、ＡＩＣなどの変数選択の拡張として捉える方がよいという人もいるかもしれません。

LASSO 推定

最小二乗推定量は，

$$\hat{\boldsymbol{\beta}}_{ols} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \left[\frac{1}{n}\sum_{i=1}^{n} (Y_i - \boldsymbol{X}_i'\boldsymbol{\beta})^2 \right]$$

で定義されるのに対して，LASSO 推定量は，

$$\hat{\boldsymbol{\beta}}_{lasso} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \left[\frac{1}{n}\sum_{i=1}^{n} (Y_i - \boldsymbol{X}_i'\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{d} |\beta_j| \right]$$

で定義されます．ただし，$\lambda > 0$ とします．

もうちょっと具体的に話をしますと、いわゆる最小二乗推定量とはどのようなものだったかというと、通常の最小二乗法と同じように、このように定義ができます。ただ、一番後ろにその推定量の絶対値の和がλ倍されたものを加えたもので定義しています。

LASSO 推定

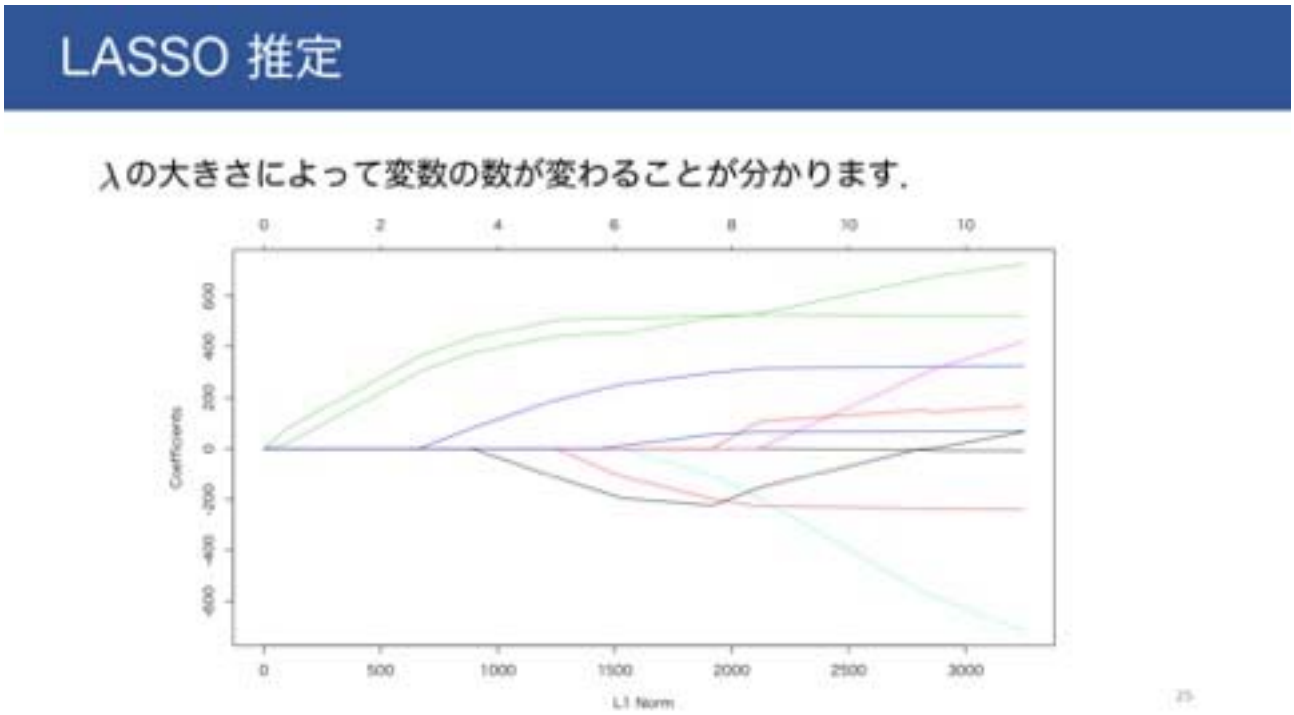
ラグランジュ的に捉えると，

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} (Y_i - \boldsymbol{X}_i\boldsymbol{\beta})^2$$

$$\text{s.t.} \sum_{j=1}^{d} |\beta_j| < \lambda$$

と書き換えられます．これは，推定量の絶対値和が
“予算” λ の中で選択されると考えることができます．

これは次のような問題として見ることもできます。そもそも絶対値の和がλより小さくなるということはどういうことなのかというのを直感的に説明すると、普通の最小二乗推定量がβ、推定量の回帰係数の和などは、マイナス無限から無限まで取れるような状況を想定しています。これは絶対値がλに収まるように、つまり「予算がλしかない中で、この推定量を振り分けてください」と考えることができます。この提案していた人、Hastie さんの本などにも書かれています。



そうすると、このλを大きくすることによって、右側に行くほど大きいのですけれども、左に行くほどλが小さくなっていて、上に書いてある小さな数字なのですけれど、それは選ばれる変数の数となりますね。すみません、左が大きいですね。λが小さくなると、選ばれる変数の数が大きくなっていって、λを小さくしていくと、すべて０になってしまう。つまり予算が０なので、全部０になってしまうということが分かります。

LASSO 推定

LASSO 推定量は関係ない変数が含まれているかもしれないというモデルを考えています.

統計学では，推定量は一致性と漸近正規性の２つの性質をもつ推定量を良い推定量としてきました.

しかしながら，LASSO 推定量は，バイアスを持ちます.
また，変数選択ができることと引き換えに一致性を失っています.

SCAD[6] 推定量のような（ある条件下では）一致性をもつLASSO のような推定量も提案されています.

[6] Fan, J., and Li, R., 2001, "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, Vol. 96, No. 456, 1348-1360.

Lasso 推定量は関係ない変数が含まれているかもしれないというモデルを考えています。統計学では、推定量は一致性と漸近性の二つの性質を持つ推定量を、良い推定量としてきました。しかしながら、Lasso 推定量はバイアスを持ちます。また変数選択ができることと引き換えに、一致性を失っています。この SCAD 推定量のように、一致性を持つようなら、Lasso のような推定量も提案されています。他にも Lasso 推定量以外に、予算の決め方というのは他にもあるのではないかということで、その他のスパース推定量について議論しています。

その他のスパース推定

LASSO 推定量の他にもさまざまなスパース推定量が提案されています.

スパース推定量を次の関数最初化する推定量とする.

$$Q(\beta) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i'\beta)^2 + p(\beta)$$

ここで，説明のために，罰則項を $p(\beta)$ と書くことする.

とりあえず分かりやすさのために、一般に、予算のところを p(β)、罰則項として書いています。

その他のスパース推定

Elastic Net

$$p(\boldsymbol{\beta}) = \lambda \left[\alpha \sum_{j=1}^{d} |\beta_j| + \frac{(1-\alpha)}{2} \sum_{j=1} \beta_j^2 \right]$$

ただし，$\lambda > 0, 0 < \alpha < 1$ とする．

$d > n$ の場合，すべての変数を取り込むことを可能にする．

[7] Zou, H., Hastie T., 2005, "Regularization and variable selection via the Elastic Net," Journal of the Royal Statistical Society, Series B, Vol. 67, 301-320.

例えばElastic Net という方法があります。Elastic Net とはどのようなときに非常に使えるかと言うと、例えばデータ数がデータの次元数より小さい場合です。そのような場合にも Elastic Net は推定することが可能です。このような場合は、従来の統計学ではどうしても扱えなかった方法なのですけれども、さらにこの二乗のペナルティを入れることによって、推定することが可能になります。

その他のスパース推定

SCAD（Smoothly Clipped Absolute Deviation）

$$p(\beta_j) = \begin{cases} \lambda|\beta_j| & \text{if } |\beta_j| \geq \lambda \\ \dfrac{\gamma\lambda|\beta_j| - (\beta_j^2 + \lambda^2)/2}{\gamma - 1} & \text{if } \lambda < |\beta_j| \leq \gamma\lambda \\ \dfrac{\lambda^2(\gamma^2 - 1)}{2(\gamma - 1)} & \text{if } \gamma\lambda < |\beta_j| \end{cases}$$

LASSO 推定量は大きな回帰係数について過剰な罰則を付加点を緩和した．

[6] Fan, J., and Li, R., 2001, "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, Vol. 96, No. 456, 1348-1360.

他にも SCAD という方法が提案されています。これはどのような方法かというと、このいわゆる Lasso 推定量というのは、絶対値をずらす推定量なのですけれども、SCAD の方は、横軸が真のパラメータなのですけれども、真のパラメータが大きくなればなるほど、きちんと同じようにズレを修正していくようなペナ

ルティとして考えられています。このように、一番上が Lasso 推定量に対応する所なのですけれども、さらに戻っていくようにすることによって、ある特定の状況においては一致性のような性質を持つような方法が提案されています。

その他のスパース推定

MC+（Minimax Concave）

$$p(\beta) = \lambda \sum_{j=1}^{d} \int_0^{|\beta_j|} \left\{1 - \frac{x}{\nu\lambda}\right\} dx$$

LASSO 推定量は大きな回帰係数について過剰な罰則を付加点を緩和した.

[8] Zhang, C. H., 2010, "Nearly unbiased variable selection under minimax concave penalty" *Annals of Statistics*, Vol. 38, 894-942.

他にもこのような方法が提案されていて、いろいろな方法が提案されています。いろいろ紹介されたは良いものの、どの方法がいいのかという話になってくると思うのですが、先ほどのデータセットを使って、Lasso、Elastic Net、SCAD、あと、MC+というような方法で比較してみました。

その他のスパース推定

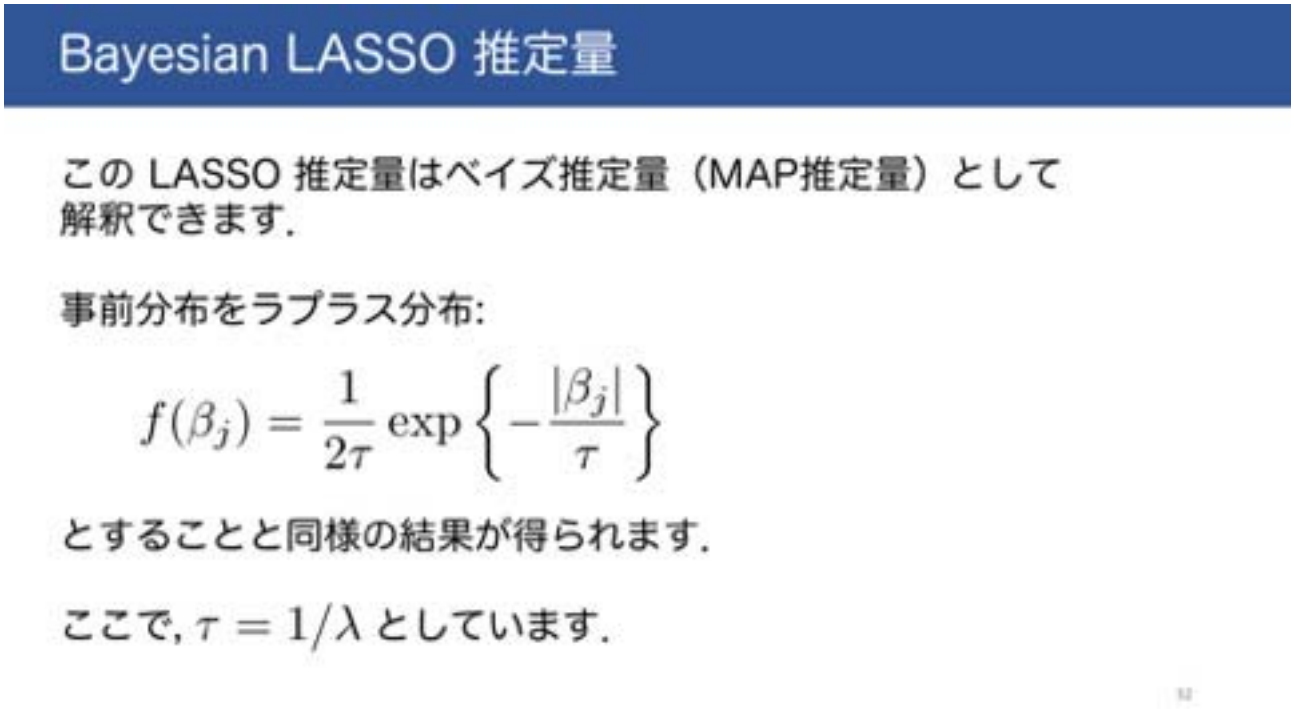
これらの推定量の比較は次の通りです.

	ols	lasso	elastic net	sacd	mc+
(Intercept)	152.13348	152.13348	152.133484	152.13348	152.1335
age	-10.01220	.	.	.	.
sex	-239.81909	-13.59236	-29.564942	-228.52846	-241.9679
bmi	519.83979	506.66309	479.262681	533.89500	518.9142
map	324.39043	199.02672	210.963526	323.96597	321.5075
tc	-792.18416	.	.	-139.93804	-618.5891
ldl	476.74584	.	.	.	351.7738
hdl	101.04457	-124.16053	-144.833396	-238.79698	.
tch	177.06418	.	.	.	128.0975
ltg	751.27932	441.69986	420.982368	550.59918	690.9516
glu	67.62539	.	9.111982	23.82546	67.1174

そうすると、どれも、例えば年齢のような変数は除いた方がいいと。ただ、この辺りやこの辺りは、あ

る方法によっては入れた方がいい、ある方法によっては入れない方がいいとなっています。なぜ、このようなことが起きるのかなということを考えたのですけれども、これはデータ数が無限大、かなり大きなデータ数があるような状況では、すべて統一された結果が得られるはずです。しかしながら、山をどの方向から登っていくのかということで、この違いが出てきているのかなと思います。

ちなみに、これはＡＩＣでもやってみたのですけれども、実はＡＩＣだとうまく働かなかったのですね。それぞれの変数についてその変数を入れるパターン、入れないパターンの２通りあります．例えば変数が 10 個あるとすると、２の 10 乗、1, 024 通りあります。従来の方法ですと 1, 024 通りのＡＩＣをやらなければいけないとなります。ＡＩＣは基本的には、多い場合はステップ関数のようなものを使ってやると思いますが、それはアルゴリズム的にうまくいくことが保証されていません。実際に今回の場合はうまくいかなかったです。スパース推定は、このように最適化の部分がしっかり議論されているので、うまくいっているのかなと思いました。



今度はベイズ的な解釈をしてみようと思います。この Lasso 推定量は、実はベイズのＭＡＰ推定量として解釈することができます。ＭＡＰ推定量に関してはちょっと後で言及します。どのようなものか。事前分布をラプラス分布とすると、同様な結果を得られます。先ほどのデータに対して、ラプラス分布を事前分布として、ＭＣＭＣによって事後分布を求めています。その事後分布における最大値と Lasso 推定量の関係性が見てとれると。

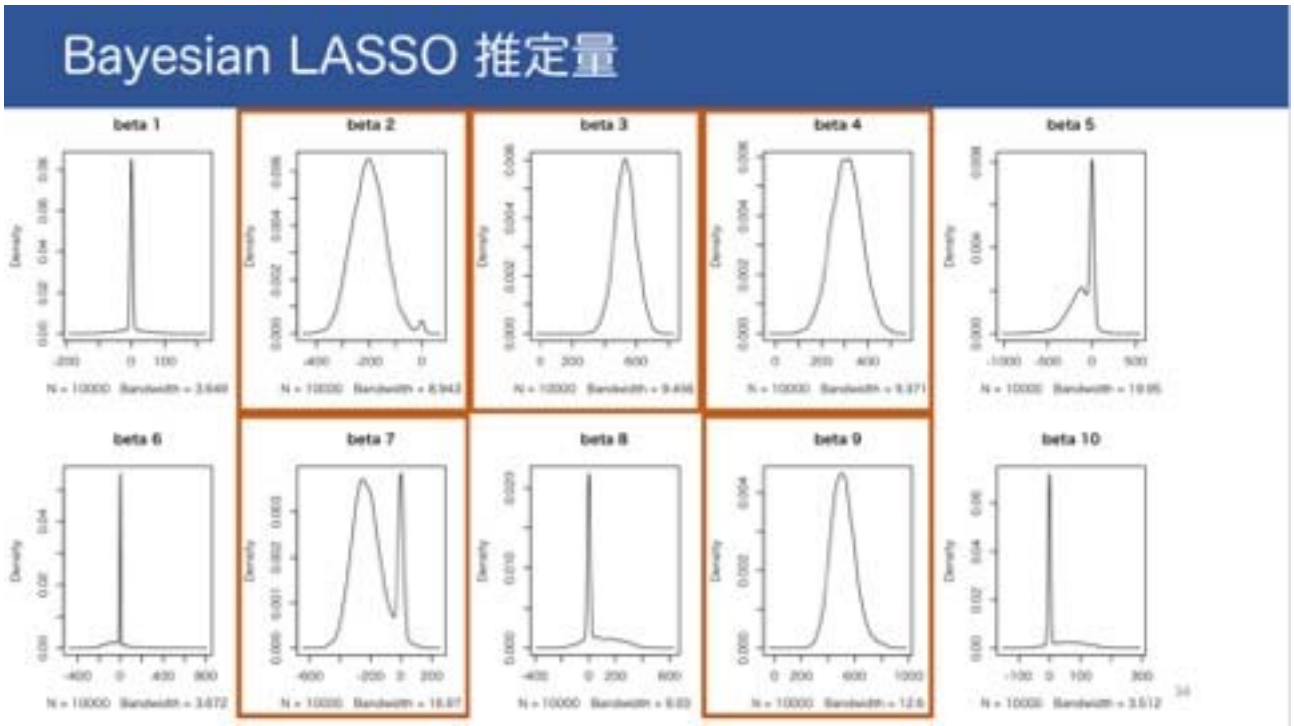
Bayesian LASSO 推定量

先程のデータに対して，ラプラス分布を事前分布として，
MCMCによって事後分布を求めてみます[9]．

その事後分布における最大値と，LASSO 推定量の関係性が
みてとれます．

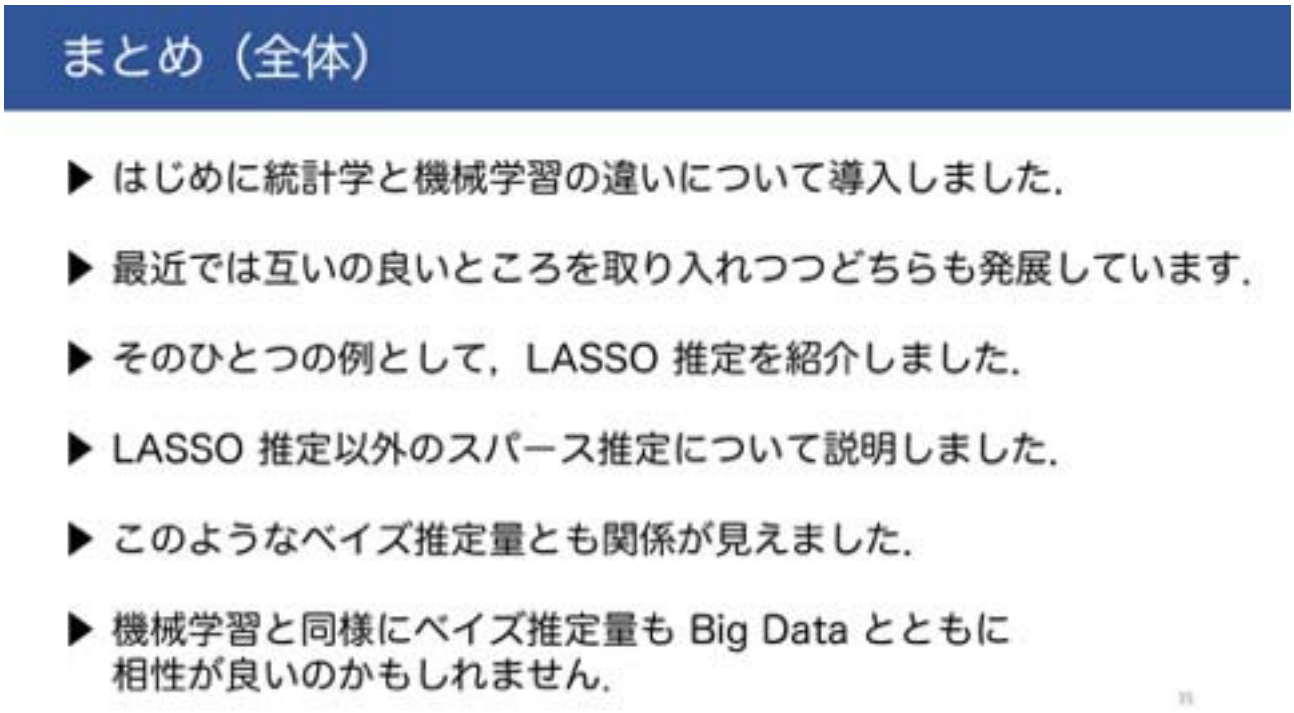
[9] Park, T., Casella, G., 2008, "The Bayesian Lasso," *Journal of the American Statistical Association*, Vol. 103, No. 482, 681-686.

ちなみにリッジ推定量の場合は、平均０の正規分布を事前分布としたときの、事後分布のＭＡＰ推定量に対応しています。



そうすると、こうなっていて、例えばこの７番目の変数などはとてもとがっていて、８番目の変数はとてもとがっていて、０の部分でとてもスパイキーな分布が得られています。

ＭＡＰ推定量というのは、この推定量の中の一番、尤度ではないですけれど、尤度のような事後分布が大きくなる所を選択しているので、０の部分でピークが来ている分布が幾つかあって、実はこれが Lasso 推定量と対応しているというのは、その Lasso 推定量が提案されたときから言及されています。



まとめとして、初めに統計学と機械学習の違いについて、導入しました。最近では互いのいい所を取り入れつつ、どちらも発展しています。その一つの例として、Lasso 推定量を紹介しました。Lasso 推定以外のスパース推定についても、議論しました。先ほどのようなベイズ推定量とも、実はとても関係があるのだということも見えてきました。機械学習と同様に、ビッグデータなど大きなデータを扱うような場合は、ベイズ推定というものも、実は Lasso 推定と同様に相性がいいのかもしれません。最後に、次の小暮先生の発表の前置きをして、今回は終わらせていただきます。ありがとうございました。

山内　小林さん、どうもありがとうございました。でも、多分聞きたいことは山ほどあると思うのですけれども、ちょっとここは心に留めていただいて、最後にＱ＆Ａの時間を設けますので、よろしくお願いします。それでは、小林さん、ありがとうございました。

それでは小暮先生、よろしくお願いします。

小暮　慶應大学の小暮と申します。どうぞよろしくお願いします。

2017年度　日本アクチュアリー会　年次大会

セッションB　15:15～16:45
「医療介護とデータサイエンスの新たな視点」

ベイズ法の視点

小暮　厚之

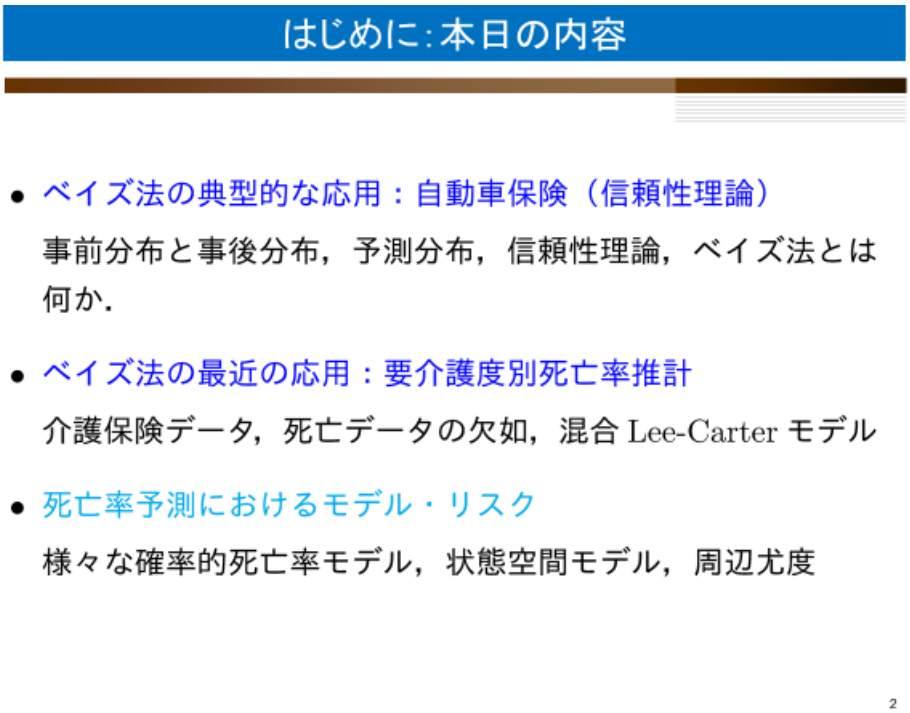
慶應義塾大学 総合政策学部

2017年11月10日

先ほど小林君からうまく言っていただいて、私はベイズという観点から、最近少し研究をしているので、さっきの機械学習もそうなのですけれども、このような新しい統計の流れというのがあって、その中で機械学習自体が私は、とても大きく言ってしまうと、ベイズの一つの形なのではないかなと思っています。それくらいベイズというのは、ある意味基本的な、データ分析の骨組みを与えるものではないかなと思っています。というわけで、今日はベイズについてお話をしたいと思います。

たくさん話したいことがあるのですけれども、少しお話を絞って話したいと思います。最初はベイズ法の典型的な応用ということで、信頼性理論というものがありますけれども、これは保険の世界でとても有名な理論です。ベイズが使われている典型的な例として、とても分かりやすいですし、皆様はもちろん専門家ですので、ベイズについてはよくご存じかと思いますけれども、少しメモリーをリフレッシュしてもらうという意味で、最初に信頼性理論の中でベイズの話をさせていただきたいと思います。

その後で、ベイズ法の最近の応用ということで、これは私が最近やっていることなのですけれども、要介護度別死亡率の推計というものを、ベイズを使ってやってみたいと思っております。その後に、もしも時間があったら、モデルリスクというような問題についても、お話しできればと思っております。以上のような感じで、進めていきたいと思います。



最初なのですけれども、ベイズ法とは何かということで、自動車保険を考えていただきたいと思います。これはベンジャミン＆クルーグマンという、あの有名なクルーグマンですが、彼が SOA で発表していた資料があったのですけれども、それを借りて、ここでお話ししたいと思います。

タロウさんとハナコさんという２人の人がいるとします。これは２人ともＡ社の自動車保険に入っています。タロウさんの方は過去５年間に２回の自動車事故を起こしているのですね。そうすると、５年で２回ですから、0.4 ですね。年間事故率のベスト・エスティメートは 0.4 です。一方、ハナコさんというのはとても優良なドライバーでして、全く事故を起こしていない。このような状況があるとします。タロウさんとハナコさんの保険料をどうしたらいいのかという問題です。ただし、Ａ社は自分の会社で 10 万人の契約者のデータを持っています。その年間の事故率は 0.05 です。このような問題です。

このときに、100％の信頼性というのは、タロウさんとハナコさんの個人的な経験だけに信頼を置くと、どうなるかと言うと、タロウさんというのは 0.4 で、全体の平均は 0.05 でしたよね。だから、それに従ってしまうと、タロウさんの翌年の保険料は８倍になってしまうのですね。そうすると、タロウさんは、「８倍にもなっちゃうんなら」と言って、もうこのような保険に入らないですよね。

一方、ハナコさんの場合、ハナコさんは過去５年間に１回も事故を起こしていないですから、ハナコさんの経験だけに頼ってしまうと、翌年の保険料は０になってしまうではないですか。そうしたら、保険会社は困ってしまいますよね。だから、100％の信頼性というのは、ちょっと無理だろうと。

では、逆に０％の信頼性も、その個人の経験は使わないで、Ａ社が持っている過去の 10 万人のデータを使ったらどうなるかと言うと、２人とも平均的な保険料を支払います。いいように見えますけれど、何とかフェアな感じはしますけれども、でもハナコさんは優良なのだから、そのような人に対して、「いやいや、うちはもっと安い料金で保険を提供できますよ」と言ったら、ハナコさんは抜けてしまいますよね。残るのはタロウさんです。タロウさんのような人たちだけが残ってしまう。だから、０％の信頼性というのもありえないわけです。

例：太郎と花子の自動社保険

- 太郎と花子は二人とも A 社の自動車保険に入っている．
  - 太郎は，過去 5 年間に 2 回の事故を経験している．この個人的な経験から，年間事故率の「ベスト・エスティメイト」は $2/5=0.4$
  - 花子は，過去 5 年間に全く事故を経験していない．この個人的な経験から，年間事故率の「ベスト・エスティメイト」はゼロ
- A 社の保有する 10 万人の契約者の年間事故率は 0.05

（Benjamin and Klugman, 2016）

4

どうしたらいいかと言うと、ちょうど中間を取ったらいいですよね。例えば、25%の信頼性を個人の経験に置きます。そうすると、タロウさんの見積値は 0.1375、ハナコさんは 0.0375、どちらもまあまあ受け入れるような額ですよね。保険会社もこれだったら、ペイしますよね。

太郎と花子の保険料

各ドライバーの個人的履歴に 25%の信頼を置いたら？

- 太郎の見積値は

$$0.25 \times 0.4 + 0.75 \times 0.05 = 0.1375.$$

となる．おそらく，太郎はこの割り増しを受け入れ，より注意深い運転に気を付けるようになるかもしれない．

- 花子の見積値は

$$0.25 \times 0 + 0.75 \times 0.05 = 0.0375.$$

となる． 花子は，この保険料の割引に満足し A 社にとどまるであろう．

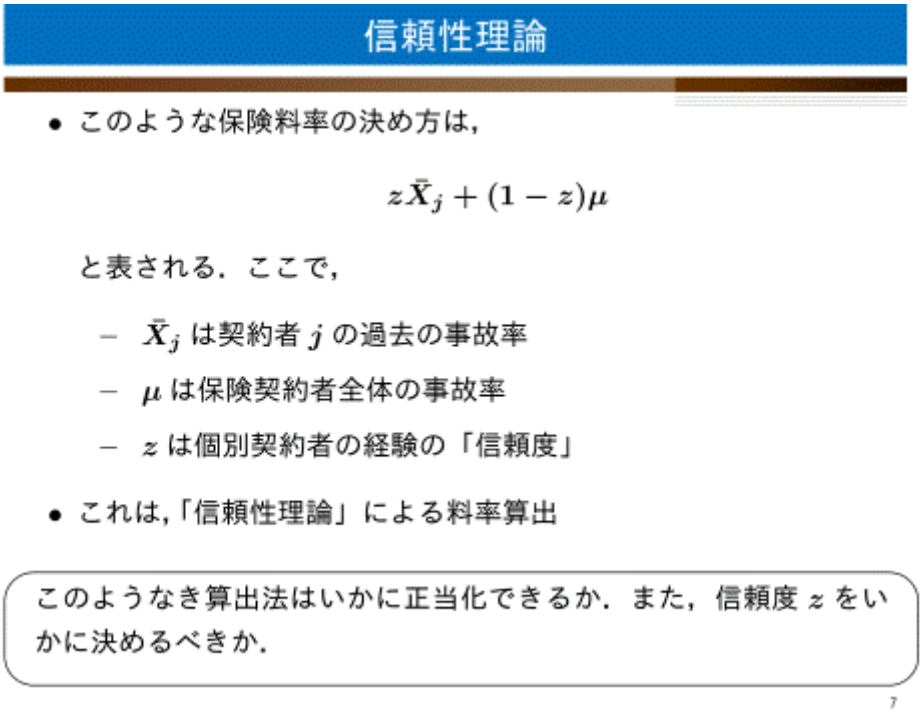
- A 社は，平均的には適正な保険料を集めることとなる．

6

これを信頼性理論というようにいいます。これをちょっと数式で書くと、このような保険料率の決め方というのは、Xバー j というのが契約者 j の過去の事故率です。Xバー j と書いてあるのは、過去５年分の平均。 j というのは j 番目の契約者という意味です。

それから、μ というのが保険契約者全体の事故率です。これを加重平均します。Z、さっき 0.25 を使いましたから、0.25 のXバー j ＋0.75 の μ というようになります。このZのことを信頼度といいます。

これが信頼性理論による料率の算出です。何かよさそうですよね。でも、よさそうだからといって、使ってしまうというわけにもいかないですよね。大体この算出法はよさそうですけれども、どのように正当化できるのか、それから、信頼性のＺはどのように決めたらいいのかということを考えなければいけないではないですか。どう考えたらいいでしょうか。



いろいろな考え方があると思いますけれども、信頼性理論の目的というのは、経験に基づいて翌年度の保険料率を算出します。いいですよね。これを統計学の言葉で言ったらどうなるかというと、過去のデータ、Ｘ１、Ｘ２、Ｘｎ、さっきはｎが５でしたけれども、これに基づいて翌年度の事故率、Ｘｎ＋１、これを予測するという、予測の問題なのです。

だとしたら、最適な予測を使えばいいですよね。予測値をｍティルドＸとします。Ｘは過去のデータです。過去のデータの関数として、予測値を作ります。一番良いのは何かというと、ターゲットのＸｎ＋１との、例えば二乗リスクを最小化するようなものです。この二乗リスクを最小化するものは、条件付き期待値です。だから、最適な予測は条件付き期待値です。これを計算すればいいわけです。どのように計算しますか。

信頼性理論と予測

- 信頼性理論の目的は「経験に基づいた翌年度の保険料率の算出」
- 統計学の用語でいえば，過去のデータ $\boldsymbol{X} = (X_1, X_2, \ldots, \quad X_n)$ に基づく次年度の事故率 X_{n+1} の予測.
- 予測値を $\widetilde{m}(\boldsymbol{X})$ とするとき，2 乗リスク

$$\mathrm{E}[(X_{n+1} - \widetilde{m}(\boldsymbol{X}))^2]$$

を最小化する最適な予測は

$$\widetilde{m}(\boldsymbol{X}) = \mathrm{E}[X_{n+1}|\boldsymbol{X}]$$

8

このときに、よく考えてみたら、保険契約者の中にはタロウさんのタイプとハナコさんのタイプと、両方いるわけではないですか。異質性があるわけです。そのような人たちを相手にしているのです。だったら、その異質性を表したいですよね。それを確率変数θで表すことにしましょう。

異質性θが与えられたときに、将来のデータと、それから過去のデータが独立だとすると、これをずっと計算していくと、最適な予測、mティルドXというのは、一番下の表現、μ(θ)というのが一番下にありますけれども、θが与えられたときの次の期の、来年の事故率の予測値を、θに関して平均を取ったものです。それがmティルドXになります。

このmティルドXが一番下になりますけれども、ちょっと一番下は分かりにくいので、ちょっと書き直してみます。異質性θをしようとしたときに、過去のデータXというのが条件付き密度 f(x|θ)、いわゆるデータのモデルに従うとします。それからθの密度を p(θ)とします。そうすると、さっきのMティルドXという、一番いい最適な予測値というのは、真ん中の辺りに積分記号で書いてあるものになります。ここで p(θ|x)いうのは、Xが与えたときのθの条件付き分布です。これのことをベイズでは事後分布といいます。事後分布と別に言ってもいいですし、例えば条件付き分布なのですね。ですので、そこにあるような表現です。

信頼性理論と異質性

- 個別契約者のリスク特性の「異質性」（太郎のタイプか，花子のタイプか）が（観測できない）確率変数 θ で表されるとする．
- 異質性 θ が与えられると，X_{n+1} と $\boldsymbol{X}$ が独立であるならば，最適な予測は

$$\begin{aligned}\widetilde{m}(\boldsymbol{X}) &= \mathrm{E}[X_{n+1}|\boldsymbol{X}] \\ &= \mathrm{E}[\mathrm{E}[X_{n+1}|\theta, \boldsymbol{X}]|\boldsymbol{X}] \\ &= \mathrm{E}[\mathrm{E}[X_{n+1}|\theta]|\boldsymbol{X}] \\ &= \mathrm{E}[\mu(\theta)|\boldsymbol{X}]\end{aligned}$$

ここで，

$$\begin{aligned}\mu(\theta) &= \mathrm{E}[X_{n+1}|\theta] \\ &= \text{ある特定の異質性 } \theta \text{ に対する翌年の事故率の予測値}\end{aligned}$$

9

p(θ)のことを事前分布という言い方をします。だから、m ティルドXという最適な予測値というのは、結局、各異質性θに対する翌年の事故率の予測値を、θの事後分布で平均したものなのです。

これはベイズの世界です。だから、これを計算してあげればいいのですよ。これを計算してあげればいいのですけれども、最初にこのようなものを考え出した人たちというのは、これを直接計算しようとはしていません。どうしてかと言うと、これは式で書けば簡単ですけれども、でも積分が入っているではないですか。その積分は一般的には難しいのです。だから、ここでどうするかと言うと、２通りのアプローチがあって、積分が簡単にできるような世界で考える。このようなものは、昔、ベイズでよくやっていました。共役分布というものを使います。でも、それはあまりよくないですよね。だって数学的に簡単だから、そういってしまうのは、よくないではないですか。

ベイズ！

- 異質性 θ を所与とするときの過去のデータ $\boldsymbol{X} = (X_1, \ldots, X_n)$ は，条件付き密度 $f(\boldsymbol{x}|\theta)$ に従うとする．また，θ の密度を $p(\theta)$ とする．
- このとき

$$\begin{aligned}\widetilde{m}(\boldsymbol{X}) &= \mathrm{E}[\mu(\theta)|\boldsymbol{X}] \\ &= \int \mu(\theta) p(\theta|\boldsymbol{X}) d\theta\end{aligned}$$

であり，

$$p(\theta|\boldsymbol{X}) = \frac{f(\boldsymbol{x}|\theta)p(\theta)}{\int f(\boldsymbol{x}|\theta)p(\theta)d\theta}$$

- ベイズ統計学では，異質性の密度 $p(\theta)$ を事前密度，データが観測された後の異質性の条件付き密度を $p(\theta|X)$ を事後密度という．
- $\widetilde{m}(\boldsymbol{X})$ は，各異質性 θ に対する翌年の事故率の予測値 $\mu(\theta)$ を θ の事後分布によって平均したもの

10

特に保険の世界ではデータは重要ですから、この f(x| θ) というのがどのような損失分布に従うかというところを、重要ではないですか。そこを大事にしたいとしたら、もう一つアプローチがあって、それはどうするかと言うと、m ティルド X という最適な予測値が線形モデルというように考えてしまいましょうと。簡単な形でしましょうと。 a ＋ b Xバー。ちょうど過去のデータから将来を予測するときに、われわれはよく分からないときは回帰モデルを考えますよね。あれは線形モデルです。あれと同じです。線形モデルで考えましょうと。

線形ベイズ推定

- 予測量のクラスを「線形モデル」 **過去のデータの平均**

$$\widetilde{m}(\boldsymbol{X}) = a + b\bar{X}$$

に限定する

- このとき，2 乗リスク $\mathrm{E}[(X_{n+1} - \widetilde{m}(\boldsymbol{X}))^2]$ を最小にする予測量は

$$\widetilde{m}(\boldsymbol{X}) = z\bar{X} + (1-z)\mu$$

という形をとる．ここで

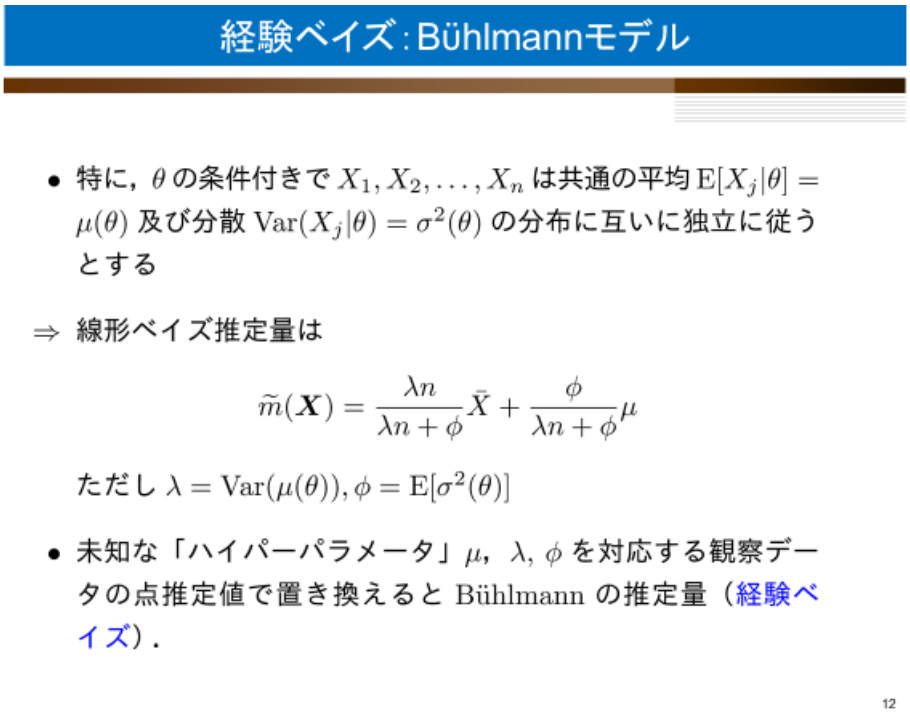
$$\mu = \mathrm{E}[\mu(\theta)], \quad z = \frac{\mathrm{Var}(\mu(\theta))}{\mathrm{Var}(\mu(\theta)) + \mathrm{E}[\mathrm{Var}(\bar{X}|\theta)]}$$

11

そうすると、二乗リスクを最小にするような予測量というのは、zx バー＋１マイナス Z μ という形となるわけです。この μ や Z というのは、そこにある形で与えられます。だから、線形モデルにすると、信頼性理論で言っているような形が出てくるわけです。特に、もうちょっと条件を加えると、ビュールマンモ

デルというものが出てきます。

ビールマンモデルはパラメータのパラメータのような、ハイパーパラメータというものを持っていて、そこをデータから推定するという形を取ります。これを経験ベイズといって、非常にプラクティカルな方法として、よく使われているということになります。これがベイズの非常に成功している、しかも分かりやすい例ではないかと思います。



この辺で、ベイズの話をこの後にしますけれども、ベイズというのは何かということを、少しまとめておきたいのですね。どうしてかと言うと、私が若いころは、もうずっと前ですけれども、このベイズ対頻度法の哲学的な論争がずっとあって、あたかも神学論争のようなことが行われていたのです。

そもそも、元々の問題は同じではないですか。統計的推測の問題は一体何かと言うと、データＸｎがあって、これがある密度関数ｇに従うわけです。データからこのｇをどのように推定するかということが、統計の問題なわけです。

頻度論はどうするかと言うと、このｇというのを直接推定するノンパラメトリックな方法もありますけれども、普通はｇというのは、ある統計モデルに従う、ある統計モデルに属しているというように仮定します。

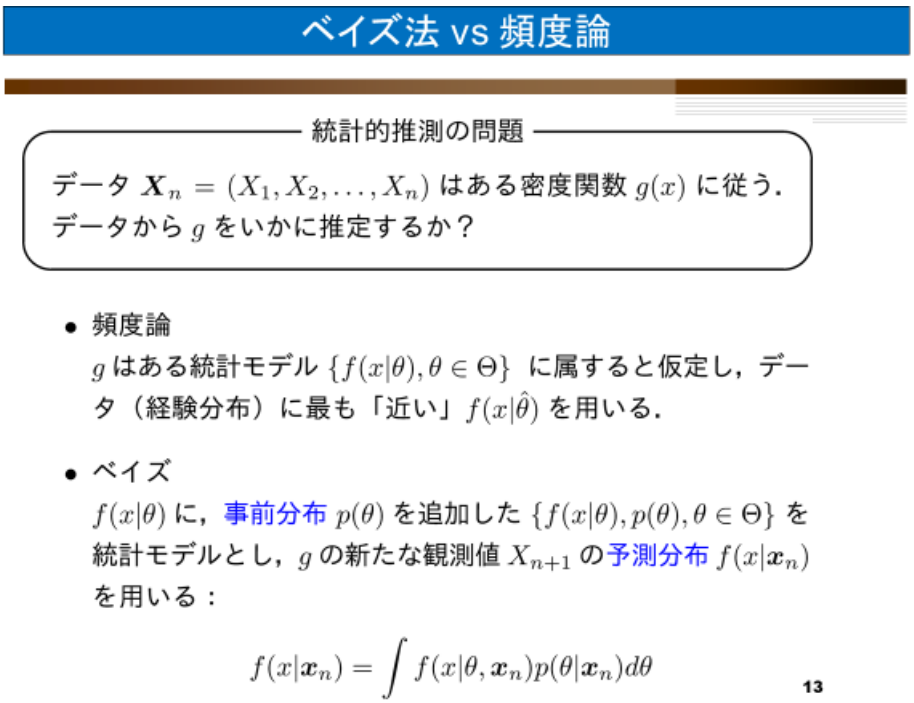
この中でデータ、データというのは経験分布ですよね。これに一番近いような統計モデルを選択します。このときに近いというのが問題になりますよね。何が近いのか。特にカルバック・ライブラー・ダイバージェンス、カルバック・ライブラーの距離というのがありますよね。あれを使うと、θハットが最推定値になります。頻度論というのは、統計モデルを考えて、そのデータに一番近いようなものを使いましょうという話です。

では、ベイズ。何ですか、ベイズというのは。ベイズは頻度論と全く違うものでしょうか。ベイズというのは、解こうとしている問題は同じなのですね。ベイズはどうやるかと言うと、同じように統計モデルを使います。ｆ(ｘ|θ)を使います。ただし、これに事前分布という、ｐ(θ)のモデルを追加します。だから、ベイズというのは、ｆ(ｘ|θ)と ｐ(θ)がペアになっている、そのようなモデルを使います。

では、そのｇをどのように推定するかと言うと、ｇから、ｇの新しい観測値、Ｘｎ＋１とします。それ

の予測分布というものを作ります。予測分布というのは、ただ単にＸｎが与えられたときの、Ｘの条件付き分布です。それは計算すると、その一番下にあるような式になります。

ベイズというのは、だから、確かに事前分布を使いますけれども、事前分布のモデルを使っているだけで、そのような意味では頻度論を拡張したモデルを使っているということが、私が言いたいことなのですね。

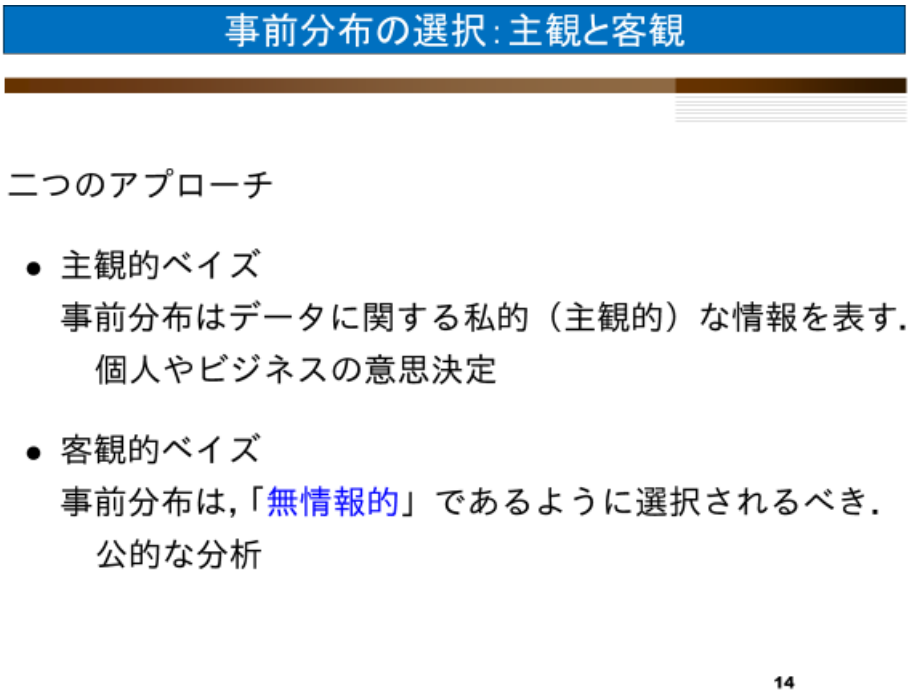


とはいえ、「じゃ、ｐ(θ)って何だよ」という話になりますよね。私からすると、「ｆ(ｘ| θ)だって何だよ」ということなのですよね。ｆ(ｘ| θ)だって主観的に選んでいるわけでしょう。ｐ(θ)だって主観的に選んで何が悪いのか、という感じはするのですけれども、そうは言っても、ｆ(ｘ| θ)の方は何とかデータのモデルなどは見えるので、少し違いますよね､ｐ(θ)とは。

この ｐ(θ)、事前分布をどのようにして選ぶかに対しては、二つのアプローチがあります。主観的ベイズというものと、客観的ベイズ。客観的ベイズはちょっと何か変な感じがするかもしれませんけれども。主観的ベイズというのは、まさにこの主観確率、事前分布は主観確率ですよというような立場です。

でも、事前分布というのは、データに関するプライベートな、必ずしも公的ではないという意味で、主観的な情報を表しています。だから、個人の意思決定やビジネスの意思決定では、主観的ベイズというのはいいかもしれません。さっきの自動車保険の例で言えば、自動車保険には過去のデータがあるわけですから、それを事前分布としてまとめて利用するというのは、当然やるべきことですよね。

もう一つ客観的ベイズというものがあります。このとき事前分布は、ちょっと今日は説明しませんけれども、情報があまりないようにします。これがある意味、公平だという意味なのですけれども。客観的ベイズというのが、最近、比較的よく使われるようになってきています。



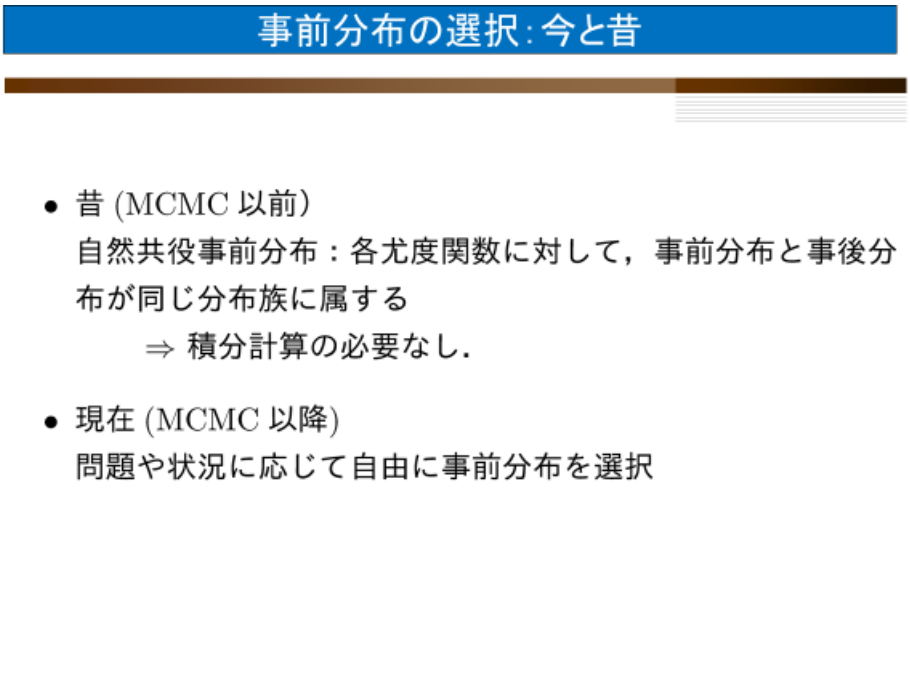
もう一つ、事前分布の選択について、今と昔を比べます。さっきは主観と客観ですけれども。昔というのは、「じゃ昔っていつですか」という話ですけれども、MCMC以前です。ベイズというのは、MCMCというシミュレーション法が使われるようになった前と後で、もう革命的に変わりました。

昔はベイズというのは、「まあいいけど、結局計算できないでしょう」という話だったのですね。ですので、ベイズを実際使うときは、自然共役事前分布という、事前分布と事後分布が同じ分布属に属するようなものに、限って使っていました。それは積分の計算の必要がありません。ある意味、リーズナブルですよね。

今はどうかと言うと、今はMCMC以降です。MCMCというのは、マルコフ連鎖モンテカルロ法という方法なのですけれども。これは元々、実験物理の世界で使われたような、シミュレーションで使われた技術なのですけれども。これを使うことによって、次元が非常に高い、さっき Lasso の話や、あるいは機械学習の話などが出ていましたけれども、機械学習の一つの特徴は、次元が非常に高いものを使うということです。変数の数が大きいということですね。そのような場合でも、積分の計算というか、乱数が発生して、積分の計算ができるようになったというのが、MCMC以降です。

MCMC以降は何がいいかと言うと、問題や状況に応じて、自由に事前分布を選択できるようになったわけです。昔は技術的な理由で「これしか使わないよ」と。例えば、さっきの自動車保険の例なら、モデルの形を線形に制約します。今はそれは必ずしもする必要はないです。もちろん実際問題としては、そうした方が便利かもしれませんけれども、事前分布は自由に選択できます。それが大きな違いなのです。

なぜベイズが今これだけはやっているかというと、まさにMCMCのおかげと言っても過言ではないのですね。時代でいうと、多分 2000 年、21 世紀になってからは、ベイズの世界というものが、今の統計学の世界に、ばっと出てきたという感じです。もちろんMCMCがすべてを解決するわけではないのですけれども、非常に大きなインパクトを与えているというのが、現在の状況ではないかなと思います。



次の話題なのですけれども、これは私自身がやっていることを、少しご紹介させていただきたいと思います。要介護状態別死亡率の予測ということです。この辺も皆さんはご専門家の方なので、あえて説明することもないかとは思いますけれども、長寿、平均余命がどんどん伸びていて、長寿リスクというのが非常に問題になっています。長寿であるということは、同時に要介護状態になるというリスクがあるわけで、だから、ロングタームケア・リスクというのが、大きな注目の的になるわけです。だから、死亡率というものを考えるときも、要介護状態に応じた死亡率というものを計算したいと、将来の死亡率を知りたいと。

だけれど、実際に見てみると、国民レベルの個票のデータでは、ある地域に限って言えば、そのようなデータが利用可能なケースもだんだん増えてきているようですけれども、国民レベルの健康状態別の死亡データというのは公開されていないという中で、どうやって要介護度別の死亡率を推定しようかということです。

要介護度別死亡率

- 過去の予想を上回る平均余命の伸長が世界的に進行中
- 平均余命の伸長は，長寿リスクとともに，要介護状態となる可能性ー長期介護リスク (LTC)ーの増大を意味する．

⇒ 将来の死亡率がいかに要介護状態と関係しているかの把握が重要

- しかし，(国民レベルの) 健康状態別の死亡データは公開されていない．

17

だから、目的としては、要介護状態別の部分集団に対する死亡率モデルを開発したい。だけれど、死亡率のデータ、死亡データはないです。あとは平均余命がどんどん伸びているというのも考えなければいけないです、ということが課題です。

課題

目的

要介護状態別の部分集合に対する死亡率モデルの開発．
ただし

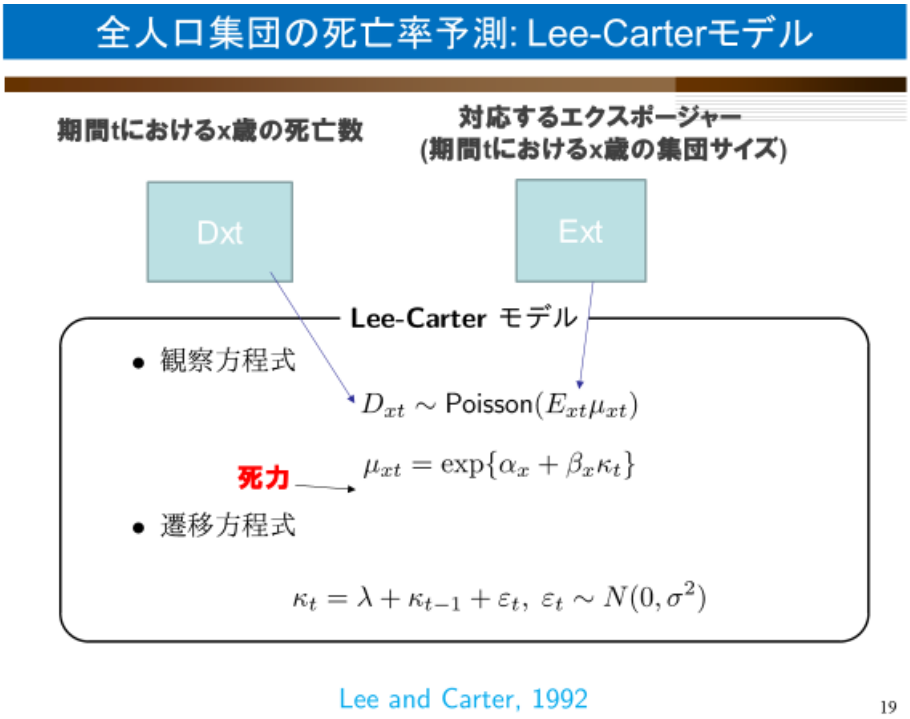
1. 要介護状態別の死亡率データは存在しない
2. 平均余命の伸長を考慮

18

これに対して、死亡率の予測と言うと、一番よく使われる代表的なモデルは、リー・カーター・モデルと呼ばれるものですよね。リー・カーター・モデルは、ここに書いてあるとおり、少しだけ復習をしますと、データがあって、Ｄｘｔというのは死亡のデータです。Ｅｘｔがエクスポージャーです。

リー・カーターは二つの方程式があって、一つはそのＤｘｔのデータのモデルで、これは観察方程式といいます。Ｄｘｔというのはポアソンで平均がＥｘｔ、エクスポージャー×死力。この死力の形がいわゆるリー・カーター・モデルの形です。ｘが年齢を表わしていて、ｔが期間です。だから、μｘｔは基本的

にはαｘというパラメータがμｘｔを決めるのだけれども、それからずれている部分というのが、βｘκｔで表されていて、κｔは一般的にどんどん死亡率が下がっているということを表わすパラメータで、ですので、このκｔについて、下にあるような状態方程式で書いてありますけれども、ランダムウォークを仮定しています。ただし、κｔもαｘもβｘも観測されないという、このようなモデルになっています。

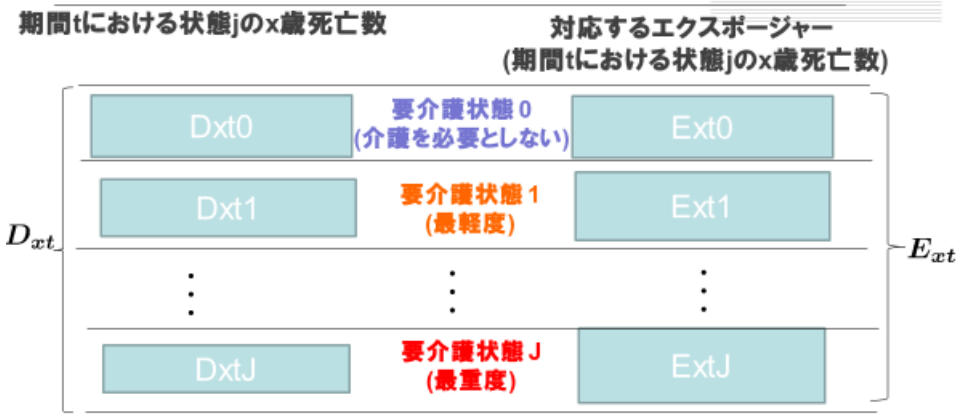


これはＤｘｔとＥｘｔがあれば、推定できますよね。だから、要介護状態による部分集団に分けて、同じことをやれば、できるはずですよね。だけれども、問題は状態０というのが要介護を必要としない状態で、1～Ｊと、Ｊが一番重度が高いような、このように部分集団に分けたとして、各部分集団について、今のリー・カーター・モデルを当てはめてやればいいではないですか。

でも、Ｄｘｔ０、Ｄｘｔ１、Ｄｘｔｊというこの右側の部分がないわけですよね。その中でどのように推定したらいいのかという問題を考えてみました。どうやるかと言うと、モデリング。モデリングに頼りましょうと。死力をモデリングします。

一番上に書いてあるのは、やはりどのような部分集団についても、リー・カーター的な構造があるでしょう。けれども、状態によって違うので、その違いを表わすようなパラメータを入れて、それをηｊとしています。ただし、ηｊというのはだんだんｊが大きくなるにつれて、大きくなるような制約を置いています。このような制約を置いた上で、推定をしましょうと。さらに、一個一個の死力の和が、加重平均が全集団の死力になるというように仮定を置きましょうと。このときの加重は人口の総対比率です。

要介護状態(健康状態)による部分集団



- 各部分集団について，D_{xtj} と E_{xtj} が利用可能ならば，各状態 j について死亡率予測モデルをフィットすればよい．
- しかし，しばしば，部分集団の死亡データ D_{xtj} は利用不可能．

20

これを仮にミクスチュア・リー・カーター・モデルと呼んで、これをさっきと同じような方式で推定するということを試みています。これは、これだけを見ると簡単そうですけれども、さっきの制約を表わすような事前分布を入れたりしなければいけないので、結構それなりに大変ということになります。

仮定:死力のモデリング

- 各部分集団の死力は

通常の Lee-Carter に η_j を追加
(Li and Lee, 2005)

$$\mu_{xtj} = \exp\{\gamma_x + \eta_j + \beta_x \kappa_t\}$$

ここで，η_j は（不）健康要因であり，

$$0 = \eta_0 < \eta_1 < \ldots < \eta_J$$

- 全集団の死力は

$$\mu_{xt} = \sum_{j=0}^{J} w_{xtj}\mu_{xtj} = \sum_{j=0}^{J} w_{xtj} \exp\{\gamma_x + \eta_j + \beta_x \kappa_t\}$$

ここで：

$$w_{xtj} = \frac{E_{xtj}}{E_{xt}} = \text{期間 } t \text{ における } j \text{ 集団の } x \text{ 歳人口の相対比率}$$

21

ということで、最尤法でやっても可能なので、実際に試みましたけれども、結構複雑で、なかなかいい解が出てきませんでした。それもあって、ベイズ法を使って、推定しました。ベイズ法のいい所はパラメータ不確実性という、その予測をするときに、モデルを推定したということの誤差まで含めてリスクを把握しなければいけないという部分も含めて、予測できるということで、ベイズ法がいいと一般に言われていて、ケアンズやブレイクなどはそのような論文を書いていますし、ベイズ法で死亡率の推計をしようというような試みは、いろいろあるというわけです。というわけで、そのような方向に沿って、ここでもベイズ法で推定したということです。

提案手法: Mixture Lee-Carter モデル

Mixture Lee-Carter model

- 観察方程式

$$D_{xt} \sim \text{Poisson}(E_{xt}\mu_{xt})$$

$$\text{with } \mu_{xt} = \sum_{j=0}^{J} w_{xtj} \exp\{\gamma_x + \eta_j + \beta_x \kappa_t\}.$$

- 状態方程式

$$\kappa_t = \lambda + \kappa_{t-1} + \varepsilon_t,\ \varepsilon_t \sim N(0, \sigma^2)$$

η_j がすべて等しい：

$$\eta_0 = \eta_1 = \eta_2 = \ldots = \eta_J = \eta,\ \text{say},$$

とき，$\alpha_x = \gamma_x + \eta$ とすれば，通常の Lee-Carter モデル.

22

死亡率予測モデルとパラメーター不確実性

- Lee-Carter モデルの通常の推定（最尤法）では，推定パラメータの誤差（パラメータ・リスク）を考慮できない.

⇒ 将来の死亡率のリスクを過小評価
Cairns et al. (2006) and Blake et al. (2008)

⇒ ベイズ法による考察
Czado et al. (2005)， Pedrosa (2006)， Kogure, Kitsukawa and Kurachi(2008)，Kogure and Kurachi (2010), etc.

- 本報告では，ベイズ法によって提案した混合 Lee-Carter モデルの推定を行う.

23

実際のデータなのですけれども、日本の介護保険のデータを応用しました。介護保険というのは、スタートした時点は要支援と、それから要介護１～要介護５の６段階に分かれていたのですけれども、途中で、2006 年に要支援の所が１と２に分かれました。その時に、要介護１とのかねあいで、要支援１・２を設定したということなので、ここでは 2001 年～2005 年までは状態１を、そこの表に書いてあるとおり、要支援＋要介護１として、2006 年以降は要支援１～２と要介護１を一つにまとめて状態１として、それ以外については、状態２～状態５というのは、それぞれ要介護２～要介護５に対応するというカテゴライズをして、状態０～状態５のときの死亡率の予測を行いました。

我が国介護保険データへの応用

- 介護保険制度は 2000 年 4 月に開始．開始時は要介護の状態に応じて受給者は 6 段階（要支援，要介護１ ～ 要介護５）に区分．
- 2006 年度から，要支援の区分は（要介護 1 との調整も行い）要支援 1 と要支援 2 に細分化．このとき経過的要介護が設けられた．

本研究では，全期間のデータの整合性のために，以下のように状態を設定：

期間	状態 0	状態 1	状態 2	状態 3	状態 4	状態 5
2001～2005	非要介護	要支援+要介護 1	要介護 2	要介護 3	要介護 4	要介護 5
2006～2008	非要介護	要支援 1～2+要介護 1 +経過的要介護	要介護 2	要介護 3	要介護 4	要介護 5
2009～2014	非要介護	要支援 1～2+要介護 1	要介護 2	要介護 3	要介護 4	要介護 5

以下では，状態 0～状態 5 の 6 状態の死亡率予測を行う．

24

要介護者受給者数

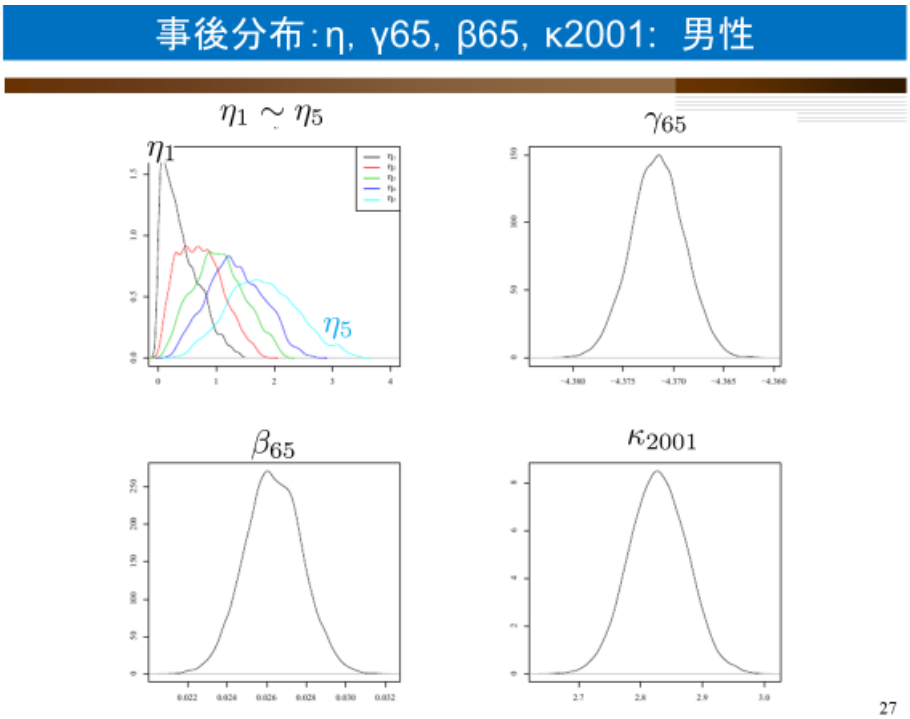
- 要介護（要支援）状態区分・性・年齢階級別の受給者数（『介護保険事業状況報告』）を要介護部分集団の人口サイズとした．
- 65 歳～89 歳までの，受給者数は 5 歳区間の年齢区分で与えられる．

$$65\sim69,\ 70\sim74,\ 75\sim79,\ 80\sim84,\ 85\sim89,$$

- これらの年齢区分の受給者数の線形補間を行うことにより，各年齢における受給者数を求め，$\{E_{xtj}\}$ を作成

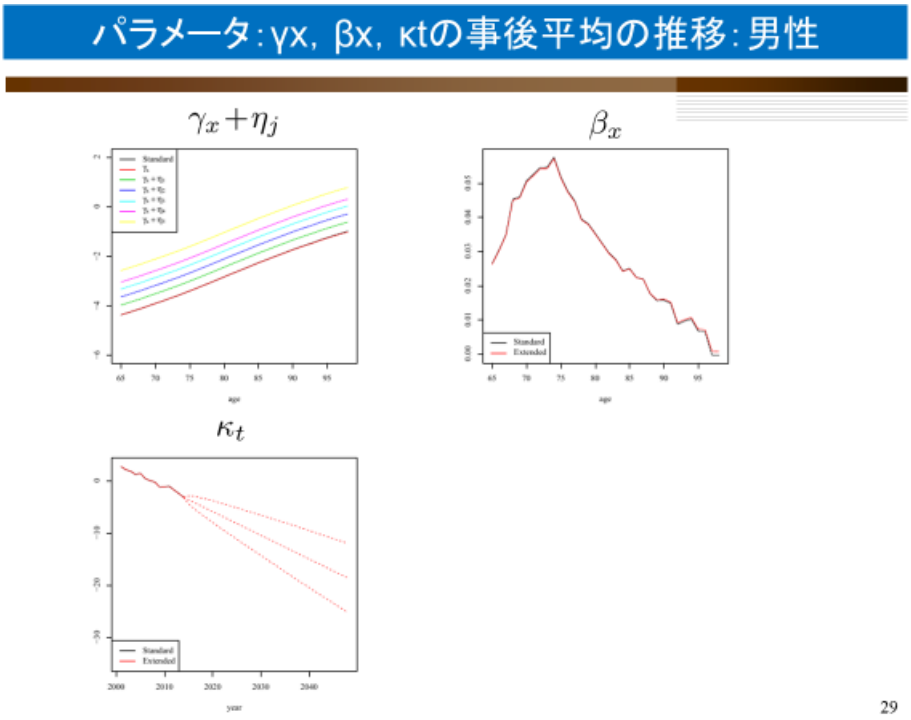
26

ベイズを使って、ＭＣＭＣを使って、予測したのですけれども。これはスキップしますね。これもちょっとスキップします。これが結果です。η１、η５というのは、さっき言った要介護状態を表わすパラメータです。η１が一番軽度で、η５が一番重度です。そうすると、ここにあるように、これはη１～η５の事後分布なのですけれども、だんだん重度が上がるごとにシフトしているということが分かると思います。他のパラメータについても、とてもきれいに推定されているということになります。



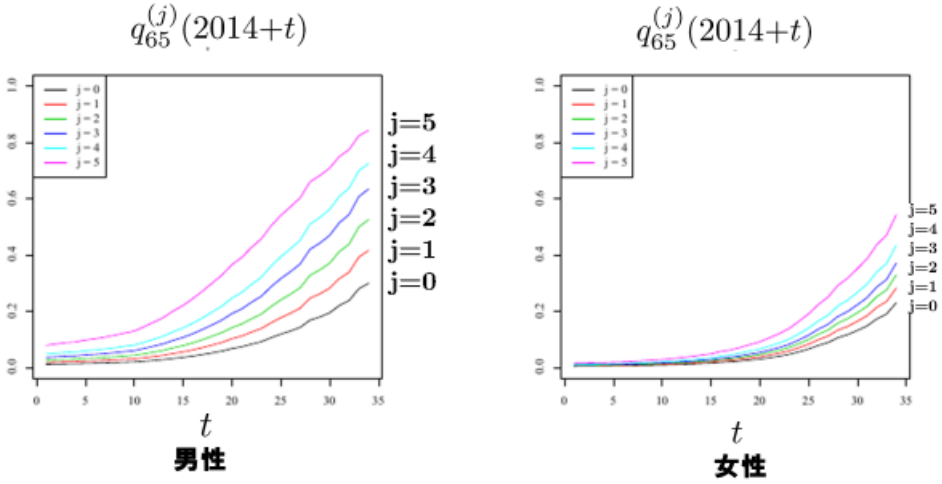
これが実際の推定結果の要約統計量なのですけれども、これを見てもη１～η５、これは平均値ですけれども、だんだん大きくなっているということで、比較的うまく推定できているのではないかと考えています。

これは各パラメータの時間的な推移、あるいは年齢に対する推移を表わしたものです。



でも、これ自体はパラメータなので、本当に関心があるのは将来の死亡率ですよね。これを予測した結果、将来の死亡率を予測するためには、将来どうなるかという部分も予測しなければいけないのですけれど、そこも予測分布というものを使って、このκのところですね。ランダムウォークのところを予測して、死亡率の推定をしてみましたというのが、この結果です。

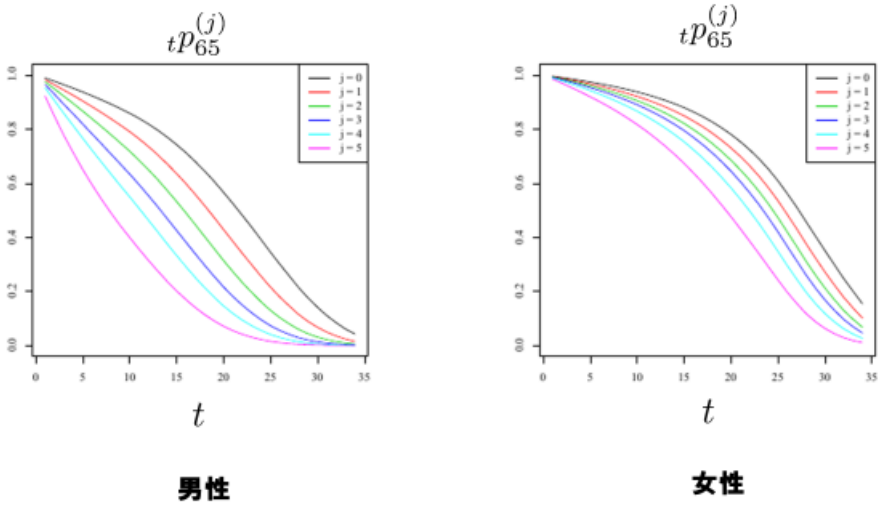
35年後までの将来死亡率の予測



31

これまでの結果というのは男性だけだったのですけれども、ここでは女性の結果も書いています。2014年に65歳であった人の将来の死亡率というのを、j＝0というのは支援を必要としない人ですね。j＝5が一番重度が高い人。その死亡率はこのようになると。このモデルが正しければ、このようになるということです。同じように、将来の生存率の予測、生存関数の予測もしています。これも65歳の人の将来の生存率がどうなるかというのを、各要介護状態別に推定したというものであります。

35年後までの将来生存率の予測



33

ただし、ちょっと戻りますけれども、これはちょっとまだプリミティブというか、プレリミナリーな研究なので、状態が全然変わらないまま死ぬという、そのように非現実的な仮定で推定しているので、実際にはその間で状態を遷移するというところも、もちろん推定して、考慮しなければいけないのですけれども。そこのところまではまだやっていないので、そのような若干非現実的な想定で計算されている部分も

ありますという、一定の考慮はあります。

このような形で、たとえデータがなくとも、うまくモデリングをして、ベイズを使うことで、比較的、何となくもっともらしい、この絵を見てもっともらしいと言っても「え？」という感じかもしれませんけれども、でも、何かもっともらしくないですか。死亡率、生存率。何か、もっともらしくないですか。もっとぐちゃぐちゃと出てくるかと思ったのですけれども、うまく出てくるのですね。そのように制約していると言われれば、そうなのかもしれませんけれども。でも、モデリングとはそのようなものですよね。ちょっと開き直りですかね。このような形で推定できますと。

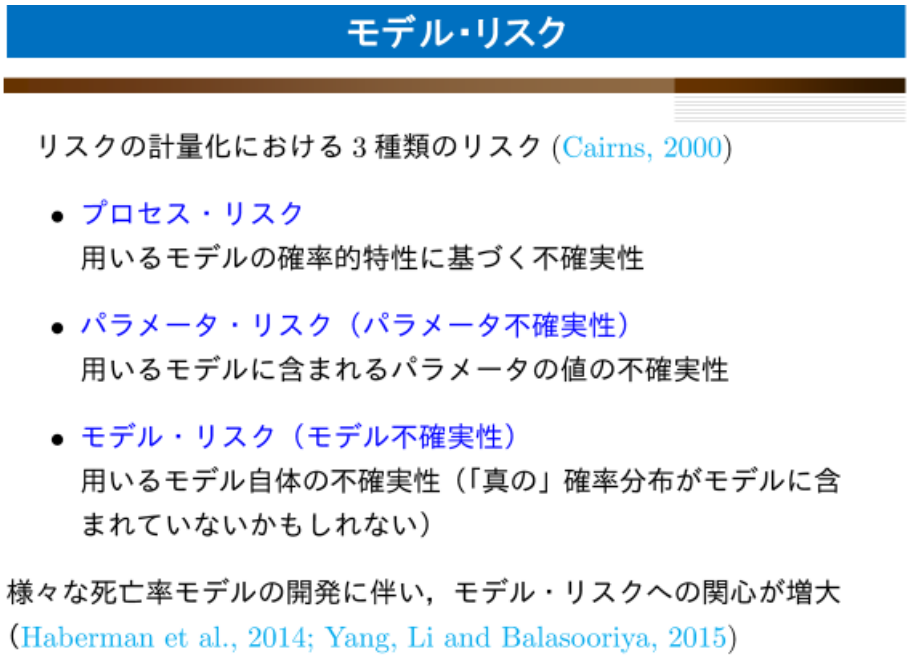
時間がちょっとなくなってきてしまったのですけれども。最後はちょっと駆け足になってしまいますけれども。ちょっと最後に言いたいといいますか、お話ししたい。２、３分いいですか。大丈夫ですか。

モデルリスクという、皆さん、言葉を聞いたことがありますかね。これはちょっと英語が小さくて読めないかもしれませんけれども、ボックスという有名な統計学者がいて、ボックスは有名なフィッシャーの娘婿なのですけれどもね。フィッシャーには娘さんが２人いて。どうでもいいかもしれませんけれども。長女と結婚したのがこのボックスで、次女と結婚したのが、私がアメリカでお世話になったアンスコム先生です。

ボックスがそのような、まさに統計学者の本流です。彼が言っている有名な言葉に、このような言葉があって、「すべてのモデルは間違ってる」。すごいですよね。「すべてのモデルは間違ってる。だけど、中には役に立つものもある」というように、ボックスは言っています。

モデルの否定のように見えますけれども、モデルの本質をとても言い当てていますよね。だって、モデルというぐらいだから、現実ではないです。間違っているに決まっています。でも、統計学で何かを語るとき、われわれはモデルが正しいかのように考えて、語ってきたことがとても多いではないですか。それがよくないというのが、例えば赤池先生のＡＩＣが出てきて、気づかされるわけですけれども。やはり、そこはとても重要だなというのは、私自身は思っています。

そのリスクを計量化するときに、ケアンズというスコットランドにいる有名なアクチュアリーの研究者がいますけれども、彼は三つのリスクを考えなければいけないと言っています。プロセスリスクは本質的な、そもそものリスクですね。それから、パラメータリスク。モデルを使うわけですから、そのモデルを推定したとして、その推定のリスク、パラメータの不確実性。さらにモデル。どのようなモデルを使ったかという、モデルリスクもあるのだということなのです。



ＡＩＣを使ってモデルを選択する。それでいいような気もしますけれども、ＡＩＣでモデルを選択するということ自体が、そもそも不確実性をはらんでいるわけです。どのようなモデルを使ったかというモデルリスクも、リスクの計量化の中に入れなければいけないよということを、ケアンズは言っています。

特に、私はこの死亡率モデルということに興味を持って、このところやっているのですけれども、最近、非常に多くの死亡率モデルというものが開発されているのですね。その中でモデルリスクというものの関心がとても増えています。死亡率のモデルを作って、それをもとに長寿リスクのデリバティブズのようなものを考えるわけですけれども、そのときに、モデルリスクというものも考慮に入れて、長寿デリバティブのプライシングをやらなければいけないというようなことを、最近イギリスのアクチュアリーの人たちの中には、そのようなことを言っている人がいるらしいです。

このような死亡率モデルというのは、状態空間モデルというもので記述できます。さっきのリー・カーター・モデルも状態空間モデルなのですね。この y の所の f という、それから ϕ の所のモデリングをいろいろ考えることで、例えば観測方程式として、リー・カーター・モデル以外にＡＰＣ、Age-Period-Cohort や、ＣＢＤ、それからリー・アンド・リーというモデルなど、いろいろモデルは考えています。

それから遷移方程式の方も、さっきのモデルのランダムウォークは、確率的トレンドモデルと言われますけれども、定常モデルや、確定的トレンドモデルや、あるいはもっと複雑な時系列モデルなど、いろいろあるわけです。

死亡率モデルと状態空間モデル

- 確率的死亡率モデルは状態空間モデルとして記述できる（Fung, Peters and Shevchenko, 2016）.

$$\begin{cases} \text{観測方程式} & \boldsymbol{y}_t \sim f(\boldsymbol{y}_t|\boldsymbol{\phi}_t, \boldsymbol{\theta}) \\ \text{遷移方程式} & \boldsymbol{\phi}_t \sim f(\boldsymbol{\phi}_t|\boldsymbol{\phi}_{t-1}, \boldsymbol{\theta}) \end{cases}$$

ここで，$\boldsymbol{y}_t$ は時系列データであり，$\boldsymbol{\phi}_t$ は潜在変数

Lee-Carter モデルでは，$\boldsymbol{y}_t = (\log m_{x_{\min}t}, \ldots, \log m_{x_{\max}t})'$, $\boldsymbol{\phi}_t = \kappa_t$

- 主要な代替モデル
  - 観測方程式
    APC モデル，CBD モデル，Li and Lee モデル
  - 遷移方程式
    確率的トレンドモデル，定常モデル，確定的トレンドモデル

36

このようなものをどうやって予測するかというときに、ＡＩＣを使うのも確かにいいかもしれないけれども、モデル選択はＡＩＣでやりましょうというような形で出てきますけれども、ベイズだとモデル選択自体が、ベイズの枠組みの中に入っているということ、これは言いたかったことなのです。周辺尤度を使うということになるのですね、モデル選択をする場合には。モデルの事後確率というのを求めて、それは周辺尤度に比例しているので。

モデル選択：周辺尤度

代替モデルを $\mathcal{M}_1, \ldots, \mathcal{M}_J$ とする.

- 各モデルはデータ $\mathcal{D}$ のモデルとパラメータ $\boldsymbol{\theta}$ の事前分布の組

$$\mathcal{M}_j = \{(f_j(\mathcal{D}|\boldsymbol{\theta}), f_j(\boldsymbol{\theta})), \boldsymbol{\theta} \in \Theta_j\}$$

- 各モデルの事前確率は等しいと仮定：

$$\Pr(\mathcal{M}_j) = \frac{1}{J} \quad (J \text{ 個のモデルのうち一つは真})$$

- モデル $\mathcal{M}_j$ の事後確率は

$$\begin{aligned} \Pr(\mathcal{M}_j|\mathcal{D}) &= \frac{f(\mathcal{D}|\mathcal{M}_j)\Pr(\mathcal{M}_j)}{f(\mathcal{D})} \\ &\propto f(\mathcal{D}|\mathcal{M}_j) = \int_{\boldsymbol{\Theta}_j} f_j(\mathcal{D} \mid \boldsymbol{\theta}) f_j(\boldsymbol{\theta}) d\boldsymbol{\theta} = \text{周辺尤度} \end{aligned}$$

⇒ 周辺尤度を最大にするモデルが望ましい.

37

その周辺尤度というのは、実は展開していくと、オッカム・ファクターと呼ばれますけれども、その複雑なモデルをちょっと罰則するような項が自動的に含まれています。ここのオッカム・ファクターの所をさらに展開すると、いわゆるＢＩＣというものになっていきます。

周辺尤度とオッカムの剃刀

- 事後分布が正規分布で近似できるとすると：

$$\text{周辺尤度} = f(\mathcal{D} \mid \mathcal{M}_j) = \int_{\Theta_j} f_j(\mathcal{D} \mid \boldsymbol{\theta}) f_j(\boldsymbol{\theta}) d\boldsymbol{\theta}$$
$$\approx \underbrace{f_j(\mathcal{D} \mid \widehat{\boldsymbol{\theta}}_j)}_{\text{データに適合させた尤度}} \underbrace{f_j(\widehat{\boldsymbol{\theta}}_j)|\Sigma_{\widehat{\boldsymbol{\theta}}_j}/2\pi|^{1/2}}_{\text{オッカム・ファクター}}$$

ここで，$\widehat{\boldsymbol{\theta}}_j$ は最大事後確率推定値であり，

$$\Sigma_{\widehat{\boldsymbol{\theta}}_j} = \left(-\frac{\partial^2 \log [f_j(\mathcal{D} \mid \boldsymbol{\theta}) f_j(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right)^{-1} \Bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_j}$$

- 「オッカム・ファクター」はモデルの複雑さに対する罰則項.

38

というわけで、ベイズのモデリングというのは、モデルリスクにも対応できるのではないかということを最後に付け加えて、私の発表としたいと思います。どうもありがとうございました。

まとめ／結論

- ベイズ法の応用として「信頼性理論」と「要介護度別死亡率の予測」を紹介した.
  - 前者は，各個人の限定されたデータと関連する外部のデータを組み合わせるために利用．事前分布は各個人の異質性を表す．主観的ベイズ.
  - 後者は，パラメータ・リスクやモデル・リスクを取り込むためにベイズ法を利用．事前分布はモデルの一部．客観的ベイズ.
- 「計算ベイズ技術」と「客観的ベイズ」の普及によって，ベイズ統計学はデータサイエンスの「実践」の重要な柱
- リスク管理の実践においても，ベイズ統計学のより一層の活用を期待したい.

39

山内　どうもありがとうございます。多分皆さん、お聞きになりたいことは山ほどあろうかと思うのですが、あと５分程度しかございませんので、お二人程度お受けしたいと思います。

岩沢　皆さん、貴重な講演、ありがとうございました。私としては、本当はディスカッションが良いのですが、ちょっと私としてはメッセージを言わないといけないところがあって。

小林さんの発表自体は別に問題はなかったと思うのですけれども、ちょっと僕は誤解を与えるところが

あると思って。できれば、小林さん、すぐ出るなら出してほしいのですけれども、出ないですか、小林さんのスライドは。

ですので、口頭で言うと思うのですけれども、Lasso の話があって。私も講演の中で機械学習から学べる手法の一つと言ったのですけれども、原論文を見れば、彼は本当は分かっているはずだと思うのですけれども、原論文は統計学の論文で出たものなのです。逆輸入なのですね。

Lasso という方式は、統計学の人が考え出したのだけれども、スパースモデリングで使えると、後で分かったのですよ。今日の発表だと、「スパースモデリングに使うのは Lasso がいいという提案が出た」という説明になっていたのですけれども、違うのです。あれは統計学のものなのです。それを「スパースモデリングで使えるよ」と分かったものだから、機械学習の人が飛びついたのです。今は学ぶのは機械学習の教科書から学ぶのです。今、学ぼうと思うと。でも、あれは統計学の手法なのですね。

ですので、実際にはそのような逆輸入のものが幾つも、実はあったりするので、そのように今日の話も捉えてもらえると良いと思います。口頭で、ちらりと最後に言ったので、もちろん彼はもう知っているわけですが、原論文にそのベイズ的な解釈が入っているのですね。Lasso の。そうですよね。そのように口頭ではおっしゃっていたのだけれども。このスライドだけを見ると、それが後から、ベイズ的にも解釈できるというように誤解されてしまうと思うのですけれども。

そのように、本当に統計学の手法が機械学習の人が先に使ってしまっているという。だから、われわれはそれを取り戻して、使うべき時期に来ているのではないかなということで、ちょっと補足として、お願いします。

山内　では、お一人、いかがでございますか。では、重原さん。

重原　すみません。紙に穴を開けてコンピュータを使った最後の世代ぐらいの重原。第一生命経済研究所の重原と申します。

今後のアクチュアリーにコーディングが必要という、あるいはコンピュータ技術が必要というときに、特に古い人だと、新しい言語を覚えれば良いのかなぐらいに思う危険性が非常にあると思うので、そこのところは恐らく違うのだということで。そこのことについて、補足でご説明していただければと。
簡単に言えば、R と Python をという、新しい言語を学べば良いという話ではないですよねと。だから、キーワードになるのは、恐らくモジュールや何かをうまく使う技術などというような話ではないかなとは、思っているのですけれども、その辺の話をお聞かせいただければ。

山内　例えば、データを投入して、ただ答えが出ればいいという話。

重原　そういったプログラムが書ければ良いのとは、ちょっと理屈が違うよねという話です。

山内　なるほど。多分、お立場によって回答が変わると思うのですけれども、では藤澤さんの方から一つお願いします。どう思われますか。

山内　いや、これはご質問というより、「どう思われますか」だと思うのですよね。

藤澤　私が使ったことがあるのはＲだけですが、Ｒを使うといろいろなパッケージがあって、データを入れると結果が出てくるというところはあって、それほどプログラムの知識がなくても、統計的なモデルが実装できるというメリットはあると思います。

ただ、それだけで本当にいいのかと言うと、やはり別の話もあって。ちょっとお答えになっているか分からないですけれども、やはり理論的なところも、どのようにして結果を解釈するのかというところは、多分とても重要で、そこを間違ってしまうと、間違ったそのモデルや、モデリングの結果を実務に使ってしまうという危険性があると思うので、そこは多分、両方必要になってくるのではないのかなと思っています。

特に言語にこだわっているというわけではなくて、ＲでもPythonでもJavaでも、何でもいいと思うのですけれども、出てきた結果を正しく解釈するというところも、必要になってくるのではないのかなというのが、実務的な観点での意見です。

山内　では、続けてお願いします。

小林　先ほどの発表でもちらっと述べましたが、何がしたいかによって、コーディングというのは、しょせん手段なので、何がしたいかによって変わってくるのかなと思いますが、どのようなことに特に興味があるのでしょうか。

もし仮に、コーディングが得意で、何も制約がないのであれば、個人的にはＣやＣ＋＋などを書けると、とてもいいのかなと思いますが、そこまで勉強するのは大変だと思うのであれば、やはりＲやPythonがとても導入しやすいのかなと思います。

重原　そのような意味では、どちらかと言うと、それ自体に興味があるというより、結果に興味がある人が、恐らくこの中にいる人はほとんどだと思うので、そのような意味では、ＲやPythonを使うということになるのでしょうけれども、あれですよね。先ほどコミュニティと言っていましたけれども、いわゆる、コミュニティに蓄積されたルーチンや何かを使っていくということが重要になってくるのではないかなと、質問の趣旨はそのようなところだったのですけれども。

小林　それもあまりいい回答ができないですけれども、自分は比較的実務とかけ離れた所にいるので、新しいものを作るには自分で全部コーディングしなければいけないという状況がかなり多いので、うまく回答も出せていないですね。ディープラーニングなどをやるのであれば、ライブラリだとPyTorchやKerasなど、そのような大学が開発しているライブラリのようなものがあるので、そのようなものをうまく利用していった方がいいというのは、そのとおりだなと思います。

山内　では小暮先生、ぜひお願いします。

小暮　私も最近脳活にいいかと思いPythonを始めたのですけれども、なかなか進んでいなくて、あれですけれども。

私はコーディングが本質的ではないというのは、確かにそうだと思うのですけれども、一方でコーディングのパワーというものに、とても実は最近思うところがあって。やはり実装できる能力ということだと

思うのですね、コーディングできるというのは。それはやはり侮れないし、これからのデータ・サイエンティストと呼ばれる人にとっては、結構本質的な部分にかかわるのではないかなと思っています。ただし、自分はもうできません。

山内　どうもありがとうございました。もうお時間ということになりました。それでは最後に、藤澤さん、それから小林さん、小暮先生、お三人に盛大な拍手をお願いいたします。