応用計量経済学：課題1*

島根哲哉

2015. 5. 8

この課題の目的はロジットモデルの特定化の方法と解釈について理解を深めることである．多くの商業ソフトウェアパッケージ，STATA, SAS, NLOGIT（もちろん R, SPSS も） にはロジットモデルの推定ルーチンが含まれている．希望があればこの課題にこれらのパッケージを使うことはかまわない．しかしながら，ここでは matlab の推定ルーチンを提供しており，これを使うこともできる．この matlab のプログラムを実行するためには，optimization toolbox をインストールされた matlab が必要である．

この課題ではカリフォルニアのおける住宅の暖房システム選択のデータを使う．観測値はカリフォルニアの 900 戸の一戸建て家族向け住宅で構成され，これらは新築であり集中空調 (central air-conditioning) がある．選択しているのは暖房システムについてである．5 つのタイプのシステムが可能であったと考えられる:

(1) gas central,

(2) gas room,

(3) electric central,

(4) electric room,

(5) heat pump.

logit を含めた離散選択モデル用のデータには 2 つの形式（”wide”と”long”）がある．ここには何れのフォーマットのデータもあり，これらの形式を確認することができる．

WIDE FORMAT wide format は各観測値（つまり一つの選択状況）にデータの一つの行をあてる．データの行は意思決定者の属性と全ての選択肢の属性からなる．ファイル `datawide.xls` は wide format のデータの excel ファイルである．ファイルを開いてデータを確認しなさい．もし excel ファイルを読むことができない場合は，データは `datawide.asc` に camma-delimited ascii(text) ファイルでもある [1]．これは任意のテキストエディタで開いて見ることができる．

各行に 19 の変数を含んだ 900 行からなる．(各列の) 変数は以下の通り:

---

*この文書は K. Train の Probrem set(ec244ps1) に含まれる”ReadMe.txt”を訳し，また講義の趣旨に沿って一部改変したものである．

[1]いわゆる CSV 形式のファイル

1. idcase: gives the observation number (1–900)
2. depvar: identifies the chosen alternative (1–5)
3. ic1: installation cost for a gas central system
4. ic2: installation cost for a gas room system
5. ic3: installation cost for a electric central system
6. ic4: installation cost for a electric room system
7. ic5: installation cost for a heat pump
8. oc1: annual operating cost for a gas central system
9. oc2: annual operating cost for a gas room system
10. oc3: annual operating cost for a electric central system
11. oc4: annual operating cost for a electric room system
12. oc5: annual operating cost for a heat pump
13. income: annual income of the household
14. agehead: age of the household head
15. rooms: number of rooms in the house
16. ncoast: identifies whether the house is in the northern coastal region
17. scoast: identifies whether the house is in the southern coastal region
18. mountn: identifies whether the house is in the mountain region
19. valley: identifies whether the house is in the central valley region

選択肢の属性 (導入費用と運転費用) は各選択肢毎に異なる値をとることに注意せよ．つまり, (一つの選択に対して）5 つの導入費用があり，5 つの運転費用がある．logit モデルを推定するためには，研究者は全ての選択肢の属性に関するデータが必要であり，選ばれた選択肢の属性だけでは十分ではない．例えば，実際に導入されたシステムへの支払いがどれだけであったかを決めるのでは十分ではない．研究者は導入されたかもしれないシステム毎に導入費用を判断する必要がある．(なぜなら）選択の過程における費用の重要性は，選ばれたシステムの費用を選ばれなかったシステムの費用と比較することを通じて決まる．.

このデータでは，仮にその住宅でそのシステムを導入した場合に要する費用の合計として費用を計算する．ここでは，住宅の特性（大きさ等)，その住宅の立地でのガスや電気の価格，その地域の気象条件を与件とする．この費用は，集中空調 (central air-conditioning)

のある住宅のものである．(これが gas central が gas room よりも導入費用が低い理由である: 集中システムは既に導入されている空調ダクトを利用することができる)

第一番目の家計は選択肢 1(gas central) を選び, 所得$70,000 であり，世帯主は 25 歳で，住宅は 7 部屋，north central 地域に立地することが見て取れるであろう．

LONG FORMAT long format は選択状況の各選択肢のデータを一つの行とする．900 戸の 5 選択肢については，long format では 4500 行のデータがある．”observation”の用語はしばしば long format のデータでは混乱をしている．計量経済学的な観点からは，それぞれの選択状況が observation であり，尤度関数の要素を一つ与える．この用法では，一つの observation はデータのいくつかの行から構成され，一つの行はそれぞれ選択肢に対応する．データ保持の観点からは，データのそれぞれの行がしばしば observation と呼ばれ，この場合，いくつかの “observation” が各選択状況に対応する．ここでは observation を計量経済学的な意味で用いるが，多くのパッケージのロジット推定ルーチンのドキュメントではデータ保持の定義で使っている．

excel ファイル `datalong.xls` は long format でのデータであり，ascii ファイルでは `datalong.asc` である．

1. Idcase: gives the observation number (1–900)
2. Idalt: gives the alternative number (1–5)
3. depvar: identifies the chosen alternative (1=chosen, 0=nonchosen)
4. ic: installation cost of system
5. oc: operating cost of the system
6. income is the annual income of the household
7. agehed is the age of the household head
8. rooms is the number of rooms in the house
9. ncostl identifies whether the house is in the northern coastal region
10. scostl identifies whether the house is in the southern coastal region
11. mountn identifies whether the house is in the mountain region
12. valley identifies whether the house is in the central valley region

long format では，従属変数はその選択肢が選ばれたか否かの 0–1 で定義されていることに注意せよ．wide format では，従属変数は選ばれた選択肢の番号 (or ラベル）を識別する 1–5 である．long format では，意思決定者の属性は全ての選択肢で（同じものが）繰り返される．

大抵のソフトウェアパッケージでは何れの format も使うことができる．通常，選択の状況毎に対応する選択肢の数が異なる場合には，long format の方が使いやすい．それは

各選択状況に対応する選択肢と同じだけの数の行を含めれば良いからである．対照的に，wide format ではどんな選択状況でも可能な選択肢それぞれに対して各属性の変数を含んでいる，(オプションとして）その選択肢がない選択状況ではそれらは欠損値もしくはゼロとする．一方で，long format は意思決定者の属性が繰り返され無駄なスペースがある．

matlab での実際では，我々のデータを wide format から long format に変換したり，その逆を行う matlab code を書くことができる．(ファイル `widetolong.m` がこれである)

EXERCISES:

問. 1 導入費用と運転費用だけを説明変数とした logit モデルを実行せよ．`logit.m` ファイルでは standerd logit モデルで特定化し，推定することができる．二つの説明変数（導入費用と運転費用）のロジットの実行が現状で設定されている．このまま実行することができ，結果は `myrun.out` に書き込まれる．この結果を `myrunKT.out` と比べて確認し，Train の結果と同じであることを確かめろ．

matlab を使ったことがない場合は，以下に手順をしるす．この ps1 フォルダを自分のマシンのフォルダに移す．Matlab を起動する．matlab の workspace ウィンドウ上で，カレントディレクトリをあなたがファイルを入れたフォルダに変更する．左側のサブウィンドウにはカレントディレクトリのすべてのファイル名が現れる．`logit.m` のファイル名のクリックすると，エディタウィンドウでそのファイルが開かれる．エディタウィンドウの右上にある Run アイコンをクリックすればプログラムが実行される．

その他のファイルが何であるかを理解しよう．`logit.m` は `check.m`, `doit.m`, `loglik.m`, `pred.m` を呼び出し，それぞれデータのチェック，最適化の呼び出し，対数尤度の計算，シェアの予測を行う．これらはユーザには透過であり，特別な理由がない限り，logit モデルの推定のためにこれらを見たり変更する必要はない．

`logit.m` を実行した推定結果を吟味せよ．特に:

(a) 推定された係数は期待した符号であるか．

(b) いずれの係数も有意にゼロと異なるか．

(c) 各暖房システムのついた住宅の実際のシェアと予測されたシェアはどの程度一致しているか．

(d) 通常係数の比率は経済的に重要な情報を与える．運転費用 1 ドル削減のための導入費用の追加的な支払い意思額（WTP）は運転費用の係数と導入費用の係数の比である．このモデルから推定された WTP は何か？それは強度（大きさ）の点で妥当なものか？

(e) 推定された WTP から割引率の推定値を得ることができる．これは運転システムの選択モデルの結果と言える．将来の運転費用の現在価値はシステムの寿命までの運転費用の割り引かれた合計である．$PV = \sum[OC/(1+r)^t]$, ここで r は割引率，$t = 1, \ldots, L$ で L はシステムの寿命である．L が増すと，PV は $(1/r)OC$ に近づく．つまり十分に寿命の長いシステムについては（これらのシステムはそうであると仮定しよう），OC 一ドルの削減が将来の運転費用の現在価値 $1/r$

の削減をする．これは，システムを選択している人が 導入費用とシステムの寿命までの運転費用を負担し，r の割引率でこの二つを合理的にトレードオフするならば，意思決定者の運転費要削減の WTP は $1/r$ になることを意味する．これらを受けて，(d) で計算された WTP の推定値から r の値を計算せよ．これは妥当なものか？

問. 2 $r = 0.12$（すなわち WTP= 8.33）の制約を加えてモデルを推定せよ．次に $r = 0.12$ の仮説を検定せよ．このためには，新しい変数を XMAT につくる必要がある．ここでは matlab のコードで新しい変数をつくり，これを使う方法を示す．導入費用の対数という新しい変数を作りたいとする．XMAT をロード（読み込み）した後に以下を加える:

```
newvar=log(XMAT(:,4));
%This creates a column vector that is the log of installation cost.
XMAT=[XMAT newvar];     %This appends the new column vector onto XMAT
```

新しい変数は XMAT の 13 番目の変数である．もし導入費用の代わりに導入費用の対数を使いたければ，説明変数を以下のように特定化する:

```
IDV=[13 5];
```

問. 3 選択肢固有の定数項を選択肢 1–4 に加えよ．選択肢 1 の選択肢固有の定数項は次のように作る:

```
alt1=(XMAT(:,2) == 1);
XMAT=[XMAT alt1];
```

ここでは 5 つの選択肢があるので，定数項は 4 つだけしか加えられないことを忘れないように．選択肢 1–4 に定数項を加えたので，選択肢 5 の定数項をゼロにするように基準化している．

(a) 推定された確率は各選択肢を選んだ顧客のシェアにどれくらいうまく一致したか？厳密に一致していることに注意せよ：ロジットモデルでの選択肢固有の定数項は，平均確率が観察されたシェアに等しくなることを保証する．

(b) 推定値から，WTP と割引率 r を計算せよ．これらは妥当か？

(c) 定数項を選択肢 1,3,4,5 について導入し，選択肢 2 についてゼロに基準化しているとする．選択肢 1 の定数項の係数の推定値はどうなるか？実際にモデルを推定するのではなく，(なぜそのような値となる) か論理的に明らかにせよ．

問. 4 さあ，社会人口学的 sociodemographic 変数 (n によって決まる変数) が入ったモデルを試してみましょう．

(a) 導入費用の代わりに導入費用を所得で割っていれなさい．この特定化では，導入費用の係数の大きさは所得の逆数と関連している，つまり高所得家計は低所得家計に比べ導入費用に関心がない．導入費用を所得で割ることはモデルを改善したか，改悪したか？

(b) 導入費用を所得で割る代わりに，選択肢固有の所得効果を入れなさい．第一の選択肢の所得変数は以下のように作る:

```
incalt1=XMAT(:,6).*(XMAT(:,2) == 1)./1000;
%Where the division by 1000 scales income to be in thousands
XMAT=[XMAT incalt1];
```

選択肢 2–4 について同様にし，選択肢 5 の係数をゼロに基準化せよ．
この推定値は，集中システムとルームシステムの選択への所得のインパクトについて何を意味するのか？これらの所得項は有意であるか？

(c) 他のモデルについて試しなさい．これらのデータから，どのモデルが最良と考えられるか決めなさい．

問. 5 logit モデルの予測への利用について考えていこう．導入費用，運転費用，選択肢固有の定数項でモデルを特定化する．このモデルを実行せよ（or 前の結果を見直せ）．以下の予測ではこのモデルを使いなさい．

問. 6 カリフォルニアエネルギー委員会 (CEC) はヒートポンプにリベートを提供するか検討している．CEC はカリフォルニアの顧客の暖房システム選択へのこのリベートの影響を予測したい．リベートは導入費用の 10%である．新しい導入費用は：

```
newic=XMAT(:,4)-XMAT(:,4).*(XMAT(:,2) == 5)./10;
 %Where the division by 10 reduces alt5's cost by 10%
```

問. 5 のモデルの推定係数を使って，オリジナルの値の代わりにこの新しい導入費用のもとで予測シェアを計算せよ．また，ヒートポンプの住宅のシェアはこのリベートによってどれだけ上昇するか？