00：00：12

小池：そろそろ、じゃあ、始めましょうか。次はSRPBの話で、Traveling Subjectのデータを使ってどうやってHarmonizeするかということです。これは脳プロのデータでも国際脳のデータでも、もしくは例えば自分たちの研究でMRIの機種を変えるなど、プロトコルを変えるときも、過去のデータが無駄にならないように計画的に撮ることもできますので、そういったもので使っていただければと思います。ではよろしくお願いします。

朱：東京大学小池研究室の朱と申します。前回もPythonのチュートリアルを担当させていただいていたので、今回はPythonを用いて、Traveling Subject法を用いた複数プロトコルデータセットのHarmonize法を担当させていただきます。

　先ほど田中先生にもたくさん紹介していただいたのですけれども、Multi-site dataは、近年は結構いろいろなプロジェクトが増えてきて、有名なのがHCPやADNI、SRPBやCRHDなどで、あとは先ほど紹介していただいたDecNefもあります。

　こういったようないろいろな施設で撮ったデータは、いろいろなスキャナーやいろいろなプロトコルが使われていて、そのままを解析するにはちょっと一つの問題点があります。あとはコホートデータなども同じ方をロング〓ティージョナル的〓にずっと撮っていくのですけれども、途中でスキャナーが変わるといったようなこともあります。

　そういったような原因、メーカーやスキャナー、撮像プロトコル、ヘッドコイルやPhase encoding、位相方向などによる影響から、実際に見たい目的変数、例えば疾患による違いや発達による変化などが見えなくなってしまって検出しにくくなる、もしくはもう検出できなくなる可能性もありますので、そういった統計的な検出力を高めるために、Harmonize法を用いてやっていきたいと思います。

　簡単にいえば、要は同じiPhoneで撮影しても、同じ環境でも同じような画像が得られるわけではありませんので、そういったようなことをHarmonize法で解決していきたいと思います。

　先ほども紹介していただいたTraveling Subjectという方法があるのですけれども、同じ人はいろいろな施設に行ってスキャナーで撮像して、いろいろなプロトコルも全てを体験して撮像していく方法になります。

　ATRの山下歩先生の論文では、こういったような報告はされていて、一番機種間差に影響されるのは、Phase encodingで、次はManufacturerで、その後にコイルとスキャナーの影響が減っていく感じになります。

　どのように補正をしていくかというと、すごくシンプルな手法なのですけれども、GLM法を使って補正していきます。開発された先生の論文の中では、Connectivityを対象にしているのですけれども、今回のこのデモが、ConnectivityをFunctional MRIのものではなくて、T1 weightのROI thicknessに置き換えてやっていきたいと思います。

　まずはこの式の中にメジャーメントバイアスというものがあって、それはもう先ほどご紹介していただいたようにPhase encodingやプロトコルなどです。さらにサンプリングバイアスも入っていて、ある施設ではうつ病の方しか集まれない、ASDの方しか集まれないという場合は、他施設間で解析した場合には、このサンプリングバイアスも取り除かなければいけません。

00：04：56

　この式の中では3つのサンプリングバイアスが入っていて、まずはヘルスコントロールです。次にはMDD、さらにスキゾフレニアのものになるのですけれども、その次にDisorder factorも入っています。さらに個人差も入っています。今回はTraveling Subjectでいろいろな機種間差を推定していくのですけれども、もちろん全部が健常者ですので、疾患のファクターがないのと、このサンプリングバイアスもありません。

　下の式は、今回落とし込んだTraveling Subjectに特化した式になります。今回はこの式の中に、後ほど紹介させていただくのですけれども、Traveling SubjectのSite1とSite2、2つのデータがありまして、ファイル名はそれぞれにTS01のCSVと、TS02のCSVで代入して計算を行います。このデータがこのセクションしか使いませんので、後ほど舞草先生にお話をしていただくCombatで補正するセクションは、これを使いませんのでご注意ください。

　ここで何が一番重要なのかというと、サイトエフェクトと疾患もそうなのですけれども、全ての要素を1-of-K binary coding schemeを使って、コーディングを使っていました。要は例えばサイトが12個ある場合は、サイトエフェクトの係数は12個になります。

　データの説明をさせていただきたいと思います。今回のデモで実際に使うデータが思春期のコホートの、男女40名のデータになります。同じ人に2回のスキャンをさせていただいています。1回目は11歳の頃で2回目は16歳の頃になります。それぞれにデータの01と02に対応させていただいています。

　全てのデータがT1 weight、FreeSurfer6.0で処理したROIのCortical thicknessになります。実際の数値からランダムに数パーセントを増減させていますので、実際のデータではあるのですけれども、若干増減させていますので、完全に架空のデータではありません。ご覧のようにSite1とSite2も、マシンが違うのとヘッドコイルとプロトコルなども全部が違います。

　こういったような違いがどうなるかというと、全く何もしていなくてもSite1とSite2の平均値を●で掛けると有意差が出てきます。補正法できれいにSite1とSite2の有意差がなくなる形になっていくのですけれども、続きをお話しさせていただきたいと思います。

　今回はTS dataをTSの01と02を用いてサイトごとの係数を求めてから、次に実際の思春期の、40名の男女のデータ01と02に関して補正していく形になります。それぞれのPython scriptが分かれている感じになります。Output filesもそれぞれに違いますので、ご注意ください。

　ではコードのほうに入っていただきたいと思います。皆さまもスクリプトを手元に用意していただくとありがたいと思います。すみません、共有がうまくいかないので少々お待ちください。

00：09：43

　ではGML\_HMD-ROI-Tutorial2020を開いていただきたいと思います。まずは各自のパスの部分を、各自のローカルのパスに書いていただきたいと思います。大丈夫ですか。このパスを書いたら、こちらがPythonのモジュールになりますので、まずはこちらのほうを全てインポートしていただくと、あとがスムーズにいきます。

　モジュールを読み込まないとPythonは何もしてくれませんので、必ず全てを最初の部分にまとめて読み込む必要はないのですが、ここはこのファンクションを使う前に読み込まないとうまくいきません。まずはファイルを読み込みたいと思いますので、TSの01と02をそれぞれに入れておきます。

　ご覧のようにこれを見ていただくと、サフィックスは付いていませんので、同じ人を2回スキャンでも、同じ名前でSub TS01として入ってしまうのですね。まずはこのデータを整形していくのですけれども、TS01のほうにサフィックスを追加していただいて、02のほうを整形していただいてから縦に連結していって、同じ変数に入れておく形になります。

　ここが重要になっていくのですけれども、1-of-K binaryを、まずは個人差を推定するための対角行列を作ります。対角行列を作った後に列名を与えてあげます。全てPandasフレームの形になっていますので、この変数を、ヘッドを出していただくと、以下のように1名から17名の対角行列がうまくできているのがお分かりかと思うのですけれども、この後に1回目撮像のコーディングと2回目撮像のコーディングを作っていきます。

　1回目撮像のコーディングではSite1が全て1になるのですけれども、Site2は0の列を作ってあげる形になります。このサイトプロットの変数なのですけれども、これはプロットのために作った変数で、実際の計算の中では使いませんのでご注意ください。

　同じ手順で2回目撮像のコーディングも作成をしていきます。作成したコーディングはこういったように同じフォルダで保存されるようになるのですけれども、ここにTSコードが来ています。ここまでできたら、もうコーディングは全てが出来上がるようになるのですけれども、ここではさらにコーディングを使ってHarmonizeをしていこうと思います。

　まずはこの先ほど作っていただいたTSのコーディングとデータをコピーする形になっていくのですけれども、ここがPythonの特別なところで、コピーをしないと、イコールだけで書くと本体も変わったりすることがありますので、PandasとNumpyではそれぞれの特性が違いますので、ご注意ください。

　基本的には、コピーをすればコピーされたもの本体は変わらないので、新しい変数のほうだけに変化を加えていきます。ここではまずは先ほどプロットのために作った変数をドロップしていくのですけれども、この列が入っていくと計算のときに余計な係数も一緒に計算されてしまうので、ドロップした後にこのデータの形を確認していきます。

　切片を求めるのですね。この切片は全てのサイトと全員の平均を求めているのですけれども、いったん切片を求めて、切片を全部のデータの形に作っていきます。作っていった後に切片をマイナスしていくだけなのですけれども、ここでは、山下先生の本文のほうでは2次計画法で解いているのですけれども、私は2次計画法にはあまり自信がなかったので、ここでは多重共線性の可能性を回避するために、ここではリッジ回帰を行っています。

00：15：21

　ここでは実際にリッジ回帰を提起してから、ここまではまだデータに関して何もしていないのですけれども、このフィットを行った後に係数を求めることになっています。ここではどれぐらいの性能で出ているかを見ている感じになります。

　求めた後に、サイトエフェクトの係数を取り出す作業があります。そのままだと全ての個人差の係数とサイトの係数を17名プラス2サイトですので、19列を掛け、69ROIの行列ができているだけですので、ここではサイトエフェクトを取り出す必要があります。

　従って、一番後ろに個人差プラスサイトエフェクトというふうに行列を作りましたので、最後の17列と18列を取り出す必要があります。Pythonをご存じない方もいらっしゃるかもしれないのですけれども、行列が0から始まるので、ここでは実際に19列なのですけれども、19列を取り出すと怒られてしまいますので、ここはきちんと－1してあげてください。

　そこからさらに行列を転置しなければいけないので、しないとこのままでは計算できません。ここではまずSite1の係数とSite1のコーディングの掛け算をして、Site2も同じような手順をしていきます。

　サイトエフェクトを求めてから、元のRawのデータからサイトエフェクトを引いていく形になるのですけれども、引いていくとHarmonizeした後のものになります。このデータからさらにHMD\_thicknessというファイルが出力されるようになります。

　ここで走らせるとこちらにHMD\_thicknessというCSVが吐き出されます。ここにFuture warningが書いてあるのですけれども、この次のバージョンでは何々の機能がなくなりますという感じですので、ここではあまり気にしないでください。

　ここからVisualizationに入っていくのですけれども、特に難しいことはしなくて、先ほどお見せしたT-Testのグラフを作っていくだけになります。まずはICVの列をなくして、さらに全体のthicknessの平均を求めていくのと、サイトプロットの係数を最初に作りましたので、これをくっつけていく形になります。

　くっつけたら、Rawの画像のSite1、Site2だけの全体の平均を求めて、T-Testを行います。次は同じ手順で、1回グラフを出力して、次の手順で同じようにHarmonizedされたものをT-Testを行って、グラフが出力される形になります。きれいにこの有意差がなくなることがお分かりかと思います。ここまでは大丈夫ですか。

　じゃあ、補正のほうに入っていきたいと思います。これが終わりましたらここにHMD\_thicknessとのファイルが吐き出されます。site coefの中にサイトごとの係数がこのように入っています。

00：19：45

　続きですね、APPLYのほうを開いていただきたいのですけれども、これも先ほどと同じように、まずはモジュールを読んでいただいてから、ご自身のデータがあるフォルダのパスに書いていただきたいと思います。

　ここのフォルダに行きましたら、まずはSite1のデータ01です。11歳頃の40名の、思春期のデータに対して補完していくのですけれども、まずはデータを読み込んでいただいて、読み込んだらcoefのところに、先ほど吐き出されたsite coefのファイルを読み込んでいただく形になります。読み込んでから、まずはSubjIDの列が入っていますので、SubjIDの列を削除して、ROIデータのみのデータフレームを作成していきます。

　ここではもう一度新しいデータフレームのサイズを確認します。ここからまた、coefのファイルの中にSite1とSite2の係数が入っていますので、まずはSite1の40名のデータに関してHarmonizeしていくのですけれども、Site1を取り出していただいて、同じ行列を作っていかないとうまく計算ができませんので、ここではまずは、この係数を人数分のものを作っていきます。人数分を作った後にサイトエフェクトを計算していきます。計算した後に、Rawのデータからサイトエフェクトを引いていく形になります。

　最後にSubjIDと元にあるヘッダーをくっつけていって、新しく計算した後の、Harmonizationした後のデータがHMD\_Data01として出力される形になります。ここまでを実行すると、ここにHMD\_Data01が吐き出されます。ここは単に私がSite1のデータの、行列の形を確認しただけですので、特に必要があるわけではないです。

　続きですね。Site2のほうの計算をしていくのですけれども、同じ手順でSite2のデータ02を読み込んでいただいて、そこからSubjIDをさよならしていきます。同じようにSite2の係数を取り出してきてSite2のエフェクトを計算した上で、引いていきます。

　最後に同じように、SubjIDとHarmonizeした後のデータをくっつけて出力させていただく形になります。これを実行していただくと、ここにHMD\_Data02のCSVが吐き出される形になります。

　最後にこれはどうHarmonizeできたか、データはどうなったのかというのを、一回ageとの相関を見てみたいと思いますので、ここにまず配布したフォルダの中にあるのだと思うのですけれども、demographicsというファイルを読み込んでいただくと、まずは1回目撮像の時のageと2回目撮像の時のageを取ってきて、縦にくっつける形になりますね。さらにまたプロットのために新しい列を作っていきます。

　ここではRawのデータとHarmonizeした後のデータをさらにここで縦にくっつけましたので、下にも同様に縦にくっつけていきます。そのまま、さらにICVとサイトプロットをさよならしていって、そのさよならした後のデータで全体のmean cortical thicknessを求めていきます。

　ここまでの手順が終わりましたら、下のほうにRegressionのプロットをしていくのですけれども、ここは実際的にRegressionの統計解析を行ったわけではありませんので、プロットをして回帰直線を書いただけです。

00：25：04

　ここから見ると、年齢に伴ってcortical thicknessが増えているように見えるのですけれども、実際にはこれは減っていくはずなのですが、きちんと補正後のデータが、年齢と負の相関が見られることになりました。

　これもそれぞれにRawと〓HMD〓した後のものを、プロットをしたものがPNGとしてここに保存されるようになります。これを実行すると、回帰直線を書くのに若干時間がかかりますので、私は何回もしているのでキャッシュが残っていて結構速いかもしれないのですけれども、皆さんのほうはもうちょっと時間がかかると思います。

　これを走らせると、最後にこの辺のthick\_mean\_hmdとthick\_mean\_rawの画像がそれぞれにプロットがされます。若干早いのですけれども、私の部分は以上となります。

小池：ありがとうございます。1つ、チャットのほうに質問が来ています。リッジ回帰のαは何を基準に決めるとよろしいでしょうか。

朱：一応今回配布させていただいたものは、αのところが0.01になっているのですけれども、これは1回グレートサーチをしてαの範囲をちょっと広めに取って、グレートサーチをしてこれがベストアンサーとなっていたので、これにしていたのですけれども。答えになりましたでしょうか。

小池：答えにはなっています。僕が決めることじゃないですね（笑）。要はそれのαというのは、データセットに依存するかどうかというところもあるかと思うのです。

朱：Pythonはデフォルト的にαの値が0.01になっていると思います。大きくなり過ぎると、リッジ回帰ではなくてラッソになってしまうらしいのですけれども、リッジとラッソの違いで、ラッソにあるたくさんの変数の中から0の変数が出てきてしまいます。

　今回は、たくさんのサイトがあったとしても0になってほしくないので、リッジ回帰をそのまま0.01に決めさせていただきました。

小池：もう一個の質問は、Traveling Subjectのための必要なサンプル数について経験的な情報というか、舞草先生がそれについて今論文を書いているので秘密ではないのですけれども、多分10人以上も撮っておけば大丈夫かなという経験値があります。

　ただ、僕たちは構造でしか確かめていなくて、ATRの方は最初にレスティングのファンクショナルコネクティビティでしか確かめていないので、全てあまねく情報がそれで集約されるかどうかは分からないのですけれども、そんなところです。

　多分30分しかないので、全部の行動を理解するのはPythonをやっている人じゃないと無理なのではないかと思うのですけれども、一つ一つ実行して出てくるファイルやアウトプットを見てくれると、やっていること自体はそんなに難しいことではありません。あとはデータ自体も、原理的には別にMRIに限らず、行列だったら何でもできる、Combatもそうですけれどもできると思うので、ちょっとやってみていただくといいのかなと思います。

　あと他に何かご質問がある人はいますか。3分ぐらい余っています。大丈夫そうですか。じゃあ、朱さん、ありがとうございました。

00：30：01