2015 Japanese Stata Users Group Meeting

タイムテーブル　午前は一橋講堂、午後は中会議場 1、2 の二部屋に分かれての発表となります。

10：30 - 11：00	開場	
11：00 - 11：15	開会の挨拶：一橋大学　北村行伸 様	
11：15 - 12：00	**「Estimating survival-time treatment effects from observational data」** **--- 喫煙習慣は心臓発作の再発間隔にどの程度の影響を与えるのか ?---** Stata Corp, Director of Econometrics, David Drukker　（Q&A 15 分） P. 1	
	ランチタイム	
12：10 - 12：40	**中会議場 1**	**中会議場 2**
	Lunch on seminar **「Stataによるベイズ統計」** 株式会社ライトストーン P. 27	ランチタイム
	ランチタイム	
12：50 - 13：30	**発表1-1**　（Q&A 10分） **「Stataによるデータ・マネージメント」** 慶應義塾大学　産業研究所 准教授 松浦寿幸 様 P. 43	**発表2-1**　（Q&A 10分） **「平成23年国民生活基礎調査－国民健康・栄養調査－ 歯科疾患実態調査のデータリンケージ状況と性・年齢の不一致について」** 国立保健医療科学院 統括研究官 安藤雄一 様 P. 139
13：40 - 14：20	**発表1-2**　（Q&A 10分） **「サンプル脱落がもたらす推計バイアスに関する考察」** 群馬大学　社会情報学部 准教授 坂本和靖 様 P. 61	**発表2-2**　（Q&A 10分） **「途上国における医療保険非利用の分析 --不完全操作変数法の応用」** 京都大学大学院　経済学研究科 准教授 高野久紀 様 P. 141
14：30 - 15：10	**発表1-3**　（Q&A 10分） **「日本の所得分配と経済成長に関する実証研究」** 龍谷大学　経済学部 准教授 大山昌子 様 P. 81	**発表2-3**　（Q&A 10分） **「Stataを用いたデータ管理法」** 自治医科大学　企画経営部医療情報部/ 臨床研究支援センター　データセンター部門 准教授 興梠貴英 様 P. 157
15：30 - 16：10	**発表1-4**　（Q&A 10分） **「信用リスクのマクロストレステストーモデルとインプリメンテーション」** 神奈川大学　経営学部・経営学研究科 准教授 菅野正泰 様 P. 97	**発表2-4**　（Q&A 10分） **「生存分析を用いた予後予測モデル：未破裂脳動脈瘤の３年後の破裂可能性」** 京都大学　医学研究科　社会健康医学系専攻 健康情報学分野 富成伸次郎 様 P. 177
16：20 - 17：00	**発表1-5**　（Q&A 10分） **「カルマンフィルターのファイナンスへの応用」** 早稲田大学大学院　ファイナンス研究科 教授 森平爽一郎 様 P. 113	**発表2-5**　（Q&A 10分） **「九州大学の医学分野及び多施設協同臨床研究における Stata の利用状況」** 九州大学病院メディカル・インフォメーションセンター 講師 徳永章二 様 P. 199
17：10 - 18：30	懇親会（3F レパスト）	

2015 Japanese Stata Group Meeting

諸注意

・午前の部で使用する一橋講堂での飲食は禁止されております。午後の部で使用する中会議場では飲食可能です。昼食のお弁当については、午前の部終了後、中会議場 2 の前で配布いたしますので、中会議場 1 または中会議場 2 の中でお召し上がりください。

・携帯電話はマナーモードに設定にしていただくか、電源をお切りください。

・会場内での録音・録画はご遠慮ください。

・会場は全館禁煙です。喫煙される場合は、2 階のエレベータ奥の喫煙室をご利用ください。

・他の会議場は別の用途で使用されているため、会場以外の会議室への立ち入りは禁止とさせていただきます。（ラウンジは利用可能です。）

・再入場の際には名札を提示。

・貴重品は常時身に着けるようにしてください。

Estimating Survival-time treatment effects from observational data

David M.Drukker
Stata Corp.

平成 27 年 8 月 20 日

概 要

この資料は 2015 年 8 月 28 日金曜日に行われる 2015 Japanese Stata Users Group Meeting における、David M.Drukkes のプレゼンテーション (英語) のポイントをまとめたものです。発表の前にご一読いただき、問題意識を整理しておくことで、ユーザ会への参加を有意義なものになさってください。

問題意識: 過去に心臓発作の経験がある人にとって、喫煙習慣があると、次の心臓発作までの時間間隔 t_i はどのくらい短くなるのか?

データの概略

`age`:被験者の年齢 (平均からの偏差)
`exercise`:?(連続変数)
`diet`:?(連続変数)
`smoke`:喫煙習慣の有無
`fail`:failure event は心臓発作の発生で 1、打ち切りの場合は 0.
`atime`:1 回目から 2 回目までの時間間隔 t
データ数:5000(t が分かっているデータ数:2,969, センサードデータが 2,031)
*事前資料には変数 `exercise` と `diet` の説明がありませんでした。

比例ハザードモデルの推定

過去の知識で分析するとしたら、次のような比例ハザードモデルを推定します。

```
.stcox smoke age excercise diet
```

$$h(t) = h_0(t) \exp(\beta_1 smoke + \beta_2 age + \beta_3 exercise + \beta_4 diet) \tag{1}$$

`smoke` のハザード比は 1.54 で z 値は 8.70 で有意となる。
ところで、

- ハザード比の単位は?

- 変数間に交互作用があるとしたら、ハザード比の解釈はどうなる?
- smoke の有無による τ の差の平均 ATE が分かれば、直感的にハザード比よりも理解しやすいのではないか?

$$\mathrm{ATE} = \mathrm{E}\left[t_i\left(smoke\right) - t_i\left(nonsmoke\right)\right] \tag{2}$$

アウトカムが単純な連続変数ならば、stata13 から用意された teffects psmatch が利用できる。しかし、上記のデータセットはセンサードデータが存在する (サバイバルタイムデータ)。これを考慮して、ATE を推定しなければならない。

回帰調整 (Regression Adjutsment) による処置効果の推定

サバイバルデータ用の傾向スコア分析コマンドとして stteffects コマンドを用意した。実際のコマンドは次のようなものである。

```
.stteffects ra (age excercise diet) (smoke)
```

これを実行すると、ATE $= -1.520$(年) となる。喫煙習慣があると再発までの期間が 1.5 年、短くなる。同時に、POmean として Nonsmoker の値が 4.057 年と推定される。つまり、喫煙習慣がなければ、再発までの期間の平均は 4.05 年と推定される。

- オッズ比と比べたとき、どちらが直感的に分かりやすいか?

IPW(Inverse-probability estimator) の場合

```
.stteffects ipw (smoke age excercise diet) (age exercise diet)
```

これは回帰調整とほぼ同じような推定値を得る。

Quantile Treatment effects(QTE)

報告者 David Drukker 作成の ado ファイル mqgamma[1]による、分位点における処置効果の推定例。

問題意識 第 1 回目の心臓発作の後に、何らかの運動 (excercise) を習慣づけた人を想定する。比較的心臓が丈夫な人は、この運動により健康になり、逆に、心臓に深刻な問題がある場合は、運動による改善は見込めないとする。このとき、再び心臓発作が起こるという事象 y の確率分布を考える。当然、喫煙習慣の有無によって、その確率分布の分布曲線 (横軸は t_i, 縦軸は再発の確率) は異なる。ここでは、2 つの分布曲線の違いから、分位点 (Quantile) ごとの ATE を求める。

[1] このコマンドは標準のコマンドではありませんので、ado ファイルとしてインストールする必要があります。

MEMO

Potential Outcome: 例えば、実際に処置を受けた人のアウトカムに対して、その人が処置を受けていない、としたときの潜在的なアウトカム。

Ratio of unconditional hazards: 処置効果の無条件ハザード比とは、

$$\frac{h_t\left(Smoking\ Potential\ Outcome\right)}{h_t\left(Nonsmoking\ Potential\ Outcome\right)}$$

以上
LightStone Corp.

Estimating survival-time treatment effects from observational data

David M. Drukker

Director of Econometrics
Stata

Japanese Stata Users Group meeting
28 August 2015

A question

- Is smoking bad for men who have already had a heart attack?
  - Too vague
- Will smoking reduce the time to a second heart-attack among men aged 45–55 who have already had a heart attack?
  - Less interesting, but more specific
  - There might even be data to help us answer this question
  - The data will be observational, not experimental
  - This question is about the time to an event, and such data are commonly known as survival-time data or time-to-event data. These data are nonnegative and, frequently, right-censored

The data

```
. use sheart2
(Time to second heart attack (fictional))
. describe
Contains data from sheart2.dta
  obs:         5,000                          Time to second heart attack
                                                (fictional)
 vars:             6                          11 Aug 2015 15:28
 size:       120,000
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
age             float   %9.0g                 Age (in decades, demeaned)
exercise        float   %9.0g                 Exercise index
diet            float   %9.0g                 Diet index
smoke           float   %9.0g      lsmoke     Smoking indicator
fail            float   %9.0g      lfail      Failure indicator
atime           float   %9.0g                 Time to second attack
-------------------------------------------------------------------------------
Sorted by:
```


The data

```
. stset atime, failure(fail)
     failure event:  fail != 0 & fail < .
obs. time interval:  (0, atime]
 exit on or before:  failure
-------------------------------------------------------------------------------
       5000  total observations
          0  exclusions
-------------------------------------------------------------------------------
       5000  observations remaining, representing
       2969  failures in single-record/single-failure data
  10972.843  total analysis time at risk and under observation
                                                at risk from t =         0
                                     earliest observed entry t =         0
                                          last observed exit t =  40.96622
. save sheart2, replace
file sheart2.dta saved
```

- 2,969 of the 5,000 observations record actual time to a second heart attack; remainder were censored

A Cox model for the treatment

- Many researchers would start by fitting a Cox model

```
. stcox smoke age exercise diet , nolog noshow
Cox regression -- no ties
No. of subjects =        5,000                 Number of obs    =        5,000
No. of failures =        2,969
Time at risk    =  10972.84266
                                               LR chi2(4)       =       271.77
Log likelihood  =   -21963.163                 Prob > chi2      =       0.0000
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
smoke	1.540071	.0764791	8.70	0.000	1.397239	1.697505
age	2.024237	.1946491	7.33	0.000	1.676527	2.444062
exercise	.5473001	.0454893	-7.25	0.000	.465026	.6441304
diet	.4590354	.0379597	-9.42	0.000	.3903521	.5398037

- Smoking increases the hazard of a second heart attack by a factor of 1.5

A Cox model for the treatment

- The Cox model models the probability that the event will occur in the next moment given that it has not yet happened as a function of covariates
  - The probability that the event will occur in the next moment given that it has not yet happened and given covariates is known as the conditional hazard function denoted by $\lambda(t|\mathbf{x})$
  - The Cox model specifies that

$$\lambda(t|\mathbf{x}) = \lambda_0(t)\exp(\mathbf{x}\boldsymbol{\beta})$$

    and only estimates $\boldsymbol{\beta}$
- Leaving $\lambda_0(t)$ unspecified increases the flexibility of the model

A Cox model for the treatment

- Does the binary treatment `smoke` affect the time to second heart attack?
- The hazard ratio reported by `stcox` indicates that smoking raises the hazard of a second heart attack by a factor of 1.5 relative to not smoking

$$\frac{\lambda(t|\mathbf{x}, \texttt{smoke} = 1)}{\lambda(t|\mathbf{x}, \texttt{smoke} = 0)} = \frac{\lambda_0(t)\exp(\beta_{smoke} + \mathbf{x}_o\boldsymbol{\beta}_o)}{\lambda_0(t)\exp(\mathbf{x}_o\boldsymbol{\beta}_o)} = \exp(\beta_{smoke})$$

where $\mathbf{x}_0\boldsymbol{\beta}_o = \texttt{age}\beta_{age} + \texttt{exercise}\beta_{exercise} + \texttt{diet}\beta_{diet}$

The effect varies

```
. stcox ibn.smoke#c.(age exercise diet) , nolog noshow
Cox regression -- no ties
No. of subjects =        5,000                 Number of obs    =        5,000
No. of failures =        2,969
Time at risk    =  10972.84266
                                               LR chi2(6)       =       223.11
Log likelihood  =   -21987.493                 Prob > chi2      =       0.0000
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
smoke#c.age						
Nonsmoker	1.714749	.1751413	5.28	0.000	1.403655	2.094791
Smoker	3.979649	1.110035	4.95	0.000	2.303673	6.874936
smoke#c.exercise						
Nonsmoker	.5514891	.0476827	-6.88	0.000	.4655224	.6533309
Smoker	.2839313	.0822003	-4.35	0.000	.1609844	.5007752
smoke#c.diet						
Nonsmoker	.4461597	.0389598	-9.24	0.000	.3759769	.5294433
Smoker	.6908017	.1785842	-1.43	0.152	.416201	1.146578

- The ratio of the smoking hazard to the nonsmoking hazard varies by age, exercise, and diet

Problems with the Cox model

- Two problems with the Cox model
  1. It is hard to understand the units of the hazard ratio
     - How bad is it that smoking raises the hazard ratio by 1.5?
  2. This interpretation is only useful if the treatment enters the $\mathbf{x}\beta$ term linearly
     - If the treatment is interacted with other covariates, the effect of the treatment varies over individuals
- The average difference in time to second heart attack when everyone smokes instead of when no one smokes
  1. is easier to interpret
  2. is easier to estimate

Doctors versus policy analysts

- What can we do when the estimated effects vary over covariate values?
- When an effect varies over the values of other covariates, you can estimate the effect for a particular type of person or estimate a population-level effect
  - Doctors use covariate specific estimates
    (They ask you many questions to learn your covariates.)
  - Policy analysts need to account for the how a policy will effect different people in the population
    The discipline of the population distribution of the effects keeps them from picking winners or losers

Effects that vary over individuals

- For each individual, the effect of the treatment is a contrast of what would happen if the individual received the treatment versus what would happen if the individual did not receive the treatment
  - A potential outcome is the outcome an individual would receive if given a specific treatment level
  - For each treatment level, there is a potential outcome for each individual

```
. use sheart2_po
(Potential outcome time to second heart attack)
. list id atime_ns atime_s smoke atime in 21/25
```

	id	atime_ns	atime_s	smoke	atime
21.	21	1.44135	.7616374	Nonsmoker	1.44135
22.	22	1.422631	1.422631	Smoker	1.422631
23.	23	4.264108	.3285356	Nonsmoker	4.264108
24.	24	1.533371	1.246619	Nonsmoker	1.533371
25.	25	.1929609	.1929609	Nonsmoker	.1929609

Ratio of unconditional hazards

- The hazard-ratio measure of the treatment effect is the ratio of the hazard of the smoking potential outcome to the hazard nonsmoking potential outcome
  - The hazard-ratio measure of the treatment effect is the ratio of the hazard from the distribution when everyone smokes to the hazard from the distribution when no one smokes
  - This ratio hazards of unconditional distributions is not the same as an average of conditional hazard ratios (See Appendix 1)

Average treatment effect

- Ratios of unconditional hazards are harder to estimate and more difficult to interpret than the average difference in time to second heart attack when everyone smokes instead of no one smokes
  - The average difference in time to second heart attack when everyone smokes instead of no one smokes is an average treatment effect (ATE)
  - $ATE = \mathbf{E}[t_i(\texttt{smoke}) - t_i(\texttt{notsmoke})]$
    $t_i(\texttt{smoke})$ is the time to event when person i smokes
    and
    $t_i(\texttt{notsmoke})]$ is the time to event when person i does not smoke
- The ATE provides a measure of the effect in the units of time in which the time to event is measured
  - In our example, the ATE is measured in years

Average treatment effect

- Recall that one of the two potential outcomes is always missing

```
. use sheart2_po
(Potential outcome time to second heart attack)
. list id atime_ns atime_s smoke atime in 21/25
```

	id	atime_ns	atime_s	smoke	atime
21.	21	1.44135	.7616374	Nonsmoker	1.44135
22.	22	1.422631	1.422631	Smoker	1.422631
23.	23	4.264108	.3285356	Nonsmoker	4.264108
24.	24	1.533371	1.246619	Nonsmoker	1.533371
25.	25	.1929609	.1929609	Nonsmoker	.1929609

- Potential outcomes are the data that we wish we had to estimate causal treatment effects
- Estimating treatment effects can be viewed as a missing-data problem

Average treatment effect

- If we had data on each potential outcome, the average difference in the (observed) potential outcomes would estimate the population average treatment effect
- The average of a potential outcome in the population is known as the potential-outcome mean (POM) for a treatment level
  - The ATE is a difference in POMs

$$\begin{aligned} ATE &= POM_{smoke} - POM_{nonsmoke} \\ &= \mathbf{E}[t_i(\texttt{smoke})] - \mathbf{E}[t_i(\texttt{notsmoke})] \end{aligned}$$

$t_i(\texttt{smoke})$ is the time to event when person i smokes
and
$t_i(\texttt{notsmoke})]$ is the time to event when person i does not smoke

Missing data

- The "fundamental problem of causal inference" (Holland (1986)) is that we only observe one of the potential outcomes
- We can use the tricks of missing-data analysis to estimate treatment effects
- For more about potential outcomes Rubin (1974), Holland (1986), Heckman (1997), Imbens (2004), (Cameron and Trivedi, 2005, chapter 2.7), Imbens and Wooldridge (2009), and (Wooldridge, 2010, chapter 21)

Random-assignment case

- If smoking were randomly assigned, the missing potential outcome would be missing completely at random
  - If the time to second heart attack was never censored and smoking was randomly assigned
    1. The average time to second heart attack among smokers would estimate the smoking POM
    2. The average time to second heart attack among nonsmokers would estimate the nonsmoking POM
    3. The difference in these estimated POMs would estimate the ATE

As good as random

- Instead of assuming that the treatment is randomly assigned, we assume that the treatment is as good as randomly assigned after conditioning on covariates
- Formally, this assumption is known as conditional independence
- Even more formally, we only need conditional mean independence (CMI) which says that after conditioning on covariates, the treatment does not affect the means of the potential outcomes

Choice of auxiliary model

- Recall that the potential-outcomes framework formulates the estimation of the ATE as a missing-data problem
- We use the parameters of an auxiliary model to solve the missing-data problem
  - The auxiliary model is how we condition on covariates so that the treatment is as good as randomly assigned
  - The auxiliary model also handles the data lost to censoring

Model		Estimator
outcome	$\longrightarrow$	Regression adjustment (RA)
treatment	$\longrightarrow$	Inverse-probability weighted (IPW)
outcome and treatment	$\longrightarrow$	IPW RA (IPWRA)

Regression adjustment estimators

- Regression adjustment (RA) estimators use predicted values from the model for the time to event to solve the missing-data problems
- RA estimators estimate the parameters of separate survival models for the outcome for each treatment level, then
  - The mean of the predicted times to second heart attack using the estimated coefficients from the model for smokers and all the observations estimates the smoking POM
  - The mean of the predicted times to second heart attack using the estimated coefficients from the model for nonsmokers and all the observations estimates the nonsmoking POM
  - The difference between the estimated smoking POM and the estimated nonsmoking POM estimates the ATE
  - Censoring is handled in the log likelihood functions of the survival models

```
. use sheart2
(Time to second heart attack (fictional))
. stteffects ra (age exercise diet) (smoke), nolog noshow
Survival treatment-effects estimation           Number of obs     =      5,000
Estimator      : regression adjustment
Outcome model  : Weibull
Treatment model: none
Censoring model: none
```

_t	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
ATE						
smoke (Smoker vs Nonsmoker)	-1.520671	.2011014	-7.56	0.000	-1.914822	-1.126519
POmean						
smoke Nonsmoker	4.057439	.1028462	39.45	0.000	3.855864	4.259014

- The average time to second heart attack is 1.5 years sooner when everyone in the population smokes instead of no one smokes
- The average time to second heart attack is 4.1 years when no one smokes


```
. stteffects ra (age exercise diet, gamma) (smoke), nolog noshow
Survival treatment-effects estimation           Number of obs     =      5,000
Estimator      : regression adjustment
Outcome model  : gamma
Treatment model: none
Censoring model: none
```

_t	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
ATE						
smoke (Smoker vs Nonsmoker)	-1.616514	.177703	-9.10	0.000	-1.964805	-1.268222
POmean						
smoke Nonsmoker	4.014823	.0988662	40.61	0.000	3.821049	4.208598

- Can model the outcome using either a gamma, exponential, or log normal distribution instead of the default Weibull distribution
- Can model the ancillary distribution parameters using `ancillary()` option

Inverse-probability-weighted estimators

- Inverse-probability-weighted (IPW) estimators:
  - IPW estimators weight observations on the observed outcome variable by the inverse of the probability that it is observed to account for the missingness process
  - Observations that are not likely to contain missing data get a weight close to one; observations that are likely to contain missing data get a weight larger than one, potentially much larger

Inverse-probability-weighted estimators

- IPW estimators use estimates from models for the probability of treatment and the probability of censoring to correct for the missing potential outcome and the observations lost to censoring
- In contrast, RA estimators model the outcome without any assumptions about the functional form for the probability of treatment model
  - RA estimators handle censoring in the log likelihood function
  - Handling censoring in the log likelihood function allows for fixed censoring times
- IPW estimators have a long history in statistics, biostatistics, and econometrics
  - Horvitz and Thompson (1952) Robins and Rotnitzky (1995), Robins et al. (1994), Robins et al. (1995), Imbens (2000), Wooldridge (2002), Hirano et al. (2003), (Tsiatis, 2006, chapter 6), Wooldridge (2007) and (Wooldridge, 2010, chapters 19 and 21)

```
. stteffects ipw (smoke age exercise diet) (age exercise diet), nolog noshow
Survival treatment-effects estimation           Number of obs     =      5,000
Estimator      : inverse-probability weights
Outcome model  : weighted mean
Treatment model: logit
Censoring model: Weibull
```

_t	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
ATE						
smoke (Smoker vs Nonsmoker)	-1.689397	.3373219	-5.01	0.000	-2.350536	-1.028258
POmean						
smoke Nonsmoker	4.200135	.2156737	19.47	0.000	3.777423	4.622848

- The average time to second heart attack is 1.7 years sooner when everyone in the population smokes instead of no one smokes
- The average time to second heart attack is 4.2 years when no one smokes


```
. stteffects ipw (smoke age exercise diet, logit)            ///
>                (age exercise diet, gamma), nolog noshow
Survival treatment-effects estimation           Number of obs     =      5,000
Estimator      : inverse-probability weights
Outcome model  : weighted mean
Treatment model: logit
Censoring model: gamma
```

_t	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
ATE						
smoke (Smoker vs Nonsmoker)	-1.922143	.4502077	-4.27	0.000	-2.804534	-1.039752
POmean						
smoke Nonsmoker	4.555551	.3345953	13.62	0.000	3.899756	5.211345

- Can model treatment by probit, logit, or heteroskedastic probit
- Can model censoring by Weibull, gamma, or log normal
  Can model ancillary parameters

Combining IPW and RA

- Inverse-probability-weighted regression-adjustment (IPWRA) estimators combine models for the outcome and the treatment to get more efficient estimates
- IPWRA estimators use the inverse of the estimated treatment-probability weights to estimate missing-data-corrected regression coefficients that are subsequently used to estimate the POMs
  - The ATE is estimated by a difference in the estimated POMs
- Censoring can be handled in the log likelihood function or by modeling the censoring process
  - Handling censoring in the log likelihood function allows for fixed censoring times
- See Wooldridge (2007) and (Wooldridge, 2010, section 21.3.4)


```
. stteffects ipwra (age exercise diet) (smoke age exercise diet) , nolog noshow
Survival treatment-effects estimation           Number of obs     =      5,000
Estimator      : IPW regression adjustment
Outcome model  : Weibull
Treatment model: logit
Censoring model: none
```

_t	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
ATE						
smoke (Smoker vs Nonsmoker)	-1.543315	.2027738	-7.61	0.000	-1.940744	-1.145885
POmean						
smoke Nonsmoker	4.064291	.1032385	39.37	0.000	3.861947	4.266634

- The average time to second heart attack is 1.5 years sooner when everyone in the population smokes instead of no one smokes
- The average time to second heart attack is 4.1 years when no one smokes

```
. steffects ipwra (age exercise diet)                    ///
>                 (smoke age exercise diet)              ///
>                 (age exercise diet)       , nolog noshow
Survival treatment-effects estimation          Number of obs     =      5,000
Estimator      : IPW regression adjustment
Outcome model  : Weibull
Treatment model: logit
Censoring model: Weibull
```

_t	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
ATE						
smoke (Smoker vs Nonsmoker)	-1.782505	.3091845	-5.77	0.000	-2.388495	-1.176514
POmean						
smoke Nonsmoker	4.233607	.2185565	19.37	0.000	3.805244	4.661969

- This example models the censoring process instead handling it in the log likelihood function for the outcome
- Additional model choices as for RA and IPW estimators

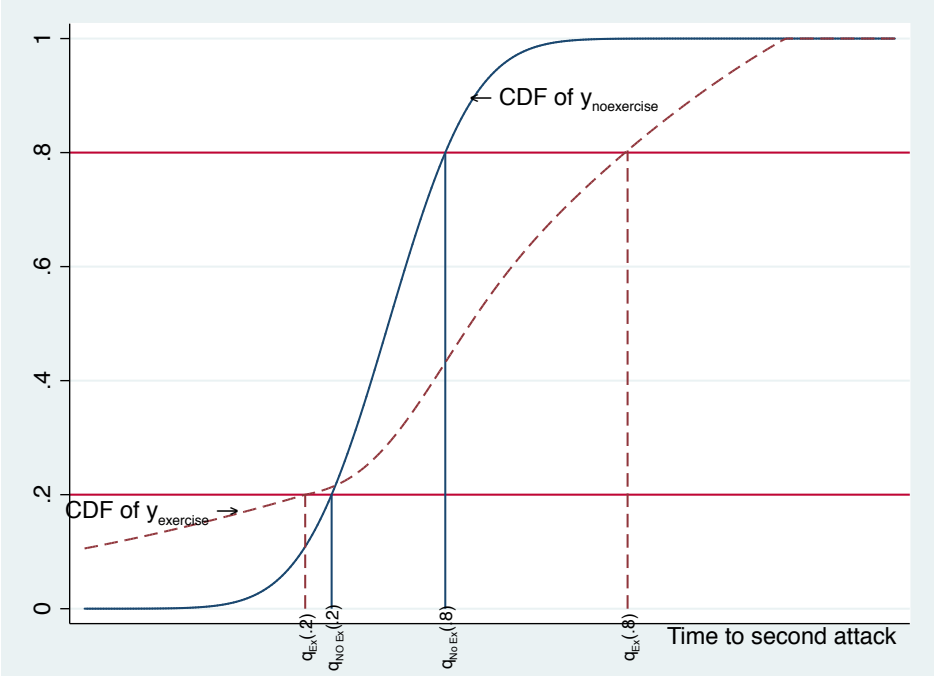
QTEs for survival data

- Imagine a study that followed middle-aged men for two years after suffering a heart attack
  - Does exercise affect the time to a second heart attack?
  - Some observations on the time to second heart attack are censored
  - Observational data implies that treatment allocation depends on covariates
  - We use a model for the outcome to adjust for this dependence

QTEs for survival data

- Exercise could help individuals with relatively strong hearts but not help those with weak hearts
- For each treatment level, a strong-heart individual is in the .75 quantile of the marginal, over the covariates, distribution of time to second heart attack
  - QTE(.75) is difference in .75 marginal quantiles
- Weak-heart individual would be in the .25 quantile of the marginal distribution for each treatment level
  - QTE(.25) is difference in .25 marginal quantiles
- our story indicates that the QTE(.75) should be significantly larger that the QTE(.25)

What are QTEs?



Quantile Treatment effects

- We can easily estimate the marginal quantiles, but estimating the quantile of the differences is harder
- We need a rank preservation assumption to ensure that quantile of the differences is the difference in the quantiles
  - The τ(th) quantile of y_1 minus the τ(th) quantile of y_0 is not the same as the τ(th) quantile of $(y_1 - y_0)$ unless we impose a rank-preservation assumption
  - Rank preservation means that the random shocks that affect the treated and the not-treated potential outcomes do not change the rank of the individuals in the population

    The rank of an individual in y_1 is the same as the rank of that individual in y_0
  - Graphically, the horizontal lines must intersect the CDFs "at the same individual"

A regression-adjustment estimator for QTEs

- Estimate the $\boldsymbol{\theta}_1$ parameters of $F(y|\mathbf{x}, t = 1, \boldsymbol{\theta}_1)$ the CDF conditional on covariates and conditional on treatment level
  - Conditional independence implies that this conditional on treatment level CDF estimates the CDF of the treated potential outcome
- Similarly, estimate the $\boldsymbol{\theta}_0$ parameters of $F(y|\mathbf{x}, t = 0, \boldsymbol{\theta}_0)$
- At the point y,

$$1/N \sum_{i=1}^{N} F(y|\mathbf{x}_i, \widehat{\boldsymbol{\theta}}_1)$$

  estimates the marginal distribution of the treated potential outcome
- The $\widehat{q}_{1,.75}$ that solves

$$1/N \sum_{i=1}^{N} F(\widehat{q}_{1,.75}|\mathbf{x}_i, \widehat{\boldsymbol{\theta}}_1) = .75$$

A regression-adjustment estimator for QTEs

- The $\widehat{q}_{0,.75}$ that solves

$$1/N \sum_{i=1}^{N} F(\widehat{q}_{0,.75}|\mathbf{x}_i, \widehat{\boldsymbol{\theta}}_0) = .75$$

  estimates the .75 marginal quantile for the control potential outcome
- $\widehat{q}_1(.75) - \widehat{q}_0(.75)$ consistently estimates QTE(.75)
- See Drukker (2014) for details

mqgamma example

- mqgamma is a user-written command documented in Drukker (2014)
- `. ssc install mqgamma`

```
. use exercise, clear
. mqgamma t active, treat(exercise) fail(fail) lns(health) quantile(.25 .75)
Iteration 0:   EE criterion =   .7032254
Iteration 1:   EE criterion =  .05262105
Iteration 2:   EE criterion =  .00028553
Iteration 3:   EE criterion =  6.892e-07
Iteration 4:   EE criterion =  4.706e-12
Iteration 5:   EE criterion =  1.604e-22
Gamma marginal quantile estimation              Number of obs      =      2000
```

t	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
q25_0 _cons	.2151604	.0159611	13.48	0.000	.1838771	.2464436
q25_1 _cons	.2612655	.0249856	10.46	0.000	.2122946	.3102364
q75_0 _cons	1.591147	.0725607	21.93	0.000	1.44893	1.733363
q75_1 _cons	2.510068	.1349917	18.59	0.000	2.245489	2.774647

mqgamma example

```
. nlcom (_b[q25_1:_cons] - _b[q25_0:_cons])       ///
>         (_b[q75_1:_cons] - _b[q75_0:_cons])
       _nl_1:  _b[q25_1:_cons] - _b[q25_0:_cons]
       _nl_2:  _b[q75_1:_cons] - _b[q75_0:_cons]
```

t	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_nl_1	.0461051	.0295846	1.56	0.119	-.0118796	.1040899
_nl_2	.9189214	.1529012	6.01	0.000	.6192405	1.218602

Appendix 1: Ratio of unconditional hazards

- The ratio hazards of unconditional (marginal) distributions is not the same as an average of conditional hazard ratio

$$\frac{\lambda_{smoke}(t)}{\lambda_{nonsmoke}(t)} = \frac{\frac{f_{smoke}(t)}{S_{smoke}(t)}}{\frac{f_{nonsmoke}(t)}{S_{nonsmoke}(t)}} \neq \mathbf{E}\left[\frac{\lambda_{smoke}(t|\mathbf{x}\beta_{smoke})}{\lambda_{nonsmoke}(t|\mathbf{x}\beta_{nonsmoke})}\right]$$

$\lambda_{smoke}(t)$	is the unconditional hazard when everyone smokes
$\lambda_{nonsmoke}(t)$	is the unconditional hazard when no one smokes
$f_{smoke}(t)$	is the unconditional density when everyone one smokes
$f_{nonsmoke}(t)$	is the unconditional density when no one smokes
$S_{smoke}(t)$	is the unconditional survival function when everyone smokes
$S_{nonsmoke}(t)$	is the unconditional survival function when no one smokes

Appendix 2: Why robust standard errors?

- Have a multistep estimator
  1. Example based on RA, same logic works for IPW and IPWRA
  2. Model outcome conditional on covariates for treated observations
  3. Model outcome conditional on covariates for not treated observations
  4. Estimate predicted mean survival time of all observations given covariates from treated model estimates
  5. Estimate predicted mean survival time of all observations given covariates from not-treated model estimates
  6. Difference in means of predicted means estimates ATE

Appendix 2: Why robust standard errors?

- Each step can be obtained by solving moment conditions yielding a method of moments estimator known as an estimating equation (EE) estimator
  - $\mathbf{m}_i(\boldsymbol{\theta})$ is vector of moment equations and $\mathbf{m}(\boldsymbol{\theta}) = 1/N \sum_{i=1}^{N} \mathbf{m}_i(\boldsymbol{\theta})$
- The estimator for the variance-covariance matrix of the estimator has the form $1/N(DMD')$ where $D = \left(\frac{1}{N}\frac{\partial m(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^{-1}$ and $M = \frac{1}{N}\sum_{i=1}^{N} \mathbf{m}_i(\boldsymbol{\theta})\mathbf{m}_i(\boldsymbol{\theta})$
- Stacked moments do not yield a symmetric D, so no simplification under correct specification

Cameron, A. Colin and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and applications*, Cambridge: Cambridge University Press.

Drukker, David M. 2014. "Quantile treatment effect estimation from censored data by regression adjustment," Tech. rep., Under review at the *Stata Journal*, http://www.stata.com/ddrukker/mqgamma.pdf.

Heckman, James J. 1997. "Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations," *Journal of Human Resources*, 32(3), 441–462.

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder. 2003. "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica*, 71(4), 1161–1189.

Holland, Paul W. 1986. "Statistics and causal inference," *Journal of the American Statistical Association*, 945–960.

Horvitz, D. G. and D. J. Thompson. 1952. "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, 47(260), 663–685.

Imbens, Guido W. 2000. "The role of the propensity score in estimating dose-response functions," *Biometrika*, 87(3), 706–710.

———. 2004. "Nonparametric estimation of average treatment effects under exogeneity: A review," *Review of Economics and statistics*, 86(1), 4–29.

Imbens, Guido W. and Jeffrey M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5–86.

Robins, James M. and Andrea Rotnitzky. 1995. "Semiparametric Efficiency in Multivariate Regression Models with Missing Data," *Journal of the American Statistical Association*, 90(429), 122–129.

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. 1994. "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association*, 89(427), 846–866.

———. 1995. "Analysis of Semiparametric Regression Models for

Repeated Outcomes in the Presence of Missing Data," *Journal of the American Statistical Association*, 90(429), 106–121.

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology*, 66(5), 688.

Tsiatis, Anastasios A. 2006. *Semiparametric theory and missing data*, New York: Springer Verlag.

Wooldridge, Jeffrey M. 2002. "Inverse probability weighted M-estimators for sample selection, attrition, and stratification," *Portuguese Economic Journal*, 1, 117–139.

———. 2007. "Inverse probability weighted estimation for general missing data problems," *Journal of Econometrics*, 141(2), 1281–1301.

———. 2010. *Econometric Analysis of Cross Section and Panel Data*, Cambridge, Massachusetts: MIT Press, second ed.

Stataによるベイズ統計

LightStone Corp

ベイズの定理

目的:モデルパラメータθの分布(期待値と分散)を知ること

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)} \quad (1)$$

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{f(y;\theta)\pi(\theta)}{m(y)} \quad (2)$$

事後分布

事前分布

$$m(y) = \int f(y;\theta)\pi(\theta)d\theta \quad (3)$$

定数(周辺尤度)

事前分布

- Stata14で利用可能な事前分布

  連続一変量
  normal,lognormal,uniform,gamma,igamma,exponential,beta,chi2,jeffreys

  連続多変量
  mvnormal,mvnormal0,zellnereg, zellnereg0, wishart,iwishart, jeffreys

  離散
  bernoulli,index, poisson

  Generic
  flat, density,logdensity

OLS

- 酸素摂取量に対する運動効果、Kuehl(2000,551)

  Stata14のPDFマニュアルから

$$change = \beta_0 + \beta_{group} group + \beta_{age} age + \epsilon$$

$$\boxed{\mathrm{E}\left(\hat{\beta}\right) = \beta}$$

OLS

$$change = \beta_0 + \beta_{group} group + \beta_{age} age + \epsilon$$

Source	SS	df	MS	Number of obs	=	12
				F(2, 9)	=	41.42
Model	647.874893	2	323.937446	Prob > F	=	0.0000
Residual	70.388768	9	7.82097423	R-squared	=	0.9020
				Adj R-squared	=	0.8802
Total	718.263661	11	65.2966964	Root MSE	=	2.7966

change	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
group	5.442621	1.796453	3.03	0.014	1.378763	9.506479
age	1.885892	.295335	6.39	0.000	1.217798	2.553986
_cons	-46.4565	6.936531	-6.70	0.000	-62.14803	-30.76498

推定量は「正規分布に従う」と考える

母数の分布

$$change = \beta_0 + \beta_{group} group + \beta_{age} age + \epsilon$$

✓ 3つの係数とデータ(change)の誤差分散をパラメータとする

$$change \sim N(X\beta, \sigma^2)$$
$$(\beta, \sigma^2) \sim \frac{1}{\sigma^2}$$

✓ noninformative Jeffreys prior

線形回帰のベイズ推定1

MCMC Standard Error*　　95%信用区間

	Mean	Std. Dev.	MCSE	Median	Equal-tailed [95% Cred. Interval]	
change						
group	5.429677	2.007889	.083928	5.533821	1.157584	9.249262
age	1.8873	.3514983	.019534	1.887856	1.184714	2.567883
_cons	-46.49866	8.32077	.450432	-46.8483	-62.48236	-30.22105
var	10.27946	5.541467	.338079	9.023905	3.980325	25.43771

ベイズ推定のコマンド1

```
.bayesmh  change group age,likelihood(normal({var}))   ///
        prior({change:},flat) prior({var},jeffreys)
```

3つの回帰係数の事前分布の定義　　事前分布の分散の定義

likelihood( )オプション：尤度、または、アウトカムの分布(normal)を入力する
prior( )オプション:事前分布
flat:密度1の分布
約束事：パラメータは大カッコ{ }で囲む
　　分散パラメータだけは自分で定義する{var}
jeffreys:分散パラメータの密度を$1/\sigma^2$とする

線形回帰のベイズ推定1

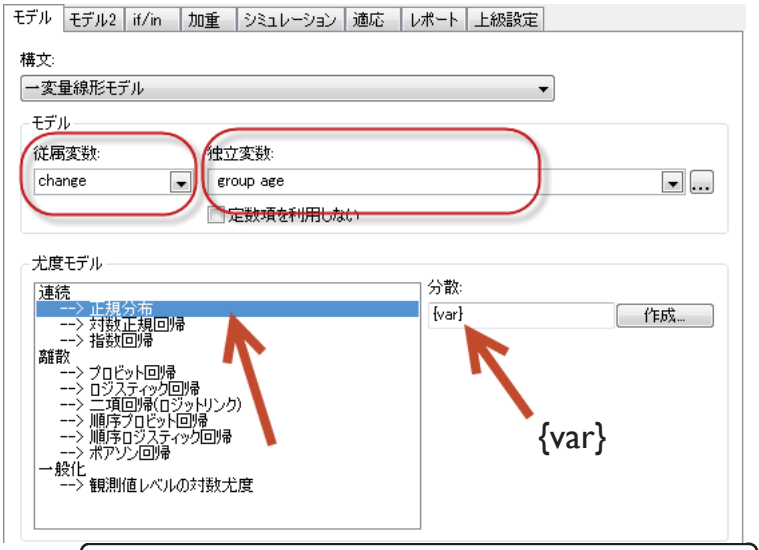
```
Bayesian normal regression                      MCMC iterations  =     12,500
Random-walk Metropolis-Hastings sampling        Burn-in          =      2,500
                                                MCMC sample size =     10,000
                                                Number of obs    =         12
                                                Acceptance rate  =      .1371
                                                Efficiency:  min =     .02687
                                                             avg =     .03765
Log marginal likelihood = -24.703776                         max =     .05724
```

受容率：MCMCアルゴリズムにより計算したパラメータの受容率。ここでは1万個の中の13.7%を受容。一般に、MH法の場合は30%を下回る。10%以下の場合は収束に問題あり。

効率性:1%以下の場合は、MCMCサンプラーを調整する。

ダイアログを利用する

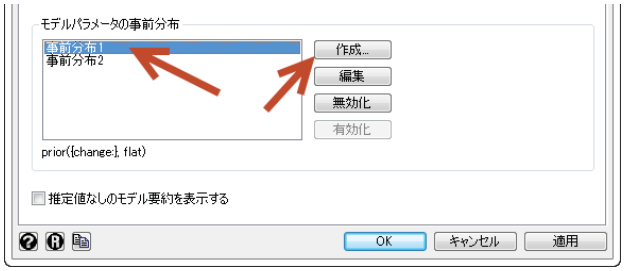
.db bayesmh

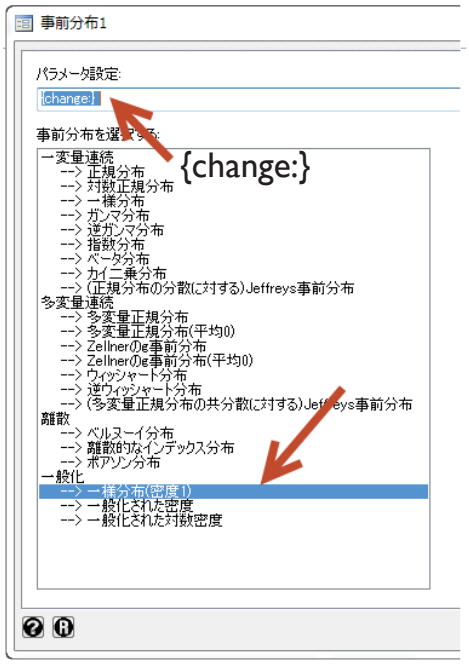


.bayesmh change group age,likelihood(normal({var})) //
prior({change:},flat) prior({var},jeffreys)

事前分布の設定

.db bayesmh

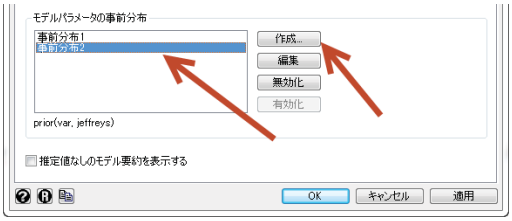


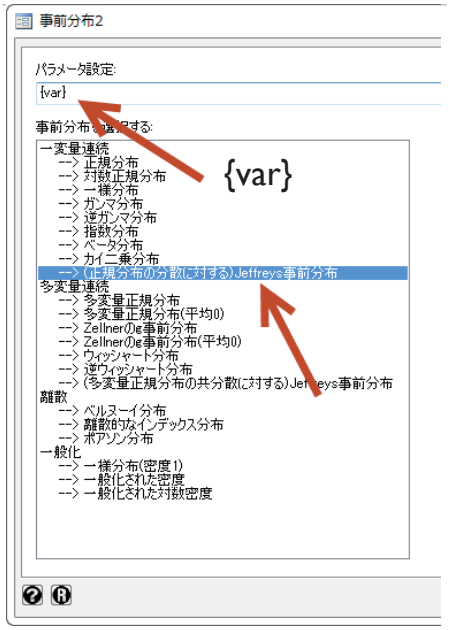


.bayesmh change group age,likelihood(normal({var})) ///
prior({change:},flat) prior({var},jeffreys)

事前分布の設定

.db bayesmh





.bayesmh change group age,likelihood(normal({var})) ///
prior({change:},flat) prior({var},jeffreys)

ベイズ推定2

$$change = \beta_0 + \beta_{group} group + \beta_{age} age + \epsilon$$

- ✓ 3つの係数推定値は正規分布
- ✓ 誤差分散は逆ガンマ分布(共役事前分布)

$$(\beta|\sigma^2) \sim \text{i.i.d. } N(0, \sigma^2)$$

$$\sigma^2 \sim \text{InvGamma}(2.5, 2.5)$$

事前分布

逆ガンマ分布

線形回帰のベイズ推定2

```
.bayesmh  change group age,likelihood(normal({var}))   ///
        prior({change:},normal(0,{var})) prior({var},igamma(2.5,2.5))
```

```
Bayesian normal regression
Random-walk Metropolis-Hastings sampling

MCMC iterations  =     12,500
Burn-in          =      2,500
MCMC sample size =     10,000
Number of obs    =         12
Acceptance rate  =      .1984
Efficiency:  min =     .03732
             avg =     .04997
             max =     .06264

Log marginal likelihood = -49.744054
```

	Mean	Std. Dev.	MCSE	Median	Equal-tailed [95% Cred. Interval]	
change						
group	6.510807	2.812828	.129931	6.50829	.9605561	12.23164
age	.2710499	.2167863	.009413	.2657002	-.1556194	.7173697
_cons	-6.838302	4.780343	.191005	-6.683556	-16.53356	2.495631
var	28.83438	10.53573	.545382	26.81462	14.75695	54.1965

ギブス・サンプラー

- bayesmhコマンドはメトロポリス-ヘイスティング(MH)アルゴリズムを利用したマルコフ連鎖モンテルロ法を実行するコマンド。
- ギブス・サンプラーはMHアルゴリズムの特別な場合。
- θを反復的に発生させ、事後分布による標本を得る。

パラメータ: $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$
観測データ: $\mathbf{y}$
パラメータの事後分布: $\pi(\boldsymbol{\theta}|\mathbf{y})$

ギブス・サンプラー

Step1. 初期値として $\boldsymbol{\theta}^{(0)} = \left(\theta_1^{(0)}, \theta_{2,}^{(0)} \ldots, \theta_p^{(0)}\right)$ を適当な分布から発生させる

Step 2. $\boldsymbol{\theta}^{(i)} = \left(\theta_1^{(i)}, \theta_{2,}^{(i)} \ldots, \theta_p^{(i)}\right)(i \geq 0)$ が得られたら

Step 2a. $\theta_1^{(i+1)}$ を $\pi\left(\theta_1|\theta_{2,}^{(i)} \ldots, \theta_p^{(i)}, y\right)$ から発生させる

Step 2b. $\theta_2^{(i+1)}$を $\pi\left(\theta_2|\theta_{1,}^{(i+1)}\theta_{3,}^{(i)} \ldots, \theta_p^{(i)}, y\right)$ から発生させる

Step 2c. $\theta_3^{(i+1)}$を $\pi\left(\theta_3|\theta_{1,}^{(i+1)}\theta_{2,}^{(i+1)}\theta_{4,}^{(i)} \ldots, \theta_p^{(i)}, y\right)$ から発生させる

これを繰り返し、$\boldsymbol{\theta}^{(i+1)} = \left(\theta_1^{(i+1)}, \theta_{2,}^{(i+1)} \ldots, \theta_p^{(i+1)}\right)$ を得たら、i+1としてStep2を繰り返す

ギブス・サンプラー

この時のθの系列は次の推移核を持つマルコフ連鎖である。

$$K\left(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}^{(i+1)} | \mathbf{y}\right)$$
$$= \pi\left(\theta_1^{i+1} | \theta_{2,}^{(i)} \ldots, \theta_p^{(i)}, \mathbf{y}\right) \prod_{j=2}^{p-1} \pi\left(\theta_j^{i+1} | \theta_{1,}^{(i+1)} \ldots, \theta_{j-1}^{(i+1)}, \theta_{j+1}^{(i)}, \ldots, \theta_p^{(i)}, \mathbf{y}\right)$$
$$\times \pi\left(\theta_p^{i+1} | \theta_{1,}^{(i+1)} \ldots, \theta_{p-1}^{(i+1)}, \mathbf{y}\right)$$

ギブス・サンプリング

ギブス・サンプラーを用いて事後分布から確率標本を得る

Step1. 初期値$\theta^{(0)}$からギブス・サンプラーを実施

Step2. Nを十分大きくとり、最初のN個を初期値に依存する期間(移動検査期間)として棄てる

Step3. $\theta^{(i)}(i = N+1, \ldots, N+m)$ を事後分布からの確率標本とする

```
MCMC iterations  =     12,500
Burn-in          =      2,500
MCMC sample size =     10,000
Number of obs    =         12
Acceptance rate  =      .1371
Efficiency:  min =     .02687
             avg =     .03765
             max =     .05724
```

MHアルゴリズム

ギブス・サンプラーと同じ設定を利用する

Step1. 現在の点が$\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}$のとき、提案密度(proposal density)$q(\boldsymbol{\theta}, \boldsymbol{\theta}')$を用いて$\boldsymbol{\theta}'$を発生させ、$\boldsymbol{\theta}^{(i+1)}$の候補とする

Step2. Step1で得た$\boldsymbol{\theta}'$を確率$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}')$で$\boldsymbol{\theta}^{(i+1)}$として受容する。棄却した場合は$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}$とする

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') = \begin{cases} \min\left(\dfrac{\pi(\boldsymbol{\theta}'|\mathbf{y}) q\left(\boldsymbol{\theta}', \boldsymbol{\theta}\right)}{\pi(\boldsymbol{\theta}|\mathbf{y}) q(\boldsymbol{\theta}, \boldsymbol{\theta}')}, 1 \right) & \pi(\boldsymbol{\theta}|\mathbf{y}) q(\boldsymbol{\theta}, \boldsymbol{\theta}') > 0 \text{のとき} \\ 1 & \pi(\boldsymbol{\theta}|\mathbf{y}) q(\boldsymbol{\theta}, \boldsymbol{\theta}') = 0 \text{のとき} \end{cases}$$

Step3. Step1にもどる

事前分布の変更

```
.set seed 14
.bayesmh  change group age,likelihood(normal({var})) prior({change:},
zellnersg0(3,12,{var})) prior({var},igamma(0.5,4))
```

```
Bayesian normal regression                      MCMC iterations  =     12,500
Random-walk Metropolis-Hastings sampling        Burn-in          =      2,500
                                                MCMC sample size =     10,000
                                                Number of obs    =         12
                                                Acceptance rate  =     .06169
                                                Efficiency:  min =      .0165
                                                             avg =     .02018
Log marginal likelihood = -35.356501                         max =     .02159
```

まだ改善したい

収束判定と効率性の診断

■ 標本経路は安定的か?

標本の時系列プロットを作り、その変動が初期値に依存せず、安定的な動きになっているか確認する。

■ サンプリングの効率性の診断

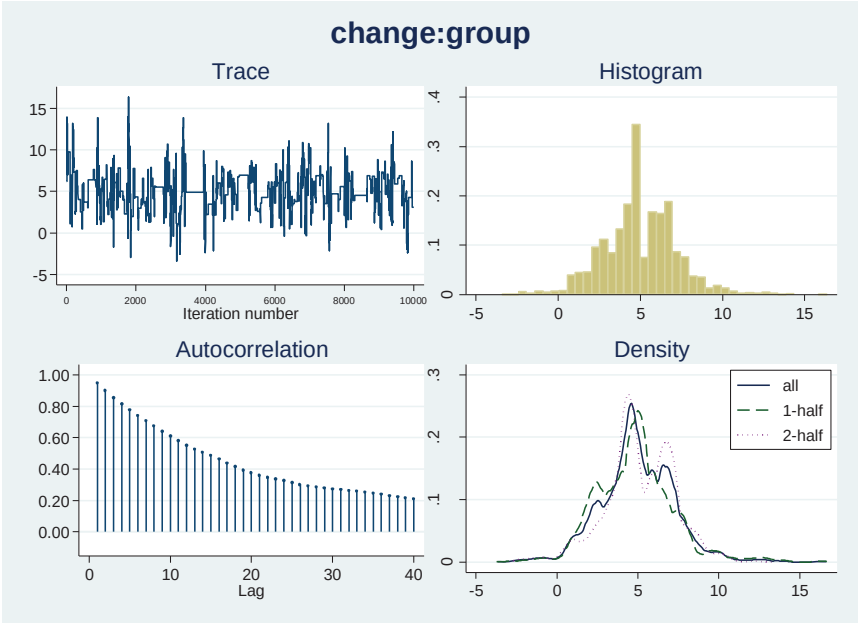
標本のコレログラムを作る。標本自己相関が急激に減衰している場合は効率的であると考える。

収束判定と効率性の診断

.bayesgraph diagnostics {change:group}

標本経路

コレログラム



効率的な標本サイズ

ESSがMCMC標本サイズに近いほど、MCMC標本の相関は弱い。
この時、パラメータの推定値はより確からしいと考える。

.bayesstats ess

Efficiency summaries MCMC sample size = 10,000

	ESS	Corr. time	Efficiency
change			
group	215.93	46.31	0.0216
age	214.39	46.64	0.0214
_cons	212.01	47.17	0.0212
var	165.04	60.59	0.0165

特に低い

*近いほど良いが、それほど近い値にはならない。特に1%以下の場合は、コマンドを再考する必要がある。

パラメータのブロック化

- MHアルゴリズムを実行する際の理想形

全パラメータが独立であること

⇩

パラメータ間の相関を最小化するように推定する

⇩

blockオプションを利用して、独立性を確保する

$$y = \{a\} + \{b\} \times x + \epsilon, \quad \epsilon \sim N(0, \{\text{var}\})$$

効率性の改良

```
.bayesmh  change group age,likelihood(normal({var})) prior({change:}, ///
          zellnersg0(3,12,{var})) prior({var},igamma(0.5,4)) block({var}) ///
          saving(agegroup_simdata)
```

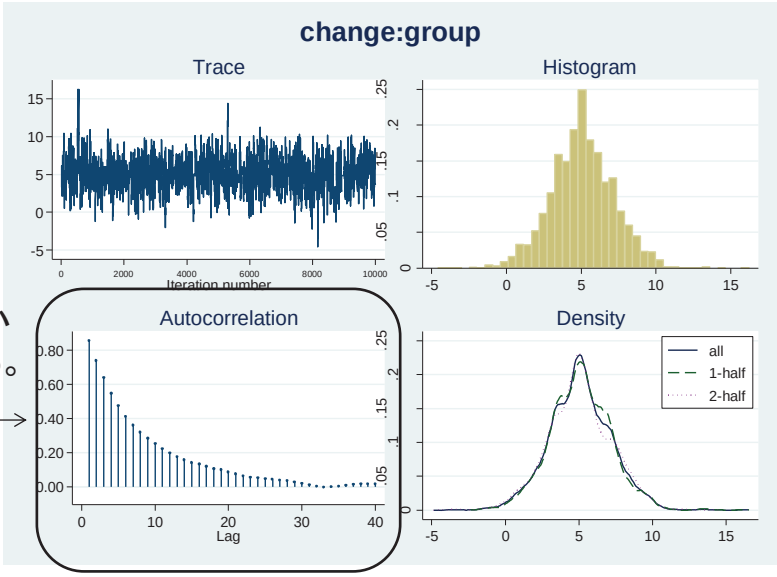
```
Bayesian normal regression                      MCMC iterations  =     12,500
Random-walk Metropolis-Hastings sampling        Burn-in          =      2,500
                                                MCMC sample size =     10,000
                                                Number of obs    =         12
                                                Acceptance rate  =      .3232
                                                Efficiency:  min =     .06694
                                                             avg =      .1056
Log marginal likelihood = -35.460606                         max =      .1443
```

かなり改善した!
効率的なMHサンプラーの場合、ARは15%から50%。

.estimates store agegroup

効率性の改良

.bayesgraph diagnostics {change:group}



自己相関が小さくなっている。

モデルの選択

- ageとgroupの交差項ageXgrを追加する

```
.set seed 14
.bayesmh change group age ageXgr, likelihood(normal({var})) prior({change:}, ///
        zellnersg0(4,12,{var})) prior({var},igamma(0.5,4)) block({var})  ///
                saving(full_simdata)
```

.

	Mean	Std. Dev.	MCSE	Median	Equal-tailed [95% Cred.	Interval]
change						
group	11.94079	16.74992	.706542	12.13983	-22.31056	45.11963
age	1.939266	.5802772	.023359	1.938756	.7998007	3.091072
ageXgr	-.2838718	.6985226	.028732	-.285647	-1.671354	1.159183
_cons	-47.57742	13.4779	.55275	-47.44761	-74.64672	-20.78989
var	11.72886	5.08428	.174612	10.68098	5.302265	24.89543

モデルの選択

. bayesstats ic full agegroup

小さいほど良い　　大きいほど良い

.

	DIC	log(ML)	log(BF)
full	65.03326	-36.73836	.
agegroup	63.5884	-35.46061	1.277756

Note: Marginal likelihood (ML) is computed
using Laplace-Metropolis approximation.

$$2 \times 1.28 = 2.56$$

*2より大きいので、agegroupの方がやや良い(mild)ことを示している。Kass and Raftery(1995)

ベイズ統計用Stataコマンド

- bayesmh
- bayesmh evaluators
- bayesmh postestimation
- bayesgraph
- bayesstats
- bayesstats ic
- bayesstats summary
- bayestest
- bayestest interval
- bayestest model

Stataによるベイズ統計

- Stata14のPDFマニュアル
  - [Bayes] Bayesian AnalysisのIntroduction
- 小西 貞則 編著、「計算統計学の方法」、朝倉書店
  - 数理統計学の知識を要する
- 松原　望 著、「入門　ベイズ統計」、東京図書
  - 医学とベイズ決定、医薬とベイズ統計学
- 中妻　照雄著、「入門　ベイズ統計学」、朝倉書店
  - ファイナンス

STATAにおける DATA MANAGEMENT

慶應義塾大学産業研究所
松浦寿幸

自己紹介

- 慶應義塾大学産業研究所　准教授　松浦寿幸
- 専門：国際貿易・海外直接投資の実証分析、産業組織論
- おもな著作
  - "Trade Liberalization in Asia and FDI Strategies in Heterogeneous Firms: Evidence from Japanese Firm-level Data," (co-authored with Kazunobu Hayakawa), *Oxford Economic Papers*, 2015, 67(2), 494-513.
  - "International Productivity Gaps and the Export Status of Firms: Evidence from France and Japan", (co-authored with Flora BELLONE, Kozo KIYOTA, Patrick MUSSO, Lionel NESTA), *European Economic Review*, 2014, 70, pp.56-74.
  - 「Stataによるデータ分析入門：経済分析の基礎からパネル分析まで」（東京図書）

Outline

- はじめに
- データの構築：ファイルの統合
  - append, merge, joinby
- 繰返し作業の省力化
  - foreach, forvalues
- 統計表、推計結果のファイル出力
  - estpost, esttab, outreg2
- 文献紹介・本報告のプログラム入手先

はじめに

- データの大規模化
  - ビックデータのなどの研究利用などが進展
- 「実証研究」の生産性改善のためには
  - データ構築のスピードアップ
  - 分析データの精度を高める
    - 外れ値の除去など
  - 再現性の確保
    - 見やすいプログラム、作業ログの保存
- Data Managementの効率化が重要

  →Stataは統計的な分析のみならず、データ構築の効率化にも有用

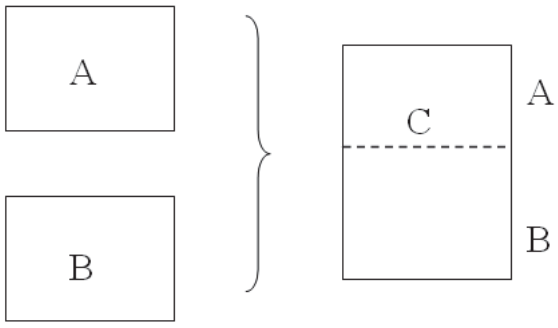
データ構築：データの統合

- 縦方向の接続：append
- 横方向の接続: merge
  - 1対1接続：one-to-one merge
  - 1対多接続：one-to-many merge
  - 多対1接続：meny-to-one merge
- Many-to-manyの接続:　joinby

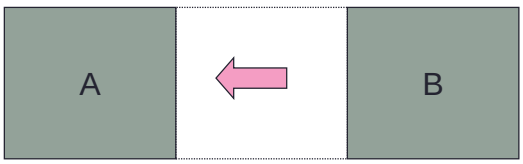
データ構築：データの統合

- データを縦方向に接続
  - appendコマンド

例）use A.dta
　　append using B.dta



- データを横方向に接続
  - merge (キー変数) using B.dta
  - 現在開いているファイルAに、ファイルBを接続する場合



Appendの例

- file_c.dtaを開いている状態でfile_dを接続

file_c.dta

firm_id	year	sales
1001	2005	2156
1002	2005	2372
1003	2005	1617
1004	2005	1779

file_d.dta

firm_id	year	sales
1001	2006	1940
1002	2006	2134
1003	2006	1455
1004	2006	1601

コマンド例

```
use file_c.dta
append file_d.dta
```

firm_id	year	sales
1001	2005	2156
1002	2005	2372
1003	2005	1617
1004	2005	1779
1001	2006	1940
1002	2006	2134
1003	2006	1455
1004	2006	1601

Mergeの例(one-to-one merge) (1)

- file_c.dtaを開いている状態でfile_dを接続

file_c.dta

firm_id	year	sales
1001	2005	2156
1001	2006	2372
1002	2005	1617
1002	2006	1779

file_d.dta

firm_id	year	emp
1001	2005	180
1001	2006	198
1002	2006	135
1002	2007	149

※キー変数は、firm_idとyear

```
use file_c.dta
merge 1:1 firm_id year  using file_d.dta
```

Mergeの例(one-to-one merge) (2)

```
. merge 1:1 firm_id year using file_d.dta

    Result                           # of obs.
    -----------------------------------------
    not matched                             2
        from master                         1  (_merge==1)
        from using                          1  (_merge==2)

    matched                                 3  (_merge==3)
    -----------------------------------------
```

	firm_id	year	sales	emp	_merge
1	1001	2005	2156	180	matched (3)
2	1001	2006	2371.6	198	matched (3)
3	1002	2005	1617	.	master only (1)
4	1002	2006	1778.7	135	matched (3)
5	1002	2007	.	148.5	using only (2)

Mergeの例 (3)

- “_merge”について: mergeコマンドを使うと自動的に作成される接続状況を示す変数
  - 1のとき:もとのファイル(A.dta)にのみ存在するデータ
  - 2のとき:接続するファイル(B.dta)にのみ存在するデータ
  - 3のとき:両方のファイル(A.dtaとB.dta)に存在する
- 例)以下の2つのファイルをmergeで接続

even.dta

number	Even
5	10
6	12
7	14
8	16

odd.dta

number	Odd
1	1
2	3
3	5
4	7
5	9

use even.dta
merge number using odd.dta

number	Even	Odd	_merge
5	10	9	3
6	12		1
7	14		1
8	16		1
1		1	2
2		3	2
3		5	2
4		7	2

Mergeの例(many-to-one merge) (4)

- file_c.dtaを開いている状態でfile_eを接続

file_c.dta

firm_id	year	sales
1001	2005	2156
1001	2006	2372
1002	2005	1617
1002	2006	1779

file_e.dta

year	price
2005	100
2006	105

※キー変数は、year

```
use file_c.dta
merge m:1 year  using file_e.dta
```


Mergeの例(many-to-one merge) (5)

```
. merge m:1 year using file_e.dta

    Result                           # of obs.
    -----------------------------------------
    not matched                             0
    matched                                 4  (_merge==3)
    -----------------------------------------
```

	firm_id	year	sales	price	_merge
1	1001	2005	2156	100	matched (3)
2	1002	2005	1617	100	matched (3)
3	1002	2006	1778.7	105	matched (3)
4	1001	2006	2371.6	105	matched (3)

Many-to-manyの接続: joinby (1)

- Joinbyコマンド:　プログラム例(A13-joingy.do)

joinby-a.dta

	year	firmid	partner	sales
1	2010	1	3	123
2	2010	2	2	180
3	2010	3	1	80
4	2011	1	2	126
5	2011	2	3	190
6	2011	3	2	89
7	2012	1	3	130
8	2012	2	3	194
9	2012	3	4	92

Mergeコマンドの場合、yearで双方のデータを特定できず、many-to-manyになってしまう！

joinby-b.dta

	year	supplier	supplier_emp
1	2010	1	120
2	2010	2	180
3	2010	3	199
4	2010	4	149
5	2011	1	122
6	2011	2	190
7	2011	3	203
8	2011	4	152
9	2012	1	124
10	2012	2	179
11	2012	3	206
12	2012	4	160

Many-to-manyの接続: joinby (2)

	year	firmid	partner	sales	supplier	supplier_emp
1	2010	1	1	123	3	199
2	2010	1	0	123	4	149
3	2010	1	0	123	1	120
4	2010	1	0	123	2	180
5	2010	2	1	180	2	180
6	2010	2	0	180	4	149
7	2010	2	0	180	3	199
8	2010	2	0	180	1	120
9	2010	3	0	80	2	180
10	2010	3	0	80	4	149
11	2010	3	1	80	1	120
12	2010	3	0	80	3	199
13	2011	1	0	126	4	203
14	2011	1	0	126	1	122
15	2011	1	1	126	2	190
16	2011	1	0	126	3	199
17	2011	2	0	190	1	120
18	2011	2	0	190	2	180
19	2011	2	1	190	3	199
20	2011	2	0	190	4	152

データの完成形：ある企業(firmid)が潜在的な供給元(supplier)からどの企業を選んだか（partner）

```
use joinby-a.dta,clear
joinby year using join-b.dta
replace partner=partner==supplier
```

繰り返し作業の省力化

- forvaluesコマンド
  - 以下の例では、1980年から1ずつ大きくなる数値time2を作成する
    → i に数値を代入する

```
gen time2=0
forvalues i=1980/2002 {
replace time2=`i'-1979 if year==`i'
}
```

A8-repeat-command.do

- forvalues i=1980(2)2002: 1つ飛ばしで数値を入力

繰り返し作業の省力化

- foreach
  - 以下の例では v のところにvarlistの後ろの変数を代入する

```
foreach v of varlist rent service age floor bus walk {
egen `v'_mean=mean(`v')
gen dif_`v'=`v'-`v'_mean
}
```

A8-repeat-command.do

- さまざまな応用が可能
- 例　foreach file in file_a.dta file_b.dta {
  append using “`file’”
  }

統計表・推計結果のファイル出力

- User Written Programの利用
  - インターネット接続環境であれば、ssc、あるいはfinditでプログラムを検索し、インストールする
- 統計表のファイル出力プログラム
  - esttpost（estoutパッケージに含まれる）
- 推計結果のファイル出力プログラム
  - esttab（estoutパッケージに含まれる）
  - outreg2

統計表のファイル出力

- estpost & esttab コマンド
  (estoutパッケージをインストール)
- sumの結果を出力したいとき

```
estpost sum (変数名)
esttab using（ファイル名）,cells(“mean sd max min”) nonumber
```

- cells: 出力したい統計表の表頭項目を列挙、sumの場合、出力したい統計量を指定
- mean, sd, max, minは平均値、標準偏差、最大値、最小値を示す
- mean(fmt(2))：統計量の小数点以下の桁数を調整したいとき
- nonumber：統計表の通し番号を表示させない

統計表のファイル出力

- tabstatの結果を出力したいとき

```
estpost tabstat (変数名)
esttab using（ファイル名）,cells(“変数名”) unstack noobs
nonumber
```

- cells: 出力したい統計表の表頭項目を列挙、
- unstack: 統計量を複数の列に並べる。これを付けないと縦に長い表になる。
- noobs: サンプル数を表示させないオプションです。これを付けない場合、表の下にサンプル数が表示されます。
- nonumber: 通り番号を付与しないオプションです。

統計表のファイル出力

- correlの結果を出力したいとき

```
estpost correl (変数名), matrix
esttab using（ファイル名）, not unstack compress nostar
noobs
```

- not: オプションはt値を表示させない
- unstack: 統計量を複数の列に並べる
- compress: 横方向の余白を小さくする
- noobs: 表の下にサンプル数を表示させない
- nostar: 有意性を示す*を付けない

統計表のファイル出力

- tabulate の統計表出力

```
estpost tabstat (変数名)
esttab using（ファイル名）, not unstack compress noobs
nolegend nonumber
```

- not: t値を表示させない
- unstack: 横長の表にする
- compress: 横方向の余白を小さくする
- nooobs: 表の下にサンプル数を表示させない
- nolegend: 表の下に***や**の説明を付与しない
- nonumber: 通し番号を付与しない

統計表のファイル出力: プログラム例

```
* sumの結果出力
estpost sum rent service if year==1999
esttab using table-estpost.csv, ///
cells("mean(fmt(2)) sd(fmt(2)) min(fmt(2)) max(fmt(2))") nonumber noobs replace
* tabstatの結果出力(1)
estpost tabstat rent service age walk,by(year) statistics(mean sd)
esttab using table-estpost.csv, ///
cells("rent(fmt(2)) service(fmt(2)) age(fmt(2)) walk(fmt(2))") noobs nonumber append
* tabstatの結果出力(2)
estpost tabstat rent service ,by(year) statistics(mean sd max min) col(stat)
esttab using table-estpost.csv, ///
cells("mean(fmt(2)) sd(fmt(2)) max(fmt(2)) min(fmt(2))") noobs nonumber append
* correlの結果出力
estpost corr rent floor age, matrix
esttab using table-estpost.csv, ///
not unstack compress noobs nonumber nostar append
* tabulate の結果出力
estpost tab auto_lock year
esttab using table-estpost.csv, ///
not unstack compress noobs nolegend nonumber append
```

A18-table-estpost.do

	mean	sd	min	max
rent	7.37	1.69	4.60	11.60
service	0.36	0.25	0.00	0.90
	rent	service	age	walk
1999				
mean	7.37	0.36	7.19	4.85
sd	1.69	0.25	4.46	4.00
2004				
mean	9.48	0.17	8.19	4.19
sd	2.96	0.22	10.74	4.00
Total				
mean	8.46	0.26	7.71	4.51
sd	2.63	0.25	8.26	3.98
	mean	sd	max	min
1999				
rent	7.37	1.69	11.60	4.60
service	0.36	0.25	0.90	0.00
2004				
rent	9.48	2.96	18.00	5.30
service	0.17	0.22	0.80	0.00
Total				
rent	8.46	2.63	18.00	4.60
service	0.26	0.25	0.90	0.00
	rent	floor	age	
rent	1			
floor	0.843	1		
age	−0.337	−0.0476	1	
	1999	2004	Total	
NO	28	21	49	
YES	6	15	21	
Total	34	36	70	

推定結果の保存

- 推定結果の論文形式→整理するのは面倒

モデル	model 1	model 2	model 3	model 4	model 5
サンプル	全サンプル	重電・家電	電子・通信機器	地方圏	都市圏
Number of obs	2031	580	1451	1207	824
log likelihood	-589.5	-90.5	-448.6	-347.1	-181.8
LR chi2:	73.9	15.9	61.5	43.4	29.8
企業属性					
売上高利益率	-7.946	-14.353	-8.904	-8.269	-9.364
	[-5.936]**	[-2.080]**	[-5.799]**	[-3.090]**	[-4.630]**
ln(従業者数)	-0.179	-0.475	-0.065	-0.175	-0.105
	[-2.122]**	[-2.328]**	[-0.645]	[-1.627]	[-0.709]
ln(平均賃金)	-0.165	-0.147	-0.140	-0.200	-0.127
	[-2.089]**	[-0.600]	[-1.608]	[-2.132]**	[-0.750]
R&D対売上比率	-11.086	11.319	-16.732	-8.099	-24.698
	[-1.893]*	[1.101]	[-2.361]**	[-1.187]	[-1.990]**
立地地域要因					
ln(経済集積)	-0.425	0.247	-0.628	-0.670	0.084
	[-2.184]**	[0.572]	[-2.710]**	[-2.353]**	[0.227]
ln(地域別賃金)	3.033	1.494	4.374	4.642	2.519
	[2.328]**	[0.505]	[2.899]**	[2.733]**	[0.877]
人的資本	-7.655	-8.983	-9.172	-17.441	-5.597
	[-1.154]	[-0.624]	[-1.202]	[-1.342]	[-0.477]
業種要因					
相対価格	-0.030	-0.054	-0.016	-0.035	-0.005
	[-2.818]**	[-1.647]*	[-1.112]	[-2.812]**	[-0.238]
ln(産業集積)	0.078	0.105	0.075	0.069	-0.009
	[0.685]	[0.440]	[0.538]	[0.502]	[-0.039]

注) 括弧内は、z値。*は10%、**は5%で統計的に有意であることを示す。

推定結果の保存:outreg2

- OUTREG2の使い方

最初だけreplace

```
reg y x1 x2
outreg2 using result1.xls,excel stats(coef tstat) replace
reg y x1 x2 x3
outreg2 using result1.xls,excel stats(coef tstat) append
```

2回目以降は、appendと記入

- Probit/Logitの場合

```
logit y x1 x2
margins, dydx(_all) post
outreg2 using result1.xls,excel stats(coef se) replace
```


推定結果の保存:outreg2

プログラム例

```
reg rent floor age distance
outreg2 using result1.xls,excel replace

reg rent floor age distance auto_lock
outreg2 using result1.xls,excel append

reg rent floor age distance auto_lock kozashibuya
shonandai mutsuai
outreg2 using result1.xls,excel append
```

chapter3-8.do

27

```
. outreg2 using result1.xls,excel append
result1.xls
dir : seeout

.
. reg rent floor age distance auto_lock kozashibuya shonandai mutsuai
```

Source	SS	df	MS	Number of obs	=	221
				F(7, 213)	=	221.07
Model	1344.78421	7	192.112031	Prob > F	=	0.0000
Residual	185.098772	213	.86900832	R-squared	=	0.8790
				Adj R-squared	=	0.8750
Total	1529.88299	220	6.95401357	Root MSE	=	.93221

rent	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
floor	.1040549	.0035133	29.62	0.000	.0971295	.1109802
age	-.0717996	.0084331	-8.51	0.000	-.0884227	-.0551766
distance	-.0361675	.0129581	-2.79	0.006	-.06171	-.010625
auto_lock	1.600204	.193175	8.28	0.000	1.219424	1.980983
kozashibuya	-.0515765	.2235953	-0.23	0.818	-.4923195	.3891665
shonandai	.9115129	.163591	5.57	0.000	.5890482	1.233978
mutsuai	.8899935	.2733351	3.26	0.001	.3512051	1.428782
_cons	3.410714	.2595725	13.14	0.000	2.899054	3.922374

```
. outreg2 using result1.xls,excel append
result1.xls
dir : seeout
```

28

	A	B	C	D
1				
2		(1)	(2)	(3)
3	VARIABLES	rent	rent	rent
4				
5	floor	0.114***	0.104***	0.104***
6		(0.00451)	(0.00380)	(0.00351)
7	age	-0.102***	-0.0709***	-0.0718***
8		(0.0106)	(0.00916)	(0.00843)
9	distance	-0.0414**	-0.0514***	-0.0362***
10		(0.0164)	(0.0134)	(0.0130)
11	auto_lock		2.059***	1.600***
12			(0.195)	(0.193)
13	kozashibuya			-0.0516
14				(0.224)
15	shonandai			0.912***
16				(0.164)
17	mutsuai			0.890***
18				(0.273)
19	Constant	4.083***	4.054***	3.411***
20		(0.295)	(0.240)	(0.260)
21				
22	Observations	221	221	221
23	R-squared	0.779	0.854	0.879
24	Standard errors in parentheses			
25	*** p<0.01, ** p<0.05, * p<0.1			

推計結果の保存:outreg2

- 主なオプション
  - stats(coef tstat)：係数とt値を表示、標準誤差→se
  - dec(x): 小数点以下の桁数を指定
  - excel: エクセル形式で出力、texと入れるとLaTex形式になる
  - addstat(Adj R-squred, `e(r2_a)', F-stat, `e(F)')
    - 自由度調整済み決定係数とF値を表示
    - e(r2_a)は直前の推計結果の自由度調整決定係数
    - e(x)の統計量は、ereturn listコマンドで確認できる

推計結果の保存:esttab

- esttabの使い方（estoutパッケージをインストール）

```
eststo: reg y x1 x2
eststo: reg y x1 x2 x3
esttab using ex-estout1.csv, se ar2 scalars(F) replace
est clear
```

- オプションなど
  - 出力はテキストファイルなのでCSVがお薦め、拡張子をtexにするとLaTex形式でも出力できる
  - r2: 決定係数、ar2:　自由度調整済み決定係数を表示
  - scalars(X): ereturn listで表示される統計量を追加
  - est clearで内部で保存している推計結果を削除

推計結果の保存:esttab

```
eststo: reg rent floor age

eststo: reg rent floor age distance auto_lock


* 例1
esttab using ex-estout1.csv, replace

* 例2
esttab using ex-estout2.csv, se ar2 scalars(F) replace

* 例3
eststo: reg rent i.year##c.floor age distance auto_lock
esttab using ex-estout3.csv, ar2 replace

eststo clear
```

A27-esttab.do

	A	B	C	D
1				
2		(1)	(2)	
3		rent	rent	
4				
5	floor	0.112***	0.104***	
6		(24.96)	(27.33)	
7				
8	age	-0.0965***	-0.0709***	
9		(-9.15)	(-7.74)	
10				
11	distance		-0.0514***	
12			(-3.83)	
13				
14	auto_lock		2.059***	
15			(10.55)	
16				
17	_cons	3.660***	4.054***	
18		(14.96)	(16.90)	
19				
20	N	221	221	
21				
22	t statistics in parentheses			
23	="* p<0.05 ** p<0.01 *** p<0.001"			

推計結果の保存:outreg2 v.s. esttab

- outreg2
  - Excel形式で出力できる
  - factor variableが利用できない
- esttab
  - factor variableが利用できる
  - 列番号が自動的に付与される
  - 各変数の間に空白行が一行ずつ入る。変数の数が多くなると縦に長い表になる

文献紹介

- Mitchell, M., 2010, *Data Management using Stata, A Practical Handbook*, Stata press

- 「逆引き事典」in 松浦寿幸著『Stataによるデータ分析入門:経済分析の基礎からパネル分析まで』(東京図書)

- 本報告のプログラム例は以下のサイトから入手可能

https://sites.google.com/site/matsuuratoshiyuki/japanese-top/text-book

サンプル脱落がもたらす推計バイアスに関する考察

A study on the estimation bias caused by sample attrition

2015年8月28日
2015 Japanese Stata Users Group Meeting

坂本和靖
（群馬大学）

国立大学法人 群馬大学
National University Corporation GUNMA UNIVERSITY

群馬大学 GUNMA UNIVERSITY

Contents

- Motivation
- Missing Data
- Mechanisms of Missing Data
- Dealing with Attrition
- Data
- Previous Research(Using KHPS)
- Empirical Example
  - Selection Model
  - Inverse Weighting Model
  - Bounds Model
- Reference

群馬大学 GUNMA UNIVERSITY

Motivation

- パネル調査(同一対象の継続的な追跡調査)における、対象者の途中脱落(Attrition)は深刻化している

The Cumulative Noninterview Rate(SIPP)

Wave	1990 Panel	1991 Panel	1992 Panel	1993 Panel	1996 Panel
1	7.3	8.4	9.3	8.9	8.4
2	12.6	13.9	14.6	14.2	14.5
3	14.4	16.1	16.4	16.2	17.8
4	16.5	17.7	18	18.2	20.9
5	18.8	19.3	20.3	20.2	24.6
6	20.2	20.3	21.6	22.2	27.4
7	21.1	21	23	24.3	29.9
8	21.3	21.4	24.7	25.5	31.3
9	–	–	26.2	26.9	32.8
10	–	–	26.6	–	34
11	–	–	–	–	35.1
12	–	–	–	–	35.5

Source: SIPP Quality Profile, 3rd Ed. (U.S. Census Bureau, 1998a).

- 政府が行っている大規模パネル調査においても、Missing Dataに関する問題、脱落が重要な瑕疵となっている

群馬大学 GUNMA UNIVERSITY

Missing Data

① “Unit nonresponse”: 調査協力しない、回答してくれない、あるいは接触できない“unit”(調査対象者の情報の多くが欠落(Missing)する場合

Ex. Attrition

②“Item nonresponse”: 回答はしてくれているが、部分的に回答してくれない場合

Ex. 回答項目に所得や支出などの数字を回答する箇所に発生しやすい

群馬大学 GUNMA UNIVERSITY

Mechanisms of Missing Data (Little and Rubin 2002)

ⅰ）Missing Completely At Random（MCAR）

脱落が完全にランダムな場合、過去（j-1 時以前）、また脱落が発生する以降（j 時以降）のどの変数にも依存しない

$$P(a_i = 1|R_{i1}, \cdots, R_{iT}; \Phi) = P(a_i = 1|\ \Phi) \text{ for all } R_{i1}, \cdots, R_{iT} \quad (1)$$

ⅱ）Missing At Random（MAR）

MAR では、脱落は観察可能な変数$R_{i1}, \cdots, R_{ij-1}$にのみ依存する。

$$P(a_i = 1|R_{i1}, \cdots, R_{iT}; \Phi) = P\left(a_i = 1|R_{i1}, \cdots, R_{ij-1}; \Phi\right)$$
$$\text{for all } R_{i1}, \cdots, R_{iT} \quad (2)$$

ⅲ）Non-ignorable、Missing At No Random(MANR)

脱落が、脱落時点以降の情報$(R_{ij}, \cdots, R_{iT})$ などの観測不可能なものに依存している場合、観測済みデータ(j-1 時点) だけでは、脱落値(j 時点) を推測することは困難

$$P(a_i = 1|R_{i1}, \cdots, R_{iT}; \Phi) = P\left(a_i = 1|R_{ij}, R_{ij+1}, \cdots, R_{iT}; \Phi\right)$$
$$\text{for all } R_{i1}, \cdots, R_{iT} \quad (3)$$

群馬大学 GUNMA UNIVERSITY

Dealing with Attrition

①Complete Case Analysis

分析で用いる変数が全て揃っているデータセットに限定して分析を行う。MCARを仮定。この場合、多くの対象者の情報を失ったままとなり、データの代表性も失う

②Single Imputation

平均値や, 前期における回答情報を代入(Last Observation Carried Forward:LOCF)、背景の似ている(同じ年齢、同じ学歴などの属性が同じ)データの値を代入(Hot-deck)。MCARを仮定。Item Nonresponseへの対応

群馬大学 GUNMA UNIVERSITY

③Model Based Solutions

Selection Model (Heckman 1974, 1976, 1979)によるサンプルセレクション2段階推定の推定・修正。第1段階で、継続回答するかどうかの選択方程式、第2段階で、継続回答されたサンプルに限定して、関心のある行動方程式(賃金関数、消費関数など)を推定。MARを仮定。

④Nonresponse Weighting

脱落などによる影響を補正する。計算方法は、継続回答確率をprobit modelで、継続回答するかどうかを被説明変数として、性別や年齢、学歴、居住地域規模などを観測可能な説明変数として推計し、その推計値を元に、その回答確率の逆数をweightとして用いる。MARを仮定。Horvitz and Thompson(1952), Robins, Rotnizky and Zhao(1995), Wooldridge(2002, 2007)

群馬大学 GUNMA UNIVERSITY

7

⑤Multiple Imputation: Rubin(1987)

重回帰やPropensity Scoreなどの(Non-)Parametricな生成により、複数の異なる値を代入し、異なる代入ごとに擬似的な完全データセットを生成。完全データセットごとに推計し、得られたM種類の解析結果を統合する。MARを仮定

⑥Partial Identification: Manski(1990, 1995, 1997)

群馬大学 GUNMA UNIVERSITY

8

Data

- 慶應義塾大学パネルデータ設計・解析センター「慶應義塾パネル調査(Keio Household Panel Survey：KHPS)」
- 対象 :全国の満20-69歳男女(層化2段無作為抽出法)
- 2004年1月～現在(年に一回)
- 調査初回(2004年)は約4,000名を、その後、1,400名(2007年)、1,000名(2012年)に新規対象を追加
- 調査項目: 構成世帯員の性別、年齢、就学就労条件、回答者の仕事や生活時間、世帯の支出、資産、住宅、健康状態など
- http://www.pdrc.keio.ac.jp/open/

慶應義塾大学 パネルデータ設計・解析センター
Panel Data Research Center at Keio University

群馬大学 GUNMA UNIVERSITY

9

継続回答率(調査全体)

	2005年 wave2	2006年 wave3	2007年 wave4	2008年 wave5	2009年 wave6	2010年 wave7	2011年 wave8	2012年 wave9
調査対象者数	4,005	3,342	2,894	4,067	3,706	3,448	3,232	3,041
うち、前年度完了数	4,005	3,342	2,887	4,062	3,691	3,422	3,207	3,030
有効回答数	3,314	2,887	2,643	3,691	3,422	3,207	3,030	2,865
うち、復活サンプル	－	0	3	0	4	7	10	10
欠票	691	455	251	371	273	222	187	175
継続回収率(%) *1	82.7	86.4	91.4	90.9	92.6	93.5	94.2	94.2

*1 (有効回答数—復活サンプル)/前年度完了数×100
参考: 慶應義塾家計パネル調査(KHPS)・日本家計パネル調査(JHPS)の概要

累積脱落率

	KHPS2004			KHPS2007			KHPS2012		
	対象者数(人)	継続回答率(%)	累積脱落率(%)	対象者数(人)	継続回答率(%)	累積脱落率(%)	対象者数(人)	継続回答率(%)	累積脱落率(%)
2004年	4,005	－							
2005年	3,314	82.75	17.25						
2006年	2,884	87.02	27.99						
2007年	2,636	91.40	34.18	1,419	－				
2008年	2,442	92.64	39.03	1,239	87.32	12.68			
2009年	2,280	93.37	43.07	1,130	91.20	20.37			
2010年	2,141	93.90	46.54	1,049	92.83	26.07			
2011年	2,037	95.14	49.14	976	93.04	31.22			
2012年	1,926	94.55	51.91	920	94.26	35.17	1,012	－	－

群馬大学 GUNMA UNIVERSITY

10

Previous Research(Using KHPS)

McKenzie et al. (2007) 宮内・他(2005) 、直井(2006)など

①対象者が回答継続するか、脱落するかは何によって決まるか？②脱落者の存在が働くかどうかの就業関数、労働時間分析、賃金関数の推計に影響するか？③賃金関数推定時に、個別の回答拒否を考慮する必要があるかどうか？

→

継続者・脱落者の属性分布の比較し、 配偶別の比較(性別・年齢階層別)有配偶の継続回答率が高い

特定項目(数字、貯蓄・負債・収入・支出)に無回答があったものは脱落する確率が高い

就業状態が無業、もしくは臨時雇用であるもの)であったものは継続回答率や項目回答確率が低いと、その後調査を拒否

群馬大学 GUNMA UNIVERSITY

11

- 直井(2006):行動方程式の被説明変数には影響しないものの、選択方程式の被説明変数に影響する変数(除外制約[Exclusion Restriction])として、(調査員訪問回数、訪問曜日への配慮、初回訪問月などの「Intensive Follow-up(Contact Effort)」を活用
- 対象者およびその世帯に関するデータ(本人票)に付随する、調査プロセスに関する情報(調査員記入票および調査員リスト)を利用
- どの調査対象者がどの調査員に当たるかは Random Assignment

群馬大学 GUNMA UNIVERSITY

12

Empirical Research

- 脱落が、政策効果(Treatment Effect)に与える影響を測定する
- 職業訓練プログラムを受講が(Treatment) Outcomeである賃金にどのような与える影響を推計
- 職業訓練状況と脱落との関係(2004-08)

		脱落	継続回答	合計
Control Group 職業訓練なし	人数	1,184	9,710	10,894
	%	10.9	89.1	100.0
Treatment Group 職業訓練あり	人数	223	1841	2064
	%	10.8	89.2	100.0
合計	人数	1,407	11,551	12,958
	%	10.9	89.1	100.0

→T.G.の継続回答率がC.Gの継続回答率より多少高い

群馬大学 GUNMA UNIVERSITY

13

- Treatment以降のOutcome(賃金)への影響

Wage

	Control Group	Treatment Group	Difference
Yt	1998.82	1897.51	−101.31
Yt+1	2036.82	2032.29	−4.53
Yt+2	2047.85	2015.83	−32.02
Yt+3	2004.04	2104.49	100.45
Yt+4	1987.99	2167.67	179.68
Yt+5	1998.84	2105.56	106.72
Yt+6	1959.28	2187.84	228.56
Yt+7	1972.08	2233.34	261.26

ln(Wage)

	Control Group	Treatment Group	Difference
Yt	7.243	7.306	0.063
Yt+1	7.272	7.358	0.086
Yt+2	7.277	7.377	0.100
Yt+3	7.272	7.407	0.135
Yt+4	7.270	7.434	0.164
Yt+5	7.275	7.421	0.147
Yt+6	7.273	7.418	0.145
Yt+7	7.275	7.440	0.166

群馬大学 GUNMA UNIVERSITY

14

- Treatment以外の説明変数

勤務年数、年齢([20],30,40,50,60歳代以上), 学歴(中学,[高校],専門他,短大・高専,大学・大学院) ,家族人数,未就学児童数, 親との同居,就業先規模([1-29人],30-99人,100-499人,500人-,官公庁),居住地域規模(大都市,[その他の市],町村),地域ブロック(8地区),年ダミー

群馬大学 GUNMA UNIVERSITY

15

Selection Model

$Y_t^*=\boldsymbol{X_t\beta}+U_1=\alpha_1+\theta T_t+\beta_1 X_t+\rho\sigma\lambda+U_1$
$S_t^*=\boldsymbol{V_{t-1}\gamma}+U_2=\alpha_2+\beta_2 T_{t-1}+\beta_3 X_{t-1}+\beta_4 Z_{t-1}+U_2$
上が行動方程式(Wage Equation)、下が選択方程式(Select Equation)
Y: Wage, X: $Age, Scale,$ *Tenure, Education and etc.* $T: Training$ (有

群馬大学 GUNMA UNIVERSITY

16

heckman Y T X, select(S= T X Z) twostep

OLSとHeckman推計

Dependent V. : Ln(wage), Independent V.: Training

	OLS			Heckman			Obs
	Coefficient	Std. Err.		Coefficient	Std. Err.		
1年後	0.03960	0.01970	**	0.03610	0.03680		8,163
2年後	0.03310	0.02090		0.02820	0.03630		7,412
3年後	0.05830	0.02180	***	0.05470	0.02220	**	6,813
4年後	0.07420	0.02250	***	0.06940	0.02740	**	6,288
5年後	0.05870	0.02650	**	0.05390	0.02680	**	4,497

* p<0.05; ** p<0.01; *** p<0.001

群馬大学 GUNMA UNIVERSITY

17

Weighting Model

$S_t^*=\alpha_2+\beta_2 T_{t-1}+\beta_3 X_{t-1}+\beta_4 Z_{t-1}+V$

からpredicted response rate $\widehat{S_t^*}$,

その逆数 $1/\widehat{S_t^*}$をweightとして用いて

$Y_t^*=\alpha_1+\theta T_t+\beta_1 X_t+U$を推計

probit S T X Z

predict pre_S

reg Y T X [pweight=1/pre_S]

群馬大学 GUNMA UNIVERSITY

18

Dependent V. : Response

	Coef.	Std. Err.			Coef.	Std. Err.	
(AGWE20)	-			regular	-0.01361	0.03622	
AGE30	0.25575	0.05404	***	retire	0.13114	0.16586	
AGE40	0.36141	0.05596	***	d_born	0.08692	0.10413	
AGE50	0.39004	0.05723	***	d_death	0.01499	0.24228	
AGE60	0.33326	0.0594	***	d_marry	0.16511	0.17603	
women	0.03424	0.0302		d_divorce	-0.3206	0.2089	
married_	0.03416	0.03774		d_separation	-0.53025	0.2542	**
nokids	0.16236	0.03561	***	d_independent	0.02155	0.15758	
(health1 good)				d_backhome	-0.25193	0.22623	
health2	0.01532	0.03913		move	-0.06478	0.07211	
health3	-0.04915	0.03624		own house	0.03567	0.03514	
health4	-0.10247	0.05052	**	visit time	-0.09585	0.01204	***
health5 bad	-0.42545	0.10534	***	first_visit_month	0.04696	0.02641	*
junior	-0.12544	0.04418	***	weekday	0.01633	0.02912	
(high)				metropolice	0.03267	0.03251	
vocational	0.01393	0.06856		(region)			
colledge	-0.00145	0.04147		rural	-0.08883	0.04077	**
university	0.02459	0.03683		household inc	-0.00011	4.1E-05	***
scale1 1-29				area block	Yes		
scale2 30-99	0.01847	0.04858		_cons	1.41357	0.11291	***
scale3 100-499	0.04669	0.04491					
scale4 500-	0.02289	0.0427					
scale5 public	-0.06298	0.06777					

Number of obs	20315
LR chi2(43)	282.45
Prob > chi2	0
Pseudo R2	0.0262
Log likelihood	-5256.23

* p<0.05; ** p<0.01; *** p<0.001

群馬大学 GUNMA UNIVERSITY

- Weighting Model(OLS, Panel)

	Weighted OLS Coefficient	Std. Err.		Obs
1年後	0.04030	0.01740	**	8,163
2年後	0.03500	0.01790	*	7,412
3年後	0.05900	0.01920	***	6,813
4年後	0.07430	0.02000	***	6,288
5年後	0.05900	0.02240	***	4,497

* p<0.05; ** p<0.01; *** p<0.001

	Non-Weighted PANEL(R.E) Coefficient	Std.Err.		Weighted PANEL(R.E) Coefficient	Std.Err.		Obs
1年後	0.06400	0.01833	***	0.06520	0.01834	***	8,163
2年後	0.05210	0.01879	**	0.05410	0.01880	**	7,412
3年後	0.07650	0.02013	***	0.07750	0.02017	***	6,813
4年後	0.11060	0.02463	***	0.11160	0.02474	***	6,288
5年後	0.08290	0.02543	**	0.08290	0.02547	**	4,497

* p<0.05; ** p<0.01; *** p<0.001

群馬大学 GUNMA UNIVERSITY

Bounds Model

これまでの推計(点識別)における様々な仮定

①被説明変数に対して、説明変数＋誤差項は線形関係にある

②説明変数の係数はすべての人にとって同一

③誤差項は説明変数群と相関しない

Manski(1989, 1990, 1995, 1997), Manski and Pepper(2000), Horowitz and Manski(2000),奥村(2015)

これまで点識別で仮定を外し、データ情報だけから、パラメータを識別する

群馬大学 GUNMA UNIVERSITY

21

- Missing Dataについて、bounds推計を用いた Horowitz and Manski(2000)

  $\underline{Y}$：Outcome Variable$Y$の取りうる最も低い値

  $\overline{Y}$：Outcome Variable$Y$の取りうる最も高い値

- パラメータを識別するのではなくて、パラメータが入りうるバウンド(幅)を識別。最初は何の仮定も起きないため、その幅がかなり大きなものとなる

群馬大学 GUNMA UNIVERSITY

22

Horowitz and Manski(2000)

Treatment Effect

upper bounds$\overline{\theta}_M$

$\overline{\theta}_M$=P$[S=1|T=1]E[y|T=1]+(1-P[S=1|T=1])\overline{Y}$
$-$P$[S=1|T=0]E[y|T=0]+(1-P[S=1|T=1])\underline{Y}$

lower bounds$\underline{\theta}_M$

$\underline{\theta}_M$=P$[S=1|T=1]E[y|T=1]+(1-P[S=1|T=1])\underline{Y}$
$-$ P$[S=1|T=0]E[y|T=0]+(1-P[S=1|T=1])\overline{Y}$

Lee(2009), DiNardo et al.(2006)

Sample Selection Indicator

- S_1 Treatment Group (職業訓練参加)に割り振られたら、観測できる(=1)　観察できない(=0)
- S_0 Control Group (職業訓練非参加)に割り振られたら、観測できる(=1)　観察できない(=0)

Type		$T=0$	$T=1$
g_{11}	$S_0=1, S_1=1$	Observed	Observed
g_{01}	$S_0=0, S_1=1$	Not Observed	Observed
g_{10}	$S_0=1, S_1=0$	Observed	Not Observed
g_{00}	$S_0=0, S_1=0$	Not Observed	Not Observed

Source : DiNardo et al.(2006) p.16

Treatment Groupに割り振られても、Control Groupに割り振られても、観測できない(g_{00}：$S_1 = 0, S_0 = 0$,)、Ex.脱落、回答拒否

$S = S_1 T + S_0(1 - T)$

→

調査対象者はTreatment Groupに割り振られ、$S_1 = 1$であるか、Control Groupに割り振られ、$S_0 = 1$であるなら、観測できる

25

- 仮定を設定することで、常に観察できるタイプg_{11}の対象者のTreatment effectの幅というか、境界を設定できます。

①Random Assignment of Treatment: Treatmentは Randomに割りつけられると仮定

②Monotonicity : treatmentの割り当てが唯一、調査対象者の継続回答するか、脱落するかに影響することを仮定

26

- 脱落しない(Non-Attriters)人々、観察可能できる人々のAverage Treatment EffectθのためのBounds
- $S_1 > S_0$を仮定

→T.G.の(継続)回答率はC.G.の(継続)回答率より高い(T.G.の脱落率は、C.G.と比べると低い)

- (Monotonicityの仮定の下で)観測されたT.G.の個人の情報は、g_{11}(常に観測できる者)とg_{01}(Treatment Groupなら観察可能。が、Control Groupなら観察できない者)の組み合わせである。
- g_{10}(Control Groupなら観察できるが、Treatment Groupなら観察できない者)は含まれていない。

群馬大学 GUNMA UNIVERSITY

Lee(2009)によるBounds

$$\overline{\theta}_L \equiv E\left[Y|T=1, S=1, Y \geq y_{(p_0)}\right] - E[Y|T=0, S=1]$$

$$\underline{\theta}_L \equiv E\left[Y|T=1, S=1, Y \leq y_{(1-p_0)}\right] - E[Y|T=0, S=1]$$

$$y_{(p_0)} \equiv {G_{S=1,T=1}}^{-1}(p_0)$$

$$p_0 \equiv \frac{P[S=1|T=1]-P[S=1|T=0]}{P[S=1|T=1]}$$

${G_{S=1,T=1}}^{-1}: T=1, S=1$時の逆累積分布関数

群馬大学 GUNMA UNIVERSITY

例：Treatment Groupの60%が継続回答し、Control Groupの50%が継続回答。ここではわれわれは、継続回答率が高いGroupの対象者数の一部を(余分なものとして)除去。除去する割合は、$\frac{P[S=1|T=1]-P[S=1|T=0]}{P[S=1|T=1]} = \frac{0.6-0.5}{0.5} = \frac{0.1}{0.5} = 0.2 = 20\%$

①Control Groupの平均値を計算

②Treatment Groupから、下位20%を削除したのちに、treatment Groupの平均を計算

③①、②で計算した両平均の差をとり、$\hat{\theta}_L$を導く

Lower Boundθ_Lの計算方法、同様に、①Treatment groupのうち、上位20%(下位80%)より高い値を取り除き、②control groupの平均と、取り除かれた後のtreatment groupの平均の差をとる

群馬大学 GUNMA UNIVERSITY

29

Narrowing bounds using covariates

- 脱落に対して、説明力がある説明変数Zは、サンプルを分割することができ、boundsを分割したもの(グループ)毎に計算できる。最終的に、分割されたもの(グループ)毎の加重平均を計算。適切な加重=継続回答確率

  DiNarido et al.(2006): Contact Effort

- 除外変数として用いられた訪問回数、訪問初回月、訪問日は平日だけかなどのうち、「訪問回数が4回以上」、「訪問日を平日だけ」をbinary variable

群馬大学 GUNMA UNIVERSITY

30

$$\overline{\theta_L}^{Z=1\text{or}0} \equiv E\left[Y|T=1,S=1,Z=1 \text{ or } 0, Y \geq y_{(p_0)}\right] - E[Y|T=0,S=1,Z=1\text{or } 0]$$

$$\underline{\theta_L}^{Z=1\text{or}0} \equiv E\left[Y|T=1,S=1,Y \leq y_{(1-p_0)},Z=1\text{or } 0\right] - E[Y|T=0,S=1,Z=1]$$

$$y_{(p_0)}{}^{Z=1} \equiv G_{S=1,T=1,Z=1}{}^{-1}(p_0)$$

$$p_0 \equiv \frac{P[S=1|T=1,Z=1]-P[S=1|T=0,Z=1]}{P[S=1|T=1,Z=1]}$$

leebounds Y T
leebounds Y T , tight(Z)

31

		Coefficient	Std.Err.	z		Number of obs.	Number of selected obs.	Trimming porportion
1年後	lower	0.0161248	0.01997	0.81		8964	8163	0.0427
	upper	0.1425082	0.01979	7.2	***			
2年後	lower	0.0044581	0.01972	0.23		8155	7412	0.0589
	upper	0.164331	0.02043	8.04	***			
3年後	lower	0.0419339	0.02148	1.95	*	7491	6813	0.0557
	upper	0.2022904	0.02142	9.44	***			
4年後	lower	0.0684036	0.02257	3.03	***	6960	6288	0.0506
	upper	0.2106991	0.02235	9.43	***			
5年後	lower	0.0531426	0.02564	2.07	**	4997	4497	0.0583
	upper	0.2096071	0.02597	8.07	***			

* p<0.05; ** p<0.01; *** p<0.001

32

Tightening

		Coefficient	Std.Err.	z		Number of obs.	Number of selected obs.	Trimming porportion
1年後	lower	0.0183732	0.02027	0.91		8964	8163	0.0427
	upper	0.1426816	0.02068	6.9	***			
2年後	lower	0.0058187	0.02009	0.29		8155	7412	0.0589
	upper	0.1636401	0.02084	7.85	***			
3年後	lower	0.0446906	0.02218	2.01	**	7491	6813	0.0557
	upper	0.2008082	0.02192	9.16	***			
4年後	lower	0.0692268	0.02334	2.97	***	6960	6288	0.0506
	upper	0.2092511	0.02396	8.73	***			
5年後	lower	0.0580053	0.02736	2.12	**	4997	4497	0.0583
	upper	0.2085591	0.02729	7.64	***			

* p<0.05; ** p<0.01; *** p<0.001

群馬大学 GUNMA UNIVERSITY

Reference

- 岩崎学(2002)『不完全データの統計解析』エコノミスト社
- 奥村綱雄(2015)「部分識別とその応用:処置効果を中心に」日本経済学会春季大会2015チュートリアルセッション
- 直井道生(2007)「家計の住居移動行動とサンプル脱落問題」樋口美雄,瀬古美喜,『日本の家計行動のダイナミズム[Ⅲ]』77-98.
- McKenzie, C. R., 直井道生,宮内環,木曽研介(2007)「労働市場における個人行動とサンプル脱落問題」樋口美雄,瀬古美喜,慶應義塾大学経商連携21世紀COE『日本の家計行動のダイナミズム[Ⅲ]』慶應義塾大学出版会,13-75.
- 宮内環,C.R. McKenzie,木村正一(2005)「回答行動の分析—調査と拒否の選択行動」樋口美雄,慶應義塾大学経商連携21世紀COE『日本の家計行動のダイナミズム[Ⅰ]』慶應義塾大学出版会,43-91.
- 宮内環, C.R.McKenzie, 木村正一(2006)「パネルデータ継続と回答行動の分析」樋口美雄,慶應義塾大学経商連携21世紀COE『日本の家計行動のダイナミズム[Ⅱ]』9-52.

群馬大学 GUNMA UNIVERSITY

- Campanelli, P., Sturgis,P. and Purdon, S.(1997)Can you hear me knocking; An investigation into the Impact of interviewers on survey response rates, The Survey Methods Centre at SCPR, London.
- DiNardo, J., J. McCrary and L. Sanbonmatsu (2006) ”Constructive Proposals for Dealing with Attrition: An Empirical Example,” University of Michigan Working Paper.
- Fitzgerald, J., P. Gottschalk and R. Moffitt, (1998a) "An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics," Journal of Human Resources, University of Wisconsin Press, vol. 33(2), 251-299.
- Fitzgerald, J., P. Gottschalk and R. Moffitt, (1998b) "An Analysis of the Impact of Sample Attrition on the Second Generation of Respondents in the Michigan Panel Study of Income Dynamics," Journal of Human Resources, University of Wisconsin Press, vol. 33(2), 300-344
- Heckman, J. J. (1976)"The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models," Annals of Economics and Social Measurement 5: 475–492.
- Heckman, J. J. (1979)"Sample selection bias as a specification error," Econometrica 47(1): 153–161.

群馬大学 GUNMA UNIVERSITY

- Horowitz J. And C. F. Manski (2000) ”Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data ,” *Journal of the American Statistical Association*, Vol. 95, No. 449, 77-84.
- Lee D.(2009)”Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects,” Review of Economic Studies, 76(3), 1071-1102.
- Lee A. Lillard and Constantijn W. A. Panis, (1998) "Panel Attrition from the Panel Study of Income Dynamics: Household Income, Marital Status, and Mortality," Journal of Human Resources, University of Wisconsin Press, vol. 33(2), 437-457.
- Lynn, P., Clarke, P., Martin, J. and Stugis, P.(2002)The effects if extended interviewer effors on nonresponse bias, in Survey Nonresponse eds. R.M. Groves, D.A. Dillman, J.L.Eltinge and R.J.A. Littile , Johm Wiley & sons, Inc, New York, 135-48.
- Manski,C.(1990) "Nonparametric Bounds on Treatment Effects," American Economic Review, American Economic Association, vol. 80(2), 319-23.
- Manski,C.(1995)Identification Problems in Social Sciences, Cambridge, M.A. Harvard University Press.
- Manski,C.(1997) "Monotone Treatment Response," Econometrica, Econometric Society, vol. 65(6), 1311-1334.

群馬大学 GUNMA UNIVERSITY

- Manski,C. and J.Pepper(2000)"Monotone Instrumental Variables: With an Application to the Returns to Schooling," Econometrica,68(4), 997–1010.
- Robins, J M., A. Rotnizky, and L. Zhao, 1994,"Estimation of Regression Coefficients When some regressors are not always observed," Journal of the American Statistical Association, 89(427), 846-866.
- Robins, J M., A. Rotnizky, and L. Zhao,, 1995,"Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," Journal of the American Statistical Association, 89(427), 846-866.
- Tauchmann, H.(2014)"Lee (2009) treatment-effect bounds for nonrandom sample selection"The Stata Journal,14(4),884-894
- U.S. Census Bureau (1998a). *Survey of Income and Program Participation Quality Profile,* 3rdEd. Washington, DC: U.S. Census Bureau.
- Wooldridge, J. M. 2002, "Inverse Probability Weighted M-estimators for Sample Selection," Portuguese Economic Journal, 1, 117-139.
- Wooldridge, J. M. 2007,"Inverse Probability Weighted Estimation for General Missing Data Problems," Journal of Econometrics, 141(2), 1281-1301.

群馬大学
GUNMA UNIVERSITY

New evidence on income distribution and economic growth in Japan

大山昌子
（龍谷大学・大阪大学）
2015/6/25

目的と貢献 (1)

- データによると、日本の所得分配は1980～2000年の間平等度が低下しているが、2000年以降は一方的に低下したとはいえない（橘木 2004,2006, 大竹 2005, 小塩 2010 等.)
- 本稿では、日本において所得分配がどのように経済成長に影響を与えたのかに関する実証研究を行った。
- このテーマでは初めて、日本の都道府県別パネルデータを用いた分析を行った。

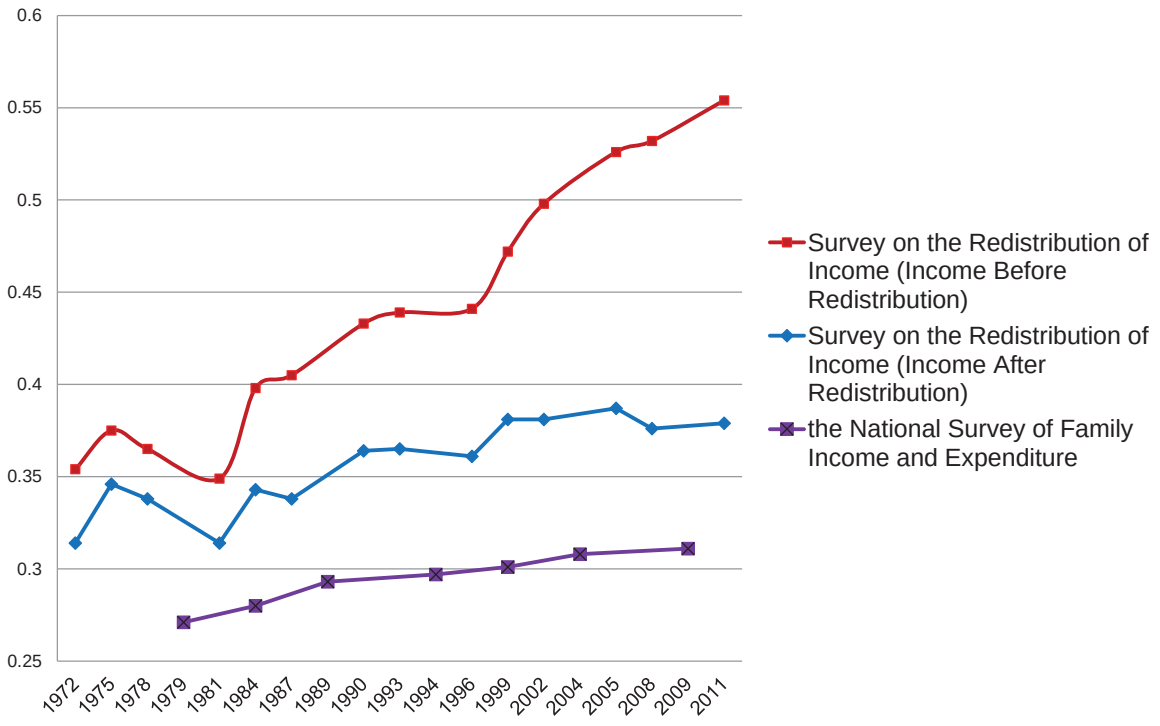
目的と貢献 (2)

- 地域別のパネルデータを用いると、国別の横断面データを用いた場合よりも、所得分配の計測方法が一貫していて比較しやすい等の長所がある。
- 異なる所得レベルでの平等度の影響を計測するために、4つの所得分配の指標を用いた。

  (1)ジニ係数(全体的な分布)

  (2)中間層, (3)高所得層 (4)低所得層

図1　日本のジニ係数



- （2000年までの平等度の変化は、高齢化と単身世帯、2人世帯の増加の影響を受けていたが、同一世代内の消費平等度の低下も観察されていた。）

5

経済成長に与える影響

- 所得分配の平等度は多くの経路を通じて経済成長に影響を与える。
- (1) ・人的資本蓄積(+)
  - ・所得再分配と効率性(+) (中位投票者定理)
  - ・（政治的安定性(+) (内閣と経済政策の安定)）
- (2) ・物的資本蓄積(-)
  - ・（高収益なプロジェクトの実現(-) ）
  - ・（R&D(-)）

6

所得平等度が経済成長に与える影響に関する既存実証研究（1）

- 既存研究では様々なデータを用いている
  (国別横断面データ、国別パネルデータ、
  一国内の地域別パネルデータ)
- 推定結果は正負の両方がある
  （データや推定方法によって異なる）
- +, -,
  ジニ係数では - でQ3では ＋ ,
  ジニ係数では+ , 高所得層の平等度は- ,
  低所得層の平等度は ＋ ,
  途上国では+, 先進国では – など

所得平等度が経済成長に与える影響に関する既存実証研究(2)

- ピケティ（2014）は、富の格差拡大は世界的に生じており、経済成長にも役立たないと指摘。

データと変数 (1)

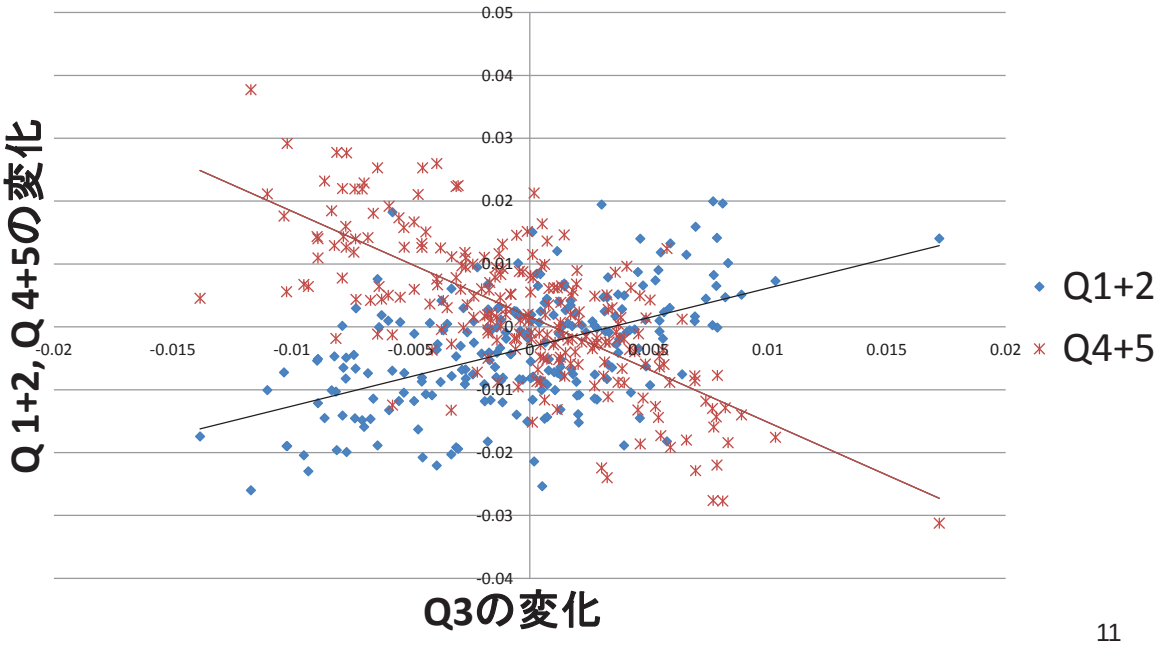
- 日本の都道府県別パネルデータ (1979~2010)
- $Growth_{(t,t+5),i}$ ：(5年間の成長率)
- $y_{t,i}$ ：(都道府県i の一人当たり県民所得の対数)
- $DISTRI_{t-1,i}$:Gini(ジニ係数)
  :Q3 (第3五分位の所得シェア)
  :90/50(第1十分位の所得シェアと第5十分位の所得シェアの比率)
  :10/50(第10十分位の所得シェアと第5十分位の所得シェアの比率)

データと変数(2)

- $X_{t,i}$ ：コントロール変数
  ・人的資本ストック (大卒者及び高卒者の県民比率)
  ・65歳以上の高齢者の比率
  ・都市化率 (都市部居住者比率)
  ・産業別労働者比率 (農業、製造業、金融・保険・不動産業、公務員)

Q3について (1)

図2. Q3の変化と Q1+2, Q4+5の変化（1979-2004）



11

Q3について (2)

- ジニ係数とQ3の相関係数は -0.378.
  (⇒Q3 は平等度の指標)

- Q3が増加する時 (減少する時),
  Q1+Q2 は増加し、(減少し) and
  Q4+Q5 は減少する (増加する)
  ⇒Q3の増加は平等度の増加を示している

12

推定

- システムGMM

$$Growth_{(t,t+5),i} = \beta y_{t,i} + \gamma DISTRI_{t-1,i} + \theta X_{t,i} + \alpha_i + \varepsilon_{t,i}$$

- (クズネッツカーブのように)経済成長は所得分配に逆方向に影響を与えるが、本研究では、所得分配が経済成長に与える影響のみを推定している。この点を明らかにするため、所得分配の変数は1年のラグを持ったものを用いている。

13

推定手法

- (1) 被説明変数（県民所得の成長率）のラグ項が説明変数に入っている ので、ダイナミック・パネル分析になっている。
- (2)データは5年間の（成長）期間が6個（30年分）である

⇒ システムGMM推定が望ましい

14

表3 システムGMM推定

	コントロール変数なし			コントロール変数あり			コントロール変数と期間ダミーあり		
Log Income	-0.314	-0.258	-0.331	-0.596	-0.598	-0.599	-0.574	-0.582	-0.575
	(.044)‡	(.024)‡	(.044)‡	(.063)‡	(.063)‡	(.065)‡	(.067)‡	(.067)‡	(.068)‡
Q3		0.386	-0.382		0.413	0.277		0.387	0.274
		(.219)†	(.350)		(.190)‡	(.326)		(.196)‡	(.331)
Gini	-0.177		-0.270	-0.125		-0.055	-0.114		-0.048
	(.069)‡		(.114)‡	(.061)‡		(.1074)	(.064)†		(.110)
High School				-0.001	-0.001	-0.001	-0.001	-0.001	-0.001
				(.0005)	(.0005)	(.0005)	(.001)	(.001)	(.001)
College				0.002	0.002	0.002	0.002	0.002	0.002
				(.001)‡	(.001)‡	(.001)‡	(.001)†	(.001)†	(.001)†
Urban				-0.001	-0.001	-0.001	-0.001	-0.001	-0.001
				(.0009)	(.0008)	(.0009)	(.0009)	(.0009)	(.0009)
Old				0.000	0.000	0.000	0.001	0.001	0.001
				(.0014)	(.0014)	(.0015)	(.0020)	(.0020)	(.0020)

Agriculture				0.002	0.001	0.001	0.002	0.001	0.002
				(.002)	(.002)	(.002)	(.002)	(.002)	(.002)
Manufactur-ing				0.003	0.003	0.003	0.002	0.002	0.002
				(.001)†	(.001)	(.001)	(.001)	(.001)	(.001)
FinanIns RealEst				0.021	0.022	0.021	0.021	0.022	0.021
				(.006)‡	(.006)‡	(.006)‡	(.006)‡	(.006)‡	(.006)‡
Governmen t				0.009	0.010	0.009	0.010	0.011	0.010
				(.009)	(.009)	(.009)	(.009)	(.009)	(.010)
Constant	1.164	0.739	1.421	1.881	1.743	1.809	1.833	1.755	1.735
	(.152)‡	(.110)‡	(.202)‡	(.238)‡	(.235)‡	(.271)‡	(.248)‡	(.240)‡	(.287)‡
N. obs.	188	188	188	188	188	188	188	188	188

Notes: standard errors in parentheses

† Denotes a parameter which is significant at 10%, ‡ at 5%.

推定結果(1)

- (1) Q3 は経済成長に正で有意の影響
- (2) ジニ係数は経済成長に負で有意の影響

⇒ 所得分配の平等度が高いと経済成長に正の影響がある

17

推定結果(2)

- 他の変数:
- 県民中の大卒者比率は経済成長に正で有意な影響があった。
- 金融･保険･不動産業の従業者数増加は経済成長に正の影響があった。

18

解釈と議論 (1)

- ジニ係数が低いこととQ3が高いことで計測された平等度の高さは成長に正の影響があった。この原因は、
  - ・(資本市場の不完全性と)人的資本投資の増加
  - ・政治的安定性(より安定した内閣と経済政策)
  - ・所得再分配の減少と効率性の上昇

- (ジニ係数は全ての人の所得分配について計測しているが、Q3は中間層(第3五分位)の所得シェアのみを計測している。)

19

解釈と議論 (2)

- 幾つかの既存研究によると、所得分配と経済成長の関係は、異なる所得レベルで異なっていた。(中間層、高所得層、低所得層)
- 従って、高所得層と低所得層における所得分配の影響について次に推定を行った。

20

表4 システムGMM推定：コントロール変数なし

	10/50	90/50	Gini and 10/50	Gini and 90/50	10/50 and 90/50	Gini, 10/50 and 90/50
LogIncome	-0.287 (.025) ‡	-0.262 (.0238) ‡	-0.335 (.0457) ‡	-0.346 (.0446) ‡	-0.287 (.0250) ‡	-0.348 (.0473) ‡
10/50	0.031 (.0344)		0.016 (.0352)		0.027 (.0342)	-0.036 (.0447)
90/50		-0.007 (.0043)		0.014 (.0090)	-0.006 (.0043)	0.020 (.0117) †
Gini			-0.143 (.0718) ‡	-0.354 (.1493) ‡		-0.461 (.1990) ‡
Constant	0.912 (.0841) ‡	0.956 (.0893) ‡	1.142 (.1534) ‡	1.329 (.1570) ‡	0.971 (.0904) ‡	1.323 (.1683) ‡
N. obs.	141	141	141	141	141	141

Notes: Robust standard errors in parentheses

† Denotes a parameter which is significant at 10%; ‡ at 5%.

表5. システムGMM推定：コントロール変数あり

	Gini	10/50	90/50	Gini and 10/50	Gini and 90/50	10/50 and 90/50	Gini, 10/50 And 90/50
LogIncome	-0.596 (.0632)‡	-0.586 (.0638)‡	-0.605 (.0632)‡	-0.596 (.0641)‡	-0.606 (.0653)‡	-0.605 (.0644)‡	-0.607 (.0670)‡
10/50		0.003 (.0321)		-0.013 (.0330)		0.005 (.0319)	0.020 (.0404)
90/50			-0.009 (.0038)‡		-0.012 (.0090)	-0.009 (.0039)‡	-0.016 (.01116)
Gini	-0.125 (.0610)‡			-0.129 (.0641)‡	0.045 (.1454)		0.114 (.1834)
HighSchool	-0.001 (.0005)	-0.001 (.0005)	-0.001 (.0005)	-0.001 (.0005)	-0.001 (.00058)	-0.001 (.0005)	-0.001 (.0006)
College	0.002 (.0009)‡	0.002 (.0009)‡	0.002 (.0008)‡	0.002 (.0009)‡	0.002 (.0009)‡	0.002 (.0009)‡	0.002 (.0009)
Urban	-0.001 (.0009)	-0.001 (.0009)	-0.001 (.0008)	-0.001 (.0009)	-0.001 (.0009)	-0.001 (.0009)	-0.001 (.0009)
Old	0.000 (.0014)	0.000 (.0014)	0.000 (.0014)	0.000 (.0015)	0.000 (.0015)	0.001 (.0014)	0.000 (.0016)

Agriculture	0.002	0.002	0.002	0.002	0.002	0.002	0.002
	(.0020)	(.00212)	(.0020)	(.0021)	(.0021)	(.0021)	(.0021)
Manufacturing	0.003	0.003	0.003	0.003	0.003	0.003	0.003
	(.0016)†	(.0016)†	(.0015)†	(.0016)†	(.0016)†	(.0016)†	(.0016)†
FinanInsReal Est	0.021	0.024	0.022	0.021	0.022	0.023	0.024
	(.0064)‡	(.0062)‡	(.0061)‡	(.0065)‡	(.0066)‡	(.0062)‡	(.0066)‡
Government	0.009	0.007	0.011	0.010	0.011	0.010	0.010
	(.0093)	(.0095)	(.0093)	(.0095)	(.0095)	(0.009)	(.0096)
Constant	1.881	1.798	1.881	1.893	1.886	1.877	1.872
	(.2380)‡	(.2389)‡	(.2355)‡	(.2442)‡	(.2423)‡	(.2409)‡	(.2459)‡
N. obs.	188	188	188	188	188	188	188

Notes: standard errors in parentheses

† Denotes a parameter which is significant at 10%, ‡ at 5%.

表6. システムGMM推定: コントロール変数と期間ダミーあり

	Gini	10/50	90/50	Gini and 10/50	Gini and 90/50	10/50 and 90/50	Gini, 10/50 and 90/50
LogIncome	-0.574	-0.569	-0.592	-0.570	-0.580	-0.591	-0.573
	(.067)‡	(.067)‡	(.067)‡	(.068)‡	(.067)‡	(.069)‡	(.069)‡
10/50		0.002		-0.013		0.004	0.027
		(.032)		(.033)		(.033)	(.041)
90/50			-0.009		-0.013	-0.009	-0.019
			(.003)‡		(.009)	(.004)‡	(.011)†
Gini	-0.114			-0.117	0.072		0.176
	(.064)†			(.067)†	(.149)		(.189)
HighSchool	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001
	(.000)	(.000)	(.000)	(.000)	(.000)	(.000)	(.000)
College	0.002	0.002	0.002	0.002	0.002	0.002	0.002
	(.001)†	(.001)	(.001)†	(.001)†	(.001)†	(.001)†	(.001)†
Urban	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001
	(.0009)	(.0009)	(.0009)	(.0009)	(.0009)	(.0009)	(.0009)
Old	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	(.0020)	(.0020)	(.0020)	(.0020)	(.0020)	(.0020)	(.0021)

Agriculture	0.002	0.001	0.001	0.002	0.002	0.001	0.002
	(.0022)	(.0022)	(.0022)	(.0022)	(.0022)	(.0022)	(.0022)
Manufactur-ing	0.002	0.003	0.003	0.002	0.002	0.003	0.002
	(.0017)	(.0017)	(.0016)	(.0018)	(.0017)	(.0017)	(.0018)
FinanIns RealEst	0.021	0.023	0.022	0.022	0.022	0.023	0.023
	(.006)‡	(.006)‡	(.006)‡	(.006)‡	(.006)‡	(.006)‡	(.006)‡
Government	0.010	0.008	0.012	0.011	0.010	0.012	0.011
	(.00983)	(.0101)	(.0098)	(.0102)	(.0100)	(.0102)	(.0102)
Constant	1.833	1.783	1.867	1.834	1.836	1.860	1.780
	(.248)‡	(.244)‡	(.245)‡	(.254)‡	(.246)‡	(.250)‡	(.250)‡
N. obs.	188	188	188	188	188	188	188

Notes: standard errors in parentheses

‡ Denotes a parameter which is significant at 10%, † at 5%.

25

推定結果 (3)

- ジニ係数は、再び経済成長に負で有意な影響があった。

⇒ 全体的な所得分配の平等さは成長に正の影響

- 第1十分位と第5十分位の所得シェアの比率（90/50）は、成長に負で有意な影響があった。

⇒ 高所得層での所得分配の平等さは成長に正の影響

- 第10十分位と第5十分位の所得シェアの比率（10/50）は、成長に有意な影響がなかった。

⇒ 低所得層での所得分配の平等さは成長に影響を与えない。

26

解釈と議論(3)

- 高所得層での平等が経済成長にプラスという結果で、低所得層の平等は成長に影響を与えないという結果は、ピケティ(2014)の議論と整合的。

27

解釈と議論(4)

- 高所得層での平等度はなぜ成長にプラスなのか？
- 高所得層での所得分配が平等になると、人々が感じる平等度が実際以上に高まり、再分配への要望が少なくなる。
- 高所得者は政治力が強く、また私立学校を利用することが多いので公的教育に支出することをあまり望まないため。
- 高所得層の所得がさらに増えると、租税回避地に移住する個人や企業が増え、効率性と税収の低下が生じるため。

28

Sensitivity analysis

- First-difference GMM 推定を行なったが、所得分配の指標の係数はシステムGMMと変わらず、推定の頑健性を確認することができた。

（表７－９）

29

結論

- 日本の都道府県別パネルデータのシステムGMM推定を用いて、様々な所得分配の指標が分配の平等度は経済成長率を高めてきたことが明らかになった。
- 全体的な平等度と、高所得層での平等度は、経済成長に正の影響があったが、
- 低所得層での平等度は成長に影響を与えていなかった。

30

STATAユーザ会資料

2015年8月28日

信用リスクのマクロストレステスト
ーモデルとインプリメンテーションー

神奈川大学経営学部准教授
京都大学博士(経済学)
菅　野　　正　泰

講　師　紹　介

菅 野 正 泰　博士(経済学)

◇神奈川大学経営学部・経営学研究科 准教授
e-mail: mkanno@kanagawa-u.ac.jp

◇専門分野:ファイナンス・金融工学・リスクマネジメント

◇外部役職
- ○日本保険・年金リスク学会(JARIP)評議員・理事
- ○日本アクチュアリー会
  - アクチュアリー基礎講座(投資理論)講師
  - ERM部会アドバイザリー
- ○ゆうちょ財団研究助成審査員
- ○金融庁金融研究センター特別研究員(～2013年8月)

その他、多数

◇主要著書
- 『信用リスク評価の実務』(中央経済社、2009)
- 『入門 金融リスク資本と統合リスク管理　第2版』(金融財政事情研究会、2010)
- 『リスクマネジメント』(ミネルヴァ書房、2011)
- 『ファイナンシャルERM』(共訳、朝倉書店)

ストレステスト

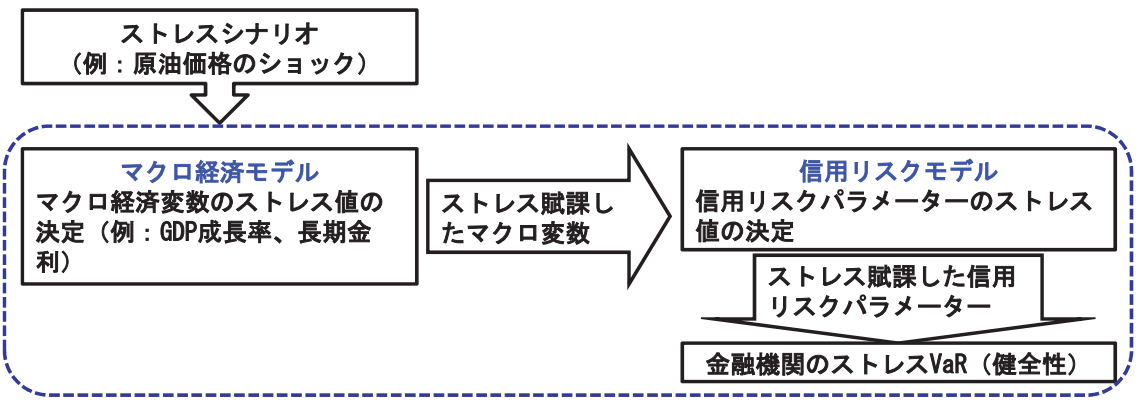
- ストレステストとは
  - 例外的ではあるが、蓋然性のあるイベントに対して金融機関がどの程度脆弱であるかを測るために用いられる手法

- ストレステストの種類
  - 感応度分析
    - ストレス・ショックは特定せず、特定のリスクパラメーターを動かしたときのインパクトを評価
  - シナリオ・テスト
    - 特定のストレス・イベントによって同時に変動する複数のリスクファクターへのインパクトを評価
    - 「ヒストリカルシナリオ」と「仮想シナリオ」

マクロストレステスト

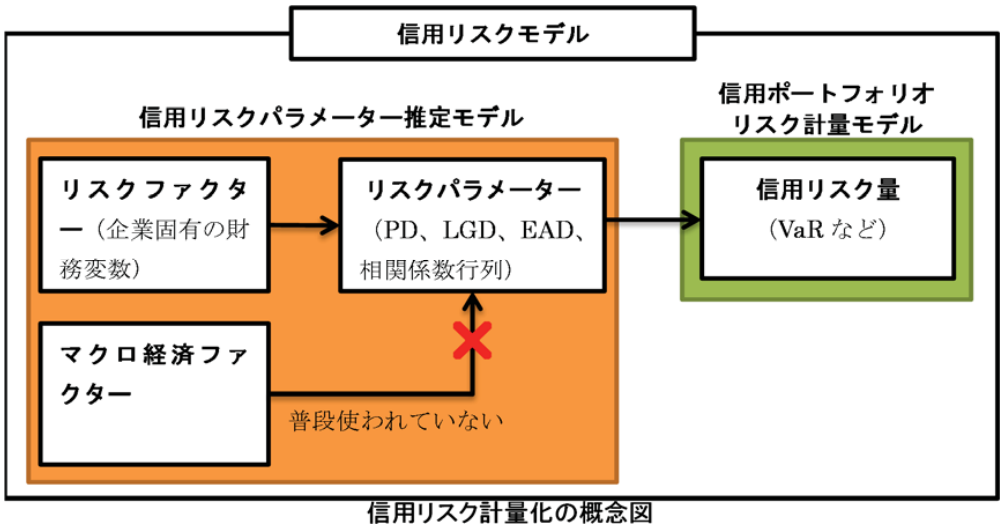
ボトムアップ型マクロストレステスト

監督当局が、あるストレスシナリオを設定し、テスト参加金融機関は、そのシナリオに対するリスクパラメーターを推定、自己のポジション（エクスポージャー）に対するシナリオのインパクトを評価。最終的に、監督当局がストレスシナリオのシステミック・インパクトを評価するために参加金融機関の評価結果を合算する手法



ストレステスト実施に伴う課題

- 金融機関が通常のリスク管理業務で使用する信用ポートフォリオリスク計量モデルは、実際のマクロ経済変数がリスクファクターとして導入されていないケースが多い。
- マクロストレスのインパクトが、通常業務にはない手続で反映されたり、あるいは、通常使用する計量モデルとはロジックが全く異なる別のモデルで計量されるため、計量される信用リスク量に一貫性がない点が課題。
  e.g. GDP成長率の単回帰分析



信用リスク計量化の概念図

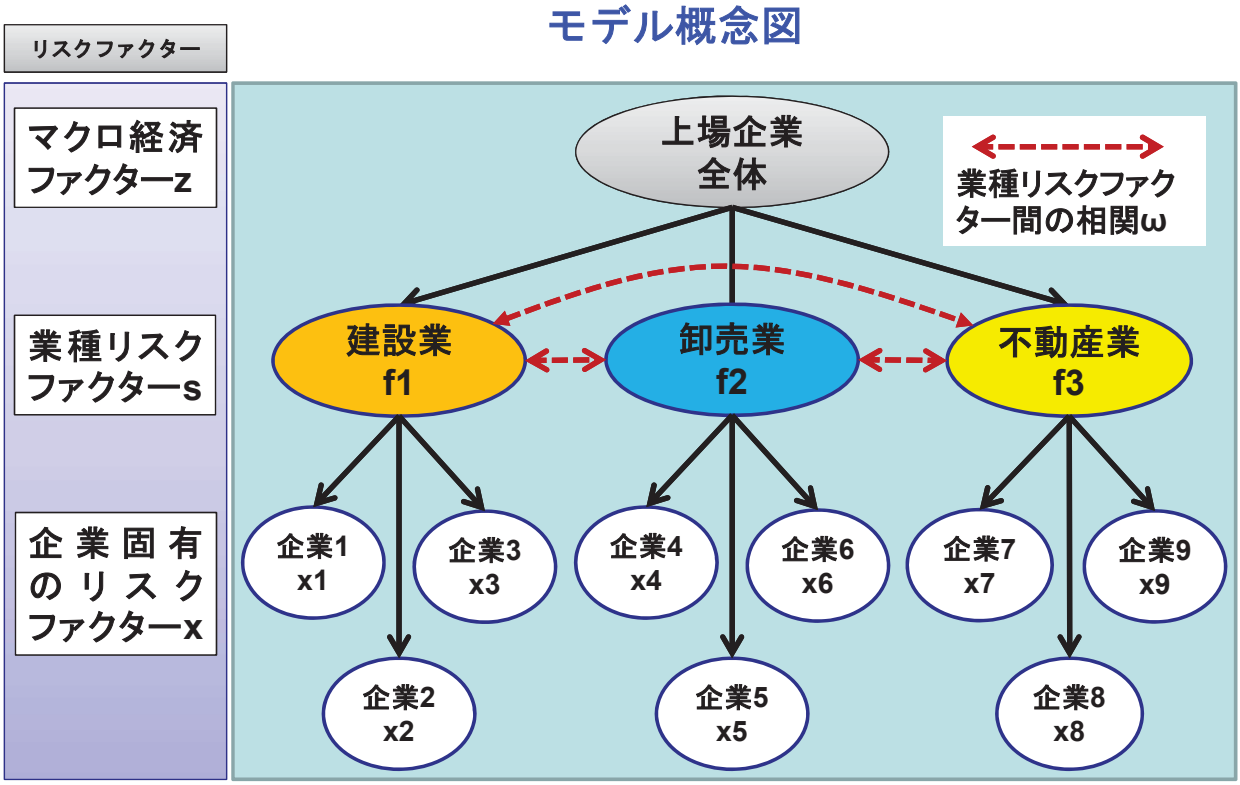
デフォルトリスクのモデリング

ねらい

- マクロストレステストの課題を踏まえ、監督当局から提示されるマクロ経済変数セットとしてのマクロストレスシナリオを、銀行が自社の内部信用ポートフォリオリスク計量モデルのリスクパラメーターに変換するためのベンチマークモデルの提案。
- 金融機関がボトムアップ型のマクロストレステストを実施する場合、「リスクファクターであるマクロ経済変数とリスクパラメーターとの関係を定式化した信用リスクパラメーター推定モデルを保有していないとき」、あるいは、「通常のリスク管理業務で使用する信用リスクモデルにリスクファクターをリスクパラメーターに変換する機能が具備されていないとき」、マクロ経済変数を信用リスクモデル（次葉以降、デフォルトリスク・モデルと呼称）に統合するアプローチが必要。

デフォルトリスク・モデル（1）

モデル概念図



デフォルトリスク・モデル（2）

信用リスクの構造モデルアプローチ

$$R_{i,t} = \beta_0 + \boldsymbol{\beta}'\mathbf{x}_{i,t-1} + \boldsymbol{\gamma}'\mathbf{z}_{t-1} + \boldsymbol{\psi}_i'\mathbf{f}_t + \varepsilon_{i,t}$$

ここで、

- $R_{i,t}\ (i = 1, \dots, N_t, t = 1, \dots, T)$:企業$i$ の期間tでの収益率。N_t:期間tの期初に生存する企業の集合。
- $x = \left(x_{i,t,1} \cdots x_{i,t,n}\right)'$: サイズ$n$,期間$t$の企業$i$固有のリスクファクター・ベクトル。個社の財務指標等観測可能な変数。
- $\mathbf{z}_t$: 期間tにおけるマクロ経済ファクター・ベクトル。
- β_0: 企業iが属する業種j固有の定数。
- $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$: 企業固有のリスクファクターとマクロファクターの係数ベクトル。
- $\mathbf{f} = \left(f_{1,t}, \dots, f_{Lt}\right)$: J種類の業種変数を含むlatentな業種リスクファクター・ベクトル。各要素は標準正規分布に従う。
- $\varepsilon_{i,t}$: 標準ロジスティック分布に従う企業i固有のリスクファクター。クロスセクション方向・時系列方向で独立。$f_{j,t}(j = 1, \dots, J)$と任意のtに関して互いに独立。
- $\boldsymbol{\psi}_i{}' = \sqrt{\psi_{i,1}}, \dots, \sqrt{\psi_{i,J}})$: $f_{j,t}$に対する係数ベクトルで以下の通り。

$$\psi_{i,j} = \begin{cases} \psi_j ; \text{企業}i\text{が業種}j\text{に属する場合} \\ 0 ; \text{それ以外} \end{cases}$$

デフォルトリスク・モデル（3）

- 業種ベクトル の相関係数行列

$$Corr(\mathbf{f}_t) := \boldsymbol{\Omega} = \begin{bmatrix} \omega_{1,1} & \cdots & \omega_{1,J} \\ \vdots & \ddots & \vdots \\ \omega_{J,1} & \cdots & \omega_{J,J} \end{bmatrix}$$

- 2社間の資産相関係数: 2社($a \in j, b \in k$)間の企業固有リスクとマクロ経済ファクターを所与）

$$Corr(R_{a,t}, R_{b,t}) = \begin{cases} \psi_j & ; j = k \\ \sqrt{\psi_j \psi_k}\,\omega_{j,k} & ; j \neq k \end{cases}$$

- Default mode方式: 資産収益率が0を下回るとその企業はデフォルトと認識

$$1_{i,t} = \begin{cases} 1; & R_{i,t} < 0 \\ 0; & \text{それ以外} \end{cases}$$

デフォルトリスク・モデル（4）

- 企業iの条件付PD（リンク関数：ロジット）

$$p_{i,t}(\mathbf{x}_{i,t-1}, \mathbf{z}_{t-1}, \mathbf{f}_t) := P(R_{i,t} < 0 | \mathbf{x}_{i,t-1}, \mathbf{z}_{t-1}, \mathbf{f}_t) = \frac{\exp(-(\beta_0 + \boldsymbol{\beta}'\mathbf{x}_{i,t-1} + \boldsymbol{\gamma}'\mathbf{z}_{t-1} + \boldsymbol{\psi}_i{}'\mathbf{f}_t))}{1 + \exp(-(\beta_0 + \boldsymbol{\beta}'\mathbf{x}_{i,t-1} + \boldsymbol{\gamma}'\mathbf{z}_{t-1} + \boldsymbol{\psi}_i{}'\mathbf{f}_t))}$$

- ポートフォリオ内のdefaultが条件付独立のとき、$\mathbf{f}_t$を所与とするポートフォリオ全体の条件付PD

$$p_t(\mathbf{x}_{i,t-1}, \mathbf{z}_{t-1}, \mathbf{f}_t) = \frac{\exp\left(-\sum_{i=1}^{N_t}(\beta_0 + \boldsymbol{\beta}'\mathbf{x}_{i,t-1} + \boldsymbol{\gamma}'\mathbf{z}_{t-1} + \boldsymbol{\psi}_i{}'\mathbf{f}_t)\, d_{i,t}\right)}{\prod_{i=1}^{N_t}(1 + \exp(-(\beta_0 + \boldsymbol{\beta}'\mathbf{x}_{i,t-1} + \boldsymbol{\gamma}'\mathbf{z}_{t-1} + \boldsymbol{\psi}_i{}'\mathbf{f}_t)))}$$

- 期間tの条件付尤度関数

$$p_t(\mathbf{x}_{i,t-1}, \mathbf{z}_{t-1}) = \iiint_{-\infty}^{\infty} \prod_{i=1}^{N_t} p_t(\mathbf{x}_{i,t-1}, \mathbf{z}_{t-1}, \mathbf{f}_t)^{d_{i,t}} (1 - p_t(\mathbf{x}_{i,t-1}, \mathbf{z}_{t-1}, \mathbf{f}_t))^{1-d_{i,t}} \phi(f_{1,t}, \ldots, f_{J,t}) df_{1,t} \cdots df_{J,t}$$

- 対数尤度関数（T期間の対数尤度の合計）

$$l = \sum_{t=1}^{T} \ln\left(\iiint_{-\infty}^{\infty} \prod_{i=1}^{N_t} p_t(\mathbf{x}_{i,t-1}, \mathbf{z}_{t-1}, \mathbf{f}_t)^{d_{i,t}} (1 - p_t(\mathbf{x}_{i,t-1}, \mathbf{z}_{t-1}, \mathbf{f}_t))^{1-d_{i,t}} \phi(\mathbf{f}_t; \mathbf{0}, \boldsymbol{\Omega}) df_{1,t} \cdots df_{J,t} \right)$$

ここで、$\phi(\mathbf{f}_t; \mathbf{0}, \boldsymbol{\Omega})$は平均ベクトル0, 相関行列$\boldsymbol{\Omega}$の$J$変量標準正規密度関数

=>STATAのgllamm関数による最尤推定

gllamm関数について

関数lの**最尤推定**

STATAの関数gllamm (Generalized linear latent and mixed models)の利用

- 数値積分：**適応ガウス求積法**
  （adaptive Gaussian quadrature method）
- 最適化数値計算：ニュートン・ラフソン法
- バージョンSE12.1で、Core i7-3770 3.4GHzのマシンを使用して、1回の計算に数時間～10数時間程度の時間を要した。

データ

データ概要

- 分析対象：わが国の上場企業
- データベース：EOL (PRONEXUS INC., Japan)
- データの種類：個社別財務データ、デフォルトデータ、マクロ経済データの3種類
- 対象業種：建設業、卸売業、および不動産業の3業種
  （デフォルト企業数が多い業種を中心に選定）

業種	延べ企業数	デフォルト企業数
建設業	1, 889	23
卸売業	3, 423	9
不動産業	1, 010	23
合計	6, 322	55

- データ項目：

自己資本（百万円）、EBITDA（百万円）、使用総資本事業利益率（ROA）（%）、売上高営業利益率（%）、売上高経常利益率（%）、売上債権回転期間（月）、流動比率（%）、当座比率（%）、固定長期適合率（%）、自己資本比率（%）、インタレスト・カバレッジ・レシオ（倍）、営業キャッシュフロー比率（キャッシュフローマージン）（%）、営業キャッシュフロー対流動負債比率（%）の13項目。

リスクファクター候補の符号

	Risk factor	Sign
Company-specific risk factors (financial ratios)	Equity capital (logarithmic transformation; million yen)	-
	EBITDA (logarithmic transformation)	-
	ROA (%)	-
	Operating margin (%)	-
	Recurring profit margin (%)	-
	Accounts receivable turnover (month)	-
	Current ratio (%)	-
	Quick ratio (%)	-
	Fixed assets to fixed liability ratio (%)	+
	Capital adequacy ratio (%)	-
	Interest coverage ratio (%)	-
	Cashflow margin (%)	-
	Cashflow to current liabilities ratio (%)	-
Macroeconomic risk factors	Real GDP growth rate (s.a., p.q., annualized %)	+/-
	CPI Core (s.a., year-to-year basis; %)	+/-
	Overall unemployment rate (s.a., year-to-year basis; %)	+/-
	Overnight call rate (month end) - latest value (%)	+/-
	10-year JGB yield (month end) - latest value (%)	+/-
	10-year JGB yield (month end) - overnight call rate (month end) (%)	+/-
	10-year long-term JGB yield to subscribers - latest value (%)	+/-
	TOPIX (first section of Tokyo Stock Exchange, month-to-month basis) latest value	+/-
	Real private final consumption expenditure growth rate (s.a., p.q.; %)	+/-
Sector-specific risk factors	Construction factor	+/-
	Wholesale trade factor	+/-
	Real estate factor	+/-

リスクファクターの推定結果（1）

モデルⅠ（マクロ経済ファクターを含む）に対するリスクファクター推定値

Variable		Estimate	S.E.	z	P>z (#)	95%C.I. [lower, upper]	
$\beta_{0,1}$	Construction intercept	-6.5393	1.3865	-4.72	0.000	-9.2569	-3.8217
$\beta_{0,2}$	Wholesale intercept	-7.5507	1.4273	-5.29	0.000	-10.3481	-4.7532
$\beta_{0,3}$	Real estate intercept	-6.8398	1.3464	-5.08	0.000	-9.4787	-4.2009
β_1	Eq. capital (log trans.)	-0.1400	0.0306	-4.57	0.000	-0.2001	-0.0800
β_2	EBITDA (log trans.)	-0.0638	0.0219	-2.92	0.004	-0.1067	-0.0209
β_3	Current ratio	-0.0281	0.0059	-4.76	0.000	-0.0396	-0.0165
β_4	FATFL ratio	0.0022	0.0008	2.64	0.008	0.0006	0.0038
γ_1	CPI Core	-1.0013	0.2902	-3.45	0.001	-1.5700	-0.4325
γ_2	Overall unemp. rate	-0.0452	0.0212	-2.13	0.033	-0.0867	-0.0037
γ_3	Overnight call rate	3.0268	0.6328	4.78	0.000	1.7865	4.2671
γ_4	JGB yield-to-sub	2.7021	0.9164	2.95	0.003	0.9060	4.4981
ψ_1	Construction factor	1.291E-18*	For reference: log-likelihood = -505.15111				
ψ_2	Wholesale trade factor	2.377E-16*					
ψ_3	Real estate factor	9.079E-14*					
$\omega_{1,2}$	Const./Wholesale	-0.9998*					
$\omega_{1,3}$	Const./Real estate	-0.5852*					
$\omega_{2,3}$	Wholesale/Real estate	0.5969*					

(#)P値を指す。*: 有意水準5%で有意。

リスクファクターの推定結果（2）

モデルⅡ（マクロ経済ファクターを含まない）に対するリスクファクター推定値

Variable		Estimate	S.E.	z	P>z (#)	95%C.I. [lower, upper]	
$\beta_{0,1}$	Construction intercept	-1.5900	0.4254	-3.74	0.000	-2.4237	-0.7562
$\beta_{0,2}$	Wholesale intercept	-2.5958	0.5108	-5.08	0.000	-3.5969	-1.5947
$\beta_{0,3}$	Real estate intercept	-1.9531	0.2991	-6.53	0.000	-2.5395	-1.3668
β_1	Eq. capital (log trans.)	-0.0821	0.0272	-3.02	0.003	-0.1355	-0.0287
β_2	EBITDA (log trans.)	-0.0648	0.0207	-3.13	0.002	-0.1054	-0.0243
β_3	Current ratio	-0.0310	0.0059	-5.26	0.000	-0.0426	-0.0195
β_4	FATFL ratio	0.0013	0.0008	1.66	0.097	-0.0002	0.0029
ψ_1	Construction factor	6.62E-20*	For reference: log-likelihood=-538.48762				
ψ_2	Wholesale trade factor	4.48E-20*					
ψ_3	Real estate factor	7.76E-20*					
$\omega_{1,2}$	Const./Wholesale	-0.7078*					
$\omega_{1,3}$	Const./Real estate	0.5778*					
$\omega_{2,3}$	Wholesale/Real estate	-0.8064*					

(#)P値を指す。*: 有意水準5%で有意。

尤度比検定

モデルⅡに対するモデルⅠの優位性の検定

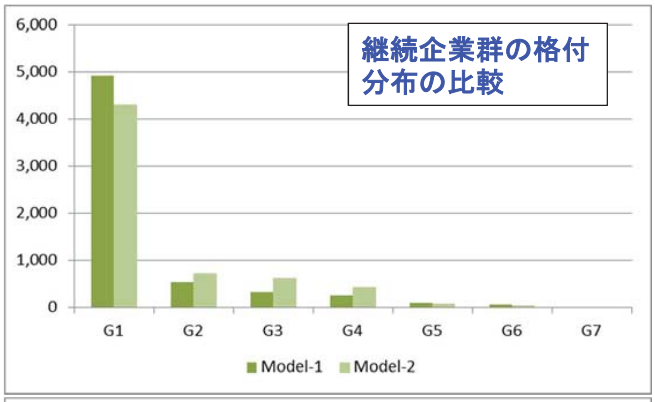
結果

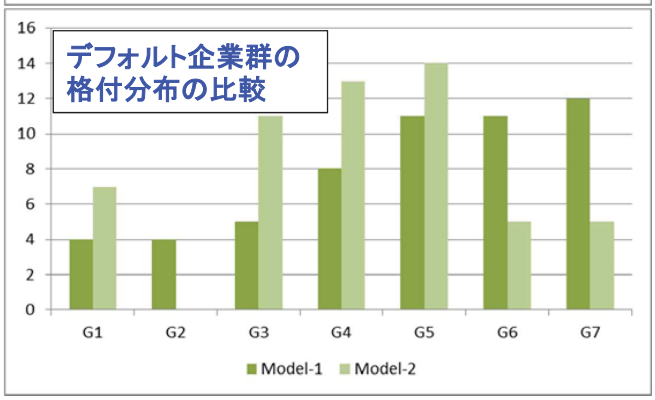
- 尤度比検定統計量（自由度5）66.67
- P値0.0000

結論

- 1%有意水準でモデルⅠのモデルⅡに対する優位性が言える。

格付分布





継続企業群とデフォルト企業群のそれぞれで格付分布を比較

継続企業群
- モデルⅠの方がモデルⅡよりも良い格付に企業が多く分布

デフォルト企業群
- モデルⅠの方がモデルⅡよりも悪い格付に企業が多く分布

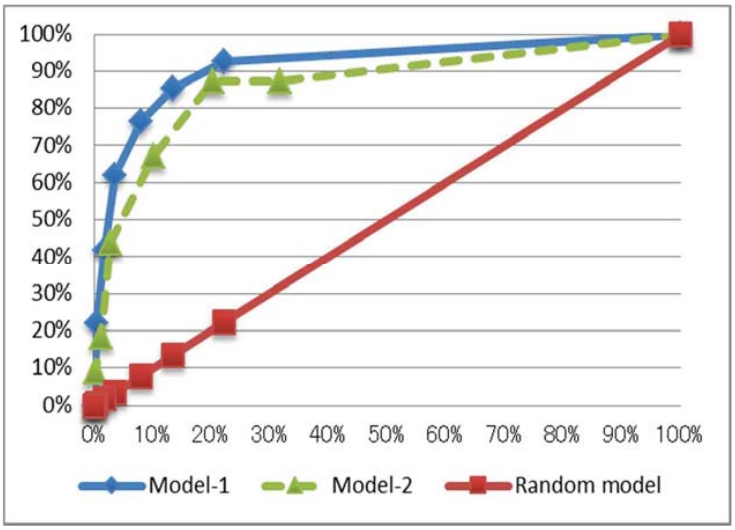
結論
マクロ経済ファクターを含む方が、継続企業群とデフォルト企業群を、より明確に峻別可能

Credit rating grade	Range of credit scores
G1	≧99.5 and ≦100
G2	≧99 and <99.5
G3	≧98 and <99
G4	≧95 and <98
G5	≧90 and <95
G6	≧68 and <90
G7	≧0 and <68

AR値

モデル別AR値

	モデルⅠ	モデルⅡ
AR値	84.1%	73.6%



モデルⅠ：マクロ経済ファクターを含む
モデルⅡ：マクロ経済ファクターを含まない

デフォルト確率の予測

ストレスシナリオの設定期間に応じて、将来期間のPDの予測値が必要。

- $\widehat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\psi}_i}$: 現時点までのデータ保有期間$T$に対応する企業$i$固有のリスクF、マクロ経済F、および業種リスクFの各係数の推定値ベクトル
- 期間$T+1$に対応する条件付PD（$\mathbf{x}_{i,T}, \mathbf{z}_T, \mathbf{f}_{T+1}$の実現値を所与）：

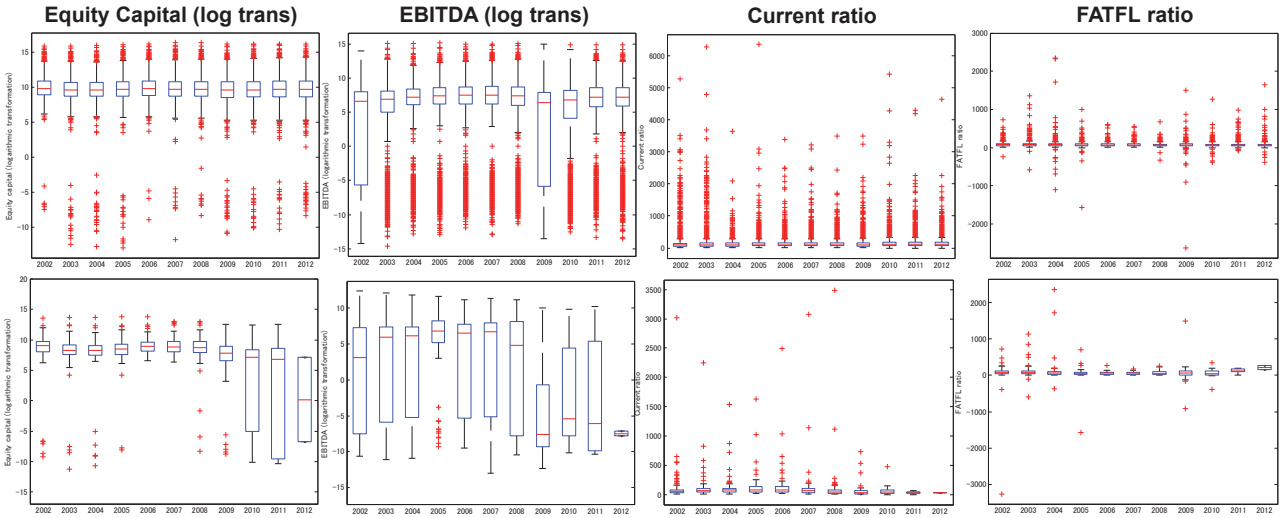
$$\hat{p}_{i,T+1}(\mathbf{x}_{i,T}, \mathbf{z}_T, \mathbf{f}_{T+1}) = \frac{\exp(-(\beta_0 + \widehat{\boldsymbol{\beta}}'\mathbf{x}_{i,T} + \widehat{\boldsymbol{\gamma}}'\mathbf{z}_T + \widehat{\boldsymbol{\psi}}_i{}'\mathbf{f}_{T+1}))}{1 + \exp(-(\beta_0 + \widehat{\boldsymbol{\beta}}'\mathbf{x}_{i,T} + \widehat{\boldsymbol{\gamma}}'\mathbf{z}_T + \widehat{\boldsymbol{\psi}}_i{}'\mathbf{f}_{T+1}))}$$

- $\hat{\mathbf{x}}_{i,T+h}, \hat{\mathbf{z}}_{T+h}, \hat{\mathbf{f}}_{T+h}$: $T+h, h = 2,3, \ldots$ に対応する企業i固有のリスクF、マクロ経済Fおよび業種リスクFの各予測値。ここで、$\hat{\mathbf{z}}_{T+h}$はストレスシナリオ、$\hat{\mathbf{x}}_{i,T+h}$は金融機関内部の方法で推定が必要。シナリオ設定期間$h = 1,2,3, \ldots$の各期間の値をヒストリカルデータから設定。$\hat{\mathbf{f}}_{T+h}$は標準正規乱数。
- 期間$T+h, h = 2,3, \ldots$に対応する条件付PD：

$$\hat{p}_{i,T+h}(\mathbf{x}_{i,T+h-1}, \mathbf{z}_{T+h-1}, \mathbf{f}_{T+h}) = \frac{\exp(-(\beta_0 + \widehat{\boldsymbol{\beta}}'\mathbf{x}_{i,T+h-1} + \widehat{\boldsymbol{\gamma}}'\mathbf{z}_{T+h-1} + \widehat{\boldsymbol{\psi}}_i{}'\mathbf{f}_{T+h}))}{1 + \exp(-(\beta_0 + \widehat{\boldsymbol{\beta}}'\mathbf{x}_{i,T+h-1} + \widehat{\boldsymbol{\gamma}}'\mathbf{z}_{T+h-1} + \widehat{\boldsymbol{\psi}}_i{}'\mathbf{f}_{T+h}))}$$

- 上記2式から、シナリオ設定期間$h = 1,2,3, \ldots$に対応したPDを予測。

財務指標のストレスシナリオ



上段：継続企業群、下段：デフォルト企業群　箱ひげ図の下端は25%点、上端は75%点

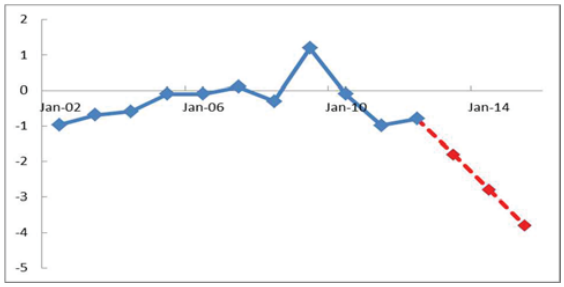
ストレスシナリオ：各予測値は、2012年3月末の各企業の財務比率に、2008年3月～2012年3月間の継続企業群の中央値に対するデフォルト企業群の中央値の比率の平均を掛けて予測。

財務指標	Equity capital (log trans)	EBITDA (log trans)	Current ratio (%)	FATFL ratio (%)
向こう3年間の増減予測（対2012年3月現在値）	0.6倍増加	−0.9倍増加	0.3倍増加	1.6倍増加

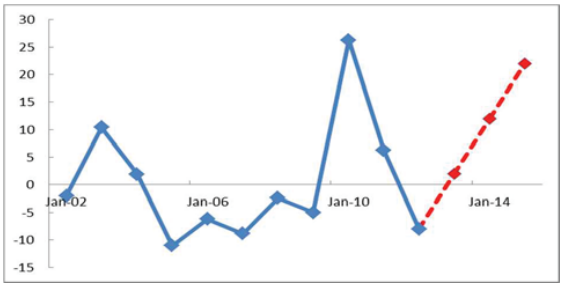
マクロ経済変数のストレスシナリオ

ストレスシナリオ：2012年、日本の金融セクターに対してIMFが実施したFSAPテスト(IMF (2012a), IMF (2012b))のシナリオ予測を参考。選択したマクロ経済変数の連続的な悪化を想定。

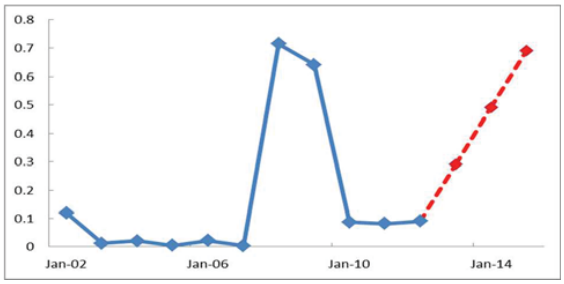
CPIコア - 成長率 (%)



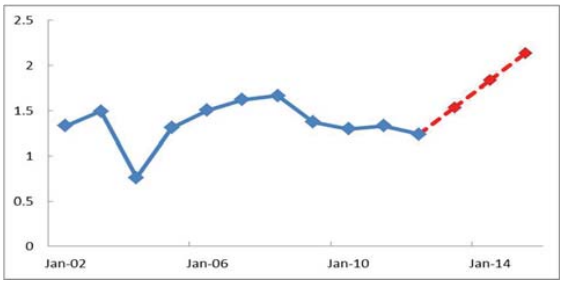
失業率 – 成長率 (%)



O/Nコールレート(%)



10年JGB応募者利回り (%)



（参考）わが国におけるFSAPストレステスト

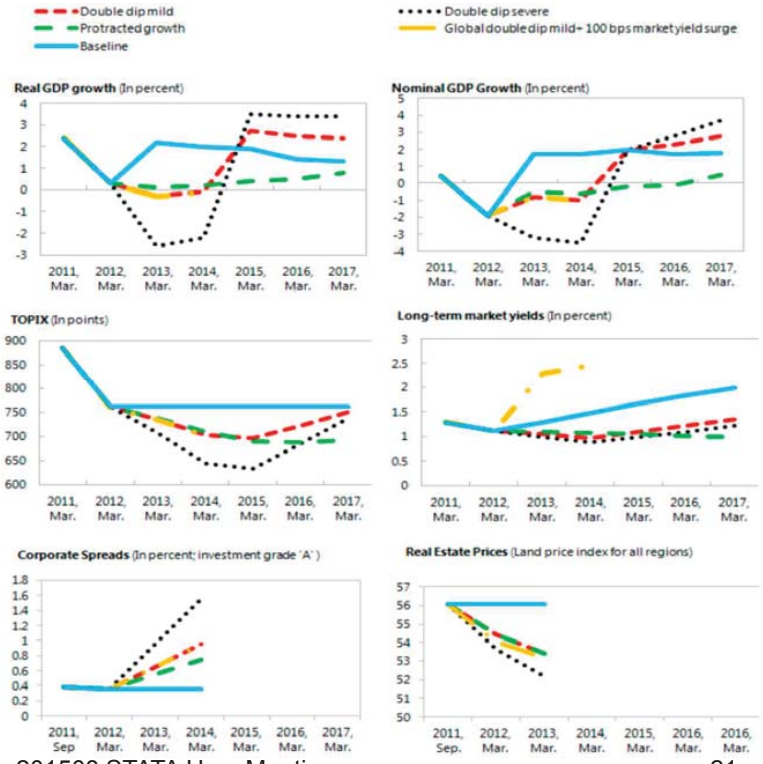
日本の主要マクロ経済変数のストレスシナリオ

基本シナリオ
2011年9月現在のWorld Economic Outlookの予測: Baseline

4つのストレスシナリオ
(a)中国における相当程度の景気後退（GDPの一標準偏差の穏やかなショック／二標準偏差の厳しめのショック）: Double dip mild, Double dip severe

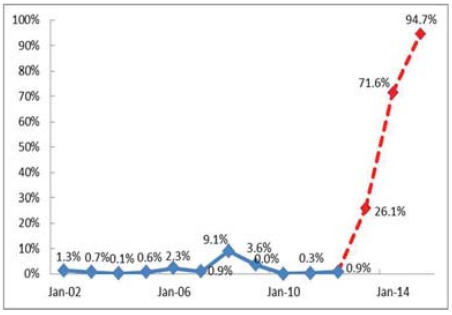
(b)デフレ圧力を伴う景気後退の長期化: Protracted growth

(c)市場利回りの急騰を伴うW型景気後退の長期化シナリオ: Global double dip mild+100bps market yield surge

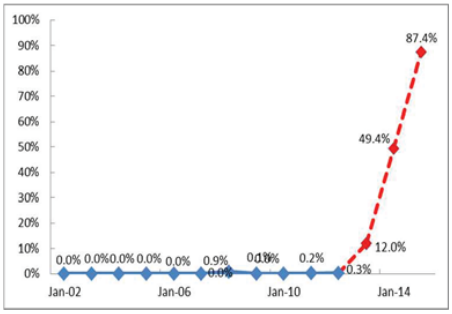


サンプル企業のPD推移

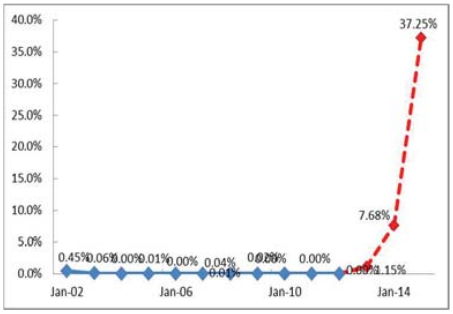
Sample 1: 建設業



Sample 2: 卸売業



Sample 3: 不動産業



実線: ヒストリカルデータ; 破線: ストレスシナリオ

ストレスVaRの計算方法

- 企業iの損失変数

$$L_{i,T+h} = D_{i,T+h} \cdot LGD_{i,T+h} \cdot EAD_{i,T+h}$$

$D_{i,T+h}$: デフォルト指標関数

$LGD_{i,T+h}$: 資産側のloss-given-default ($ELGD_{i,T+h} = 45\%$,定義域$[0, 2ELGD_{i,T+h}]$の対称な三角分布)

$EAD_{i,T+h}$:資産側のexposure-at-default (=1)

- ポートフォリオ損失

$$L_{T+h} = \sum_{i \in \Omega_{T+h}} L_{i,T+h}$$

Ω_{T+h}:$[T+h-1, T+h], h = 1,2,3, \ldots$の期初に存在する継続企業群

- 損失割合

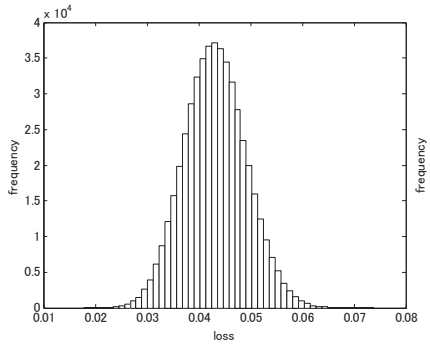
$$L_{T+h}{}^{*} = \frac{1}{\#(\Omega_{T+h})} \sum_{i \in \Omega_{T+h}} L_{i,T+h}$$

$\#(\Omega_{T+h})$: Ω_{T+h}の要素数

- 信頼水準：片側99.9%
- シミュレーション回数：50万回
- 想定ポートフォリオ：2012年3月31日現在継続している企業（上場）402社

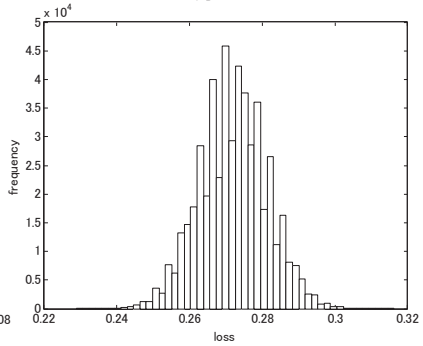
ストレスVaR

予測期間1年目



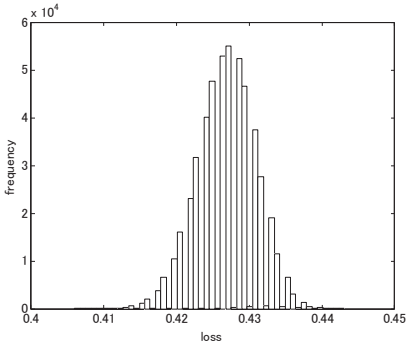
Stressed VaR=0.0626

予測期間2年目



Stressed VaR=0.3006

予測期間3年目



Stressed VaR=0.4386

PDの増加と共に増加

まとめ（1）

- 本研究は、ボトムアップ型のマクロストレステスト実施のために、マクロ経済変数を信用リスクモデルに変換するためのデフォルトリスクモデルのベンチマークとテスト実施手法を提示。

- STATA12のgllamm関数を使用して、一般化線形混合モデルにより、マクロス手レステストのためのデフォルトリスクモデルを開発。

  gllamm関数に関する参考文献・サイト

  - http://www.gllamm.org/
  - Rabe-Hesketh, S. and Skrondal, A. (2012). *Multilevel and Longitudinal Modeling Using Stata*, Third Edition, Vol I and II., Stata Press, College Station, TX.

まとめ（2）

- ボトムアップ型マクロストレステストでは、金融機関は監督当局の提示するストレス賦課したマクロ経済シナリオを信用リスクパラメーターに変換する計量モデルが必要がある。
- この際、提示された一部のデータだけを使用するのではなく、極力、データ全てを使用することが肝要。
- 同時に、従来のデフォルトリスクモデルは、財務変数を主に使用しているため、財務変数とマクロ経済変数を同時推定する方法を本研究では提案した。

参考文献

Masayasu Kanno (2014), “Macro Stress Test for Credit Risk,” submitted to a journal.

ご清聴ありがとうございました

質問・意見等はE-mailで、

mkanno@kanagawa-u.ac.jp

「xxxxx」受講生限り ©森平爽一郎

Fitting state space model is notoriously difficult. STATA sspce p.14

線形の状態空間モデル
（Linear SSPM）

カルマンフィルターのファイナンスへの応用

Stata Group Meeting in Tokyo 2015

2015年8月28日（金曜日）
神田　一橋大学 一橋講堂

早稲田大学大学院　ファイナンス研究科
森平　爽一郎

1

カルマンフィルターの応用：工学

お掃除ロボット

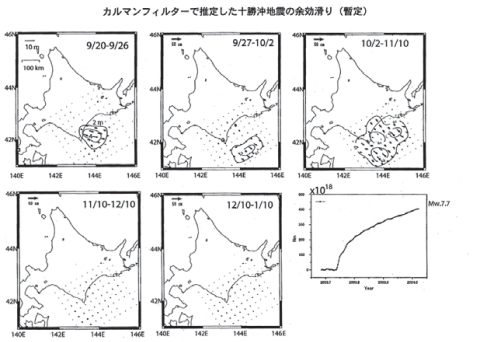
監視カメラと画像解析

KFのファイナンスへの応用

2

カルマンフィルターの応用：工学

Googleによる自動操縦



地震解析（国土地理院）

その他多方面への応用がある

報告内容

1. 状態空間モデルSSMとは？
2. STATAが想定する状態空間モデルの特徴
3. STATA　sspaceを用いた「簡単な」応用例
4. 重要な問題：初期値設定をSATAでどの様に？
5. Sspaceに関連するその他のモジュール

ファイナンス研究科での社会人への教育経験かから

STATAが想定する状態空間モデル

わかりやすいように、行列でなく、スカラー表現で説明する。

観測可能な変数
（従属変数）

外生変数
(独立変数)

誤差項

観測方程式

状態方程式

$$\tilde{y}_t = d\tilde{z}_t + fw_t + g\tilde{v}_t$$

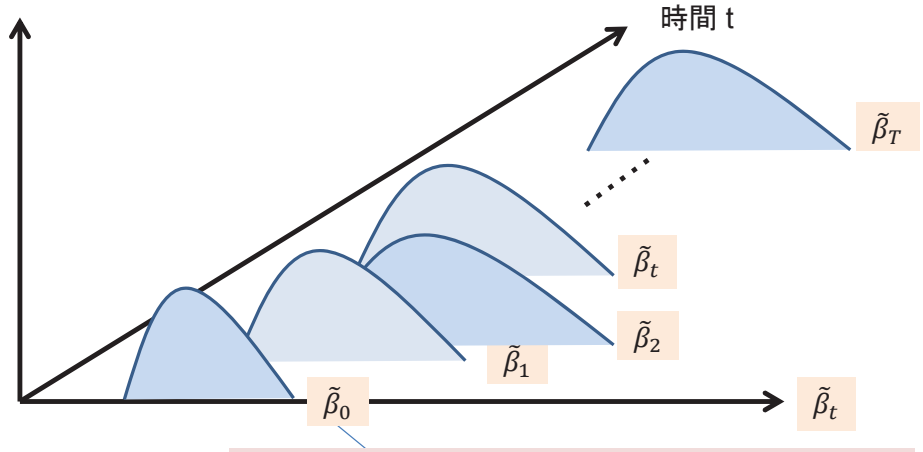
$$\tilde{z}_t = a\tilde{z}_{t-1} + bx_t + c\tilde{\varepsilon}_t$$

Z：状態変数。t期、t-1でも観察できない確率変数

1. a,b,c,d,f,gは推定すべきパラメータ（回帰係数？）。制約条件を置くことができる
2. これらは時間に依存市内固定パラメータ（強い制約）
3. 誤差項vとεは互いに独立で、互いに独立は平均ゼロ、分散一定の正規分布している。
4. 誤差項vとεにgとcの係数がかかっていることに注意

線形のカルマンフィルタでは、
―状態変数β_tの確率分布を―

注：　カルマンフィルターでは、各期のβ_tは正規分布すると仮定。あるいは、
各期のβ_tは平均と分散だけで記述出来ると仮定している
ただし、定常性（分散と平均が時間を通じてい等しく）は仮定する必要がない。



時点ゼロの状態変数すら、正確な値はわからない。
その分散と平均値を初期値として与える必要がある。

事例1
株式の1要因モデル+未知の要因

$$\tilde{r}_t - r_{F,t} = \alpha + \beta\left(\tilde{r}_{M.t} - r_{F,t}\right) + \varepsilon_t$$

一要因モデル　復習

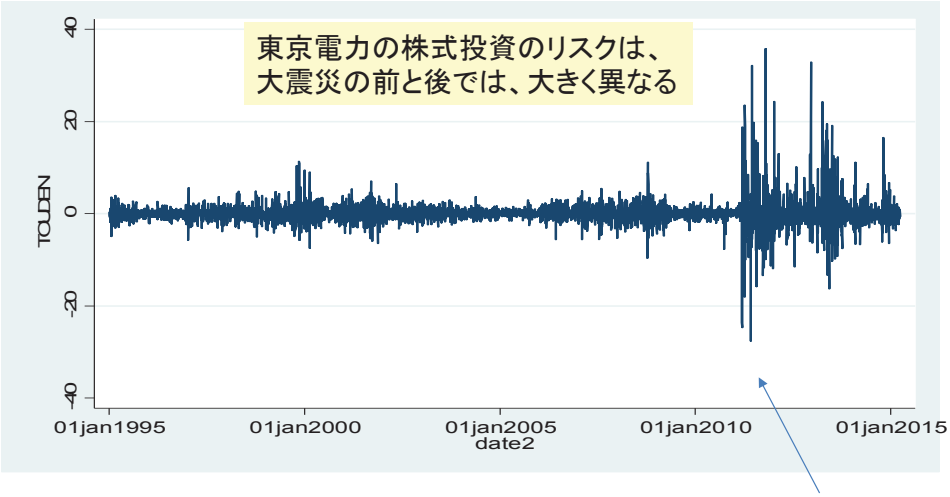
資産価格の1要因モデルとは

$$\underbrace{\tilde{r}_t - r_{F,t}}_{\text{資産の超過リターン}} = \alpha + \beta\underbrace{\left(\tilde{r}_{M.t} - r_{F,t}\right)}_{\text{市場の超過リターン}} + \varepsilon_t$$

$\tilde{r}_t$ = t期の株式投資収益率（rate of reurn on investment)
r_F = t期のリスクフリーレート
$\tilde{r}_{M.t}$ = t期の市場収益率の代理変数
ε_t = t期の誤差項

東京電力(TEPCO)の超過収益率

$$\tilde{r}_t - r_{F,t}$$



1995年1月4日から2015年3月31日までの、4997日の営業日

$$\tilde{r}_t - r_{F,t} = \alpha + \beta\left(\tilde{r}_{M.t} - r_{F,t}\right) + f_t + \varepsilon_t$$

観測方程式

$$f_t = bf_{t-1} + e_t$$

状態方程式

未知のファクターfの挙動を推定する。

1. 未知のファクターf_tは一階の自己回帰過程（AR(1)）に従うと仮定
2. もしbが1なら、の挙動を推定する。
3. 未知ファクターfはランダムウォークしている。

```
constraint define 1 [Level] L.Level = 1
constraint define 2 [touden] Level = 1
sspace (Level L.Level, state noconstant) (touden rm_rf Level), constraints(1 2)
```

推定結果

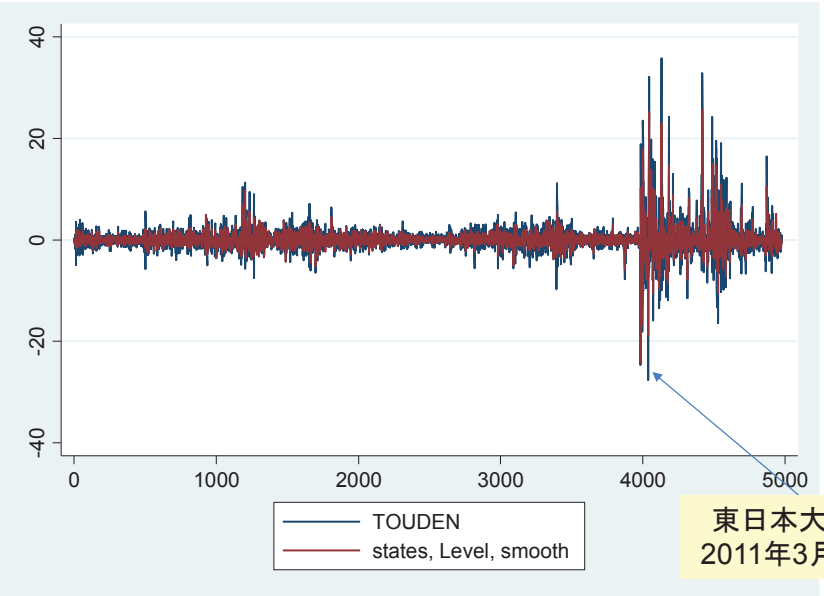
State-space model

Sample: 1 - 4977 — Number of obs = 4977
Wald chi2(3) = 417.52
Log likelihood = -11633.217 — Prob > chi2 = 0.0000

touden	Coef.	OIM Std. Err.	z	P>|z|	[95% Conf.	Interval]
Level						
Level						
L1.	.3433921	.0943849	3.64	0.000	.1584011	.5283832
touden						
Level	.4288969	49.31593	0.01	0.993	-96.22856	97.08635
rm_rf	.5237199	.0262751	19.93	0.000	.4722216	.5752182
_cons	-.0036098	.0410136	-0.09	0.930	-.0839949	.0767753
var(Level)	9.267277	2131.161	0.00	0.498	0	4186.267
var(touden)	4.418089	.605491	7.30	0.000	3.231349	5.60483

東電の超過収益率と未知ファクターの推定値

```
predict fac if e(sample), states smethod(smooth) equation(Level)
.tsline touden fac, xtitle("") legend(rows(2))
```



βの不安定性

定数と仮定

$$\tilde{r}_t - r_{F,t} = \alpha + \beta\left(\tilde{r}_{M.t} - r_{F,t}\right) + \varepsilon_t$$

1. ベータは市場全体との連動度合いを示し、
2. 東電の株のシステマチック・リスク（分散投資をしても除去できないリスクの大きさ）を示す重要なリスク尺度である

カルマンフィルターの、経済学やファイナンス理論への、重要な応用に、
回帰モデルの係数の不安定性を推定することがある。
回帰係数（この場合は、βが未知のファクターであり、確率的に時系列変動している

カルマンフィルターによる定式化

システマチックリスクの尺度が
不確実に変化する！

$$r_t - r_{F,t} = \alpha + \tilde{\beta}_t\left(r_{M.t} - r_{F,t}\right) + \varepsilon_t$$

観測方程式

$$\tilde{\beta}_t = \tilde{\beta}_{t-1} + e_t$$

状態方程式

質問：　STATAの状態空間（カルマンフィルター）　sspaceで推定ができるか？

答え：　できない。外生変数に対するパラメータは固定されている。
時間（t）あるいは外生変数の関数であってはいけない

ベータが確率的に変動する1要因モデル

単一指標モデル

$$\tilde{r}_{i,t} - r_F = \alpha_i + \beta_{i,t}\left(r_{M,t} - r_{F,t}\right) + \tilde{\varepsilon}^{r}_{i,t}$$

この時

ベータが不確実、
かつ、平均回帰する

$$\boxed{\Delta\tilde{\beta}_{i,t} = a_i\left(\bar{\beta}_i - \beta_{i,t}\right) + \tilde{\varepsilon}^{\beta}_{i,t}}$$

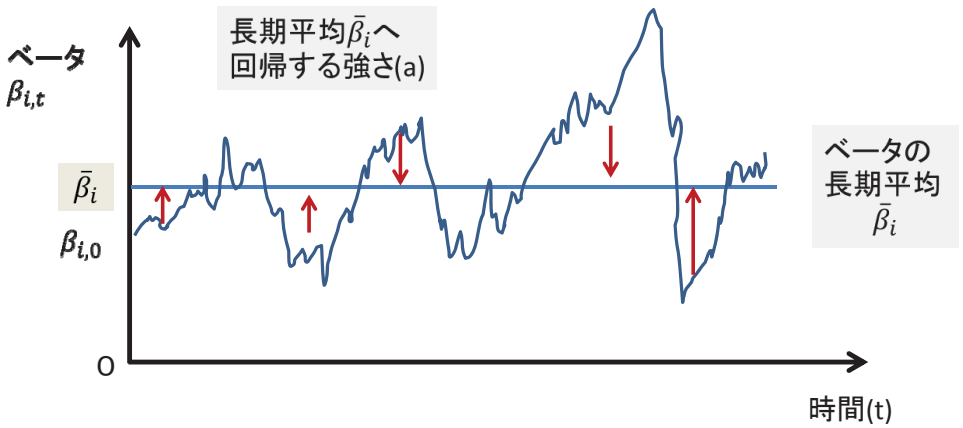
線形状態空間モデル（カルマンフィルター）表示で書き直すと

観測（信号）方程式 $\tilde{r}_{i,t} - r_{F,t} = \alpha_i + \beta_{i,t}\left(r_{M,t} - r_{F,t}\right) + \tilde{\varepsilon}^{r}_{i,t}$

状態（推移）方程式 $\tilde{\beta}_{i,t} = a_i\bar{\beta}_i + \left(1 - a_i\right)\beta_{i,t-1} + \tilde{\varepsilon}^{\beta}_{i,t}$

Vasicek, Oldrich A.." A Note on Using Cross-Sectional Information In Bayesian Estimation of Security Betas," *Journal of Finance*, 1973, 28(5), 1233-1239.

不確実かつ平均回帰をするβitを推定

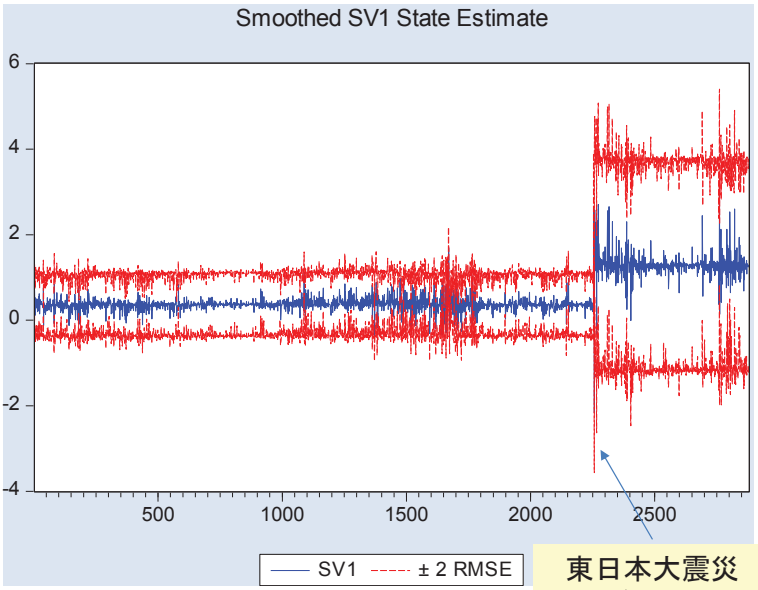


$$\Delta\tilde{\beta}_{i,t} = a_i\left(\bar{\beta}_i - \beta_{i,t}\right) + \tilde{\varepsilon}^{\beta}_{i,t} \;\Leftrightarrow\; \tilde{\beta}_{i,t} = a_i\bar{\beta}_i + \left(1 - a_i\right)\beta_{i,t-1} + \tilde{\varepsilon}^{\beta}_{i,t}$$

東京電力の確率ベータ

スムージング結果：$\beta_{t|T}$

Eviewsによる推定結果

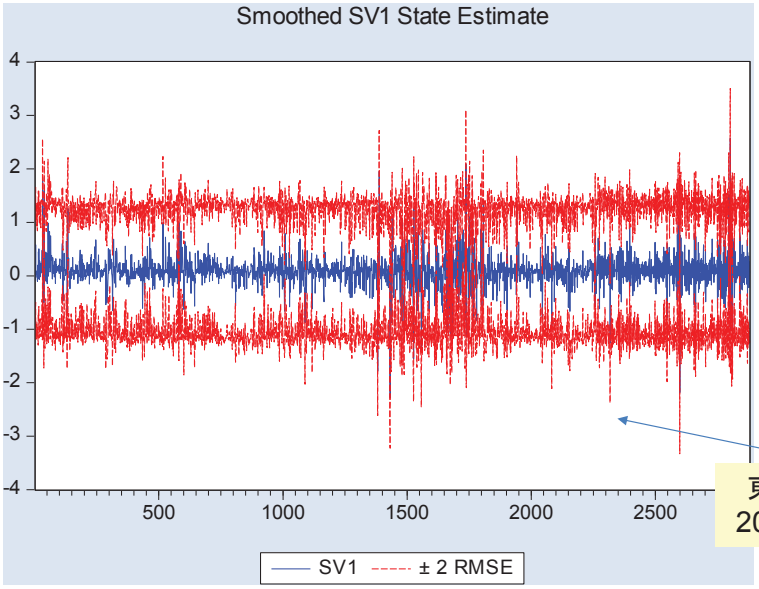


東日本大震災
2011年3月11日

沖縄電力の確率ベータ

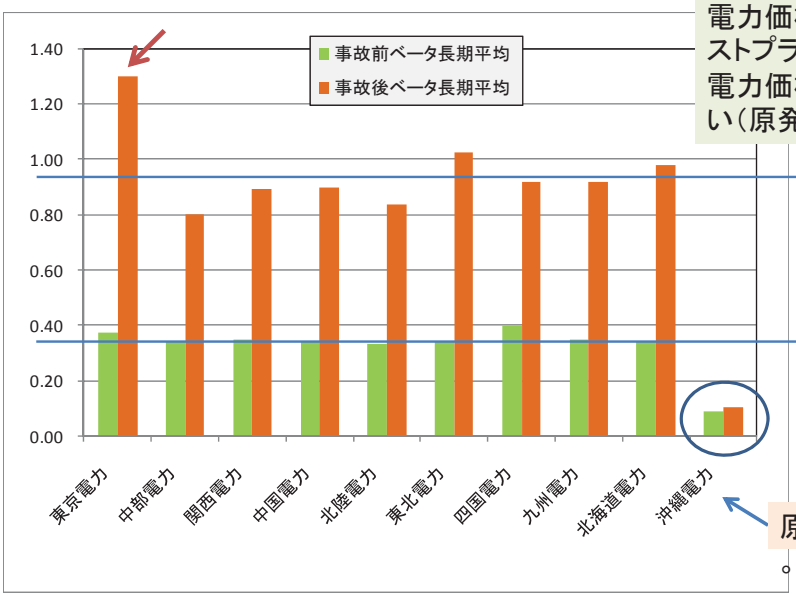
－スムージング結果：$\beta_{t|T}$

Eviewsによる推定結果



東日本大震災
2011年3月11日

電力会社βの長期平均：$\bar{\beta}_i$



電力価格は総括原価方式（コストプラス法）で決定。
電力価格の価格弾力性は低い（原発事故前は？！）

極めて低い
長期β値　$\bar{\beta}_i$

原発は保有していない

カルマンフィルターによる定式化：1

リスクを最小にするヘッジ比率は、回帰式の係数として求めることができた

$$\left(\tilde{P}_{t+1}-P_t\right)=\alpha+h\left(\tilde{F}_{t+1}-F_t\right)+e_t$$

この場合ヘッジ比率hは時間に関わらず一定で。これに対して

ヘッジ比率を時間共に変わり得る、時変パラメータh_tとして推定する、

$$\left(\tilde{P}_{t+1}-P_t\right)=\alpha+h_t\left(\tilde{F}_{t+1}-F_t\right)+e_t$$

時間とともに変わるヘッジ比率を推定するためには、
カルマンフィルターを用いることができよう。

動的な先物ヘッジ比率の推定
ーカルマンフィルターによる定式化:2 ー

$$\Delta\tilde{P}_t = \alpha + h_t\left(F_{t+1} - F_t\right) + e_t$$ 観測方程式

$$\Delta h_t = a\left(b - h_t\right) + \varepsilon_t$$ 状態方程式

ヘッジ比率は平均回帰すると考えた場合、

$$\Delta F_t \equiv F_t - F_{t-1}$$
$$\Delta P_t \equiv P_t - P_{t-1}$$

$$\Delta\tilde{P}_t = \alpha + h_t \Delta F_t + e_t$$ 観測方程式

$$h_t = ab + \left(1-a\right)h_{t-1} + \varepsilon_t$$ 状態方程式

ここで、$\Delta F_t \equiv F_t - F_{t-1}$は外生変数、先物ヘッジ比率$h_t$は状態変数
観測方程式と状態方程式の誤差項の標準誤差、パラメータα、a、bは、カルマンフィルター推定時に最尤法で推定する。

先物価格Fから未知の現物価格X=lnSを推定
Schwartz[1997]のモデル1

1. 観測方程式

STATAでは推定できない

$$y_t = Z_t X_t + d_t + \tilde{\varepsilon}_t$$

ここで

$$y_t \equiv \ln F\left(t,T\right), \quad Z_t \equiv e^{-a(T-t)},$$

時間の関数

$$d_t \equiv b^*\left(1 - e^{-a(T-t)}\right) + \frac{\sigma^2}{4a}\left(1 - e^{-2a(T-t)}\right), \quad \tilde{\varepsilon}_t \sim N\left(0, \sigma_\varepsilon^2 \Delta t\right)$$

2. 状態方程式

$$\tilde{X}_t = Q X_{t-1} + c + \tilde{e}_t$$

パラメータに関して非線形

ここで

$$c_t \equiv b^* a \Delta t, \quad Q \equiv \left(1-a\right)\Delta t, \quad \tilde{e} \sim N\left(0, \sigma^2 \Delta t\right)$$

インフレ期待の推定

物価上昇率π_tは平均回帰する。

$$\Delta\tilde{\pi}_t = a\left(\mu - \pi_{t-1}\right) + \tilde{\varepsilon}_t \qquad \tilde{\varepsilon}_t \sim N\left(0, \sigma_\varepsilon^2\right)$$

μはインフレ期待（インフレ率の長期平均）
aは期待に「回帰」する強さ。もしa=1なら即座に（この場合1月で）π_tはμに回帰する。

もしμが一定ならパラメータμとaはOLSで推定可能

カルマンフィルターによる定式化

$$\Delta\tilde{\pi}_t = a\left(\tilde{\mu}_t - \pi_{t-1}\right) + \tilde{\varepsilon}_t$$

観測方程式(πtのみが観察できる)

$$\tilde{\mu}_t = c + d\,\mu_{t-1} + e_t$$

状態方程式(インフレ期待:μtの確率的な振舞い)

```
*=====================================================================
*   inflation expectation mu Stata
*=====================================================================

use inflation.dta, clear
generate t = d(1980m1) + _n-1
format t %tm
tsset t, m

constraint define 1 [ D.cpif0161b ]mu = [ D.cpif0161b ]cpif0161b

sspace (mu L.mu, state noconstant ) (D.cpif0161b  mu cpif0161b, noconstant) in 253/422 , constraints(1)
iterate(300) difficult technique(nr) nolog

predict fac if e(sample) in 253/422, states smethod(smooth) equation(mu)
. tsline cpif0161b fac in 253/422, xtitle("") legend(rows(2))
```

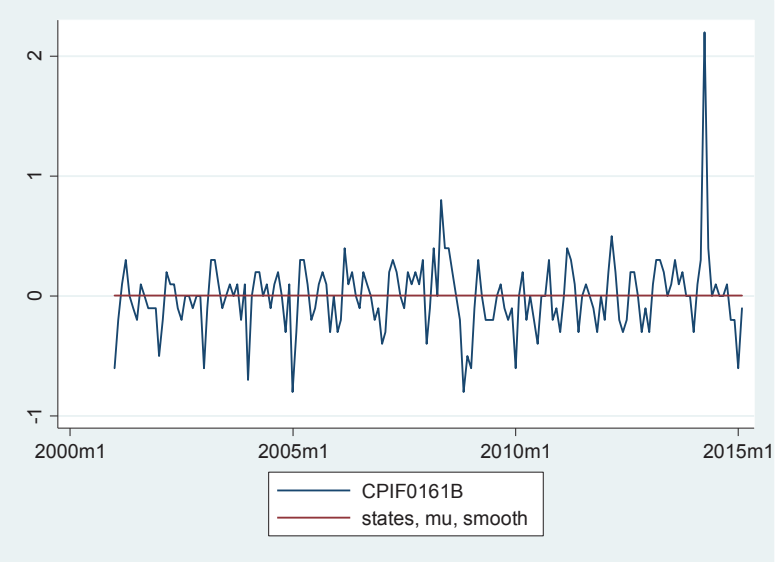

推定結果

```
State-space model

Sample: 2001m1 - 2015m2                         Number of obs   =       170
                                                Wald chi2(2)    =     89.93
Log likelihood = -29.956949                     Prob > chi2     =    0.0000
 ( 1)  [D.cpif0161b]mu - [D.cpif0161b]cpif0161b = 0
```

cpif0161b	Coef.	OIM Std. Err.	z	P>|z|	[95% Conf.	Interval]
mu						
mu						
L1.	.9927763	41.9052	0.02	0.981	-81.13992	83.12547
_cons	.0000278	.1615065	0.00	1.000	-.316519	.3165747
D.cpif0161b						
mu	.6916874	.0729376	9.48	0.000	.5487324	.8346425
cpif0161b	.6916874	.0729376	9.48	0.000	.5487324	.8346425
var(mu)	2.83e-16	.	.	.	.	.
var(D.cpif0161b)	.0832895	.0090341	9.22	0.000	.065583	.100996

異次元緩和は成功していない

インフレ期待μ_tの推定値



インフレ期待はほとんど変化していない

事前(Rx-ante)の実質金利の推定
（Ex ante real interest rate）

Hamilton, J.D: Time Series Analysis, 1994. Princeton University Press pp.376

Fisher方程式とその含意

Fisher, Irving (1977) [1930]. *The Theory of interest*. Philadelphia: Porcupine Press。

$$(1+i)=(1+r)(1+E_t\tilde{\pi}_{t+1})$$
$$1+i=1+r+E_t\tilde{\pi}_{t+1}+r\times E_t\tilde{\pi}_{t+1}\approx 1+r+E_t\tilde{\pi}_{t+1}$$
$$i\approx r+E_t\tilde{\pi}_{t+1}$$

名目金利(i)＝実質金利(r)＋期待インフレ率($E_t\pi_{t+1}$)

「事後的」な実質金利(r^e) = 名目金利(i) ー 実現した物価上昇率(π)

事前の実質金利（r 観察できない）

$$\pi_t=\frac{P_{t+1}}{P_t}$$

P_tは物価水準

「事後的」な実質金利(r^e) = 名目金利(i) ー 実現した物価上昇率(π)

$$\begin{aligned} r_t &= i_t-\pi_t \\ &= i_t-E_t\tilde{\pi}_{t+1}+E_t\tilde{\pi}_{t+1}-\pi_t \\ &= f_t+\varepsilon_t \end{aligned}$$

ここで $f_t=i_t-E_t\tilde{\pi}_{t+1}$

観察することができない、「事前」の実質金利

$$\varepsilon_t\equiv E_t\tilde{\pi}_{t+1}-\pi_t\sim N\left(0,\sigma_\varepsilon^2\right)$$

インフレ期待（予測）誤差、は平均はゼロで、観察できない分散をもって分布していると仮定

カルマンフィルターによる定式化

観測方程式

状態方程式

$$r_t = \alpha + f_t + \varepsilon_t$$

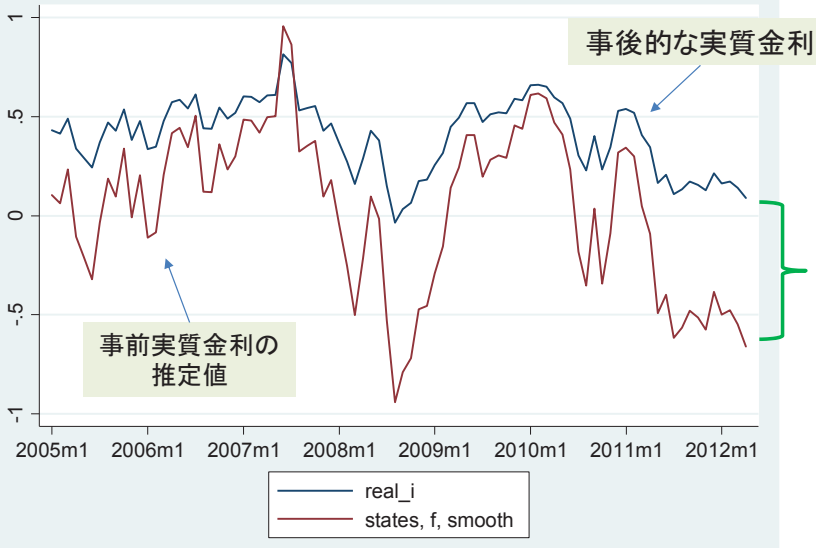
$$f_t = \beta f_{t-1} + e_t$$

r_t: 観察できる事後的な実質金利
（名目金利から観察可能な事後的な物価上昇率を差し引く）I

観察できない事前の実質金利;f

事前の実質金利の推定

2005年1月から2012年4月



カルマンフィルターによる推定結果

State-space model

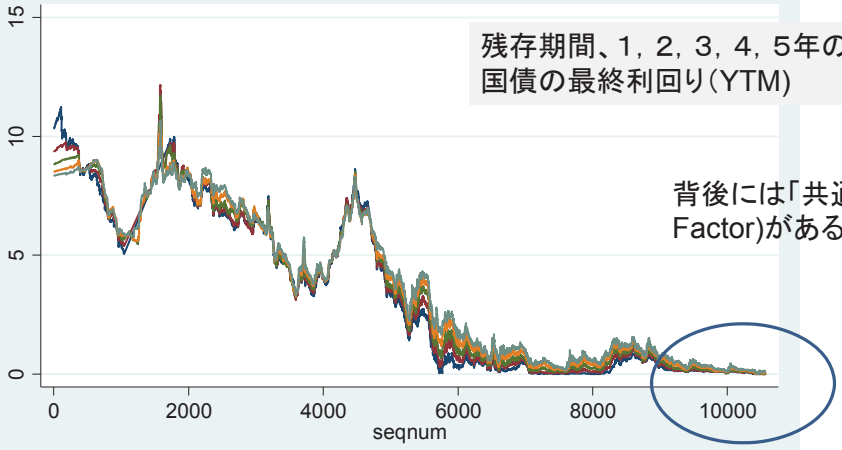
Sample: 2005m1 - 2012m4 — Number of obs = 88
Wald chi2(2) = 263.91
Log likelihood = 84.047142 — Prob > chi2 = 0.0000

real_i	Coef.	OIM Std. Err.	z	P>|z|	[95% Conf.	Interval]
f						
f L1.	.8643745	.0532104	16.24	0.000	.760084	.968665
real_i						
f	.4471855	2.594091	0.17	0.863	-4.63714	5.531511
_cons	.3859764	.0682333	5.66	0.000	.2522416	.5197113
var(f)	.042672	.4950604	0.09	0.466	0	1.012973
var(real_i)	9.96e-14	.	.	.	.	.

動的ファクターモデル
Dynamic Factor Model of term structure of interest rate

資産価格の背後にある
共通要因を見つけ出す。

Dynamic Factor Model of term structure of interest rate



残存期間、1，2，3，4，5年の国債の最終利回り(YTM)

背後には「共通」の要因(Factor)がある。

直近のデータで推定する。

Dynamic Factor Model

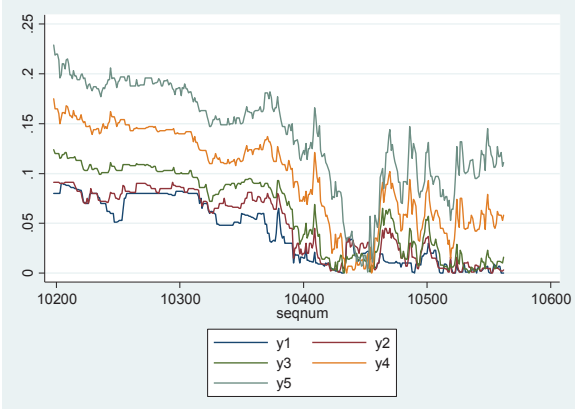
5つの観測方程式

$$y_{1.t} = a_1 + b_1 \tilde{f}_t + \varepsilon_{1,t}$$
$$y_{2,t} = a_2 + b_2 \tilde{f}_t + \varepsilon_{2,t}$$
$$y_{3,t} = a_3 + b_3 \tilde{f}_t + \varepsilon_{3,t}$$
$$y_{4,t} = a_4 + b_4 \tilde{f}_t + \varepsilon_{4,t}$$
$$y_{5,t} = a_5 + b_5 \tilde{f}_t + \varepsilon_{5,t}$$

$$\tilde{f}_t = \alpha + \beta \tilde{f}_{t-1} + e_t$$

一つの状態方程式

1年債から5年債の最終利回りを。その背後にかくれている未知の一つのファクターで説明する。
18個のパラメータを推定する。



```
*=============================================================
*  dynamic factor model of term structure of interest rate
*=============================================================
use JGBYTMDaily.dta, clear
tsset se

sspace (f L.f, state) (y1 f) (y2 f) (y3 f) (y4 f) (y5 f) in 10198/10562, iterate(300)
ltolerance( 0.001) difficult technique(nr)

predict fac if e(sample), states smethod(smooth) equation(f)
. tsline y1 y2 y3 y4 y5  fac, xtitle("") legend(rows(2))
```


推定結果

State-space model

Sample: 10199 - 10562 — Number of obs = 364; Wald chi2(5) = 11499.53; Log likelihood = 4905.1564; Prob > chi2 = 0.0000

	Coef.	OIM Std. Err.	z	P>|z|	[95% Conf.	Interval]
f						
fac						
L1.	.1834083	.	.	.	.	.
_cons	.1593938	.	.	.	.	.
y1						
fac	.4337846	.0150406	28.84	0.000	.4043055	.4632637
_cons	.0489079	.0009499	51.49	0.000	.0470462	.0507696
y2						
fac	.4528927	.015949	28.40	0.000	.4216331	.4841522
_cons	.0574193	.0010072	57.01	0.000	.0554451	.0593934
y3						
fac	.5947267	.0168015	35.40	0.000	.5617964	.627657
_cons	.0742647	.0010611	69.99	0.000	.072185	.0763443
y4						
fac	.7415323	.0133016	55.75	0.000	.7154616	.767603
_cons	.1116079	.00084	132.86	0.000	.1099614	.1132543
y5						
fac	.8267522	.0111473	74.17	0.000	.8049039	.8486004
_cons	.1549362	.000704	220.08	0.000	.1535564	.156316
var(f)	4.73e-06	.	.	.	.	.
var(y1)	.0002955	.0000219	13.49	0.000	.0002525	.0003384
var(y2)	.0003322	.0000246	13.49	0.000	.000284	.0003805
var(y3)	.0003687	.0000273	13.49	0.000	.0003151	.0004223
var(y4)	.0002311	.0000171	13.49	0.000	.0001975	.0002647
var(y5)	.0001623	.000012	13.49	0.000	.0001387	.0001859

Note: Tests of variances against zero are one sided, and the two-sided confidence intervals are truncated at zero.

確率ボラティリティモデル

ボラティリティはボラタイルだ！
(変動性は変動する)

観測方程式

Step1： 対数logリターンを定義

$$\tilde{y}_t \equiv \Delta \ln x_t = \ln x_{t+1} - \ln x_t$$

あるいは、通常の収益率(配当込みが望ましい)

Step2: 観測方程式の設定
この対数リターン$\tilde{y}_t$のボラティリティが確率的に変化する様子を捉えたい
対数リターン の不確実な振る舞いは、

$$\tilde{y}_t = \sigma \tilde{\varepsilon}_t \exp\left\{\frac{\tilde{h}_t}{2}\right\}$$

確率的に変動するボラティリティ部分（tがついていることに注意）

なぜ指数関数を用いるのか？ なぜ2で割っているのか？

$$\tilde{\varepsilon}_t \sim N(0,1)$$

ボラティリティの固定効果部分（通常の標準偏差にあたるもの）

観測方程式の線形化

$$\tilde{y}_t = \sigma\tilde{\varepsilon}_t \exp\left\{\frac{\tilde{h}_t}{2}\right\}$$

Step1: 両辺を2乗する

$$\tilde{y}_t^2 = \sigma^2\tilde{\varepsilon}_t^2\left(\exp\left\{\frac{\tilde{h}_t}{2}\right\}\right)^2 = \sigma^2\tilde{\varepsilon}_t^2\exp\{\tilde{h}_t\}$$

Step2: 両辺の対数をとる

$$\ln\tilde{y}_t^2 = \ln\sigma^2 + \ln\tilde{\varepsilon}_t^2 + \tilde{h}_t$$

注意:

$\ln\tilde{\varepsilon}_t^2$は平均が-1.2704，分散が$\pi^2/2$の対数χ2乗分布をする(εは平均ゼロ、分散$\sigma^2$の正規分布をするので、その2乗は自由度1のχ2乗分布に従う)ので、この式の誤差項の平均をゼロにするために、右辺で、1.2704を加減すると、

カルマンフィルターによる推定

観測方程式

$$\ln\tilde{y}_t^2 = \left(\ln\sigma^2 - 1.2703\right) + \tilde{h}_t + \xi_t$$

状態方程式
確率ボラティリティ

$$\tilde{h}_t = \alpha + \beta\tilde{h}_{t-1} + \tilde{e}_t \quad \tilde{e}_t \sim N\left(0,\sigma_h^2\right)$$

ただし

$$\xi_t \equiv \ln\tilde{\varepsilon}_t^2 - (-1.2703)$$

本来は自由度1のカイ二乗分布をする

正規分布で近似？

$$\xi_t \sim N\left(0,\frac{\pi^2}{2}\right)$$

分散=4.9532(既知！)

注意:最尤法によるパラメータ推定は、α、βとσ_hの2つ.
もし、状態方程式に定数項を考えれば3つ＋

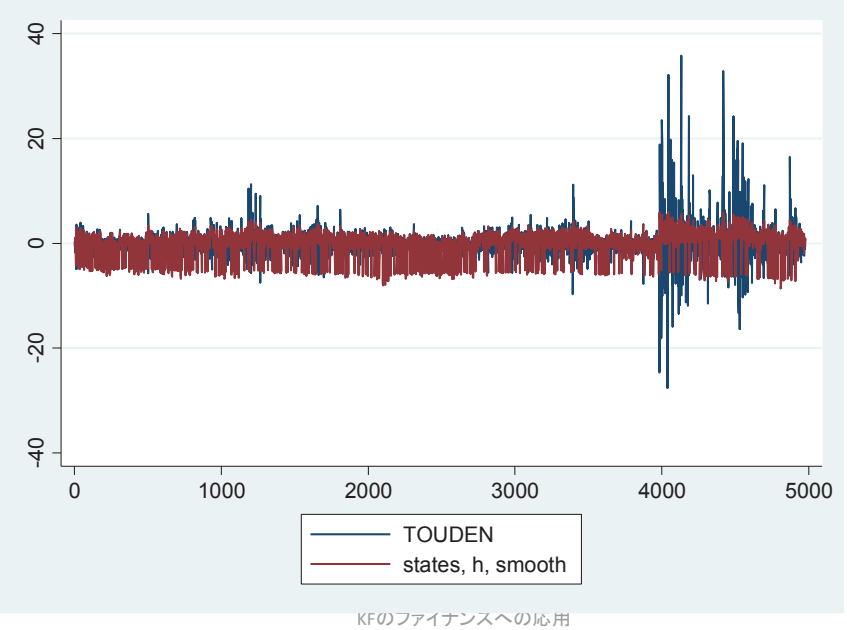
STATAによる推定

```
* ==================================
* random volatility model
* ==================================
use Price_elct_gas.dta
tsset seqnum
generate ztouden = ln(touden^2)

constraint drop _all
constraint define 2 [ ztouden ]h = 1
constraint define 3 [ var(ztouden) ]_cons = 4.9532

sspace ( h L.h, state) (ztouden h), constraints(2 3) iterate(300) ltolerance(
0.01) difficult technique(nr) nolog

predict fac if e(sample), states smethod(smooth) equation(h)
tsline touden fac, xtitle("") legend(rows(2))
```


推定結果

```
State-space model

Sample: 1 - 4977                                    Number of obs   =       4977
                                                    Wald chi2(1)    =     163.49
Log likelihood = -13420.513                         Prob > chi2     =     0.0000
 ( 1)  [ztouden]h = 1
 ( 2)  [var(ztouden)]_cons = 4.9532
------------------------------------------------------------------------------
             |                 OIM
     ztouden |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
h            |
           h |
         L1. |    .297613   .0232757    12.79   0.000     .2519935    .3432324
             |
       _cons |   .1276827   .0450152     2.84   0.005     .0394545    .2159108
-------------+----------------------------------------------------------------
ztouden      |
           h |          1  (constrained)
       _cons |  -1.371315          .        .       .            .           .
-------------+----------------------------------------------------------------
       var(h) |   7.553575   .2680313    28.18   0.000     7.028243    8.078907
 var(ztouden) |     4.9532  (constrained)
------------------------------------------------------------------------------
```

推定結果
東電の確率ボラティリティ：σ_t



初期値問題

Fitting state space model is notoriously difficult.

STATA sspce p.14

最も難しい問題
ー固定パラメータの初期値ー

パラメータの初期値は

1. STATAがうまく与える（事が出来る場合もある）
2. うまく行かない場合に、何が起きるか？
    1. 推定結果がおかしい。
        1. パラメータが有意でない。
        2. パラメータの有意性が極めて高い（z値が100とか1,000のオーダー！
        3. 観測方程式、状態方程式の誤差項がゼロになる！
    2. 尤度関数が収束しない

どうしたらよいか。 状態空間モデリングの最も困難な問題

1. パラメータの初期値を、理論や問題の背景をよく考えた上で与える（重要）、その上で、
2. 真の初期値が漠然としかわからいのであるから
    1. 初期値の分布を考え（どのような分布を想定するのか？）
    2. その分布から乱数を発生させ、何回もsspaceによる推定を繰り返し、
    3. 最大尤度を与える固定パラメータと状態変数の推定値をえる。

初期値問題：続き

STATAは、こうした方法を用いて初期値を推定している。しかし、それでも収束しない場合がある。次のようなエラーメッセージに注意

random initial values failed; use the from() option to supply initial estimates

どのようにして、from() を用いて、研究者が初期値を与えるのか？

```
matrix b0 = e(b)    /// 推定結果を取り出し、それをb0と言う行列に格納
matrix list b0      /// b0の中身を確認する。パラメータの順番と中身を確かめる

matrix b0[1,1] = 0.40       /// ユーザが考える初期値を与える。
matrix b0[1,2] = 1.0
matrix b0[1,3] = 0.84
matrix b0[1,4] = 0.001
matrix b0[1,5] = 0.09

constraint 1 [real_i] f = 1.0
constraint 2 [f] L.f = 1
sspace (f L.f, state noconstant) (real_i f) in 1/88 , covstate( diagonal ) covob( diagonal )
constraints(1) from(b0) iterate(300) nolog
```

初期値問題：続き

```
matrix b0[1,1] = uniform()*0.6+0.2
```

1番目のパラメータの価が0.2から0.6の間のあるという革新があれば、

乱数を、数百から数千回あたえて、sspace計算をくりかえし、収束しかつ最大尤度の結果をあたえる。

その他、最大化にかんするオプションを良く理解することが大事

STATAで利用できるその他の状態空間モデリング

1. dfactor： dynamic Factor models
2. ucm: unobsereved components model

Sspaceの特別な場合で、sspaceでも計算することは可能であるが、ucmを用いたほうが便利

その他、sspaceモジュールをつかえば、
ベクトル回帰やARMAなどの計算ができるが、それらは専門の計算で行うほうがよい。

参考文献

1. 谷崎久志.『状態空間モデルの経済学への応用—可変パラメータ・モデルによる日米マクロ計量モデルの推定』. 日本評論社, 1993.
2. コマンダー、クープマンズ、『状態空間時系列分析入門』、2008年、CAP出版、数学を余り使わない、経済学への応用に関する入門書、
3. ハーベー、A.C.『時系列モデル入門』、1985年、東京大学出版界（絶版）、特に第4章「状態空間モデルとカルマンフィルター」、東京大学出版会
4. ダービン、クープマンズ、『状態空有巻モデリングによる時系列分析入門』、和合・松田訳、2004年、CAP出版、絶版
5. G.ペトリス、S.ペトローネ、P.カンパニョーリ、和合肇 監訳；萩原淳一郎訳、『 Rによるベイジアン動的線形モデル (統計ライブラリー) 』、2013、朝倉書店 （原書：**Dynamic Linear Models with R**）
6. 足立修一・丸田一郎 『カルマンフィルターの基礎、東京電機大学出版局。2012、カルマンフィルターの導出について、詳しく説明したよい入門書

平成 26 年度厚生労働科学研究費補助金（地域医療基盤開発推進研究事業）
歯科疾患の疾病構造の変化を踏まえた歯科口腔保健の実態把握のための評価項目と必要客体数に関する研究（H26－医療－一般－007）研究代表者：三浦宏子

2011 年歯科疾患実態調査、国民健康・栄養調査、国民生活基礎調査のリンケージデータを用いた解析結果

研究分担者　安藤雄一（国立保健医療科学院・生涯健康研究部）

研究要旨

【目的】本研究では、歯科疾患実態調査の参加率を国民健康・栄養調査の参加情報別に検討することを主目的とした。また、国民健康・栄養調査の歯科関連項目について歯科疾患実態調査の参加有無別に差を検討することと、歯科医院の通院状況別に歯の保有状況を比較することを副次的目的とした。

【方法】2011年の歯科疾患実態調査（以下「歯調」）、国民健康・栄養調査（以下「栄調」、国民生活基礎調査（以下「基調」）について厚労省の担当窓口に目的外利用を申請し利用許可を得たの個票データを用い、IDによるデータリンケージを行い、各調査間で性・年齢が一致しないデータを除いた13,351件のデータを用いて解析を行った。

【結果】「歯調」の参加率について、「基調」に対する参加率を「栄調」を構成する各調査と比較したところ、参加率は血液検査と酷似していた。さらに「栄調」の参加状況別に「歯調」の参加率をみたところ、血液検査を受けた人では100%近くが「歯調」に参加していたのに対し、血液検査を受けていない人では数%と著しい差を示した。「歯調」参加有無別に「栄調」および「基調」の歯科関連調査項目の基礎統計量を比較したところ、歯の保有状況、歯科健診・口腔ケアの受診頻度において高齢者層で有意差が認められた。「基調」で調査された歯科通院の有無別に歯の保有状況を比較したところ、高齢者層において通院者で良好な傾向が認められた。

【考察】血液検査の参加有無による「歯調」参加率の顕著な差は、受診者が「栄調」の身体状況調査を「歯調」の前に受けるため、血液検査に参加しない人たちが「歯調」受診の働きかけを受ける前に帰ってしまうためと考えられた。「歯調」の参加率を上げるためには、血液検査の不参加者に対して「歯調」参加の声がけを必ず行うことや、血液検査の参加率を上げる取り組みなどが必要である。「歯調」参加有無別に認められた差異より、高齢者の歯の保有状況に関する国民真の代表値は「歯調」で報告された値よりも少し低めの値とみなすのが妥当と考えられた。

キーワード：　歯科疾患実態調査、国民健康・栄養調査、国民生活基礎調査、身体状況調査、血液検査、データリンケージ

表1．データリンケージを行った調査の一覧

調査年	調査名	調査票名	レコード件数
2011（平成23）年	国民生活基礎調査	世帯票	46,099（世帯数）注1
2011（平成23）年	国民健康・栄養調査	栄養摂取量票	8,761（人数）注2
		食品群別摂取量票	
		身体状況・生活習慣票	
2011（平成23）年	歯科疾患実態調査	―	4,253（人数）

注1：人数＝118,955人

注2：国民健康・栄養調査における各調査票ごとの調査人数は実際のところ異なっている。

http://www.mhlw.go.jp/bunya/kenkou/eiyou/dl/h23-houkoku-02.pdf

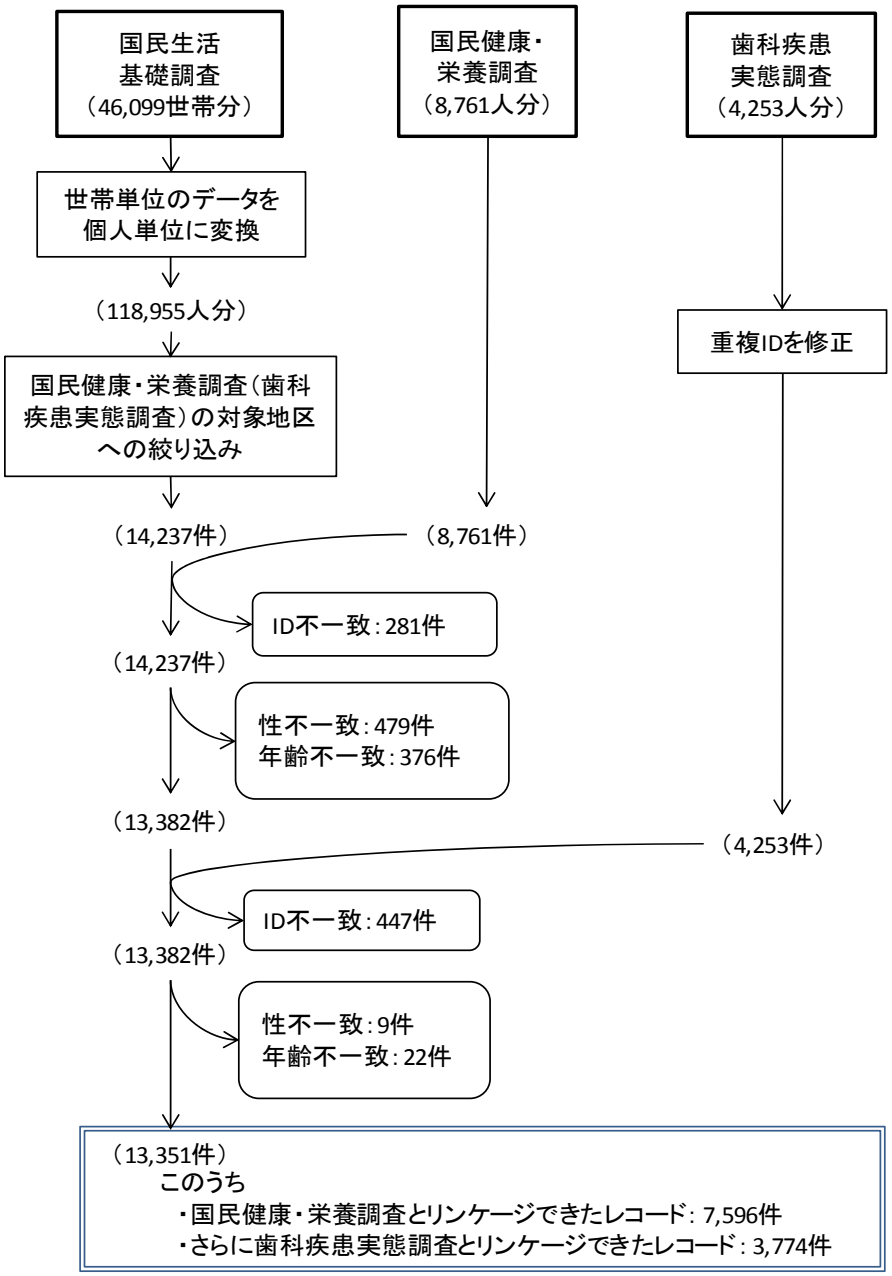


図1．データリンケージ結果

Determinants and Impacts of Health Insurance Nonutilization

Hisaki Kono

Graduate School of Economics
Kyoto University

Stata Conference, August 2015

Motivation and Research Question

- Expansion of health insurance programs to cover the poor and disadvantaged group .
- Mixed results on the impact of health insurance on service utilization and OOP payment:
  - Service utilization:
    - **(+)** Waters (1999, Ecuador), Trujillo et al. (2005, Colombia), Yip and Berman (2005, Egypt), Wagstaff et al. (2009, China)
    - **insig.** Thornton et al. (2010, Nicaragua), Waters (1999, Ecuador), Wagstaff (2010, Vietnam)
  - OOP payment:
    - **(-)** Thornton et al. (2010, Nicaragua), Yip and Berman (2005, Egypt), Wagstaff (2010, Vietnam)
    - **insig.** Wagstaff et al. (2009, China)
- The impact on health outcomes will depend both on the demand and supply side factors, making it context-dependent.
  - demand side: price elasticity
  - supply side: accessibility (distance, waiting time, staff presence), quality

Motivation and Research Question

- An issue missing in the literature: discrimination against the insurance holders and nonutilization of insurance.
- Insurance is helpful only when it is utilized.
- A story of a patient visiting a central cancer hospital in Hanoi (Ha, 2011):
  - If he would use it, it would take two days for him to be fully checked.
  - When he paid by himself, it finished in a morning.
- Vietnam Household Living Standard Survey 2006
  - insurance utilization (free health insurance card): 0.80 for inpatient treatment, 0.55 for outpatient treatment
  - Three most cited reasons for the nonutilization
    - cumbersome procedures (21.9%)
    - lower quality of health care service when using the insurance (11.9%)
    - little prospect of receiving reimbursement (9.0%)
  - But no info when and where they chose not to utilize insurance.
- We have detailed info on insurance utilization at each facility visit.
- Need to correct endogeneity of the facility choice (and sample selection).

Overview of the Result

- Insurance utilization tends to be lower at the public central hospital.
  - Reflecting low incentives of the public, especially crowded, healthcare facility to increase the number of patients?
  - Limitation: Lack of data to infer the problem in each facilities.
- Less prospect of utilizing health insurance at the central hospital discourages people from visiting there.
- The average effect of insurance nonutilization on the healthcare expenses per visit is around 1 million VND, which is 1.5 times larger than the average healthcare expenses per visit.

Health Insurance System in Vietnam

- Compulsory health insurance (CHI) and voluntary HI.
  - Health Care Fund for the Poor (HCFP) in 2002: free HI for the poor, the disadvantaged, and ethnic minorities..
  - HCFP integrated into the CHI system in 2005.
  - As of 2010, 69.8% of the target population enrolled the CHI, and 58.5% of the total population enrolled the CHI or VHI.
  - For the poor and the priority group, >95% enrollment as it is free.
- Quite generous insurance package.
  - Covers medical consultations; diagnosis & treatment; x-ray & lab tests, functional exam, ...; drugs approved by MOH; surgery; antenatal exam & delivery; transport costs for the poor referred.
  - Covers 100% of exam & treatment costs for meritorious people & all treatment at CHCs; Covers 95% of exam & treatment costs for beneficiaries of social allowances, the poor & ethnic minority. Others are reimbursed 80% of exam & treatment costs.
  - Exclusion: early detection of pregnancy, medical check-ups, family planning services & infertility treatment, prosthesis & aesthetic surgery...

Discrimination against insurance holders

- (Anecdotes) insurance utilization often results in longer waiting time or lower quality of healthcare service.
- This will be likely to happen in crowded public hospitals.
  - Private facilities concern attracting many patients to raise profit, and have incentives to ensure insurance utilization to attract more patients.
  - In public facilities, they may not care for attracting more patients as it will not increase their income. Rather, more patients implies more work, and insurance utilization will cause extra admin works.
  - In crowded facility already facing a heavy workload may have more incentives to discriminate the insurees.
  - Public facilities with the best quality may be in stronger position to ask bribe given their monopolistic technology.
- In Vietnam, public facilities are the main service providers, making the insurance nonutilization problem important.
  - Hopefully have data on crowdedness (e.g. patients/staff ratio) at each facility.

Data

- 2,673 adults aged $\geq$ 18 in 60 communes, by the multi-stage stratified random sampling.
- Health insurance is quite familiar. 68.7% have some health insurance, 16.2% do not have but did have.

Table: Summary statistics of demographic variables

	(1) Mean	
gender: female	0.576	(0.016)
Age of the respondent	43.201	(0.553)
Respondent is an ethnic minority	0.236	(0.030)
Years in school	8.548	(0.322)
log(1+household income per capita)	13.729	(0.065)
ln(work hour)	3.020	(0.082)
can read Kinh	0.963	(0.006)
chronic disease	0.320	(0.015)
last illness: severity (1-4)	2.329	(0.021)
have HI	0.687	(0.027)
does not have but did have HI	0.162	(0.014)
Observations	2426	

Standard errors in parentheses

Insurance holdings

- The majority of health insurance is health insurance for the poor.
- In the analysis, we exclude those who hold voluntary government health insurance, the second largest group.

(B) Insurance holdings across insurance types

	(1) Mean	
student HI	0.042	(0.007)
Poor household HI	0.326	(0.043)
priority group HI	0.131	(0.013)
obliged gov. HI	0.176	(0.030)
voluntary gov. HI	0.287	(0.031)
private HI	0.023	(0.006)
other HI	0.022	(0.005)
Observations	1929	

Standard errors in parentheses

Pattern of Health Insurance Utilization

- Use the healthcare facility visit level data for the last disease in the last 12 months.
- Information on the type of the healthcare facility visited, insurance utilization, and the amount of healthcare expenses for each visit is available.
- Use observations which have health insurance, experienced illness in the last 12 months, and visited any healthcare facilities.
- Insurance utilization rate is slightly lower for the ethnic minority and female.

Table: Insurance Utilization

	(1) Full sample			(1) Full sample	
last illness: use HI			last illness: use HI		
Kinh	0.782	(0.028)	male	0.803	(0.029)
Ethnic_Minority	0.744	(0.034)	female	0.753	(0.028)
Observations	1177		Observations	1177	
Standard errors in parentheses			Standard errors in parentheses		

Pattern of Health Insurance Utilization across Facilities

- Private facilities: often no contract with the insurance agency, especially small clinics.
  - Excluded in the analysis.
- Highest insurance utilization rates in the district hospitals/policlinics.
  - Set as the reference category.
- The central hospital has a lower utilization rates.

Table: Insurance Utilization across healthcare facilities

	(1) Treatment or examinations only	
last illness: use HI		
Central	0.789	(0.047)
Province	0.949	(0.030)
District	0.960	(0.016)
Commune	0.886	(0.022)
Private_hospital	0.712	(0.140)
Private_clinic	0.034	(0.028)
Observations	1025	

Econometric Issues

- Sample selection
  - Insurance utilization decision is observed only when the individual (1) had an insurance and (2) chose to visit public health facilities.
  - Free insurance might have been provided to the poor, who were unfamiliar with insurance and hence less likely to utilize insurance.
  - Individual may not go to public health facilities because they expect that they will not utilize the insurance.
- Endogeneity of facility choice
  - Healthcare facility visit choice will depend on Pr(utilize insurance | facility).
  - People who visited the central hospital might have higher prospect of utilizing insurance.
  - The insurance utilization rate at the central hospital itself does not tell us how difficult to utilize insurance there.

Econometric Specification

Insurance utilization decision:

$$y_i = 1[\mathbf{x}_i\beta + \mathbf{w}_i\gamma + u_i > 0], \tag{1}$$

where $1[\cdot]$ is the indicator function.

- $y_i = 1$ if observation i utilized insurance and zero otherwise.
- $\mathbf{x}_i$ are indicators for the healthcare facility visited.
  - Reference category is the district hospital/policlinic (highest insurance utilization)
- $\mathbf{w}_i$ are exogenous control variables.
- $s_{1i} = 1$ if i held the health insurance card and zero otherwise.
- $s_{2i} = 1$ if i visited any public facilities and zero otherwise.
- We observe $(\mathbf{x}_i, \mathbf{w}_i, y_i)$ only when $s_i = s_{1i} \cdot s_{2i}$ is one.

Correction for the sample selection problem

- Sample selection problem: $E[u_i|s_i = 1] \neq 0$.
- Use the IPW
  - Estimate $\hat{p}_i = \Pr(s_i = 1|\mathbf{q}_i)$, where $\mathbf{q}_i$ are variables predicting selection, and

    $\Pr(s_i = 1|\mathbf{q}_i) = \Pr(s_{1i} = 1, s_{2i} = 1|\mathbf{q}_i) = \Pr(s_{1i} = 1|\mathbf{q}_i) \Pr(s_{2i} = 1|s_{1i} = 1$
  - Weight each observation by $1/\hat{p}_i$.
- Required assumptions:

$$\Pr(s_i = 1|\mathbf{r}_i, \mathbf{q}_i) = \Pr(s_i = 1|\mathbf{q}_i) \equiv p(\mathbf{q}_i)$$
$$p(\mathbf{q}_i) > 0 \text{ for all the possible values of } \mathbf{q}_i$$

$\mathbf{q}$ should predict the sample selection sufficiently well in the sense that once $\mathbf{q}$ are controlled, the probability of observing obs i is uncorrelated with the unobservable component u_i.

Correction for the endogenous facility choice

- Endogeneity problem: $E[\mathbf{x}_i u_i] \neq 0$.
  - Individual might choose a facility because it is easier to utilize insurance there.
- Use the village average of the health facility utilization, $\mathbf{z}_i$, which would reflect the distance to each facility, as the instruments.
- Might not satisfy the exclusion restriction $E[\mathbf{z}_i' \mathbf{u}_i] = 0$, but would satisfy $E[\mathbf{z}_i' \mathbf{u}_i] \geq 0$ and $|Corr(z_{ji}, u_i)| < |Corr(x_{ji}, u_i)|$.
  - An individual who lives in a village with greater utilization of a health facility is not less likely to utilize insurance at that facility.
- Use Nevo and Rosen (2012)'s Imperfect IV (IIV) to obtain the conservative estimates on β.

Nevo and Rosen (2012)'s IIV

- Consider a linear model only consisting of a dummy for the central hospital, x_i:

$$\Pr(y_i = 1) = \alpha + x_i\beta + u_i.$$

- plim of the OLS and IV estimators with an instrument z_i is

$$\beta^{OLS} = \frac{\sigma_{xy}}{\sigma_x^2} = \beta + \frac{\sigma_{xu}}{\sigma_x^2} \quad (2)$$

$$\beta_z^{IV} = \frac{\sigma_{zy}}{\sigma_{zx}} = \beta + \frac{\sigma_{xu}}{\sigma_{zx}} \quad (3)$$

- If $\sigma_{xz} < 0 \rightarrow$ a two-sided bound: $\beta_z^{IV} \leq \beta \leq \beta^{OLS}$.
  If $\sigma_{xz} > 0 \rightarrow$ a one-sided bound: $\beta \leq \min\{\beta^{OLS}, \beta_z^{IV}\}$.
  - when $\sigma_{xu}, \sigma_{zu} \geq 0$.
  - $\beta < 0$ as we set the facility with highest insurance utilization as the reference category.
- Can narrow the band by using an IV $v_i = \sigma_x z_i - \sigma_z x_i$:
  $\beta_z^{IV} \leq \beta \leq \beta_v^{IV}$ if $\sigma_{xz} < 0$; $\beta \leq \min\{\beta_v^{IV}, \beta_z^{IV}\}$ if $\sigma_{xz} > 0$.

Procedures for additional regressors and multiple endogenous variables

1. Regress **x**, **z** and y on **w** to obtain the residuals $\tilde{\mathbf{x}}$, $\tilde{\mathbf{z}}$ and $\tilde{y}$.
   - Partial out the effect of **w**.
   - $\tilde{y}_i = \tilde{\mathbf{x}}_i\beta + u_i$ (no constant term)
2. Demean **z** to obtain $\bar{\mathbf{z}}$.
3. Define $\bar{v}_{ij} \equiv \sigma_{x_j}\bar{z}_{ij} - \sigma_{z_j}\tilde{x}_{ij}$.
4. Obtain the IV estimators (1) using $\bar{\mathbf{z}}$ as the IVs, and (2) $\bar{v}_j \equiv (\bar{v}_1, \ldots, \bar{v}_{k_x})$ as the IVs:

$$\beta_{\bar{\mathbf{z}}}^{IV} = E[\bar{\mathbf{z}}'\tilde{x}]^{-1}E[\bar{\mathbf{z}}'\tilde{y}]$$

$$\beta_{\bar{v}}^{IV} = E[\bar{v}'\tilde{y}]^{-1}E[\bar{v}'\tilde{y}]$$

An informal test for the effect of the low prospect of insurance utilization on the healthcare facility choice

- The endogeneity problem will arise mainly because the healthcare facility choice will depend on the expectation how likely they will use the insurance there.
- $\beta \leq \min\{\beta_v^{IV}, \beta_z^{IV}\}$.
- If $|\beta_j^{IV}| > |\beta_j^{OLS}|$ against the null $|\beta_j^{IV}| = |\beta_j^{OLS}|$, which can be tested by Durbin-Wu-Hausman test, it is implied that the low prospect of utilizing insurance at a given healthcare facility actually discourages people from visiting that healthcare facility.
  - Required assumption: measurement errors ignorable.

	(1) OLS	(2) Probit	(3) IV	(4) OLS	(5) IV
Central hospital	-0.143***	-0.135***	-0.278**	-0.135***	-0.277**
	(0.051)	(0.031)	(0.134)	(0.050)	(0.131)
Province/City hospital	0.008	0.022	0.188	0.026	0.172
	(0.037)	(0.054)	(0.147)	(0.039)	(0.137)
Commune health center	-0.048	-0.046	-0.151	-0.080*	-0.069
	(0.033)	(0.029)	(0.113)	(0.040)	(0.102)
poor HH HI	-0.056	-0.061	-0.017	-0.061	-0.051
	(0.055)	(0.046)	(0.056)	(0.047)	(0.051)
priority group HI	0.086**	0.103***	0.108**	0.063*	0.086**
	(0.036)	(0.023)	(0.044)	(0.037)	(0.043)
female	-0.026	-0.038*	-0.021	0.007	0.006
	(0.028)	(0.022)	(0.030)	(0.023)	(0.024)
Ethnic_Minority	-0.041	-0.042	-0.025	-0.027	-0.037
	(0.043)	(0.031)	(0.049)	(0.034)	(0.037)
ln(HH income per capita)	-0.040	-0.047**	-0.040	-0.034	-0.029
	(0.029)	(0.021)	(0.028)	(0.022)	(0.021)
ln(work hour)	-0.017	-0.044*	-0.000	-0.026	-0.010
	(0.016)	(0.027)	(0.021)	(0.017)	(0.018)
can read Kinh	0.105	0.105**	0.143*	0.049	0.044
	(0.074)	(0.049)	(0.078)	(0.077)	(0.080)
health	No	No	No	Yes	Yes
Observations	704	704	704	698	698

Use time/distance to central hosp and district hosp as the additional IVs

	(1) IV:Time	(2) IV:Time	(3) IV:Distance	(4) IV:Distance
Central hospital	-0.282**	-0.279**	-0.282**	-0.283**
	(0.128)	(0.127)	(0.129)	(0.131)
Province/City hospital	0.172	0.161	0.168	0.150
	(0.139)	(0.127)	(0.140)	(0.133)
Commune health center	-0.146	-0.067	-0.150	-0.068
	(0.115)	(0.104)	(0.113)	(0.103)
poor HH HI	-0.020	-0.053	-0.020	-0.055
	(0.057)	(0.052)	(0.056)	(0.051)
priority group HI	0.106**	0.084**	0.106**	0.083**
	(0.042)	(0.042)	(0.043)	(0.042)
female	-0.022	0.006	-0.022	0.007
	(0.030)	(0.024)	(0.030)	(0.024)
Ethnic_Minority	-0.027	-0.038	-0.027	-0.038
	(0.049)	(0.037)	(0.049)	(0.037)
ln(HH income per capita)	-0.040	-0.029	-0.040	-0.029
	(0.028)	(0.021)	(0.028)	(0.021)
ln(work hour)	-0.001	-0.010	-0.001	-0.011
	(0.020)	(0.018)	(0.020)	(0.018)
can read Kinh	0.139*	0.042	0.139*	0.039
	(0.080)	(0.082)	(0.081)	(0.083)
health	No	Yes	No	Yes

Imperfect instrumental variables

	(1) IV Z	(2) IV V	(3) IV Z: outpatient	(4) IV V: outpatient	(5) est5
Central hospital	-0.277**	-0.083	-0.277**	-0.083	-0.277
	(0.131)	(0.163)	(0.131)	(0.163)	(0.13
Province/City hospital	0.172	-0.005	0.172	-0.005	0.17
	(0.137)	(0.148)	(0.137)	(0.148)	(0.13
Commune health center	-0.069	-0.082	-0.069	-0.082	-0.06
	(0.102)	(0.099)	(0.102)	(0.099)	(0.10
Observations	698	698	698	698	698

- Correcting the endogenous healthcare facility choice (Column (5)) substantially increases the magnitude of the coefficient of central hospital and commune health center.
  - If a patient visits the central hospital, the probability of utilizing insurance will be 31 percentage points lower than visiting the district hospitals.
- Durbin-Wu-Hausman tests rejects the null hypothesis that these facility choice variables are exogenous (p-value < 0.001).
  - The low prospect of health insurance utilization at the central hospital actually discourages people from visiting there.
- Patient with chronic disease are more likely to utilize insurance; patient with severe illness are less likely to utilize the insurance, suggesting that in the case of the severe illness, the patients prioritize the treatment, and insurance utilization is the second matter. However, if the illness is very severe, in which case the medical expenses could be huge, patients seem to try to find a way to utilize the insurance.
- The low insurance utilization at the central hospital is not due to the health characteristics of the patients. Rather, once we control the health and illness status, the insurance utilization at the central hospital is even lower.

Coding tips (1)

- Use macro to reduce programming errors and complication which often occur in revising the code.
- Use LaTeX to automatically update the results in the manuscript.

```
global xvar x1 x2 x3     // "$xvar" is read as "x1 x2 x3"
global ivset z1 z2 z3
global control w1 w2 w3 w4 w5 i.district
* global control $control w6
* global control w1 w2 w3 w4 w6
global ifs "if sample"
global sumopt "cells("mean(fmt(2)) sd(fmt(2)) /*
   */ min(fmt(1)) max(fmt(1))") tex replace label nomtitle"
global meanopt "b(%8.3f) se(%8.3f) tex replace label wide nostar"

** Summary statistics
cd $output
est clear
 estpost tabstat y $xvar $ivset $control $ifs, s(mean sd max min) c(s)
esttab using sum_a.tex, $sumopt

est clear
 eststo: mean y $xvar $ivset $control $ifs, noheader
esttab using sum_b.tex, $meanopt
```

Coding tips (2)

```
** Regression
global vce vce(cluster district)   // Define variance estimator
global dropv *district _cons
global etopt r drop($dropv, relax) omit nobase label compress nogaps /
 */ b(%8.3f) se(%8.3f) star(* 0.05 ** 0.01) i("district= *district")

est clear
 eststo: reg y $xvar $control $ifs, $vce    // OLS

 probit y $xvar $control $ifs, $vce    // Probit
    eststo: margins, dydx(*) post

 eststo: ivregress 2sls y ($xvar = $ivset) $control $ifs, $vce  // IV
esttab * using reg1.tex, $etopt mtitles("OLS" "Probit" "IV")
```

- Use `input` (for table) or `include` (for graph) command in LaTex

```
\begin{table}[htbp]
  \caption{Table title here}
  \centering
  \footnotesize
  \input{../analysis/output/reg1.tex}
  \label{t reg1}
\end{table}
```


Implementing IIV in STATA: A coding example (1)

```
global xvar x1 x2 x3     // define endogeneous variables
global wvar w1 w2    // define control variables
global zvar z1 z2 z3    // define imperfect IVs

capture program drop iiv
program iiv
   qui reg $yvar $wvar [pweight=weight] if sample
      tempvar t_$yvar
      predict double `t_$yvar' if sample, resid   // get residuals
   local xzvar $xvar $zvar
   drop _iiv_*
   foreach x of local xzvar {
      qui reg `x' $wvar [pweight=weight] if sample
         predict double _iiv_`x' if regsample, resid   // get residual
      local varlabel: variable label `x'    // get the variable label
         label variable _iiv_`x' "`varlabel'"
      qui corr `x' _iiv_`x' if sample, covariance
         scalar sigma_`x' = sqrt(r(Var_1))    // obtain sigma_x
         scalar sig_t_`x' = sqrt(r(Var_2))     // obtain sigma_til_x
         scalar cov_`x'_t = r(cov_12)     // obtain sigma_x_tilde_x
   }
```

Implementing IIV in STATA: A coding example (2)

```
    global endvar ""
    global IIV ""
    global V_IV ""
    local j=1
    foreach x of global xvar {
       global endvar $endvar _iiv_`x'
       local zk : word `j' of $zvar
       global IIV $IIV _iiv_`zk'
       tempvar V_x`j'
       gen `V_x`j'' = sig_t_`x' * _iiv_`zk' - sig_t_`zk' * _iiv_`x'
          global V_IV $V_IV `V_x`j''
       local ++j
    }
    eststo: ivregress 2sls `t_$yvar' ($endvar =$IIV) [pweight=weight] i
    estat firststage, all
    eststo: ivregress 2sls `t_$yvar' ($endvar =$V_IV) [pweight=wtad] if
    estat firststage, all
end
```


The magnitude of the effect of insurance nonuntilization

- How important is the insurance nonutilization in terms of the health expenditure?
- For examining the effect of health insurance nonutilization on the health expenditure, we use the self-reported health care expenditure which is elicited by the question “In total, how much did you pay for this visit?”.
- The drawback of this question is that it is not clear if the respondent reports the health care expenditure net of health insurance reimbursement.
- It is likely that some respondents report the expenditure net of reimbursement, and others report the gross health care expenditure.
- We use a second outcome variable which is obtained by subtracting the amount of health insurance reimbursement elicited by the question “How much were you paid by health insurance in total at that time?”. The true value of the expenses should be somewhere between the first outcome variable and the second one.

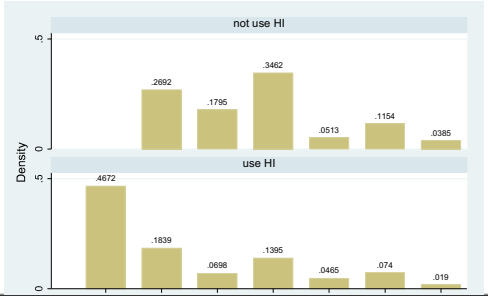
The amount of the self-reported health care expenditure for the visit at the last illness

	count	mean	sd
not use HI			
last illness: pay for the visit	78	1037730.8	3735050.8
last illness: pay for examination	81	84252.1	406606.1
last illness: pay for medication	83	639999.9	2731860.4
last illness: pay for incentive money for health staff	81	28888.8	176167.5
last illness: pay for transport	82	52073.1	233920.9
last illness: pay for other category	83	156024.0	690499.9
use HI			
last illness: pay for the visit	473	544564.5	3147995.1
last illness: pay for examination	372	64089.8	578916.0
last illness: pay for medication	387	217278.7	1317363.0
last illness: pay for incentive money for health staff	385	13506.4	77600.2
last illness: pay for transport	383	58018.1	185152.1
last illness: pay for other category	386	237745.9	1114456.1
Total			
last illness: pay for the visit	551	614377.5	3238389.6
last illness: pay for examination	453	67694.9	551730.2
last illness: pay for medication	470	291929.5	1661084.3
last illness: pay for incentive money for health staff	466	16180.1	101716.5
last illness: pay for transport	465	56969.7	194371.7
last illness: pay for other category	469	223283.4	1051787.0

Motivation and Research Question Health Insurance System in Vietnam Data Empirical Results on Insurance Utilization The magnitude of the effect of insu

The amount of the self-reported health care expenditure for the visit at the last illness

- While almost half of those who used insurance paid nothing for the visit, more than a third of those who did not use insurance paid 100,001 - 500,000 VND.
- The ratio of those who paid more than 1 million VND for the visit is also higher for those who did not use health insurance.
- More than a forth of those who did not use insurance eventually pay no more than 50,000 VND, in which case there is not so much benefit for insurance utilization.



Econometric analysis

- The characteristics of those who used the health insurance and those who did not are different.
- To estimate the average treatment effect on the treated of insurance nonutilization, we use the nearest neighbor matching and the regression adjustment with propensity score weighting.
  - The latter is doubly robust in the sense that consistency of the estimators only requires either the conditional mean model or the propensity score model to be correctly specified, not both.
  - Only use the observations with $0.1 < \hat{p} < 0.9$.
- Use the same set of the covariates in the insurance nonutilization analysis.
- These may not suffice to explain the difference between the two groups, but it is likely that the estimates are conservative ones.
  - If there remains some unobservables which affect the expenses, a major factor will be “selection” and “moral hazard”.
  - People do not utilize insurance because their expenses are not so high.
  - People who use insurance will receive more healthcare service as the costs are discounted.

Impact of Insurance Nonutilization on Healthcare Expenditure

- Columns (1) and (2) use the self-reported healthcare expenditure (“In total, how much did you pay for this visit?”).
- Columns (3) and (4) use a second outcome variable which is obtained by subtracting the amount of health insurance reimbursement elicited by the question “How much were you paid by health insurance in total at that time?”.
- The estimated effect:: around 0.9-1.1 million VND (1.5 times higher than the average of the healthcare expenses)

Table: Health care expenses for the visit

	(1) NNM	(2) Reg Adj PSW	(3) NNM	(4) Reg Adj PSW
HI nonutilization	843084.5***	1084540*	974463.9***	1181121*
	(198194.2)	(571180.2)	(187435.4)	(659005.8)
Observations	549	124	549	124

Conclusion

- Expanding the coverage of the health insurance will not be enough. Some occurrence of insurance nonutilization.
- In Vietnam, insurance nonutilization is actually an issue. The likelihood of insurance utilization is significantly and substantially lower at the public central hospital, which in turn discourages people from visiting there.
  - Probably due to the lack in incentives in overcrowded public hospitals?
- The average effect of insurance nonutilization on the healthcare expenses per visit is around 1 million VND, which is 1.5 times larger than the average healthcare expenses per visit of our sample.
- Barriers to insurance utilization may partly explain the mixed results of health insurance on healthcare utilization and OOP payment across countries.
- Need to improve supply side as well.
- Self-targeting?? Target for those whose time costs are low.

2015 Japanese Stata Users Group Meeting

@一橋大学一橋講堂(東京) 2015/8/28

Stataを用いたデータ管理法

自治医科大学
企画経営部医療情報部
臨床研究支援センターデータセンター部門
興梠　貴英

臨床試験におけるデータ管理の問題
KYOTO HEART Studyは何が問題だったのか

Valsartan試験に関して当初指摘された問題点

表. 開始時および達成時の血圧値

血圧値 (mmHg)	バルサルタン群		対照群※	
	開始時	達成時	開始時	達成時
Jikei Heart Study				
SBP平均値（SD）	139.2 (11)	132.0 (14)	138.8 (11)	132.0 (14)
DBP平均値（SD）	81.4 (11)	76.7 (8)	81.4 (11)	76.6 (9)
Kyoto Heart Study				
SBP平均値（SD）	157 (14)	133 (14)	157 (14)	133 (14)
DBP平均値（SD）	88 (11)	76 (11)	88 (11)	76 (10)
Valsartan Amlodipine Randomized Trial				
SBP平均値（SD）	158 (19)	135 (13)	158 (18)	135 (14)
DBP平均値（SD）	93 (13)	80 (10)	94 (13)	80 (10)

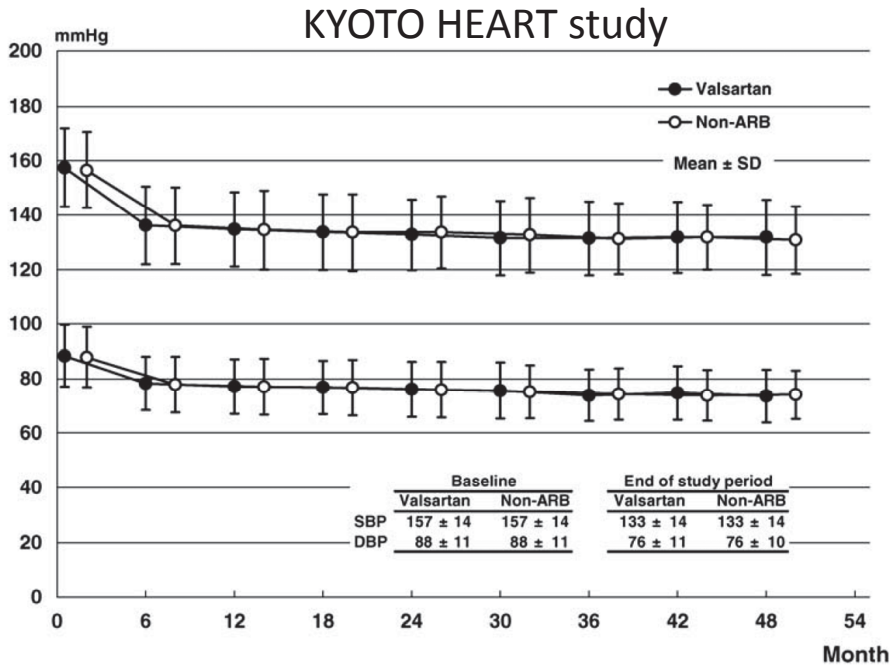
SBP：収縮期血圧，DBP：拡張期血圧，SD：標準偏差

※ Jikei Heart Study, Kyoto Heart Studyでは
ARBなしの従来治療群，VARTではアムロジピン群

（出典：Lancet 2012; 379: e48）

http://mtpro.medical-tribune.co.jp/mtpronews/1204/1204042.html

経過中の血圧も二群間でよく一致



Eur Heart J (2009) 30:2461-2469, doi: 10.1093/eurheartj/ehp363

糖尿病新規発症患者に注目したサブ解析

Circulation Journal
Official Journal of the Japanese Circulation Society
http://www.j-circ.or.jp

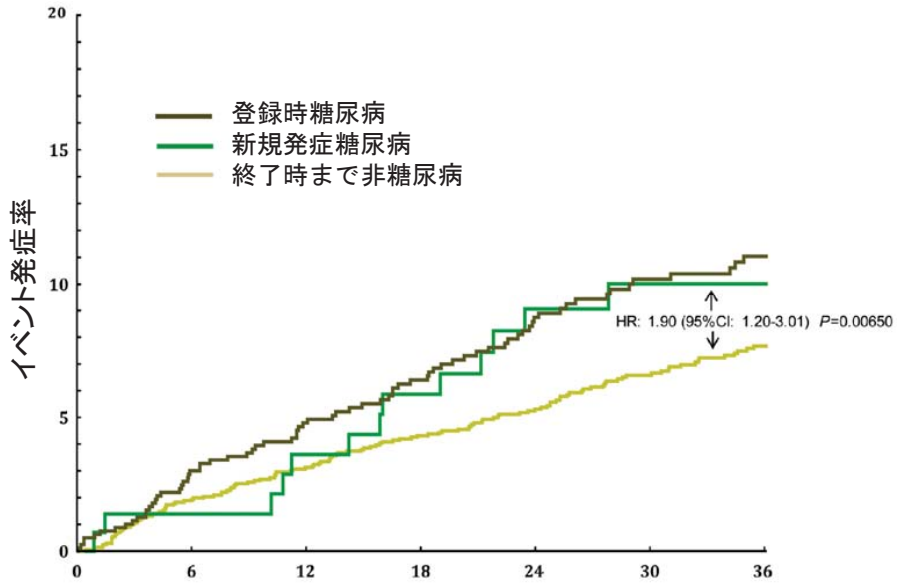
Effects of Valsartan on Cardiovascular Morbidity and Mortality in High-Risk Hypertensive Patients With New-Onset Diabetes Mellitus

– Sub-Analysis of the KYOTO HEART Study –

Shinzo Kimura, MD, PhD; Takahisa Sawada, MD, PhD; Jun Shiraishi, MD, PhD;
Hiroyuki Yamada, MD, PhD; Hiroaki Matsubara, MD, PhD
for the KYOTO HEART Study Group

Circ J. 2012. pii: DN/JST.JSTAGE/circj/CJ-12-0387

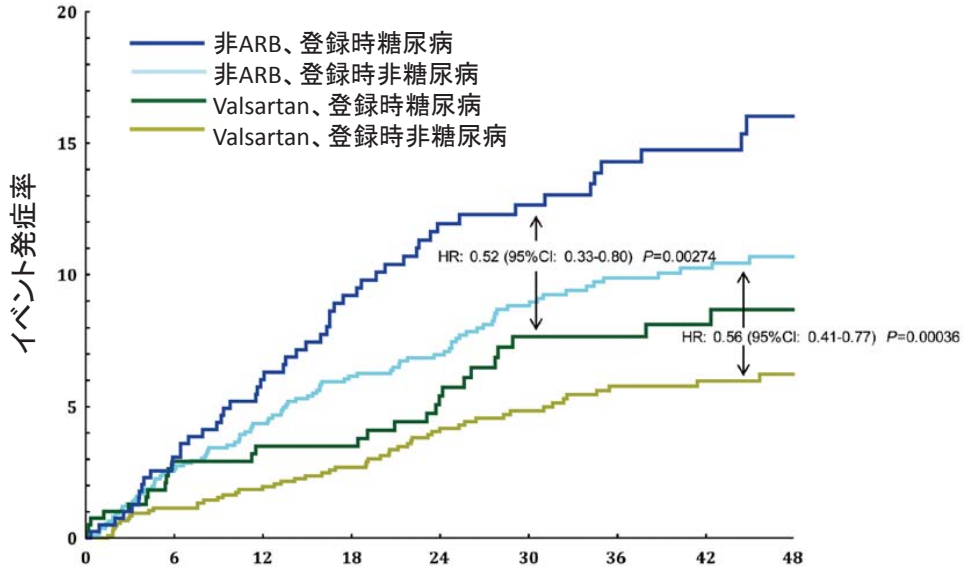
主要な結果-1



ランダム化後の時間(月)

Circ J. 2012. pii: DN/JST.JSTAGE/circj/CJ-12-0387

主要な結果-2



Circ J. 2012. pii: DN/JST.JSTAGE/circj/CJ-12-0387

問題は背景データ

Table 1. Baseline Characteristics for Baseline Diabetes and Baseline Non-Diabetes

	Baseline diabetes			Baseline non-diabetes			
	All patients (n=807)	Valsartan (n=401)	Non-ARB (n=406)	All patients (n=2,224)	Valsartan (n=1,116)	Non-ARB (n=1,108)	
Age (years)	65±11	65±11	65±11	65±11	65±11	65±11	
Gender (M/F)	487/320	247/154	240/166	1,241/983	614/502	627/481	
Systolic blood pressure (mmHg)	156±13	156±13	156±13	157±15	158±15	157±14	
Diastolic blood pressure (mmHg)	85±11	85±10	85±11	89±12	89±11	89±12	
Heart rate (beats/min)	70±17	69±18	71±17	70±17	70±17	71±16	
Body-mass index (kg/m²)	25±4	25±4	25±4	24±4	24±4	24±4	
Echocardiography							
Ejection fraction (%)	62±9	62±9	63±9	63±10	63±10	63±10	
Fractional shortening (%)	38±7	38±7	38±8	38±25	37±30	38±19	
LV mass index	173±50	175±49	172±52	175±55	176±56	175±54	
LDL-cholesterol (mg/dl)	118±31	119±31	117±32	124±42	123±51	125±30	
HDL-cholesterol (mg/dl)	53±14	54±15	52±14	56±16	55±15	56±16	
Trigriceride (mg/dl)	149±95	146±103	152±86	152±94	151±93	153±96	
HbA_{1c} (%)	7.2±1.5	7.2±1.7	7.1±1.4	5.4±1.8	5.5±2.5	5.4±0.5	***
Fasting plasma glucose (mg/dl)	164±67	164±68	163±65	106±20	107±20	106±20	***
Serum creatinine (mg/dl)	0.9±0.4	0.8±0.3	0.9±0.5	0.8±0.3	0.8±0.3	0.8±0.3	
Serum sodium (mmol/L)	146±13	141±13	150±16	143±30	143±41	142±13	
Serum potassium (mmol/L)	4.5±3.9	4.5±5.0	4.4±2.2	4.4±6.8	4.5±9.3	4.3±2.2	
Risk factors, n (%)							
Current smokers	155 (19%)	83 (21%)	72 (18%)	518 (23%)	258 (23%)	260 (23%)	*
Diabetes mellitus	807 (100%)	401 (100%)	406 (100%)	0 (0%)	0 (0%)	0 (0%)	
Dyslipidemia	510 (63%)	252 (63%)	258 (64%)	1,634 (73%)	814 (73%)	820 (74%)	***
Obesity	341 (42%)	174 (43%)	167 (41%)	836 (38%)	419 (38%)	417 (38%)	*
LV hypertorophy	202 (25%)	104 (26%)	98 (24%)	601 (27%)	295 (26%)	306 (28%)	
Cerebrovascular disease	25 (3%)	13 (3%)	12 (3%)	98 (4%)	45 (4%)	53 (5%)	
Coronary heart disease	212 (26%)	98 (24%)	114 (28%)	495 (22%)	257 (23%)	238 (21%)	
Heart failure	47 (6%)	18 (4%)	29 (7%)	146 (7%)	66 (6%)	80 (7%)	
Medication, n (%)							
CCB	459 (57%)	227 (57%)	232 (57%)	1,198 (54%)	598 (54%)	600 (54%)	
ACE inhibitor	228 (28%)	107 (27%)	121 (30%)	366 (16%)	182 (16%)	184 (17%)	***
β-blocker	154 (19%)	74 (18%)	80 (20%)	387 (17%)	190 (17%)	197 (18%)	
α-blocker	31 (4%)	14 (3%)	17 (4%)	65 (3%)	31 (3%)	34 (3%)	
Thiazide	23 (3%)	12 (3%)	11 (3%)	74 (3%)	40 (4%)	34 (3%)	
Anti-aldosterone	19 (2%)	10 (2%)	9 (2%)	38 (2%)	21 (2%)	17 (2%)	
Other diuretics	61 (8%)	26 (6%)	35 (9%)	101 (5%)	50 (4%)	51 (5%)	
Statin	329 (41%)	165 (41%)	164 (40%)	665 (30%)	326 (29%)	339 (31%)	***
Fibrate	25 (3%)	14 (3%)	11 (3%)	40 (2%)	21 (2%)	19 (2%)	
Other anti-hyperlipidemic agents	23 (3%)	18 (4%)	5 (1%)	51 (2%)	27 (2%)	24 (2%)	
Sulfonyl urea	342 (42%)	173 (43%)	169 (42%)	0 (0%)	0 (0%)	0 (0%)	
Other oral hyperglycemic agents	275 (34%)	137 (34%)	138 (34%)	0 (0%)	0 (0%)	0 (0%)	
Insulin	80 (10%)	37 (9%)	43 (11%)	0 (0%)	0 (0%)	0 (0%)	

CAD/非CAD比較

Cardio-Cerebrovascular Protective Effects of *Valsartan* in High-Risk Hypertensive Patients With Coronary Artery Disease (from the Kyoto Heart Study)

Jun Shiraishi, MD, PhD[a,*], Takahisa Sawada, MD, PhD[b], Masahiro Koide, MD, PhD[b], Hiroyuki Yamada, MD, PhD[b], and Hiroaki Matsubara, MD, PhD[b], for the Kyoto Heart Study Group

Am J Cardiol. 2012 May 1;109(9):1308-14.

血清Na値のSD

Table 1
Baseline characteristics

Characteristics	With CAD				Without CAD				With vs Without CAD
	All (n = 707)	Valsartan (n = 355)	Non-ARB (n = 352)	p Value	All (n = 2,324)	Valsartan (n = 1,162)	Non-ARB (n = 1,162)	p Value	
Age (years)	70 ± 9	70 ± 9	70 ± 9	0.3001	65 ± 11	65 ± 11	65 ± 11	0.6964	<0.0001
Men/women	475/232	236/119	239/113	0.7477	1,253/1,071	625/537	628/534	0.9337	0.0001
Systolic blood pressure (mm Hg)	156 ± 14	157 ± 15	154 ± 13	0.1226	157 ± 14	157 ± 14	157 ± 14	0.6936	0.0959
Diastolic blood pressure (mm Hg)	85 ± 11	85 ± 11	85 ± 11	0.7651	89 ± 11	89 ± 11	89 ± 12	0.1549	0.0001
Heart rate (beats/min)	70 ± 15	70 ± 15	70 ± 15	0.8514	70 ± 18	70 ± 18	71 ± 17	0.1815	0.9772
Body mass index (kg/cm^2)	24 ± 3	24 ± 4	24 ± 3	0.7621	25 ± 4	25 ± 4	25 ± 4	0.5037	0.0522
Waist size (cm)	86 ± 10	86 ± 10	86 ± 10	0.7385	85 ± 11	86 ± 10	85 ± 11	0.5763	0.4057
Cardiothoracic ratio (%)	51 ± 6	50 ± 6	52 ± 6	0.1606	51 ± 5	51 ± 6	51 ± 5	0.6805	0.8049
Electrocardiography (S wave in lead V_1 + R wave in lead V_5, mm)	30 ± 11	30 ± 11	30 ± 12	0.6086	31 ± 11	31 ± 10	30 ± 11	0.4892	0.2326
Echocardiographic ejection fraction (%)	60 ± 11	60 ± 11	60 ± 11	0.8230	64 ± 9	64 ± 9	64 ± 9	0.9215	<0.0001
Low-density lipoprotein cholesterol (mg/dl)	113 ± 30	112 ± 30	115 ± 29	0.1448	125 ± 42	125 ± 51	125 ± 31	0.8562	<0.0001
High-density lipoprotein cholesterol (mg/dl)	52 ± 15	53 ± 15	51 ± 15	0.1572	56 ± 16	56 ± 15	56 ± 16	0.7673	<0.0001
Triglycerides (mg/dl)	144 ± 78	143 ± 80	144 ± 75	0.8137	153 ± 99	152 ± 100	155 ± 98	0.4060	<0.0001
Hemoglobin A1c (%)	6.2 ± 2.5	6.2 ± 3.3	6.2 ± 1.3	0.9316	6.0 ± 1.7	6.1 ± 2.1	5.9 ± 1.2	0.1237	0.0361
Fasting plasma glucose (mg/dl)	123 ± 42	120 ± 40	125 ± 44	0.1332	122 ± 47	122 ± 49	121 ± 46	0.4399	0.5669
Serum creatinine (mg/dl)	0.9 ± 0.3	0.9 ± 0.3	0.9 ± 0.3	0.2332	0.8 ± 0.3	0.8 ± 0.2	0.8 ± 0.3	0.7180	0.0777
Estimated glomerular filtration rate (ml/min/1.73 m^2)	64 ± 20	64 ± 20	63 ± 20	0.5389	70 ± 20	70 ± 19	70 ± 20	0.7773	<0.0001
Serum sodium (mEq/L)	142 ± 17	141 ± 8	143 ± 22	0.1043	144 ± 83	143 ± 40	145 ± 110	0.5119	0.6242
Serum potassium (mEq/L)	4.2 ± 0.4	4.2 ± 0.4	4.2 ± 0.4	0.9832	4.5 ± 7.0	4.6 ± 9.6	4.3 ± 2.5	0.4293	0.4029
Current smokers	152 (21%)	76 (21%)	76 (22%)	0.9741	521 (22%)	265 (23%)	256 (22%)	0.6907	0.6433
Obesity	231 (33%)	111 (31%)	120 (35%)	0.4715	946 (41%)	482 (41%)	464 (40%)	0.4729	0.0001
Diabetes mellitus	212 (30%)	98 (28%)	114 (32%)	0.1919	595 (26%)	303 (26%)	292 (25%)	0.6346	0.0238
Dyslipidemia	484 (68%)	236 (66%)	248 (70%)	0.2907	1,660 (71%)	830 (71%)	830 (71%)	0.9634	0.1408
Cerebrovascular disease	23 (3%)	11 (3%)	12 (3%)	0.9835	100 (4%)	47 (4%)	47 (4%)	0.6093	0.2586
Heart failure	80 (11%)	33 (9%)	47 (13%)	0.1133	113 (5%)	51 (4%)	62 (5%)	0.3348	0.0666

Am J Cardiol. 2012 May 1;109(9):1308-14.

肥満患者のサブ解析

Contents lists available at SciVerse ScienceDirect

International Journal of Cardiology

journal homepage: www.elsevier.com/locate/ijcard

Cardio-cerebrovascular protective effects of valsartan in high-risk hypertensive patients with overweight/obesity: A post-hoc analysis of the KYOTO HEART Study

Hidekazu Irie [a,1], Jun Shiraishi [a,1,*], Takahisa Sawada [b], Masahiro Koide [b], Hiroyuki Yamada [b], Hiroaki Matsubara [b] and for the KYOTO HEART Study Group

[a] Department of Cardiology, Kyoto First Red Cross Hospital, Honmachi, Higashiyama-ku, Kyoto 605-0981, Japan
[b] Department of Cardiovascular Medicine, Kyoto Prefectural University School of Medicine, Kawaramachi-Hirokoji, Kamigyo-ku, Kyoto 602-8566, Japan

Int J Cardiol. 2012 Jul 12.

血清Na値のSD

Table 1
Baseline characteristics of the population with and without overweight/obesity at baseline.

Characteristics	With overweight/obesity			Without overweight/obesity			Comparison between overweight/obesity; +/−
	All n=1177	Valsartan n=593	Non-ARB n=584	All n=1854	Valsartan n=924	Non-ARB n=930	
Age (years)	64±11	64±12	64±11	67±11	67±11	68±10	*<0.0001*
Gender (men/women)	630/547	312/281	318/266	1098/756	614/504	594/516	0.0023
Systolic blood pressure (mm Hg)	156±14	157±14	156±14	157±14	158±15	157±14	*0.1552*
Diastolic blood pressure (mm Hg)	89±11	89±11	89±11	88±12	88±11	87±12	*0.2680*
Heart rate (beats/min)	70±17	69±18	71±16	71±17	71±18	71±17	*0.2063*
Body mass index (kg/m^2)	27±3	27±3	27±4	23±3	23±3	23±3	*<0.0001*
Overweight/obesity	1000/ 177	504/ 89	496/ 88				
Waist size (cm)	92±9	92±9	92±9	81±9	81±9	81±9	*<0.0001*
Cardiothoracic ratio (%)	52±10	51±6	52±5	50±11	50±6	50±6	*0.2249*
Electrocardiography (Sv1+Rv5 mm)	29±12	29±10	29±10	31±14	32±11	31±11	*0.1470*
Echocardiography							
Ejection fraction (%)	63±9	63±9	63±9	63±10	63±10	63±10	*0.5869*
LDL-cholesterol (mg/dL)	125±51	126±64	124±31	121±31	120±31	122±31	*0.0079*
HDL-cholesterol (mg/dL)	53±14	53±15	53±14	56±16	56±16	56±16	*<0.0001*
Triglyceride (mg/dL)	163±105	163±114	163±94	144±86	141±80	147±92	*<0.0001*
HbA1c (%)	6.1±2.0	6.2±2.5	6.1±1.3	6.0±1.9	6.0±2.3	6.0±1.2	*0.1039*
Fasting plasma glucose (mg/dL)	124±49	124±48	124±49	120±44	121±46	120±43	*0.1409*
Serum creatinine (mg/dL)	0.8±0.3	0.8±0.3	0.8±0.3	0.9±0.4	0.8±0.3	0.9±0.4	*0.1178*
Serum sodium (mEq/L)	143±40	144±55	142±12	143±88	141±8	146±123	*0.9038*
Serum potassium (mEq/L)	4.3±2.9	4.4±4.1	4.2±0.4	4.5±7.5	4.6±10.3	4.4±2.7	*0.3628*
Risk factor n (%)							
Current smokers	210 (18%)	109 (18%)	101 (17%)	463 (25%)	232 (25%)	231 (25%)	*<0.0001*
Diabetes mellitus	341 (29%)	174 (29%)	167 (29%)	466 (25%)	227 (25%)	239 (26%)	*<0.0001*
Dyslipidemia	784 (67%)	390 (66%)	394 (67%)	1360 (73%)	672 (73%)	688 (74%)	*<0.0001*
Coronary heart disease	231 (20%)	111 (19%)	120 (21%)	476 (26%)	244 (26%)	232 (25%)	*<0.0001*
Cerebrovascular disease	33 (3%)	18 (3%)	15 (3%)	90 (5%)	40 (4%)	50 (5%)	*0.0071*
Heart failure	57 (5%)	22 (4%)	35 (6%)	136 (7%)	62 (7%)	74 (8%)	*0.0078*

Int J Cardiol. 2012 Jul 12.

CKDサブ解析

Contents lists available at SciVerse ScienceDirect

International Journal of Cardiology

journal homepage: www.elsevier.com/locate/ijcard

Enhanced cardio-renal protective effects of valsartan in high-risk hypertensive patients with chronic kidney disease: A sub-analysis of KYOTO HEART Study

Katsuya Amano [a,1,2], Jun Shiraishi [b,*,1,2], Takahisa Sawada [a,1], Masahiro Koide [a,1], Hiroyuki Yamada [a,1], Hiroaki Matsubara [a,1]

[a] Department of Cardiovascular Medicine, Kyoto Prefectural University School of Medicine, Kawaramachi-Hirokoji, Kamigyo-ku, Kyoto 602-8566, Japan
[b] Department of Cardiology, Kyoto First Red Cross Hospital, Honmachi, Higashiyama-ku, Kyoto 605-0981, Japan

Int J Cardiol. 2012 Feb 13.

Table 1
Baseline characteristics of the population with and without CKD at baseline. Comparison between with/without CKD: *p<0.05; **p<0.01; ***p<0.001. Comparison between valsartan/non-ARB: †p<0.01.

	With CKD			Without CKD		
	Valsartan	Non-ARB	All	Valsartan	Non-ARB	All
n=	480	501	981	989	959	1948
Age (years)	70±10	70±10	70±10	63.9±11.1	63.8±10.8	63.8±11***
Gender (men/women)	273/207	276/225	549/432	563/426	561/398	1124/824
Blood pressure (mm Hg)						
Systolic blood pressure	158±15	158±15	158±15	157±14	156±13	156±14**
Diastolic blood pressure	88					89±11**
Heart rate (beats/min)	71					70±17
Body mass index (kg/m²)	24				8	24.6±3.8
Cardiothoracic ratio (%)	51				9	50.4±5.1***
Electrocardiogram						
Sv1+Rv5	30				0.7	30.3±10.4
Echocardiography						
Ejection fraction (%)	62					63±9
HbA1c(%)	6.3					6.0±1.3*
LDL-cholesterol (mg/dL)	123±32	123±31	123±31	122±53	123±31	122±43
HDL-cholesterol (mg/dL)	54±15	54±16	54±16	56±15	55±15	55±15
Triglyceride (mg/dL)	151±82	152±86	152±84	149±102	154±97	151±99
Fasting plasma glucose (mg/dL)	124±49	122±44	123±47	121±45	122±46	121±46
Serum creatinine (mg/dL)	0.8±0.2	0.9±0.3	0.9±0.3	0.8±0.2	0.8±0.2	0.8±0.2***
eGFR	48±9	48±10	48±9	78±15	79±16	78±16***
Sodium (Eq/L)	144±16	142±14	143±14	141±6	146±11	144±8
Potassium (Eq/L)	4.3±0.4	4.3±1.0	4.3±0.8	4.6±1.0	4.3±2.5	4.5±0.8
Risk factor						
Obesity	189 (39%)	179 (36%)	368 (38%)	382 (39%)	384 (40%)	766 (39%)
Dyslipidemia	349 (73%)	378 (75%)	727 (74%)	687 (69%)	669 (70%)	1356 (70%)
Diabetes	124 (26%)	148 (30%)	272 (28%)	260 (26%)	241 (25%)	501 (26%)
Current smoking	87 (18%)	87 (17%)	174 (18%)	244 (25%)	232 (24%)	476 (24%)***
Coronary artery disease	143 (30%)	152 (30%)	295 (30%)	200 (20%)	182 (19%)	382 (20%)***
Cerebrovascular disease	19 (4%)	26 (5%)	45 (5%)	39 (4%)	37 (4%)	76 (4%)
Congestive heart failure	35 (7%)	67 (13%)†	102 (10%)	45 (5%)	36 (4%)	81 (4%)***

	CKD群	非CKD群
血清Cre	0.9±0.3	0.8±0.2
eGFR	48±9	78±16

異常データの検証

	Baseline diabetes			Baseline non-diabetes		
	All patients (n=807)	Valsartan (n=401)	Non-ARB (n=406)	All patients (n=2,224)	Valsartan (n=1,116)	Non-ARB (n=1,108)
Serum sodium (mmol/L)	146±13	141±13	150±16	143±30	143±41	142±13
Serum potassium (mmol/L)	4.5±3.9	4.5±5.0	4.4±2.2	4.4±6.8	4.5±9.3	4.3±2.2

Circ J. 2012 Sep 12.

全体の個数n、平均μ、SDがσの集団を個数がそれぞれn1、n2の二群に分けたときに、平均をμ1、μ2、SDをσ1、σ2とする。このとき、

$n \times \mu = n1 \times \mu1 + n2 \times \mu2$

が成り立つ。さらに、μ、μ1、μ2がおおよそ等しいとき、

$n \times \sigma^2 \fallingdotseq n1 \times \sigma1^2 + n2 \times \sigma2^2$

が成り立つ。

平均の検証～Baseline diabetes群のNaで

146 x 807 = 117822

141 x 401 + 150 x 406 = 117441

SDの検証～Baseline non-diabetesのKで

2224 x 6.8 x 6.8 = 102837.76

1116 x 9.3 x 9.3 + 1108 x 2.2 x 2.2 = 101885.56

オリジナル論文

European Heart Journal
doi:10.1093/eurheartj/ehp363

FASTTRACK
ESC HOT LINE

Effects of valsartan on morbidity and mortality in uncontrolled hypertensive patients with high cardiovascular risks: KYOTO HEART Study

Takahisa Sawada[1]*, Hiroyuki Yamada[1], Björn Dahlöf[2], and Hiroaki Matsubara[1], for the KYOTO HEART Study Group

[1]Department of Cardiovascular Medicine, Kyoto Prefectural University School of Medicine, Kajiicho 465, Kamigyoku, Kyoto 602-8566, Japan; and [2]Department of Medicine, Sahlgrenska University Hospital/ Östra, Göteborg, Sweden

Received 4 August 2009; accepted 13 August 2009

Eur Heart J. 2009 Oct;30(20):2461-9. doi: 10.1093

Table 1 Baseline characteristics

	Valsartan, n = 1517	Non-ARB, n = 1514
Age	66 (11)	66 (11)
Men/women	861/656 (57/43%)	867/647 (57/43%)
Current smoker	341 (22%)	332 (22%)
Obesity BMI ≥25	593 (39%)	584 (39%)
Coronary artery disease	355 (23%)	352 (23%)
Cerebrovascular disease	58 (4%)	65 (4%)
Heart failure	84 (6%)	109 (7%)
Diabetes	401 (26%)	406 (27%)
Dyslipidaemia	1065 (70%)	1079 (71%)
LVH by electrocardiogram	122 (8%)	129 (9%)
Systolic blood pressure (mmHg)	157 (14)	157 (14)
Diastolic blood pressure (mmHg)	88 (11)	88 (11)
Heart rate (b.p.m.)	70 (18)	70 (16)
EF (%)	63 (10)	63 (9)
HDL cholesterol (mg/dL)	55 (15)	55 (15)
LDL cholesterol (mg/dL)	121 (31)	123 (31)
Triglyceride (mgl/dL)	147 (83)	150 (84)
Fasting plasma glucose (mg/dL)	121 (43)	121 (43)
HbA1c (%)	6.1 (2.3)	6.0 (1.3)
Serum creatinine (mg/dL)	0.87 (0.35)	0.84 (0.38)
Sodium (mEq/L)	142 (2.7)	142 (2.5)
Potassium (mEq/L)	4.5 (2.2)	4.3 (2.2)

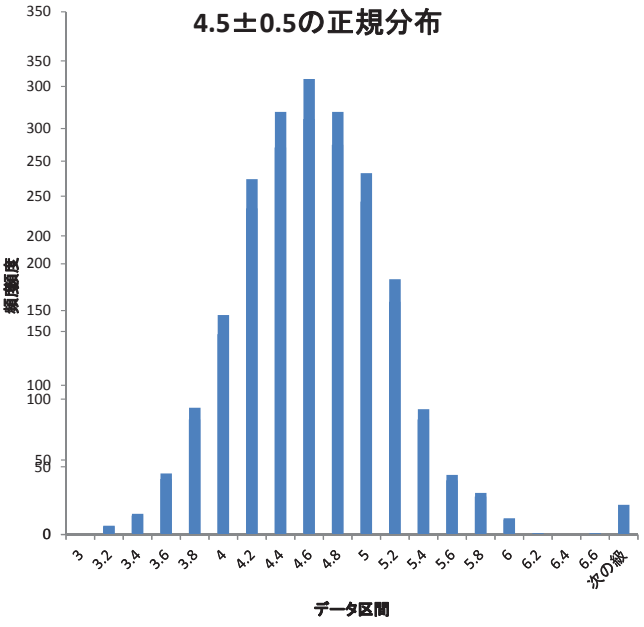
考えうるメカニズム

- KやHbA1cの値は通常小数点1桁のデータ
- データ入力時に小数点を誤って入力し忘れると…

シミュレーションデータによる検証

平均=4.5、1SD=0.5で正規分布する2000個の数列を生成

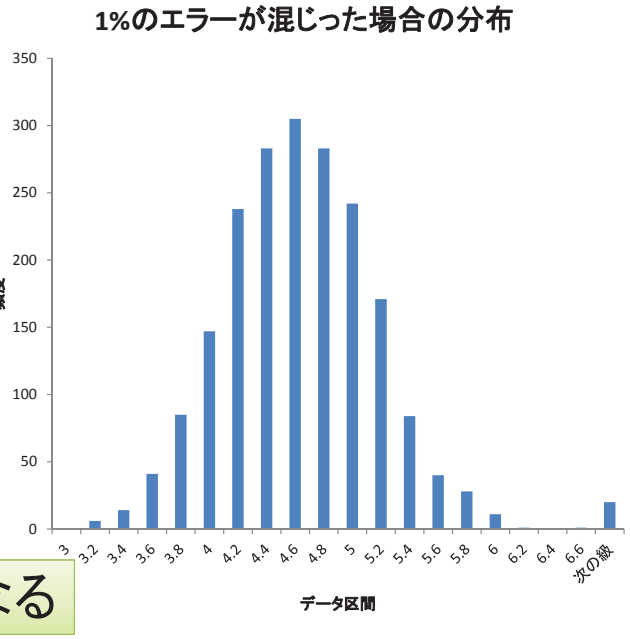
4.0, 4.8, 4.4, 4.6, 3.8, 4.7, 4.9, 4.7, 4.7, 4.8, 4.1, 4.7, 4.5, 4.8, 5.3, 4.2, 4.3, 4.3, 5.1, 4.6, 4.3, 4.6, 4.3, 5.4, 5.2, 4.7, 5.4, 4.6, 4.4, 4.6, 4.9, 5.2, 4.8, 5.0, 3.7, 5.0, 5.1, 4.2, 4.1, 4.5, 4.1, 4.5, 4.2, 4.6, 5.3, 4.6, 5.2, 4.2, 4.2, 4.4, 3.9, 4.3, 4.0, 4.5, 4.5, 4.2, 4.4, 4.1, 4.3, 4.8, 4.3, 4.2, 5.2, 4.9, 4.1, 3.7, 5.5, 4.7, 5.5, 4.9, 3.6, 3.9, 3.7, 5.4, 6.0, 4.2, 4.8, 4.1, 4.1, 4.3, 4.6, 4.9, 4.8, 4.3, 5.7, 5.0, 4.3, 4.9, 4.1, 4.9, 4.1, 4.3, 5.5, 5.1, 4.8, 4.1, 4.4, 4.1, 3.8, 3.7, 5.1, 5.7------



シミュレーションデータによる検証

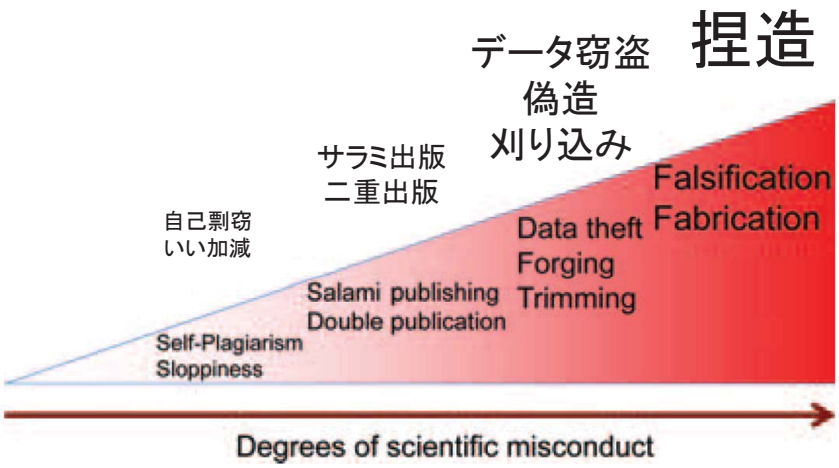
最初の20個(1%)の数字の小数点を落とすと…

40, 48, 44, 46, 38, 47, 49, 47, 47, 48, 41, 47, 45, 48, 53, 42, 43, 43, 51, 46, 4.3, 4.6, 4.3, 5.4, 5.2, 4.7, 5.4, 4.6, 4.4, 4.6, 4.9, 5.2, 4.8, 5.0, 3.7, 5.0, 5.1, 4.2, 4.1, 4.5, 4.1, 4.5, 4.2, 4.6, 5.3, 4.6, 5.2, 4.2, 4.2, 4.4, 3.9, 4.3, 4.0, 4.5, 4.5, 4.2, 4.4, 4.1, 4.3, 4.8, 4.3, 4.2, 5.2, 4.9, 4.1, 3.7, 5.5, 4.7, 5.5, 4.9, 3.6, 3.9, 3.7, 5.4, 6.0, 4.2, 4.8, 4.1, 4.1, 4.3, 4.6, 4.9, 4.8, 4.3, 5.7, 5.0, 4.3, 4.9, 4.1, 4.9, 4.1, 4.3, 5.5, 5.1, 4.8, 4.1, 4.4, 4.1, 3.8, 3.7, 5.1, 5.7------



平均=4.9、1SD=4.1となる

科学上の「不正」の序列

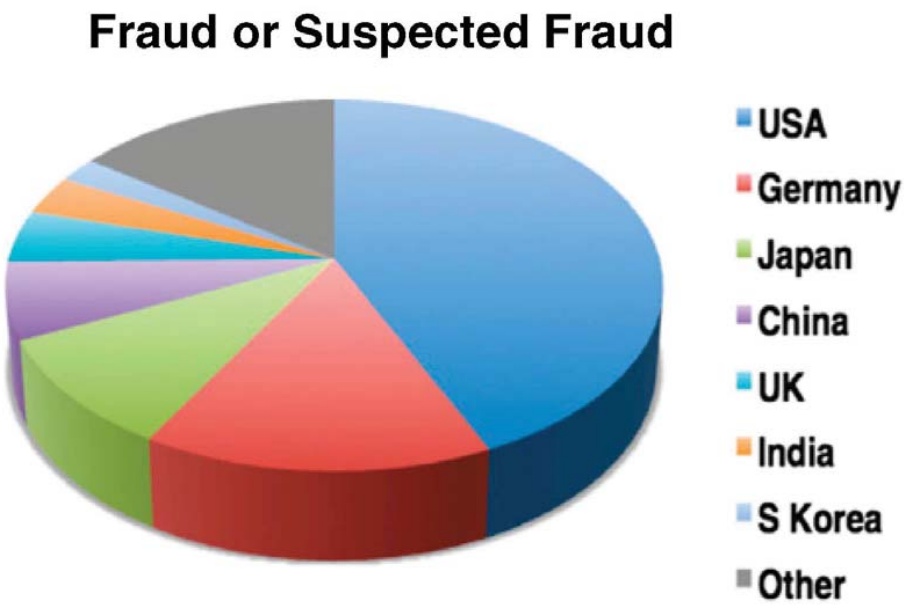


Eur Heart J. 2013 Apr;34(14):1018-23

ずさん?故意の捏造?

- 背景データのデータの扱い方、解析の仕方からは血圧のデータのみ非常に精密にコントロールされているのは奇妙である
- Circulation Journal編集委員長等からの調査・対応依頼に対し、京都府立医大は2013/1/31に故意の捏造はないとの回答をしたが、2013/2/15に元データに踏み込んだ調査するよう依頼があり、カルテデータと解析データセットの照合を行ったところ、バルサルタンに有意な方向でのデータ操作があったことが判明した

世界的に不正論文は増えている



Proc Natl Acad Sci U S A. 2012 Oct 16;109(42):17028-33

臨床試験とは

「ずさん」≒「エラー」

- データには数多くのエラーが入り込む機会がある
- 測定時…プロトコールは「空腹時血糖」となっているが、実際は食後の血糖を測定している
- データ登録時…入力ミスや転記ミス
- 解析データセット作成時…カテゴリ値への変換ミスなど

データ入力時のエラーと対策

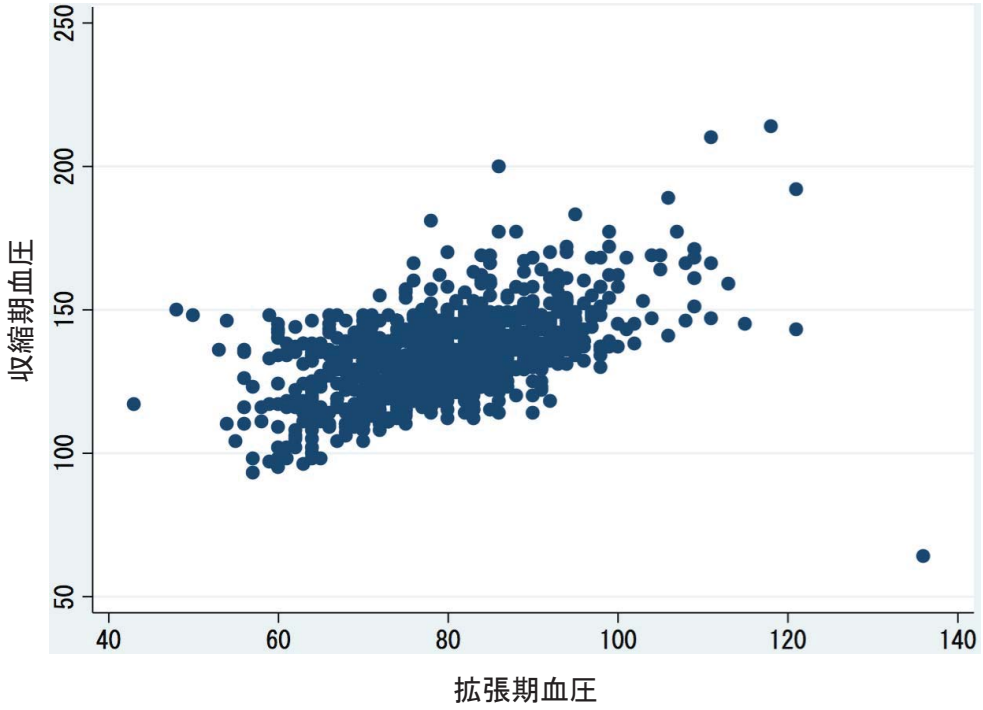
- 最近はデ ... ェック
(範囲チェ ... が多
い→Kや ... いの
データは
- しかし、チェック範囲内のエラーは検出できない(145を134と誤って入力するなど)
- データ発生時のエラーまでは必ずしも検出できない

> 身長 [ ] cm
> もし　身長 > 220 なら　警告
> もし　身長 < 100 なら　警告
> それ以外はデータ登録

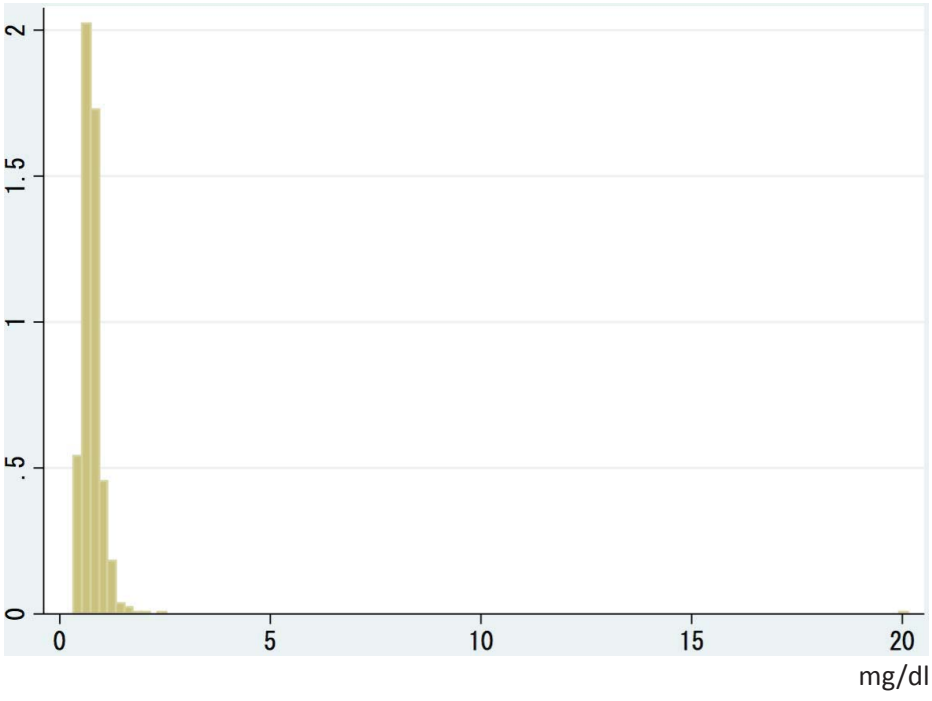
データ再チェックの重要性

- 入力時の範囲チェックは万能ではない
- 個別にはありえても組み合わせるとあり得ないデータもある(身長と体重、イベント発生日 etc.)
- そのため解析データセットを作成後は、直ちに統計解析に移るのではなく、基本統計を出してデータの分布等を評価することが重要

血圧値


収縮期血圧
250
200
150
100
50
40
60
80
100
120
140
拡張期血圧

血清Cre値


2
1.5
1
.5
0
0
5
10
15
20
mg/dl

データ変換時のエラー

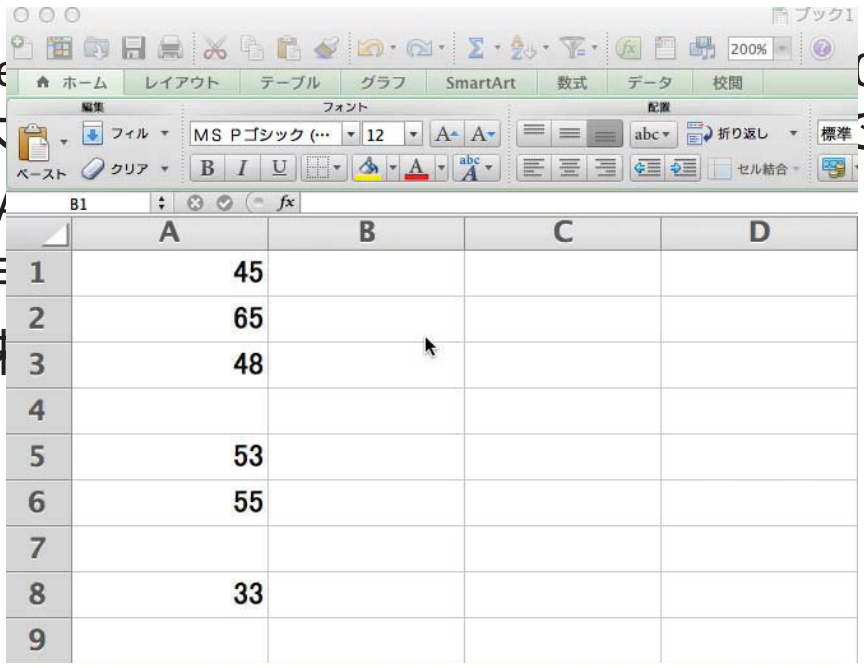
Excel上で連続値をカテゴリ化したい場合、50より大きい場合は1、50以下は0にするつもりで、

=if(A1>50,1,0)

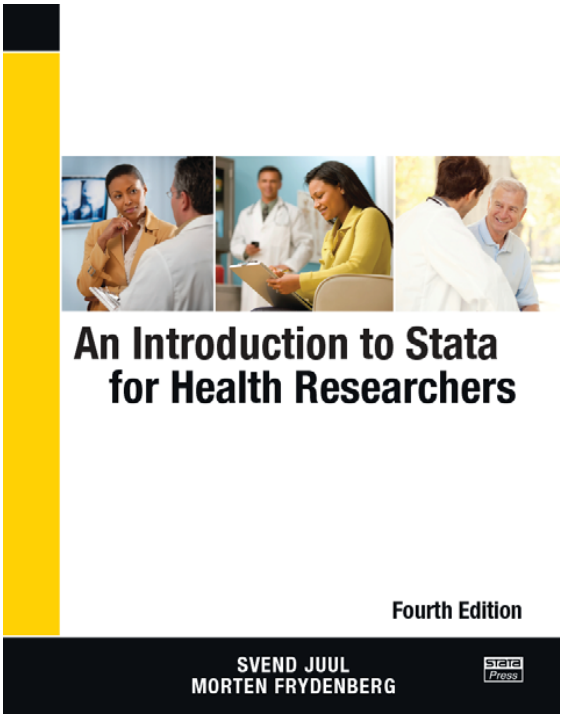
を用いると一見論理上は正しいようだが…

本来は欠損値が値が割り当てられてしまう

データ変換時のエラー



データ・解析の管理の詳細手順



目次

管理不足によって生じる問題例

- 公表した結果を再現できない。
- データセット中の参加者が何故157人だったのか説明できない。159人だったはず。
- 4週間の休みの後に自分自身のデータを理解するのに苦労する。
- データをどのようにコード化するかを間違えた。そのため間違った結果にたどり着く。
- 間違ったデータセットで作業する。
- 若干異なるバージョンのデータが異なるフォルダに入っていて、結果が一貫しない。
- 現存するテープ装置では読めないテープに保存されているために保管したデータを復元することができない。
- 注意深くバックアップを作成したが、バックアップ媒体をコンピュータと同じ研究室に保管し、後にその部屋が火事で燃えた。

データファイルの管理

表 10.1　複数の発生源からのデータセット

	インタビュー	臨床検査 1	臨床検査 2
入力データ	a_disx1.dta	b_disx1.dta	c_disx1.dta
ラベル追加	a_disx2.dta	b_disx2.dta	c_disx2.dta
エラー修正	a_disx3.dta	b_disx3.dta	c_disx3.dta
加工変数追加	a_disx4.dta	b_disx4.dta	c_disx4.dta
統合データセット		abc_disx1.dta	

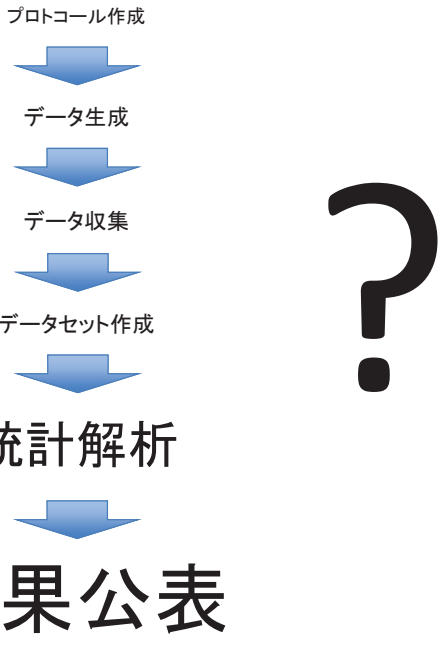
コードー覧表

変数	情報源	意味	コード、有効域	入力形式 [a]
pid	Q1	個人 ID		S10
qid	Q2	アンケート番号	1-750	3.0
sex	Q3	回答者の性別	1 男性 2 女性 9 回答無し	1.0
byear	Q4	誕生年	1890-1990 -2[b] 回答無し	4.0
schooled	Q5	教育レベル	1 9 年修了前 2 9 年修了 3 10 年修了 4 高校修了 5 その他 9 回答無し	1.0
children	Q6	子供の数	0-10 -2 回答無し	2.0
voced	Q7	職業訓練	1 なし 2 肉体労働, 3 年未満 3 肉体労働, 3 年以上 4 非肉体労働, 3 年未満 5 非肉体労働, 3-4 年 6 非肉体労働, 5 年以上 7 分類不能 9 回答無し	1.0
init	Q8	氏名の頭文字	最大 3 文字	S3

全体履歴の管理

プロジェクト: 疾病 X の治療 作業フォルダ: C:\docs\disx			
インタビューデータ			
Do-ファイル	入力データ	出力データ	コメント
	a_disx1a.rec a_disx1b.rec (EpiData ファイル)	a_disx1.dta	12oct2009 二つの修正されたデータ入力 ファイルの最終比較。 一致は a_disx1_comp.txt に記録
gen_a_disx2.do	a_disx1.dta	a_disx2.dta	13oct2009 a_disx1.dta にラベル追加。
errorfind.do	a_disx2.dta		13oct2009 エラー検索; errorfind.log 参照。
gen_a_disx3.do	a_disx2.dta	a_disx3.dta	15oct2009 発見したエラーの修正; gen_a_disx3.log 参照。
errordinf.do	a_disx3.dta		15oct2009 修正は正しい？そうでなければ gen_a_disx.do を修正。それと errorfind.do を再実行。
gen_a_disx4.do	a_disx3.dta	a_disx4.dta	16oct2009 変数 hrqol opage opagr を新たに 作成。正しいことを確認。 gen_a_disx4.log を参照。
臨床検査データ (b_disx1.dta; ここには表示しておらず)			
カルテからのデータ (c_disx1.dta; ここには表示しておらず)			
三つの情報源からのデータの併合			
gen_abc_disx1.do	a_disx4.dta b_disx4.dta c_disx4.dta	abc_disx1.dta	16oct2009 3 つの情報源からのデータを併合。 正しいことを確認。 gen_abc_disx1.log を参照。

臨床試験とは



まとめ

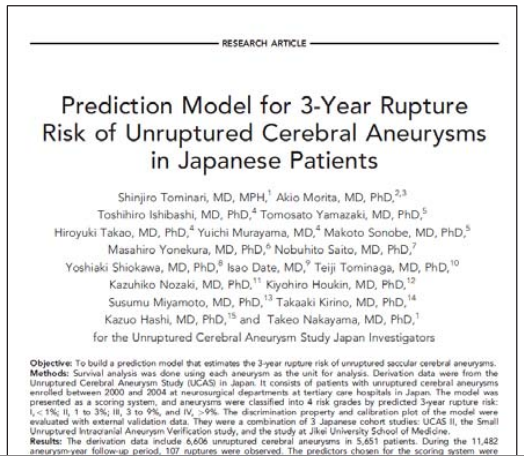
- 日本の臨床試験はその実施体制を含め、様々な問題点がある
- しかし、あまり注目されておらず、また実地における方法論も普及していないのがデータ管理の部分
- 意図的な捏造をせずともデータ管理が杜撰だと誤った結果を得ることとなる
- Stataを用いて履歴管理をしっかり行うことにより、系統的なデータ管理を行うことができる

生存分析を用いた予後予測モデル:
未破裂脳動脈瘤の3年後の破裂可能性

京都大学大学院　医学研究科
社会健康医学系専攻　健康情報学分野
富成伸次郎

本日の発表内容について

- 使用バージョン: Stata/IC 13.1
- Prediction model for 3-Year Rupture Risk of Unruptured Cerebral Aneurysm in Japanese Patients. (Ann Neurol 2015;77:1050-1059) で行った解析を、説明のため簡略化・改変したものです。



RESEARCH ARTICLE

Prediction Model for 3-Year Rupture Risk of Unruptured Cerebral Aneurysms in Japanese Patients

Shinjiro Tominari, MD, MPH,[1] Akio Morita, MD, PhD,[2,3] Toshihiro Ishibashi, MD, PhD,[4] Tomosato Yamazaki, MD, PhD,[5] Hiroyuki Takao, MD, PhD,[4] Yuichi Murayama, MD,[4] Makoto Sonobe, MD, PhD,[5] Masahiro Yonekura, MD, PhD,[6] Nobuhito Saito, MD, PhD,[7] Yoshiaki Shiokawa, MD, PhD,[8] Isao Date, MD,[9] Teiji Tominaga, MD, PhD,[10] Kazuhiko Nozaki, MD, PhD,[11] Kiyohiro Houkin, MD, PhD,[12] Susumu Miyamoto, MD, PhD,[13] Takaaki Kirino, MD, PhD,[14] Kazuo Hashi, MD, PhD,[15] and Takeo Nakayama, MD, PhD,[1] for the Unruptured Cerebral Aneurysm Study Japan Investigators

Objective: To build a prediction model that estimates the 3-year rupture risk of unruptured saccular cerebral aneurysms. **Methods:** Survival analysis was done using each aneurysm as the unit for analysis. Derivation data were from the Unruptured Cerebral Aneurysm Study (UCAS) in Japan. It consists of patients with unruptured cerebral aneurysms enrolled between 2000 and 2004 at neurosurgical departments at tertiary care hospitals in Japan. The model was presented as a scoring system, and aneurysms were classified into 4 risk grades by predicted 3-year rupture risk: I, < 1%; II, 1 to 3%; III, 3 to 9%, and IV, >9%. The discrimination property and calibration plot of the model were evaluated with external validation data. They were a combination of 3 Japanese cohort studies: UCAS II, the Small Unruptured Intracranial Aneurysm Verification study, and the study at Jikei University School of Medicine. **Results:** The derivation data include 6,606 unruptured cerebral aneurysms in 5,651 patients. During the 11,482 aneurysm-year follow-up period, 107 ruptures were observed. The predictors chosen for the scoring system were

多変量回帰分析による予測モデル

（予測モデル・・診断や予後を推測するための数式）

重回帰分析

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i$$

ロジスティック回帰分析

$$\log\{(p/1\text{-}p)\} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i$$

Cox比例ハザード回帰分析

$$\log\{h(t)/h_0(t)\} = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i$$

3

Centor criteria（診断予測モデル）

(Centor 1981)

- tonsillar exudates
- swollen tender anterior cervical nodes
- lack of a cough
- fever history （≥101°F)

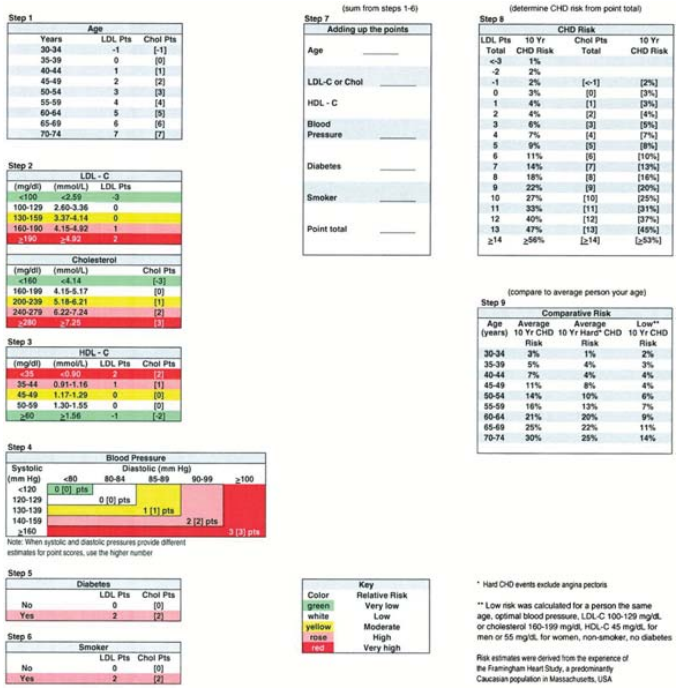
Table 3. Predictive Model for Positive Culture

Number of Variables Present	Probability of Positive Culture
4	55.7%
3	30.1–34.1%
2	14.1–16.6%
1	6.0–6.9%
0	2.5%

$$P = \frac{e^x}{1+e^x}, \text{ and } x = a + b_1(y_1) + \cdots + b_n(y_n), \text{ or}$$
$$X = -2.69 + 1.04(exudtons) + 1.00(swolacn) - 0.95(cough) + 0.89(fevhist).$$

4

Framingham heart study（予後予測モデル）



Step 1

Age		
Years	LDL Pts	Chol Pts
30-34	-1	[-1]
35-39	0	[0]
40-44	1	[1]
45-49	2	[2]
50-54	3	[3]
55-59	4	[4]
60-64	5	[5]
65-69	6	[6]
70-74	7	[7]

Step 2

LDL - C		
(mg/dl)	(mmol/L)	LDL Pts
<100	<2.59	-3
100-129	2.60-3.36	0
130-159	3.37-4.14	0
160-190	4.15-4.92	1
≥190	≥4.92	2

Cholesterol		
(mg/dl)	(mmol/L)	Chol Pts
<160	<4.14	[-3]
160-199	4.15-5.17	[0]
200-239	5.18-6.21	[1]
240-279	6.22-7.24	[2]
≥280	≥7.25	[3]

Step 3

HDL - C			
(mg/dl)	(mmol/L)	LDL Pts	Chol Pts
<35	<0.90	2	[2]
35-44	0.91-1.16	1	[1]
45-49	1.17-1.29	0	[0]
50-59	1.30-1.55	0	[0]
≥60	≥1.56	-1	[-2]

Step 4

Blood Pressure					
Systolic (mm Hg)	Diastolic (mm Hg) <80	80-84	85-89	90-99	≥100
<120	0 [0] pts				
120-129		0 [0] pts			
130-139			1 [1] pts		
140-159				2 [2] pts	
≥160					3 [3] pts

Note: When systolic and diastolic pressures provide different estimates for point scores, use the higher number

Step 5

Diabetes		
	LDL Pts	Chol Pts
No	0	[0]
Yes	2	[2]

Step 6

Smoker		
	LDL Pts	Chol Pts
No	0	[0]
Yes	2	[2]

(sum from steps 1-6)

Step 7

Adding up the points	
Age	_____
LDL-C or Chol	_____
HDL - C	
Blood Pressure	_____
Diabetes	_____
Smoker	_____
Point total	_____

Key	
Color	Relative Risk
green	Very low
white	Low
yellow	Moderate
rose	High
red	Very high

(determine CHD risk from point total)

Step 8

CHD Risk			
LDL Pts Total	10 Yr CHD Risk	Chol Pts Total	10 Yr CHD Risk
<-3	1%		
-2	2%		
-1	2%	[<-1]	[2%]
0	3%	[0]	[3%]
1	4%	[1]	[3%]
2	4%	[2]	[4%]
3	6%	[3]	[5%]
4	7%	[4]	[7%]
5	9%	[5]	[8%]
6	11%	[6]	[10%]
7	14%	[7]	[13%]
8	18%	[8]	[16%]
9	22%	[9]	[20%]
10	27%	[10]	[25%]
11	33%	[11]	[31%]
12	40%	[12]	[37%]
13	47%	[13]	[45%]
≥14	≥56%	[≥14]	[≥53%]

(compare to average person your age)

Step 9

Comparative Risk			
Age (years)	Average 10 Yr CHD Risk	Average 10 Yr Hard* CHD Risk	Low** 10 Yr CHD Risk
30-34	3%	1%	2%
35-39	5%	4%	3%
40-44	7%	4%	4%
45-49	11%	8%	4%
50-54	14%	10%	6%
55-59	16%	13%	7%
60-64	21%	20%	9%
65-69	25%	22%	11%
70-74	30%	25%	14%

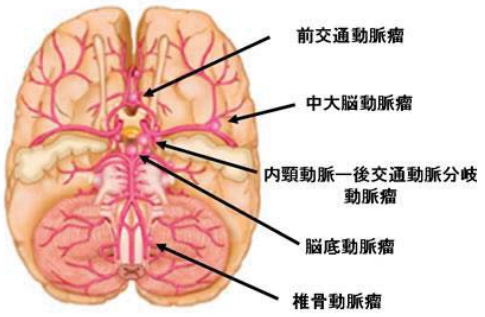
* Hard CHD events exclude angina pectoris

** Low risk was calculated for a person the same age, optimal blood pressure, LDL-C 100-129 mg/dL or cholesterol 160-199 mg/dl, HDL-C 45 mg/dL for men or 55 mg/dL for women, non-smoker, no diabetes

Risk estimates were derived from the experience of the Framingham Heart Study, a predominantly Caucasian population in Massachusetts, USA

Peter W. F. Wilson et al. Circulation. 1998;97:1837-1847

未破裂脳動脈瘤とは



破裂すると・・・

くも膜下出血（予後不良）

6

目的

- 患者背景や動脈瘤の性状などの臨床データから、未破裂脳動脈瘤の3年間の破裂可能性を計算する予後予測モデルの作成、およびその妥当性の検証。

Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD)

- 医学分野における予測モデル研究の報告の質を高めるための推奨
- 22 Itemsのchecklistから成る
- Ann Intern Med, BMJ, Circulation, J Clin Epidemiology などに掲載された
- www.tripod-statement.orgから入手可能

TRIPOD Checklist: Prediction Model Development and Validation

Section/Topic	Item		Checklist Item	Page
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.	
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	
	5b	D;V	Describe eligibility criteria for participants.	
	5c	D;V	Give details of treatments received, if relevant.	
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.	
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.	
Sample size	8	D;V	Explain how the study size was arrived at.	
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.	
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	
	10c	V	For validation, describe how the predictions were calculated.	
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.	
Risk groups	11	D;V	Provide details on how risk groups were created, if done.	
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	
Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.	
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	
	15b	D	Explain how to the use the prediction model.	
Model				

9

解析手順

1. 生存分析を用いたモデルの作成
   （Development）
2. モデルの妥当性の検証
   （Validation）
3. モデルの提示
   （Presentation）

10

解析手順

1. **生存分析を用いたモデルの作成（Development）**
2. モデルの妥当性の検証（Validation）
3. モデルの提示（Presentation）

11

モデル作成用データ

- UCAS (Unruptrued Cerebral Aneurysm Study) Japanによるコホート
- 未破裂動脈瘤を3年間フォローするプロトコル
- 本研究では6606動脈瘤のデータを使用

Survival-time data
（観察期間の開始と終了があり、打ち切りかアウトカム発生で終わる）

12

Data Editor (Browse) - [dtucas]

File Edit View Data Tools

dataid[1] UA02780

Snapshots

	dataid	size	dau	sex	sah	smoking	age	oc	day	censorbyrx	htn	loc
1	UA02780	6	0	0	0	0	64	0	1060	0	0	ICA
2	UA03753	3	0	0	0	1	45	0	8	1	0	ICA
3	UA03795	3	0	0	0	1	54	0	204	0	0	ICA
4	UA07982	3	0	0	0	1	53	0	63	1	0	ICA
5	UA01685	3	0	0	0	1	61	0	96	1	0	ICA
6	UA07854	5	0	0	0	0	69	0	1163	0	0	ICA
7	UA06482	4	0	0	0	0	38	0	876	0	0	ICA
8	UA02916	5	0	0	0	0	65	0	467	1	0	ICA
9	UA07022	5	0	0	0	1	50	0	36	1	0	ICA
10	UA04399	4	0	0	0	0	42	0	1086	0	0	ICA
11	UA03012	3	0	0	0	0	58	0	1099	0	0	ICA
12	UA02349	4	0	0	0	0	53	0	3117	0	0	ICA
13	UA00305	3	0	0	0	0	39	0	1095	0	0	ICA
14	UA05167	3	0	0	1	0	56	0	1152	0	0	ICA
15	UA04866	4	0	0	0	0	42	0	1096	0	0	ICA
16	UA05474	4	0	0	0	1	48	0	20	1	0	ICA
17	UA03785	4	0	0	0	1	45	0	94	0	0	ICA
18	UA07758	5	0	0	0	1	37	0	67	1	0	ICA
19	UA07184	3	0	0	0	0	50	0	1100	0	0	ICA
20	UA05238	3	0	0	0	0	43	0	1098	0	0	ICA
21	UA06158	3	0	0	0	1	57	0	27	1	0	ICA
22	UA04940	4	0	0	0	0	50	0	18	1	0	ICA
23	UA05912	5	0	0	0	0	60	0	1029	0	0	ICA
24	UA06158	3	0	0	0	1	57	0	1171	0	0	ICA
25	UA04111	6	0	0	0	1	54	0	35	1	0	ICA
26	UA03705	3	0	0	0	1	64	0	1116	0	0	ICA
27	UA05362	6	0	0	0	0	35	0	22	1	0	ICA
28	UA04130	4	0	0	0	0	67	0	157	0	0	ICA

13

// Declare data to be survival-time data

stset day ,failure(oc) scale(365.25)

day…観察日数
oc…アウトカム
oc=1: 破裂
oc=0: 打ち切り
（治療、破裂以外による死亡、
最後のフォローアップ）

14

予測因子

患者の属性

- Age (years old)
- Sex (male/female)
- Smoking (yes/no)
- Hypertension (yes/no)
- Diabetes (yes/no)
- History of SAH (yes/no)
- Number of aneurysms (yes/no)

動脈瘤の特性

- Size (mm)
- Location
  (ACA, ACOM, MCA, ICA, IC-PCOM, BA, VA)
- Daughter sac* (yes/no)

  (*Irregular protrusion of aneurysm wall)

15

Cox比例ハザード回帰による Risk estimateの計算式

Risk estimate = $1 - S(t)^{\exp(\beta X)}$

S(t) = Baseline survivor function when all covariates are zero

β = Regression coefficients

X = Individual values of the risk factors

※**β**、**X**はベクトル　$\beta X = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i$

(Sullivan 2004; Iasonos 2008)

16

モデルの作成手順

1. 欠測値への対応（省略） TRIPOD Item 9
2. 予測因子のCoding（省略） TRIPOD Item 10a
   - 連続変数のカテゴリ化など
3. 予測因子の選択 TRIPOD Item 10b
4. 回帰分析でβを計算 TRIPOD Item 10b
5. 回帰分析の前提の確認（省略）
   - Schoenfeld residuals, log-log plotなど
6. S(t)を計算

17

// Stepwise selection

xi: sw stcox age70 sex htn dm smoking sah lsize/*
*/ (i.loc) dau num ,pr(0.2) nohr

_t	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age70	.2536673	.200245	1.27	0.205	-.1388056	.6461403
lsize	1.988876	.1642172	12.11	0.000	1.667016	2.310736
sex	.4205425	.2296954	1.83	0.067	-.0296522	.8707372
htn	.2553019	.1986919	1.28	0.199	-.1341271	.644731
loc						
1	1.447078	.4444233	3.26	0.001	.5760246	2.318132
2	.8709635	.4329189	2.01	0.044	.0224581	1.719469
3	1.447933	.4225284	3.43	0.001	.6197922	2.276073
4	1.225519	.4614973	2.66	0.008	.321001	2.130037
5	.3758657	.8044577	0.47	0.640	-1.200842	1.952574
6	.5502358	.8063582	0.68	0.495	-1.030197	2.130669
dau	.3947133	.212769	1.86	0.064	-.0223062	.8117329

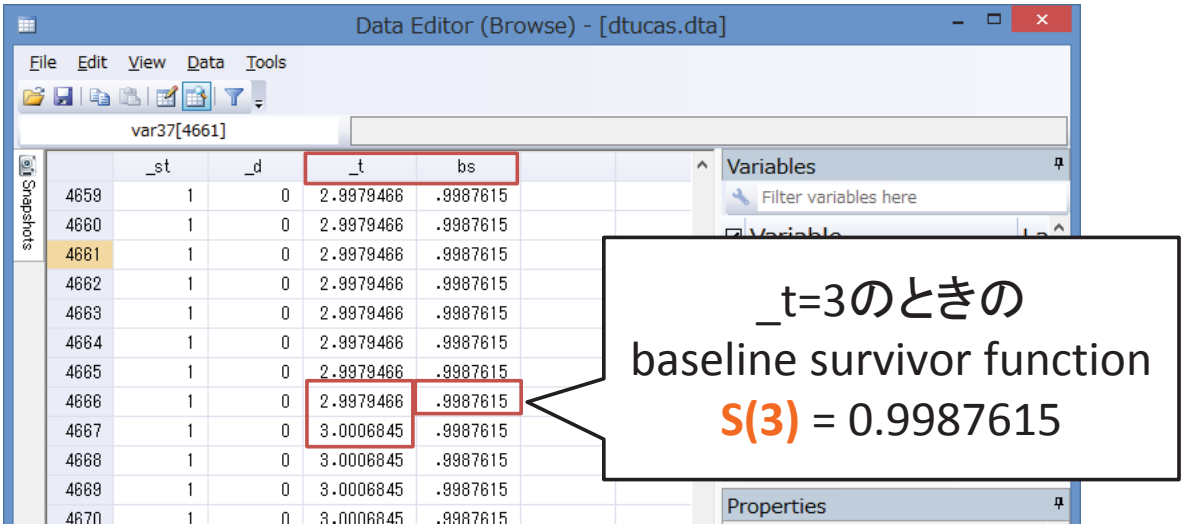
β

18

// Calculate baseline survivor function S(t)

```
stcox age70 lsize sex htn i.loc dau
predict bs ,basesurv
sort _t
```

予測因子が0のときの
baseline survivor functionを計算



	_st	_d	_t	bs
4659	1	0	2.9979466	.9987615
4660	1	0	2.9979466	.9987615
4661	1	0	2.9979466	.9987615
4662	1	0	2.9979466	.9987615
4663	1	0	2.9979466	.9987615
4664	1	0	2.9979466	.9987615
4665	1	0	2.9979466	.9987615
4666	1	0	2.9979466	.9987615
4667	1	0	3.0006845	.9987615
4668	1	0	3.0006845	.9987615
4669	1	0	3.0006845	.9987615
4670	1	0	3.0006845	.9987615

_t=3のときの
baseline survivor function
S(3) = 0.9987615

// Display baseline survivor function (optional)

```
stcox age70 lsize sex htn i.loc dau
predict bs, basesurv
sort _t
local N=_N
forvalues i = 1/`N' {
                    if _t[`i']<=3 {
                                local b3=bs[`i']
                                }
                    }
di "Baseline survival at 3yrs ="`b3'
```

local … マクロを宣言
_N … データ内のサンプル数
var[n] … 変数varのn番目のデータ

作成した予測モデル

3-year rupture risk estimate = $1 - S(3)^{\exp(\beta X)}$

$= 1 - 0.9987615 \wedge \exp$

$(0.254*X_{Age\geq 70} + 0.421*X_{Sex} + 0.255*X_{Hypertension}$
$+ 1.989*X_{log(size)} + 0.871*X_{MCA} + 1.447*X_{ACOM}$
$+ 1.448*X_{IC\text{-}PCOM} + 1.226*X_{BA} + 0.376*X_{VA}$
$+ 0.550*X_{ACA} + 0.395*X_{Daugter\ sac})$

解析手順

1. 生存分析を用いたモデルの作成（Development)
2. **モデルの妥当性の検証（Validation）**
3. モデルの提示（Presentation)

TRIPOD Item 10d

Validationの種類(1)

Internal validation

- モデル作成に使用したデータを、
  モデルがどのくらい正確に予想するか

External validation

- モデル作成に使用したのとは別の外部データを、
  モデルがどのくらい正確に予測するか

23

TRIPOD Item 10d

Validationの種類(2)

Discrimination

- Harrell's C index (Harrell 1982)

Calibration

- Calibration plot

  Calibration is preferably reported graphically with predicted outcome probabilities (on the *x*-axis) plotted against observed outcome frequencies (on the *y*-axis).
  This plot is commonly done by tenths of the predicted risk.

  (TRIPOD)

24

// Discrimination (Harrell's C index) on derivation data

stcox age70 lsize sex htn i.loc dau

estat concordance

```
. estat concordance

       failure _d:  oc
 analysis time _t:  day/365.25
               id:  dummyid

Harrell's C concordance statistic

Number of subjects (N)                =      6606
Number of comparison pairs (P)        =    383220
Number of orderings as expected (E)   =    311958
Number of tied predictions (T)        =       795

       Harrell's C = (E + T/2) / P =    .8151
                           Somers' D =    .6302
```


// Confidence interval of C index by bootstrapping

stcox age70 lsize sex htn i.loc dau
estimate store A

```
program define b_ci, rclass
estimate restore A
estat concordance
return scalar harrellc=r(C)
end
```

r(harrellc)に結果を返すコマンドb_ciを作成

bs c_index=r(harrellc), reps(100): b_ci

// Discrimination (Harrell's C index) on external data

stcox sex htn age70 lsize i.loc dau

use dtext ,clear
estimate esample:
estat concordance

UCASとは異なる1661動脈瘤の
survival-time data

27

// Draw calibration plot - 1

stcox age70 lsize sex htn i.loc dau

predict lp ,xb
xtile ptile=lp ,nq(10)
tabstat lp ,by(ptile) stat(mean) save

xbオプション…
linear prediction（βXに相当）を計算

saveオプション…
行列r()に結果を保存

28

```
. tabstat lp ,by(ptile) save

Summary for variables: lp
     by categories of: ptile (10 quantiles of lp)
```

ptile	mean
1	.6373263
2	1.280797
3	1.622222
4	1.921806
5	2.213566
6	2.507699
7	2.828371
8	3.214074
9	3.691053
10	4.677921
Total	2.451787

```
. return list

macros:
             r(name10) : "10"
              r(name9) : "9"
              r(name8) : "8"
              r(name7) : "7"
              r(name6) : "6"
              r(name5) : "5"
              r(name4) : "4"
              r(name3) : "3"
              r(name2) : "2"
              r(name1) : "1"

matrices:
             r(Stat10) :  1 x 1
              r(Stat9) :  1 x 1
              r(Stat8) :  1 x 1
              r(Stat7) :  1 x 1
              r(Stat6) :  1 x 1
              r(Stat5) :  1 x 1
              r(Stat4) :  1 x 1
              r(Stat3) :  1 x 1
              r(Stat2) :  1 x 1
              r(Stat1) :  1 x 1
          r(StatTotal) :  1 x 1
```


// Draw calibration plot - 2

matrix define r=/*

/r(Stat1)¥r(Stat2)¥r(Stat3)¥r(Stat4)¥r(Stat5)¥/

*/r(Stat6)¥r(Stat7)¥r(Stat8)¥r(Stat9)¥r(Stat10)

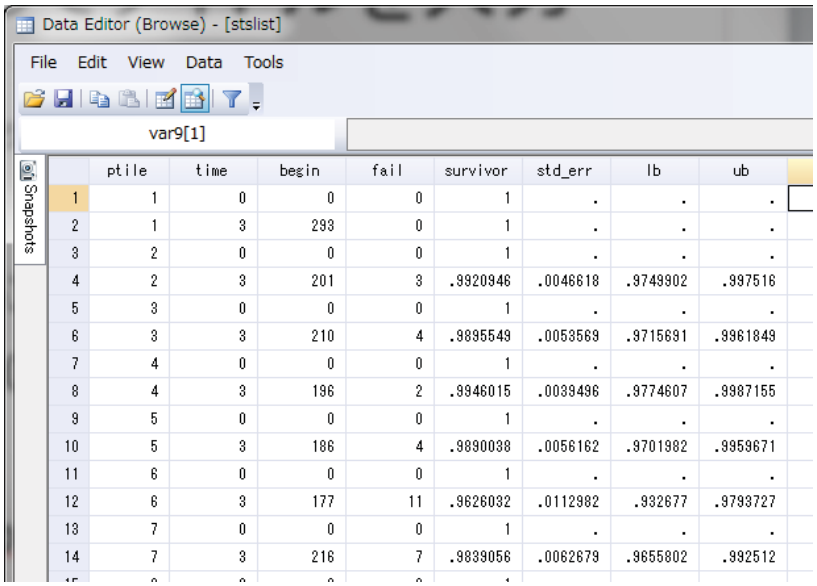
tabstatの結果を
行列rにまとめた

```
. mat list r

r[10,1]
                lp
mean   .63732627
mean    1.280797
mean   1.6222221
mean   1.9218062
mean   2.2135661
mean   2.5076993
mean    2.828371
mean   3.2140737
mean   3.6910528
mean    4.677921
```

// Draw calibration plot - 3

sts list ,at (0 3) by(ptile) saving(“stslist" ,replace)

stslist.dta
の中身



	ptile	time	begin	fail	survivor	std_err	lb	ub
1	1	0	0	0	1	.	.	.
2	1	3	293	0	1	.	.	.
3	2	0	0	0	1	.	.	.
4	2	3	201	3	.9920946	.0046618	.9749902	.997516
5	3	0	0	0	1	.	.	.
6	3	3	210	4	.9895549	.0053569	.9715691	.9961849
7	4	0	0	0	1	.	.	.
8	4	3	196	2	.9946015	.0039496	.9774607	.9987155
9	5	0	0	0	1	.	.	.
10	5	3	186	4	.9890038	.0056162	.9701982	.9959671
11	6	0	0	0	1	.	.	.
12	6	3	177	11	.9626032	.0112982	.932677	.9793727
13	7	0	0	0	1	.	.	.
14	7	3	216	7	.9839056	.0062679	.9655802	.992512

31

// Draw calibration plot - 4

use stslist, clear

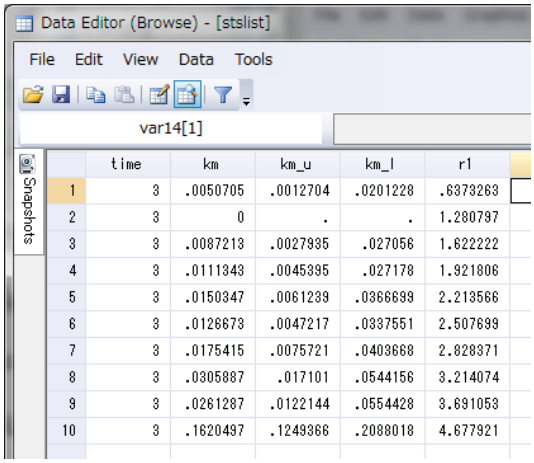
keep if time==3

gen km=(1-survivor)*100

gen km_u=(1-lb)*100

gen km_l=(1-ub)*100

svmat r



	time	km	km_u	km_l	r1
1	3	.0050705	.0012704	.0201228	.6373263
2	3	0	.	.	1.280797
3	3	.0087213	.0027935	.027056	1.622222
4	3	.0111343	.0045395	.027178	1.921806
5	3	.0150347	.0061239	.0366699	2.213566
6	3	.0126673	.0047217	.0337551	2.507699
7	3	.0175415	.0075721	.0403668	2.828371
8	3	.0305887	.017101	.0544156	3.214074
9	3	.0261287	.0122144	.0554428	3.691053
10	3	.1620497	.1249366	.2088018	4.877921

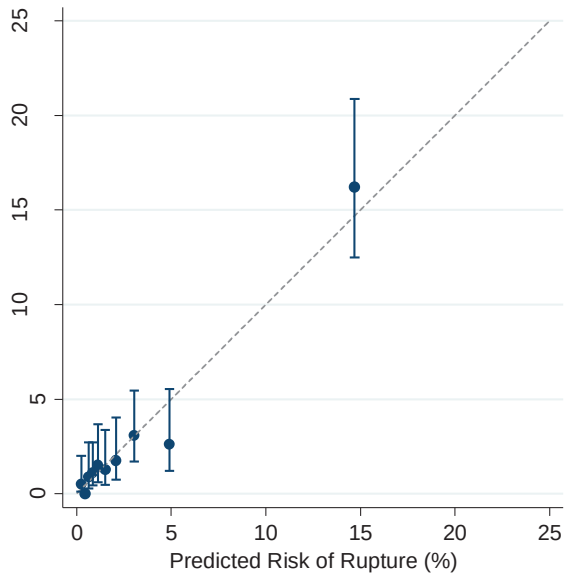
replace r1=(1-0.9987615^exp(r1))*100

$1 - S(3)^{\exp(\beta X)}$

32

// Draw calibration plot - 5

twoway scatter km r1 || rcap km_u km_l r1



33

解析手順

1. 生存分析を用いたモデルの作成（Development）
2. モデルの妥当性の検証（Validation）
3. **モデルの提示（Presentation）**

34

モデルの提示

- スコア表による提示
- 予測因子のβ係数を2倍して四捨五入し、整数値としたものをスコアとした。
- 各動脈瘤のスコアの合計値ごとに、３年間の破裂可能性とその信頼区間を算出。

35

// Scoring - 1

stcox age70 lsize sex htn i.loc dau

matrix scoring = $\frac{e(b)}{\beta}$

回帰分析を行うとβ係数が行列e(b)に格納される（変数名が列名）

```
. mat list e(b)

e(b)[1,12]
                                                          0b.          1.          2.
        age70       lsize         sex         htn         loc         loc         loc
y1  .25366733  1.9888762  .42054254  .25530192           0  1.4470783  .87096352

            3.          4.          5.          6.
           loc         loc         loc         loc         dau
y1  1.4479326   1.225519  .37586566   .5502358  .39471332
```

36

// Scoring - 2

```
local N=colsof(scoring)
forvalues i=1/`N‘{
        matrix scoring[1,`i']=round(scoring[1,`i']*2)
        }
matrix list scoring
```

```
. mat list scoring

scoring[1,12]
                                        0b.       1.       2.       3.       4.       5.       6.
     age70    lsize      sex      htn      loc      loc      loc      loc      loc      loc      loc      dau
y1       1        4        1        1        0        3        2        3        2        1        1        1
```

37

スコア表 (1)

Factor		Score
Age (years)	<70	0
	70≤	1
Sex	Male	0
	Female	1
Hyper-tension	No	0
	Yes	1

Factor		Score
Size (mm)	3≤size<7	0
	7≤size<10	2
	10≤size<20	5
	20≤size	8
Location	ICA	0
	ACA or VA	1
	MCA or BA	2
	ACOM or IC-PCOM	3
Daughter sac	No	0
	Yes	1

38

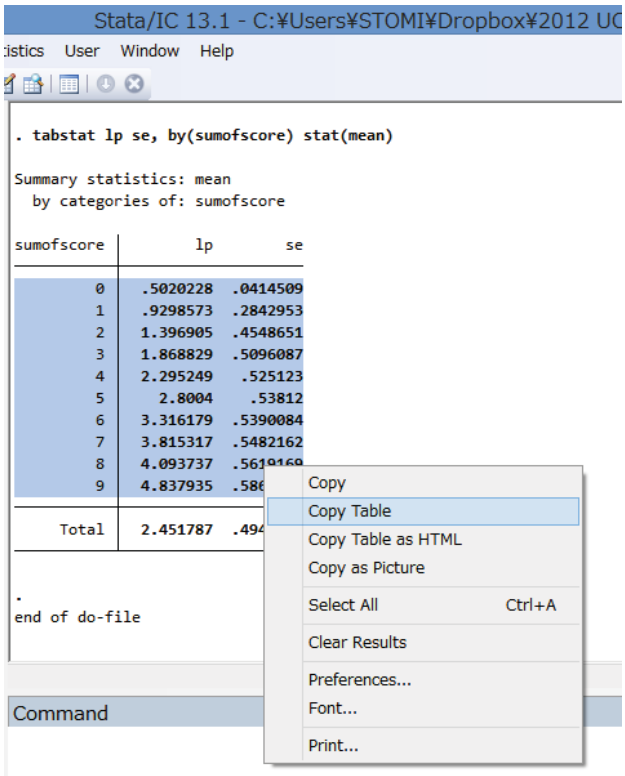
// Scoring - 3

predict lp, xb

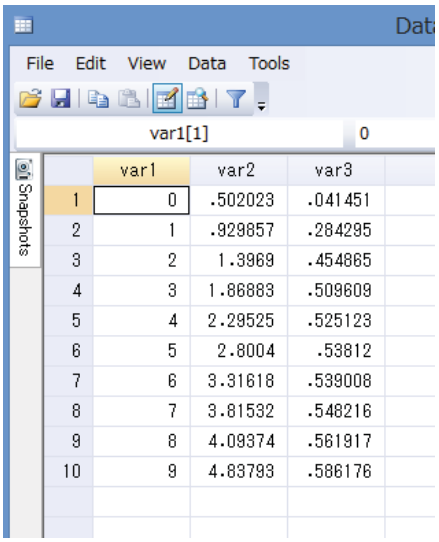
predict se, stdp

matrix score sumofscore = scoring

tabstat lp se, by(sumofscore) stat(mean)



空のDBにペースト



※またはcalibration plotの時のように行列を用いればすべてDo-file内で完結する

// Scoring - 4

```
rename var1 score
gen p3     =(1-.9987615^exp(var2))*100
gen p3_l   =(1-.9987615^exp(var2-1.96*var3))*100
gen p3_u  =(1-.9987615^exp(var2+1.96*var3))*100
list score p3 p3_l p3_u
```

```
. list score p3 p3_l p3_u

     +-------------------------------------------+
     | score        p3       p3_l       p3_u |
     |-------------------------------------------|
  1. |     0   .229037   .204837   .257491 |
  2. |     1   .354509   .211225   .662501 |
  3. |     2   .565161   .242564    1.5114 |
  4. |     3   .908331   .356699   2.43908 |
  5. |     4   1.37825   .514084    3.7597 |
     |-------------------------------------------|
  6. |     5   2.26438    .81713   6.28478 |
  7. |     6    3.6577   1.32714   10.0528 |
  8. |     7    5.7446   2.07445   15.7513 |
  9. |     8   7.59688   2.67429   20.8169 |
 10. |     9   17.2446   6.39048   39.9153 |
     +-------------------------------------------+
```

41

スコア表 (2)

Sum of scores	Risk of rupture in 3 years (%) [95% CI]
0	0.2 [0.2-0.3]
1	0.4 [0.2-0.7]
2	0.6 [0.2-1.5]
3	0.9 [0.2-2.4]
4	1.4 [0.5-3.8]
5	2.3 [0.8-6.3]
6	3.7 [1.3-10]
7	5.7 [2.1-16]
8	7.6 [2.7-21]
9≤	17 [6.4-40]

42

References

- Collins, G.S., et al., Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. Annals of Internal Medicine, 2015. **162**(1): p. 55-63.
- Moons, K.G.M., et al., Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Annals of Internal Medicine, 2015. **162**(1): p. W1-W73.
- Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York, NY: Springer, 2009.

TRIPOD Checklist: Prediction Model Development and Validation

Section/Topic	Item		Checklist item	Page
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.	
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	
	5b	D;V	Describe eligibility criteria for participants.	
	5c	D;V	Give details of treatments received, if relevant.	
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.	
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.	
Sample size	8	D;V	Explain how the study size was arrived at.	
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.	
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	
	10c	V	For validation, describe how the predictions were calculated.	
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.	
Risk groups	11	D;V	Provide details on how risk groups were created, if done.	
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	
Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.	
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	
	15b	D	Explain how to the use the prediction model.	
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.	
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).	
Discussion				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.	
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.	
Other information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.	

*Items relevant only to the development of a prediction model are denoted by D, items relating solely to a validation of a prediction model are denoted by V, and items relating to both are denoted D;V. We recommend using the TRIPOD Checklist in conjunction with the TRIPOD Explanation and Elaboration document.

九州大学の医学分野及び多施設協同臨床研究における Stata の利用状況

九州大学病院メディカル・インフォメーションセンター
徳永章二

2015 Japanese Stata users group meeting
2015/8/28 一橋大学一橋講堂（東京）

九州大学（医学）における Stata

- 九州大学における 医学分野での Stata の利用
  - 3人の Users
  - 分野
- 臨床試験、観察研究と Stata
  - 臨床試験、臨床疫学における統計家の関与
  - 臨床研究での Stata の活用例
    - 統計学的研究デザイン：必要症例数の推定
    - 生存時間解析における必要症例数の推定
    - 観察研究における必要症例数の推定

九州大学（医学分野）での Stata の利用

- 3人の users（現職）
  - 清原千香子
    予防医学分野
  - 錦谷まりこ
    持続可能な社会のための決断科学センター
  - 徳永章二
    ﾒﾃﾞｨｶﾙ･ｲﾝﾌｫﾒｰｼｮﾝｾﾝﾀｰ
- Stata を使った論文は10年間で合計 101 本以上

清原千香子（予防医学）

1）利用している分野
- ゲノム疫学

2）活用している統計解析手法
- logistic regression
- meta-analysis
- ROC analysis etc.

3）Stata を使って書いた論文
最初にSTATAを使ったのが、Kiyohara C, Shirakawa T, Otsu A, Fukuda S, Hopkin JM. Genetic polymorphisms and lung cancer susceptibility: a review. Lung Cancer, 37: 241-256, 2002.
Stata Ver7 です。以来すっとSTATAです。

清原千香子（予防医学）

著書

Kiyohara C, Washio M, Horiuchi T. Chapter 6.Modifying effect of bmoking on the association between SLE and the genetic polymorphisms involved in ROS production. In:A dvances in Genetics Reseach. Volume13. Urbano KV (ed) Nova Science Publishers Inc. NY, pp. 131-153, 2014.

原著（最近のごく一部）

Kiyohara C, Washio M, Horiuchi T, Asami T, Ide S, Atsumi T, Kobashi G, Tada Y, Takahashi H, the Kyushu Sapporo SLE (KYSS) Study Group. Dietary patterns and systemic lupus erythematosus in a Japanese population: the Kyushu Sapporo SLE (KYSS) Study. Int Med J, in press.

Tanaka A, Tsukamoto H, Mitoma H, Kiyohara C, Ueda N, Ayano M, Ohta S, Kimoto Y, Akahoshi M, Arinobu Y, Niiro H, Tada Y, Horiuchi T, Akashi K. Serum progranulin levels are elevated in dermatomyositis patients with acute interstitial lung disease, predicting prognosis. Arthritis Res Therapy, in press.

Oryoji K, Kiyohara C, Horiuchi T, Tsukamoto H, Nakagawa M, Niiro H, Akashi K, Yanase T. Reduced carotid intima-media thickness in systemic lupus erythematosus patients treated with cyclosporine A. Mod Rheumatol, 24:86-92, 2014.

Otake T, Fukumoto J, Abe M, Takemura S, Pham NM, Mizoue T, Kiyohara C. Linking between lifestyle factors and insulin resistance based on fasting plasma insulin and HOMA-IR in middle-aged Japanese men: A cross-sectional study. Scand J Clin Lab Invest, 74: 536-545, 2014.

Kiyohara C, Horiuchi T, Takayama K, Nakanishi Y. Genetic polmorphisms involved in the inflammatory response and lung cancer risk: A case-control study in Japan. Cytokine, 65:88-94, 2014.

Kiyohara C, Washio M, Horiuchi T, Asami T, Ide S, Atsumi T, Kobashi G, Takahashi H, Tada Y, the Kyushu Sapporo SLE (KYSS) Study Group. The modifying effect of NAT2 genotype on the association between systemic lupus erythematosus and consumption of alcohol and caffeine-rich beverages. Arthritis Care Res, 66:1048-1056, 2014

Furukawa M, Kiyohara C, Horiuchi T, Tsukamoto H, Mitoma H, Kimoto Y, Uchino A, Nakagawa M, Oryoji K, Nakashima K, Akashi K, Harada M. Prevalence of and risk factors for vertebral fracture in Japanese female patients with systemic lupus erythematosus. Mod Rheumatol, 23: 765-773, 2013.

Kiyohara C, Miyake Y, Koyanagi M, Fujimoto T, Shirasawa S, Tanaka K, Fukushima W, Sasaki S, Tsuboi Y, Yamada T, Oeda T, Shimada H, Kawamura N, Sakae N, Fukuyama H, Hirota Y, Nagai M for the Fukuoka Kinki Parkinson's Disease Study Group. MDR1 rs1045642 polymorphism and Parkinson's disease in a Japanese population: Interaction with smoking, alcohol consumption and pesticide use. Drug Metab Pharmacokin, 28: 138-143, 2013

Kiyohara C, Washio M, Horiuchi T, Asami T, Ide S , Atsumi T, Kobashi G, Tada Y, Takahashi H and the Kyushu Sapporo SLE (KYSS) Study Group. Cigarette smoking, alcohol consumption and the risk of systemic lupus erythematosus: a case-control study in a Japanese population. J Rheumatol, 39:1363-1370, 2012.

Tanaka A, Tsukamoto H, Mitoma H, Kiyohara C, Ueda N, Ayano M, Ohta S, Inoue Y, Arinobu Y, Niiro H, Horiuchi T, Akashi K. Serum progranulin levels are elevated in patients with systemic lupus erythematosus, reflecting disease activities. Arthritis Res Therapy, 14: R244, 2012.

清原千香子（予防医学）

4）Stata への思い入れ

浮気はしません。

5）その他 Stata の感想等

徳永先生のおすすめの統計解析ソフトでしかもSASに較べて安価なのでとても気に入っています。

Department of Health and Human Services
Centers for Disease Control and Prevention

Human Genome Epidemiology Network (HuGENet™)

アメリカ疾患予防管理センター (Centers for Disease Control and Prevention, CDC) は、ゲノム疫学研究の躍進のためにHuGE reviewの執筆を専門家に依頼している。

Kiyohara C, et al., Genet Med 2005; 7: 463-478.
Kiyohara C, et al., Epidemiol 2006; 17: 89-99.
.

HuGE Reviews

ゲノム疫学研究を効率よく行うためには、HuGENet™により提供される情報 (HuGE Reviewsなど) を上手に活用する必要がある

私はSTATAを活用しています！

清原（予防医学）

ILCCO Project

DNA修復と細胞周期制御に関わる遺伝子と肺がんリスク：プール分析 (NCI R03 grant (CA 119704))（症例=8,454、対照=9,344）

遺伝子多型	major/major	major/minor	minor/minor
APEX1 Asp148Glu	1.0（基準）	0.89 (0.81 - 0.99)	0.91 (0.78 - 1.06)
OGG1 Ser326Cys	1.0（基準）	0.94 (0.84 - 1.07)	1.34 (1.01 - 1.79)
XRCC3 Thr241Met	1.0（基準）	0.89 (0.79 - 0.99)	0.84 (0.71 - 1.00)
ERCC2 Lys751Gln	1.0（基準）	0.99 (0.89 - 1.10)	1.19 (1.02 - 1.39)
TP53 Arg72Pro	1.0（基準）	1.14 (1.00 - 1.29)	1.20 (1.02 - 1.42)

Hung RJ, Kiyohara C, et al. Cancer Epidemiol Biomarkers Prev 2008; 17: 3081-3089.

18の DNA修復関連遺伝子のうち上記5 の遺伝子多型が肺がんリスクと関連

私の仲間もSTATAを活用しています！！！！！

清原（予防医学）

錦谷まりこ (決断科学センター)

Graduate education and research training program in DECISION SCIENCE for a sustainable society 九州大学決断科学大学院プログラム

1）利用している分野

産業保健学、社会疫学（公衆衛生学）、
実験・記録データの解析なども（生理学、環境学）

2）活用している統計解析手法

F検定、t検定、Wilcoxon順位和検定、ペアt検定、Wilcoxon符号付順位和検定、分散分析、共分散分析、
相関分析（ピアソン、スピアマン相関係数など）、
カイ二乗検定、フィッシャーの正確性検定、
重回帰分析、ランダム効果モデルによる最尤法、intrinsic estimatorを用いたAPCモデル（重回帰分析）、
ロジスティック回帰分析、一般化線形モデル、
マルチレベル分析

錦谷まりこ (決断科学センター)

Graduate education and research training program in DECISION SCIENCE for a sustainable society 九州大学決断科学大学院プログラム

3）Stata を使って書いた論文のリスト（過去 10 年位）

- Nishikitani M, Nakao M, Karita K, Nomura K, Yano E. Influence of overtime work, sleep duration, and perceived job characteristics on the physical and mental status of software engineers. *Ind Health* **43**:623-629, 2005.
- Nishikitani M, Yano E. Differences in the lethality of occupational accidents in OECD countries. *Safety Sci* **46**: 1078-1090, 2008.
- Nishikitani M, Inoue S, Yano E. Competition or complement: relationship between judo therapists and physicians for elderly patients with musculoskeletal disease *Environ Health Prev Med* **13**:123-129, 2008.
- Nishikitani M, Tsurugano S, Inoue M, Yano E. Effect of unequal employment status on workers' health: Results from a Japanese national survey. *Soc Sci Med* 75:439-451, 2012.
- Nishikitani M, Nakao M, Tsurugano S, Yano E. The possible absence of a healthy-worker effect: a cross-sectional survey among educated Japanese women. *BMJ Open* **2(5)**:e000958(1-10), 2012.
- Umihara J, Nishikitani M. Emergent use of Twitter in the 2011 Tohoku earthquake. *Prehosp Disaster Med*, 28(5)、1-13, 2013.
- Umihara J, Nishikitani M. Effect of perceived economic status on knowledge about cancer prevention, healthy behaviors, and cancer check-up rate in Japan. *Ningen Dock International,* 1:47-53, 2014.
- Umihara J, Kubota K, Nishikitani M. Rapport between cancer patients and physicians is a critical issue for patient satisfaction with treatment decisions. J Nippon Med School (in press).

錦谷まりこ (決断科学センター)

Graduate education and research training program in DECISION SCIENCE for a sustainable society
九州大学決断科学大学院プログラム

4）Stata への思い入れ

一昔前に比べ日本語の情報も増えて、とても使いやすくなりました。統計学者の個人ブログの他、STATA（本社）の作成している教育動画などもあり、とても助かります。大学院生の時に、version8の英語テキストで独学しましたが、数年後の留学先の大学院では授業で使われており、仲間に会えた！と、さらに理解が深まりました。一応、SPSS、SAS、JMPも使えますが、私にとってはSTATAが最も理解しやすく使いやすかったです。

5）その他 Stata の感想等

長いプログラム（Doファイル）を作って、エラーなく走らせることができると感動します。単純な解析から、複雑な解析へと使用手法の範囲が増えてきましたが、ケースが何万件と多くなり、解析もマルチレベル分析などになると、本当に時間がかかり（何日もかかることもあり）、よく働いているなぁと（STATAとPCの両方に）感謝しています。

徳永章二(ﾒﾃﾞｨｶﾙ･ｲﾝﾌｫﾒｰｼｮﾝｾﾝﾀｰ)

1）利用している場所・分野

九大病院・多施設共同研究（肺がん、消化器がんなど）

臨床試験・観察研究の研究デザイン（症例数推定）

臨床試験・観察研究の統計解析

2）活用している統計解析手法

power analysis (sample size) , Monte Carlo simulation, ralloc
logistic regression, Cox regression,
Wilson confidence interval, Kappa statistics,
random effects linear regression, ROC curve
などなど

徳永章二(ﾒﾃﾞｨｶﾙ･ｲﾝﾌｫﾒｰｼｮﾝｾﾝﾀｰ)

3）Stata を使って書いた論文のリスト（一部）

Shinohara N, Nomura N, Eto M, Kimura G, Minami H, Tokunaga S, Naito S. A randomized multicenter phase II trial on the efficacy of a hydrocolloid dressing containing ceramide with a low-friction external surface for hand-foot skin reaction caused by sorafenib in patients with renal cell carcinoma. Annals of Oncology 2014;25:472-476.

Ideno N, Ohtsuka T, Kono H, Fujiwara K, Oda Y, Aishima S, Ito T, Ishigami K, Tokunaga S, Ohuchida K, Takahata S, Nakamura M, Mizumoto K, Tanaka M.Intraductal Papillary Mucinous Neoplasms of the Pancreas With Distinct Pancreatic Ductal Adenocarcinomas Are Frequently of Gastric Subtype. Ann Surg. 2013 Jul;258(1):141-51.

Takeuchi S, Saeki H, Tokunaga S, Sugaya M, Ohmatsu H, Tsunemi Y, Torii H, Nakamura K, Kawakami T, Soma Y, Gyotoku E, Hide M, Sasaki R, Ohya Y, Kido M, Furue M. A randomized, open-label, multicenter trial of topical Tacrolimus of the treatment of pruritis in patients with atopic dermatitis. Ann Dermatol. 2012 May;24(2):144-50.

Tsukimori K(*), Tokunaga S(*), Shibata S, Uchi H, Nakayama D, Ishimaru T, Nakano H, Wake N, Yoshimura T, Furue M. (* Equal contribution) Long-term effects of polychlorinated biphenyls and dioxins on pregnancy outcomes in women affected by the Yusho incident. Environmental Health Perspectives. 116:626-630, 2008.

4）Stata への思い入れ

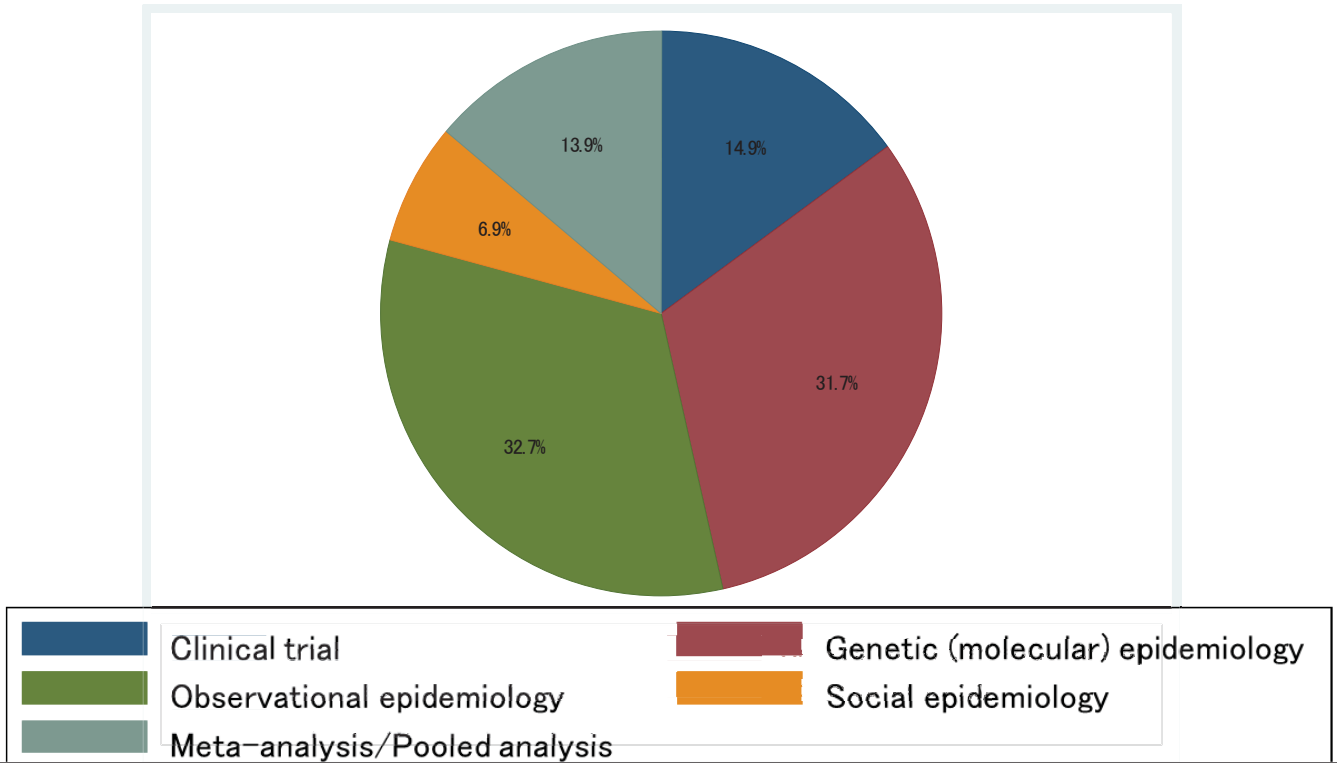
London School of Hygiene and Tropical Medicine の標準統計ソフト

多機能、高機能、安価

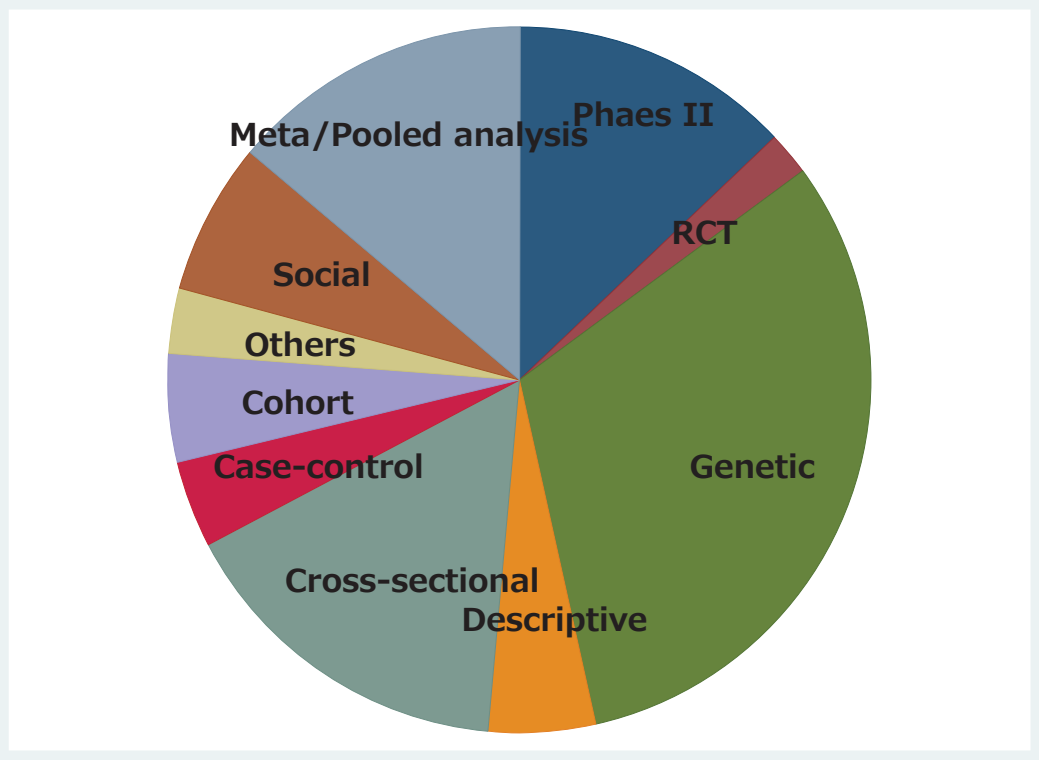
柔軟なプログラミングが可能（マクロの利用、ado file）

使いやすい Help, インターネット機能.....

Stata を利用した論文の分野



Stata を利用した論文の分野



九州大学（医学）における Stata

- 九州大学における 医学分野での Stata の利用
  - ➢3人の Users
  - ➢分野

- 臨床試験、観察研究と Stata
  - ➢臨床試験、臨床疫学における統計家の関与
  - ➢臨床研究での Stata の活用例
    - 統計学的研究デザイン：必要症例数の推定
    - 生存時間解析における必要症例数の推定
    - 観察研究における必要症例数の推定

臨床試験、臨床疫学における統計家の関与

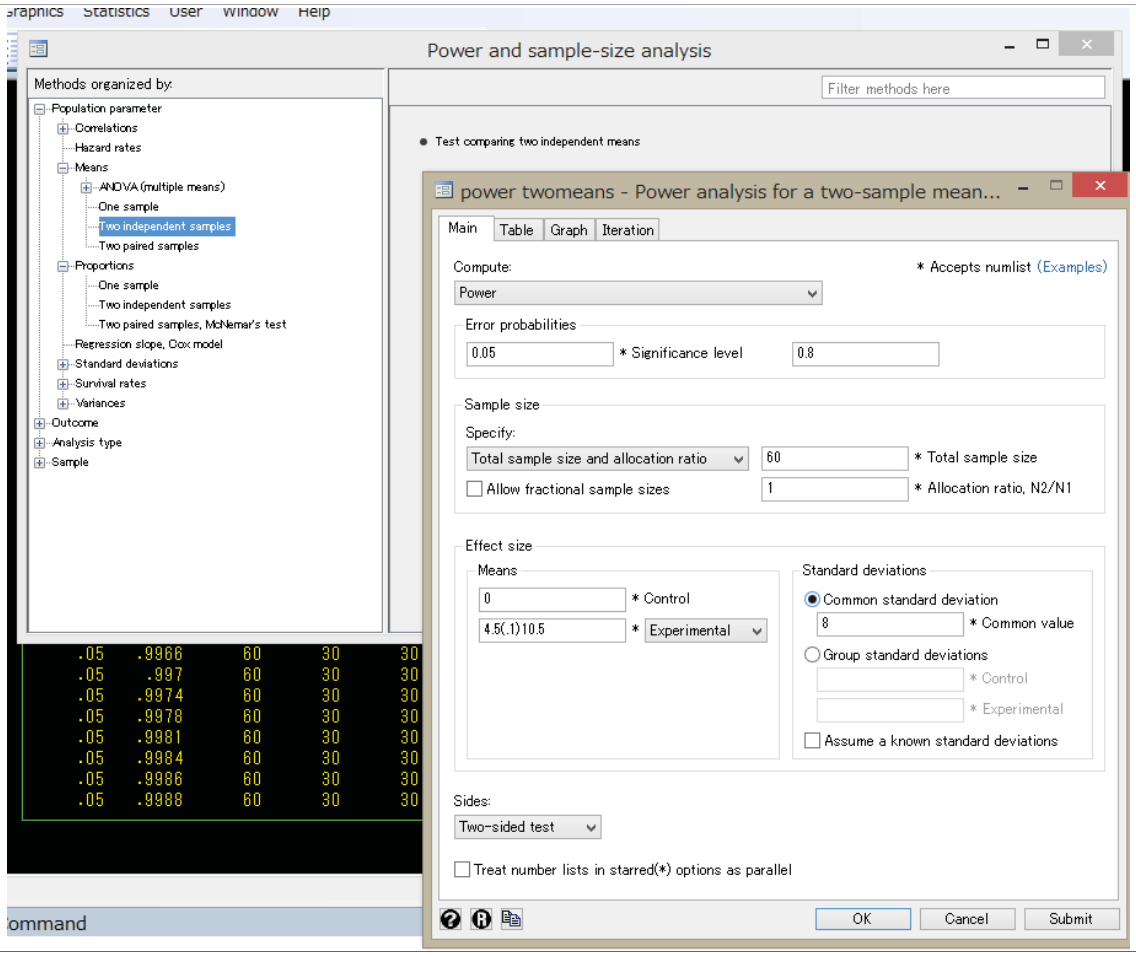
- 一部の研究者に未だにしぶとく生き残る誤解、悪癖
  - 「臨床研究はコンセプトの策定時から統計家と共同作業」という常識が無い
  - データが集まってから統計家に相談する
  - 解析結果を見ながら研究のコンセプトを考える
  - 有意な結果を見るまで手を変え品を変え何回も解析を依頼する
- あるべき姿と現実
  - 研究のコンセプトを考える段階から統計家と相談する
  - 統計家の主な仕事は統計学的デザイン策定、統計解析計画書の作成、論文の一部執筆、査読者への対応
  - 解析、図表作成はﾌﾟﾛｸﾞﾗﾏｰとﾒﾃﾞｨｶﾙﾗｲﾀｰ（に任せたい…）
  - 現状では統計解析や図表作成の多くとﾃﾞｰﾀﾏﾈｰｼﾞﾝｸﾞの一部を統計家が担う（支援人材の不足のため）

臨床研究における Stata の活用例

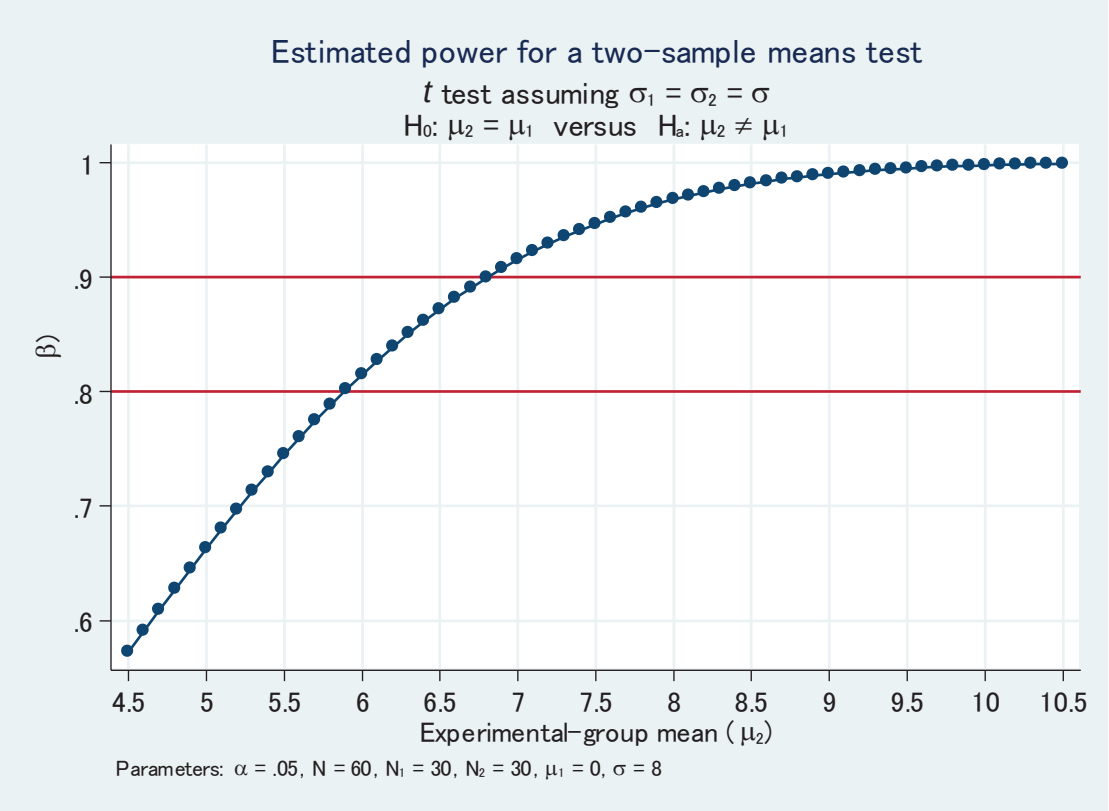
- **統計学的研究デザイン：必要症例数の推定**
  - **ランダム化比較試験**
  - **Phase II study**
- **生存時間解析における必要症例数の推定**
- **観察研究における必要症例数の推定**

Stata による必要症例数の推定（例）

- 連続変量、2群比較
- 症例集積能力は 60 例の見込み
- 見込まれる平均値の差を変化させて検出力を算出する
- power twomeans 0 (4.5(0.1)10.5), sd(7.9) n(60) table graph(ydimension(power) xdimension(m2) plotdimension(sd) yline(.8 .9) xlabel(4.5(.5)10.5) legend(off))



連続変量、2群比較、60症例



Stata による必要症例数の推定（例）

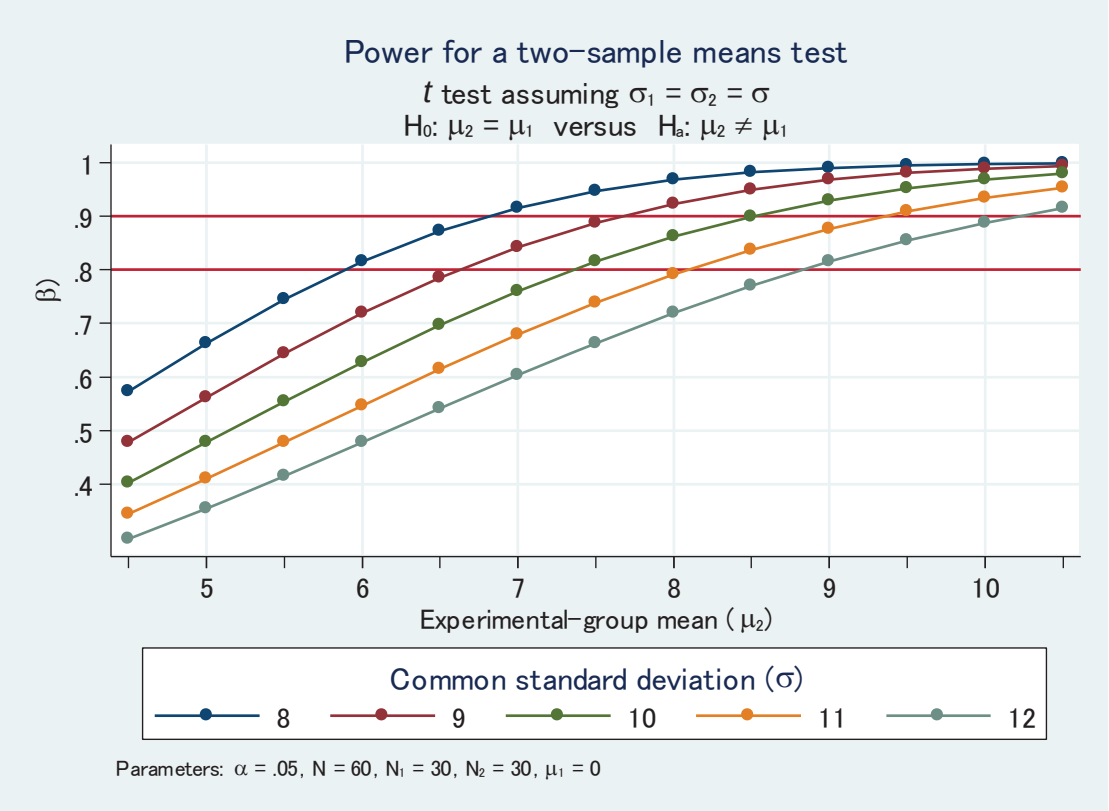
- 連続変量、2群比較
- 症例集積能力は 60 例の見込み
- 見込まれる平均値の差を変化させて検出力を算出する

```
power twomeans 0 (4.5(.1)10.5), sd(8) n(60) table graph(yline(.8 .9)
   xlabel(4.5(.5)10.5))
```

- 異なった SD で検出力を算出する

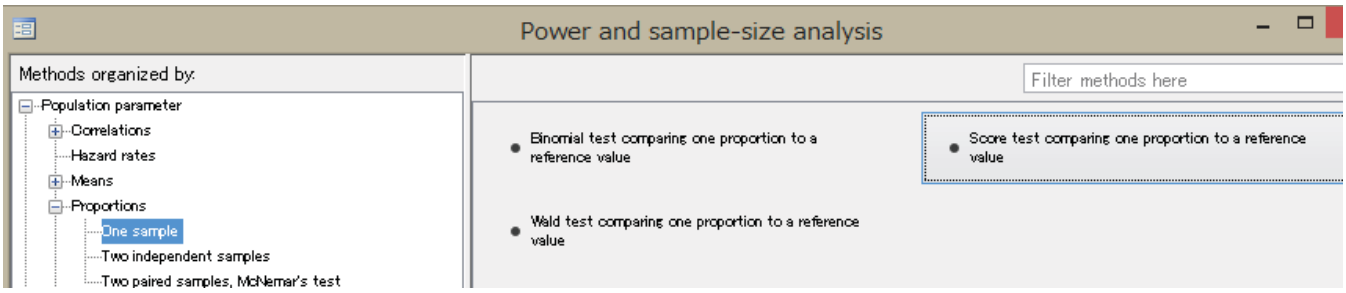
```
power twomeans 0 (4.5(0.5)10.5), sd(8(1)12) n(60) table
   graph(ydimension(power) xdimension(m2) plotdimension(sd)
   ylab(.4(.1)1) yline(.8 .9)  xlabel(5(1)10) xtick(4.5(.5)10.5) legend(on
   rows(1)))
```

連続変量、2群比較、60症例、SDも変化

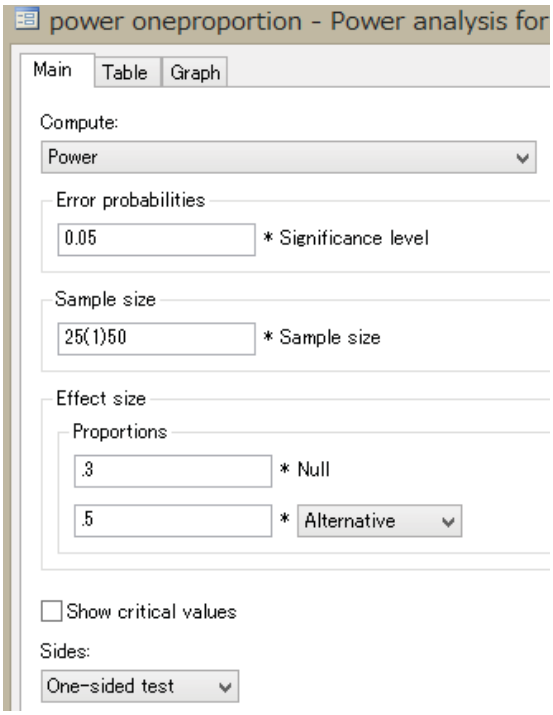


Single-arm phase II study

- 奏効割合、1群、片側検定
- 帰無割合 vs 期待割合
- 症例数を変化させて検出力を算出する
- Binomial test vs. Score test



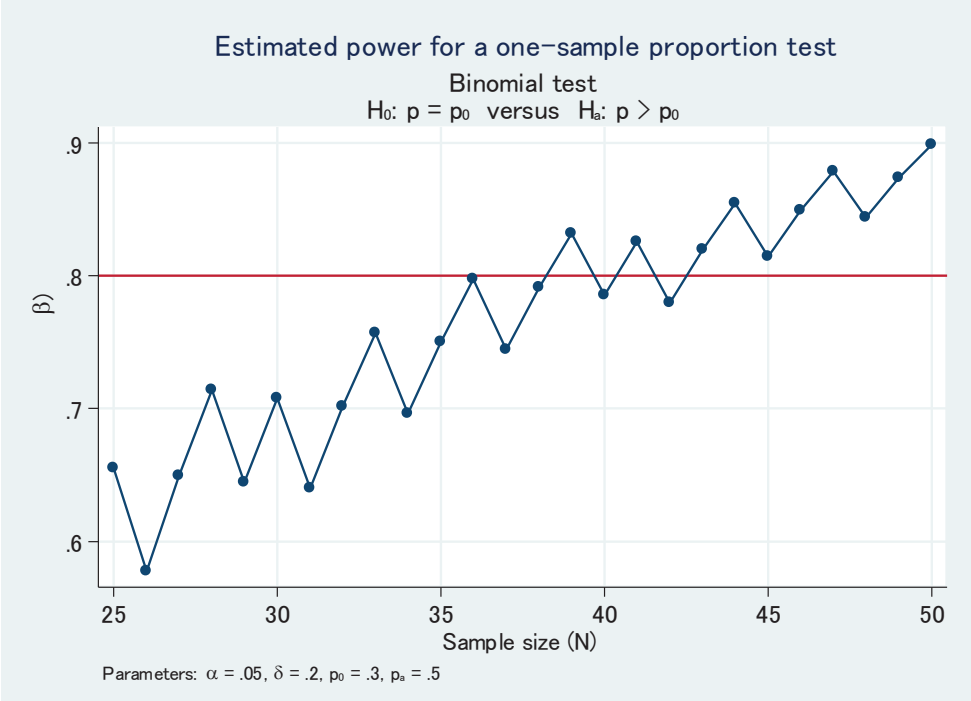
Single-arm phase II study



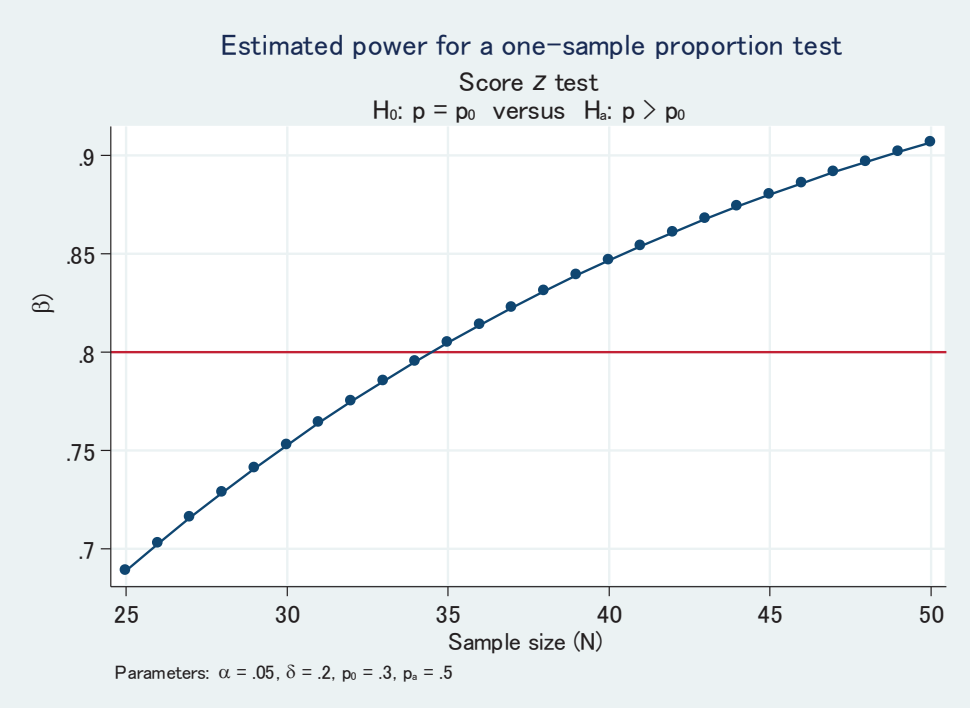
- Binomial test
- 帰無奏効割合 = 0.3
- 期待奏効割合 = 0.5
- 25 ～ 50 症例
- グラフを描く

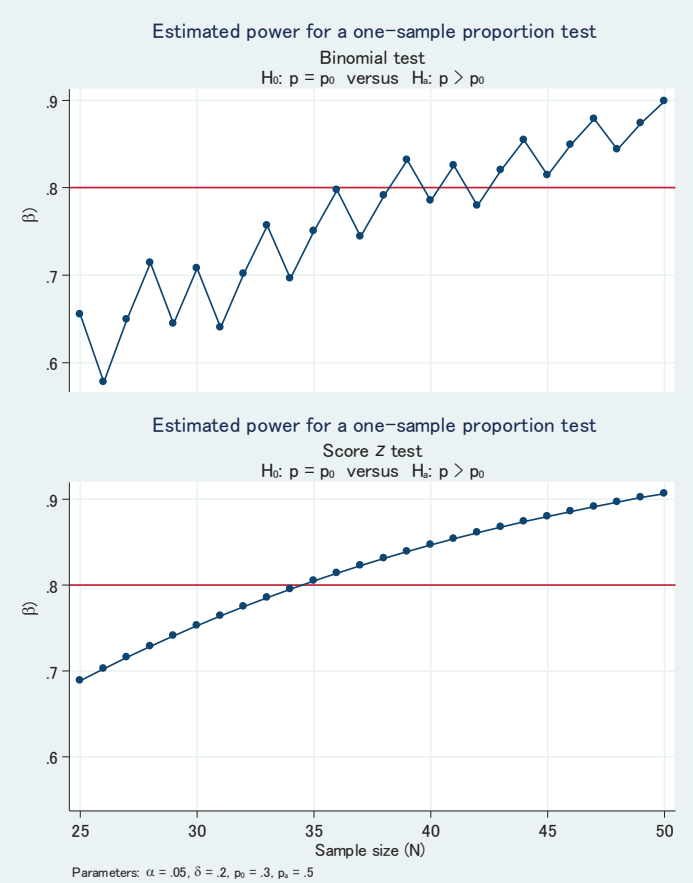
power oneproportion .3 .5, test(binomial) n(25(1)50) onesided table graph(yline(.8)

Single-arm phase II study
Binomial test



Single-arm phase II study
Score test





- Binomial test
- Exact confidence interval (Binomial CI)

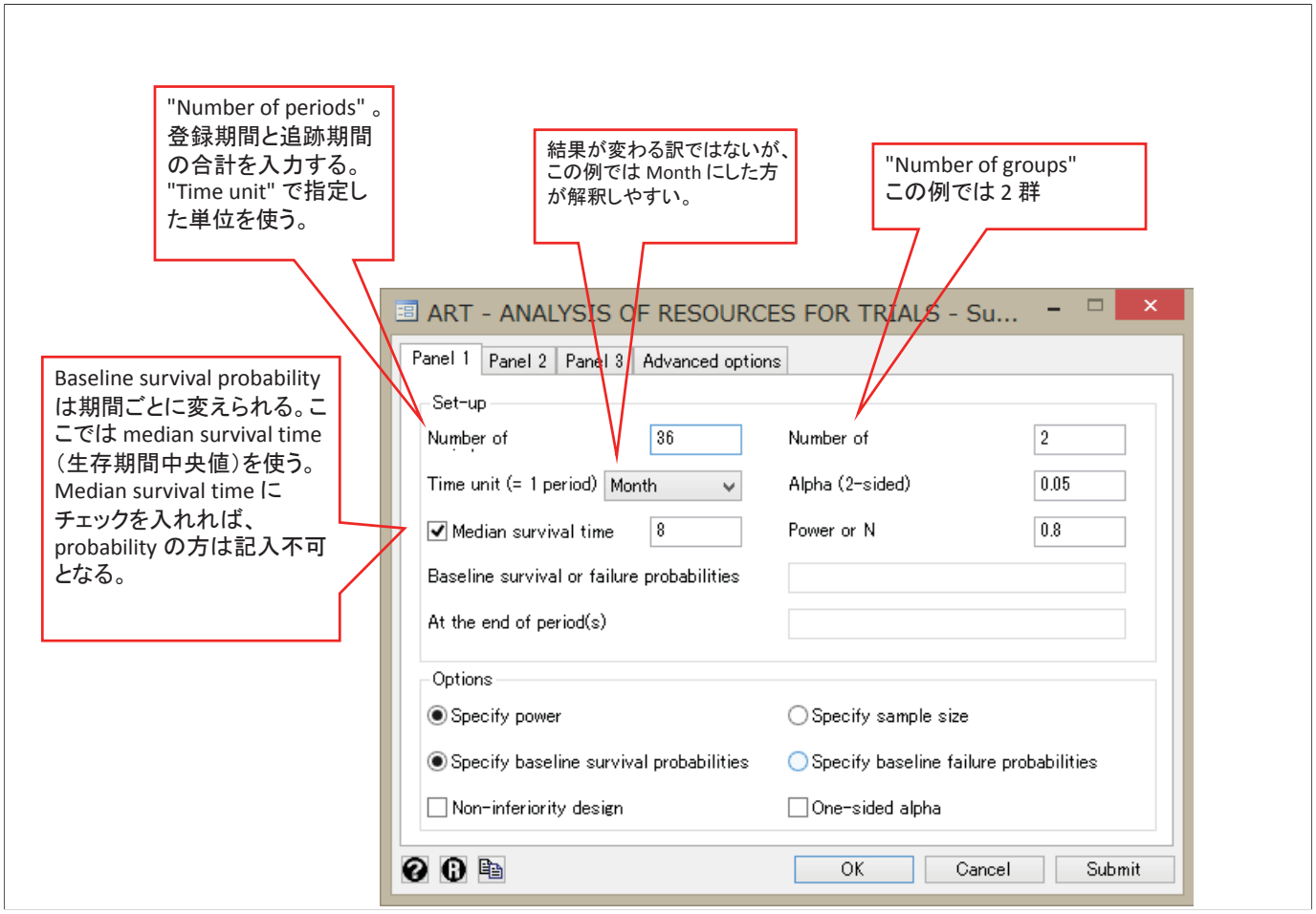
- Score z-test
- Wilson confidence interval

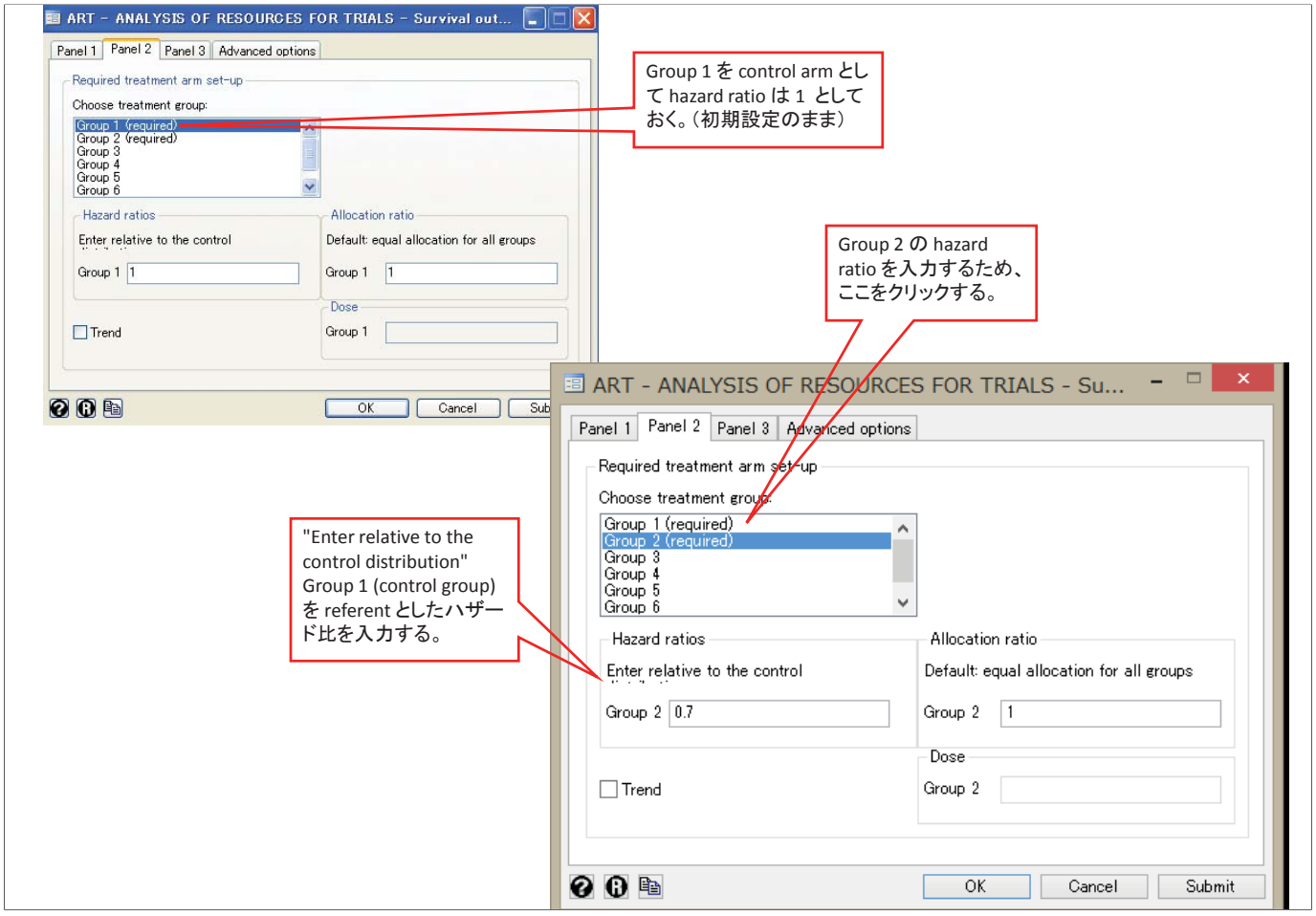
Survival time analysis

- stpower で可能であるが.....必要 event 数の推定
- 現実には、観察期間中に event が起こらない症例が存在する
- 必要 event 数 ≠ 必要登録数
- artsurv.ado
- Barthel, F. M.-S., P. Royston, and A. Babiker. A menu-driven facility for complex sample size calculation in randomized controlled trials with a survival or a binary outcome: Update. Stata Journal 2005;5: 123-129.
- Barthel, F. M.-S., A. Babiker, P. Royston, and M. K. B. Parmar. Evaluation ofsample size and power for multi-arm survival trials allowing for non-uniform accrual, non-proportional hazards, loss to follow-up and cross-over. Statistics in Medicine 2006;25:2521-2542.

artsurv

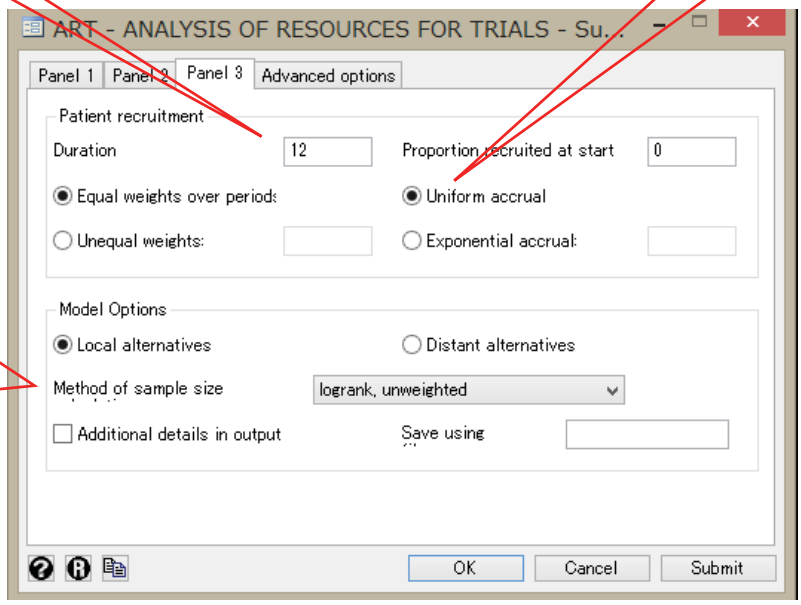
- artsurv は多機能なので artsurv.dlg で対話的に実行させると便利
- artmenu on → user menu に追加される
- 日本語環境では一部のメッセージが読めない（フォントのため？）
- 例として、登録期間1年（12ヶ月）、追跡期間2年 (24ヶ月）、同サイズの2群、優越性試験、Baseline の生存期間中央値 8ヶ月、ハザード比 0.7、登録は 0 人から時間あたり一定の率、両側 α = 0.05、検出力 80%、脱落・クロスオーバー無し、の場合を示す。


"Number of periods" 。登録期間と追跡期間の合計を入力する。"Time unit" で指定した単位を使う。
結果が変わる訳ではないが、この例では Month にした方が解釈しやすい。
"Number of groups" この例では 2 群
Baseline survival probability は期間ごとに変えられる。ここでは median survival time（生存期間中央値）を使う。Median survival time にチェックを入れれば、probability の方は記入不可となる。
ART - ANALYSIS OF RESOURCES FOR TRIALS - Su...
Panel 1
Panel 2
Panel 3
Advanced options
Set-up
Number of
36
Time unit (= 1 period)
Month
Median survival time
8
Baseline survival or failure probabilities
At the end of period(s)
Number of
2
Alpha (2-sided)
0.05
Power or N
0.8
Options
Specify power
Specify sample size
Specify baseline survival probabilities
Specify baseline failure probabilities
Non-inferiority design
One-sided alpha
OK
Cancel
Submit

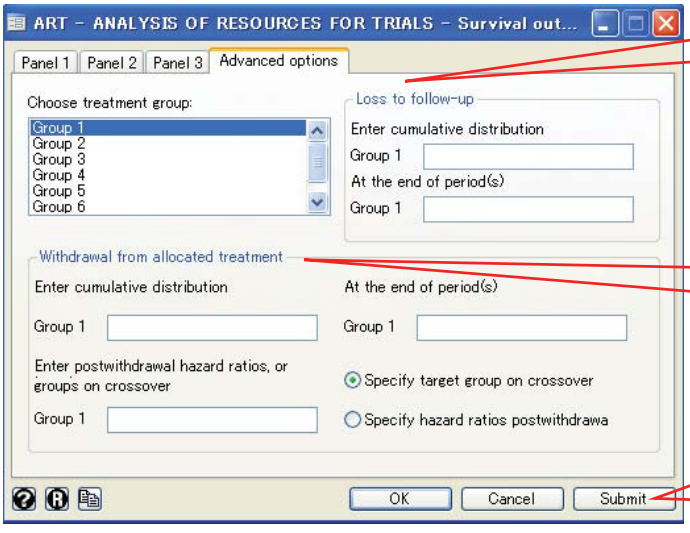

ART - ANALYSIS OF RESOURCES FOR TRIALS - Survival out...
Panel 1
Panel 2
Panel 3
Advanced options
Required treatment arm set-up
Choose treatment group:
Group 1 (required)
Group 2 (required)
Group 3
Group 4
Group 5
Group 6
Hazard ratios
Enter relative to the control
Group 1
1
Trend
Allocation ratio
Default: equal allocation for all groups
Group 1
1
Dose
Group 1
OK
Cancel
Group 1 を control arm として hazard ratio は 1 としておく。（初期設定のまま）
Group 2 の hazard ratio を入力するため、ここをクリックする。
"Enter relative to the control distribution" Group 1 (control group) を referent としたハザード比を入力する。
ART - ANALYSIS OF RESOURCES FOR TRIALS - Su...
Panel 1
Panel 2
Panel 3
Advanced options
Required treatment arm set-up
Choose treatment group:
Group 1 (required)
Group 2 (required)
Group 3
Group 4
Group 5
Group 6
Hazard ratios
Enter relative to the control
Group 2
0.7
Trend
Allocation ratio
Default: equal allocation for all groups
Group 2
1
Dose
Group 2
OK
Cancel
Submit

登録期間を入力する。(Panel 1 を参照のこと)

時間あたり同じ率で登録される場合(初期設定)



"Method of sample size calculation"。デフォルトの logrank 検定が一般的。



Loss to follow-up も詳しく指定できる。

Withdrawal や crossover も詳しく指定できる。

最後に Submit ボタンをクリックすると artsurv が実行される。

```
. artsurv, method(I) nperiod(36) ngroups(2) fp(0) median(8) hratio(1, 0.7) alpha(0.05)
power(0.8) aratios(1 1 ) recrt(12 0, 1, 0 ) distant(0) detail(0) onesided(0) ni(0)
tunit(4) trend(0)

ART - ANALYSIS OF RESOURCES FOR TRIALS        (version 1.0.7, 19 October 2009)
-------------------------------------------------------------------------------
A sample size program by Abdel Babiker, Patrick Royston & Friederike Barthel,
MRC Clinical Trials Unit, London NW1 2DA, UK.
-------------------------------------------------------------------------------
Type of trial                           Superiority - time-to-event outcome
Statistical test assumed                Unweighted logrank test (local)
Number of groups                        2
Allocation ratio                        Equal group sizes

Total number of periods                 36
Length of each period                   One month

Baseline median survival time           8 months
Survival probs per period (group 1)     0.917 0.841 0.771 0.707 0.648 0.595
                    中略
Survival probs per period (group 2)     0.941 0.886 0.834 0.785 0.738 0.695
                    中略

Number of recruitment periods           12
Number of follow-up periods             24
Method of accrual                       Uniform
Recruitment period-weights              1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0
                                         0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Hazard ratios as entered (groups 1,2)   1, 0.7
Alpha                                   0.050 (two-sided)
Power (designed)                        0.800

Total sample size (calculated)          282
Expected total number of events         248
-------------------------------------------------------------------------------
2群（両腕）で 282 症例必要と推定された。
```

Observational studies

- 前向きコホート研究 (prospective cohort study) を想定する
- 曝露群間でイベントが観察された時間を比較する（hazard ratio の推定)
- Hazard ratio から必要症例数の推定、あるいは、集積可能な登録症例数で検出可能な hazard ratio の推定
- 算定方法の基本は臨床試験（介入研究）の場合と同じだが、適用に注意すべき点がある

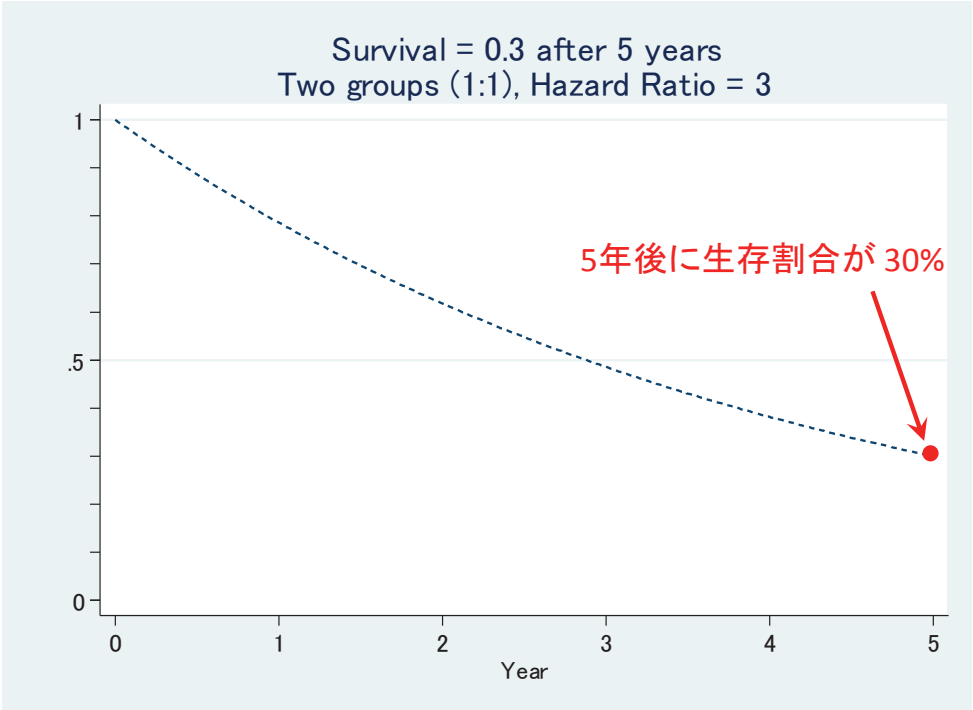
Observational studies

- 臨床試験（介入研究）における必要症例数推定との違い（一般的な疫学研究計画を想定）
  - コホート研究では複数の要因（曝露、背景など）に注目する
  - それぞれの曝露群の集団中の割合 (prevalence) は不明
  - 集団全体の cumulative incidence が報告されている事は多いが....
  - 特定の要因に曝露された群の hazard (incidence) は分からないのが一般的

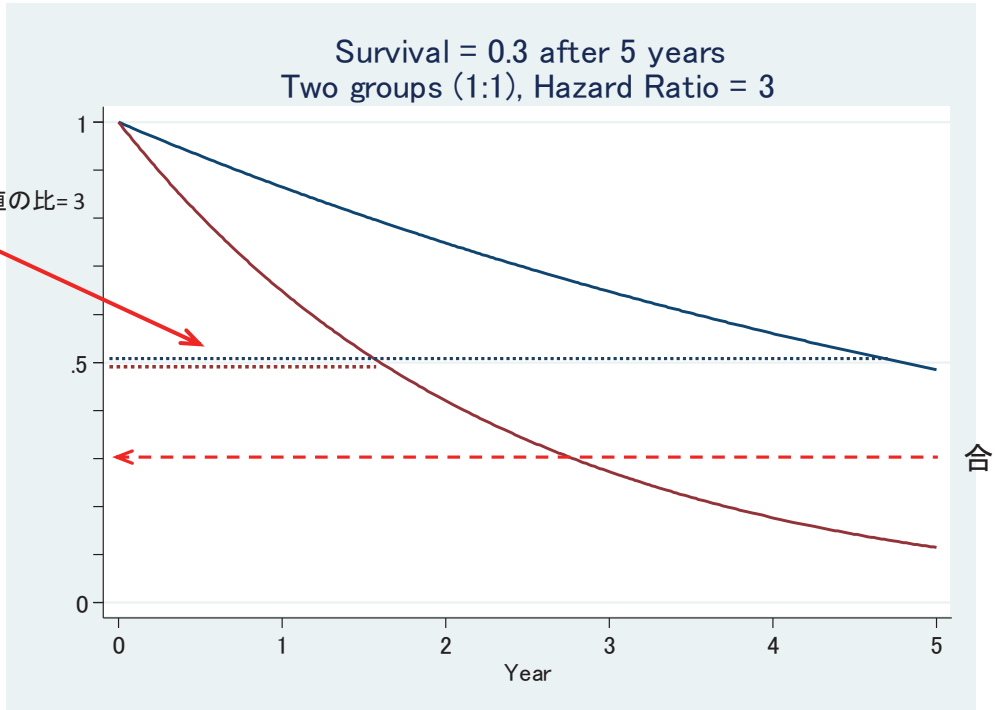
Observational studies

- 臨床試験（介入研究）における必要症例数推定との違い（一般的な疫学研究計画を想定）
  - それぞれの曝露群の集団中の割合 (prevalence) は不明
    → 介入研究の割付比に相当
  - 特定の要因に曝露された群の hazard (incidence) は分からないのが一般的
    → 介入研究で baseline hazard に相当
  - 集団全体の hazard と hazard ratio から baseline hazard を推定する必要がある

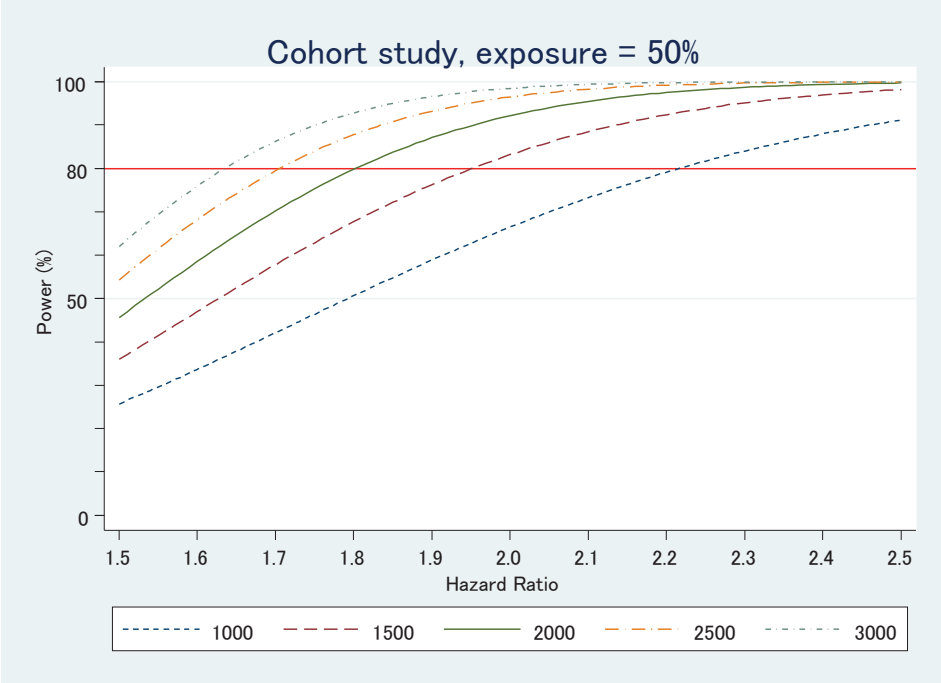
集団全体の hazard と hazard ratio から baseline hazard を推定(例)



集団全体の hazard と hazard ratio から baseline hazard を推定(例)



あるコホート研究でのHR と検出力の関係



まとめ

- 九州大学の医学分野における Stata の利用例を紹介した。
- Stata は疫学の様々な分野や臨床試験、メタアナリシスで活用され、Stata を利用して多くの論文が発表されている。
- 臨床研究における Stata の活用例として、必要症例数の推定を取り上げた。
- Power コマンドによる図の作成は研究者とのコミュニケーションに有用である。
- がんの単腕第Ⅱ相試験、生存時間解析、観察疫学研究における必要症例数の推定方法を説明した。