

平成 28 年度社会医学実習-実施例

王 超辰

2016 年 6 月 30 日

- 1 例:日本人 (40 歳以上) 肝がん罹患の年齢,出生コホート,時期効果分析
- 2 必要となるパッケージ:

```
install.packages("XLConnect") # エクセルファイルのデータを読み込む用
install.packages("ggplot2") # グラフ作成用
install.packages("reshape2") # データ操作する用
# ダウンロードされたエクセルファイルの 4 番目の sheet"rate" だけ抽出し,
# 名前を付けてデスクトップに保存する. 例: "cancer_incidence(1975-2011)rate.xls"
setwd("C://Users/xxxxxx/Desktop") # 自分の PC のデスクトップのアドレスに書き換える.
# 通常は"C://Users/自分のユーザー名/Desktop/", (円マーク"¥" をスラッシュ"/" に変更)
```

3 データのマネージメント

```
library(XLConnect) # パッケージのローディング

rate.all <- readWorksheetFromFile("cancer_incidence(1975-2011)rate.xls",
                                sheet = 1)

# さっき抽出された, がん罹患データを R に読み込む
# 変数名の変更:
names(rate.all) <- gsub("X", "age", names(rate.all))
# ここから漢字の前に半角の空白" " が出てくるが, 無視して (半角空白なしで) 書くことにする.
names(rate.all) <- gsub(" 歳", "", names(rate.all))
names(rate.all) <- gsub(" 以上", "plus", names(rate.all))
names(rate.all) <- gsub(" 診断年", "Dia_yr", names(rate.all))
names(rate.all) <- gsub("\\\\.", "_", names(rate.all))
# チェックするために, 最初の 6 行を示す
head(rate.all)

##   コード   部位   ICD_10   性別   Dia_yr   粗率   age0_4   age5_9   age10_14
## 1      1  全部位   C00-C96   男女計   1975  184.6549  13.24920  8.021910  7.944879
## 2      1  全部位   C00-C96   男女計   1976  185.0981  13.14640  7.311146  7.597889
## 3      1  全部位   C00-C96   男女計   1977  189.4090  13.86778  7.400872  7.299964
## 4      1  全部位   C00-C96   男女計   1978  194.5231  13.11493  7.108764  6.130980
## 5      1  全部位   C00-C96   男女計   1979  206.5175  13.38973  6.660657  5.638117
## 6      1  全部位   C00-C96   男女計   1980  214.4543  14.02163  7.107233  6.551611
##   age15_19 age20_24 age25_29 age30_34 age35_39 age40_44 age45_49 age50_54
```

```
## 1 8.882128 13.18414 24.54009 41.92178 71.88043 128.1969 211.0058 300.2575
## 2 8.891426 11.32447 25.69828 38.42129 70.01396 125.0030 205.6629 296.9398
## 3 10.551438 10.42322 26.78686 38.13225 72.73761 125.9141 206.6744 297.4311
## 4 10.116003 10.40606 23.60261 38.56832 75.24708 126.1025 205.0247 308.5769
## 5 9.655429 11.11250 22.93809 41.57780 84.30307 127.7256 211.8893 326.0116
## 6 8.643361 12.09025 20.82652 43.88338 81.21430 127.6162 213.8217 327.8601
## age55_59 age60_64 age65_69 age70_74 age75_79 age80_84 age85plus
## 1 430.9053 639.0219 884.1888 1173.113 1377.386 1360.574 1087.513
## 2 417.5249 618.8536 876.7602 1147.714 1389.089 1407.351 1121.981
## 3 405.1523 625.2349 895.6720 1114.264 1376.906 1431.072 1137.963
## 4 406.5789 621.4057 903.8451 1125.717 1400.840 1400.000 1208.894
## 5 438.9733 656.5278 910.9551 1170.201 1436.000 1442.207 1258.635
## 6 456.2925 656.3355 911.2209 1235.173 1448.579 1500.307 1314.581
```

4 肝がん罹患率のグラフを作成する

```
# 男女計のデータだけを抽出する:
rate.hepatic <- subset(rate.all, 部位 == " 肝臓" & 性別 == " 男女計")
# データの形を変更する wide -> long
library(reshape2)
rate.hepatic.melt <- melt(data= rate.hepatic,
                        measure.vars = names(rate.hepatic)[grep("age",
                                                                names(rate.hepatic))],
                        variable.name = "Age_Range",
                        value.name = "Incidence_Rate")
# アンダーバーをハイフオンに変更
names(rate.hepatic.melt$Age_Range) <- gsub("_", "-",
                                           as.character(rate.hepatic.melt$Age_Range))
# チェックするために、最初の 6 行を示す
head(rate.hepatic.melt)
```

```
##   コード 部位 ICD_10 性別 Dia_yr      粗率 Age_Range Incidence_Rate
## 1     8 肝臓   C22 男女計  1975  9.679323   age0_4      0.6699593
## 2     8 肝臓   C22 男女計  1976 10.232036   age0_4      0.7111653
## 3     8 肝臓   C22 男女計  1977 10.302749   age0_4      0.6767309
## 4     8 肝臓   C22 男女計  1978 11.002483   age0_4      0.5081630
## 5     8 肝臓   C22 男女計  1979 11.981952   age0_4      0.3948111
## 6     8 肝臓   C22 男女計  1980 12.930932   age0_4      0.2818418
```

```
# 診断された年を 10 年ごとにカテゴリ化する
rate.hepatic.melt$Dia_yr10 <- cut(rate.hepatic.melt$Dia_yr, dig.lab=10,
                                right = FALSE,
                                breaks = seq(from = 1975, to = 2015, by = 10))
# 各年齢群カテゴリに下限値の年齢として名前を付ける
```

```

rate.hepatic.melt$age <- seq(from = 0, to = 85,
                             by = 5)[rate.hepatic.melt$Age_Range]

# 生まれた年の計算
rate.hepatic.melt$Birth_yr <- with(rate.hepatic.melt,
                                   Dia_yr - age)

# 生まれた年（出生コホート）を 10 年ごとカテゴリ化する
rate.hepatic.melt$Birth_yr10 <- cut(rate.hepatic.melt$Birth_yr,dig.lab=10,
                                   right = FALSE,
                                   breaks = seq(from = 1890, to = 2020, by = 10))

# 40 歳以上に限定する:
rate.hepatic.melt.40 <- subset(rate.hepatic.melt,
                               (as.numeric(Age_Range) > 8))

# データをチェックするために、最初の 20 行を示す
head(rate.hepatic.melt.40, 20)

```

##	コード	部位	ICD_10	性別	Dia_yr	粗率	Age_Range	Incidence_Rate
## 297	8	肝臓	C22	男女計	1975	9.679323	age40_44	4.924569
## 298	8	肝臓	C22	男女計	1976	10.232036	age40_44	5.930886
## 299	8	肝臓	C22	男女計	1977	10.302749	age40_44	6.303419
## 300	8	肝臓	C22	男女計	1978	11.002483	age40_44	6.097706
## 301	8	肝臓	C22	男女計	1979	11.981952	age40_44	5.703971
## 302	8	肝臓	C22	男女計	1980	12.930932	age40_44	5.145427
## 303	8	肝臓	C22	男女計	1981	14.055343	age40_44	5.926623
## 304	8	肝臓	C22	男女計	1982	15.245212	age40_44	5.765704
## 305	8	肝臓	C22	男女計	1983	16.551309	age40_44	6.106017
## 306	8	肝臓	C22	男女計	1984	18.210172	age40_44	5.381647
## 307	8	肝臓	C22	男女計	1985	19.452466	age40_44	5.199807
## 308	8	肝臓	C22	男女計	1986	20.603754	age40_44	4.986181
## 309	8	肝臓	C22	男女計	1987	22.210135	age40_44	5.149740
## 310	8	肝臓	C22	男女計	1988	22.936400	age40_44	5.812983
## 311	8	肝臓	C22	男女計	1989	23.603099	age40_44	6.205938
## 312	8	肝臓	C22	男女計	1990	26.152168	age40_44	7.759218
## 313	8	肝臓	C22	男女計	1991	27.076901	age40_44	7.302123
## 314	8	肝臓	C22	男女計	1992	28.361939	age40_44	7.240903
## 315	8	肝臓	C22	男女計	1993	28.694175	age40_44	5.587029
## 316	8	肝臓	C22	男女計	1994	28.142745	age40_44	5.603539
##		Dia_yr10	age	Birth_yr	Birth_yr10			
## 297		[1975,1985)	40	1935	[1930,1940)			
## 298		[1975,1985)	40	1936	[1930,1940)			
## 299		[1975,1985)	40	1937	[1930,1940)			
## 300		[1975,1985)	40	1938	[1930,1940)			
## 301		[1975,1985)	40	1939	[1930,1940)			

```

## 302 [1975,1985) 40      1940 [1940,1950)
## 303 [1975,1985) 40      1941 [1940,1950)
## 304 [1975,1985) 40      1942 [1940,1950)
## 305 [1975,1985) 40      1943 [1940,1950)
## 306 [1975,1985) 40      1944 [1940,1950)
## 307 [1985,1995) 40      1945 [1940,1950)
## 308 [1985,1995) 40      1946 [1940,1950)
## 309 [1985,1995) 40      1947 [1940,1950)
## 310 [1985,1995) 40      1948 [1940,1950)
## 311 [1985,1995) 40      1949 [1940,1950)
## 312 [1985,1995) 40      1950 [1950,1960)
## 313 [1985,1995) 40      1951 [1950,1960)
## 314 [1985,1995) 40      1952 [1950,1960)
## 315 [1985,1995) 40      1953 [1950,1960)
## 316 [1985,1995) 40      1954 [1950,1960)

```

```

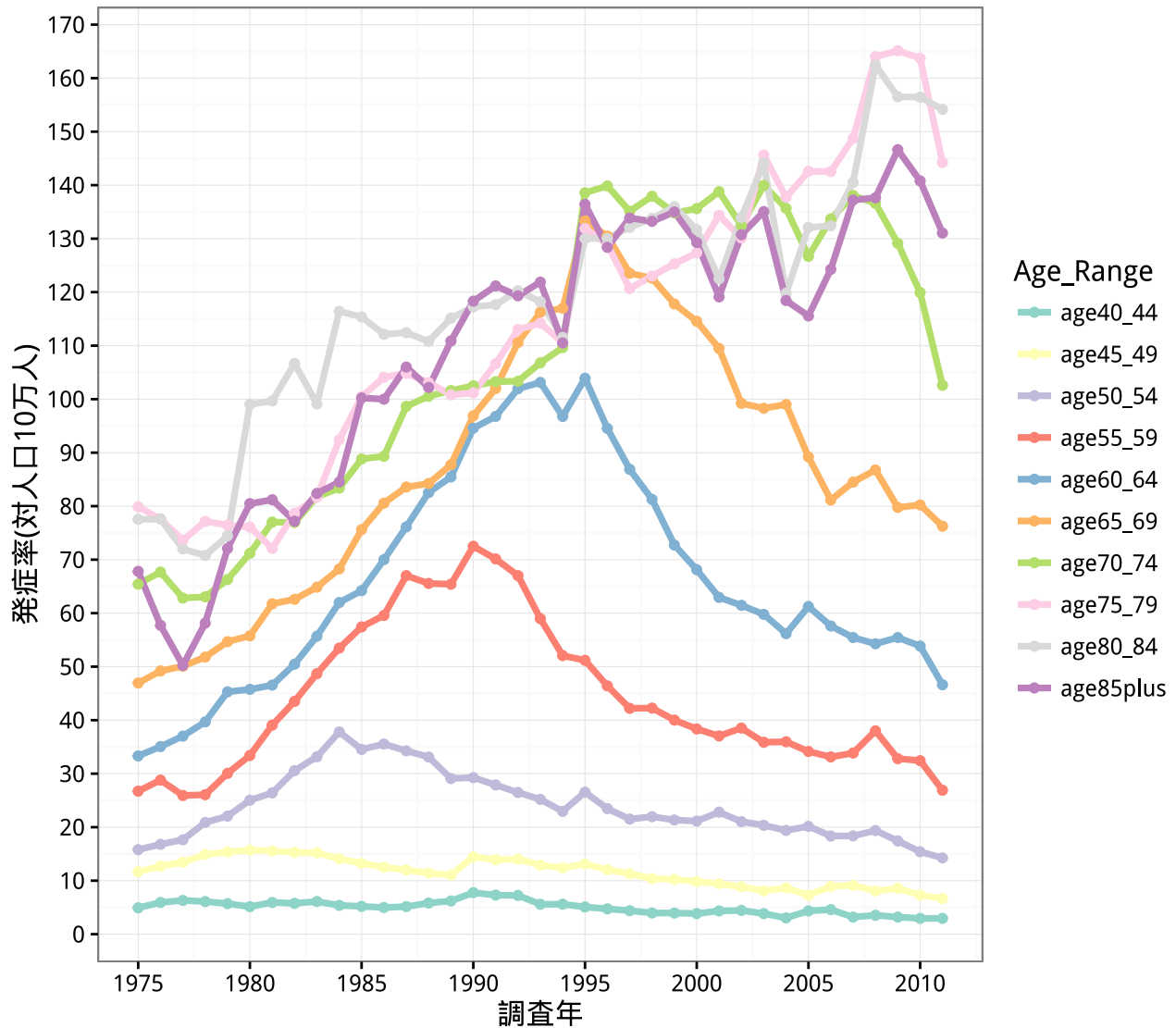
# グラフ 1 作成
library(ggplot2)

# 横軸に調査年, 5歳ごと年齢階級別 (Age_Range) 罹患率

ggplot(data = rate.hepatic.melt.40,
        mapping = aes(x = Dia_yr, y = Incidence_Rate,
                       color = Age_Range)) +
  geom_line(size = 1.3) +
  geom_point() +
  xlab(" 調査年") +
  ylab(" 発症率 (対人口 10 万人)") +
  labs(title =
        " グラフ1: 5 歳ごと年齢階級別罹患率\n(40 歳以上)") +
  theme_bw() +
  theme(legend.key = element_blank(),
        axis.text.x = element_text(angle=0, vjust=1)) +
  scale_y_continuous(breaks = seq(0,170, 10)) +
  scale_x_continuous(breaks = seq(1890, 2020, 5)) +
  scale_color_brewer(palette = "Set3")

```

グラフ 1：5歳ごと年齢階級別罹患率
(40歳以上)



```
# グラフ 2 作成
```

```
# 網掛けデータ
```

```
rect <- data.frame(xmin = 1930, xmax = 1935, ymin=-Inf, ymax=Inf)
```

```
# 横軸に生まれた年, 診断された年齢 (Age_Range) により可視化
```

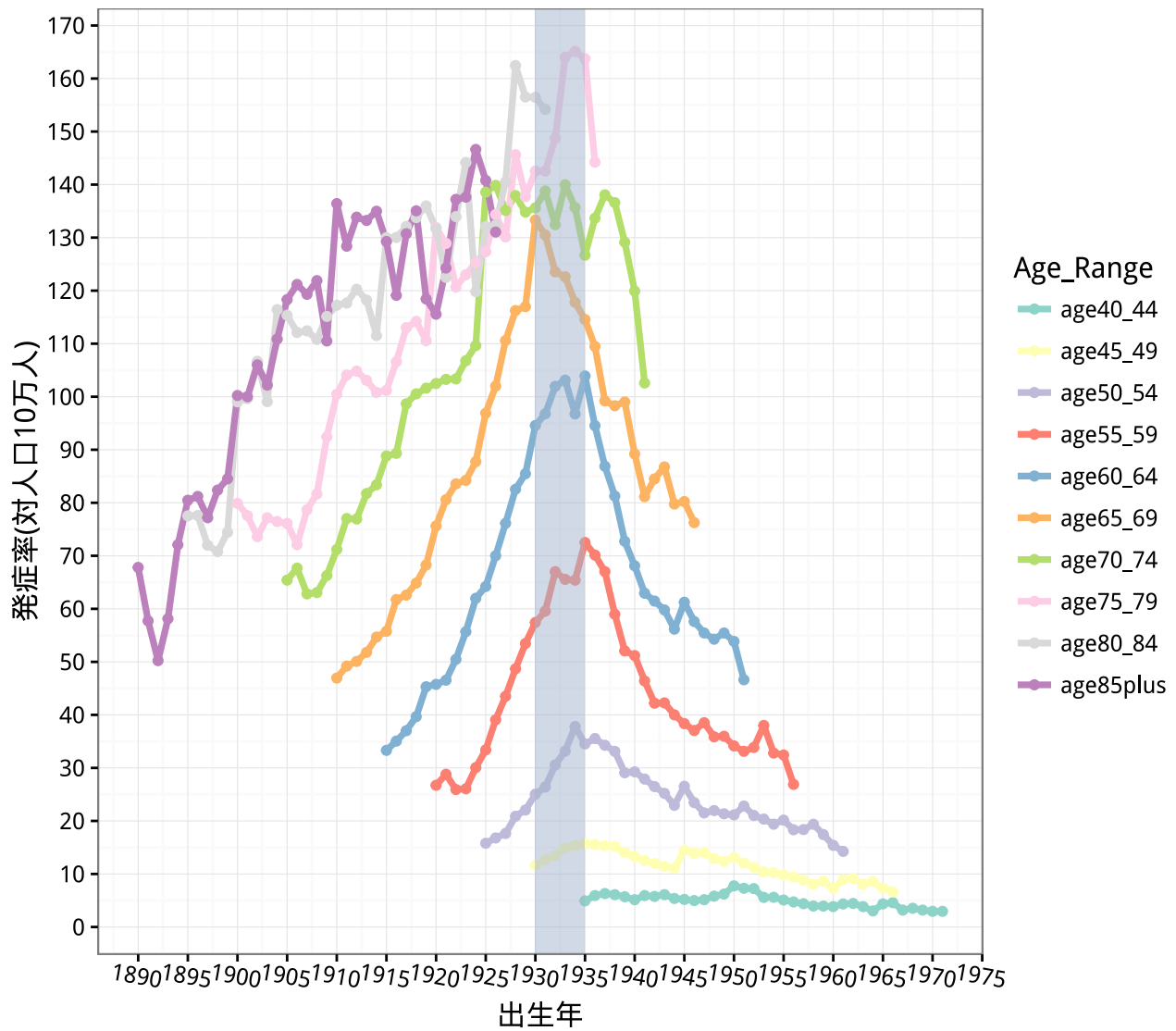
```
ggplot(data = rate.hepatic.melt.40,
       mapping = aes(x = Birth_yr, y = Incidence_Rate,
                    color = Age_Range)) +
  geom_line(size = 1.3) +
  geom_point() +
  xlab(" 出生年") +
```

```

ylab(" 発症率 (対人口 10 万人)") +
labs(title =
      " グラフ2：出生年・年齢別肝がん罹患率\n(40 歳以上)" ) +
theme_bw() +
theme(legend.key = element_blank(),
      axis.text.x = element_text(angle=-10, vjust=1)) +
scale_y_continuous(breaks = seq(0,170, 10)) +
scale_x_continuous(breaks = seq(1890, 2020, 5)) +
scale_color_brewer(palette = "Set3") +
geom_rect(data = rect, aes(xmin=xmin, xmax=xmax, ymin=ymin, ymax=ymax),
          fill = "#A2B5CD", inherit.aes = FALSE, alpha=0.5)

```

グラフ2：出生年・年齢別肝がん罹患率
(40歳以上)



- 4.1 グラフ2: 肝臓がんの罹患率は、1930年代前半生まれにピークがあることがわかる(濃い網掛け)。この年代生まれは、C型肝炎ウイルスの陽性割合の高い世代と一致している。

```
# グラフ 3 作成
# 横軸に診断された時年齢の5歳ごと階級 (Age_Range), 出生コホートにより可視化

ggplot(data = rate.hepatic.melt.40,
        mapping = aes(x = Age_Range, y = Incidence_Rate, group=factor(Birth_yr),
                       color = Birth_yr10)) +
  geom_line(size=1.3) +
  geom_point() +
  xlab(" 診断年齢") +
  ylab(" 発症率 (対人口 10 万人)") +
  labs(title =
        " グラフ3: 日本人の肝がん罹患率\n (出生コホートによる)") +
  theme_bw() +
  theme(legend.key = element_blank(),
        axis.text.x = element_text(angle=10, vjust=1)) +
  scale_y_continuous(breaks = seq(0,170, 10))+
  scale_color_brewer(palette = "Paired")
```

グラフ 3：日本人の肝がん罹患率
(出生コホートによる)

