数理手法 I 演習
第 1 回 (全 5 回)

準備:
- Mac でなく Windows を起動してください
  (Mac の Excel は分析ツールを使えないため)
- 演習用テキスト:
  「Excel による統計入門 Excel 2007 対応版」
  この演習では第 9–10 章は扱わない
  今日は第 1 章から第 4 章をカバーする
- データをダウンロードしてください
  http://asakura.co.jp/
  books/isbn/978-4-254-12172-8/

第 1 章 Excel 入門

概要

- データ入力, 数式, 関数, 表示形式の変更
- ファイル保存, 読み込み

データの入力 (1)

テキスト 1–9 ページ

- 「リボン」と呼ばれるインターフェース
  - タブ (ファイル, ホーム, 挿入, ...)
    - グループ (クリップボード, フォント, 配置, ...)
      - コマンド (クリックするボタン)
- セル, セル番地, アクティブセル, 数式バー

データの入力 (2)

- 半角英数, かな入力切り替え
  教室の端末では
  英数キー (スペースキーの左隣)
  を押した後は半角英数入力ができ,
  Kana キー (スペースキーの右隣)
  を押した後はローマ字入力ができる
  コマンドキー + スペースキーでは
  切り換えられないことに注意する

データの入力 (3)

- 漢字変換 (スペース), カタカナ変換 (F7, Ctrl + i), 文字区切り変更 (Shift + ←, Shift + →)
  Ctrl キー (左の方, a の左隣) を押しながら i を押すということ
- 入力後のカーソル移動 (矢印キー, Enter キー, Tab キー)
- セルの内容の変更 (ダブルクリック, 数式バー)

セルの表示の変更

テキスト 9–14 ページ

- 文字位置の変更 (左揃え, 中央, 右揃え)
- 小数点表示桁上げ, 桁下げ
- コンマ表示
- パーセント表示
- 指数表示
- 列幅の変更

数式の入力

テキスト 14 ページ

$2+3$	$\Rightarrow$	=2+3 と入力
$(2+3)^2/2$	$\Rightarrow$	=(2+3)^2/2
$\log_{10} 5$	$\Rightarrow$	=log10(5) または =log(5)
$\log 5$	$\Rightarrow$	=ln(5)
e^5	$\Rightarrow$	=exp(5)
$\sqrt{2}$	$\Rightarrow$	=sqrt(2)

特殊な記号の入力

- 1/2 と入力
  $\Rightarrow$ 日付として扱われ, 加算, 減算が可能
- '1/2 と入力
  $\Rightarrow$ 単なる文字列として扱われる

行や列の挿入・削除

アクティブセルを移動して
  ホーム → セル → 挿入
  または
  ホーム → セル → 削除
または
  セルを右クリックして出てきた
  メニューから行・列を挿入・削除

ファイルの保存

- ファイル → 名前を付けて保存として
  表示されるダイアログで,
  保存したい場所とファイル名を指定して
  現在作業中のファイルを保存する
  (ファイル名は EX1 としておく)

ファイルの呼び出し

一旦ファイルを閉じる (ファイル → 閉じる)

- ファイル → 開くとして表示される
  ダイアログで, 呼び出したいファイルを
  クリック

ここで第 1 章は終了
一旦 Excel を終了する

第 2 章 Excel による表計算

概要

- 表計算ソフトの仕事
  - 式と関数の入力
  - 相対セル番地と絶対セル番地
  - 式の複写と値の複写
- 実例: 人口増加率の計算

式の入力

A1 から B5 (A1:B5 と書く) にデータを入力

1	6
2	7
3	8
4	9
5	10

C1 に =A1+B1 と入力

行和の計算と相対セル番地

Ctrl キー (左の方, a の左隣) を押しながら c を押すということ

C1 をアクティブセルにして,
  コピーボタンをクリックまたは
  編集 → コピーまたは Ctrl + C
C2:C5 を選択して, Enter キー押下または
  貼り付けボタンをクリックまたは
  編集 → 貼り付けまたは Ctrl + C
C2 の内容を確認 ⇒ "=A2+B2" となってる

単純にセル番地を指定すると
相対セル番地で記録される

関数による計算

列和 (列ごとの和) の計算
  セルの個数が多いから関数を使う

1. A6 を選択 (アクティブセルにする)
   して $\sum$ をクリック
   または
   A6 に =sum(A1:A5) を入力
2. A6 を B6:C6 に複写

割合の計算と絶対セル番地 (1)

総計 (C6= 55) に対する割合を
A11:C16 に計算
  A11: =A1/$C$6
  A11 を A11:C16 に複写
  %表示, 小数桁調整

"$" を付けてセル番地を指定すると
絶対セル番地で記録される

割合の計算と絶対セル番地 (2)

行和 (列 C) に対する割合を E1:G6 に計算

E1: =A1/$C1

E1 を E1:G6 に複写

列和 (行 6) に対する割合を E11:G16 に複写

E11: =A1/A$6

E11 を E11:G16 に複写

値の複写

総計に対する割合の表 (A11:C16) を
別の場所 (A21:C26) に複写したい

A11:C16 を選択 → コピー

A21 を選択 → 貼り付けボタンの▼を
クリック → 形式を選択して貼り付け → 値

この操作を値の複写と呼ぶ

一旦ファイルを閉じる (ファイル名は EX2 としておく)

人口割合の計算

朝倉書店のサイトからダウンロードして
おいたファイルを呼び出す

H1: 表 2　地域別の人口割合の推移

A2:A16 を H2:H16 に複写

B2:F2 を I2:M2 に複写

I3: =B3/B$3

I3 を I3:M16 に複写

小数桁調整

人口増加率の計算 (1)

P_0 : 基準年の人口

P_t : t 年後の人口

r : 年当たりの人口増加率 (一定)

とすれば,

$$P_t = P_0 \cdot (1 + r)^t$$

だから人口増加率 r は

$$r = (P_t/P_0)^{1/t} - 1$$

人口増加率の計算 (2)

A31: 表 3　地域別人口増加率の推移

A32:A46 に地域名を複写

B32:F32 に 1950, 1990, 2000, 2025,2050 と入力

B33: =(C3/B3)^(1/(C$32-B$32))-1

B33 を B33:E46 に複写

B33:E46 を同じ場所に値の複写

B32:F32 を 1950-1990 年等に変更

一旦ファイルを保存 (上書き保存で良い)

第 3 章 グラフの作成

概要

- 棒グラフの作成
- 散布図 (X-Y グラフ) の作成

棒グラフの作成

ファイル → 新規作成 → 空白のブック
新しいワークシートの
A1:C5 にデータを入力

年	民間消費	政府消費
1994 年	267	72
2000 年	283	85
2003 年	282	89
2006 年	288	91

簡単な棒グラフの作成

A1:B5 を選択

挿入タブ → グラフグループ中の
縦棒ボタン → 2-D 縦棒の集合縦棒

グラフの移動, 拡大縮小

レイアウトタブ → ラベルグループ中の
各ボタンで グラフタイトル, 軸ラベル,
凡例 を変更

グラフ以外の場所をクリック (再度グラフを
編集したくなったらグラフをクリック)

複数のデータを使った棒グラフ

A1:C5 を選択
(A1 を選択した状態で, Shift を押下したまま C5 をクリックしても良い)
上と同様にして, 集合縦棒グラフを描く
(タイトルは図 2　民間消費と政府消費)
同じデータ範囲を選択
挿入 → 縦棒 → 積み上げ縦棒
(タイトルは図 3　民間消費と政府消費 (積み重ねグラフ))

連続しないデータを使ったグラフ

A1:A5,C1:C5 を選択
(A1:A5 を選択した状態で, Ctrl を押下したまま C1 から C5 までドラッグする)
(テキスト図 3.21)
上と同様にして, 集合縦棒グラフを描く
(タイトルは図 4　政府消費の推移)

散布図 (X-Y グラフ) の作成

A1:C5 を A11 へ複写
A12:A15 に半角英数で 1994, 2000, 2003, 2006 と入力
A11:B15 を選択
挿入 → 散布図
　→ 散布図 (直線とマーカー)
位置と大きさ, タイトル, 軸ラベル を適当に編集
ファイルを閉じる (ファイル名は EX3)

第 4 章 第 2 節 人口増加率の計算

人口密度と人口増加率の計算
ダウンロードしたファイルの
Table4.1 というワークシートを選択
- H2: 人口密度
- I2: 人口増加率
- H3:H80: 2000 年時点の人口密度の式
- I3:I80: 2000–2050 年の平均人口増加率の式

上書き保存して終了 (またはファイル名をテキストの通り pop2 としても良い)
次回はこの計算結果を使う

次回予定

次回はテキスト第 5–6 章をカバーする

数理手法 I 演習
第 2 回 (全 5 回)

準備:
・Mac でなく Windows を起動してください
(Mac の Excel は分析ツールを使えないため)
・演習用テキスト:
今日は第 5 章と第 6 章をカバーする
・分析ツールの組み込み
ファイルタブ → オプション → アドイン
→ 設定 → Analysis ToolPak をチェック
→ OK をクリック
データタブの Analysis グループに
Data Analysis があることを確認

第 5 章 データの並べ替え・抽出

概要

- データの並べ替え
- データの抽出
- データベース関数による統計量計算

使用するデータ

前回の演習で作業したファイルを使う

ファイル名は
1) ダウンロードしたものに
上書き保存していたら 36
2) 第 4 章にしたがって
名前を付けて保存していたら pop2

Table4.1 というワークシートを選択
H 列に人口密度, I 列に人口増加率が
計算してある

用語

フィールド名:
列見出し, データ表の一番上の行
リスト:
フィールド名とその下のデータを
合わせたもの (周囲のセルとは, 空白の
行と列で区切られている)

データの並べ替え

2000 年人口が降順になるように並べ替える

Table4.1 のリストを選択
(リスト (A2:I80) に含まれるセルのどれか
1 つをアクティブセルにする)
データ → 並べ替えとフィルタ → 並べ替え
最優先されるキーの▼をクリック
→ 2000 年人口を選択
順序の▼をクリック → 降順を選択

データの抽出

2000 年人口が 5000 万人以上の国を抽出する

リストを選択
データ → 並べ替えとフィルタ → フィルタ
(⇒ フィールド名の右に下向き矢印が付く)
2000 年人口の▼をクリック
→ 数値フィルタ → 指定の値以上
抽出条件を 50000 以上とする

抽出条件の追加 (1)

さらにアジアまたはアフリカの国にしぼる

地域の▼をクリック
→ テキストフィルタ → 指定の値に等しい
抽出条件を
アジアと等しい, OR, アフリカと等しい
とする

抽出条件の追加 (2)

さらに 2000 年人口 1 億人未満の国にしぼる

同じ手順でやると,

- 抽出条件が多いほど手順が増加する
- 抽出後に作業ミスを確認できない

⇓

抽出条件を別記する手法

一旦フィルタを解除する
(データタブのフィルタをクリック)

条件を別記した抽出 (1)

アジア・アフリカで 2000 年人口が
5000 万以上 1 億未満の国を抽出する

K2:M4 に抽出条件を記述

地域	2000 年人口	2000 年人口
アジア	>=50000	<100000
アフリカ	>=50000	<100000

同一行の条件 ⇒ AND (論理積,「かつ」)
別の行の条件 ⇒ OR (論理和,「または」)

条件を別記した抽出 (2)

リストを選択

データ → 並べ替えとフィルタ → 詳細設定

リスト範囲を確認 ⇒ "A2:I80" となってる

検索条件範囲に K2:M4 を指定
(直接入力またはセルを選択)

抽出先の指定した範囲をチェック
→ 抽出範囲に K6 を指定

データベース関数による計算

抽出の操作を省略して直接
アジアの国の 2000 年人口の合計を計算する

K17:K18 に抽出条件を記述:

地域
アジア

リストにデータ 1 と名前を付ける
(A2:I80 を選択 (Ctrl + 矢印キーが便利)
数式 → 定義された名前 → 名前の定義
名前にデータ 1 と入力)

検索条件に条件 1 と名前を付ける

K20: =DSUM(データ 1,"2000 年人口", 条件 1)
(=DSUM(A2:I80,F2,K17:K18) でも良い)

データベース関数

データベース関数の使い方:

関数名 (リスト,
"フィールド名" またはそのセル番地,
抽出条件)

関数名	説明	関数名	説明
DAVERAGE	平均	DMIN	最小値
DCOUNT	数値のセル数	DPRODUCT	積
DCOUNTA	空白でないセル数	DSTDEV	標準偏差
DGET	条件に合うある値	DSUM	合計
DMAX	最大値	DVAR	分散

第 6 章 度数分布表によるデータの分析

- 度数分布表
- 階級, 階級値
- 度数, 相対度数
- 累積度数, 累積相対度数

度数分布表の作成 (1)

新しいワークシートの挿入
(ホーム → セル → 挿入の▼
→ シートの挿入 (テキスト p.48))

以下, 新しく挿入したシートの名前が
Sheet1 であるとする

Table4.1 の A1:I80 を Sheet1 の A1 に複写

Sheet1 の F3:F80 に人口 1 と名前を付ける
(データ範囲のみで, フィールド名は含まない)

度数分布表の作成 (2)

2000 年人口の最大値と最小値を計算

K1: 2000 年人口
K2: 最小値
K3: 最大値
L2: =MIN(人口 1) ⇒ 223 万
L3: =MAX(人口 1) ⇒ 12 億 7398 万

6 つの階級を作る (等間隔ではない)

度数分布表の作成 (3)

K6:K11 に階級上限を入力

10000
50000
100000
500000
1000000
1500000

度数分布表の作成 (4)

分析ツールで度数を計算させる

データ → Analysis → Data Analysis
(なければ分析ツールを組み込む)
Histogram を選択 → OK
  Input Range に人口 1 と入力 (入力範囲)
  Bin Range に K6:K11 を指定 (データ区間)
  Output Range をチェックして K13 を指定 (出力先)

度数分布表の作成 (5)

見出しを作成する

N1: 表 1　2000 年国別人口の度数分布表
N2:T2 に階級下限, 階級上限, 階級値, 度数, 相対度数, 累積度数, 累積相対度数と入力
N3:N8 に階級下限 1000, 10000, 50000, 100000, 500000, 1000000 を入力
O3:O8 に階級上限 10000, 50000, 100000, 500000, 1000000, 1500000 を入力
P3: =(N3+O3)/2 → P3:P8 に複写

度数分布表の作成 (6)

度数から各数値を計算する

L14:L19 を Q3 に複写 (次の級の値は不要)
Q9: =SUM(Q3:Q8)
R3: =Q3/$Q$9 → R3:R8 に複写
S3: =Q3
S4: =Q4+S3 → S4:S8 に複写
S3:S8 を T3:T8 に複写
セルの幅, 文字位置, 数字の表示, 罫線等の修飾で見やすくする

ヒストグラムの作成 (1)

Q2:Q8 (度数) を選択
挿入 → グラフ → 縦棒
  → 2-D 縦棒の集合縦棒
デザイン → データ → データの選択
  横 (項目) 軸ラベルの編集ボタンをクリック
  軸ラベルの範囲に P3:P8 (階級値) を指定
  OK をクリック
OK をクリック

ヒストグラムの作成 (2)

第 3 章と同様に, グラフのタイトルを
図 1　2000 年国別人口のヒストグラムとして,
横軸のラベルを人口 (単位千人) として,
縦軸のラベルを度数とする
凡例はなしにする
位置, 大きさ, フォント等を適当に変える

※本来, ヒストグラムは棒グラフではなく, 横軸で各階級の幅を表し, その幅の長方形の面積で度数を表した図 (だが, 今は妥協する)

累積度数グラフの作成

N11: 対数人口
O11: 累積度数
N12: =LOG10(N3)
N13: =LOG10(O3) → N13:N18 に複写
O12: 0
S3:S8 (累積度数) を O13 に値の複写
N11:O18 を選択 → 挿入
  → グラフ 散布図 → 散布図 (直線)
グラフのタイトルを図 2　国別人口の累積度数グラフとして, 軸のオプション等を編集

累積相対所得グラフの作成 (1)

ローレンツ曲線を描く (所得の偏りを見る)

D2:D80 (一人当 GDP) を A85 に複写
F2:F80 (2000 年人口) を B85 に複写
新しいリスト (A85:B163) を選択 → データ
  → 並べ替えとフィルタ → 並べ替え
  一人当 GDP で昇順にする (所得が低い順)
元の GDP を計算 (10 億ドル単位)
  C85: GDP
  C86: =A86*B86/1E6 → C86:C163 に複写

累積相対所得グラフの作成 (2)

D85: 累積人口
D86: =B86/1000
D87: =D86+B87/1000 → D87:D163 に複写
E85: 累積所得
E86: =C86
E87: =E86+C87 → E87:E163 に複写
F85: 累積相対人口
G85: 累積相対所得
F86: =D86/D$163 → F86:G163 に複写

累積相対所得グラフの作成 (3)

F85:G163 を選択
挿入 → 散布図 → 散布図 (直線)

タイトルを図 3　国別人口に基づくローレンツ曲線とする

軸のオプション, ラベル等を編集

上書き保存して終了

次回予定

次回はテキスト第 7–8 章をカバーする

数理手法 I 演習
第 3 回 (全 5 回)

準備:
・演習用テキスト:
今日は第 7 章と第 8 章をカバーする

第 7 章 代表値・散らばりの尺度の計算

概要

- 分布の代表値
- 散らばりの尺度
- 元のデータからの計算
- 度数分布表からの計算

用語

分布の代表値:
平均, 中央値, 分位点
散らばりの尺度:
範囲, 四分位偏差, 分散, 標準偏差

それぞれの定義と特徴は講義の通り
ここで中央値と分位点についてのみ復習する

中央値

中央値は観測値を昇順にしたときの中央

$x_1, x_2, \ldots, x_n$: n 個の観測値
$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$: 昇順にした観測値 $(x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)})$
と書くと, 中央値 x_M:

$$x_M = \begin{cases} x_{((n+1)/2)} & (n \text{ が奇数}) \\ \frac{x_{(n/2)} + x_{(n/2+1)}}{2} & (n \text{ が偶数}) \end{cases}$$

例) $n = 5$ のとき $x_M = x_{(3)}$
$n = 4$ のとき $x_M = (x_{(2)} + x_{(3)})/2$

分位点 (1)

分位点は中央値を一般化したもの
(← 中央値を 50% 分位点として考える)

p パーセント分位点 $x_{p\%}$:

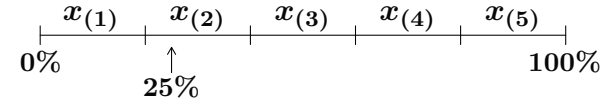
$$x_{p\%} = \begin{cases} x_{(m^*)} & (m \text{ が整数でない}) \\ \frac{x_{(m^*)} + x_{(m^*+1)}}{2} & (m \text{ が整数}) \end{cases}$$

$m := n \cdot p/100$
m^* : $m \leq m^*$ を満たす最小の整数

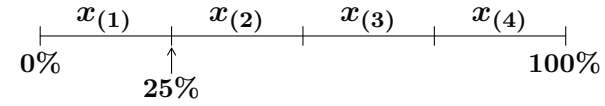
後で出るパーセント点と混同しないよう注意

分位点 (2)

$n = 5$ のとき $x_{25\%} = x_{(2)}$ $(m = 1.25)$



$n = 4$ のとき $x_{25\%} = (x_{(1)} + x_{(2)})/2$



$\{x_i\}$ を降順にしたときの 75%分位点と一致することは図より明らか

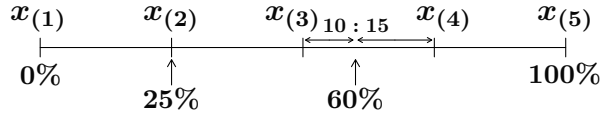
分位点 (3)

分位点 (percentile, quantile) の定義は
他にもあり, 主流といえるものはない

例えば
Excel の PERCENTILE(データ, 割合) と
QUARTILE(データ, 番号) という関数の
定義はテキストとは異なる

分位点 (4)

例えば $n = 5$ のとき PERCENTILE 関数は
下のような計算をすると思われる



$$x_{0\%} = x_{(1)}$$
$$x_{100\%} = x_{(5)}$$
$$x_{25\%} = x_{(2)}$$
$$x_{60\%} = x_{(3)} + (x_{(4)} - x_{(3)})\frac{60-50}{75-50}$$

度数分布表からの分位点計算 (1)

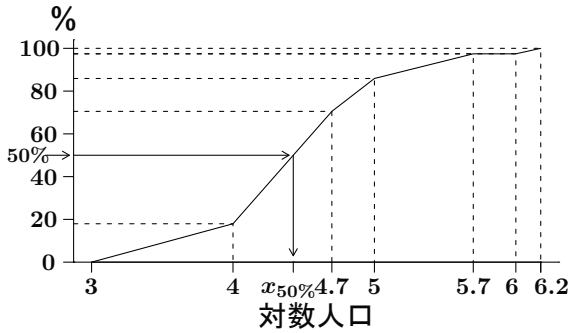
データが度数分布表としてのみ入手可能な場合は分位点として次の値を計算する

$$x^*_{p\%} = x_L + (x_U - x_L)\frac{p - R_0}{R_1 - R_0}$$

R_1 : $x^*_{p\%}$ の階級までの累積相対度数 (%)
R_0 : 下の階級までの累積相対度数 (%)
x_L : 階級下限
x_U : 階級上限

例えば前回のローレンツ曲線作成で使った度数分布表では $x^*_{50\%}$ は次頁のようになる

度数分布表からの分位点計算 (2)



図より $x^*_{50\%}$ は 2 番目の階級にあるから

$$x^*_{50\%} = 4 + (4.7 - 4)\frac{50 - 17.9}{70.5 - 17.9}$$

もとのデータからの代表値の計算 (1)

まずは Excel 関数を使わず定義通り計算する

前回のファイルを呼び出して
新しいワークシートを挿入する
(ホーム → セル → 挿入の▼ → シートの挿入)
2000 年人口をフィールド名を含めて
新しいシート (Sheet2) の A1:A79 へ複写
A2:A79 を選択して人口 2 と名前を付ける
昇順に並べ替える
(リストを選択 → データ → 並べ替え)
E1: 元のデータから

もとのデータからの代表値の計算 (2)

E2:E9 に平均, 25% 分位点, 中央値,
75% 分位点, 範囲, 四分位偏差, 分散,
標準偏差 と入力
F2: =AVERAGE(人口 2)
(=SUM(人口 2)/COUNT(人口 2) でも良い)
F3: =A21 ⇐ 下から 20 番目
F4: =(A40+A41)/2
F5: =A60 ⇐ 下から 59 番目 (上から 20 番目)
F6: =A79-A2
F7: =(F5-F3)/2

もとのデータからの代表値の計算 (3)

B1: 偏差
C1: 偏差の二乗
B2: =A2-$F$2 → B2:B79 に複写
C2: =B2^2 → C2:C79 に複写
C80: =SUM(C2:C79) → 指数表示
F8: =C80/COUNT(人口 2)
(COUNT 関数はデータの個数を計算する)
F9: =SQRT(F8)
(=F8^0.5 でも良い)

もとのデータからの代表値の計算 (4)

同じものを Excel 関数で計算してみる

G3: =QUARTILE(人口 2,1)
G4: =MEDIAN(人口 2)
G5: =QUARTILE(人口 2,3)
G6: =MAX(人口 2)-MIN(人口 2)
G7: =(G5-G3)/2
G8: =VARP(人口 2)
G9: =STDEVP(人口 2)

度数分布表からの代表値の計算 (1)

前回作成した度数分布表 (Sheet1!N1:T8) を
E11 に値の複写

平均:
L12: 階級値*相対度数
L13: =G13*I13 → L13:L18 に複写
L19: =SUM(L13:L18)

分散:
M12: 偏差の二乗*相対度数
M13: =(G13-$L$19)^2*I13 → M13:M18 に複写
M19: =SUM(M13:M18)

度数分布表からの代表値の計算 (2)

E21:E24 に四分位点, 25%, 50%, 75% と入力
F22: =E14+(F14-E14)*(0.25-K13)/(K14-K13)
F23: =E14+(F14-E14)*(0.5-K13)/(K14-K13)
F24: =E15+(F15-E15)*(0.75-K14)/(K15-K14)
H1: 度数分布表から
H2:H9 に値を複写
(範囲は
(階級上限値の最大値 − 階級下限値の最小値)
として標準偏差は分散の平方根とする)
→ 3 通りの計算結果を各自観察する

第 8 章 二次元のデータの整理・分析

- 散布図と分割表
- 共分散と相関係数

散布図と分割表

散布図
- X 軸 (横方向に分布) 原因となる変数
- Y 軸 (縦方向に分布) 説明される変数

分割表
- 表側 (縦方向に分布) 原因となる変数
- 表頭 (横方向に分布) 説明される変数

p.105 の表と
p.101 の図

	表頭 (列ラベル)	
表側 (行ラベル)	第 1 行第 1 列	第 1 行第 2 列
	第 2 行第 1 列	第 2 行第 2 列

散布図の作成 (1)

1 人当たり GDP と人口増加率の散布図を描く

ワークシートを挿入

1 人あたり GDP をフィールド名を含めて
新しいシート (Sheet3) の A1:A79 へ複写

B1: 対数 GDP
B2: =LOG10(A2) → B2:B79 に複写

人口増加率 をフィールド名を含めて
新しいシートの C1:C79 へ値の複写

散布図の作成 (2)

B1:C79 を選択

挿入 → グラフ → 散布図
→ 散布図 (マーカーのみ)

(タイトルは図 1　一人当たり GDP と
人口増加率の散布図, X 軸ラベルは対数
一人当たり GDP, Y 軸ラベルは人口増加率)

見た目を整える

分割表の作成 (1)

所得と人口増加率の分割表を作成する

D1: 所得
D2: =IF(A2<7500,"低","高")
(IF(条件,X,Y) は条件が真なら X, 偽なら Y)
E1: 人口増加
E2: =IF(C2<0.005,"低",IF(C2<0.014,"中","高"))
F1: 度数
F2: 1
D2:F2 を D3:F79 に複写

分割表の作成 (2)

リスト (A1:F79) を選択

挿入 → テーブル → ピボットテーブル
分析するデータのテーブル/範囲が
A1:F79 であることを確認
配置する場所の既存のワークシートを
チェック → 場所に A86 を指定 → OK

画面右側にピボットテーブルのフィールド
リストが現れる

分割表の作成 (3)

ピボットテーブルのフィールドリストで
所得を下方の行ラベルにドラッグ
人口増加を下方の列ラベルにドラッグ
度数を下方の Σ 値にドラッグ

(A86 以下に表が現れる)

適当な空白セルをクリックして
ピボットテーブルの編集を終了
(アクティブセルをピボットテーブル内に
移動すると再編集できる)

A85: 表 1　所得と人口増加の分割表

ピボットテーブルを使った集計

所得別人口増加率の平均の表を作成する

リストを選択
挿入 → テーブル → ピボットテーブル
配置する場所 に A102 を指定 → OK
ピボットテーブルのフィールドリストで
所得を下方の行ラベルにドラッグ
人口増加率を下方の Σ 値にドラッグ
Σ 値の合計/人口... ▼ → 値フィールドの設定
集計方法の値フィールドの集計に
平均を選択 (テーブル編集を終了)

共分散と相関係数の計算 (1)

対数 GDP と人口増加率の共分散と相関係数:
まずは Excel 関数を使わず定義通り計算する
B1:C79 を A111 に値の複写
A190: =AVERAGE(A112:A189)
→ B190 に複写
C111: x の偏差
D111: y の偏差
E111: 偏差積
C112: =A112-A$190 → C112:D189 に複写
E112: =C112*D112 → E112:E189 に複写
E190: =SUM(E112:E189)

共分散と相関係数の計算 (2)

A195:A198 に x の標準偏差, y の標準偏差,
共分散, 相関係数 と入力
B195: =STDEVP(A112:A189)
(第 7 章で標準偏差の定義通りの計算は
一度やったので Excel 関数を使う)
B196: =STDEVP(B112:B189)
B197: =E190/78
B198: =B197/(B195*B196)

共分散と相関係数の計算 (3)

同じものを Excel 関数で計算してみる

B201: =COVAR(A112:A189,B112:B189)
B202: =CORREL(A112:A189,B112:B189)
A201:A202 に見出しを付ける
上書き保存して終了

次回予定

次回はテキスト第 11 章をカバーする

数理手法 I 演習
第 4 回 (全 5 回)

準備:
- Mac でなく Windows を起動してください
  (Mac の Excel は分析ツールを使えないため)
- 演習用テキスト:
  今日は第 11 章と第 12 章をカバーする
- 分析ツールの組み込み
  ファイルタブ → オプション → アドイン
  → 設定 → Analysis ToolPak をチェ¿ ック
  → OK をクリック
  データタブの Analysis グループに
  Data Analysis があることを確認

第 11 章 正規母集団に関する推定と検定

概要

- 点推定と区間推定
- 仮説検定
- 2 つの母集団の同一性の検定
- 課題 (国別人口データを使った演習)

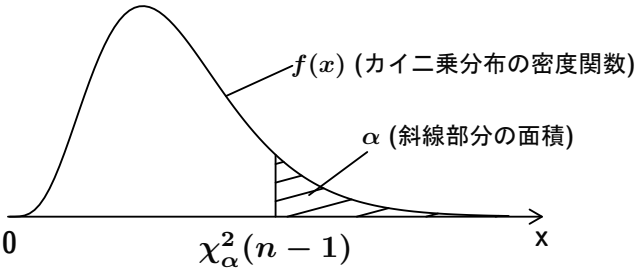
用語

母集団と標本:
  母集団の分布, 標本の分布, 確率分布

母数:
  母集団の分布がいくつかの数 (母数) で
  特定される確率分布にしたがうと仮定する
  → 分析手法が単純化
  この章では母集団の分布が母平均 μ,
  母分散 σ^2 の正規分布の場合を考える

統計量, 推定量, 推定値

パーセント点とパーセント分位点



$\chi^2_\alpha(n-1)$ は, 自由度 $n-1$ のカイ二乗分布の
上側 α ($\times 100$) パーセント点と読む

母平均と母分散の推定

1. 人口増加率の母平均の点推定, 区間推定
2. 人口増加率の母分散の点推定, 区間推定

前回のファイルを呼び出して
ワークシートを挿入
人口増加率と所得をフィールド名を含めて
新しいシート (Sheet4) の A1:B79 へ値の複写
A2:A79 を選択し人口増加率と名前を付ける

母平均の点推定, 区間推定 (1)

F1:F6 に平均, 標準偏差, 標本の大きさ (n),
自由度 (n-1), 信頼係数, パーセント点と入力
G1: =AVERAGE(人口増加率) (点推定)
G2: =STDEV(人口増加率)
G3: =COUNT(人口増加率)
G4: =G3-1
G5: 95% ($=1-\alpha$)
G6: =TINV(1-G5,G4) (他の xxxINV と異なる)
(Excel 2010 以降は xxx.INV として統一された)

母平均の点推定, 区間推定 (2)

F9: 母平均の信頼区間
F10: 幅*1/2
F11: 下限
F12: 上限
G10: =G2*G6/SQRT(G3)
G11: =G1-G10
G12: =G1+G10
セルの数字の表示を見やすくする

母分散の点推定, 区間推定 (1)

F15: カイ二乗分布
F16: 下側パーセント点
F17: 上側パーセント点
F18: 偏差の二乗和
G16: =CHIINV(1-(1-G5)/2,G4)
G17: =CHIINV((1-G5)/2,G4)
G18: =DEVSQ(人口増加率)

母分散の点推定, 区間推定 (2)

F21: 分散 (点推定)
G21: =VAR(人口増加率)
F24: 母分散の信頼区間
F25: 下限
F26: 上限
G25: =G18/G17
G26: =G18/G16

仮説検定の用語

帰無仮説: 母集団に関する仮説
対立仮説: 帰無仮説が成り立たないときは
必ず成り立つと仮定する排反な仮説
Maintained hypothesis: 検証せずに用いる仮定
(帰無仮説と対立仮説の和集合 (全事象) を定める)
棄却: 帰無仮説の下で標本が得られる確率が低い
(低さの基準を有意水準という)
「帰無仮説が正しければ」の意味
採択: 帰無仮説を棄却しない
※背理法 (帰謬法) と同じ非対称性
第 1 種の誤り: 正しい帰無仮説を棄却すること
第 2 種の誤り: 正しくない帰無仮説を棄却しないこと

母平均の両側検定 (1)

国別人口増加率の母平均が 0.7%と等しいか?
(100 年で 2 倍になる増加率)
帰無仮説 H_0: $\mu = 0.7\%$
対立仮説 H_1: $\mu \neq 0.7\%$

I1: 母平均の検定
I3:I8 に検定 1 (両側検定), 帰無仮説,
対立仮説, 有意水準, 検定統計量 t,
パーセント点 と入力

母平均の両側検定 (2)

J4: 0.7%
J5: $\neq$ 0.7% (または $<>0.7\%$)
J6: 5%
J7: =SQRT(G3)*(G1-J4)/G2
J8: =TINV(J6,G4) $\rightarrow$ |J7| $<$ J8 で採択

I9: |t|$<$ パーセント点
I10: 検定結果
J10: 帰無仮説を採択

母平均の片側検定 (1)

国別人口増加率の母平均が 1.78%より
小さいか? (1950–2000 年までの増加率)
帰無仮説 H_0: $\mu = 1.78\%$
対立仮説 H_1: $\mu < 1.78\%$

I12:I17 に検定 2 (片側検定), 帰無仮説,
対立仮説, 有意水準, 検定統計量 t,
パーセント点 と入力

母平均の片側検定 (2)

J13: 1.78%
J14: $<$1.78%
J15: 5%
J16: =SQRT(G3)*(G1-J13)/G2
J17: =TINV(2*J15,G4) $\rightarrow$ J16$<-$J17 で棄却

I18: t$<-$ パーセント点
I19: 検定結果
J19: 帰無仮説を棄却

母分散の両側検定 (1)

国別人口増加率の母分散が
$(1\%)^2 = 0.0001$ と等しいか?
帰無仮説 H_0: $\sigma^2 = (1\%)^2$
対立仮説 H_1: $\sigma^2 \neq (1\%)^2$

I22: 母分散の検定
I24:I30 に検定 3 (両側検定), 帰無仮説,
対立仮説, 有意水準, 検定統計量 chi2,
下側パーセント点, 上側パーセント点 と入力

母分散の両側検定 (2)

J25: 0.0001
J26: $\neq$ 0.0001
J27: 5%
J28: =G18/J25
J29: =CHIINV(1-J27/2,G4)
J30: =CHIINV(J27/2,G4)

I31: 下側パーセント点 $<$chi2$<$ 上側パーセント点
I32: 検定結果
J32: 帰無仮説を採択

2 つの母集団の同一性

第 1 の母集団からの標本:
$X_1, X_2, \ldots, X_m \sim N(\mu_1, \sigma_1^2)$
第 2 の母集団からの標本:
$Y_1, Y_2, \ldots, Y_n \sim N(\mu_2, \sigma_2^2)$
と仮定する

2 つの母集団が同一とは,
$\mu_1 = \mu_2, \quad \sigma_1^2 = \sigma_2^2$
ということ

2 つの母集団の同一性の検定

1. 母平均の同一性の検定
   H_0: $\mu_1 = \mu_2$ vs. H_1: $\mu_1 \neq \mu_2$
   (または H_0: $\mu_1 = \mu_2$ vs. H_1: $\mu_1 > \mu_2$)
   (a) $\sigma^2 = \sigma_1^2 = \sigma_2^2$ の場合
   (b) $\sigma_1^2 \neq \sigma_2^2$ の場合
2. 母分散の同一性の検定
   H_0: $\sigma_1^2 = \sigma_2^2$ vs. H_1: $\sigma_1^2 \neq \sigma_2^2$
   (または H_0: $\sigma_1^2 = \sigma_2^2$ vs. H_1: $\sigma_1^2 > \sigma_2^2$)

母平均の同一性の検定 (1)

低所得国 (母集団 1) の人口増加率は
高所得国 (母集団 2) のものより高いか?

1. (a) $\sigma^2 = \sigma_1^2 = \sigma_2^2$ の場合に
   $H_0 : \mu_1 = \mu_2$
   $H_1 : \mu_1 > \mu_2$
   を検定する

母平均の同一性の検定 (2)

第 5 章と同様に低所得国と高所得国を抽出
A85: 所得
A86: 低
人口増加率のリストを選択 (A1:B79)
データ → 並べ替えとフィルタ → 詳細設定
  リスト範囲を確認 ⇒ "A1:B79" となっている
  検索条件範囲に A85:A86 を指定
  (直接入力またはセルを選択)
  抽出先の指定した範囲をチェック
  → 抽出範囲に A90 を指定
A86: 高
として C90 に高所得国を同様に抽出

母平均の同一性の検定 (3)

データ → Analysis → Data Analysis
(なければ分析ツールを組み込む)
t-Test (Equal Variances) を選択 → OK
  Variable 1 Range に A91:A142 を指定
  Variable 2 Range に C91:C116 を指定
  Alpha を 0.01 に変更
  Output Range をチェックして F90 を指定
→ 片側検定で棄却 (低所得国の方が高い)

母平均の同一性の検定 (4)

1. (b) $\sigma_1^2 \neq \sigma_2^2$ の場合に
   $H_0 : \mu_1 = \mu_2$
   $H_1 : \mu_1 > \mu_2$
   を検定する (ウェルチの検定)
データ → Analysis → Data Analysis
t-Test (Unequal Variances) を選択 → OK
  Variable 1 Range に A91:A142 を指定
  Variable 2 Range に C91:C116 を指定
  Alpha を 0.01 に変更
  Output Range をチェックして F110 を指定
→ 片側検定で棄却 (低所得国の方が高い)

母分散の同一性の検定 (1)

2. $H_0 : \sigma_1^2 = \sigma_2^2$
   $H_1 : \sigma_1^2 \neq \sigma_2^2$
   を検定する
データ → Analysis → Data Analysis
F-Test (for Variances) を選択 → OK
  Variable 1 Range に A91:A142 を指定
  Variable 2 Range に C91:C116 を指定
  Alpha が 0.05 であることを確認
  Output Range をチェックして F129 を指定
← 両側検定用の値が自動的に計算されない

母分散の同一性の検定 (2)

F141: 下側パーセント点
F142: 上側パーセント点
G141: =FINV(97.5%,51,25)
G142: =FINV(2.5%,51,25)
→ 帰無仮説を棄却
上書き保存して第 11 章終了

第 12 章 2 変数間の関係の検定

第 11 章の概要
- 1 つの正規母集団の検定 (平均, 分散)
- 2 つの正規母集団の同一性検定 (平均, 分散)

第 12 章の概要
- 適合度検定 (標本と任意の分布との適合度)
  - 1 次元データと既知の分布との適合度検定
  - 2 次元データの独立性の検定 (分割表)
- 一元配置分散分析
  3 つ以上の正規母集団の同一性検定 (平均)
- 2 次元正規分布データの独立性の検定
  (相関係数を使った検定)

適合度検定

k: 母集団のカテゴリー数
帰無仮説 H_0: 「$p_1, \ldots, p_k$ は既知の値」
(p_i: 各カテゴリーの母比率 ($\sum_{i=1}^{k} p_i = 1$))
各カテゴリーの観測数が大きければ H_0 の下で

$$\sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

(f_i: 観測度数, e_i: 期待度数)
が自由度 $\nu_1 - \nu_0$ のカイ二乗分布にしたがう
(ν_0, ν_1 は帰無仮説と対立仮説の未知母数の数)
(証明略 (Rao, 1973, Ch. 6))

分割表

	人口増加			
所得	高	中	低	計
高	1	7	18	26
低	18	22	12	52
計	19	29	30	78

周辺度数

観測度数から相対度数を計算して
帰無仮説の下での期待度数を推定

$$\sum_i \sum_j \frac{(f_{ij} - e_{ij})^2}{e_{ij}} \to \chi^2((3-1)(2-1))$$

適合度検定による独立性検定 (3)

A20: 相対誤差
A12:D15 を A21:D24 に値の複写 (見出し)
B23: =(B4-B14)^2/B14 → B23:D24 に複写
A27:A30 に検定統計量 chi2, 自由度, 有意水準, パーセント点と入力
B27: =SUM(B23:D24)
B28: 2
B29: 5%
B30: =CHIINV(B29,B28)
B31: chi2> パーセント点
B32: 帰無仮説は棄却される

用語

2 次元データ:
各観測対象 ($i = 1, 2, \ldots, n$) について
2 つの変数を観測したデータ
(2 変数同時確率分布からの標本とみなす)

条件付き確率分布と周辺確率分布

独立と無相関
2 変数は独立 $\not\Rightarrow$ 2 変数は無相関

適合度検定による独立性検定 (1)

所得水準と人口増加率に関係はないか?
(所得水準が高いほど人口増加率が低いか?
という片側検定ではない)
帰無仮説 H_0: 両者の階級は独立に決まる
対立仮説 H_1: 独立でなく何らかの関係あり
前回のファイルを呼び出して
ワークシートを挿入
第 8 章で作成した所得と人口増加の分割表
(Sheet3!A85:E90) を
新しいシート (Sheet5) の A1:E6 へ値の複写

適合度検定による独立性検定 (2)

A11: 表 2　独立の場合の期待度数
C12: 人口増加
A13: 所得
A14: 高
A15: 低
B13: 高
C13: 中
D13: 低
B14: =$E4*B$6/$E$6 → B14:D15 に複写

一元配置分散分析 (1)

3 つ以上の正規母集団の母平均の同一性検定
(第 11 章の 2 つの正規母集団の同一性検定のうち,
1. (a) $\sigma^2 = \sigma_1^2 = \sigma_2^2$ の場合に対応)
地域により人口増加率が異なるか?
(各地域の母平均の同一性)
帰無仮説 H_0: $\mu_1 = \mu_2 = \cdots = \mu_s = \mu$
対立仮説 H_1: どれか 1 つは等しくない
以前のワークシート (Sheet 1) から地域と
人口増加率を Sheet5 の H1:I79 へ値の複写
(フィールド名を含める)

一元配置分散分析 (2)

K1: 地域
K2: アフリカ
M1: 人口増加率 → N1:O1 に複写
地域と人口増加率のリストを選択
データ → 並べ替えとフィルタ → 詳細設定
  リスト範囲を確認 ⇒ "H1:I79"
  検索条件範囲に K1:K2 を指定
  抽出先の指定した範囲をチェック
  → 抽出範囲に M1 を指定
M1: アフリカ (上書き)

一元配置分散分析 (3)

同様に K2 を変えてアジアとヨーロッパを
それぞれ N1 と O1 から下のセルに抽出
データ → Analysis → Data Analysis
(なければ分析ツールを組み込む)
Anova: Single Factor を選択 → OK
  Input Range に M2:O23 を指定
  Alpha が 0.05 であることを確認
  Output Range をチェックして M30 を指定
→ F>F crit だから棄却
  (地域により人口増加率は異なる)

相関係数を使った独立性検定 (1)

人口増加率と 1 人あたり GDP の
相関係数 ρ は負か?
(所得水準が高いほど人口増加率が低いか?)
帰無仮説 H_0: $\rho = 0$
対立仮説 H_1: $\rho < 0$
R1: 人口増加率と対数一人当たり GDP
R2: 標本相関係数 r
第 8 章の計算結果を Sheet3!B198 から S2 に
値の複写 (なければ Sheet3 で
B202: =CORREL(B2:B79,C2:C79) → 複写)

相関係数を使った独立性検定 (2)

R3:R7 に標本の大きさ (n), 有意水準,
自由度 (n-2), 検定統計量 t, パーセント点と入力
S3: 78
S4: 1%
S5: =S3-2
S6: =SQRT(S3-2)*S2/SQRT(1-S2^2)
S7: =TINV(2*S4,S5)
→ S6< −S7 だから棄却 (負の相関がある)
上書き保存して終了

次回予定

次回はテキスト第 13 章をカバーする

数理手法 I 演習
第 5 回 (全 5 回)

準備:
- Mac でなく Windows を起動してください
  (Mac の Excel は分析ツールを使えないため)
- 演習用テキスト:
  今日は第 13 章をカバーする
- 分析ツールの組み込み
  ファイルタブ → オプション → アドイン
  → 設定 → Analysis ToolPak をチェック
  → OK をクリック
  データタブの Analysis グループに
  Data Analysis があることを確認

第 6 章 回帰分析

概要

- 線形重回帰モデル
- 最小二乗法と最尤法による推定
- t 検定, F 検定 (尤度比検定はパス)
- あてはまりの良さと説明変数の選択
  (AIC とカルバック・ライブラー情報量)
- ダミー変数

重回帰モデル

$$Y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$$

「Y を $X_2, \ldots, X_k$ に回帰」という $(i = 1, \ldots, n)$

古典的線形回帰モデルの標準的仮定 (緑本 p.122):

6.5. $X_{1i}, \ldots, X_{ki}$ は確定した値をとる
(通常は確率変数とするが簡単のため)

6.6. $E(u_i) = 0,\ i = 1, \ldots, n$

6.7. $\forall i \neq j,\ \mathrm{Cov}(u_i, u_j) = E(u_i u_j) = 0$

6.8. $V(u_i) = E(u_i^2) = \sigma^2,\ i = 1, \ldots, n$

6.9. 説明変数は他の説明変数の線形関数で
表せない (完全な多重共線性がない)

E, V, Cov はそれぞれ期待値, 分散, 共分散を表す

最尤法とは (復習)

表の確率 p, 裏の確率 $q = 1 - p$ のコイン

p は未知 (p を推定したい)

5 回投げて (表, 表, 裏, 表, 表) を得た
$(Y_1 = 1, Y_2 = 1, Y_3 = 0, Y_4 = 1, Y_5 = 1)$

p も所与のとき, この標本を得る確率は

$$\begin{aligned} L(Y_1, \ldots, Y_5, p) &= \prod_{i=1}^{5} p^{Y_i} (1-p)^{1-Y_i} \\ &= p^4 (1-p) \end{aligned}$$

と書ける

最尤法とは (復習)

$p = 0.2$ と $p = 0.8$ では

$L(Y_1, \ldots, Y_5, 0.2) = 0.2^4 \cdot 0.8 = 0.00128,$

$L(Y_1, \ldots, Y_5, 0.8) = 0.8^4 \cdot 0.2 = 0.08192$

より $p = 0.8$ の方がもっともらしい

→ 尤度 $L(p|Y_1, \ldots, Y_5) = L(Y_1, \ldots, Y_5, p)$
(関数の変数を 2 組に分けて書いただけ)

尤度を最大にするものを
推定量とするのが最尤法

通常, その値が
所与である変数を
“|” の右に書く

p は推定したい係数 (パラメーターともいう),
$Y_1, \ldots, Y_5$ の値はデータである

回帰モデルにおける最尤法 (1)

重回帰方程式:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i,$$
$$i = 1, 2, \ldots, n$$

誤差項の正規性の仮定を追加 (緑本 p.122, 125):

6.5. $X_{1i}, \ldots, X_{ki}$ は確定した値をとる
(通常は確率変数とするが簡単のため)

6.6. $E(u_i) = 0,\ i = 1, \ldots, n$

6.7. $\forall i \neq j,\ \mathrm{Cov}(u_i, u_j) = E(u_i u_j) = 0$

6.8. $V(u_i) = E(u_i^2) = \sigma^2,\ i = 1, \ldots, n$

6.9. 説明変数は他の説明変数の線形関数で
表せない (完全な多重共線性がない)

6.2.3.b. u_i は正規分布にしたがう (正規性の仮定)

回帰モデルにおける最尤法 (2)

単回帰 $(k = 2)$ の例

u_i の分布が各 i について独立で

$$u_i = Y_i - (\beta_1 + \beta_2 X_i) \sim N(0, \sigma^2)$$

つまり Y_i の分布が各 i について独立で

$$Y_i \sim N(\beta_1 + \beta_2 X_i, \sigma^2)$$

と仮定する

X_i を確率変数でもなく
係数でもないと仮定したから
書かなくてよい (仮定 6.5)

β_1, β_2, σ^2 が既知のとき Y_i の確率密度関数は

$$\begin{aligned} f(Y_i|X_i, \beta_1, \beta_2, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{(Y_i - \beta_1 - \beta_2 X_i)^2}{2\sigma^2} \right\} \end{aligned}$$

と書ける (正規分布の密度関数)

回帰モデルにおける最尤法 (3)

尤度関数は

$$\begin{aligned} &L(\beta_1, \beta_2, \sigma^2 | X_1, \ldots, X_n, Y_1, \ldots, Y_n) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ \frac{(Y_i - \beta_1 - \beta_2 X_i)^2}{2\sigma^2} \right\} \end{aligned}$$

暗黙に $\sigma > 0$ を仮定している
($\sqrt{2\pi\sigma^2}$ と書いてもよい)

対数尤度関数は

$$\begin{aligned} &\log L(\beta_1, \beta_2, \sigma^2 | X_1, \ldots, X_n, Y_1, \ldots, Y_n) \\ &= -n(\log\sqrt{2\pi} + \log\sigma) \\ &\quad - \sum_{i=1}^{n} \left\{ \frac{(Y_i - \beta_1 - \beta_2 X_i)^2}{2\sigma^2} \right\} \end{aligned}$$

回帰モデルにおける最尤法 (4)

・最小二乗法
残差の平方和 (二乗和) を最小化
・最尤法
対数尤度を最大化
β の推定量を最小二乗法と同じ形で与える
同時に誤差項の分散 σ^2 の推定量も与える

$$\begin{aligned}\widehat{\beta_2} &= \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2} = b_2\\ \widehat{\beta_1} &= \bar{Y}-\widehat{\beta_2}\bar{X} = b_1\\ \widehat{\sigma^2} &= \frac{\sum e_i^2}{n}\end{aligned}$$

推定量なので "^" を付けて
真の係数 $\beta_1, \beta_2, \sigma^2$ と区別する

$\log L(\widehat{\beta_1}, \widehat{\beta_2}, \widehat{\sigma^2}|Y_1, \ldots, Y_n)$ を対数最大尤度という

重回帰分析における検定 (1)

・ t 検定
帰無仮説は, 単独の係数に関する線形制約

帰無仮説 H_0: $\beta_j = a$ の下で検定統計量

$$t = \frac{b_j - a}{\text{s.e.}(b_j)}$$

「帰無仮説が正しければ」の意味

は自由度 $(n-k)$ の t 分布にしたがう (証明略)
ここで　緑本第 7 章に書かれているべきだが抜けている
β_j: j 番目の係数の真の値
s.e.(b_j): β_j の最小二乗推定量 b_j の標準誤差
$(= s^2(X'X)^{-1}$ の第 j 対角成分の平方根$)$

重回帰分析における検定 (2)

・ F 検定
帰無仮説は, 複数の線形制約からなる

帰無仮説が正しいときの残差平方和, つまり
帰無仮説を制約とした最小二乗推定の
残差平方和は, 制約が無い場合の残差平方和と
比べて大幅に大きくはないはず

⇓

残差平方和を比較する

「残差平方和の差が大きければ帰無仮説は疑わしい」
ということ

重回帰分析における検定 (3)

(F 検定続き) 帰無仮説 H_0 の下で検定統計量

$$F = \frac{(S_0 - S_1)/p}{S_1/(n-k)}$$

は自由度 $(p, n-k)$ の F 分布にしたがう (証明略)
ここで
S_0: H_0 の下での残差平方和
S_1: H_1 の下での残差平方和
n: 観測数
k: 係数の数
p: 帰無仮説に含まれる式の数
(原理上, 片側検定のような対立仮説は考えられない)

説明変数選択とモデルの当てはまり (1)

誤ったモデルで推定する 2 つの例

1. 説明変数が足りないモデルで推定した場合

真のモデル: $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$
推定したモデル: $Y_i = \gamma_1 + \gamma_2 X_{2i} + v_i$
$\to \widehat{\gamma_1}, \widehat{\gamma_2}$ は不偏でない ("^" は推定量を表す)

2. 説明変数が多いモデルで推定した場合

真のモデル: $Y_i = \beta_1 + \beta_2 X_{2i} + u_i$
推定したモデル: $Y_i = \gamma_1 + \gamma_2 X_{2i} + \gamma_3 X_{3i} + v_i$
$\to \widehat{\gamma_i}$ は不偏だが $V(\widehat{\gamma_i}) \geq V(b_i)$ $(i = 1, 2)$

説明変数選択とモデルの当てはまり (2)

(2. の続き)
パラメーター θ (スカラー) の推定量 $\widehat{\theta}$ について
$\widehat{\theta}$ の偏り (bias):

$$E(\widehat{\theta} - \theta) = E(\widehat{\theta}) - \theta$$

$\widehat{\theta}$ の平均二乗誤差 (Mean Squared Error):

$$\text{MSE} = E((\widehat{\theta} - \theta)^2)$$

上式は次のように変形できる

$$\text{MSE} = (E(\widehat{\theta}) - \theta)^2 + V(\widehat{\theta})$$

⇒ 不偏であっても分散が大きいと MSE が大きい

1 と 2 より最適な説明変数を選ぶことが重要といえる

モデル選択と (修正) 決定係数 (1)

決定係数:
最小二乗推定のあてはめ値 $\widehat{Y}_i$ と残差 e_i に関して

$$\sum(Y_i - \bar{Y})^2 = \sum(\widehat{Y}_i - \bar{Y})^2 + \sum e_i^2$$

TSS (Total) = ESS (Explained) + RSS (Residual)

が成り立つことより決定係数 R^2 を

$$R^2 = 1 - \frac{\sum e_i^2}{\sum(Y_i - \bar{Y})^2} = \frac{\sum(\widehat{Y} - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$$

と定義した ($0 \leq R^2 \leq 1$ がいえる)
R^2 はモデルとデータのあてはまりのよさを表し,
説明変数を追加すれば必ず増加する
$\to$ モデル選択に不適

モデル選択と (修正) 決定係数 (2)

修正決定係数:
説明変数の数の違いを考慮するために
説明変数の数の増加にペナルティを課した $\bar{R}^2$

$$\bar{R}^2 = 1 - \frac{\sum e_i^2/(n-k)}{\sum(Y_i - \bar{Y})^2/(n-1)}$$

(記号の印象と異なり, 何かの 2 乗ではない)
上式より

$$\bar{R}^2 \text{ 最大} \iff s^2 \text{ 最小}$$

$\to$ モデル選択に使う根拠にはならない
(σ^2 を最も小さく推定するモデルが最良?)

モデル選択と AIC, BIC

最適な説明変数の組み合わせを選ぶ基準としては
AIC (赤池の情報量基準) と
BIC (Schwartz の Bayes 情報量基準)
の利用が一般的
次のように定義されることが多い

$\text{AIC} := -2\log L + 2v$

$\text{BIC} := -2\log L + (\log n)v$

$\log L$：対数最大尤度

v：モデルに含まれる未知パラメーターの数

AIC または BIC を最小にするモデルを選択する

ダミー変数 (1)

ダミー変数: 0 または 1 をとる変数

例えば女性を 0, 男性を 1 として性別を表せる

$$D_i = \begin{cases} 0 & \text{(観測 } i \text{ が女性のものの場合)} \\ 1 & \text{(観測 } i \text{ が男性のものの場合)} \end{cases}$$

← 男性ダミーという

賃金と勤続年数の関係のモデル

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

Y_i：賃金

X_i：勤続年数

において 3 通りの導入例を示す

ダミー変数 (2)

例 1)

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 D_i + u_i$$

D_i：上で定義した男性ダミー

これは, 男女の初任給の差を導入したモデル

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad \text{(女性)}$$

$$Y_i = \beta_1^* + \beta_2 X_i + u_i \quad \text{(男性)}$$

と同一である

ダミー変数は通常の変数と同じに扱える

「$\text{H}_0 : \beta_3 = 0$」 として賃金差を検定できる

ダミー変数 (3)

例 2)

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 D_i X_i + u_i$$

これは, 男女で初任給は等しく, その後の
賃金上昇率が異なるモデル

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad \text{(女性)}$$

$$Y_i = \beta_1 + \beta_2^* X_i + u_i \quad \text{(男性)}$$

と同一である

$Z_i = D_i X_i$ なる説明変数を追加して推定できる

「$\text{H}_0 : \beta_3 = 0$」 として賃金上昇率の差を
検定できる

ダミー変数 (4)

例 3)

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 D_i + \beta_4 D_i X_i + u_i$$

これは, 初任給, 賃金上昇率ともに異なるモデル

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad \text{(女性)}$$

$$Y_i = \beta_1^* + \beta_2^* X_i + u_i \quad \text{(男性)}$$

と同一である

つまり, 男女別々に推定を行うのと同じ

「$\text{H}_0 : \beta_3 = \beta_4 = 0$」 として F 検定により
男女間の差を検定できる

ダミー変数 (5)

ダミー変数の使用に関する注意

上の性別ダミーの例では誤りにくいが, ダミー変数で
表したい区分が多いとき次のような誤りを犯しやすい

誤りの例

四季を区別するために春夏秋冬の 4 つの
ダミー変数を説明変数に入れる

⇓⇓

春夏秋ダミーにより冬ダミーを表現できてしまう

これは回帰の標準的な仮定 6.9 に反する

(「説明変数に完全な多重共線性がある」という)

($X'X$ が特異行列になり推定不能になる (緑本 7 章))

演習: 人口増加率の単回帰分析 (1)

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

Y_i：人口増加率$_i$

X_i：$\log_e$(一人当 GDP$_i$)

$$\beta_2 = \frac{\partial E(Y_i)}{\partial X_i} = \frac{\partial E(Y_i)}{\partial \text{一人当 GDP}_i} \frac{\partial \text{一人当 GDP}_i}{\partial X_i}$$

$$= \frac{\partial E(Y_i)}{\partial \text{一人当 GDP}_i / \text{一人当 GDP}_i}$$

← 一人当 GDP 1%増加に対する人口増加率の変化

前回のファイルを呼び出してワークシートを挿入
Sheet1 から人口増加率, 一人当 GDP, 人口密度,
地域の順に新しいシート (Sheet6) の A1:D79 へ
値の複写 (フィールド名を含む)

人口増加率の単回帰分析 (2)

E1: 対数一人当 GDP

E2: =LN(B2) → E2:E79 に複写

データ → Analysis → Data Analysis
(なければ分析ツールを組み込む)

Regression を選択 → OK

- Input Y Range に A1:A79 を指定
- Input X Range が E1:E79 を指定
- Labels をチェック ← フィールド名あり
- Output Range をチェックして A85 を指定

演習: 人口増加率の単回帰分析 (3)

推定結果 (かっこ内は標準誤差)

$$\underset{(0.003431)}{Y_i = 0.03010} \underset{(0.000423)}{- 0.002818 X_i}$$

→ 1 人あたり GDP が 1%増加すると
人口増加率は約 0.28% ポイント減少する
予想では人口増加率は所得水準と逆に動くから,
帰無仮説 H_0: $\beta_2 = 0$ ＼テキスト誤り
対立仮説 H_1: $\beta_2 < 0$
有意水準 α: 1%
として片側検定を行うと =TINV(2%,78-2)
$-6.666 < -t_\alpha(n-k) = -2.376$ より棄却

演習: 人口増加率の重回帰分析 (2)

データ → Analysis → Data Analysis
Regression を選択 → OK
Input Y Range に A1:A79 を指定
Input X Range が E1:G79 を指定
Labels をチェック ← フィールド名あり
Output Range をチェックして A105 を指定
推定結果 (かっこ内は標準誤差)

$$\underset{(0.005201)}{Y_i = 0.02503} \underset{(0.000514)}{- 0.001979 X_{2i}} + \underset{(0.000579)}{0.000717 X_{3i}} \underset{(0.00214)}{- 0.00516 X_{4i}}$$

$s = 0.006474$ (=B111=SQRT(C117/B117))

演習: 人口増加率の重回帰分析 (4)

2. すべての説明変数が Y を説明しないという
帰無仮説の検定 (有意水準 α は 1%)
H_0: $\beta_2 = \beta_3 = \beta_4 = 0$
$F = 19.112$ (=E116)
$F_\alpha(k-1, n-k) = 4.058$ (=FINV(1%,4-1,78-4))
より棄却 (つまり説明変数は Y を説明している)

AIC と Kullback-Leibler 情報量 (1)

・確率分布の「近さ」
あるコインを投げたときの表の確率を予想する
予想 A は 0.7, 予想 B は 0.5 とする

真の確率が 0.4 のとき, B の方が良いとしてよさそう

それでは
真の確率が 0.6 のときはどちらを良しとするべきか?

真のモデルからのずれを表す基準が欲しい
⇓⇓
確率分布の情報量という概念を使う

演習: 人口増加率の重回帰分析 (1)

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$$

Y = 人口増加率
$X_2 = \log_e$(一人当 GDP)
$X_3 = \log_e$(人口密度)
X_4 = アフリカが 1, それ以外がゼロのダミー

F1: 対数人口密度
F2: =LN(C2) → F2:F79 に複写
G1: アフリカダミー
G2: =IF(D2="アフリカ",1,0) → G2:G79 に複写
(IF(条件,a,b): 条件が真なら a, 偽なら b を計算)

演習: 人口増加率の重回帰分析 (3)

1. β_2, β_3, β_4 に関する個別の帰無仮説の検定
H_0: $\beta_2 = 0$, H_1: $\beta_2 < 0$
H_0: $\beta_3 = 0$, H_1: $\beta_3 < 0$
H_0: $\beta_4 = 0$, H_1: $\beta_4 > 0$
有意水準 α は 5%とする
$t_2 = -3.851$, $t_3 = -1.238$, $t_4 = 2.408$,
$t_\alpha(n-k) = 1.666$ (=TINV(10%,78-4))
より順に棄却, 採択, 棄却となる

演習: 人口増加率の重回帰分析 (5)

3. 人口密度とアフリカダミーの有効性の検定
H_0: $\beta_3 = \beta_4 = 0$
H_1: 少なくともどちらか 1 つはゼロでない
有意水準 α は 5%とする C97 C117
帰無仮説の下では, 説明変数が対数一人当 GDP の
単回帰モデルと同じだから, 前の推定結果から
$F = \frac{(S_0 - S_1)/p}{S_1/(n-k)} = \frac{(0.003474 - 0.003102)/2}{0.003102/74} = 4.441$,
$F > F_\alpha(p, n-k) = 3.120$ (=FINV(5%,2,78-4))
より棄却 (S_0, S_1 は帰無仮説と対立仮説の下での
残差平方和, p は帰無仮説の制約の数)

AIC と Kullback-Leibler 情報量 (2)

事象の情報量と確率分布の情報量
・20 の扉: 二択の質問を繰り返して答えを絞っていき,
20 回以内の質問で当てるゲーム
「それは食べられる」 v.s. 「それはお菓子である」
前者より後者の方が多くの情報を含むという直感と
整合的な「情報の量」を定義する
つまり事象の情報量を「起こりにくさ」で決め,
確率分布の情報量を「平均的起こりにくさ」で決める
具体的には次の定義が導かれる (足せるように対数)
確率 p_i の事象の情報量 : $-\log p_i$
(離散) 分布の平均情報量 : $E(-\log p_i) = -\sum_{i=1}^{n} p_i \log p_i$

AIC と Kullback-Leibler 情報量 (3)

カルバック・ライブラー (KL) 情報量 KL:

$$\mathrm{KL} = E_g \log \frac{g}{f} = E_g \log g - E_g \log f$$

g：真のモデルの密度関数
f：分析対象とするモデルの密度関数

ただし, E_g は g による期待値を表す
$E_g \log f$ は平均対数尤度と呼ばれる

このとき以下が成り立つ (証明略)

(i) $\mathrm{KL} \geq 0$
(ii) $\mathrm{KL} = 0 \Leftrightarrow$「$f$ と g が一致する」

$\longrightarrow$ KL 情報量は真のモデルからのずれを表す

AIC と Kullback-Leibler 情報量 (4)

例えば上のコイン投げの例で KL 情報量を使えば,
表が出る真の確率が 0.6 のとき

$$\mathrm{KL_A} = 0.6 \log \frac{0.6}{0.7} + 0.4 \log \frac{0.4}{0.3} \approx 0.0226$$

$$\mathrm{KL_B} = 0.6 \log \frac{0.6}{0.5} + 0.4 \log \frac{0.4}{0.5} \approx 0.0201$$

より, 予想 B の方が真のモデルに近いといえる

AIC と Kullback-Leibler 情報量 (5)

KL 情報量の定義において $E_g \log g$ は一定だから

KL を最小化 $\Leftrightarrow$ 平均対数尤度 $E_g \log f$ を最大化

g を知らないから平均対数尤度を計算できないが,
平均対数尤度を対数最大尤度で推定すればどうか?

$\longrightarrow$ バイアスがあるが n が大きいとき補正できる

モデル f が v 個の自由なパラメーターを持ち
それらを適当にとれば真の分布を表せると仮定すると

$$E_g \log f \approx (\log L - v)/n = \mathrm{AIC}/(-2)$$

とバイアスを補正して近似できる (証明略)

つまり AIC 最小のモデルは KL 情報量最小といえる

参考文献

- 坂元慶行・石黒真木夫・北川源四郎 (1983)「情報量統計学」共立出版.
- 縄田和満 (2009)「EViews による計量経済分析入門」朝倉書店.