計量経済学各論 (開発経済の計量分析) 課題 1

新保一成

2016 年 *5* 月 *27* 日

Contents

課題

下の表に示すように，合計特出生率の低下とともに平均寿命が上昇してきたことが確認される．その背景には，開発途上国における 5 歳未満児童死亡率の急激な減少があった．

地域	平均寿命		合計特殊出生率		5 歳未満児童死亡率	
	1950-55	2010-15	1950-55	2010-15	1950-55	2010-15
世界	46.8	70.5	4.96	2.51	215.1	49.6
先進国	64.7	78.3	2.82	1.67	77.7	6.3
開発途上国	41.5	68.6	6.08	2.65	247.7	54.3

以下に 1950-55 年から 2010-15 年の期間における平均寿命，合計特殊出生率，5 歳未満児童死亡率の変化をグラフで示すとともに，平均寿命 (e_0) を非説明変数，5 歳未満児童死亡率 (q_5) を説明変数とする回帰モデル (u は誤差項で独立に $N(0, \sigma^2)$ に従うものとする)

$$\log e_0 = \alpha + \beta \log q_t + u$$

を最小 2 乗法で推定して，5 歳未満児童死亡率の 1% の低下が平均寿命を何% 減少させるかを示す係数 β の値を推定する方法を説明する．

国連経済社会局人口部は国連世界人口予測 (UN World Population Prosepect) において 2100 年までの人口予測を公表している．その中位推計において，平均寿命，合計特殊出生率，5 歳未満児童死亡率の関係がどのようになっているかを 2015-2050 年の期間と 2050-2100 年の期間に分けて，以下に示す方法と同様の方法で分析し，2 つの期間における予測の違いについて記述しなさい．

提出期日と提出方法

- 提出期日: 2016 年 6 月 16 日 (木) まで
- keio.jp から PDF ファイルで提出
- PDF ファイルは 1 つにまとめる．複数の PDF ファイルを提出しない．
- 用紙サイズは A4 とする．

分析方法

必要なパッケージの導入

```
library(sqldf)
library(dplyr)
library(ggplot2)
```

データ

新保研 Web データのページから以下のファイルを RStudio のプロジェクト・ディレクトリ (フォルダ) にダウンロードする．

- 「国・地域」(ファイル番号 6) のデータファイル WPP2015_F01_Locations.dsv
- 「期間指標」(ファイル番号 1) のデータファイル WPP2015_DB01_Period_Indicators.dsv

データの読み込み

データファイルには，世界，先進国，開発途上国，アジア，ヨーロッパなどの地域と国別のデータが収録され，それは国・地域コード (LocID) で識別できる．今回の課題では国別データのみを使用する．国・地域ファイルの LocationTypeCode が 4 のとき，それが国であることがわかる．最初に国・地域ファイルを読み込み，そこから LocationTypeCode = 4 の LocID を抜き出し，それを条件として期間指標ファイルを読み込む．

国・地域情報の読み込み

read.table で国・地域ファイルのデータを全て読み込む．

```
location <- read.table(" WPP2015_F01_Locations.dsv" , sep = " |" ,
                       header = TRUE, stringsAsFactors = FALSE)
```

このファイルの 1 行目は，変数名からなるヘッダなので head = TRUE を指示する．また国名などの文字列を R の因子オブジェクトに自動変換しないように stringsAsFactors = FALSE を指示する．

期間指標ファイルの読み込み

sqldf パッケージの read.csv.sql 関数で WPP2015_DB01_Period_Indicators.dsv から使用するデータを指定して読み込む．データの抽出条件は以下のとおり．- 予測シナリオは中位推計 (VarID = 2) - 国・地域ファイルにおいて LoccationTypeCode = 4 の LocID に対応する国 2 番めの条件を SQL に指示するためには where 以下の LocID in() のカッコ内に対応する LocID を列挙しなければならない．国・地域ファイルに含まれる国数 (LocationTypeCode = 4) は 233 もあるのに対し，地域数 (LocationTypeCode != 4) は 40 しかない．

```
nrow(filter(location, LocationTypeCode == 4))
```

```
## [1] 233
```

```
nrow(filter(location, LocationTypeCode != 4))
```

```
## [1] 40
```

233 の LocID を列挙するよりも，40 だけ列挙する方が楽であり，条件が少ないほうがデータを読み込む効率もよい．LocID in (LocationTypeCode == 4 の LocID) は，LocID not in(LocationTypeCode != 4) と同じである．

まず，国・地域情報から LocationTypeCode が 4 でないものを抜き出す．

```
notCountry <- filter(location, LocationTypeCode != 4)
```

次に以下のようにして，期間指標ファイルから LocationTypeCode == 4 の国の中位推計データを読み込む．

```
locID <- paste(notCountry$LocID, collapse = ",")
sql <- paste("select LocID, Location, Time, MidPeriod, TFR, LEx, Q5 from file",
             "where VarID = 2 and LocID not in(", locID, ")")
wpp <- read.csv.sql("WPP2015_DB01_Period_Indicators.dsv",
                    sep = "|", eol = "\n", sql = sql)
```

1. 1 行目と 2 行目で使っている **paste** 関数は，文字列を結合する関数．
2. notCountry$LocID は，40 個の整数からなるベクトルである．これを paste 関数に渡して，collaps = "," を指定することで，ベクトルの要素をカンマで区切られた 1 つの文字列に潰してしまう．
3. 2 行値で SQL を作る．ここでは 4 つの文字列を paste で結合している．1 番目，2 番目と 3 番めは二重引用符で囲まれた文字列で，3 番めが 1 行で作った LocID からなる文字列である．こうすることで LocID no in() のカッコ内に，LocTypeCode != 4 の LocID を書き込むことができる．
4. 読み込む指標は以下のとおり．
    - LocID(国コード)
    - Location(国名)
    - Time(期間，"1950-1955" から "2095-2100" までの 5 年期間の文字列)
    - MidPeriod(期間の年央，"2010-2015" の場合 2013
    - TFR(合計特殊出生率)
    - LEx(平均寿命)
    - Q5(5 歳未満児童死亡率, 生存 1000 人あたり死亡数)

読み込んだデータを確認する

```
head(wpp)
```

```
##   LocID    Location      Time MidPeriod  TFR   LEx      Q5
## 1     4 Afghanistan 1950-1955      1953 7.45 28.59 406.850
## 2     4 Afghanistan 1955-1960      1958 7.45 31.11 374.505
## 3     4 Afghanistan 1960-1965      1963 7.45 33.39 346.918
## 4     4 Afghanistan 1965-1970      1968 7.45 35.58 321.696
## 5     4 Afghanistan 1970-1975      1973 7.45 37.83 296.874
## 6     4 Afghanistan 1975-1980      1978 7.45 40.39 269.943
```

dplyr でデータを整理する

欠損値の処理

全ての国について，合計特殊出生率，平均寿命，5 歳未満児童死亡率が利用可能とは限らない．全ての期間について，3 指標が利用できる国のデータにする．そのために 3 指標のうちいずれかが利用できない国のリストを作成する．

```
omitLocID <- wpp %>%
  filter(TFR == 0 | LEx == 0 | Q5 == 0) %>%
  distinct(LocID) %>%
  select(LocID)
```

- 2 行目で TFR(合計特殊出生率) または LEx(平均寿命) または Q5(5 歳未満児童死亡率) がゼロであるレコードのみを残している．
- 残ったレコードには同じ国の複数期間のデータが含まれているので，2 行目で LocID の重複をなくし，3 行目で LocID だけをデータフレームの残すことを指示している．

wpp$LocID が読み込んだデータに含まれる全ての国コードで，omitCountry$LocID が取り除く国コードであるから，その差集合が残すべき国コードになる．それは，setdiff(wpp$LocID, omitLocID$LocID) で得られる．以下，- 2 行目で全期間についてデータが利用可能な国に限定し，- 3 行目で国情報に含まれている DeVGrp(1 = 開発途上国，2 = 先進国) を LocID をキーにしてマージしている．

```
wpp <- wpp %>%
  filter(LocID %in% setdiff(wpp$LocID, omitLocID$LocID)) %>%
  left_join(select(location, LocID, DevGrp))
```

データを観測期間 (1950-2015 年) と予測期間 (2015-2050 年と 2050-2100 年) に分割する

2015-2050 年の予測期間は年央 (MidPeriod) が 2013 le MidPeriod le 2048 のレコードで (1 行目)，2050-2100 年の予測期間は年央 (MidPeriod) が 2048 ge MidPeriod のレコードである (2 行目)．当初 wpp には全期間のレコードが収録されていたが，3 行目で観測期間 (MidPeriod le 2013) だけのレコードからなるデータフレームにしている．

```
wpp2050 <- filter(wpp, MidPeriod >= 2013 & MidPeriod <= 2048)
wpp2100 <- filter(wpp, MidPeriod >= 2048)
wpp <- filter(wpp, MidPeriod <= 2013)
```

合計特殊出生率，平均寿命，5 歳未満児童死亡率の関係を視覚化する

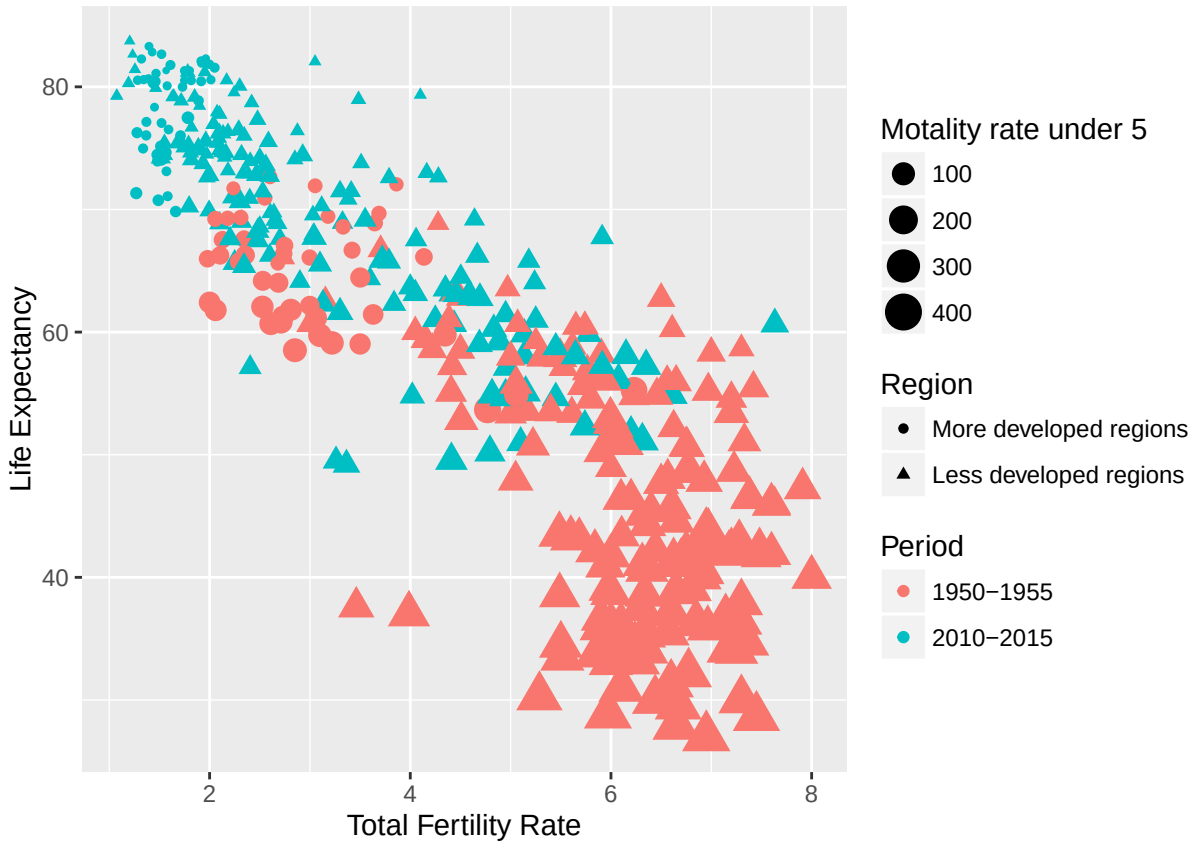
3 つ指標を同時に視覚化するために，合計特殊出生率 (x 軸)，平均寿命 (y 軸) の散布図の点の大きさが 5 歳未満児童死亡率に比例するようにグラフを描いてみよう．3 つの異なる期間についてグラフが描けるように，内容の異なる同じ形式のデータを視覚化する関数を作る．plotTfrLexQ5 <- function(wpp) で，plotTfrLexQ5 は wpp を引数とする関数であることを宣言し，{} で囲まれた範囲に関数の定義を与える．この関数では，- データフレーム wpp に含まれる全ての期間のうち，最初と最後の期間について (2 行目)，- x 軸に TFR(合計特殊出生率)，y 軸に LEx(平均寿命)，- 色の違いで期間を区別し (colour = Time)，- 点の形で開発途上国か先進国かを区別し (shape = factor(DevGrp)，- 点の大きさを Q5(5 歳未満児童死亡率) に比例させて (size = Q5)(3 行目)，- 散布図を描く (4 行目) ように指示している．

```
plotTfrLexQ5 <- function(wpp) {
  wpp <- filter(wpp, MidPeriod == min(MidPeriod) | MidPeriod == max(MidPeriod))
  ggplot(wpp, aes(x = TFR, y = LEx, colour = Time, shape = factor(DevGrp), size = Q5)) +
    geom_point() +
    labs(x = " Total Fertility Rate" , y = " Life Expectancy" ,
         shape = " Region" , colour = " Period" ,
     size = " Motality rate under 5" ) +
    scale_shape_discrete(labels=c(" More developed regions" , " Less developed regions" ))
}
```

1950–2015 年の観測期間についてグラフを描くには次のようにする．

```
plotObs <- plotTfrLexQ5(wpp)
print(plotObs)
```

```
## Warning: 使われていないコネクション 6 (WPP2015_DB01_Period_Indicators.dsv)
## を閉じます
```



グラフを “plotObs.pdf” というファイル名の PDF ファイルに保存するには以下のようにする.

```
ggsave(" plotObs.pdf" )
```

```
## Saving 6.5 x 4.5 in image
```

回帰分析

平均寿命 (e_0) を非説明変数，5 歳未満児童死亡率 (q_5) を説明変数とする回帰モデル (u は誤差項で独立に $N(0, \sigma^2)$ に従うものとする)

$$\log e_0 = \alpha + \beta \log q_t + u$$

を最小 2 乗法で推定するためには，以下のようにする.

```
olsObs <- summary(lm(log(LEx)~log(Q5), data = wpp))
```

推定結果の詳細が olsObs に保存されるので，それを出力する.

```
olsObs
```

```
##
## Call:
## lm(formula = log(LEx) ~ log(Q5), data = wpp)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.08230 -0.05914  0.02464  0.07647  0.21065
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.778209   0.007414  644.48   <2e-16 ***
## log(Q5)     -0.166652   0.001754  -95.03   <2e-16 ***
## ---
## Signif. codes:  0 ’ ***’  0.001 ’ **’  0.01 ’ *’  0.05 ’ .’  0.1 ’  ’  1
##
## Residual standard error: 0.105 on 2611 degrees of freedom
## Multiple R-squared:  0.7757, Adjusted R-squared:  0.7756
## F-statistic:  9031 on 1 and 2611 DF,  p-value: < 2.2e-16
```