演習の解答

この文書は『Stata 統計解析ハンドブック』の章末にある演習の解答です．当書籍の原文である『A Gentle Introduction to Stata, Fourth Edition』(Stata 社出版 (2014)) のために，Stata 社が作成し，配布する資料 (英文) に即した内容です．

目次

概要

本解答は，単に正しい数値だけの留めませんでした．多くの設問で，実際にメニューから適切な選択をした画像や，出力された結果への文章による説明を加えました．このため最近，数学や統計学の教科書で演習問題を解いた方は，はじめのうちは本文書に違和感を覚えるかもしれません．統計分析では，特定の相関係数，有意検定結果，回帰係数といった数値を並べるだけでは分析結果を十分報告できません．出てきた数値に対して解釈を与えることは欠かせません．この点を踏まえ，本文書でも出来るだけ多くの記述を掲載するようにしました．

表記法

本文書の表記法は，本体である『Stata 統計解析ハンドブック』の 1.1 節「本書での表記法」と同じ方法を用いています．

新出の機能

演習の解答には，発展学習の意味合いもこめて，本体の書籍では登場しなかった機能が含めています．

データとプログラム

各章の解説の最後には do ファイルのプログラムを掲載しています．

演習で使用するデータは，本体の書籍用のウェブサイトから取得できます (http://www.stata-press.com/data/acock4.html)．本文書に掲載された do ファイルと上手く連動させる場合は，データの保管場所を，Windows 版では `C:\data` とし，Mac 版ではある特定のディレクトリとして，操作の際の作業ディレクトリも同じ場所に設定してください．作業ディレクトリは，Command ウィンドウの下に表示になります．尚，本文書に掲載の do ファイルプログラムは Windows 版用に作成されたものです．

```
/************** Begin do-file ************/
* chapter8.1.do
use "C:\data\gss2006_chapter8.dta", clear
correlate educ hrs1
by sex, sort: correlate educ hrs1
regress hrs1 educ
```

```
by sex, sort: regress hrs1 educ, beta
/************** End do-file **************/
```

ファイルのパス名に関するコマンド文は，作業ディレクトリに応じて変更する必要があります．たとえば，Mac 版をお使いの方は，`Users\acocka` というディレクトリを作業ディレクトリにした場合は，`C:\data\gss2006_chapter8.dta` を `\Users\acocka\gss2006_chapter8.dta` へと，プログラムの記述を置き換えてから実行してください．

Stata での分析は，do ファイルを作成しながら，実行内容を常に記録して行うことを推奨します．これは，ミスの早期発見にも大いに役に立ちます．

アクティブな学習を

演習解答の確認は，自力で解答を導いた後の方が活発な演習となります．解答を眺めるだけでも一定の有益性はありますが，自力で解答を作成した時ほどの効果は望めません．このため，本文書はできるだけすべての演習について解答を記述してあります．どれだけアクティブに取り組むかは読者自身の手に委ねられています．

Windows 版と Mac 版について

本文書は Windows 版 Stata ユーザを念頭に置いています．Mac 版で演習する場合は，若干の読み替えをする必要があります．たとえば，Ctrl+c でのコピーの指示では，Cmnd+c，Page Up は Fn+↑ などです．慣れている方には，問題なく置き換えができる程度の読み替えであると思います．

第 1 章 (1.7 節, pp.22-23) の解答

1. YouTube に投稿された Stata チュートリアルでは，実際に Stata を操作している動画を多数用意しています (https://www.youtube.com/user/statacorp)．ほとんどが高度なトピックを扱ったものですが，この機会にいくつか閲覧して，Stata の雰囲気やトピックの種類を見てみてください．

2. SPSS 経験者にとって，演習に示された，Stata での Results ウィンドウからの Word などの文書ドキュメントへのコピー法は，少し違和感が感じられるものだったかもしれません．Stata では，分析結果が固定長フォントの平文で表示であるのに対して，一般的に文書で用いられる書式は，広い文字（w, m など）と狭い文字（i, f など）の入り混じる可変幅フォントです．可変幅フォントは読みやすい一方，縦のラインを揃えることが難しいフォントです．SPSS では，可変幅フォントでの文字がグラフィックボックスとして文書ドキュメントに挿入されますが，画像を貼り付けたときのように，後から編集できないという制限があります．

   演習に記述したように，Stata の分析結果をコピーする際は，文書ドキュメントのフォントを Courier New などの固定長フォントに設定してから，コピー& ペーストします．コピー以外の部分は可変幅フォントでも構いません．コピーを実行した後で，貼り付けた部分を選択し，フォントを変えるという方法でも構いません．文書エディタによっては全角文字と半角文字の調整を行わないようにする設定が必要な場合もあります．

   貼り付けた部分は，固定長フォントの使用の他に，フォントサイズを小さくしたり，文書の余白を狭めたりする必要が出る場合が多くあります．通常，フォントサイズを 8 ポイントにすると首尾よく収まりますが，それでも入りきらない場合は，余白を狭くします．サイズを 9 ポイントか 10 ポイントに留め，その分，余白で調整するのも一つの方法です．編集上の変更を加えて，大きいサイズのフォントのままにするのも一つの方法です．

   ヒント：Stata 内の表示フォントは変更できます．お使いのディスプレイの大きさや解像度に応じて調整するとよいでしょう．たとえば，Results ウィンドウのフォントを変更する場合は，ウィンドウ内で右クリックをしてフォントを変更します．使用可能なフォントはいくつかありますが，いずれも固定長フォントです．フォントの変更で，分析結果が見やすくなることもあると思います．do ファイルでも同様にフォントの変更ができます．do ファイルで変更する場合には，右クリックの後，Preferences...（ユーザ設定...）を選択します．

3. 演習に記載の操作に続き,Chapter 25 - Working with categorical data and factor variables まで下へスクロールし，censusfv.dta の右にある use をクリックすると，データセットが読み込まれます．既に別のデータセットを開いていた場合は，事前に clear を実行する必要があるかもしれません．この演習で問題とするのは divorcert です．コマンドの実行結果は以下のようになると思います．

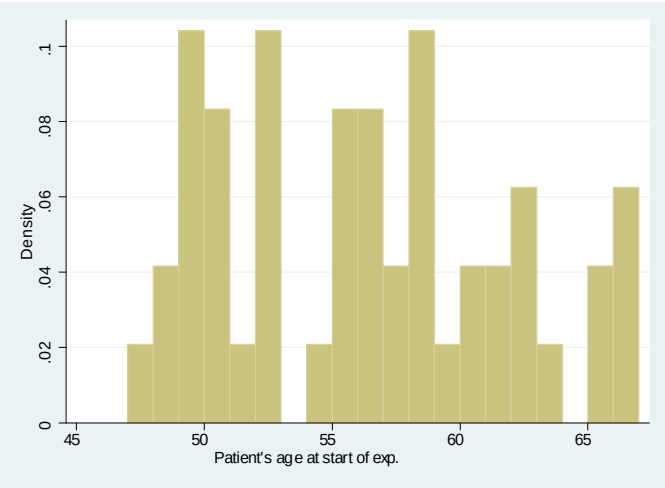
```
. describe

Contains data from http://www.stata-press.com/data/r13/censusfv.dta
  obs:            50                          1980 Census data by state
 vars:            16                          26 Jun 2012 18:08
 size:         3,500
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
statestr        str14   %-14s                 State
state2          str2    %-2s                  Two-letter state abbreviation
region          int     %-8.0g     cenreg     Census region
pop             long    %12.0gc               Population
poplt5          long    %12.0gc               Pop, < 5 year
pop5_17         long    %12.0gc               Pop, 5 to 17 years
pop18p          long    %12.0gc               Pop, 18 and older
pop65p          long    %12.0gc               Pop, 65 and older
popurban        long    %12.0gc               Urban population
medage          float   %9.2f                 Median age
death           long    %12.0gc               Number of deaths
marriage        long    %12.0gc               Number of marriages
divorce         long    %12.0gc               Number of divorces
state           long    %13.0g     st         State
marriagert      long    %12.0g                Marriages per 100,000
divorcert       long    %12.0g                Marriages per 100,000
-------------------------------------------------------------------------------
Sorted by:

. summarize
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    statestr |          0
      state2 |          0
      region |         50        2.66    1.061574          1          4
         pop |         50     4518149     4715038     401851   2.37e+07
      poplt5 |         50    326277.8    331585.1      35998    1708400
-------------+---------------------------------------------------------
     pop5_17 |         50    945951.6    959372.8      91796    4680558
      pop18p |         50     3245920     3430531     271106   1.73e+07
      pop65p |         50    509502.8    538932.4      11547    2414250
    popurban |         50     3328253     4090178     172735   2.16e+07
      medage |         50       29.54    1.693445       24.2       34.7
-------------+---------------------------------------------------------
       death |         50    39474.26    41742.35       1604     186428
    marriage |         50     47701.4    45130.42       4437     210864
     divorce |         50    23679.44    25094.01       2142     133541
```

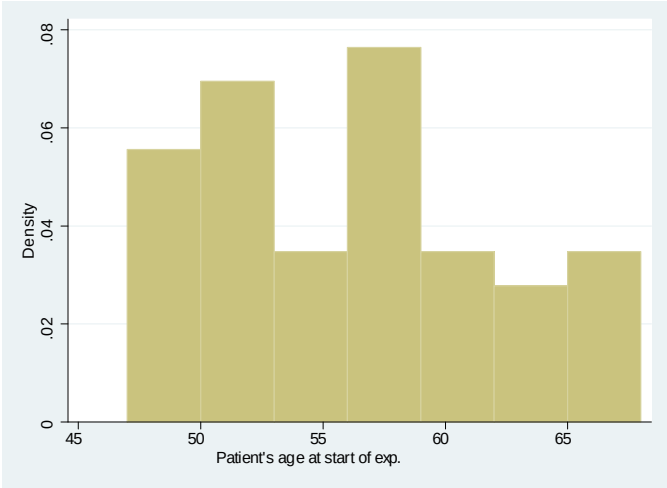
state	50	25.5	14.57738	1	50
marriagert	50	1331.72	1881.18	746	14282
divorcert	50	565.9	224.7591	294	1729

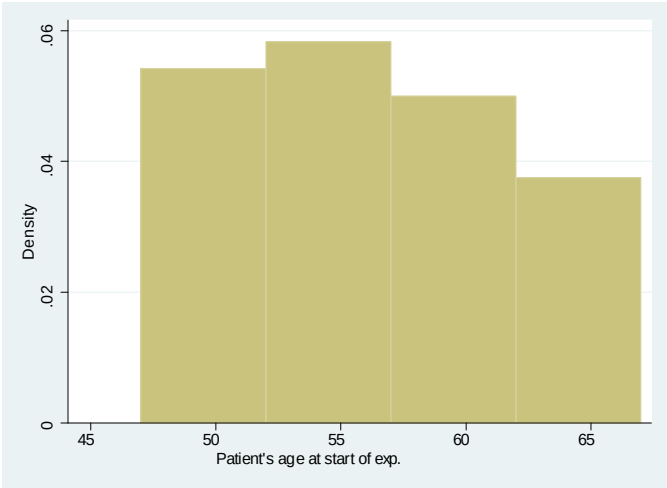
describe コマンドの結果から，変数 divorcert が 10 万人当たりの離婚件数であると分かります．変数の値が 1 万人当たり，10 万人当たり，100 万人当たりのどれなのかによって，summarize コマンドの結果に表示される平均値や標準偏差の解釈が変わってくるため，変数ラベルに記述を残すことは大切です．summarize の結果で，divorcert に関する Mean 列の値は 565.9 です．従って，全米 50 州の平均離婚件数は，10 万人当たり 565.9 件となります．

4. Stata をインストールすると，いくつかのデータセットも一緒に自動的にインストールされます．cancer.dta もそのうちの一つです．これを読み込むには，単純に sysuse cancer を実行します．

ヒント：Stata のインストール時には，PDF の Stata マニュアルもまたインストールされます．Stata マニュアルを，Help ▷ PDF Documentation から開き，“Example datasets” というキーワードで検索した先のリンクをクリックすると，このインストールされるデータセットの一覧を見ることができます．これらのデータセットは sysuse コマンドで開くことができます．
ヒストグラムは次のようになります．







ヒストグラムの作成法についての詳細は，本体書籍の 15-21 ページを参照してください．

グラフを文書ドキュメントにコピーする場合は，たとえば，Stata が表示したグラフの上で右クリックをして Copy（コピー）を選択し，その後，ドキュメントの任意の位置へと貼り付けします．場合によって，他の画像と同じような方法でサイズ変更をします．

ヒストグラムの瓶の幅を広げるに従い，分布は一様分布へと近づきます．瓶幅を 1 年にすると，特にデータが小規模なときなど，必要以上に変動が見えるようになります．一方で，瓶幅を 5

年にすると，十分に変動を示せないかもしれません．最終的に，瓶幅の選択は重要であり，選択次第では分布の大切な特徴を誇大にも見えなくもします．分布の見え方が，調査の実施者の誤った選択によっていかに歪められてしまう可能性があるかを，この例はよく示しています．

5. 動画の視聴は有用です．数章を読み終えた後に改めて見直すと，さらに得られるものがあることと思います．動画によっては，進行が非常に速かったり，後半の章にある知識をベースにしたりします．一旦，全章を読み終えれば，こうした動画の視聴がよい復習となり，さらなる発展的な内容を吸収していけることと思います．

第 1 章 (1.7 節, pp.22-23) の do-file

演習 1.3

```
/************** Begin do-file ************/
* chapter1.3.do
clear
use http://www.stata-press.com/data/r13/censusfv.dta
describe
summarize
/************** End do-file **************/
```

演習 1.4

```
/************** Begin do-file ************/
* chapter1.4.do
clear
sysuse cancer
describe
summarize
histogram age
histogram age, width(1)
histogram age, width(3)
histogram age, width(5)
/************** End do-file **************/
```

第 2 章 (2.10 節, pp.53-54) の解答

1. 1 番目と 4 番目のデータについて，コードを文章に戻すと以下のようになります．

```
Observation 1:
1 2 15 4 5 4 2
```

1 は ID 番号です．つまり，このデータが最初の観測データです．
2 は性別です．表 2.1 から，この回答者が女性であると分かります．
15 は回答者が今までに教育を受けた合計年数です．
4 は回答者が居住する州の公立学校を Good と評価していることを示します．
5 は自分の通った公立学校を Very good と評価していることを示します．
4 は懲役刑を Too harsh と評価していることを示します．
2 は回答者が Liberal であることを示します．

```
Observation 4:
4 1 8 -9 1 -9 5
```

4 は ID 番号です．このデータは 4 番目の観測データです．
1 は性別です．表 2.1 に照らし合わせると，この回答者は男性です．
8 は回答者が今までに教育を受けた合計年数です．
-9 は回答者が居住する州の公立学校の評価について，回答しなかったことを示します．
1 は自分の通った公立学校を Very poor と評価していることを示します．
-9 は回答者が懲役刑の評価について，無回答だったことを示します．
5 は回答者が Very conservative であることを示します．

2. この演習のポイントは Variables Manager を開いた後，いずれかの変数の上で右クリックをし，Manage Notes for Dataset...（データセットのメモを管理...）を選択することです．開いたダイアログボックス（*Notes for Data*）で Add（追加）をクリックし，メモを手入力し，Submit（適用）をクリックします．すると Result ウィンドウに以下のコマンドが表れることと思います．

```
. notes _dta: This is the dataset created for chapter 2 of AGIS, 4th edition
```

Command ウィンドウにカーソルを移動させ，Page up を押すと，このコマンドを表示できます．さらに Enter を押すと，同じ文が書かれた 2 つ目の注記 (メモ) が追加になります．

注記の削除も，同様の操作で行います．*Note for Data* ダイアログボックスで削除する注記を選択し，Delete を押します．

1 か月振りにデータセットを開いたときなど，`notes` とだけコマンド入力することで，内容の説明を参照できます．これがデータセットに注記を残す利点です．

3. Variables Manager で各変数に注記 (メモ) を添付すると，Result ウィンドウには以下のようなコマンドが表れます．

```
. notes conserv: A higher score is more conservative
```

その後，`notes` コマンドを実行すると，変数に適用した注記の一覧が表示になります．注記からは，変数の変更履歴など，単なる変数名，変数ラベル，値ラベルの利用では得られない詳細な情報を獲得できます．

4. コマンドは `codebook, compact` です．実行すると，次のような結果が得られます．

```
. codebook, compact
Variable   Obs Unique   Mean  Min  Max  Label
------------------------------------------------------------------------
id          20     20   10.5    1   20  Respondent´s identification
gender      20      2    1.5    1    2  Participant´s gender
education   20     10  14.45    8   20  Years of education
sch_st      20      5    2.2   -9    5  Ratings of schools in your state
sch_com     20      5    3.5    1    5  Ratings of schools in your community...
prison      20      6   1.35   -9    5  Ratings of prison sentences
conserv     20      6   2.55   -9    5  Conservatism/liberalism
------------------------------------------------------------------------
```

変数 `gender` では，観測数が 20 であり，値の種類が 2 種類です．平均が 1.5，最小値が 1，最大値が 2 です．“Participant’s gender”(参加者の性別) という変数ラベルが添えられています．変数 `education` では，やはり観測数が 20 ですが，値の種類は 10 種類です．値の範囲が 8 から 20 までの自然数で，平均は 14.45 です．変数 `conserv` では，観測数 20，値の種類は 6 です．平均が 2.35，最小値が 5，最大値が 9 です．ところで，このデータセットの変数名やコードとしての値，変数ラベルには，改善の余地があります．たとえば，`conserv` の変数ラベルは，“Conservatism from 1 to 5 being more conservative”(保守傾向の度合いを 1 から 5 までで評価 (5 が最大)) とした方が情報としてより有用です．いくつかの変数の最小値に見られる-9 という値は，欠損値に変換して平均の計算からは除外するべきでしょう．変数名には大きい値が示すものを用いるという原則に立てば，変数 `gender` は `female` と変更した方がいいかもしれません．同変数で，

最大値である 2 は女性です．同様の原則を変数 `conserv` について検証すると，たしかに最大値 5 が最も強い保守傾向を表しています．

5. 各自

6. `codebook`，`notes` を実行すると，`codebook` 出力が注記 (メモ) 付きで表示になります．

7. 簡単なアンケートの結果を管理する場合において，2 ステップに分かれた Stata の値ラベル作成プロセスは無駄に長いのではと感じるかもしれません．値ラベルを名称つきで定義した後，さらに変数への適用が必要です．しかし，規模が大きくなれば，yes/no で答える質問が 50 個に上る場合などが表れるため，この分かれたプロセスの有用性が実感できるようになります．yes に 1 を，no に 2 を割り当てる定義作業を一度行えば，残るはその値ラベルを 50 個の変数に適用する作業だけになります．

第 3 章 (3.9 節, pp.81-82) の解答

1. はじめに，本体書籍の p.vii にある手順に従って，書籍用のデータセットをすべてダウンロードしてください．データセットを開くにはメニューから File ▷ Open...（ファイル (F) ▷ 開く...(O)）を選択し，開きたいファイル `relate.dta` を選択します．ほかにも，以下のコマンドを実行する方法があります．

```
. use "C:\data\relate.dta", clear
```

データセットのダウンロード先は，書籍にある場所以外でも構いません．ただし，その場合は上記のコマンドでも調整が必要です．作業ディレクトリは Stata 画面の最下段に表示があります．

ヒント：上記コマンドには, 二重引用符があります．二重引用符が必要になるのは, My Documents や My Documents\Stata Data など，パス名にスペースがあるときだけです．しかし，それ以外の場合に二重引用符をつけても，エラーにはなりません．従って，二重引用符を常につけるようにしたのが簡単だと言えます．

次に，目的の変数に対し `tabulate` を実行します．実行結果は以下のようになります．

```
. tabulate R3828700
SYMBOL!KEY! |
   SEX 1999 |      Freq.     Percent        Cum.
------------+-----------------------------------
         -5 |        775        8.63        8.63
          1 |      4,170       46.42       55.04
          2 |      4,039       44.96      100.00
------------+-----------------------------------
      Total |      8,984      100.00
```

結果のように，変数には内容を説明したラベルがありません．まず，-5 という値が，性別の情報が欠損していることを Stata に入力しなければなりません．`mvdecode` で，-5 を欠損値へ変換します．

```
. mvdecode R3828700, mv(-5 = .)
   R3828700: 775 missing values generated
```

次に，男性を 1，女性を 2 とする値ラベルを定義します．

```
. label define sex 1 "male" 2 "female"
```

定義した値ラベルを変数 `R3828700` へ適用します．

```
. label values R3828700 sex

. sjlog clse
clse invalid
r(198);
```

これで，tabulate の結果がラベルと共に分かりやすい表示になりました．

```
. tabulate R3828700
SYMBOL!KEY!
   SEX 1999 |      Freq.     Percent        Cum.
------------+-----------------------------------
       male |      4,170       50.80       50.80
     female |      4,039       49.20      100.00
------------+-----------------------------------
      Total |      8,209      100.00
```

表でパーセント数の合計は 100 になるよう計算されます．性別に関する情報が存在しているデータだけについて言えば，上記の結果は的確です．表に欠損値の度数を含めたいときは，tabulate コマンドで missing オプションを指定します．実行結果は，次のようになります．

```
. tabulate R3828700, missing
SYMBOL!KEY!
   SEX 1999 |      Freq.     Percent        Cum.
------------+-----------------------------------
       male |      4,170       46.42       46.42
     female |      4,039       44.96       91.37
          . |        775        8.63      100.00
------------+-----------------------------------
      Total |      8,984      100.00
```

表に欠損値を含めると，パーセント数も欠損値を考慮した値で計算されます．標本中の男性の割合を求めたい場合，欠損値を含めずに計算します．

2. 本体書籍のデータは，以下の URL からダウンロードできます．
http://www.stata-press.com/data/agis4/
ダウンロードしたデータから relate.dct を開きます．relate.dct は平文 (ASCII) で書かれたテキストファイルなので，メモ帳などで閲覧できます．Microsoft の Word からも閲覧することが可能ですが，そこで改めて保存すると，編集中に見えることのないデータも一緒に保存されます．こうなると，Stata からファイルを正しく読むことができませんので注意してください．
dct ファイルを開くと，変数に関する情報があり，続いてデータセットとしての生データがあるのが分かります．次のような行を見てください．

```
R0000100 %4f "PUBID - YTH ID CODE 1997"
```

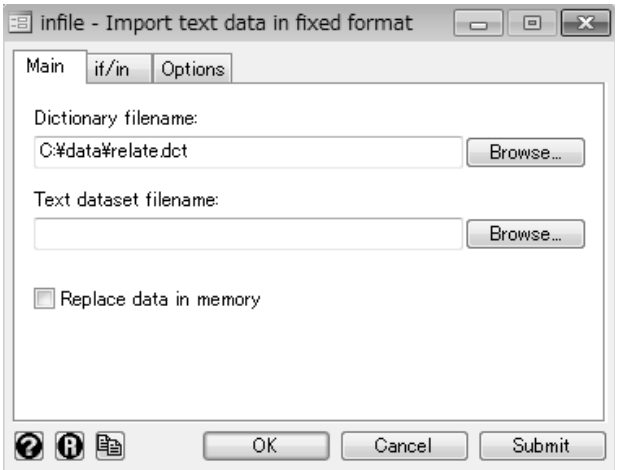
最初の R0000100 は変数名です．次の%4f はフォーマット情報で，生データの最初から 4 桁分の

固定長に変数値が埋め込まれていることを意味します．次の“PUBID - YTH ID CODE 1997”は変数ラベルです．

次に，生データの最初の行を見てください．

```
1-4-4-4-4-4-4-4-418 2
```

読み取りづらいかもしれませんが，この行の最初には 3 桁のスペースがあり，次に数字の 1 があります．Stata での読み込みは，ここを「最初の観測について ID 番号．値は 1」と認識します．先ほどの変数名と変数ラベルの定義と合わせ，変数名 `R0000100` で，変数ラベル“PUBID - YTH ID CODE 1997”の変数の最初の観測データを 1 と認識します．同様にして，(先ほど見た生データの 1 に続く 2 桁の領域から)2 番目の変数の値を-4 と認識します．2 番目の変数の変数ラベルは“MOTH PRAISES R DOING WELL 1999”と認識します．.dct という拡張子を持つディクショナリファイル (定義付の固定型形式ファイル) からデータセットを構築するには，メニューから File ▷ Import ▷ Text data in fixed format with a dictionary（ファイル (F) ▷ インポート (I) ▷ 定義付の固定型形式ファイル）を選択します．ダイアログボックスでは以下のように入力します．



Submit（適用）をクリックすると，`relate.dta` が作成されます．この段階では，欠損値の作成と値ラベルの適用が行われていないため，まだ保存はしません．このデータセットでは，-1 から-5 までの負の値で欠損値を示しています (p.58-59 参照)．それぞれ別の欠損値で置き換えるよう，以下のコマンドを実行します．

```
. mvdecode _all, mv(-5=.a\-4=.b\-3=.c\-2=.d\-1=.e)
```

次に新たな値ラベル often を定義し (p.62-63 参照)，変数 R3483600 へ適用します (p.63 参照). tabulate コマンドを実行し，その後，欠損値も含めて表示する tabulate コマンドを実行すると，以下のような結果になります.

```
. tabulate R3483600
MOTH PRAISES
R DOING WELL
        1999 |      Freq.     Percent        Cum.
-------------+-----------------------------------
       Never |       118        2.61        2.61
      Rarely |       235        5.20        7.81
   Sometimes |       917       20.30       28.12
     Usually |     1,546       34.23       62.34
      Always |     1,701       37.66      100.00
-------------+-----------------------------------
       Total |     4,517      100.00

. tabulate R3483600, missing
MOTH PRAISES
R DOING WELL
        1999 |      Freq.     Percent        Cum.
-------------+-----------------------------------
       Never |       118        1.31        1.31
      Rarely |       235        2.62        3.93
   Sometimes |       917       10.21       14.14
     Usually |     1,546       17.21       31.34
      Always |     1,701       18.93       50.28
Noninterview |       775        8.63       58.90
  Valid skip |     3,669       40.84       99.74
  Don´t know |         3        0.03       99.78
     Refused |        20        0.22      100.00
-------------+-----------------------------------
       Total |     8,984      100.00
```

3. 演習 2 の結果では，Never，Rarely などのラベルが表示になりますが，対応する数値コードは表示されません．データの確認をするときは，ラベルと数値コードを一緒に表示する方が便利なときがあります．これを行うには，numlabel，add を実行します．実行すると，以降の tabulate では値ラベルと数値コードが一緒に表示されるようになります．この設定を無効にするには，numlabel，remove を実行します．コマンドの実行結果は次のようになります.

```
. numlabel, add
. tabulate R3483600
  MOTH PRAISES R
 DOING WELL 1999 |      Freq.     Percent        Cum.
-----------------+-----------------------------------
```

```
       0. Never |       118        2.61        2.61
      1. Rarely |       235        5.20        7.81
  2. Sometimes  |       917       20.30       28.12
    3. Usually  |     1,546       34.23       62.34
     4. Always  |     1,701       37.66      100.00
----------------+-----------------------------------
          Total |     4,517      100.00
. numlabel, remove
. tabulate R3483600, missing
MOTH PRAISES
R DOING WELL
       1999     |      Freq.     Percent        Cum.
----------------+-----------------------------------
       Never    |       118        1.31        1.31
      Rarely    |       235        2.62        3.93
   Sometimes    |       917       10.21       14.14
     Usually    |     1,546       17.21       31.34
      Always    |     1,701       18.93       50.28
Noninterview    |       775        8.63       58.90
  Valid skip    |     3,669       40.84       99.74
  Don´t know    |         3        0.03       99.78
     Refused    |        20        0.22      100.00
----------------+-----------------------------------
       Total    |     8,984      100.00
```

missing オプションは，欠損値の入力が意図通り行われたかを確認する際に便利です．

4. ((訂正) 本演習については本体書籍の記述に誤植がありました．以下のように訂正するとともに，心からお詫び申し上げます．)

   (訂正文) 4. relate.dta は 3 年に及ぶ調査の 3 年目に回収したデータです．このため，1997 年の調査開始時には 18 歳未満でも，終了時までに 18 歳以上になった青少年からの回答が混じっています．まず，変数 R3828100 について，-5 という値を欠損値に変更してください．その後，1999 年時点で 18 歳以上である青少年の回答データを，行単位で削除してください．R0000100，R3483600，R3483800，R3485200，R3485400，R3828700 を残して，ほかの変数を削除してください．データセットを positive.dta として新たに保存してください．

まずはじめに，relate.dta を読み込みます．メニューで File ▷ Open...（ファイル (F) ▷ 開く...(O)）を選択し，開いたダイアログボックスでファイルを選択するか，もしくは以下のコマンドを実行します．

```
. use "C:\data\relate.dta", clear
```

次に，-5 という値を Stata での欠損値にします．変数 R3828100 には欠損値は 1 種類しかない

ので，以下のようなコマンドで対応できます．

```
. mvdecode R3828100, mv(-5=.)
   R3828100: 775 missing values generated
```

欠損値を変換をしたら，度数分布表で確認します．tabulate コマンドを missing オプション付きで実行すると，以下のように欠損値が.(ドット) で表示になります．

```
. tabulate R3828100, missing
SYMBOL!KEY!
   AGE 1999 |      Freq.     Percent        Cum.
------------+-----------------------------------
         14 |        109        1.21        1.21
         15 |      1,664       18.52       19.74
         16 |      1,632       18.17       37.90
         17 |      1,728       19.23       57.13
         18 |      1,622       18.05       75.19
         19 |      1,387       15.44       90.63
         20 |         67        0.75       91.37
          . |        775        8.63      100.00
------------+-----------------------------------
      Total |      8,984      100.00
```

一部の変数と 18 歳未満の青少年の回答データのみ残すには 2 つの操作を行います．一つは，指定のあった変数を残す操作，もう一つは，18 歳未満の参加者の回答データを残す操作です．本体書籍の p.67-68 に記載のあるダイアログボックスで操作を行い，以下のコマンドを生成します．

```
. keep if R3828100 < 18
(3851 observations deleted)
. keep R0000100 R3483600 R3483800 R3485200 R3485400 R3828700
```

演習では要求されていませんが，こうして抽出したデータセットに注記を添えても良いでしょう．以下は注記の一例です．

```
. notes: positive.dta includes only participants under 18
. describe
Contains data from relate.dta
  obs:         5,133
 vars:             6                          14 Feb 2008 09:39
 size:       123,192                          (_dta has notes)
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
R0000100        float   %9.0g                 PUBID - YTH ID CODE 1997
R3483600        float   %9.0g                 MOTH PRAISES R DOING WELL 1999
R3483800        float   %9.0g                 MOTH HELPS R WITH WHAT IMPT TO R
                                                1999
R3485200        float   %9.0g                 FATH PRAISES R DO WELL 1999
R3485400        float   %9.0g                 FATH HELP WITH WHAT IMPT TO R
```

```
                                                  1999
R3828700        float   %9.0g                 SYMBOL!KEY!SEX 1999
-------------------------------------------------------------------------------
Sorted by:
     Note:  dataset has changed since last saved
. notes list _dta
_dta:
  1.  positive.dta includes only participants under 18
```

最後に，出来上がったデータセットを positive.dta という名前で保存します．

```
. save "C:\data\positive.dta", replace
```

5. 本演習では，演習 4 で作成した positive.dta を使用します．positive.dta を別のフォルダに保存した場合，下記の use コマンドでパス名を調整してください．

大規模な調査では，多くの場合，それぞれの決まりに従って変数名を決めています．たとえば，変数名の 1 文字目を，回答者 (respondent) を示す R や，観測者 (observer) を示す O を最初の文字にしています．2 文字目以降には数字を続けて，アンケートの質問番号などを示しています．こうした変数名は，設問の構成が複雑であったり，全ての変数を使用したりする場合には便利です．しかし，本演習のように，一部の回答のみを使用する場合，R3483600 とするよりは mopraise という変数名にした方が分析がやりやすいと思います．
まずはじめに，データセットを開きます．

```
. use "C:\data\positive.dta", clear
```

次にデータがどのようにコード化されているのか，度数分布表で確認します．tabulate で変数一つ一つを表示する代わりに，tab1(一元配置クロス表) で一度に表示します．

```
. tab1 R3483600 R3483800 R3485200 R3485400, missing
```

コード化の確認には，codebook コマンドも便利です．

```
. codebook R3483600 R3483800 R3485200 R3485400
```

複数の種類の欠損値を再コード化するため，mvdecode を実行します．-1 から-5 までのコードを以下のコマンドで変換します．

```
. mvdecode _all, mv(-5=.a4=.b3=.c2=.d1=.e)
```

既存の変数のコピーを作成する方法には 2 種類あります．ここでは，本体書籍で用いなかった generate を利用して実行します．

```
. generate mopraise = R3483600
. generate mohelp = R3483800
. generate fapraise = R3485200
. generate fahelp = R3485400
. tab1 mopraise mohelp fapraise fahelp, missing
```

ヒント：値ラベルなども含めた完全な形で変数をコピーしたいときは，clonevar を利用できます．ここでは，値ラベルがなかったため，generate と clonevar の両方で同じ結果が得られます．値ラベルがある場合であれば，以下のように clonevar を実行してラベルを含めたコピーをする方が簡単です．

```
. clonevar mopraise = R3483600
. clonevar mohelp = R3483800
. clonevar fapraise = R3485200
. clonevar fahelp = R3485400
. tab1 mopraise mohelp fapraise fahelp, missing
```

6. 本演習では，演習 4 で作成した positive.dta を使用します．positive.dta を別のフォルダに保存した場合，下記の use コマンドでパス名を調整してください．

まず，演習 4 で作成した positive.dta を開きます．

```
. use "C:\data\positive.dta", clear
```

次に，欠損値を確認するため，各変数について tabulate(または目的の全変数について tab1)を実行します．欠損値の変換後は，tab1 を用いて結果を確認します．

```
. tab1 R3483600 R3483800 R3485200 R3485400, missing
-> tabulation of R3483600
       MOTH |
  PRAISES R |
 DOING WELL |
       1999 |      Freq.     Percent        Cum.
------------+-----------------------------------
         -4 |        623       12.14       12.14
         -2 |          3        0.06       12.20
         -1 |         20        0.39       12.59
          0 |        118        2.30       14.88
          1 |        233        4.54       19.42
          2 |        912       17.77       37.19
          3 |      1,533       29.87       67.06
          4 |      1,691       32.94      100.00
------------+-----------------------------------
      Total |      5,133      100.00
-> tabulation of R3483800
 MOTH HELPS |
R WITH WHAT |
  IMPT TO R |
```

```
       1999 |      Freq.     Percent        Cum.
------------+-----------------------------------
         -4 |        623       12.14       12.14
         -2 |          3        0.06       12.20
         -1 |         21        0.41       12.60
          0 |        121        2.36       14.96
          1 |        297        5.79       20.75
          2 |        905       17.63       38.38
          3 |      1,550       30.20       68.58
          4 |      1,613       31.42      100.00
------------+-----------------------------------
      Total |      5,133      100.00

-> tabulation of R3485200

       FATH |
  PRAISES R |
    DO WELL |
       1999 |      Freq.     Percent        Cum.
------------+-----------------------------------
         -4 |      1,764       34.37       34.37
         -2 |          2        0.04       34.40
         -1 |         15        0.29       34.70
          0 |        134        2.61       37.31
          1 |        320        6.23       43.54
          2 |        747       14.55       58.09
          3 |      1,114       21.70       79.80
          4 |      1,037       20.20      100.00
------------+-----------------------------------
      Total |      5,133      100.00

-> tabulation of R3485400

  FATH HELP |
  WITH WHAT |
  IMPT TO R |
       1999 |      Freq.     Percent        Cum.
------------+-----------------------------------
         -4 |      1,764       34.37       34.37
         -2 |          4        0.08       34.44
         -1 |         15        0.29       34.74
          0 |        189        3.68       38.42
          1 |        367        7.15       45.57
          2 |        826       16.09       61.66
          3 |      1,042       20.30       81.96
          4 |        926       18.04      100.00
------------+-----------------------------------
      Total |      5,133      100.00

. mvdecode _all, mv(-5=.a\-4=.b\-3=.c\-2=.d\-1=.e)
    R3483600: 646 missing values generated
    R3483800: 647 missing values generated
    R3485200: 1781 missing values generated
    R3485400: 1783 missing values generated

. tab1 R3483600 R3483800 R3485200 R3485400, missing

-> tabulation of R3483600

       MOTH |
  PRAISES R |
 DOING WELL |
       1999 |      Freq.     Percent        Cum.
```

```
          0 |       118        2.30        2.30
          1 |       233        4.54        6.84
          2 |       912       17.77       24.61
          3 |     1,533       29.87       54.47
          4 |     1,691       32.94       87.41
         .b |       623       12.14       99.55
         .d |         3        0.06       99.61
         .e |        20        0.39      100.00
------------+-----------------------------------
      Total |     5,133      100.00
-> tabulation of R3483800

 MOTH HELPS |
R WITH WHAT |
  IMPT TO R |
       1999 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |       121        2.36        2.36
          1 |       297        5.79        8.14
          2 |       905       17.63       25.77
          3 |     1,550       30.20       55.97
          4 |     1,613       31.42       87.40
         .b |       623       12.14       99.53
         .d |         3        0.06       99.59
         .e |        21        0.41      100.00
------------+-----------------------------------
      Total |     5,133      100.00
-> tabulation of R3485200

       FATH |
  PRAISES R |
    DO WELL |
       1999 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |       134        2.61        2.61
          1 |       320        6.23        8.84
          2 |       747       14.55       23.40
          3 |     1,114       21.70       45.10
          4 |     1,037       20.20       65.30
         .b |     1,764       34.37       99.67
         .d |         2        0.04       99.71
         .e |        15        0.29      100.00
------------+-----------------------------------
      Total |     5,133      100.00
-> tabulation of R3485400

  FATH HELP |
  WITH WHAT |
  IMPT TO R |
       1999 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |       189        3.68        3.68
          1 |       367        7.15       10.83
          2 |       826       16.09       26.92
          3 |     1,042       20.30       47.22
          4 |       926       18.04       65.26
         .b |     1,764       34.37       99.63
```

```
         .d |          4        0.08       99.71
         .e |         15        0.29      100.00
------------+-----------------------------------
      Total |      5,133      100.00
```

両親との関わり合いの度合いについて，男子の平均を計算するため，egen を実行します．平均値は parents_boys という変数に格納し，変数名で男子の平均であることを明確にします．コマンドが 2 行以上に及ぶかもしれませんが，途中で Enter を押さずに，すべて入力し終えてから押してください．

```
. egen parents_boys = rowmean(R3483600 R3483800 R3485200 R3485400) if R3828700
> == 1
(2714 missing values generated)
```

同様に女子についても実行します．

```
. egen parents_girls = rowmean(R3483600 R3483800 R3485200 R3485400) if R3828700
>  == 2
(2874 missing values generated)
```

その後，度数分布表で結果を確認します．

```
. tab1 parents_boys parents_girls, missing
-> tabulation of parents_boys
parents_boy
          s |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         12        0.23        0.23
        .25 |          2        0.04        0.27
         .5 |         15        0.29        0.56
        .75 |          7        0.14        0.70
          1 |         51        0.99        1.69
       1.25 |         27        0.53        2.22
        1.5 |         87        1.69        3.92
       1.75 |         59        1.15        5.07
          2 |        191        3.72        8.79
       2.25 |        135        2.63       11.42
   2.333333 |          1        0.02       11.44
        2.5 |        278        5.42       16.85
   2.666667 |          1        0.02       16.87
       2.75 |        184        3.58       20.46
          3 |        384        7.48       27.94
       3.25 |        145        2.82       30.76
        3.5 |        318        6.20       36.96
   3.666667 |          1        0.02       36.98
       3.75 |        158        3.08       40.05
          4 |        363        7.07       47.13
          . |      2,714       52.87      100.00
------------+-----------------------------------
      Total |      5,133      100.00
-> tabulation of parents_girls
```

parents_girls	Freq.	Percent	Cum.
0	18	0.35	0.35
.25	4	0.08	0.43
.5	20	0.39	0.82
.75	12	0.23	1.05
1	68	1.32	2.38
1.25	43	0.84	3.21
1.5	90	1.75	4.97
1.75	74	1.44	6.41
2	183	3.57	9.97
2.25	112	2.18	12.16
2.333333	1	0.02	12.18
2.5	251	4.89	17.07
2.75	170	3.31	20.38
3	321	6.25	26.63
3.25	148	2.88	29.51
3.5	297	5.79	35.30
3.75	111	2.16	37.46
4	336	6.55	44.01
.	2,874	55.99	100.00
Total	5,133	100.00	

ヒント：`egen` は多彩な機能を持つコマンドです．`help egen` を実行して，開いたウィンドウから機能の一覧を見ることができます (より詳細な情報については『Data Management Reference Manual』参照)．また，さらに多彩な `egenmore` という Stata ユーザが作成したコマンドもあります．`findit egenmore` を実行すると，リンクと詳細の掲載されたウィンドウを表示できます．

第 3 章 (3.9 節, pp.81-82) の do-file

演習 3.1

```
/************** Begin do-file ************/
* chapter3.1.do
clear
use "C:\data\relate.dta", clear
tabulate R3828700
mvdecode R3828700, mv(-5 = .)
label define sex 1 "male" 2 "female"
label values R3828700 sex
tabulate R3828700
tabulate R3828700, missing
/************** End do-file **************/
```

演習 3.2

```
/************** Begin do-file ************/
* chapter3.2.do
clear
infile using "C:\data\relate.dct", clear
mvdecode _all, mv(-5=.a\4=.b\3=.c\2=.d\1=.e)
label define often 0 "Never" 1 "Rarely" 2 "Sometimes" ///
  3 "Usually" 4 "Always" .a "Noninterview" .b "Valid skip" ///
  .c "Invalid skip" .d "Don't know" .e "Refusal"
label values R3483600 often
tabulate R3483600
tabulate R3483600, missing
/************** End do-file **************/
```

演習 3.3

```
/************** Begin do-file ************/
* chapter3.3.do
clear
infile using "C:\data\relate.dct", clear
tabulate R3483600
tabulate R3483600, missing
numlabel, add
tabulate R3483600
numlabel, remove
tabulate R3483600, missing
/************** End do-file **************/
```

演習 3.4

```
/************** Begin do-file ************/
* chapter3.4.do
clear
use "C:\data\relate.dta", clear
mvdecode R3828100, mv(-5=.)
tabulate R3828100, missing
keep if R3828100 < 18
keep R0000100 R3483600 R3483800 R3485200 R3485400 R3828700
notes: positive.dta includes only participants under 18
describe
notes list _dta
save "C:\data\positive.dta", replace
/************** End do-file **************/
```

演習 3.5

```
/************** Begin do-file ************/
* chapter3.5.do
clear
use "C:\data\positive.dta", clear
tab1 R3483600 R3483800 R3485200 R3485400, missing
codebook R3483600 R3483800 R3485200 R3485400
mvdecode _all, mv(-5=.a\4=.b\3=.c\2=.d\1=.e)
generate mopraise = R3483600
generate mohelp = R3483800
generate fapraise = R3485200
generate fahelp = R3485400
tab1 mopraise mohelp fapraise fahelp, missing
clonevar mompraise = R3483600
clonevar momhelp = R3483800
clonevar dadpraise = R3485200
clonevar dadhelp = R3485400
tab1 mompraise momhelp dadpraise dadhelp, missing
/************** End do-file **************/
```

演習 3.6

```
/************** Begin do-file ************/
* chapter3.6.do
clear
use "C:\data\positive.dta", clear
tab1 R3483600 R3483800 R3485200 R3485400, missing
mvdecode _all, mv(-5=.a\4=.b\3=.c\2=.d\1=.e)
tabulate R3828700, m
egen parents_boys = rowmean(R3483600 R3483800 R3485200 R3485400) if R3828700==1
egen parents_girls = rowmean(R3483600 R3483800 R3485200 R3485400) if R3828700==2
tab1 parents_boys parents_girls, missing
/************** End do-file **************/
```

第 4 章 (4.7 節, p.99) の解答

1. まず，演習に記載のある `firstsurvey_chapter4.dta` を開きます．ここでは，ウェブページに置かれたデータセットを取得する方法で開いてみます．この方法を用いれば，データセットをウェブページに上げておくことで，ほかのユーザへの配布が可能になります．以下のようなコマンドを実行して，ウェブページからデータセットを開きます．

```
. use http://www.stata-press.com/data/agis4/firstsurvey_chapter4.dta
```

2. どんなコメントが適切なのか．それは，do ファイルの使用目的によって変わってきます．しかし，ファイルの先頭には，ファイル名をコメントとして記しておくと便利です．もし印刷された do ファイルから，ディスク上の保存場所を検索する場合などに，すぐファイル名が分かります．また，do ファイルの使用目的もコメントでまとめておくと便利です．今後 Stata の使用を継続していくなかで，作業に応じて，同じファイルを再利用したり，改良を加えて保存することがあると思います．

3. 演習 3 の実行結果は，以下のようになります．実行結果を Word 文書に写すときには，実行結果を範囲選択しコピーを行い，Word 文書に貼り付けます．その後フォントの種類を Courier New，サイズを 10 ポイントなどにします．十分なスペースがある場合，もう少し大きいフォントを用いても構いません．

```
. use http://www.stata-press.com/data/agis4/firstsurvey_chapter4, clear
. describe
Contains data from http://www.stata-press.com/data/agis4/firstsurvey_chapter4.d
> ta
  obs:            20
 vars:             7                          15 Feb 2008 09:57
 size:           160
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
id              int     %8.0g                 Identification number
gender          byte    %8.0g      sex        1 if male 2 if female
education       byte    %8.0g                 Years of education
sch_st          byte    %9.0g      approve    Rating of schools in your state
sch_com         byte    %9.0g      approve    Ratings of schools in your
                                                community of origin
prison          byte    %16.0g     length     Rating of prison sentences
conserv         byte    %17.0g     con        Conservativism/liberalism
-------------------------------------------------------------------------------
Sorted by:
```

```
. summarize education, detail
                      Years of education
-------------------------------------------------------------
      Percentiles      Smallest
 1%            8              8
 5%          9.5             11
10%         11.5             12       Obs                  20
25%           12             12       Sum of Wgt.          20
50%         14.5                      Mean              14.45
                        Largest       Std. Dev.      2.946452
75%         16.5             17
90%           18             18       Variance       8.681579
95%           19             18       Skewness      -.1636124
99%           20             20       Kurtosis       2.522208
```

4. 演習 3 の解答に同じです．
5. ログの記録の開始するには，メニューから File ▷ Log ▷ Begin...（ファイル (F) ▷ ログ (L) ▷ 開始...(B)）を選択します．ログの記録を終了するには，File ▷ Log ▷ Close（ファイル (F) ▷ ログ (L) ▷ 終了 (C)）を選択します．ログファイルの保存形式は 2 種類あります．一つは，`.smcl` という拡張子の付く既定の形式で，Stata Markup and Control Language(SMCL) と言います．Stata で閲覧が可能である特別な形式です．Word など，ほかの文書エディタで開くと，この形式の特有の文字列も共に表示され，読みにくいものになってしまいます．Word などの文書エディタで閲覧する場合には，もう一つの形式 (`.log` という拡張子の付く形式) で保存します．こちらは，ASCII の平文で書かれたテキストファイルです．

   ログの記録も do ファイルに含めるには，記録を取りたいコマンド（のまとまり）の前後に，ログの開始と終了のコマンドをそれぞれ挿入します．

   次節にある do ファイルでは，ログファイルを外部記憶装置 (F ドライブ) の所定の場所に保存するプログラムになっています．ログファイルを Word 文書に挿入するには，挿入位置にカーソルを合わせ，メニューから挿入 ▷ ファイルからテキスト...（Insert ▷ File...）を選択します．開いたウィンドウで，ファイルの形式（Files of type）をすべてのファイル (*.*)（All files (*.*)）に設定し，*.log ファイルが表示されるようにします．その後，`4results.log` を選択します．フォントの変更，行数，半角文字と全角文字の文字幅の調整の無効化など，別途必要な場合もあります．表示を整えると，ログは以下のようになります．

```
-------------------------------------------------------------------------------
      name:  <unnamed>
       log:  C:¥Users¥lightstone¥Documents¥stata¥data¥4results.log
  log type:  text
 opened on:  26 Mar 2015, 13:49:16

. describe

Contains data from http://www.stata-press.com/data/agis4/firstsurvey_chapter4.d
> ta
  obs:            20
 vars:             7                          15 Feb 2008 09:57
 size:           160
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
id              int     %8.0g                 Identification number
gender          byte    %8.0g      sex        1 if male 2 if female
education       byte    %8.0g                 Years of education
sch_st          byte    %9.0g      approve    Rating of schools in your state
sch_com         byte    %9.0g      approve    Ratings of schools in your
                                                community of origin
prison          byte    %16.0g     length     Rating of prison sentences
conserv         byte    %17.0g     con        Conservativism/liberalism
-------------------------------------------------------------------------------
Sorted by:

. summarize

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          id |        20        10.5     5.91608          1         20
      gender |        20         1.5    .5129892          1          2
   education |        20       14.45    2.946452          8         20
      sch_st |        18    3.444444    1.149026          2          5
     sch_com |        20         3.5    1.395481          1          5
-------------+--------------------------------------------------------
      prison |        17    3.176471    1.550617          1          5
     conserv |        19    2.947368    1.544657          1          5

. summarize education, detail

                      Years of education
-------------------------------------------------------------
      Percentiles      Smallest
 1%            8              8
 5%          9.5             11
10%         11.5             12       Obs                  20
25%           12             12       Sum of Wgt.          20

50%         14.5                      Mean              14.45
                        Largest       Std. Dev.      2.946452
75%         16.5             17
90%           18             18       Variance       8.681579
95%           19             18       Skewness      -.1636124
99%           20             20       Kurtosis       2.522208

. log close
      name:  <unnamed>
       log:  C:¥Users¥lightstone¥Documents¥stata¥data¥4results.log
  log type:  text
 closed on:  26 Mar 2015, 13:50:17
-------------------------------------------------------------------------------
```

第 4 章 (4.7 節, p.99) の do-file

演習 4.1

```
/************** Begin do-file ************/
* 4-1.do
use http://www.stata-press.com/data/agis4/firstsurvey_chapter4.dta, clear
summarize
/************** End do-file **************/
```

演習 4.2

```
/************** Begin do-file ************/
* 4-2.do
* This was 4-1.do but has expanded the comments
* This do-file opens the dataset called firstsurvey_chapter4.dta.
* It then runs a summary of the dataset to show descriptive statistics
* for all the variables.
use http://www.stata-press.com/data/agis4/firstsurvey_chapter4.dta, clear
summarize
/************** End do-file **************/
```

演習 4.3

```
/************** Begin do-file ************/
* 4-3.do
* This was 4-2.do but has the describe command added and a summary of
* education to obtain the median score.
* This do-file opens the dataset called firstsurvey_chapter4.dta.
* It describes the dataset.
* It runs a summary of the dataset to show descriptive statistics
* for all the variables.
* It then runs a summary of education to obtain the median value for this variable
use http://www.stata-press.com/data/agis4/firstsurvey_chapter4.dta, clear
describe
summarize
summarize education, detail
/************** End do-file **************/
```

演習 4.5

```
/************** Begin do-file ************/
* 4-5.do
* This was 4-3.do but has the log command added to create a log file.
* This do-file opens the dataset called firstsurvey_chapter4.dta.
* It opens the log file called 4results.log.
* It describes the dataset.
```

```
* It runs a summary of the dataset to show descriptive statistics
*    for all the variables.
* It then runs a summary of education to obtain the median value for this variable
* It closes the log file called 4results.log.
use "C:\data\firstsurvey_chapter4.dta", clear
log using "F:\StataBook\Exercises\4results.log", replace
describe
summarize
summarize education, detail
log close
/************** End do-file **************/
```

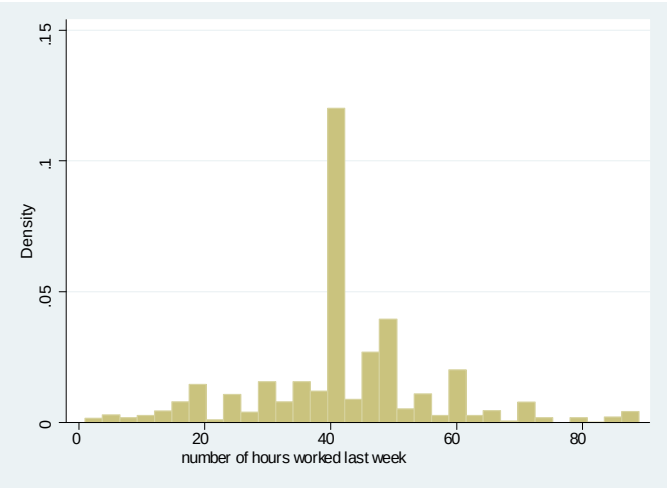
第 5 章 (5.8 節, pp.128-130) の解答

1. descriptive_gss.dta を開き，以下のコマンドを実行します (detail オプションは忘れがちなので注意しましょう).

```
. summarize hrs1, detail
                number of hours worked last week
-------------------------------------------------------------
      Percentiles      Smallest
 1%            6              1
 5%           16              2
10%           21              2       Obs                1729
25%           36              2       Sum of Wgt.        1729
50%           40                      Mean           41.77675
                        Largest       Std. Dev.      14.62304
75%           50             89
90%           60             89       Variance       213.8332
95%           68             89       Skewness       .2834814
99%           88             89       Kurtosis       4.310339
```

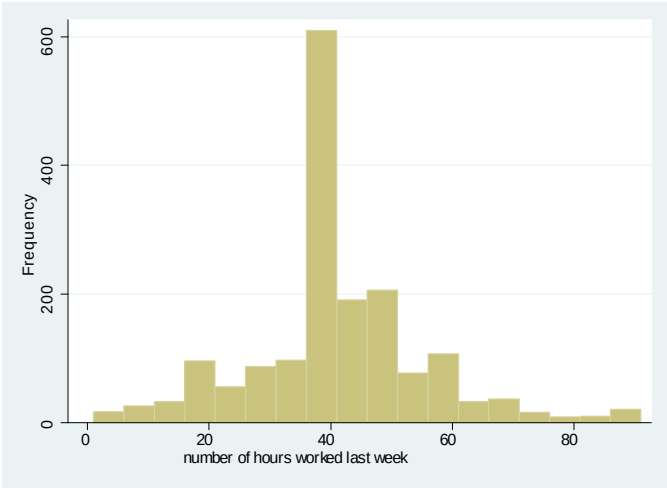
次に，Graphics ▷ Histogram（グラフィックス (G) ▷ ヒストグラム）を選択して histogram ダイアログボックスを開きます．Main（メイン）タブの変数に hrs1 を入力して OK または Submit（適用）をクリックすると，以下のようなヒストグラムが表れます．

```
. histogram hrs1
(bin=32, start=1, width=2.75)
```



このヒストグラムは，Stata が階級の数，幅，開始値を自動で判断したものです．Results ウィンドウに表示の出力から，階級の開始値が 1，幅が 2.75 であると分かります．階級幅は整数の方が望ましいところです．Main（メイン）タブで *Width of bins*（ビンの幅）に 5 と入力します．また，縦軸も *density*（密度）から *frequency*（度数）へ変更しましょう．

```
. histogram hrs1, width(5) frequency
(bin=18, start=1, width=5)
```



ヒストグラムで分布が平均 (41.78) を中心としてほぼ左右対称であることは，`summarize hrs1, detail` の結果で歪度がほぼゼロ (0.28) である点にも表れています．それでも，グラフにも見られる少数の極端なケースが，歪度をわずかに (正方向に) ずれる要因になっています．平均がメディアン (中央値) よりも少し大きいことも，歪度がゼロから正方向にわずかにずれていることと一貫性があります．

演習では尖度が正の値である理由を尋ねる問いがあります．尖度は 4 次モーメントです．

$$\text{尖度}\ (kurtosis) = \frac{\sum (x-\mu)^4}{N\sigma^4}$$

正規分布の尖度は 3.0 です．今回の分布のように，尖度が 3.0 より大きいと，通常，分布が正規分布よりも尖っています．先週の労働時間は，正規分布での想定よりも，40 時間付近に大きく集中しています．これには，調査を実施した米国が，法定労働時間を週 40 時間としていることが一因にあります．尖度については，SAS や SPSS で，上式から 3 を引いた値を報告すること

があり，扱いに注意が必要です．ソフトウェアによっては，分布の尖度が，4.31 ではなく 1.31 で報告されます．この −3 をする尖度は過剰尖度と呼ばれることがありますが，分野によって別の名称で呼ばれることもあります．通常どちらの式を用いるのかについても，分野によって様々です．

2. descriptive_gss.dta を開いたら，satjob7 についての詳細な要約表を表示します．コマンドは，summarize satjob7, detail です．numlabel satjob7, add を実行した後，satjob7 の度数分布表を表示します．numlabel コマンドを事前に実行することで，度数分布表には数値コードと値ラベルがともに表示になります．両方を同時に見ると便利なときがあります．数値コードを非表示にするには numlabel satjob7, remove を実行します．

```
. summarize satjob7, detail
                 job satisfaction in general
-------------------------------------------------------------
      Percentiles      Smallest
 1%            1              1
 5%            1              1
10%            1              1       Obs                 820
25%            2              1       Sum of Wgt.         820
50%            2                      Mean           2.676829
                        Largest       Std. Dev.      1.302463
75%            3              7
90%            5              7       Variance        1.69641
95%            5              7       Skewness       1.097978
99%            7              7       Kurtosis       4.238774
. numlabel satjob7, add
. tabulate satjob7
          job satisfaction in general |      Freq.     Percent        Cum.
------------------------------------+-----------------------------------
              1. completely satisfied |        127       15.49       15.49
                    2. very satisfied |        289       35.24       50.73
                   3. fairly satisfied |        264       32.20       82.93
4. neither satisfied nor dissatisfied |         53        6.46       89.39
               5. fairly dissatisfied |         47        5.73       95.12
                 6. very dissatisfied |         29        3.54       98.66
           7. completely dissatisfied |         11        1.34      100.00
------------------------------------+-----------------------------------
                                Total |        820      100.00
. numlabel, remove
```

メディアンは，very satisfied に相当する 2 であり，平均は，やや fairly satisfied よりな 2.68 です．このことから，平均的な人の仕事満足度について，平均値を用いると，メディアンを用いるときよりも，やや低い報告になります．

メディアンで報告する意義は何でしょう．分布に歪みがあっても，メディアンはその影響を受

けづらいことがあります．今回の分布は dissatisfied から completely dissatisfied にかけて，分布が歪んでいます．従って，より歪みの影響の少ないメディアンの方が平均値よりもよく平均的な人を表している，と言えます．

平均値で報告する意義は何でしょう．確かに今回のように歪んだ分布ではメディアンでの報告が推奨されますが，回答人数の最も多い 2 つのレベルの差はわずかであり，両レベルがともに分布の中心を代表する値である，と言えます．度数分布表からも，very satisfied と fairly satisfied の 2 レベルに多くの人数が集まっているのが分かります．従って，両レベルの間の値である平均値が分布の中心をより的確に反映しています．

3. `descriptive_gss.dta` を開いたら，`deckids` についての度数分布表を表示します．

```
. tabulate deckids
    who makes decision about how to
                   bring up children |      Freq.     Percent        Cum.
-------------------------------------+-----------------------------------
                           mostly me |         81       15.06       15.06
                     mostly my spouse |         44        8.18       23.23
sometimes me or sometimes my spouse |        102       18.96       42.19
                  we decide together |        305       56.69       98.88
                        someone else |          6        1.12      100.00
-------------------------------------+-----------------------------------
                               Total |        538      100.00
```

次に，メニューから Statistics ▷ Summaries, tables, and tests ▷ Frequency tables ▷ One-way table（統計 (S) ▷ 要約/表/検定 ▷ 度数分布表 ▷ 一元配置表）を選択してダイアログボックスを開きます．Main（メイン）タブで，変数に `deckids` と指定します．by/if/in タブで，*Repeat command by groups*（グループごとにコマンドを実行する）を選択し，*Variables that define groups*（グループ変数）に `sex` と指定します．

```
. by sex, sort: tabulate deckids
-------------------------------------------------------------------------------
-> sex = male
    who makes decision about how to
                   bring up children |      Freq.     Percent        Cum.
-------------------------------------+-----------------------------------
                           mostly me |         10        4.26        4.26
                     mostly my spouse |         27       11.49       15.74
sometimes me or sometimes my spouse |         53       22.55       38.30
                  we decide together |        140       59.57       97.87
                        someone else |          5        2.13      100.00
-------------------------------------+-----------------------------------
                               Total |        235      100.00
-------------------------------------------------------------------------------
```

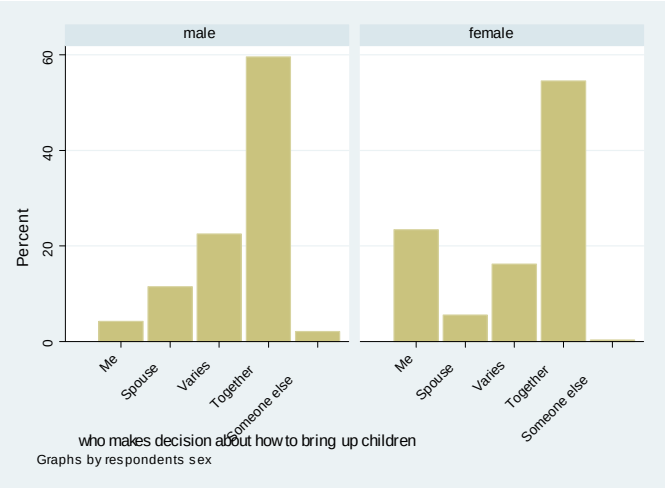
```
-> sex = female
    who makes decision about how to
                  bring up children |      Freq.     Percent        Cum.
------------------------------------+-----------------------------------
                          mostly me |        71       23.43       23.43
                       mostly my spouse |        17        5.61       29.04
sometimes me or sometimes my spouse |        49       16.17       45.21
                  we decide together |       165       54.46       99.67
                        someone else |         1        0.33      100.00
------------------------------------+-----------------------------------
                              Total |       303      100.00
```

deckids の平均，メディアン，標準偏差について，標本全体での値を報告してもあまり意味がありません．それらが男女間で異なっていることに加え，deckids に連続性が全くないという，さらに重要なことがあるからです．deckids で，1 はほぼすべて自らが決定，2 はほぼすべて配偶者が決定，3 はそれぞれが決定するなど，項目と順番には関連がありません．

この点は，分布を棒グラフで表すと明確になります．メニューから Graphics ▷ Histogram（グラフィックス (G) ▷ ヒストグラム）を選択します．Main（メイン）タブで，変数に deckids を指定します．*Data*（データ）セクションで *Data are discrete*（離散データ）を選択し，*Y axis*（y 軸）セクションで *Percent*（パーセント）を選択します．Bar properties（棒のプロパティ）を選択し，開いたダイアログで *Bar gap*（棒の間隔）に 10 を入力して，棒の間に間隔があるようにします．By（by 条件）タブで，*Draw subgraphs for unique values on variables*（変数のユニーク値ごとのサブグラフを作成する）を選択し，*Variables*（変数）で sex を指定します．X axis（x 軸）タブで，*Major tick/label properites*（主目盛/ラベルのプロパティ）を選択し，開いたダイアログで *Custom*（カスタム）を選択し，*Custom rule*（カスタムルール）に `1 "Me" 2 "Spouse" 3 "Varies" 4 "Together" 5 "Someone else"`を入力します．これにより，ラベル付けが行われ，グラフが見やすくなります．さらに同じダイアログの Label（ラベル）タブで，*Angle*（角度）に 45 degrees を指定します．

生成されるコマンドとグラフは以下のようになります．

```
. histogram deckids, discrete percent gap(10) xlabel(1 "Me" 2 "Spouse" 3 "Varie
> s" 4 "Together" 5 "Someone else", angle(forty_five)) by(sex)
```



開いたグラフで右クリックをし，Start Graph Editor（グラフエディタの開始）を選択します．グラフは今のままでも十分見やすいので，特に大きく変更する必要はありませんが，強いて言えば文字列で，最初を大文字するとさらに良くなることと思います．グラフの下の方にあるグラフのタイトルの上でクリックし，*Text* ボックスで文字列を変更します．同様に，各グラフの上部にある `male` と `female` を `Men`，`Women` に変更します．グラフのタイトルは，下の方でなく，上の方にある方がよいかもしれません．上の方にタイトルを表示するには，右側にある Object Browser で，title を選択し，*Text* ボックスに文字列を入力します．左下の脚注に表れている文はより有益な情報にできないでしょうか．同 Object Browser で note を選択し，Text ボックスで `Data from the General Social Survey` へと文字列を変更します．

4. `descriptive_gss.dta` を開いたら，`strsswrk` についての度数分布表と要約表を表示します (`tabulate` と `summarize`)．`tabulate` ダイアログボックスを開くには，メニューから Statistics ▷ Summaries, tables, and tests ▷ Frequency tables ▷ One-way table（統計 (S) ▷ 要約/表/検定 ▷ 度数分布表 ▷ 一元配置表）を選択します．Main（メイン）タブで，`strsswrk` を指定します．

```
. tabulate strsswrk
   job is rarely stressful |      Freq.     Percent        Cum.
------------------------------------------------------------------
            strongly agree |        73        7.30        7.30
                     agree |       249       24.90       32.20
neither agree nor disagree |       198       19.80       52.00
                  disagree |       329       32.90       84.90
         strongly disagree |       151       15.10      100.00
```

```
                      Total |      1,000      100.00
```

summarize ダイアログボックスを開くには，メニューから Statistics ▷ Summaries, tables, and tests ▷ Summary and descriptive statistics ▷ Summary statistics（統計 (S) ▷ 要約/表/検定 ▷ 記述統計量 ▷ 記述統計量）を選択します．Main（メイン）タブで，strsswrk を指定します．

```
. summarize strsswrk
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    strsswrk |       1000       3.236    1.191522          1          5
```

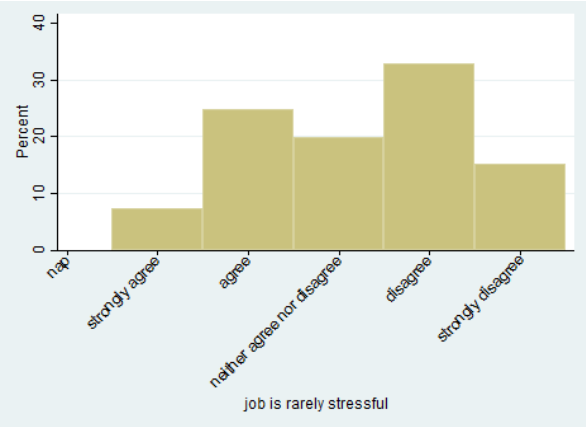
次に，by/if/in タブでの指定も加えて，再び 2 つのダイアログボックスを使用します．by/if/in タブで，*Repeat command by groups*（グループごとにコマンドを実行する）を選択し，*Variables that define groups*（グループ変数）に sex と指定します．

```
. by sex, sort: tabulate strsswrk
-------------------------------------------------------------------------------
-> sex = male
   job is rarely stressful |      Freq.     Percent        Cum.
-------------------------+-----------------------------------+------------
            strongly agree |         38        8.58        8.58
                     agree |        106       23.93       32.51
neither agree nor disagree |         97       21.90       54.40
                  disagree |        139       31.38       85.78
         strongly disagree |         63       14.22      100.00
-------------------------+-----------------------------------+------------
                     Total |        443      100.00
-------------------------------------------------------------------------------
-> sex = female
   job is rarely stressful |      Freq.     Percent        Cum.
-------------------------+-----------------------------------+------------
            strongly agree |         35        6.28        6.28
                     agree |        143       25.67       31.96
neither agree nor disagree |        101       18.13       50.09
                  disagree |        190       34.11       84.20
         strongly disagree |         88       15.80      100.00
-------------------------+-----------------------------------+------------
                     Total |        557      100.00
. by sex, sort: summarize strsswrk
-------------------------------------------------------------------------------
-> sex = male
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    strsswrk |        443    3.187359     1.19714          1          5
-------------------------------------------------------------------------------
-> sex = female
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
```

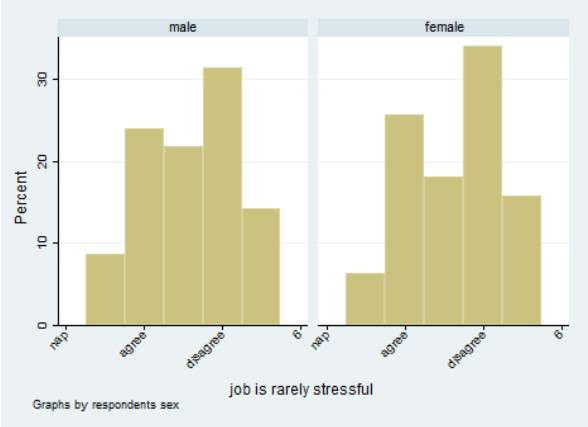
```
    strsswrk |       557    3.274686    1.186687          1          5
```

次に，Graphics ▷ Histogram（グラフィックス (G) ▷ ヒストグラム）を選択してヒストグラムを作成します．Main（メイン）タブで，`strsswrk` を指定し，*Data are discrete*（離散データ），*Percent*（パーセント）をそれぞれ選択します．

X axis（x 軸）タブで，*Major tick/label properites*（主目盛/ラベルのプロパティ）をクリックし，開いたダイアログの Label（ラベル）タブで，*Angle*（角度）に `45 degrees` を指定し，*Use value labels*（値ラベルを使用する）を選択します．この操作により，値ラベルが 45 度傾いて，ヒストグラムに収まりよく表示されます．



最後に，ヒストグラムを男女別にするため，By（by 条件）タブで，*Draw subgraphs for unique values on variables*（変数のユニーク値ごとのサブグラフを作成する）を選択し，*Variables*（変数）で `sex` を指定します．



各グラフのそれぞれの棒にパーセント値を表示したい場合，Main（メイン）タブで，*Add height to bars*（棒に高さのラベルを追加する）を選択します．

strongly disagree を意味する 1 から strongly agree を意味する 5 までで，仕事に対するストレスの度合いのメディアンは 3.0 です．男女を問わず，等しく 3.0 です．平均は，女性で 3.19，男性で 3.27 です．男性の方に高い値が示された一方，女性との差はわずかです．仕事に対するストレスの度合いは，男女ともに同じくらいだと結論づけできます．

5. `descriptive_gss.dta` を開いたら，メニューから Statistics ▷ Summaries, tables, and tests ▷ Frequency tables ▷ Multiple one-way tables（統計 (S) ▷ 要約/表/検定 ▷ 度数分布表 ▷ 複数の度数分布表）を選択して `tab1` ダイアログボックスを開きます．変数を指定して Submit（適用）をクリックします．

```
. tab1 trustpeo wantbest advantge goodlife
-> tabulation of trustpeo
there are only a few ppl r |
      can trust completely |      Freq.     Percent        Cum.
-----------------------------+-----------------------------------
            strongly agree |       445       39.35       39.35
                     agree |       455       40.23       79.58
neither agree nor disagree |       106        9.37       88.95
                  disagree |       107        9.46       98.41
         strongly disagree |        18        1.59      100.00
-----------------------------+-----------------------------------
                     Total |     1,131      100.00
-> tabulation of wantbest
  r is sure that other ppl |
```

want the best for r	Freq.	Percent	Cum.
strongly agree	158	14.06	14.06
agree	542	48.22	62.28
neither agree nor disagree	268	23.84	86.12
disagree	131	11.65	97.78
strongly disagree	25	2.22	100.00
Total	1,124	100.00	

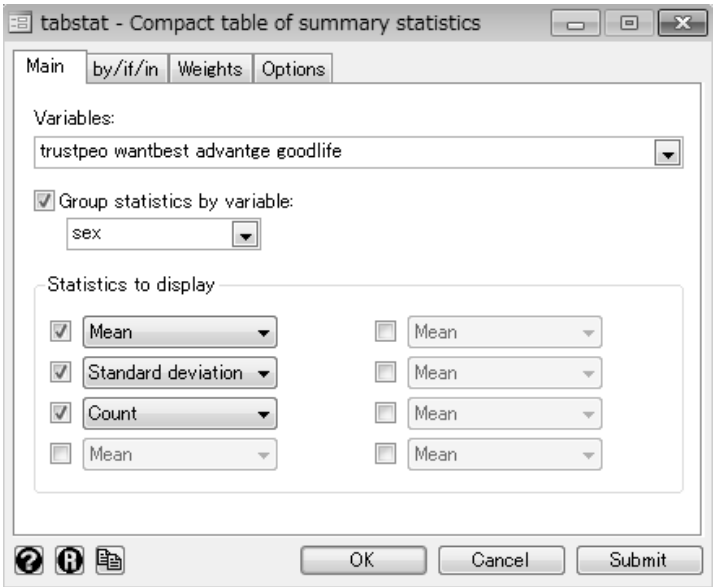
-> tabulation of advantge

other ppl will take advantage of r if not careful	Freq.	Percent	Cum.
strongly agree	280	24.80	24.80
agree	514	45.53	70.33
neither agree nor disagree	180	15.94	86.27
disagree	143	12.67	98.94
strongly disagree	12	1.06	100.00
Total	1,129	100.00	

-> tabulation of goodlife

standard of living of r will improve	Freq.	Percent	Cum.
strongly agree	231	25.41	25.41
agree	447	49.17	74.59
neither	96	10.56	85.15
disagree	118	12.98	98.13
strongly disagree	17	1.87	100.00
Total	909	100.00	

tabstat ダイアログボックスを開くには，Statistics ▷ Summaries, tables, and tests ▷ Other tables ▷ Compact tables of summary statistics（統計 (S) ▷ 要約/表/検定 ▷ その他の表 ▷ 簡易型要約統計表）を選択します．Main（メイン）タブで，以下のような指定を行います．



実行結果は，以下のようになります．

```
. tabstat trustpeo wantbest advantge goodlife, statistics( mean sd count ) by(s
> ex)
Summary statistics: mean, sd, N
  by categories of: sex (respondents sex)
    sex |  trustpeo  wantbest  advantge  goodlife
--------+----------------------------------------
   male |  1.886239  2.504604  2.154128   2.08933
        |  .9634371  .9544405   .930541  .9367884
        |       545       543       545       403
--------+----------------------------------------
 female |  1.984642  2.297762  2.236301  2.229249
        |  1.036811  .9205387   1.03406  1.061856
        |       586       581       584       506
--------+----------------------------------------
  Total |  1.937224  2.397687  2.196634  2.167217
        |  1.002891  .9423422  .9858697  1.010181
        |      1131      1124      1129       909
----------------------------------------------------
```

ヒント：tabstat のような特殊なダイアログボックスは，Stata のバージョンによってメニュー内の配置が幾分異なることがあります．tabstat というコマンド名を覚えておけば早いのですが，それでもコマンドシンタックスなど，確認したいことがあるときは，help tabstat を実行すると便利です．help ウィンドウでは，右上にある dialog（ダイアログ）メニューからダイアログボックスを開くことができます．

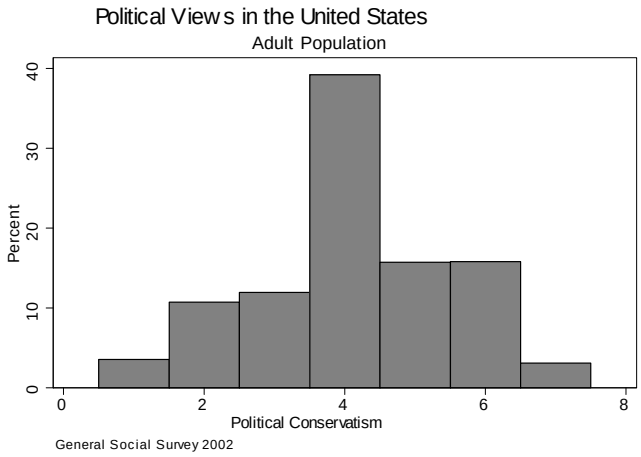
6. descriptive_gss.dta を開いたら，polviews についての度数分布表を表示します (tabulate または fre)．tabulate ダイアログボックスを開くには，メニューから Statistics ▷ Summaries, tables, and tests ▷ Frequency tables ▷ One-way table（統計 (S) ▷ 要約/表/検定 ▷ 度数分布表 ▷ 一元配置表）を選択します．Main（メイン）タブで，strsswrk を指定し，*Treat missing values like other values*（欠損値を他の値と同様に扱う）を選択します．

```
. tabulate polviews, missing
    think of self as
          liberal or
        conservative |      Freq.     Percent        Cum.
-------------------+-----------------------------------
   extremely liberal |         47        1.70        1.70
             liberal |        143        5.17        6.87
    slightly liberal |        159        5.75       12.62
            moderate |        522       18.88       31.50
slghtly conservative |        209        7.56       39.06
        conservative |        210        7.59       46.65
extrmly conservative |         41        1.48       48.14
                   . |      1,434       51.86      100.00
-------------------+-----------------------------------
               Total |      2,765      100.00
```

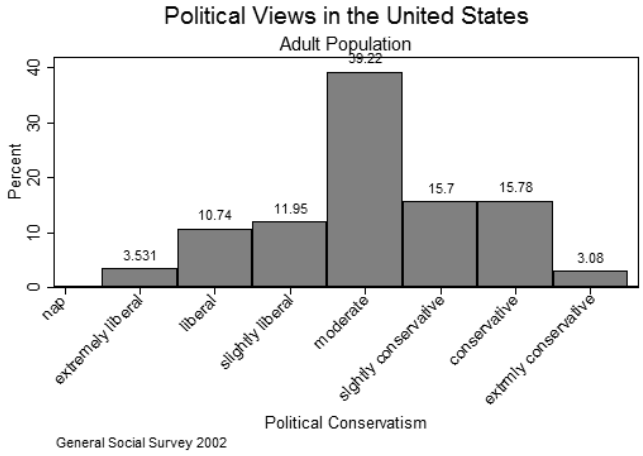
次に，同じく polviews に対して fre を実行してみます．

```
. fre polviews
polviews — think of self as liberal or conservative
-------------------------------------------------------------------------
                                  |     Freq.    Percent      Valid       Cum.
----------------------------------+--------------------------------------
Valid   1 extremely liberal       |        47       1.70       3.53       3.53
        2 liberal                 |       143       5.17      10.74      14.27
        3 slightly liberal        |       159       5.75      11.95      26.22
        4 moderate                |       522      18.88      39.22      65.44
        5 slghtly conservative    |       209       7.56      15.70      81.14
        6 conservative            |       210       7.59      15.78      96.92
        7 extrmly conservative    |        41       1.48       3.08     100.00
        Total                     |      1331      48.14     100.00
Missing .                         |      1434      51.86
Total                             |      2765     100.00
-------------------------------------------------------------------------
```

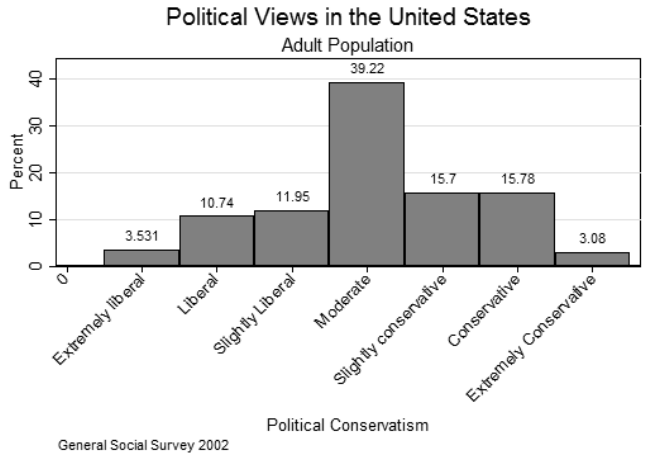
次に，棒グラフを作成します．作成の手順は本体書籍の p.119 とほぼ同じです．本体書籍で作成するグラフ (図 5.9) は，以下のようになると思います．



このグラフで，さらに x 軸で 8 つの値ラベルを 45 度傾けて表示するように設定し，それぞれの棒の頭にパーセント値を表示するようにすると，以下のようになります.



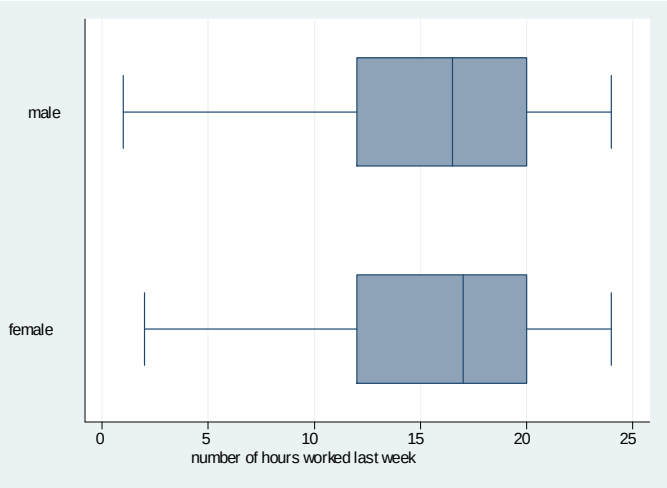
グラフはこの状態でも問題ありませんが，左端に nap(not applicable) という値ラベルのついた項目が見えます．このように，当てはまる観測が全く無くても，変数の値ラベルが存在することで，Stata がグラフにもその項目を含めることがあります．このため，新たな値ラベルを定義し，`polviews` に適用する必要があります．さらに体裁を若干、調整した最終的なグラフは，以下のようになると思います.



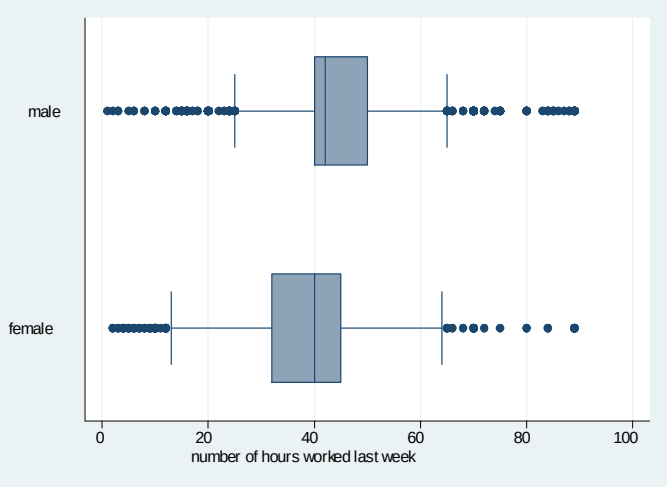
グラフはこの状態でも問題ありませんが，棒の間隔を 10% に設定すると，さらに棒グラフらしくなります．本体書籍の図 5.9 と比較すると，値ラベルの使用により，それぞれの棒の示す項目が明確になったことが挙げられます．棒の頭へのパーセント値の表示の是非ついては，好みに依るところでしょう．値が明確になり，理解しやすいと思える面もあり，表示が混雑するという見方もできます．

最後に，0 パーセントのカテゴリを非表示にする方法ですが，一つの方法としては，X axis（x 軸）タブの *Major tick/label properites*（主目盛/ラベルのプロパティ）で，*Range/Delta*（範囲/増分）を選択し，*Minimum value*（最小値）に 1 を，*Maximum value*（最大値）に 7 を，*Delta*（増分）に 1 を指定する方法があります．

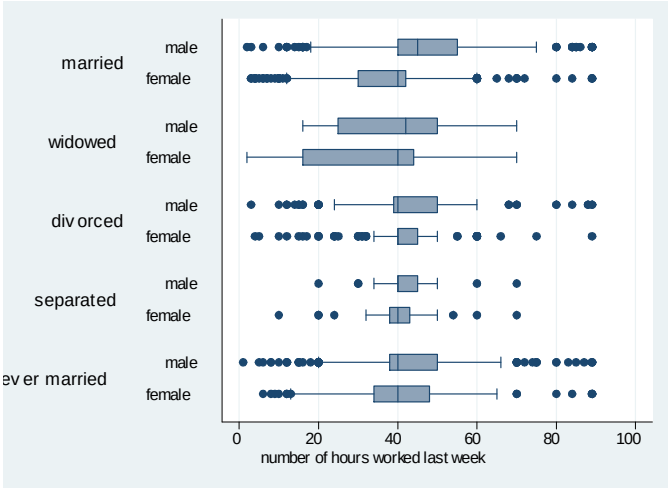
7. `descriptive_gss.dta` を開きます．図 5.13 を作成するときの手順で，`wwwhr` を `hrs1` に代えて箱ひげ図を作成します．`graph box` ダイアログボックスの Main（メイン）タブでは，忘れずに *Horizontal*（水平）を選択してください．



if 条件式にある `hrs1 < 25` を取り除くと，箱ひげ図は次のようになると思います．



さらに，グループ 2 に `marital` を指定すると，箱ひげ図は次のようになると思います．



8. 無作為抽出でシードを指定するのは，同じ抽出プログラムで同じ結果を再現させるためです．具体的には，シードの指定により抽出の開始位置が決定します．逆にシードを指定しなければ，do ファイルの実行の度に，(たとえば 10 個の) 異なる観測値が抽出されます．次節の do ファイルでも，シードを 153 に指定します (シードはこれ以外の値でも問題ありませんが，開始位置が異なってきます).

   データセットの全標本 ($N = 2,753$) で求めると，歪度と尖度は共に有意となります．大きい標本では，正規分布からのずれが比較的重要でない一方，小さい標本では，ずれが重要さの度合いが比較的大きくなります．本演習のように標本が 10 個のみの小さい標本では，正規分布からのずれは重要ですが，その有意性は見られにくくなります．

9. ヒストグラム作成の操作法は演習 4 と同様です．ここでのポイントは，Stata が自動で選択する階級の上限値と下限値を，小数から整数に修正する点です．教育年数は 1 年刻みでの階級分けが重要になる場合が多くあります．階級数は増えてしまいますが，階級幅を 1 としてヒストグラムを作成します．階級幅の設定は Main（メイン）タブで行います．

   作成したヒストグラムはそれほど有用なものではありません．女性の教育年数は男性よりも 12 年である傾向が高いことが分かります．12 年というのは高校までの教育課程を修了した場合に相当します．一方，男性の教育年数は 17 年になる傾向が高くなっています．17 年は一般的に学士号を取得した場合に相当します．全体としては，両ヒストグラムは似通ったものです．

10. graph box ダイアログボックスを開くには，メニューから Graphics ▷ Box plot（グラフィックス (G) ▷ 箱ひげ図）を選択します．Main（メイン）タブで，*Vertical*（垂直）を選択し，*Variables*（変数）に educ を指定します．By（by 条件）タブで，*Draw subgraphs for unique values of variables*（変数のユニーク値ごとのサブグラフを作成する）を選択し，*Variables*（変数）に sex を指定します．実行すると，基本となる箱ひげ図が表示になります．演習 9 と同様の方法で，さらに編集していくことができます．箱ひげ図を見ると，メディアンが男性と女性でほぼ同じ値であることが分かります．男性は女性と比較して，分布の下側で観測値がわずかに多く (下部のひげが長い)，また，大卒付近の観測値も多い (箱の上部) ことが分かります．

11. tabstat ダイアログボックスを開くには，Statistics ▷ Summaries, tables, and tests ▷ Other tables ▷ Compact tables of summary statistics（統計 (S) ▷ 要約/表/検定 ▷ その他の表 ▷ 簡易型要約統計量）を選択します．Main（メイン）タブで，*Variables*（変数）に educ を指定し，*Group statistics by variable*（変数で統計量をグループ分けする）を選択し，変数として sex を指定します．*Statistics to display*（表示する統計量）セクションで，チェックボックスを 6 つ選択し，Mean(平均値), Median(中央値), Standard deviation(標準偏差), Skewness(歪度), Kurtosis(尖度)，Interquartile range(四分位範囲）を選択します．そして，Options（オプション）タブで，*Use as columns*（列として配置）に Statistics（統計量）を選択して見やすい表示にします．

実行結果は，次のようになります．

```
. tabstat educ, statistics( mean median sd skewness kurtosis iqr ) by(sex) colu
> mns(statistics)
Summary for variables: educ
     by categories of: sex (respondents sex)
```

sex	mean	p50	sd	skewness	kurtosis	iqr
male	13.37827	13	3.135253	-.6232506	4.872827	4
female	13.35252	13	2.839165	-.2607622	4.590152	3
Total	13.36397	13	2.973924	-.4491471	4.786138	4

上記の結果は，演習 9，10 で作成したグラフと一貫するものですが，特定の統計量について，正確な値を示しています．教育年数について，平均とメディアン (p50) は男女間に実質的に同じです．一方，標準偏差と四分位範囲 (iqr) は男性の方が高く，ばらつきが大きいと読み取れます．また，男性の分布は女性ものに比べてより左に歪んでいるとできます．

グラフと統計量は，共に同じストーリーを伝える一方で，互いに補い合っており，この点を本演習で確認しておきましょう．

第 5 章 (5.8 節, pp.128-130) の do-file

演習 5.1

```
/************** Begin do-file ************/
* chapter5.1.do
* This program opens the dataset called descriptive_gss.dta.
* It then runs a detailed summary of hrs1 and creates a histogram.
clear
use "C:\data\descriptive_gss.dta", clear
summarize hrs1, detail
histogram hrs1, width(5)
histogram hrs1, frequency width(5)
/************** End do-file **************/
```

演習 5.2

```
/************** Begin do-file ************/
*chapter5.2.do
clear
use "C:\data\descriptive_gss.dta", clear
summarize satjob7, detail
numlabel satjob7, add
tabulate satjob7
numlabel satjob7, remove
/************** End do-file **************/
```

演習 5.3

```
/************** Begin do-file ************/
* 4-3.do
* This was 4-2.do but has the describe command added and a summary of
* education to obtain the median score.
* This do-file opens the dataset called firstsurvey_chapter4.dta.
* It describes the dataset.
* It runs a summary of the dataset to show descriptive statistics
* for all the variables.
* It then runs a summary of education to obtain the median value for this variable
use http://www.stata-press.com/data/agis4/firstsurvey_chapter4.dta, clear
describe
summarize
summarize education, detail
/************** End do-file **************/
```

演習 5.4

```
/************** Begin do-file ************/
```

```
* chapter5.4.do
* This program opens the dataset called descriptive_gss.dta.
* It then runs a tabulation with and without the numlabel turned on.
clear
use "C:\data\descriptive_gss.dta", clear
tabulate strsswrk
summarize strsswrk, detail
by sex, sort: tabulate strsswrk
by sex, sort: summarize strsswrk, detail
histogram strsswrk if strsswrk >0, discrete percent xlabel(, ///
valuelabel angle(forty_five))
histogram strsswrk if strsswrk >=1 & strsswrk<=5, ///
discrete percent xlabel(, valuelabel angle(forty_five)) by(sex)
/************** End do-file **************/
```

演習 5.5

```
/************** Begin do-file ************/
* chapter5.5.do
use "C:\data\descriptive_gss.dta", clear
tab1 trustpeo wantbest advantge goodlife
tabstat trustpeo wantbest advantge goodlife, ///
statistics(mean sd count) by(sex) columns(variables)
/************** End do-file **************/
```

演習 5.6

```
/************** Begin do-file ************/
*chapter5.6.do
use "C:\data\descriptive_gss.dta", clear
tabulate polviews, missing
fre polviews
* Create simple chart 5_6_a
histogram polviews, discrete percent gap(10)
* Add labels at 45 degree angle and percentages at top of bars
histogram polviews, discrete percent gap(10) addlabel ///
xlabel(#7, angle(forty_five) valuelabel ticks)
label define new 1 "Extreme liberal" 2 "Liberal" ///
3 "Slight liberal" 4 "Moderate" ///
5 "Slight conservative" 6 "Conservative" 7 "Extreme conservative"
label values polviews new
histogram polviews, discrete percent gap(10) addlabel ///
xlabel(#7, angle(forty_five) valuelabel ticks)
histogram polviews, discrete percent gap(10) addlabel ///
xlabel(1(1)9, angle(forty_five) valuelabel ticks)
/************** End do-file **************/
```

演習 5.7

```
/************** Begin do-file ************/
* chapter5.7.do
use "C:\data\descriptive_gss.dta", clear
graph hbox hrs1 if hrs1 < 25, over(sex)
graph hbox hrs1 if hrs1, over(sex) over(marital)
/************** End do-file **************/
```

演習 5.8

```
/************** Begin do-file ************/
* chapter5.8.do
use "C:\data\descriptive_gss.dta", clear
set seed 153
summarize educ, detail
sktest educ
histogram educ
preserve
sample 10, count
summarize educ, detail
sktest educ
histogram educ
restore
/************** End do-file **************/
```

演習 5.9

```
/************** Begin do-file ************/
*chapter5.9.do
clear
use "C:\data\descriptive_gss.dta", clear
by sex, sort: tabulate educ
histogram educ, percent by(sex) width(1)
/************** End do-file **************/
```

演習 5.10

```
/************** Begin do-file ************/
*chapter5.10.do
clear
use "C:\data\descriptive_gss.dta", clear
graph box educ, by(sex)
/************** End do-file **************/
```

演習 5.11

```
/************** Begin do-file ************/
*chapter5.11.do
clear
use "C:\data\descriptive_gss.dta", clear
tabstat educ, statistics(mean median sd kurtosis iqr skewness) by(sex) ///
columns(statistics)
/************** End do-file **************/
```

第 6 章 (6.11 節, pp.157-158) の解答

1. ((訂正) 本演習については本体書籍の記述に誤植がありました．以下のように訂正するとともに，心からお詫び申し上げます．

   (誤)gss2006chapter6.dta

   (正)gss2006_chapter6.dta

   (訂正終わり))

   gss2006_chapter6.dta を開きます．Data ▷ Describe data ▷ Describe data contents (codebook)（データ (D) ▷ データの内容表示 ▷ データの内容表示 (codebook)）を選択し，Main（メイン）タブで，pornlaw を指定して変数 pornlaw の codebook 出力を表示します．

   ヒント：codebook ダイアログボックスでは，Options（オプション）タブの *Display compact report on the variables*（変数の簡易版レポートを表示する）を選択すると，変数 1 つにつき 1 行の短縮された codebook 出力を表示します．初めて開くデータセットなどに対して，この方法で内容の感触を掴むと便利です．

```
. codebook pornlaw
-------------------------------------------------------------------------------
pornlaw                                          FEELINGS ABOUT PORNOGRAPHY LAWS
-------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  PORNLAW

                 range:  [1,3]                        units:  1
         unique values:  3                        missing .:  2543/4510

            tabulation:  Freq.   Numeric  Label
                           775         1  ILLEGAL TO ALL
                          1125         2  ILLEGAL UNDER 18
                            67         3  LEGAL
                          2543         .
```

   私たちは，性別がポルノの合法化への意見によっては決まらない一方，意見が性別に依存することはあると知っているので，性別が独立変数であると判断できます．女性の方がポルノに対して否定的な意見を持つと予想します．独立変数は，通常，クロス表で行変数として縦に並ぶように配置します．パーセント数は各独立変数内で計算するので，本演習でも行内のパーセントを計算します．クロス表を作成するには，メニューから，Statistics ▷ Summaries, tables, and tests ▷ Frequency tables ▷ Two-way tables with measures of association（統計 (S) ▷ 要約/表/検定 ▷ 度数分布表 ▷ 二元配置の表/統計値）を選択します．Main（メイン）タブの *Row variable*

（行の変数）で，sex を指定し，*Column variable*（列の変数）で，pornlaw を指定します．行内のパーセント数を計算するので，*Within-row relative frequenies*（行内の相対度数）を選択します．

```
. tabulate sex pornlaw, row
```

Key
frequency *row percentage*

Gender	FEELINGS ABOUT PORNOGRAPHY LAWS ILLEGAL T	ILLEGAL U	LEGAL	Total
MALE	234 27.86	568 67.62	38 4.52	840 100.00
FEMALE	541 48.00	557 49.42	29 2.57	1,127 100.00
Total	775 39.40	1,125 57.19	67 3.41	1,967 100.00

もし表を外部に公表するのであれば，今の表のラベルのままでは意味が通じません．作成者である私たちは，“ILLEGAL T”が“ILLEGAL TO ALL”であると知っていますが，codebook 出力を見ない限りその判断はつきません．クロス表で文字を表示できるスペースは限られています．ここは少しクリエイティブになる必要があります．たとえば，変数ラベルを“Pornography should be legal to(ポルノグラフィを合法とする年齢)”とします．そして，値ラベルとして 1 を“Nobody”，2 を“Over 18”，3 を“All”とする手があります．

```
. label variable pornlaw "Pornography should be legal to"
. label define newporn 1 "Nobody" 2 "Over 18" 3 "All"
. label values pornlaw newporn
```

クロス表を再び表示すると，以下のようになります．

```
. tabulate sex pornlaw, row
```

Key
frequency *row percentage*

Gender	Pornography should be legal to Nobody	Over 18	All	Total
MALE	234	568	38	840

	27.86	67.62	4.52	100.00
FEMALE	541	557	29	1,127
	48.00	49.42	2.57	100.00
Total	775	1,125	67	1,967
	39.40	57.19	3.41	100.00

2. 前回の演習のデータセットとダイアログボックスを引き続き使用します．Main（メイン）タブで，*Pearson's chi-squared*（ピアソンのカイ二乗）と *Cramer's V*（クラメールの V）を選択します．

```
. tabulate sex pornlaw, chi2 row V
```

Key
frequency *row percentage*

	Pornography should be legal to			
Gender	Nobody	Over 18	All	Total
MALE	234	568	38	840
	27.86	67.62	4.52	100.00
FEMALE	541	557	29	1,127
	48.00	49.42	2.57	100.00
Total	775	1,125	67	1,967
	39.40	57.19	3.41	100.00

```
Pearson chi2(2) =  82.8157   Pr = 0.000
     Cramr´s V =   0.2052
```

χ^2 値は 82.82 です．従って，性別とポルノへの見解の関連性は $\chi^2 = 82.82, p < 0.001$ により有意に存在すると報告できます．

クラメールの V は 0.21 であり，性別とポルノへの見解に中程度の関連性があることを示しています．
多くの調査が，2 変数の関連性の強さを連関係数で示していますが，今回の場合の $V = 0.21$ など，一つの値だけでは，関連性に関するある側面のみしか示せません．ここでは，パーセント数による吟味が大変有用です．行ごとの相対度数を表示したら，それを列ごとに比較します．女性では 48% がポルノの完全非合法を支持しているのに対して，男性は 28% に留まります．数字から，そのような意見を女性が持つ確率は男性の 2 倍近いと言えます．

3. データセットを開いたら，クロス表を表示します．メニューから，Statistics ▷ Summaries, tables, and tests ▷ Frequency tables ▷ Two-way tables with measures of association（統計 (S) ▷ 要約/表/検定 ▷ 度数分布表 ▷ 二元配置の表/統計値）を選択します．独立変数と考えられる `pres00` を *Row variable*（行の変数）に，もう一方の `pres04` を *Column variable*（列の変数）にそれぞれ指定します．行内のパーセント数を計算するので，*Pearson's chi-squared*（ピアソンのカイ二乗），*Within-row relative frequenies*（行内の相対度数），*Cramer's V*（クラメールの V），*Treat missing values like other values*（欠損値を他の値と同様に扱う）を選択します．表れるクロス表は大変大きく表示が乱れてしまうので，ここには記載しません．

```
. codebook pres00 pres04
-------------------------------------------------------------------------------
pres00                                                  VOTE FOR GORE, BUSH, NADER
-------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  PRES00

                 range:  [1,6]                        units:  1
         unique values:  5                        missing .:  2,740/4,510

            tabulation:  Freq.   Numeric  Label
                           813         1  GORE
                           903         2  BUSH
                            26         3  NADER
                            19         4  OTHER (SPECIFY)
                             9         6  DIDNT VOTE
                         2,740         .
-------------------------------------------------------------------------------
pres04                                                 VOTE FOR KERRY, BUSH, NADER
-------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  PRES04

                 range:  [1,6]                        units:  1
         unique values:  4                        missing .:  1,566/4,510

            tabulation:  Freq.   Numeric  Label
                         1,434         1  KERRY
                         1,446         2  BUSH
                            47         3  NADER
                            17         6  DIDNT VOTE
                         1,566         .
```

演習では続けて，2000 年と 2004 年のどちらの選挙も，民主党と共和党のどちらかの候補者に投票した有権者以外を除外するようにとあります．この除外によりクロス表はすっきりします．（ペロー氏の支持者に起きた事柄に関心のある政治学者はこのアプローチに抵抗感を覚えるかもしれませんが．）先ほどクロス表を作成したときのダイアログボックスに戻ります．Main（メイン）タブで，*Treat missing values like other values*（欠損値を他の値と同様に扱う）のチェックを外します．by/if/in タブで，*Restrict observations*（観測値の制限）に `(pres00 == 1 | pres00`

== 2) & (pres04 == 1 | pres04 == 2) と入力します．指定する値がこれらで良いことは，既に codebook の結果から分かっています．（ここで，pres00 <= 2 & pres04 <=2 と入力しても問題ありません．）結果は以下のようなシンプルなものになります．

```
. tabulate pres00 pres04 if (pres00 == 1 | pres00 == 2) & (pres04 == 1 | pres04 == 2)
> , chi2 row V
```

Key
frequency *row percentage*

VOTE FOR GORE, BUSH, NADER	VOTE FOR KERRY, BUSH, NADER KERRY	BUSH	Total
GORE	680 92.02	59 7.98	739 100.00
BUSH	65 7.85	763 92.15	828 100.00
Total	745 47.54	822 52.46	1,567 100.00

```
Pearson chi2(1) =  1.1e+03   Pr = 0.000
     Cramr´s V =   0.8413
```

この表を公開するときは，既に表から除外されている NADER を変数ラベルからも除外した方がよいでしょう．

両選挙とも民主党あるいは共和党の候補者だけに絞って調べると，2000 年と 2004 年における有権者の投票先には極めて有意な関連があることが分かります．$\chi^2(1)$ の値は Stata に用意されたスペースに収まり切らないため，1.1e+03 という指数表現で示さていれます．1.1 の小数点を右に+03 ずらすことにより，$\chi^2(1) = 1,100$, $p <\ 0.001$ を得ます．クロス表が 2×2 なので，クラメールの V は係数 ϕ と等しくなります．0.84 という ϕ の値は強い関連を示します．党候補者に対する投票行為にはかなりの一貫性があると言えそうです．パーセント数を見ると，ゴア氏（民主党）に投票した有権者の 92% がケリー氏（民主党）に投票し，残り 8% がブッシュ氏（共和党）したことが分かります．2000 年に共和党（のブッシュ氏）に投票した有権者は 2004 年にも高い確率でブッシュ氏に投票した（92% 対 8% ）ことも分かります．

4. データセットを開いたら，Statistics ▷ Summaries, tables, and tests ▷ Frequency tables ▷ Two-way tables with measures of association（統計 (S) ▷ 要約/表/検定 ▷ 度数分布表 ▷ 二元配置の表/統計値）を選択してダイアログボックスを開きます．Main（メイン）タブで，*Row variable*（行の変数）に polviews と入力し，*Column variable*（列の変数）に premarsx と入力します．

Pearson's chi-squared（ピアソンのカイ二乗），*Goodman and Kruskal's gamma*（Goodman と Kruskal のガンマ），*Kendall's tau-b*（ケンドールのタウ b）を選択します．また，*Within-row relative frequencies*（行内の相対度数）を選択します．

```
. tabulate polviews premarsx, chi2 gamma row taub
```

Key
frequency *row percentage*

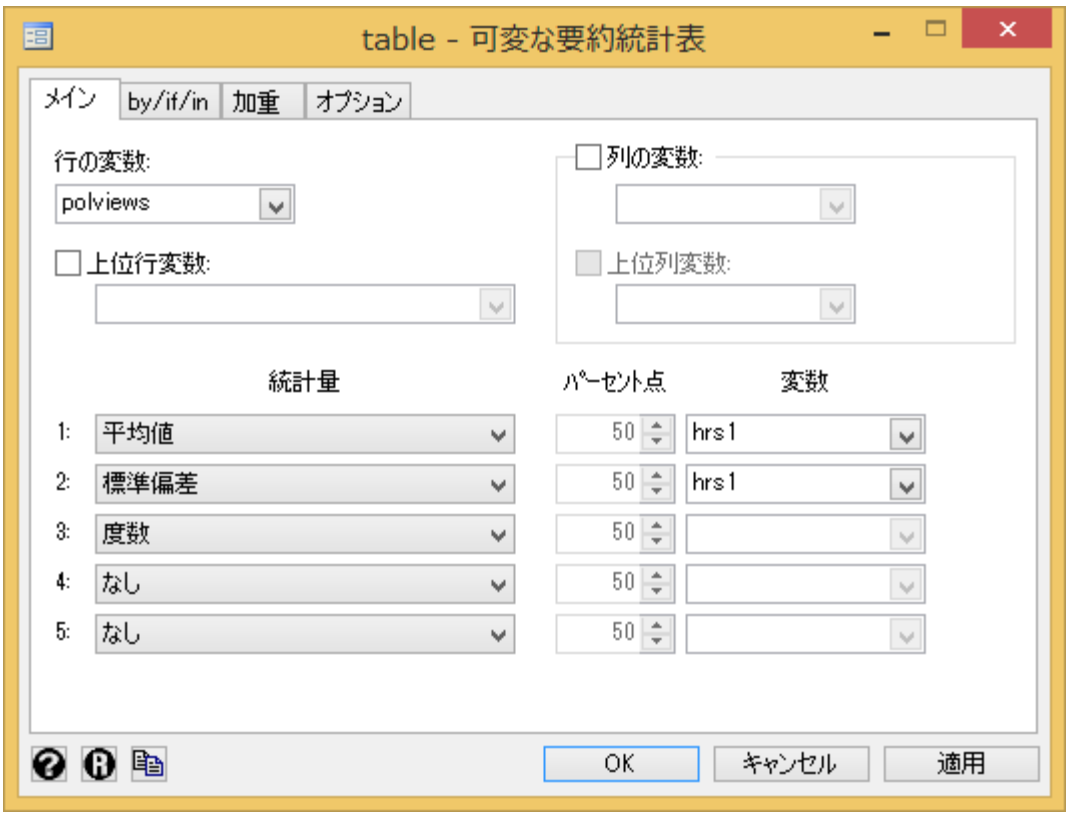
THINK OF SELF AS LIBERAL OR CONSERVATIVE	SEX BEFORE MARRIAGE ALWAYS WR	ALMST ALW	SOMETIMES	NOT WRONG	Total
EXTREMELY LIBERAL	17	2	7	43	69
	24.64	2.90	10.14	62.32	100.00
LIBERAL	28	15	40	144	227
	12.33	6.61	17.62	63.44	100.00
SLIGHTLY LIBERAL	30	15	54	128	227
	13.22	6.61	23.79	56.39	100.00
MODERATE	170	57	140	346	713
	23.84	7.99	19.64	48.53	100.00
SLGHTLY CONSERVATIVE	63	30	65	109	267
	23.60	11.24	24.34	40.82	100.00
CONSERVATIVE	135	38	55	78	306
	44.12	12.42	17.97	25.49	100.00
EXTRMLY CONSERVATIVE	43	5	7	17	72
	59.72	6.94	9.72	23.61	100.00
Total	486	162	368	865	1,881
	25.84	8.61	19.56	45.99	100.00

```
Pearson chi2(18) = 191.2380   Pr = 0.000
           gamma =  -0.3157  ASE = 0.025
 Kendall´s tau-b =  -0.2307  ASE = 0.019
```

政治的な意見と婚前交渉に対する考えには統計的に有意な関連が見られます．ガンマとタウ b の解釈では符号の解釈が必要です．マイナスの符号は，両変数間の負の関連を示します．表を見ると，極めてリベラルであることは，婚前交渉を間違いと考えないことと連関し，極めて保守的であることは，婚前交渉を常に間違いと考えることと連関しているのが見て取れます．負のガンマと負のタウ b は，リベラル度が強いほど婚前交渉に反対する意見が少ないという傾向を示します．政治的にリベラルな人ほど，婚前交渉を間違いとは思わないという関連の存在を，ガンマとタウ b を基にした結果の解釈から見出すことができます．

パーセント数は，この関連の性質についてより詳細な情報を伝えます．極めてリベラルな人は婚前交渉を常に間違いと考える人は少なく，およそ 25% に留まります．この数字は極めて保守的な人の間では，およそ 60% に増加します．一方，婚前交渉を全く間違いでないと考える人の割合を見ると，逆のパターンが見られます．極めて保守的な人たちの間では 24% に留まるのに対し，極めてリベラルでは 62% に増加します．

5. クロス表で，行数は 7，列数は 4 でした．従って，自由度は（行数 − 1）×（列数 − 1）= 18 です．
6. データセットを開き，Statistics ▷ Summaries, tables, and tests ▷ Other tables ▷ Flexible table of summary statistics（統計 (S) ▷ 要約/表/検定 ▷ その他の表 ▷ 要約統計量）を選択して要約統計量のダイアログボックスを開きます．Main（メイン）タブで，*Row variable*（行の変数）に `polviews` と入力します．下の方にある *Statistics*（統計量）で *Mean*（平均値）を選択し，右の *Variable*（変数）で `hrs1` を入力します．さらに次の行において，*Statistics*（統計量）で *Standard Deviation*（標準偏差）を選択し，右の *Variable*（変数）で `hrs1` を入力します．さらに下の行で，*Statistics*（統計量）で *Frequency*（度数）を選択し，右の *Variable*（変数）は空欄にします．ダイアログボックスは以下のようになります．



適用をクリックすると，次を得ます．

```
. table polviews, contents(mean hrs1 sd hrs1 freq )
```

think of self as liberal or conservative	mean(hrs1)	sd(hrs1)	Freq.
extremely liberal	37.8889	16.82565	47
liberal	43.8387	16.46359	143
slightly liberal	42.8364	11.87977	159
moderate	42.1548	13.45617	522
slghtly conservative	42.3643	16.22543	209
conservative	43.896	15.49392	210
extrmly conservative	37	10.25914	41

演習で指示された gss2002_chapter6.dta でなく gss2006_chapter6.dta を使用した場合は違う結果が得られます．あるいは，2002 年と 2006 年の間でどれだけ関連が異なるかを見てみて

も良いかもしれません．

表からは，週労働時間の平均が最も短かったグループは両派ともに急進グループであることを示しています．極めてリベラル（37.89 時間），極めて保守（37.00 時間）共に週労働時間の最も短かったグループです．

7. グラフは次のコマンドにより得られます．

```
. graph bar (mean) hrs1, over(polviews, label(angle(forty_five)))
```

グラフを公開する前に，読み手に分かりやすく加工しましょう．たとえば，すべて大文字の値ラベルは変更したほうが良いです．

8. Statistics ▷ Summaries, tables, and tests ▷ Frequency tables ▷ Table calculators（統計 (S) ▷ 要約/表/検定 ▷ 度数分布表 ▷ 分割表計算）を選択して `tabi` ダイアログボックスを開きます．マスの値（余白の小計は含めない）を行ごとに入力します．

```
234 568 38 \ 541 557 29
```

各数字はスペース区切り，改行は \ で入力します．さらに *Pearson's chi-squared*（ピアソンのカイ二乗），*Cramer's V*（クラメールの V），*Within-row relative frequencies*（行内の相対度数）を選択します．

関連には $\chi^2(2) = 82.82, p < 0.001$ と高い有意性が示されます．クラメールの V は 0.21 と中程度の関連を示します．行内のパーセント数からは，ポルノの非合法化に賛成の割合が男性で 28% ，女性で 48% であると分かります．

9. 変数 `health` の再コード化には専用のダイアログボックスが利用できます．Data ▷ Create or change data ▷ Other variable-transformation commands ▷ Recode categorical variable（データ ▷ データの作成または変更 ▷ その他の変数変換コマンド ▷ カテゴリ変数を再コード化）を選択して `recode` ダイアログボックスを開きます．Main（メイン）タブで，*Variables*（変数）に `health` と入力し，*Required*（必須）に `(1/2=2 satisfactory) (3/4=1 unsatisfactory)` と入力します．Options（オプション）タブで，*Generate new variables*（変数を新規作成する）を選択し，新たな変数名として `health2` を入力します．何かしらの変更を加える際は常に新たな変数を作成するようにします．こうすることで，間違ったときのために元のデータを保存しておけます．健康状態が良いほうを高い得点 2，悪いほうを低い得点 1 として記録しました．では，確認してみます．`tabulate health health2, missing` を実行すると，クロス分割表を表示して，変更を意図通り行えたか確認ができます．

次に表を作成します．Statistics ▷ Summaries, tables, and tests ▷ Frequency tables ▷ Two-way

tables with measures of association（統計 (S) ▷ 要約/表/検定 ▷ 度数分布表 ▷ 二元配置の表/統計値）を選択してダイアログボックスを開きます．*Row variable*（行の変数）に独立変数 `sex` を入力し，ます．下の方にある *Statistics*（統計量）で *Mean*（平均値）を選択し，*Column variable*（列の変数）に `health2` と入力します．*Test statistics*（検定統計量）の中から *Pearson's chi-squared*（ピアソンのカイ二乗），*Cramer's V*（クラメールの V）を選択します．*Cell contents*（セルの内容）の中からは，行に独立変数があるので *Within-row relative frequencies*（行内の相対度数）を選択します．

```
. tabulate sex health2, chi2 row V

+-------------------+
| Key               |
|-------------------|
|   frequency       |
| row percentage    |
+-------------------+

             |   RECODE of health
             |  (CONDITION OF HEALTH)
      Gender |  unsatisfa  satisfact |     Total
-------------+----------------------+----------
        MALE |       370      1,199 |     1,569
             |     23.58      76.42 |    100.00
-------------+----------------------+----------
      FEMALE |       529      1,418 |     1,947
             |     27.17      72.83 |    100.00
-------------+----------------------+----------
       Total |       899      2,617 |     3,516
             |     25.57      74.43 |    100.00
          Pearson chi2(1) =   5.8777   Pr = 0.015
               Cramr´s V =  -0.0409
```

不満の残る健康状態となるオッズ比は男性で $370/1199 = 0.309$，女性で $529/1418 = 0.373$ です．オッズ比は男性対女性で $0.309/0.373 = 0.828$ です．オッズ比が 1 より小さいときは，$100 \times (1 - OR) = 100 \times (1 - 0.828) = 17.2\%$ を計算します．男性が健康状態に不満と答えるオッズは，女性より 17.2% 低くなりました．この結果は，男性のほうが女性より健康という意味ではなく，あくまで調査での回答の話です．この図式は健康状態が不満な男性が 23.6% に留まるのに対し，女性が 27.2% に上ることに表れています．χ^2 値は有意であり，男女間の差は有意であることを示しています．クラメールの $V = -0.04$ は，弱い関連を示しています．調査員により連関を示すのに使用する統計量はクラメールの V など様々ですが，中でもパーセント数とオッズ比は明確な全体像を示してくれます．ここでは，男性が健康状態に不満と答えるオッズは，女性より 17.2% 低いという構図です．

第 6 章 (6.11 節, pp.157-158) の do-file

演習 6.1

```
/************** Begin do-file ************/
* chapter6.1.do
use "C:\data\gss2006_chapter6.dta", clear
codebook pornlaw
codebook, compact
tabulate sex pornlaw, row
label variable pornlaw "Pornography should be legal to"
label define newporn 1 "Nobody" 2 "Over 18" 3 "All"
label values pornlaw newporn
tabulate sex pornlaw, row
/************** End do-file **************/
```

演習 6.2

```
/************** Begin do-file ************/
* chapter6.2.do
use "C:\data\gss2006_chapter6.dta", clear
tabulate sex pornlaw, chi2 row V
/************** End do-file **************/
```

演習 6.3

```
/************** Begin do-file ************/
* chapter6.3.do
use "C:\data\gss2006_chapter6.dta", clear
codebook pres00 pres04
tabulate pres00 pres04, chi2 row V
tabulate pres00 pres04, chi2 row V miss
tabulate pres00 pres04 if (pres00 == 1 | pres00 == 2) & ///
(pres04 == 1 | pres04 == 2), chi2 row V
/************** End do-file **************/
```

演習 6.4

```
/************** Begin do-file ************/
* chapter6.4.do
use "C:\data\gss2006_chapter6.dta", clear
tabulate polviews premarsx, chi2 gamma row taub
/************** End do-file **************/
```

演習 6.6

```
/************** Begin do-file ************/
* chapter6.6.do
use "C:\data\gss2002_chapter6.dta", clear
table polviews, contents(mean hrs1 sd hrs1 freq)
/************** End do-file **************/
```

演習 6.7

```
/************** Begin do-file ************/
* chapter6.7.do
use "C:\data\gss2002_chapter6.dta", clear
graph bar (mean) hrs1, over(polviews, label(angle(forty_five)))
/************** End do-file **************/
```

演習 6.8

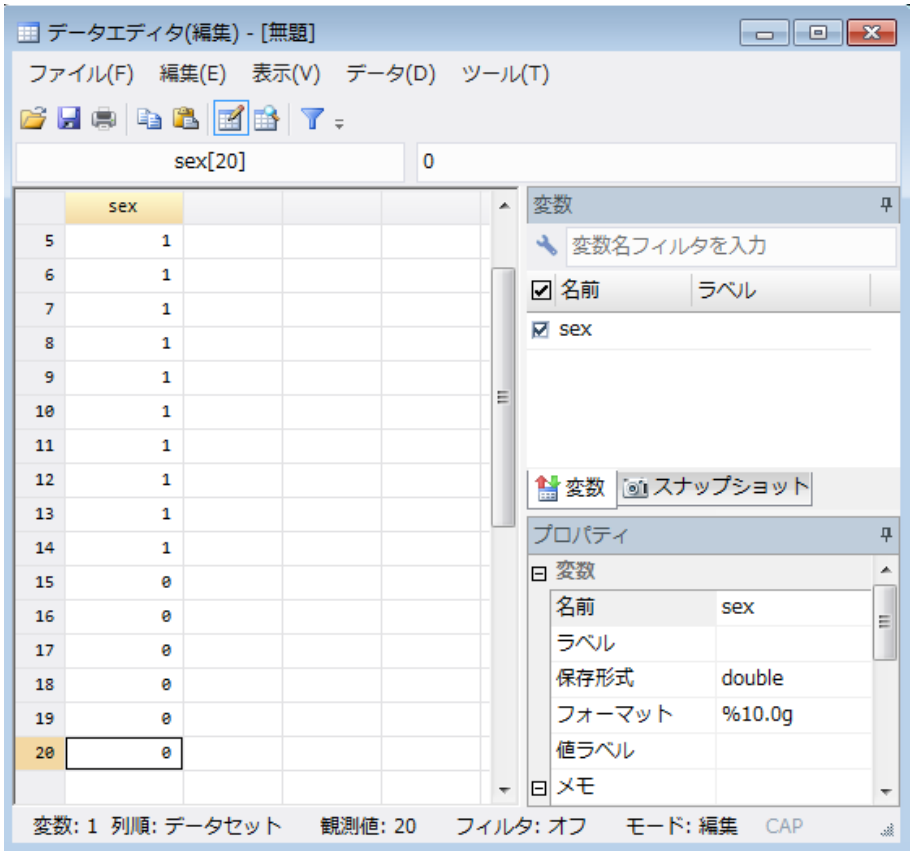
```
/************** Begin do-file ************/
* chapter6.8.do
tabi 234 568 38  541 557 29, chi2 row V
/************** End do-file **************/
```

演習 6.9

```
/************** Begin do-file ************/
*chapter6.9.do
clear
use "C:\data\gss2006_chapter6.dta", clear
recode health (1/2=2 satisfactory) (3/4=1 unsatisfactory), generate(health2)
tabulate health health2, missing
tabulate sex health2, row V chi2
/************** End do-file **************/
```

第 7 章 (7.13 節, pp.196-198) の解答

1. まず，データセットを作成する必要があります．データエディタで作成すると，以下のようになります．



ここでは変数名を sex としています．データの入力が完了したら，データをメモリ上に展開するためにデータエディタを閉じなければなりません．データエディタウィンドウの右上にある **X** をクリックし，Stata のメインウィンドウに戻ります．これでデータがメモリ上に置かれ，解析が可能になります．Stata は RAM 上で動作するので，データはまだハードディスクには保存されていません．作成したデータは，後で読み出せるようにファイルに保存しておきます．コマンドウィンドウで，`save "C:\data\c7_1.dta"`を実行します．作成する do ファイルでの先頭は，このデータセットを読み込むステップとします．

一群の比率検定を行うために，Statistics ▷ Summaries, tables, and tests ▷ Classical tests of hypotheses ▷ Proportion test（統計 (S) ▷ 要約/表/検定 ▷ 古典的仮説検定 ▷ 比率検定）を選択します．Main（メイン）タブで，*Variable*（変数）に sex と入力し，*Hypothesized proportion*（比率 (仮定)）に.52 と入力します．Submit（適用）をクリックします．結果は以下のようになります．

```
. prtest sex == 0.52
One-sample test of proportion                    sex: Number of obs =       20
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.                     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |          .7   .1024695                      .4991635    .9008365
------------------------------------------------------------------------------
    p = proportion(sex)                                           z =   1.6113
Ho: p = 0.52
    Ha: p < 0.52                 Ha: p != 0.52                 Ha: p > 0.52
 Pr(Z < z) = 0.9464         Pr(|Z| > |z|) = 0.1071          Pr(Z > z) = 0.0536
```

結果では下部中央にある両側検定の結果を使用します．少し上にある帰無仮説は $p = 0.52$，対立仮説は $p \neq 0.52$ です．計算された z 値は $z = 1.61$ で p 値は 0.11 です．従って，今のサンプルは，ランダムに集められたときのサンプルとの間に顕著な違いはないと結論付けることができます．今のサンプルと同数の女性が偶然に集まる確率は 11% で，有意水準 0.05 においては帰無仮説を棄却できません．

2. たとえば以下のようなプログラムが書けます．

```
* c7_exercise_2.do
clear
set obs 30
generate id = _n
list
set seed 100
sample 15, count
list
```

まず，clear でメモリ上のデータセットを一掃します．次に，set obs 30 コマンドで 30 観測分のスペースを作ります．今問題になっている人数は合計で 30 人です．generate id = _n コマンドは id という名前の変数を作成します．変数 id の値は，list の結果からも分かるように，1 から 30 までの行番号です．結果に再現性を持たせるため，set seed 100 を実行します．ここでシード値に違う数字を指定すると，得られる結果も変わってきます．sample 15, count で，観測を 15 個，復元なしに無作為抽出します．選択したほうを処理群，そうでないほうを対照群としても，またはその逆としても，どちらでも構いません．以下は，上記のプログラムで選んだ 15 個の観測です．

```
. list

     +-----+
     |  id |
     |-----|
  1. |   5 |
  2. |  26 |
  3. |  13 |
  4. |  20 |
  5. |  16 |
     |-----|
  6. |   8 |
  7. |   2 |
  8. |   4 |
  9. |  18 |
 10. |  28 |
     |-----|
 11. |  14 |
 12. |   7 |
 13. |  11 |
 14. |  21 |
 15. |   9 |
     +-----+
```

3. この演習の作業は，演習 2 での do ファイルの `sample 15, count` を `bsample 15` に代えることで実施できます．（`bsample` コマンドでは `count` オプションを指定しません．）確認のコマンドも変数 `id` に関する `tabulate` に変更します．その結果，以下が得られます．

```
. tabulate id
         id |      Freq.     Percent        Cum.
------------+-----------------------------------
          2 |          2       13.33       13.33
          4 |          1        6.67       20.00
          6 |          2       13.33       33.33
         11 |          1        6.67       40.00
         13 |          2       13.33       53.33
         16 |          1        6.67       60.00
         21 |          1        6.67       66.67
         22 |          1        6.67       73.33
         23 |          1        6.67       80.00
         24 |          1        6.67       86.67
         25 |          1        6.67       93.33
         29 |          1        6.67      100.00
------------+-----------------------------------
      Total |         15      100.00
```

上記のように，母集団に対してそれなりの大きさのあるサンプルを，復元ありで抽出する場合，問題が発生します．`id` が 2,6,13 に対応する人は 2 回選ばれています．

4. 以下，コマンドです．

```
* c7_exercise_4.do
clear
set obs 200
```

```
generate id = _n
set seed 953
sample 10, count
list
```

そして，10 個の観測は以下です．

```
. list

     +-----+
     |  id |
     |-----|
  1. |  75 |
  2. |  47 |
  3. | 154 |
  4. |  50 |
  5. | 157 |
     |-----|
  6. | 156 |
  7. | 180 |
  8. |  52 |
  9. |  66 |
 10. | 124 |
     +-----+
```

シード値に 953 を指定したので，何度実施しても同じ結果が得られます．

5. この演習で用いるのはグループ間における t 検定です．Statistics ▷ Summaries, tables, and tests ▷ Classical tests of hypotheses ▷ t test (mean-conparison test)（統計 (S) ▷ 要約/表/検定 ▷ 古典的仮説検定 ▷ t 検定 (平均比較検定)）を選択します．Main（メイン）タブで，*Two-sample using groups*（グループ変数による二群）を選択します．*Variable name*（変数名）に `hh18_97` と入力し，*Group variable name*（グループ変数名）に `ethnic97` と入力します．このダイアログボックス操作により，以下のような結果を得ます．

```
. ttest hh18_97, by(ethnic97)
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
Non-Hisp |   7,061    2.384365    .0148875    1.250996    2.355181    2.413549
Hispanic |   1,899    2.708268    .0314648    1.371158    2.646558    2.769977
---------+--------------------------------------------------------------------
combined |   8,960    2.453013    .0135665     1.28417     2.42642    2.479607
---------+--------------------------------------------------------------------
    diff |           -.3239027    .0330206               -.3886307   -.2591747
------------------------------------------------------------------------------
    diff = mean(Non-Hisp) - mean(Hispanic)                        t =  -9.8091
Ho: diff = 0                                     degrees of freedom =     8958
    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

この結果は仮説を支持しています．18 歳未満の子供の数の平均はヒスパニック系世帯で 2.71 人

であるのに対し，非ヒスパニック系世帯で 2.38 人です．平均間に見られる差異は -0.32 です．（対立仮説 H_a を (非ヒスパニック系の平均) $<$ (ヒスパニック系の平均) とする）片側検定のときは，結果で左下にある表記を用いて $t(8958) = -9.809, p < 0.001$ とし，差は有意性があることを示します．（対立仮説 H_a を (非ヒスパニック系の平均) $\neq$ (ヒスパニック系の平均) とする）両側検定のときは，結果で下部中央にある表記を用いてやはり $t(8958) = -9.809, p < 0.001$ となり，検定結果は有意となります．

標本の大きさが大変大きいので，実質的には重要でない小さな差も統計的に有意となり得ます．効果量の計算には，本体書籍の第 7 章に記載のあるコーエンの d の式を用います．計算時では，カッコの位置が正しくなるよう注意してください．以下，計算のコマンドとその結果です（コマンドウィンドウで入力し，数式をすべて入力し終えてから Enter を押してください）．

```
. display "Cohen´s d: = " (2.384-2.708)/sqrt((7061*(1.2510)^2 +
>  1899*(1.3712)^2)/8958)
Cohen´s d: = -.25360786
```

コーエンの d では 0.25 超の値は大きい効果量とされることは既に学びました．従って，ヒスパニック系世帯と非ヒスパニック世帯の間に見られる子供の数の平均差は，大きい効果量と分類します．

6. Statistics ▷ Summaries, tables, and tests ▷ Summary and descriptive statistics ▷ Summary statistics（統計 (S) ▷ 要約/表/検定 ▷ 記述統計量 ▷ 記述統計量）を選択して `summarize` ダイアログボックスを開きます．Main（メイン）タブで，*Variable name*（変数名）に `hh18_97` と入力し，*Options*（オプション）で *Display additional statistics*（追加の統計量を表示する）を選択します．by/if/in タブで，*Repeat command by groups*（グループごとにコマンドを実行する）を選択し，*Variables that define groups*（グループ変数）に `ethnic97` と指定します．また，*If*（条件式）ボックスに `ethnic97 < .` と入力します．

```
. by ethnic97, sort : summarize hh18_97 if ethnic97 < ., detail
-----------------------------------------------------------------------
-> ethnic97 = Non-Hisp
                 # in household < age 18 1997
-------------------------------------------------------------
      Percentiles      Smallest
 1%            1              0
 5%            1              0
10%            1              1       Obs               7,061
25%            2              1       Sum of Wgt.       7,061
50%            2                      Mean           2.384365
                        Largest       Std. Dev.      1.250996
75%            3              9
```

```
90%            4              9       Variance       1.564991
95%            5              9       Skewness       1.362753
99%            7             10       Kurtosis       5.984799
-------------------------------------------------------------
-> ethnic97 = Hispanic
                # in household < age 18 1997
-------------------------------------------------------------
      Percentiles      Smallest
 1%            1              0
 5%            1              1
10%            1              1       Obs               1,899
25%            2              1       Sum of Wgt.       1,899
50%            2                      Mean           2.708268
                        Largest       Std. Dev.      1.371158
75%            3              8
90%            5              8       Variance       1.880074
95%            5             12       Skewness       1.112575
99%            7             12       Kurtosis       5.562286
-------------------------------------------------------------
-> ethnic97 = .a
                # in household < age 18 1997
-------------------------------------------------------------
no observations
-------------------------------------------------------------
-> ethnic97 = .b
                # in household < age 18 1997
-------------------------------------------------------------
no observations
```

平均は異なりましたが，メディアン（中央値）を見ると，18 歳未満の子供の数のメディアンは，ヒスパニック系世帯で 2 人，非ヒスパニック系世帯で同じく 2 人でした．差はなさそうに見えます．`tabulate` を実行すると，以下を得ます．

```
. by ethnic97, sort: tabulate hh18_97 if ethnic < .
-------------------------------------------------------------------------------
-> ethnic97 = Non-Hisp
       # in
household <
age 18 1997 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |          2        0.03        0.03
          1 |      1,724       24.42       24.44
          2 |      2,670       37.81       62.26
          3 |      1,586       22.46       84.72
          4 |        676        9.57       94.29
          5 |        242        3.43       97.72
          6 |         73        1.03       98.75
          7 |         50        0.71       99.46
          8 |         31        0.44       99.90
          9 |          6        0.08       99.99
         10 |          1        0.01      100.00
------------+-----------------------------------
```

```
      Total |      7,061      100.00
-------------------------------------------------------------------------------
-> ethnic97 = Hispanic
       # in |
household <
age 18 1997 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |          1        0.05        0.05
          1 |        341       17.96       18.01
          2 |        628       33.07       51.08
          3 |        475       25.01       76.09
          4 |        255       13.43       89.52
          5 |        130        6.85       96.37
          6 |         49        2.58       98.95
          7 |         11        0.58       99.53
          8 |          7        0.37       99.89
         12 |          2        0.11      100.00
------------+-----------------------------------
      Total |      1,899      100.00
-------------------------------------------------------------------------------
-> ethnic97 = .a
no observations
-------------------------------------------------------------------------------
-> ethnic97 = .b
no observations
```

結果は両分布とも正に歪んでいることを示しています．Statistics ▷ Summaries, tables, and tests ▷ Nonparametric tests of hypotheses ▷ K-sample equality-of-medians test（統計 (S) ▷ 要約/表/検定 ▷ ノンパラメトリック仮説検定 ▷ k 標本中央値相等検定）を選択してメディアン検定をするダイアログボックスを開きます．Main（メイン）タブで，*Variable*（変数）に `hh18_97` と入力し，*Grouping variable*（グループ変数）に `ethnic97` と入力し，さらに *How to handle values equal to the median*（中央値と等しい値の扱い）で *Split equally between two groups*（両方のグループに等しく分ける）を選択します．結果は以下になります．

```
. median hh18_97, by(ethnic97) medianties(split)
Median test
   Greater |
  than the | youth ethnicity 1997
    median | Non-Hispa   Hispanic |     Total
-----------+----------------------+----------
        no |     3,067        650 |     3,717
       yes |     3,994      1,249 |     5,243
-----------+----------------------+----------
     Total |     7,061      1,899 |     8,960
          Pearson chi2(1) =  52.2619   Pr = 0.000
   Continuity corrected:
          Pearson chi2(1) =  51.8833   Pr = 0.000
```

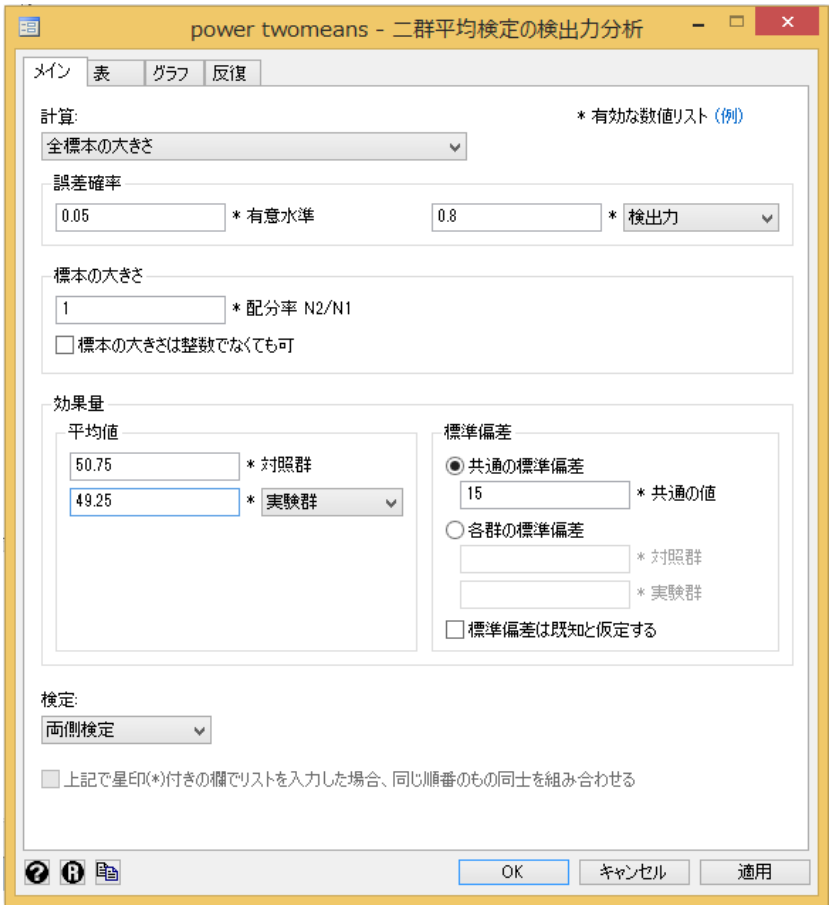
困ったことに，$\chi^2(1) = 52.26$, $p < 0.001$ となり，メディアン間に有意差が示されました．一

方で，両グループのメディアンは共に 2 であることに変わりありません．ここで問題なのは，メディアンが整数であり，且つ多くの観測が 2 という同じ値で同点となっていることです．平均値は 2.1 や 2.2 など小数になり得ますが，ここでのメディアンは整数しかあり得ません．`tabulate` の結果からヒスパニック系世帯では，2 人以下の子供を持つ割合が 51% であるのに対し，非ヒスパニック系世帯では 62% です．もし同点の観測が全くないとしたら，2 という値が 50 パーセント値により近いグループはヒスパニック系のほうになるはずです．3 人以上の子供を持つ世帯の割合はヒスパニック系で 49% であるのに対し，非ヒスパニック系は 38% です．今回のように同点が数多く存在するケース（子供の数が 2 人である世帯について，その世帯数は非ヒスパニック系で 2,670）ではメディアン検定は問題含みとなります．

7. 調査の得点について，平均を 50 付近，標準偏差を 15 と仮定します．こうした仮定は事前の調査結果，または同じ尺度を用いた別の調査結果を基にします．ここでは $\alpha = 0.05$ を有意水準とします．コーエンの d について，着目する最小の値を 0.10 とします．0.10 は中程度の効果としては最小の値です．コーエンの d が 0.10 のとき，共和党員と民主党員の得点が標準偏差 0.10 個分だけ異なる，つまり $0.10 \times 15 = 1.5$ だけ異なるということになります．従って，ここで見たい最小の平均差は 1.5 ということになります．

   以上の情報を踏まえ，検出力分析を行います．全体の平均が 50，二群の平均差が 1.5 であることから，暫定的に片方の平均を 49.25，もう片方を 50.75 と設定します．（$50.75 - 49.25 = 1.5$ となるからです．）

   Statistics ▷ Power and sample size（統計 (S) ▷ 検出力/標本の大きさ）を選択して `power` ダイアログボックスを開きます．開いたダイアログボックスの左枠で Means（平均値）を選択し，右枠で Test comparing two independent means（2 つの独立した平均を比較する）を選択します．開いたダイアログボックスで先ほど取り決めた情報と知りたい情報を入力します．入力後のダイアログボックスは以下のようになります．



ダイアログでは，*Total sample size*（全標本の大きさ）が選択され，*Signficance level*（有意水準）が 0.05，*Power*（検出力）が 0.8 の状態を維持します．両群の大きさは同じものと仮定し，*Allocation ratio, N2/N1*（配分率 N2/N1）は 1 のままとします．$d = 0.10$（平均差に換算して 1.5）というかなり小さな効果量を検出したいので，*Means*（平均値）には 49.25 と 50.75 を入力します．両群の標準偏差は 15.0 で共通とします．最下部では，*Two-sided test*（両側検定）が選択されたままにします．

OK をクリックすると，以下の結果が得られます．

```
. power twomeans 49.25 50.75, sd(15)
Performing iteration ...
Estimated sample sizes for a two-sample means test
```

```
t test assuming sd1 = sd2 = sd
Ho: m2 = m1  versus  Ha: m2 != m1
Study parameters:
        alpha =    0.0500
        power =    0.8000
        delta =    1.5000
           m1 =   49.2500
           m2 =   50.7500
           sd =   15.0000
Estimated sample sizes:
            N =      3142
  N per group =      1571
```

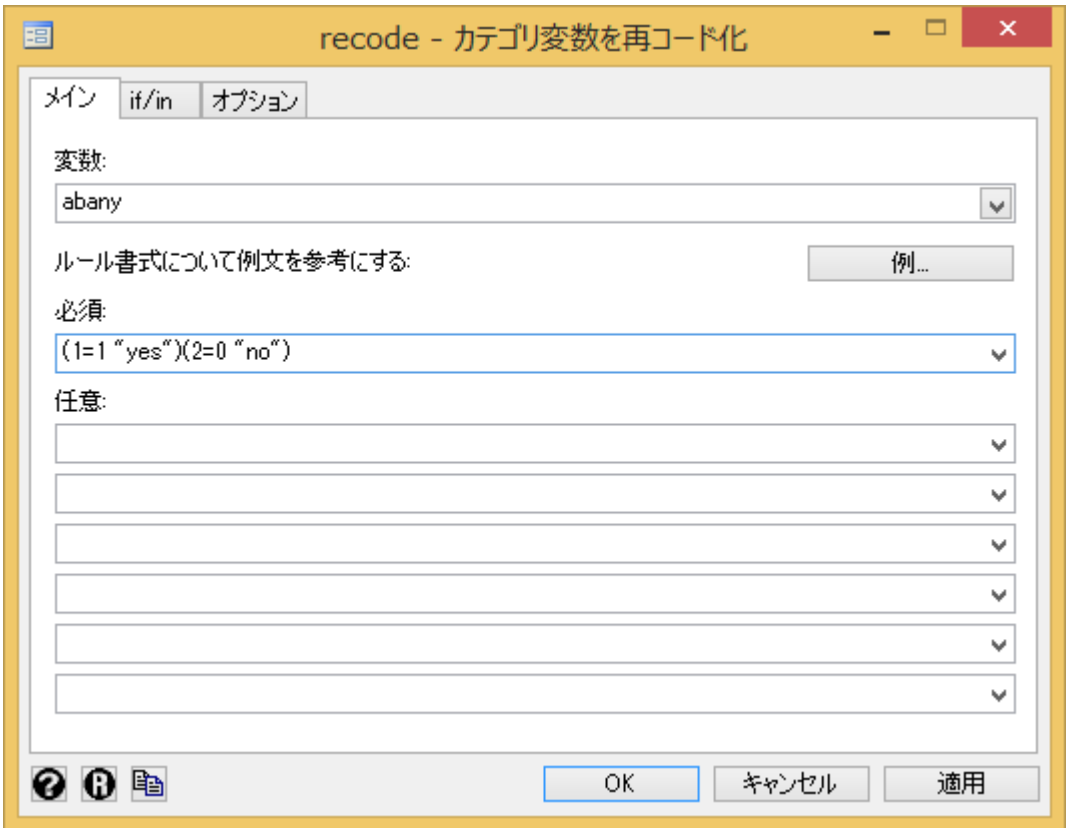
共和党員と民主党員に共に 1,571 人，計 3,142 人が必要になることが分かります．母数の比率がどちらかに傾いている場合，配分率を調整します．

0.90 の検出力が必要な場合，各 2,103 人，計 4,206 人が必要になり，0.99 の検出力が必要な場合，各 3,676 人，計 7,352 人が必要になります．

ヒント：検出力分析は有用ですが，分析者が設定する仮定次第で，結果が大きく変わってきます．ここでは 0.80，0.90，0.99 などとした検出力の設定が，必要な標本の大きさに極めて劇的な影響を与えました．効果量については 0.10，平均差に換算して 1.5 というかなり小さな効果量を検出できるようにしました．これがもし 0.50，すなわち標準偏差の半分といった大きな差にのみに関心がある場合どうなるのでしょう．この場合，各群の平均を 46.25 と 53.75 にし，その差が $0.50 \times 15 = 7.5$ となるようにします．結果，必要な標本の大きさは各群で 148，計 296 となります．このように，仮定する値の変更が必要な標本の大きさに大きな影響が及ぶことがあります．

8. どのような種類の検定をすべきか，その決定のため変数にはどのような値が格納されているのかを確認する必要があります．Data ▷ Describe data ▷ Describe data contents (codebook)（データ (D) ▷ データの内容表示 ▷ データの内容表示 (codebook)）を選択し，`codebook sex abany` コマンドを生成します．コマンドの実行結果より，変数 `sex` では男性が 1，女性が 2 でコード化されており，変数 `abany` では人工中絶に賛成の意見が 1，反対が 2 でコード化されているのが分かります．どちらも 2 値変数なので，比率検定が行えます．ここで知りたいのは，人工中絶に賛成する割合が男性より女性のほうが高いかどうかです．`abany` は 1 と 2 でコード化されていますが，0 と 1 でコード化される必要があります．`abany` のコード系を 0,1 から 1,2 へと変更する必要があります．変更の方法はいくつかあります．ひとつは，メニューで Data ▷ Create or change data ▷ Other variable-transformation commands ▷ Recode categorical variable（データ ▷ データの作成または変更 ▷ その他の変数変換コマンド ▷ カテゴリ変数を再コード化）

を選択して変更する方法です．ダイアログボックスが開きますので，Main（メイン）タブで，*Variables*（変数）に abany と入力し，再コード化のルールを入力します．以下の画像では正しい値ラベルの指定も共に行っています．



もしダイアログボックスで Submit（適用）をクリックした場合，Stata では既存の変数の上書き時にラベルの再定義を許可していないため，エラーメッセージが出ます．また，データセットの元の値も変更しないようにしたいところです．新たなコードは，新たに変数を作成して格納します．Options（オプション）タブで，*Generate new variables*（変数を新規作成する）を選択し，新たな変数名として abortok を入力します．tabulate で変換結果を確認すると，以下のようになります．

```
. recode abany (1=1 "yes") (2=0 "no"), generate(abortok)
```

```
(513 differences between abany and abortok)
. tabulate abany abortok
  abortion |     RECODE of abany
  if woman |  (abortion if woman
 wants for |  wants for any reason)
any reason |        no        yes |     Total
-----------+----------------------+----------
       yes |         0        387 |       387
        no |       513          0 |       513
-----------+----------------------+----------
     Total |       513        387 |       900
```

再コード化を行うもうひとつの方法は，やはり本体書籍の第 7 章に記載のあるやり方です．まず，abany と同じ値を持つ変数を作成し，その後，値を変更します．コマンドは，以下です．

```
. generate abrtok = abany
. replace abrtok = 0 if abany == 1
```

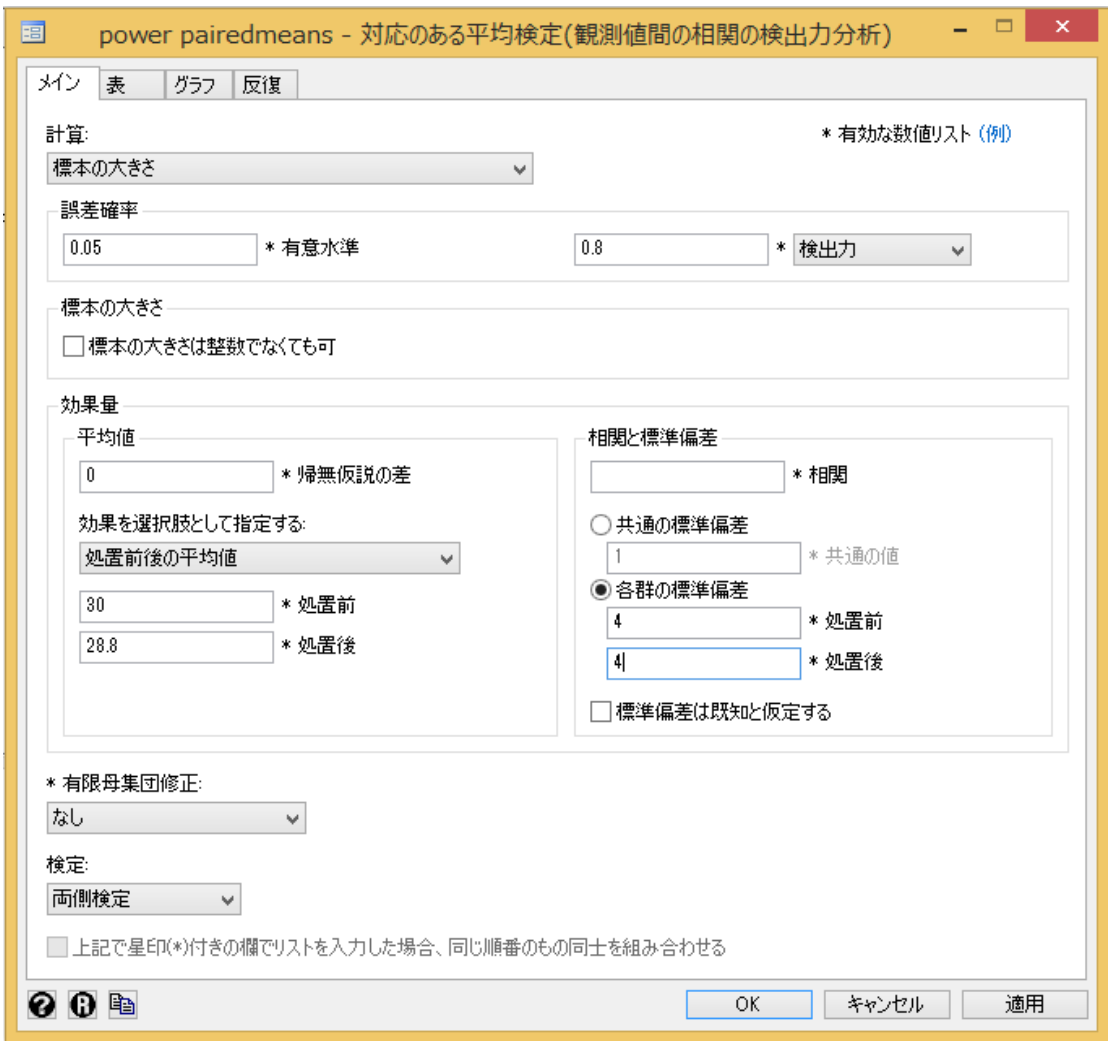
ここでは実際に両例を示すために先ほどと異なる変数名を用いています．abrtok に対し tabulate を実行すると，abany や abortok に付いているような値ラベルがないことが分かります．abrtok に値ラベルを割り当てても良いのですが，通常は 1 というコードを変数名が真となるほうに割り当てるので，1 が人工中絶を OK とするほう，0 が OK としないほうであると判断できます．比率検定をするためには，二標本比率検定なのかグループ比率検定なのかを決める必要があります．ここでは，変数が 1 つ (abortok) と，グループ変数が 1 つ (男性か女性か) のデータがあるため，後者であるグループ比率検定を行います．Statistics ▷ Summaries, tables, and tests ▷ Classical tests of hypotheses ▷ Proportion test（統計 (S) ▷ 要約/表/検定 ▷ 古典的仮説検定 ▷ 比率検定）を選択して，適切なダイアログを開きます．ダイアログボックスで，*Two-group using groups*（グループ変数による二群）を選択します．*Variable name*（変数名）に abortok と入力し，*Group variable name*（グループ変数名）に sex と入力します．結果は，以下になります．

```
. prtest abortok, by(sex)
Two-sample test of proportions                    male: Number of obs =      484
                                                female: Number of obs =      416
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .4442149   .0225854                      .3999484    .4884814
      female |   .4134615   .0241446                      .3661391     .460784
-------------+----------------------------------------------------------------
        diff |   .0307533   .0330614                     -.0340459    .0955526
             |  under Ho:   .0330997     0.93   0.353
------------------------------------------------------------------------------
        diff = prop(male) - prop(female)                          z =   0.9291
    Ho: diff = 0
```

```
     Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
  Pr(Z < z) = 0.8236         Pr(|Z| > |z|) = 0.3528          Pr(Z > z) = 0.1764
```

当初，理由の如何によらず人口中絶を支持する比率は，男性よりも女性のほうが高いと予想しました．もし片側検定を行った場合，対立仮説 H_a は (女性での比率) $>$ (男性での比率) です．ところが，計算された比率は男性で 0.44 なのに対し，女性で 0.41 であるため，上記の対立仮説は採択されないことになります．計算結果は，当初の予想とは逆向きでした．もし両側検定を行った場合，対立仮説 H_a は (女性での比率) $\neq$ (男性での比率) となり，`prtest` の結果で下部中央にある記述を用いることになります．$z = 0.93, p = 0.35$ となり，結果は統計的に有意とはなりません．

9. この演習で用いる対応のある t 検定の検出力分析は，独立な 2 つの平均の差の例までを記述する本体書籍の内容を超えるものです．まず，Statistics ▷ Power and sample size（統計 (S) ▷ 検出力/標本の大きさ）を選択して `power` ダイアログボックスを開きます．開いたダイアログボックスの左枠で Means（平均値）を選択し，右枠で Paired test comparing two correlated means, specify correlation between paired observations（相関のある 2 つの平均を比較する対応のある検定，対応のある観測値間の相関を設定する）を選択します．開いたダイアログボックスで，以下のように相応しい情報を入力します．



入力では，α の値を 0.05，検出力を 0.8 とします．演習の質問文では，平均 30，標準偏差は 4 であるとしています．さらに効果量は中程度としています．Cohen (1988) では，$d = 0.1$ を弱い効果量，$d = 0.3$ を中程度の効果量，$d = 0.5$ を強い効果量としています．そこで，中程度を示す $d = 0.3$ を標準偏差の 4 とかけ算して，検出する検査結果の差の最小値を得ます．従って，後の検査の平均は $30 - 1.2$ より 28.8 となります．

Effect size（効果量）セクションでは，*pre and posttreatment means*（処置前後の平均値）を選択したままとし，*Pretreatment*（処置前）に 30，*Posttreatment*（処置後）に 28.8 と入力しま

す．*Correlation and standard deviations*（相関と標準偏差）セクションでは，前後の検査間の相関として 0.6 を入力します．この値は，今回の状況でかなり強い相関が期待されるため，妥当な値です．もし独立な 2 つの平均の t 検定を行う場合，0.0 とすべきところです．相関に関する入力の下では，*Group standard deviations*（各群の標準偏差）を選択し，*Pretreatment*（処置前）と *Posttreatment*（処置後）それぞれに 4 と入力します．上記の情報の入力後，Submit（適用）をクリックすると，以下の結果を得ます．

```
. power pairedmeans 30 28.8, corr(0.6) sd1(4) sd2(4)
Performing iteration ...
Estimated sample size for a two-sample paired-means test
Paired t test
Ho: d = d0  versus  Ha: d != d0
Study parameters:
        alpha =    0.0500          ma1 =   30.0000
        power =    0.8000          ma2 =   28.8000
        delta =   -0.3354          sd1 =    4.0000
           d0 =    0.0000          sd2 =    4.0000
           da =   -1.2000         corr =    0.6000
         sd_d =    3.5777
Estimated sample size:
            N =        72
```

従って，前後の検査で参加を必要とする人数はそれぞれに 72 人です．

10. 調査で男性と女性は独立と仮定します．つまり，男女間で結婚，同棲など，対関係はないとします．満足度を測る 20 項目を合計すると，20（すなわち 20×1）から 100(すなわち 20×5) までに分布します．すべての項目に満足あるいは不満足である人はほぼいないと考えられます．ここでは，大部分 (95%) の人は 30 から 90 までに分布すると仮定してみます．正規分布はシグマが 4 つ分に相当する範囲に全体の 95% が分布するので，標準偏差を $(90-30)/4=15$ と概算できます．ここでは小さい効果量を検出したいので，$d=0.2$ となり，得点に換算する場合 $d\times$SD，すなわち $0.2\times 15=3.0$ です．その差が 3.0 であれば，2 つの平均はどんな値でも大丈夫です．ここでは，70 と 67 としましょう．（ほかにも 50 と 53，あるいは 84 と 81 としても同じ結果になります．）また，男女は同じ人数ずついるとしましょう．上記をまとめると，以下のようになります．

```
alpha = 0.05
power = 0.90
mean1 = 70
mean2 = 67
sd1 = 15
sd2 = 15
n2/n1 = 1.0 (same N for women and men)
```

Statistics ▷ Power and sample size（統計 (S) ▷ 検出力/標本の大きさ）を選択して power ダイアログボックスを開きます．開いたダイアログボックスの左枠で Means（平均値）を選択し，右枠で Test comparing two independent means（2 つの独立した平均を比較する）を選択します．開いたダイアログボックスで先ほど取り決めた情報と知りたい情報を入力します．入力後のダイアログボックスは以下のようになります．
実行結果は以下です．

```
. power twomeans 70 67, sd(15) power(0.9)
Performing iteration ...
Estimated sample sizes for a two-sample means test
t test assuming sd1 = sd2 = sd
Ho: m2 = m1  versus  Ha: m2 != m1
Study parameters:
        alpha =    0.0500
        power =    0.9000
        delta =   -3.0000
           m1 =   70.0000
           m2 =   67.0000
           sd =   15.0000
Estimated sample sizes:
            N =      1054
  N per group =       527
```

男女それぞれ 527 人，計 1,054 人という大人数が必要なことが分かりました．

11. この演習では何をどう仮定すれば良いでしょうか．使用するダイアログボックスは演習 10 と同じで，有意水準は 0.05，検出力は 0.80 とし，両側検定を行います．テスト時の平均は 80 点付近だったので，平均の一つは 80 とできます．中程度またはそれ以上の効果量に注目するので，$d = 0.5$ となります．標準偏差は両群で共に 10 とできますので，得点へ換算すると $d\times$SD により $0.5 \times 10 = 5.0$ となります．従って，もう一つの平均は $80 + 5$ で 85 です．最後に，両群は共に同じ人数を有することにし，$N2/N1 = 1.0$ とします．

```
alpha = 0.05 (two-side)
power = 0.80
mean1 = 80
mean2 = 85
sd1 = 10
sd2 = 10
n2/n1 = 1.0
```

Statistics ▷ Power and sample size（統計 (S) ▷ 検出力/標本の大きさ）を選択して power ダイアログボックスを開きます．開いたダイアログボックスの左枠で Means（平均値）を選択し，右枠で Test comparing two independent means（2 つの独立した平均を比較する）を選択します．

開いたダイアログボックスで先ほど取り決めた情報と知りたい情報を入力します．入力後のダイアログボックスは以下のようになります．

実行結果は以下です．

```
. power twomeans 80 85, sd(10)
Performing iteration ...
Estimated sample sizes for a two-sample means test
t test assuming sd1 = sd2 = sd
Ho: m2 = m1  versus  Ha: m2 != m1
Study parameters:
        alpha =    0.0500
        power =    0.8000
        delta =    5.0000
           m1 =   80.0000
           m2 =   85.0000
           sd =   10.0000
Estimated sample sizes:
            N =       128
  N per group =        64
```

有意水準 0.05，検出力 80% で中程度の効果量を検出するには，各群に 63 人ずつの生徒が必要になります．

12. 片側検定と両側検定でそれぞれに必要な標本の大きさを分析するには，演習 11 で用いたダイアログボックスを使用します．ダイアログボックスの左下には，両側検定か片側検定を選択するオプションがあります．先ほどの演習 11 では各群 64 人，計 128 人が必要だった両側検定も，正当な理由のもと，片側検定に変更すれば，各 51 人，計 102 人で済みます．

    処置群の人数には，コスト面から調整が難しいという問題があります．処置群の人数が対照群に比べて少なくなるような状況は，高く信頼性で平均値を推定したい通常の調査では好ましくありません．

    より費用の掛からない群に対し，費用の掛かる群の 2 倍の人数を割り当てたい場合，配分率に 2 と入力します．演習 11 では，各群 64 人，計 128 人でしたが，配分率を 2 にすると，全体の人数は N=144 へと増加しますが，処置群だけを見ると，48 人へと必要人数が減少します．こうした事柄を踏まえ，対費用効果が最も良くなる人数と配分率を探ることもできます．

    上記に加え，片側検定を行うことにすると，必要人数は全体的に減少します．（処置群に 38 人，全体で 114 人）

第 7 章 (7.13 節, pp.196-198) の do-file

演習 7.1

```
/************** Begin do-file ************/
* chapter7.1.do
use "C:\data\c7_1.dta"
prtest sex == .52
/************** End do-file **************/
```

演習 7.2

```
/************** Begin do-file ************/
* chapter7.2.do
clear
set obs 30
generate id = _n
list
set seed 100
sample 15, count
list
/************** End do-file **************/
```

演習 7.3

```
/************** Begin do-file ************/
* chapter7.3.do
clear
set obs 30
generate id = _n
list
set seed 100
bsample 15
list
tabulate id
/************** End do-file **************/
```

演習 7.4

```
/************** Begin do-file ************/
* chapter7.4.do
clear
set obs 200
generate id = _n
set seed 953
sample 10, count
list
```

```
/************** End do-file **************/
```

演習 7.5

```
/************** Begin do-file ************/
* chapter7.5.do
use "C:\data\nlsy97_chapter7.dta", clear
ttest hh18_97, by(ethnic97)
display "Cohen's d: = " (2.384-2.708) / sqrt((7061*(1.2510)^2 + ///
1899*(1.3712)^2)/8958)
/************** End do-file **************/
```

演習 7.6

```
/************** Begin do-file ************/
* chapter7.6.do
use "C:\data\nlsy97_chapter7.dta", clear
by ethnic97, sort: summarize hh18_97 if ethnic97<., detail
median hh18_97, by(ethnic97) medianties(split)
/************** End do-file **************/
```

第 8 章 (8.9 節, pp.223-224) の解答

1. 相関を計算するには，Statistics ▷ Summaries, tables, and tests ▷ Summary and descriptive statistics ▷ Correlations and covariances（統計 (S) ▷ 要約/表/検定 ▷ 記述統計量 ▷ 相関と共分散）を選択して，適切なダイアログを開きます．*Variable*（変数）ボックスに `educ` と `hrs1` を入力し，以下のような結果を得ます．

```
. correlate educ hrs1
(obs=6,387)
               |     educ     hrs1
---------------+------------------
        educ   |   1.0000
        hrs1   |   0.0835   1.0000
```

相関を男女別に計算するには，先ほどのダイアログの by/if/in タブで，*Repeat command by groups*（グループごとにコマンドを実行する）を選択し，`sex` と指定します．すると，以下の結果を得ます．

```
. by sex, sort : correlate educ hrs1
-------------------------------------------------------------------------------
-> sex = MALE
(obs=3,195)
               |     educ     hrs1
---------------+------------------
        educ   |   1.0000
        hrs1   |   0.0681   1.0000
-------------------------------------------------------------------------------
-> sex = FEMALE
(obs=3,192)
               |     educ     hrs1
---------------+------------------
        educ   |   1.0000
        hrs1   |   0.1220   1.0000
```

結果からは，教育期間と労働時間には弱い相関しかありませんが，男女間で比べると女性のほうが男性よりも強い相関であることが分かります．教育は，男性の労働時間よりも女性の労働時間のほうに強い連関があります．

回帰を行うには，Statistics ▷ Linear models and related ▷ Linear regression（統計 (S) ▷ 線形モデル他 ▷ 線形回帰）を選択します．*Dependent variable*（従属変数）に `hrs1` と入力し，*Independent variable*（独立変数）に `educ` と入力します．さらに Report（レポート）タブで *Standardized beta coefficients*（標準化済み β 係数）を選択します．これらの手続きにより，以下の結果を得ます．

```
. regress hrs1 educ, beta
      Source |       SS           df       MS      Number of obs   =     6,387
-------------+----------------------------------   F(1, 6385)      =     44.86
       Model |  9117.1584         1  9117.1584   Prob > F        =    0.0000
    Residual |  1297645.62     6,385  203.233456   R-squared       =    0.0070
-------------+----------------------------------   Adj R-squared   =    0.0068
       Total |  1306762.77     6,386  204.62931   Root MSE        =    14.256

------------------------------------------------------------------------------
        hrs1 |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
        educ |   .4120023   .0615131     6.70   0.000                .0835279
       _cons |   36.41293   .8661046    42.04   0.000                       .
------------------------------------------------------------------------------
```

男女別に回帰を行うには，線形回帰ダイアログボックスに戻り，by/if/in タブをクリックし，*Repeat command by groups*（グループごとにコマンドを実行する）を選択し，*Variables that define groups*（グループ変数）に sex と指定します．結果は，以下です．

```
. by sex, sort : regress hrs1 educ, beta
-------------------------------------------------------------------------------
-> sex = MALE
      Source |       SS           df       MS      Number of obs   =     3,195
-------------+----------------------------------   F(1, 3193)      =     14.86
       Model |  3051.83297         1  3051.83297   Prob > F        =    0.0001
    Residual |  655855.542     3,193  205.404179   R-squared       =    0.0046
-------------+----------------------------------   Adj R-squared   =    0.0043
       Total |  658907.375     3,194  206.295359   Root MSE        =    14.332

------------------------------------------------------------------------------
        hrs1 |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
        educ |   .3189366   .0827425     3.85   0.000                .0680563
       _cons |   40.80528   1.160458    35.16   0.000                       .
------------------------------------------------------------------------------

-------------------------------------------------------------------------------
-> sex = FEMALE
      Source |       SS           df       MS      Number of obs   =     3,192
-------------+----------------------------------   F(1, 3190)      =     48.17
       Model |   8733.9907         1   8733.9907   Prob > F        =    0.0000
    Residual |  578446.896     3,190  181.331315   R-squared       =    0.0149
-------------+----------------------------------   Adj R-squared   =    0.0146
       Total |  587180.887     3,191   184.01156   Root MSE        =    13.466

------------------------------------------------------------------------------
        hrs1 |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
        educ |   .6076469   .0875551     6.94   0.000                .1219608
       _cons |   30.57771   1.237583    24.71   0.000                       .
------------------------------------------------------------------------------
```

全体での回帰の結果を見ると，educ に対し，$B = 0.41$ であり，教育期間が 1 年増加するごとに，見積もりで 0.41 時間の週労働時間の増加に繋がります．これに対し，男性というサブグループ

だけで回帰を行うと，B は有意性が保たれるものの，先ほどより小さい $B = 0.32$ と，より弱い教育の効果となります．一方，女性については，$B = 0.32, p < 0.001$ となり，より強い教育の効果となります．この調査結果は，演習の本文に書かれた主張に反するもののようです．実は，教育期間の増加が週労働時間に与える効果は，女性のほうが男性よりも大きくなります．たとえば，教育期間が 4 年増加すると，男性では労働時間が $4 \times 0.32 = 1.28$ 時間 (約 90 分) 増加するのに対し，女性では $4 \times 0.61 = 2.44$ 時間 (約 140 分) 増加するという予測が導かれます．

2. `describe` を使用して，配偶者についての先週の労働時間が `sphrs1` に収まっていることを理解します．演習 1 と同じダイアログボックスを使用し，by/if/in タブも使用して，全体の分析と変数 `sex` を入力した男女別の分析を行います．以下，全体の分析結果です．

```
. correlate hrs1 sphrs1
(obs=2,393)
             |     hrs1   sphrs1
-------------+------------------
        hrs1 |   1.0000
      sphrs1 |   0.0177   1.0000
```

男女別に実施すると，以下を得ます．

```
. by sex, sort : correlate hrs1 sphrs1
-------------------------------------------------------------------------------
-> sex = MALE
(obs=1,136)
             |     hrs1   sphrs1
-------------+------------------
        hrs1 |   1.0000
      sphrs1 |   0.0974   1.0000
-------------------------------------------------------------------------------
-> sex = FEMALE
(obs=1,257)
             |     hrs1   sphrs1
-------------+------------------
        hrs1 |   1.0000
      sphrs1 |   0.1922   1.0000
```

全体の分析結果ではとても弱い相関しか見られません．一方，男女別では女性に男性より強い相関が見られました．女性から報告される自身と夫の労働時間の間には，男性から報告されるものよりも高い類似性があります．

演習 1 と同様に回帰を行うと，全体の結果は以下になります．

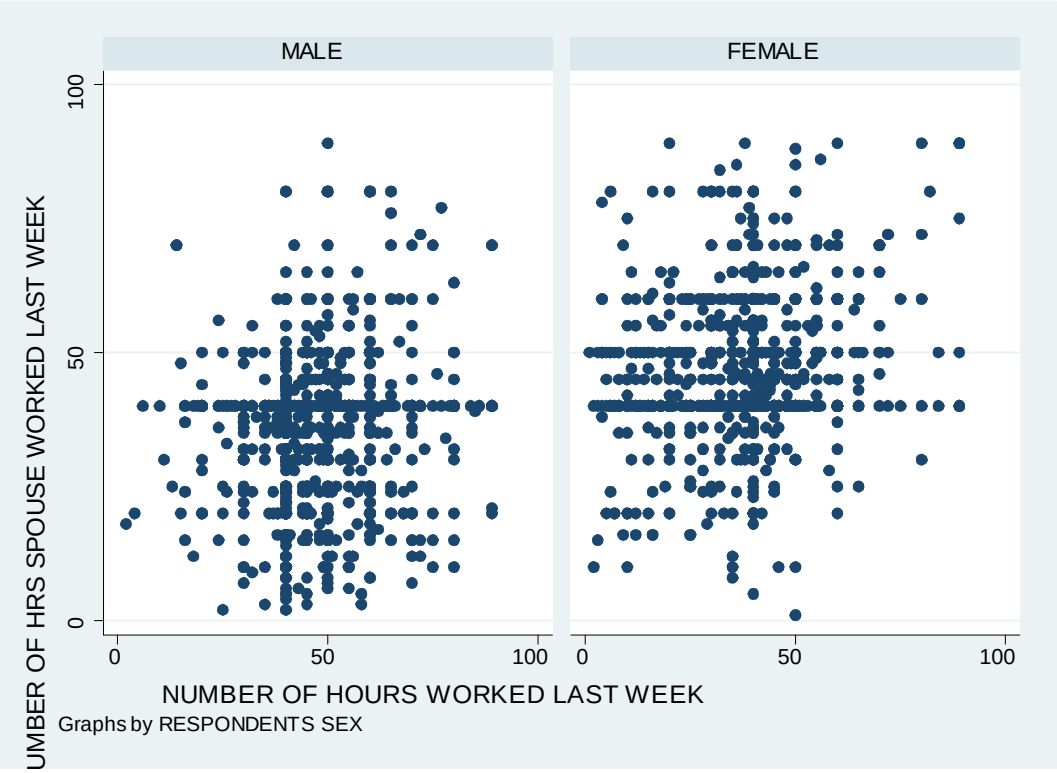
```
. regress sphrs1 hrs1, beta
      Source |       SS           df       MS      Number of obs   =     2,393
-------------+----------------------------------   F(1, 2391)      =      0.75
```

```
      Model |  130.596748         1  130.596748   Prob > F        =    0.3872
   Residual |  417393.396     2,391  174.568547   R-squared       =    0.0003
-------------+------------------------------------ Adj R-squared   =   -0.0001
      Total |  417523.992     2,392  174.550164   Root MSE        =    13.212

------------------------------------------------------------------------------
     sphrs1 |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
       hrs1 |   .0164893   .0190642     0.86   0.387                 .0176858
      _cons |    40.7701   .8522983    47.84   0.000                        .
------------------------------------------------------------------------------
```

男女別については，以下を得ます．

```
. by sex, sort : regress sphrs1 hrs1, beta
-------------------------------------------------------------------------------
-> sex = MALE
      Source |       SS           df       MS      Number of obs   =     1,136
-------------+------------------------------------ F(1, 1134)      =     10.86
      Model |  1752.10969         1  1752.10969   Prob > F        =    0.0010
   Residual |  182950.073     1,134  161.331634   R-squared       =    0.0095
-------------+------------------------------------ Adj R-squared   =    0.0086
      Total |  184702.183     1,135  162.733201   Root MSE        =    12.702

------------------------------------------------------------------------------
     sphrs1 |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
       hrs1 |   .0991445   .0300848     3.30   0.001                 .0973968
      _cons |   31.98229   1.472278    21.72   0.000                        .
------------------------------------------------------------------------------

-------------------------------------------------------------------------------
-> sex = FEMALE
      Source |       SS           df       MS      Number of obs   =     1,257
-------------+------------------------------------ F(1, 1255)      =     48.16
      Model |   6765.1294         1   6765.1294   Prob > F        =    0.0000
   Residual |  176296.727     1,255   140.47548   R-squared       =    0.0370
-------------+------------------------------------ Adj R-squared   =    0.0362
      Total |  183061.857     1,256  145.749886   Root MSE        =    11.852

------------------------------------------------------------------------------
     sphrs1 |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
       hrs1 |   .1644874   .0237025     6.94   0.000                 .1922379
      _cons |   39.55875   .9600615    41.20   0.000                        .
------------------------------------------------------------------------------
```

散布図を描くには，Graphics ▷ Twoway graph (scatter, line, etc.)（グラフィックス (G) ▷ 二元グラフ (散布図/折れ線など)）を選択します．*Create...*（作成...）を選択し，*Basic plots*（基本的なプロット）から *Scatter*（散布図）を指定します．*Y variable*（y 変数）に `sphrs1` を指定し，*X variable*（x 変数）に `hrs1` を指定します．OK をクリックします．By（by 条件）タブで，*Draw subgrahs for unique values of variables*（変数のユニーク値ごとにサブグラフを作成する）を選択し，*Variable*（変数）に `sex` を入力します．



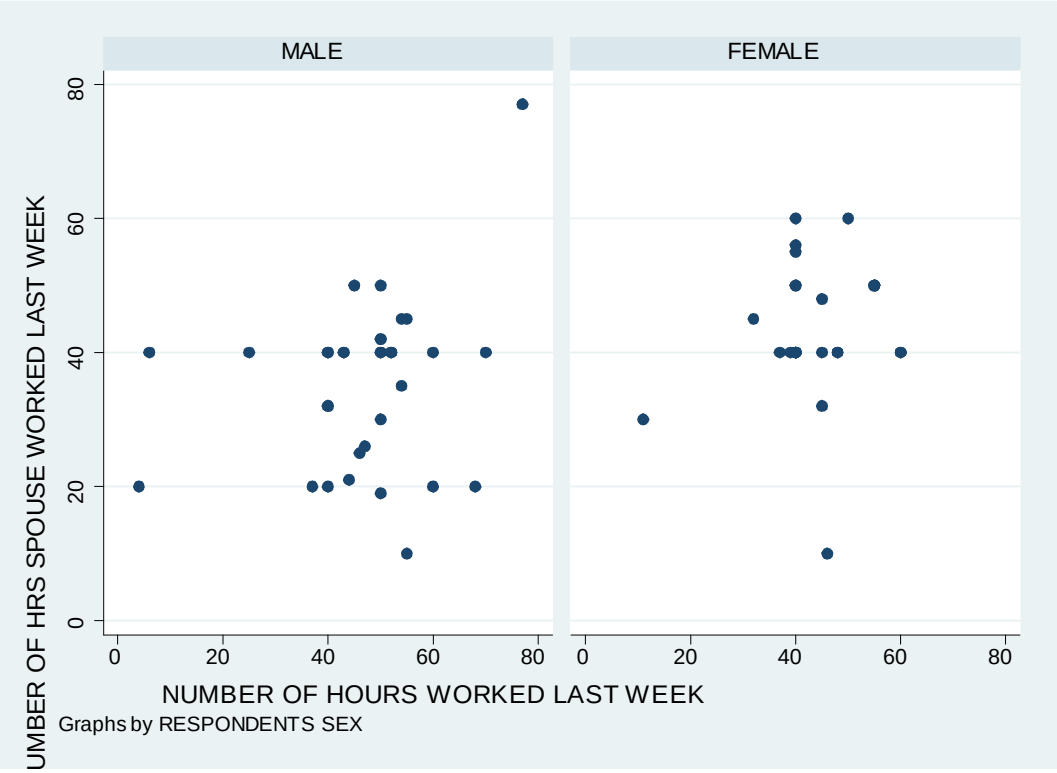
データが多いと，散布図の解釈も容易ではありません．ですが，女性からのデータである右側の図を見ると，見た目には何ら関係性を示さない左側の図に比べ，いくらかの関係を示していることが感じ取れます．

3. まずシード値を設定し，その後 250 人の参加者のデータを集めます．コマンドは以下になります．

```
. set seed 111
. sample 250, count
```

次に，演習 2 の `twoway` ダイアログボックスを使用し，250 人のデータに対して同様の処理を行います．

```
. twoway (scatter sphrs1 hrs1), by(sex)
```



観測データをランダムに削減すると，グラフが明解になることが時々あります．データを全て使った場合，たとえば週労働時間での 40 時間など，同値のデータが多数あって，全体像が描かれないといった問題が生じることがあります．しかし，ここでのデータでは 2 変数間に弱い相関しかないので，ここで描かれた男女別のグラフからは役立つ情報は得られません．

4. `correlate` で計算を行うには, Statistics ▷ Summaries, table, and tests ▷ Summary and descriptive statistics ▷ Correlations and covariances（統計 (S) ▷ 要約/表/検定 ▷ 記述統計量 ▷ 相関と共分散）を選択してダイアログボックスを開きます．*Variables*（変数）に `happy`，`hapmar`，`health` と入力し，Submit（適用）をクリックします．すると，以下のような結果を得ます．

```
. correlate happy hapmar health
(obs=2,481)
```

	happy	hapmar	health
happy	1.0000		
hapmar	0.4536	1.0000	
health	0.2159	0.1615	1.0000

コマンドの表示のすぐ下に，(obs=2481) とあります．これは，今行われた計算がリストワイズ除去（ケースワイズ除去とも呼ばれる）を用いていることを示しており，3 つの質問すべてに回答した人は 2,481 人だったことを表しています．

次にペアワイズ相関分析を行うため，Statistics ▷ Summaries, table, and tests ▷ Summary and descriptive statistics ▷ Pairwise correlations（統計 (S) ▷ 要約/表/検定 ▷ 記述統計量 ▷ 対相関）を選択します．開いたダイアログボックスで *Variables*（変数）に先ほどと同様の変数（happy，hapmar，health）を入力し，Submit（適用）をクリックします．

```
. pwcorr happy hapmar health
             |   happy   hapmar   health
-------------+---------------------------
       happy |  1.0000
      hapmar |  0.4547   1.0000
      health |  0.2437   0.1619   1.0000
```

上記の結果で，各数字はペアワイズ相関であり，先ほどのリストワイズ相関と異なります．その理由は，ペアワイズ相関が，分析の対象となった 2 つの質問に両方とも回答した人のデータを使用して計算しているからです．happy と hapmar のペアワイズ相関は 0.45 です．この数字は，全般的な幸福度，また結婚生活の幸福度の両方の質問に回答した人すべてのデータに基づいて計算されています．その中には，健康に関する質問に回答していない人も含まれていますが，このペアワイズ相関においては問題としません．

相関行列での問題点は，各相関の計算で用いた回答者数が全く分からない点です．また，各相関の有意性の有無も知りたいところです．これらの情報を得るには，Main（メイン）タブで *Print number of observations for each entry*（各エントリーに観測数を表示する），*Print significance level for each entry*（各エントリーに有意水準を表示する）を選択します．さらに *Use listwise deletion to handle missing values*（欠損値をリストワイズ除去する）も選択すると，以下を得ます．

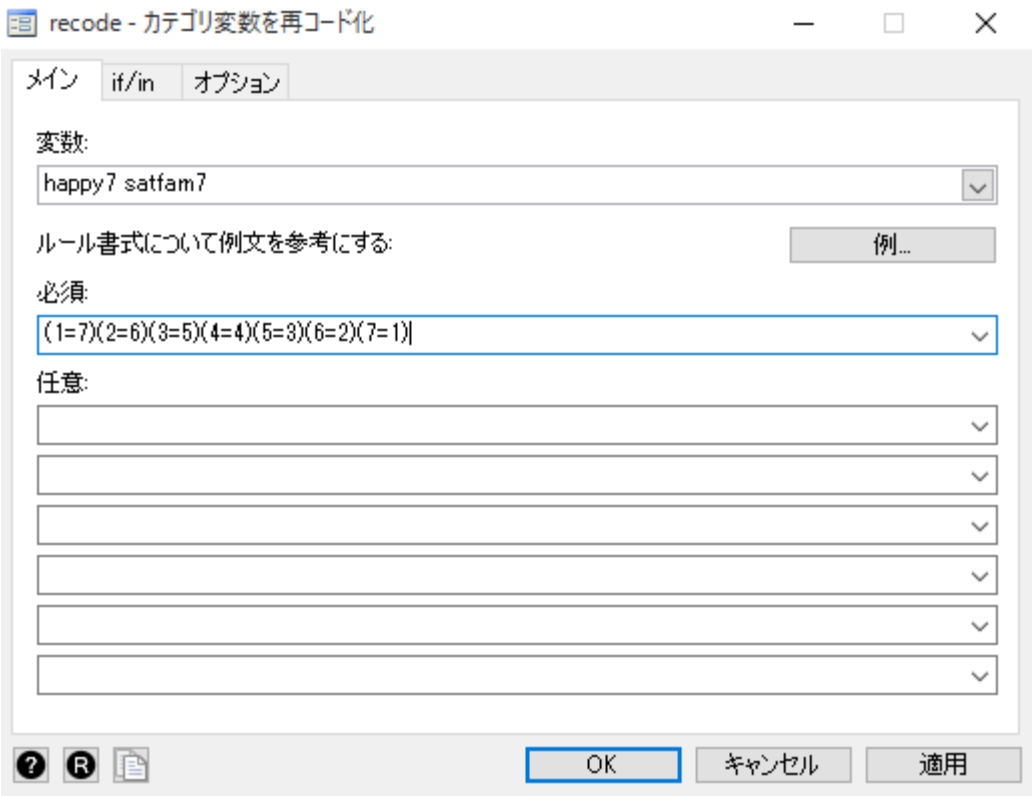
```
. pwcorr happy hapmar health, obs sig listwise
             |   happy   hapmar   health
-------------+---------------------------
       happy |  1.0000
             |
             |    2481
             |
      hapmar |  0.4536   1.0000
             |  0.0000
             |    2481     2481
             |
      health |  0.2159   0.1615   1.0000
             |  0.0000   0.0000
```

```
     |      2481      2481      2481
```

上記の試行から，リストワイズ/ケースワイズ除去とペアワイズ除去で相関が異なる理由が読み取れます．結婚生活の幸福度について回答した人は 3,392 人のみです．さらに全般的な幸福度にも回答した人については 3,386 人のみです．どちらのペアワイズ相関も $p < 0.001$ で統計的に有意です (上記のコマンド例で，`listwise` を除くと計算できます).

相関分析では時々，リストワイズ/ケースワイズ除去を行い且つ有意性を検定したい場合があります．これは，`pwcorr` コマンドで `listwise` オプションを指定すると達成できます．

5. 変数を再コード化する方法は数多くあります．ここでは `recode` を用います．Data ▷ Create or change data ▷ Other variable-transformation commands ▷ Recode cateorical variables（データ (D) ▷ データの作成または変更 ▷ その他の変数変換コマンド ▷ カテゴリ変数を再コード化）を選択してダイアログボックスを開きます．Main（メイン）タブで，再コード化のルールを入力します．



Options（オプション）タブで，*Generate new variables*（変数を新規作成する）を選択します．その後，どのように再コード化されたかを変数ラベルに記述しておくこともできます．たとえば，以下のような実施できます．

```
. recode happy7 satfam7 (1=7)(2=6)(3=5)(4=4)(5=3)(6=2)(7=1),
> generate(happynew satfamnew)
. label variable happynew
> "global happiness: 1 completely dissatisfied to 7 completely satisfied"
. label variable satfamnew
> "family satisfaction: 1 completely dissatisfied to 7 completely satisfied"
```

今度は回帰分析を行います．ここではコマンドと結果のみ掲載します．

```
. regress happynew satfamnew, beta
      Source |       SS           df       MS      Number of obs   =     1,143
-------------+----------------------------------   F(1, 1141)      =    840.15
       Model |  466.941968         1  466.941968   Prob > F        =    0.0000
    Residual |  634.146396     1,141  .555781241   R-squared       =    0.4241
```

```
                                                  Adj R-squared   =    0.4236
       Total | 1101.08836     1,142   .96417545   Root MSE        =    .74551

    happynew |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
   satfamnew |   .5559163   .0191792    28.99   0.000                 .651209
       _cons |    2.36949   .1110445    21.34   0.000                       .
```

regress の実行結果には，相関係数 r は表示されませんが，R-squared が 0.4241 と計算されています．説明変数が一つの場合，r は R^2 の平方根に等しくなります．また，同様に説明変数が一つの場合，β が r に等しくなります．

平方根の計算は，以下のようにして Stata 内蔵の計算機を用いて求めることもできます．

```
. display sqrt(.4241)
```

これにより $r = 0.651$ を得ます．説明変数が一つなので，β の値からも r が求まります．

回帰式は以下のように書けます．

$$Y \text{ の推定値} = B_0 + B_1 X$$

ここで Y は happynew の推定値，B_0 は定数または切片，B_1 は satfamnew（標準化前の傾き）です．

$$\texttt{happynew}\text{の推定値} = 2.369 + 0.556(\texttt{satfamnew})$$

分野によっては，式の中で定数（または切片）を後ろに置くこともあります．

定数（切片）は，X がゼロのときの Y の推定値です．家庭での幸福度をゼロと回答したとする人に対し，一般的な幸福度は 2.369 と推定されます．しかし，家庭での幸福度の最小値は 1（値の範囲は 1 から 7 まで）なので，この定数値に意味はありません．

傾きは 0.556 です．結婚生活の幸福度が 1.0 ポイント高ければ，全般的な幸福度が 0.556 上昇すると推定されます．家庭に全く満足しない人 (1) と家庭に全く満足している人 (7) を比較すれば，その間の関係を見ることができます．家庭に全く満足しない人に対する全般的な幸福度の推定は，$2.369 + 0.556(1) = 2.925$ となります．一方，家庭に全く満足している人に対する全般的な幸福度の推定は，$2.369 + .556(7) = 6.261$ となります．ここから明らかになるのは，結婚に満足か不満かが，全般的な幸福度に大きな影響を与えるということです．

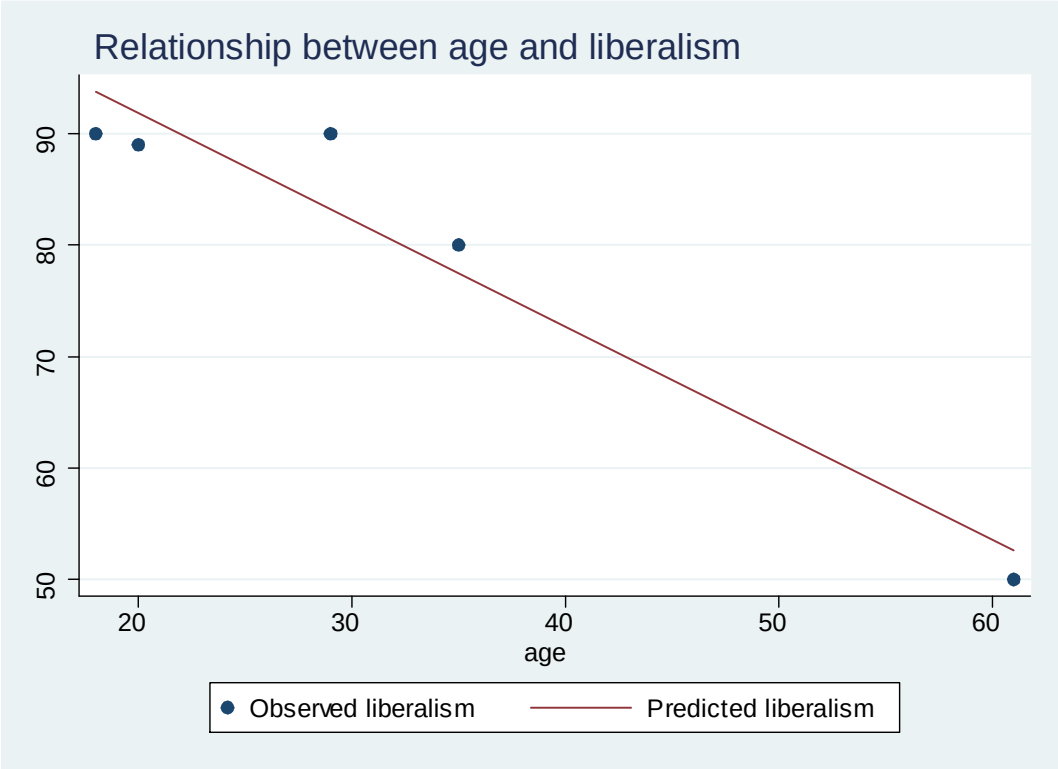
6. 散布図と回帰線を重ねて描くには，Graphics ▷ Twoway graph (scatter, line, etc.)（グラフィックス (G) ▷ 二元グラフ (散布図/折れ線など)）を選択します．*Create...*（作成...）を選択し，*Basic*

plots（基本的なプロット）から *Scatter*（散布図）を指定します．*Y variable*（y 変数）に `liberal` を指定し，*X variable*（x 変数）に `age` を指定します．OK をクリックして元のダイアログボックスに戻ります．二つ目のグラフを描くべく，再び *Create...*（作成...）を選択します．*Fit plots*（フィットプロット）から *Linear prediction*（線形予測）を指定します．*Y variable*（y 変数）と *X variable*（x 変数）にそれぞれ `liberal` と `age` を指定します．OK をクリックします．ここで Title（タイトル）タブで，グラフのタイトルを入力しても良いでしょう．また，Legend（タイトル）タブで，凡例を読みやすく変更しても良いでしょう．同タブの *Override default keys*（既定のキーを一時的に変更する）を選択し，ボックスに以下を入力します．

```
1 "Observed liberalism" 2 "Predicted liberalism"
```

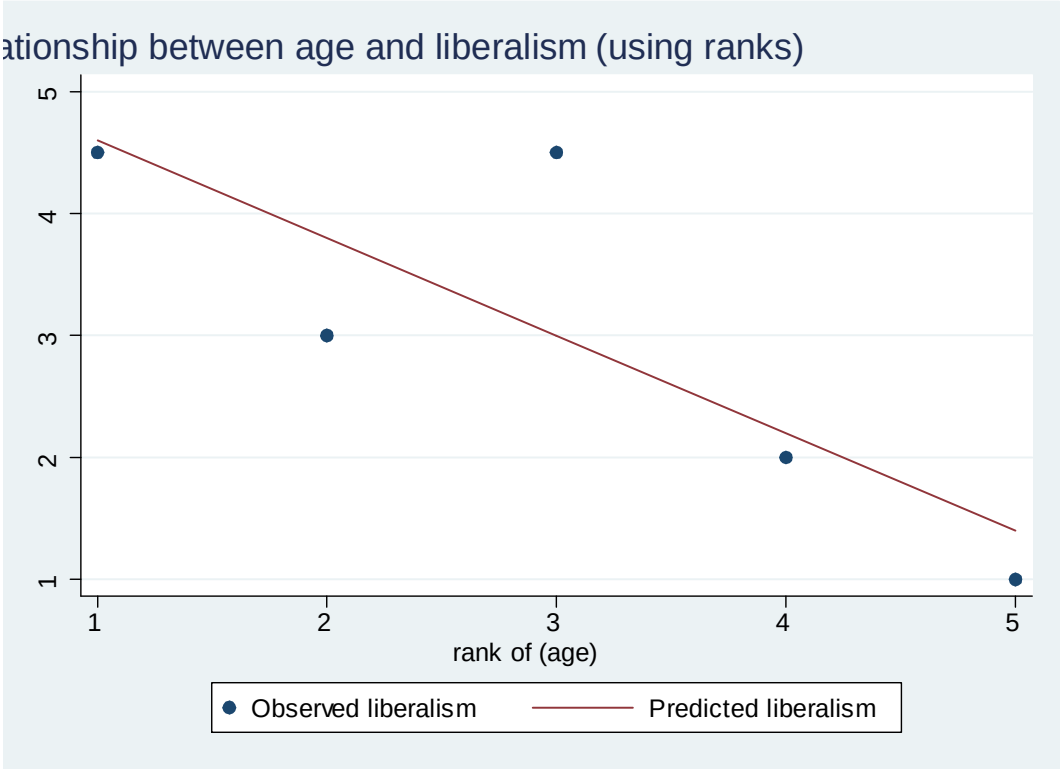
上記の指定によりプロット 1 に“Observed liberalism”，プロット 2 に“Predicted liberalism”と凡例が表示されます．コマンドとその結果は以下です．

```
. twoway (scatter liberal age) (lfit liberal age),
>  title("Relationship between age and liberalism")
>  legend(order(1 "Observed liberalism" 2 "Predicted liberalism"))
```



同様のプロセスを ranklib と rankage についても行います.

```
. twoway (scatter ranklib rankage) (lfit ranklib rankage),
>  title("Relationship between age and liberalism (using ranks)")
>  legend(order(1 "Observed liberalism" 2 "Predicted liberalism"))
```



このグラフを見ると少し戸惑います．同点の存在により，4 位および 5 位が不在となる代わりに 4.5 位が存在し，グラフにもそれらの順位に点がプロットされています．

```
. list
```

	age	liberal	rankage	ranklib
1.	18	90	1	4.5
2.	29	90	3	4.5
3.	35	80	4	2
4.	61	50	5	1
5.	20	89	2	3

次に，`liberal` と `age` の相関およびスピアマンの ρ を計算します．以下のように，それぞれ -0.965，-0.821 となります．

```
. correlate age liberal
(obs=5)
```

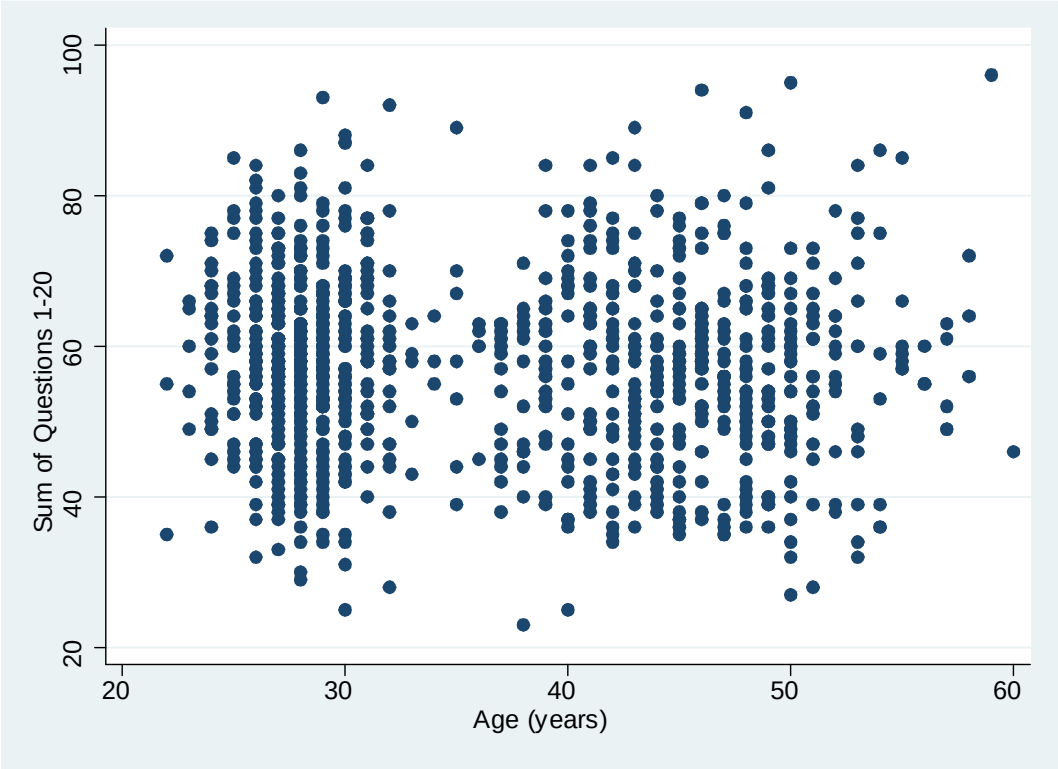
```
             |       age  liberal
-------------+------------------
         age |    1.0000
     liberal |   -0.9650   1.0000

. spearman age liberal

 Number of obs =       5
Spearman´s rho =      -0.8208

Test of Ho: age and liberal are independent
    Prob > |t| =       0.0886
```

7. ((訂正) 本演習については本体書籍のコマンドの記述に誤植がありました．以下のように訂正するとともに，心からお詫び申し上げます．)

   (訂正コマンド) . `use http://www.stata-press.com/data/r13/depression.dta`

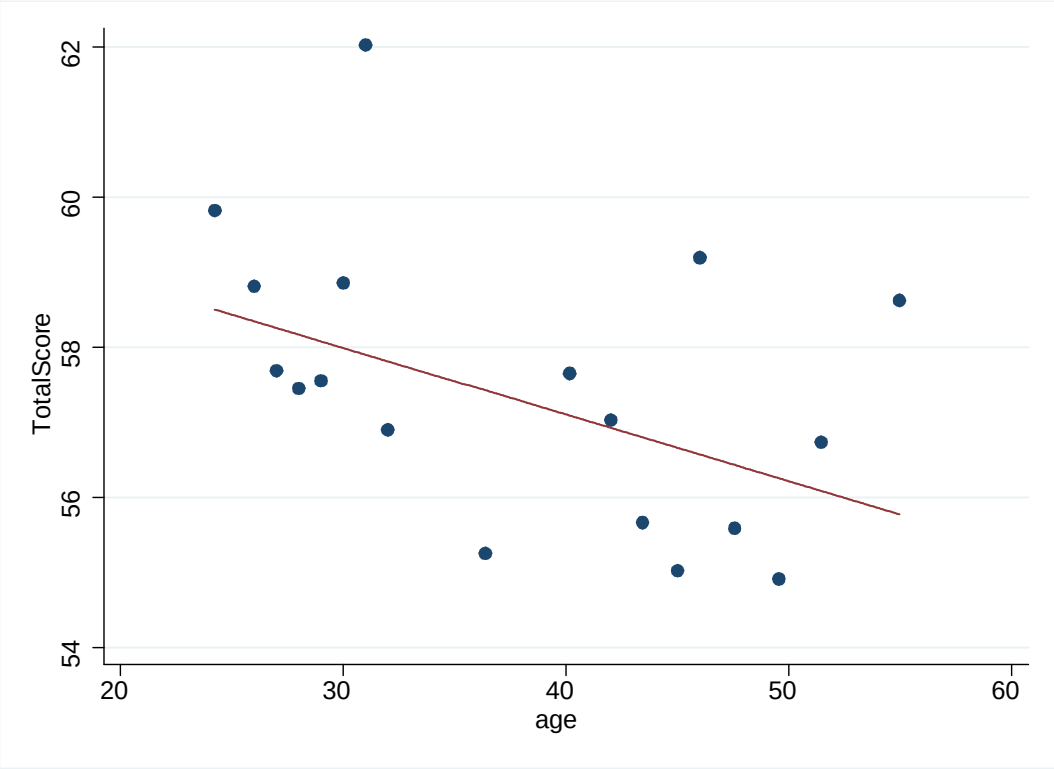
   ここではメニューではなくコマンドを使用しましょう．まず，セッションをクリアし，データセットを開いて散布図を描きます．

```
. clear

. use http://www.stata-press.com/data/r13/depression.dta
. twoway (scatter TotalScore age, sort)
```



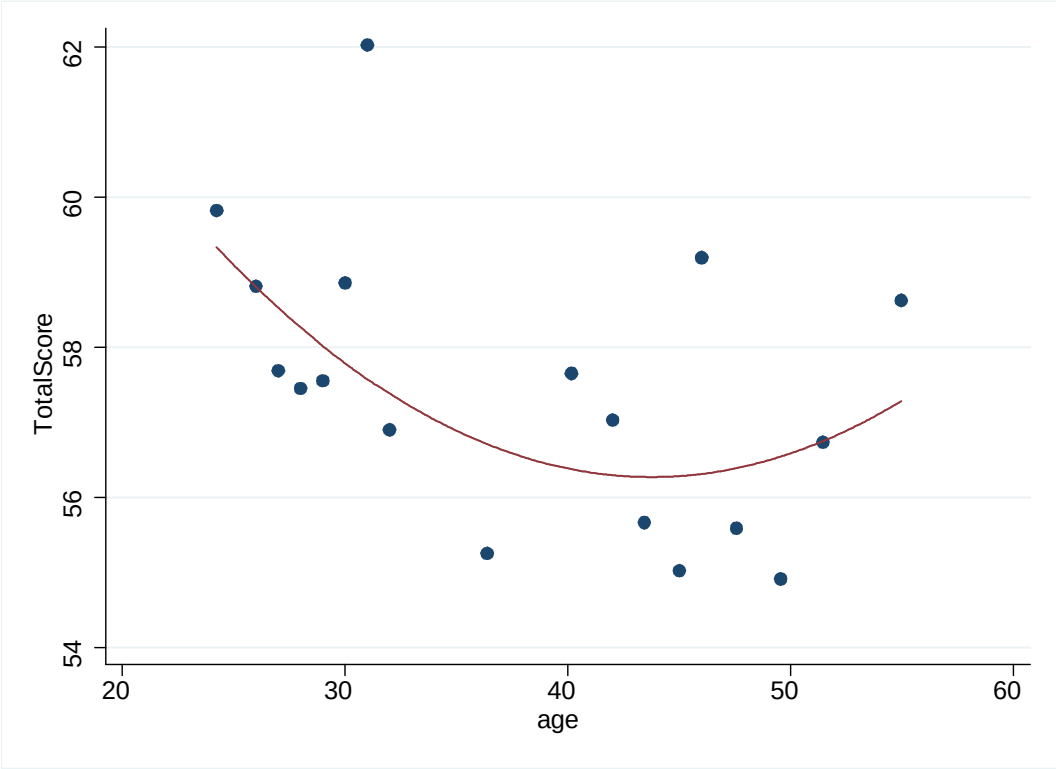
散布図が抱える問題点は，標本の大きさが大きいと，プロットされる点が多すぎて，グラフの特徴が見え辛くなることです．しかし，`binscatter` を用いれば連続な独立変数 `age` をカテゴリへ区分けすることができます．その後さらにカテゴリごとの従属変数の平均を求めてプロットすることができます．結果，以下を得ます．

```
. binscatter TotalScore age
```



binscatter により，鬱病に関する合計点 (TotalScore) と年齢の間の特徴が明確に浮かび上がりました．

8. 2 次モデルを用いて簡単な曲線を作成します．実行するコマンドは binscatter TotalScore age, line(qfit) で，ここでの qfit は 2 次フィットを行います．結果，以下を得ます．



この binscatter グラフからは，年齢が 25 才から 40 才までの区間のあたりでは一般的な下落傾向が見られるのに対し，その後横ばいとなり，40 才台後半には上昇に転じることが示されています．後述の第 10 章では，演習 11 などでモデルにおける 2 乗項の必要性について検定します．

第 8 章 (8.9 節, pp.223-224) の do-file

演習 8.1

```
/************** Begin do-file ************/
* chapter8.1.do
use "C:\data\gss2006_chapter8.dta", clear
correlate educ hrs1
by sex, sort: correlate educ hrs1
regress hrs1 educ
by sex, sort: regress hrs1 educ, beta
/************** End do-file **************/
```

演習 8.2a

```
/************** Begin do-file ************/
* chapter8.2.do
use "C:\data\gss2006_chapter8.dta", clear
codebook, compact
describe
by sex, sort: correlate hrs1 sphrs1
by sex, sort: regress sphrs1 hrs1, beta
twoway (scatter sphrs1 hrs1), by(sex)
/************** End do-file **************/
```

演習 8.3

```
/************** Begin do-file ************/
* Chapter8.3.do
use "C:\data\gss2006_chapter8.dta", clear
set seed 123
sample 250, count
twoway (scatter sphrs1 hrs1), by(sex)
scatter sphrs1 hrs1 if sex==1, jitter(3)
scatter sphrs1 hrs1 if sex==2, jitter(3)
* or
scatter sphrs1 hrs1, by(sex) jitter(3)
/************** End do-file **************/
```

演習 8.4

```
/************** Begin do-file ************/
* chapter8.4.do
use "C:\data\gss2006_chapter8.dta", clear
correlate happy hapmar health
pwcorr happy hapmar health
pwcorr happy hapmar health, obs sig
```

```
pwcorr happy hapmar health, listwise obs sig
/************** End do-file **************/
```

演習 8.5

```
/************** Begin do-file ************/
* chapter8.5.do
use "C:\data\gss2002_chapter8.dta", clear
codebook happy7 satfam7
recode happy7 satfam7 (1=7)(2=6)(3=5)(4=4)(5=3)(6=2)(7=1), ///
generate(happynew satfamnew)
label variable happynew ///
   "global happiness: 1 completely dissatisfied to 7 completely satisfied"
label variable satfamnew ///
   "family satisfaction: 1 completely dissatisfied to 7 completely satisfied"
regress happynew satfamnew, beta
/************** End do-file **************/
```

演習 8.6

```
/************** Begin do-file ************/
* chapter8.6.do
use "C:\data\spearman.dta", clear
twoway (scatter liberal age) (lfit liberal age), ///
   title(Relationship between age and liberalism) ///
   legend(order(1 "Observed liberalism" 2 "Predicted liberalism"))
twoway (scatter ranklib rankage) (lfit ranklib rankage), ///
   title(Relationship between age and liberalism ///
   (Using Ranks)) legend(order(1 "Observed liberalism" ///
   2 "Predicted liberalism"))
list
correlate age liberal
spearman rankage ranklib
/************** End do-file **************/
```

演習 8.7

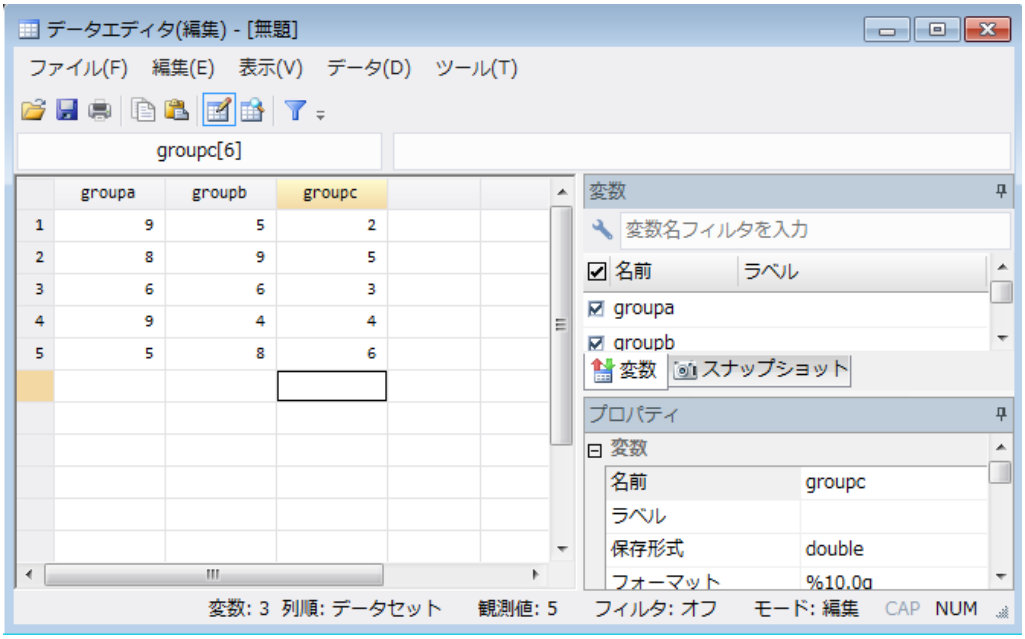
```
/************** Begin do-file ************/
* chapter8.7.do
clear
use http://www.stata-press.com/data/r13/depression.dta
twoway (scatter TotalScore age, sort)
binscatter TotalScore age
/************** End do-file **************/
```

演習 8.8

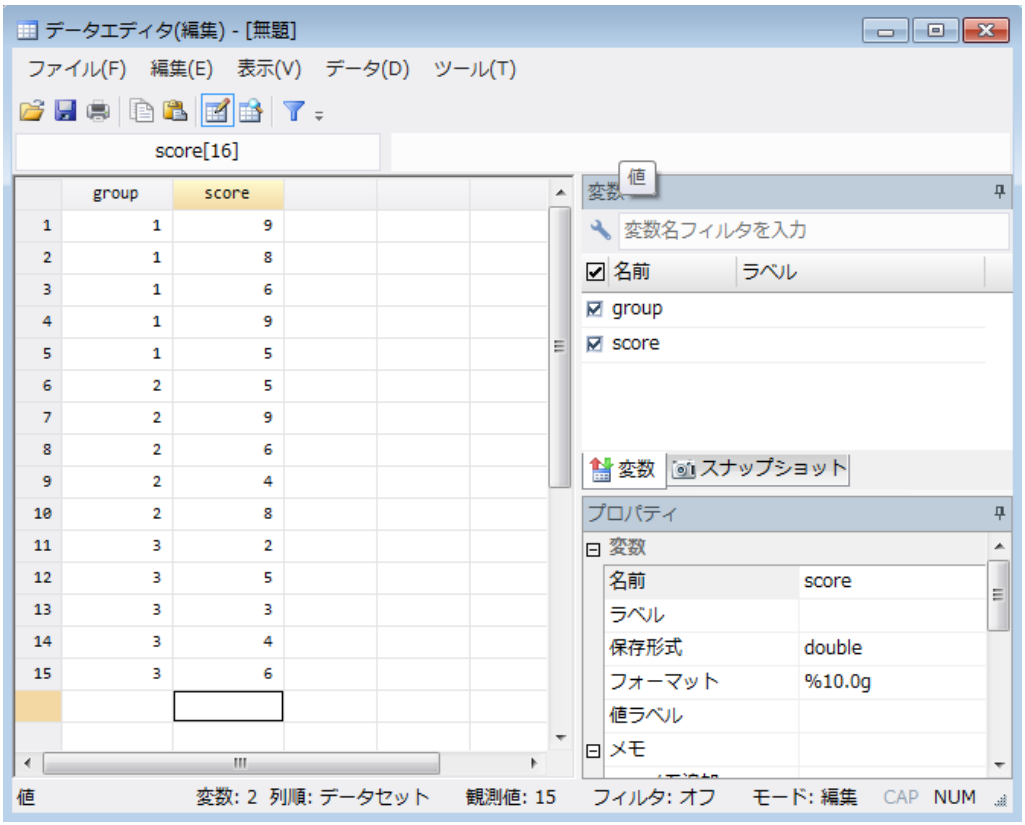
```
/************** Begin do-file ************/
* chapter8.8.do
use http://www.stata-press.com/data/r13/depression.dta, clear
binscatter TotalScore age, line(qfit)
/************** End do-file **************/
```

第 9 章 (9.11 節, pp.278-279) の解答

1. データをワイド形式で入力すると，以下のようになります．



データをロング形式で入力すると，以下のようになります．



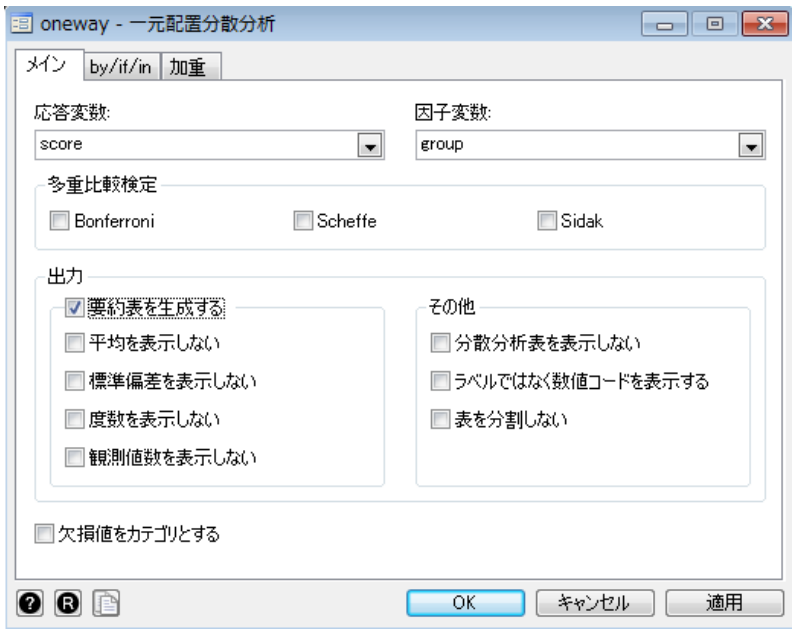
今後の演習で使用できるよう，データを保存しておきます．File ▷ Save as...（ファイル (F) ▷ 名前を付けて保存 (A)）を選択し，`C:\data\agis9.dta` として保存します．コマンドは以下です．

```
. save "C:\data\agis9.dta"
```

2. まず演習 1 で保存したデータセットを開きます．File ▷ Open...（ファイル (F) ▷ 開く... (O)）を選択し，`C:\data\agis9.dta` を選択します．コマンドは以下です．

```
. use "C:\data\agis9.dta"
```

ANOVA を実施するには，今回の分析が一元配置分散分析であるため，`oneway` が使用できます．また，ANOVA 実施後に多重比較も行います．では，Statistics ▷ Linear models and related ▷ ANOVA/MANOVA ▷ One-way ANOVA（統計 (S) ▷ 線形モデル他 ▷ ANOVA/MANOVA ▷ 一元配置分散分析）を選択します．Main（メイン）タブで，以下のように入力します．



Response variable（応答変数）に `score` を入力し，*Factor variable*（因子変数）に `group` を入力します．さらに，*Produce summary table*（要約表を生成する）を選択して，平均と標準偏差を載せた要約表を出力します．結果は以下です．

```
. oneway score group, tabulate
                    Summary of score
      group |        Mean   Std. Dev.       Freq.
------------+------------------------------------
          1 |         7.4   1.8165902           5
          2 |         6.4   2.0736441           5
          3 |           4   1.5811388           5
------------+------------------------------------
      Total |   5.9333333   2.2509257          15
                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      30.5333333      2   15.2666667      4.53     0.0341
 Within groups            40.4     12   3.36666667
------------------------------------------------------------------------
    Total           70.9333333     14   5.06666667
Bartlett´s test for equal variances:  chi2(2) =   0.2626  Prob>chi2 = 0.877
```

平均と標準偏差を比較すると，グループ 3（処置 C）が他の 2 グループに比べて平均値が低くなっているのが分かります．標準偏差にばらつきがありますが，量的にも割合的にもそれほど

大きなばらつきではなく，Bartlett 検定結果も 3 群の分散に有意差を示していません．ただ，グループ 3 の平均が他のいずれのグループに対しても，平均から標準偏差で一つ分以上低い値となっていることは大きな発見です．

$F(2,12) = 4.53, p < 0.05$ から，平均値の間には有意水準 5% の統計的な有意差があることが示されています．群間の分散の推定値は，群内の分散の推定値に対し 4.53 倍であることになります．

`pwmean` の実行コマンドと実行結果は以下です．実行したコマンドでは Bonferroni 多重比較を行っています．これは Stata に備わる 8 種類の比較手法のうち，多重比較であることを考慮する 7 種類の比較手法のうちの一つです．別の比較手法が良い場合は，そちらを指定しても構いません．

```
. pwmean score, over(group) effects cimeans mcompare(bonferroni)
Pairwise comparisons of means with equal variances
over         : group
```

score	Mean	Std. Err.	Unadjusted [95% Conf.	Interval]
group				
1	7.4	.8205689	5.612134	9.187866
2	6.4	.8205689	4.612134	8.187866
3	4	.8205689	2.212134	5.787866

	Number of Comparisons
group	3

score	Contrast	Std. Err.	Bonferroni t	Bonferroni P>|t|	Bonferroni [95% Conf.	Interval]
group						
2 vs 1	-1	1.16046	-0.86	1.000	-4.225466	2.225466
3 vs 1	-3.4	1.16046	-2.93	0.038	-6.625466	-.1745335
3 vs 2	-2.4	1.16046	-2.07	0.183	-5.625466	.8254665

結果には，先ほど `oneway` で得たものと同じ平均値が表示されます．多重比較を行っているのは，最下部の表です．グループ 3（$M = 4.0$）とグループ 2（$M = 6.4$）の比較では，差異（コントラスト）が -2.4 です．この比較での t 値は -2.07 であり，Bonferroni 調整後の p 値は 0.183 であるため，有意水準 5% の下では有意とはなりません．もし調整なしで多重比較を行うには，`mcompare(noadjust)` オプションを指定します．実施すると，平均，コントラスト，t 値は同値

を得ますが，p 値は 0.061 となり，Bonferroni 調整が行われたときと比べ，有意性を示す値にだいぶ近づいています．

3. 一元配置分散分析を行うには，`oneway` コマンドを使用します．ダイアログボックスは演習 2 のものと同じです．今回は *Response variable*（応答変数）に `tvhours` を入力し，*Factor variable*（因子変数）に `marital` を入力します．実行結果は，以下です．

```
. oneway tvhours marital, bonferroni tabulate
                  | Summary of hours per day watching
                  |                tv
     marital      |
      status      |       Mean   Std. Dev.       Freq.
------------------+------------------------------------
     married      |  2.5984655   1.8186774         391
     widowed      |  3.9767442   2.5758002          86
    divorced      |  3.1171875   2.6698111         128
   separated      |  2.7575758   2.2504208          33
   never mar      |  3.1910112   2.7153549         267
------------------+------------------------------------
       Total      |  2.9834254   2.3613666         905
                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      158.277593      4   39.5693983      7.29     0.0000
 Within groups      4882.47379    900   5.42497088
------------------------------------------------------------------------
    Total           5040.75138    904   5.57605241
Bartlett´s test for equal variances:  chi2(4) =  61.4829  Prob>chi2 = 0.000
          Comparison of hours per day watching tv by marital status
                                (Bonferroni)
Row Mean-
Col Mean  |     married    widowed   divorced   separate
----------+------------------------------------------------
 widowed  |     1.37828
          |       0.000
          |
divorced  |     .518722   -.859557
          |       0.290      0.083
          |
separate  |      .15911   -1.21917   -.359612
          |       1.000      0.107      1.000
          |
never ma  |     .592546   -.785733    .073824    .433435
          |       0.014      0.066      1.000      1.000
```

結果では，$F(4, 900) = 7.29, p < 0.001$ であることから，一日当たりのテレビ視聴時間は，婚姻区分別で見たとき，統計的な有意差があることが示されています．平均値の比較の詳細を見ると，パートナーに先立たれたグループが最も長いテレビ視聴時間であると見られるほか，婚姻中のグループと別居中のグループが最も短い視聴時間であることが伺われます．離婚したグループと未婚のグループの視聴時間は中程度です．観測数が群間で異なるため（別居中が $n = 33$,

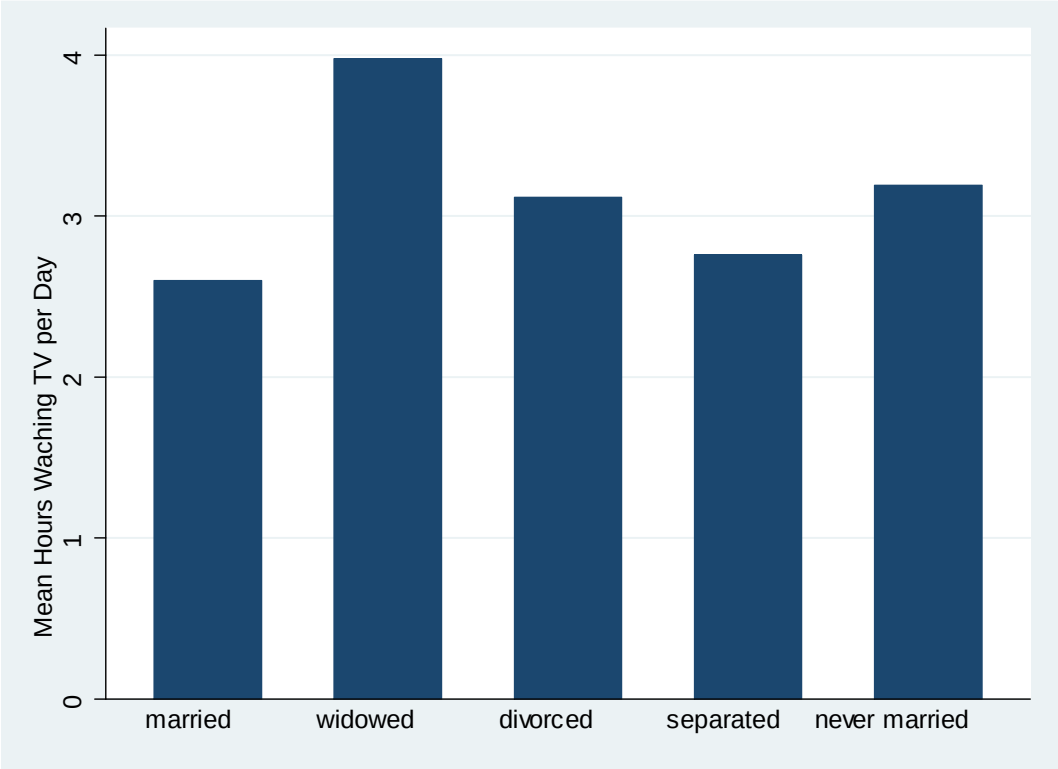
婚姻中が $n = 391$），等分散の仮定が重要な問題になります．等分散性を検定する Bartlett 検定の結果が $\chi^2(4) = 61.48, p < 0.001$ と高い統計的有意性を示していることから，等分散の仮定が採用できません．ここが手持ちのデータの限界です．しかし，標準偏差同士はそれほどばらついているとは言えません．Bartlett 検定には，実際の差異が小さくても，標本の大きさが大きい（$N = 905$）ときには統計的有意差を示しやすい傾向があります．

平均の多重比較では，明らかに有意差が認められる比較が 2 つ，有意差に近い（marginally significant）差が認められる比較が 2 つあります．（p 値が 0.05 を超え，且つ 0.10 未満の場合，marginally significant と表現されることがあります．）

テレビ視聴時間について，配偶者を亡くしたグループ（$M = 3.98$）は婚姻中のグループ（$M = 2.60$）に対し，高い値を持つことが $p < 0.001$ の有意性の下に示されています．未婚のグループ（$M = 3.19$）も婚姻中のグループ（$M = 2.60$）に対し，高い値を持つことが $p < 0.05$ の有意性で示されています．統計的な有意性と実質的な有意性の差は重要です．配偶者を亡くしたグループは離婚したグループに対し，統計的有意性はないものの，-1.22 の差があります．一方で，婚姻中のグループと未婚のグループの間には，その半分以下の差しかないにもかかわらず，その差は統計的に有意となっています．なぜでしょう．手元のデータには，別居中のグループに 33 人，配偶者を亡くしたグループに 86 人のデータしかないのに対し，婚姻中のグループに 391 人，未婚のグループに 267 人のデータがあります．標本の大きさが大きいと，たとえ観測された差が小さくても，その差に対する信頼性は高くなります．

Graphics ▷ Bar chart（グラフィックス (G) ▷ 棒グラフ）を選択して，棒グラフを作成します．デフォルトで *Graph by calculating summary statistics*（記述統計量をグラフにする）と，*Vertical*（垂直）が選択されていますので確認してください．*Mean*（平均値）の変数として，`tvhours` を指定します．Categories（カテゴリ）タブで，*Group 1*（グループ 1）を選択し，*Grouping variable*（グループ変数）ボックスに `marital` を指定します．Y axis（y 軸）タブで，*Title*（タイトル）ボックスで y 軸のタイトルを指定しても構いません．実行結果とグラフは以下のようになります．

```
. graph bar (mean) tvhours, over(marital)
> ytitle(Mean Hours Waching TV per Day)
```



4. partyid についてのクロス表を表示すると，以下になります．

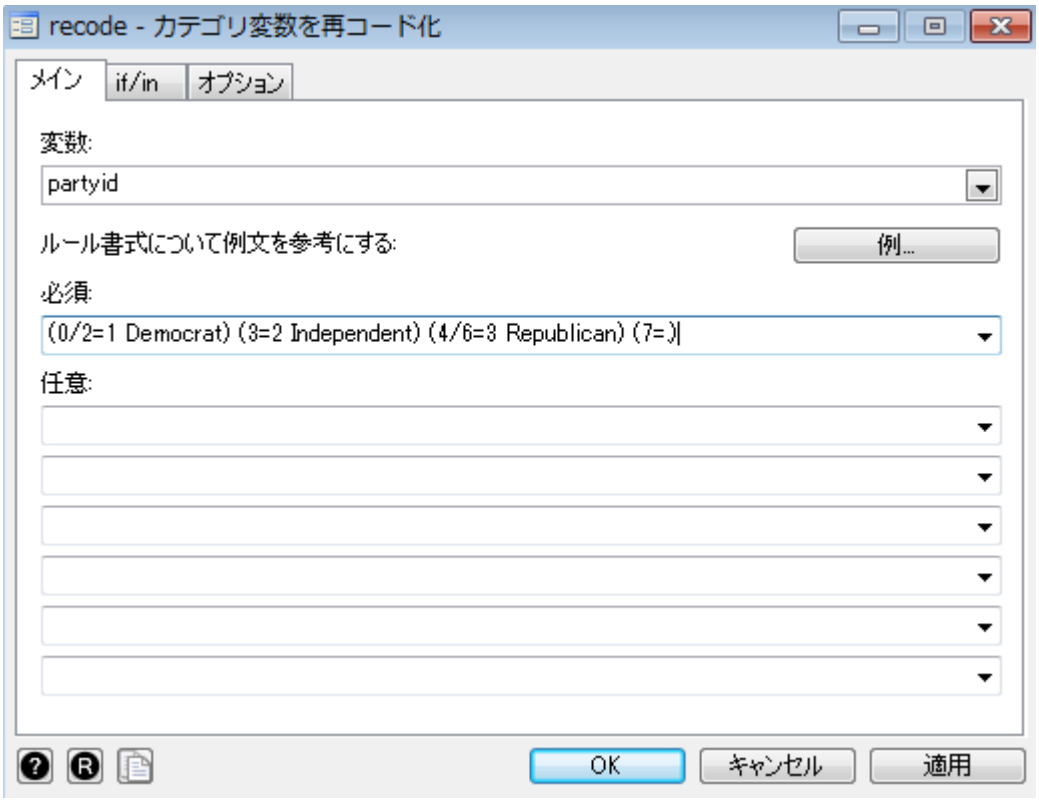
```
. tabulate partyid
   political party
       affiliation |      Freq.     Percent        Cum.
-------------------+-----------------------------------
     strong democrat |       408       14.95       14.95
   not str democrat |       515       18.87       33.82
       ind,near dem |       267        9.78       43.61
        independent |       528       19.35       62.95
       ind,near rep |       199        7.29       70.25
 not str republican |       449       16.45       86.70
  strong republican |       315       11.54       98.24
        other party |        48        1.76      100.00
-------------------+-----------------------------------
              Total |     2,729      100.00
```

その他の政党に区別された人は 48 人おり，この人たちの分のデータを分析から外します．また，共和党と民主党にある複数のカテゴリを党ごとにまとめます．最終的には，民主党，無党派，共和党の 3 つのカテゴリを得ます．再コード化をするためには，元のコードでどの数値が

何を意味するのかを把握する必要があります．方法は 2 つあります．`numlabel partyid, add` を実行した後，`tabulate` コマンドを実行するか，またはよりシンプルと思われる以下のコマンドを実施します．

```
. codebook partyid
-------------------------------------------------------------------------------
partyid                                                  political party affiliation
-------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  partyid

                 range:  [0,7]                        units:  1
         unique values:  8                        missing .:  36/2,765

            tabulation:  Freq.   Numeric  Label
                           408         0  strong democrat
                           515         1  not str democrat
                           267         2  ind,near dem
                           528         3  independent
                           199         4  ind,near rep
                           449         5  not str republican
                           315         6  strong republican
                            48         7  other party
                            36         .
```

次に `recode` コマンドでカテゴリを再コード化します．0，1，2 の 3 つのコードは民主党として一つに統合します．また，4，5，6 の 3 つは共和党として統合します．さらに，7 は欠損値にします．Data ▷ Create or change data ▷ Other variable-transformation commands ▷ Recode cateorical variables（データ (D) ▷ データの作成または変更 ▷ その他の変数変換コマンド ▷ カテゴリ変数を再コード化）を選択して該当するダイアログボックスを開きます．



Required（必須）での指定は，少し手が込んでいます．0 から 2 までを 1 へ変換し，Democrat とラベルを付けます（もしラベルにスペースが含まれていた場合，全体を二重引用符で囲います）．3 は 2 へ変換し，Independent とラベルを付けます．4 から 6 までは 3 へ変換し，Republican とラベルを付けます．最後に，7 は.（ドット）へ変換し欠損値とします．Options（オプション）タブで，*Generate new variables*（変数を新規作成する）を選択し，新規変数名として `party` を入力します．元のデータを上書きすると，元に戻すことができなくなります．元の変数はそのまま残し，新たに変数を作成してそちらに新しいデータを格納することを覚えておいてください．実行されるコマンドと実行結果は以下です（変換後，確認のために新たな変数に対して `tabulate` を実行しています）．

```
. recode partyid (0/2=1 Democrat) (3=2 Independent) (4/6=3 Republican) (7=.), generat
> e(party)
(2214 differences between partyid and party)
. tabulate party
```

```
  RECODE of |
    partyid |
 (political |
      party |
affiliation |
          ) |      Freq.     Percent        Cum.
------------+-----------------------------------
   Democrat |      1,190       44.39       44.39
Independent |        528       19.69       64.08
 Republican |        963       35.92      100.00
------------+-----------------------------------
      Total |      2,681      100.00
```

これで一元配置分散分析を実施するの準備が整いました．演習 2 と同じダイアログボックスを用いて実行したコマンドと結果は以下です．

```
. oneway tvhours party, bonferroni tabulate
  RECODE of |
    partyid |
 (political |
      party |   Summary of hours per day watching
affiliation |                   tv
          |         Mean   Std. Dev.       Freq.
------------+------------------------------------
   Democrat |   3.0645995   2.1900535         387
  Independe |   3.3011364   3.1085502         176
  Republica |   2.7047619   2.0218101         315
------------+------------------------------------
      Total |   2.9829157    2.356665         878
                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups       44.776093      2   22.3880465      4.06     0.0176
 Within groups      4825.96764    875   5.51539159
------------------------------------------------------------------------
    Total           4870.74374    877   5.55386971
Bartlett´s test for equal variances:  chi2(2) =  49.7607  Prob>chi2 = 0.000
                   Comparison of hours per day watching tv
              by RECODE of partyid (political party affiliation)
                                (Bonferroni)
Row Mean-
Col Mean    |   Democrat   Independ
------------+----------------------
Independ    |    .236537
            |      0.805
            |
Republic    |   -.359838   -.596374
            |      0.131      0.021
```

平均と標準偏差の表から，無党派が最も長くテレビを視聴し，民主党派，共和党派はより短い時間視聴することが分かります．$F(2, 875) = 4.06, p < 0.05$ より平均差は有意差です．群平均の比較を見ると，1 組だけ有意であることが見て取れます．共和党派は無党派に対し，テレビ

視聴時間が短いことが，有意（$p < 0.05$（実際には p は 0.021））であると示されています．民主党派 vs 無党派，また民主党派 vs 共和党派の比較に関しては，有意ではありません．

5. クラスカル・ウォリス検定を行うには，Statistics ▷ Summaries, tables, and tests ▷ Nonparametric tests of hypotheses ▷ Kruskal-Wallis rank test（統計 (S) ▷ 要約/表/検定 ▷ ノンパラメトリック仮説検定 ▷ Kruskal-Wallis 順位検定）を選択します．*Outcome variable*（アウトカム変数）に `tvhours` を指定し，*Variable defining groups*（グループを定義する変数）に `marital` を指定します．結果は，以下です．

```
. kwallis tvhours, by(marital)
Kruskal-Wallis equality-of-populations rank test
```

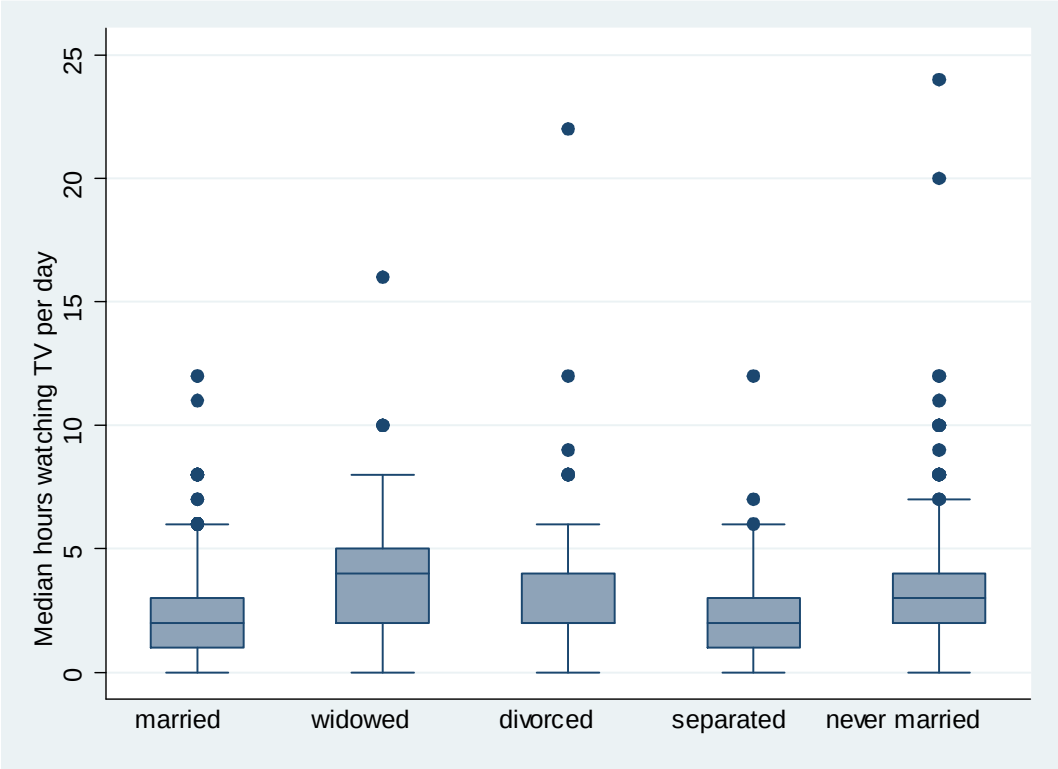
marital	Obs	Rank Sum
married	391	161540.50
widowed	86	49512.50
divorced	128	59309.50
separate	33	13887.50
never ma	267	125715.00

```
chi-squared =    29.991 with 4 d.f.
probability =     0.0001
chi-squared with ties =    31.136 with 4 d.f.
probability =     0.0001
```

結果では，婚姻区分別のテレビ視聴時間のメディアンの間に統計的有意差が示されています（$\chi^2(4) = 31.136, p < 0.001$）．データには従属変数である視聴時間に関し，同じ値の人が数多くいる，すなわち同じ視聴時間を答えた人が数多くいるので，2 つある chi-squared 値のうち，同点を考慮した chi-squared with ties のほうを使用します．

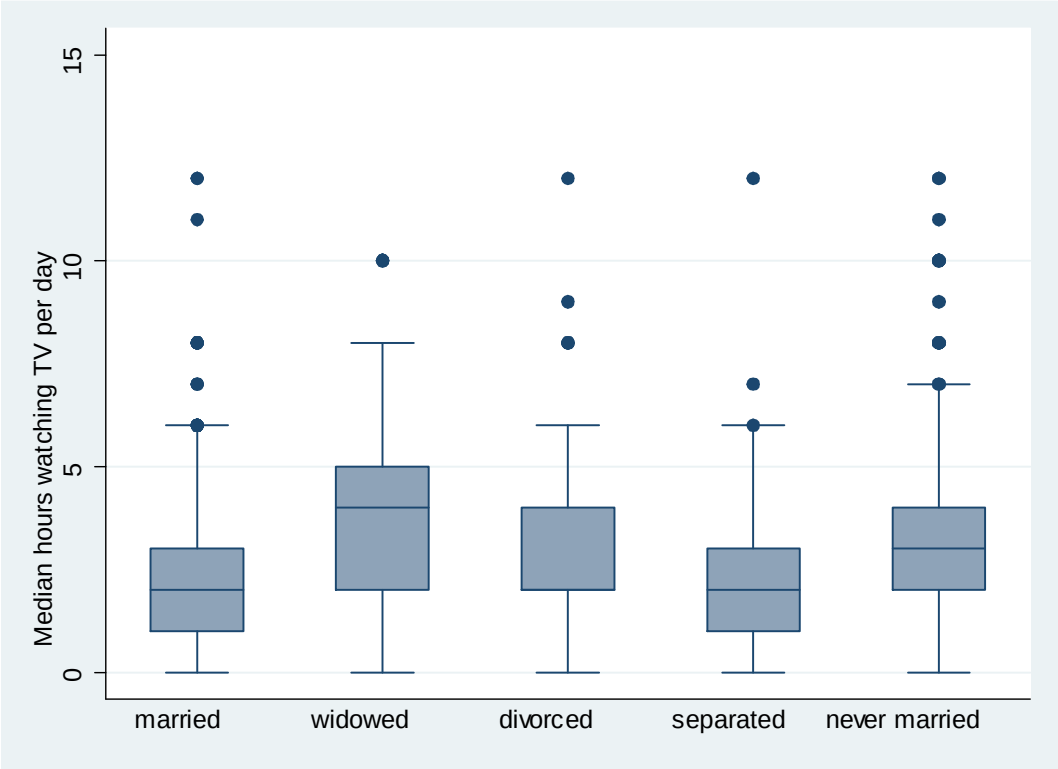
結果の解釈をしやすくするために，箱ひげ図を描きます．Graphics ▷ Box plot（グラフィックス (G) ▷ 箱ひげ図）を選択します．Main（メイン）タブで，*Variables*（変数）ボックスに `tvhours` を指定します．Categories（カテゴリ）タブで，*Group 1*（グループ 1）を選択し，*Grouping variable*（グループ変数）ボックスに `marital` を指定します．Y axis（y 軸）タブで，*Title*（タイトル）ボックスで `Median hours watching TV per day` と入力しても構いません．以上の手続きによるコマンドと結果のグラフは以下のようになります．

```
. graph box tvhours, over(marital)
> ytitle(Median hours watching TV per day)
```



上記のグラフは外れ値の影響で，必ずしも見やすいグラフとは言えない状況です．配偶者と死別したグループに，テレビ視聴が 15 時間を超える人が 1 人，離婚したグループに 20 時間を超える人が 1 人，未婚のグループに 20 時間を超える人が 2 人います．これらの外れ値があるため，y 軸が延伸され，描画が小さくなってしまっています．ここは `tvhours` $>$ `15` の条件を満たす人を除いた方が良いでしょう．if/in タブでの指定によりこれらのデータをグラフから除くと，以下のようなコマンドとグラフを得ます．

```
. graph box tvhours if tvhours < 15, over(marital)
> ytitle(Median hours watching TV per day)
```



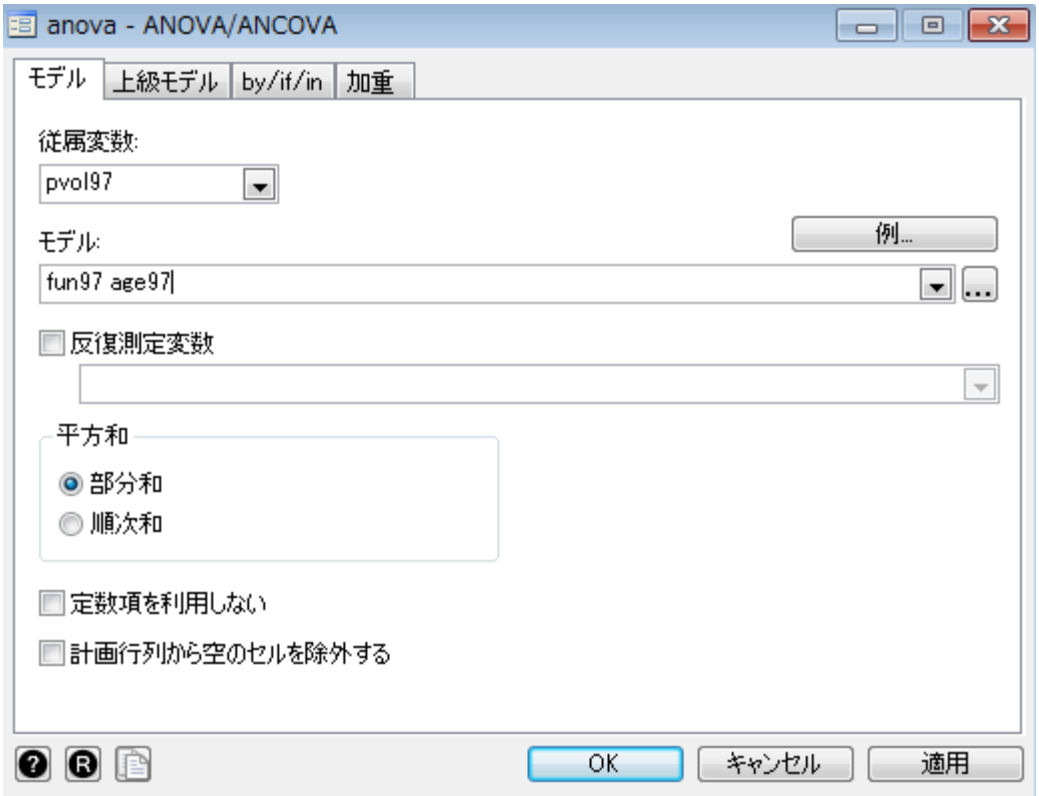
このグラフからは，単に各群の平均とメディアンを並べたときよりも，さらに詳細に分布の様子を把握することができます．メディアンとデータの集まりがともに類似していることから，婚姻中のグループと別居中のグループは同じような分布をしています．一方，配偶者を亡くしたグループと未婚のグループには，この中では長めのテレビ視聴をする傾向があります．

婚姻区分がテレビの視聴時間に関係するという結論はクラスカル・ウォリス順位検定でも ANOVA でも同じでした．しかし，ANOVA では情報がより多く得られ，平均の多重比較検定まで行えました．一方，クラスカル・ウォリス検定と箱ひげ図の組み合わせからは，平均値の比較に関して明確な検定結果が得られません．

6. ((訂正) 本演習については本体書籍のデータセット名の記述に誤植がありました．以下のように訂正するとともに，心からお詫び申し上げます．)

   (訂正データセット名) `nlsy97_selected_variables.dta`

共分散分析を行うため，anova ダイアログボックスを開きます．Statistics ▷ Linear models and related ▷ ANOVA/MANOVA ▷ Analysis of variance and covariance（統計 (S) ▷ 線形モデル他 ▷ ANOVA/MANOVA ▷ ANOVA/ANCOVA）を選択します．以下のように入力します．*Dependent variable*（従属変数）に pvol97 を指定し，その下にある *Model*（モデル）ボックスに予測変数 fun97 と age97 を指定します．その下で *Partial*（部分和）が選択されていることを確認してください．入力後のダイアログボックスは以下です．



コマンドと結果は，以下です．

```
. anova pvol97 fun97 age97
                       Number of obs =      5,278    R-squared     =  0.0411
                       Root MSE      =    1.11729    Adj R-squared =  0.0391
                Source | Partial SS         df         MS        F    Prob>F
                -------+----------------------------------------------------
                 Model |  281.44008         11   25.585462     20.50  0.0000
                       |
```

```
      fun97 |  167.31302          7    23.90186     19.15  0.0000
      age97 |  91.926517          4   22.981629     18.41  0.0000
            |
   Residual |  6573.7255      5,266   1.2483337
-----------+-----------------------------------------------------
      Total |  6855.1656      5,277   1.2990649
```

ダイアログボックスで *Partial*（部分和）を指定したにもかかわらず，コマンドにはそのことが表示になっていない点が気になるかも知れませんが，これは部分和を計算することが anova の既定の状態であるためです．

ここでの関心事は，age97 の効果をコントロールしたときの fun97 の影響です．結果からは，家族と楽しく過ごす頻度が友人のボランティア活動参加率に有意な影響を与えていることが分かります（$F(7, 5266) = 19.15, p < 0.001$）．結果で問題なのは，家族団欒の各レベルにおける友人ボランティア参加率の平均，さらにより重要な年齢をコントロールしたときのそれらの推定値が表示されないことです．そうした平均値を得る場合は，事後推定コマンド margins を使用します．Statistics ▷ Postestimation ▷ Marginal means and predictive margins（統計 (S) ▷ 推定後の分析 ▷ 周辺平均効果，限界効果，基礎的分析）を選択します．Main（メイン）タブで，予測変数（因子変数）fun97 を共変量として指定します．これにより age97 が平均値の仮想的な状態に対し，fun97 の水準ごとに pvol97 の予測平均値が計算されます．コマンドと結果は，以下です．

```
. margins fun97
Predictive margins                              Number of obs     =      5,278
Expression   : Linear prediction, predict()
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       fun97 |
           0 |   2.036278   .0378039    53.86   0.000     1.962167    2.110389
           1 |   1.953236   .0381784    51.16   0.000      1.87839    2.028081
           2 |   2.116734   .0355515    59.54   0.000     2.047038     2.18643
           3 |   2.158206   .0391742    55.09   0.000     2.081409    2.235004
           4 |   2.211448   .0451649    48.96   0.000     2.122906    2.299989
           5 |   2.435613   .0470144    51.81   0.000     2.343445    2.527781
           6 |   2.458464   .0850116    28.92   0.000     2.291806    2.625122
           7 |   2.545288   .0566935    44.90   0.000     2.434145    2.656431
------------------------------------------------------------------------------
```

age97 の調整なしに fun97 の各水準で pvol97 の予測平均値を計算する場合，table コマンドが活用できます．Statistics ▷ Summaries, tables, and tests ▷ Other tables ▷ Flexible table of summary statistics（統計 (S) ▷ 要約/表/検定 ▷ その他の表 ▷ 要約統計量）を選択します．*Row*

variable（行の変数）に `fun97` を指定します．最下部で，1 行目の *Statistics*（統計量）を `Mean`（平均値）に設定し，*Variable*（変数）に `pvol97` を指定します．
コマンドと結果は，以下です．

```
. table fun97, contents(mean pvol97 )
```

# days/wk fun as a family 1997	mean(pvol97)
0	2.02743
1	1.94639
2	2.10718
3	2.15356
4	2.22186
5	2.44523
6	2.48555
7	2.57179
.a	2.85714
.b	2.08333
.c	4
.d	1.97006

7. ((訂正) 本演習については本体書籍の記述に誤植がありました．以下のように訂正するとともに，心からお詫び申し上げます．)

   (誤) 肝細胞
   (正) 幹細胞

二元配置分散分析をするには，`anova` コマンドを使用します．Statistics ▷ Linear models and related ▷ ANOVA/MANOVA ▷ Analysis of variance and covariance（統計 (S) ▷ 線形モデル他 ▷ ANOVA/MANOVA ▷ ANOVA/ANCOVA）を選択します．*Dependent variable*（従属変数）に `stemcell` を指定します．*Model*（モデル）に交互作用項を含む予測変数，すなわち `partyid female partyid#female` を指定します（古いバージョンの Stata では交互作用項を `partyid*female` と指定します）．もし，交互作用項の指定法を忘れてしまったときは，Example...（例...）をクリックすると基本的なモデルでの指定法を示した簡単な例が閲覧できます．*Sum of squares*（平方和）では *Partial*（部分和）が指定されていることを確認します．
コマンドと結果は，以下です．

```
. anova stemcell partyid female partyid#female
                        Number of obs =         46    R-squared     =  0.5897
```

```
                  Root MSE      =    1.45805    Adj R-squared =  0.5141
        Source | Partial SS         df         MS        F    Prob>F
----------------+----------------------------------------------------
         Model |  116.08544          7   16.583634     7.80  0.0000
               |
       partyid |  81.889492          3   27.296497    12.84  0.0000
        female |  .06179973          1   .06179973     0.03  0.8655
partyid#female |  23.023181          3   7.6743938     3.61  0.0218
               |
      Residual |  80.784127         38   2.1258981
----------------+----------------------------------------------------
         Total |  196.86957         45   4.3748792
```

他の変数をコントロールした上で各変数の周辺平均を確認することはとても有用です．`margins` を使用すると，そうした平均値を得ることができます．

```
. margins partyid female partyid#female
Predictive margins                              Number of obs     =         46
Expression   : Linear prediction, predict()
------------------------------------------------------------------------------
                             |            Delta-method
                             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------------------+------------------------------------------------
                     partyid |
                    democrat |   7.819876   .4303547    18.17   0.000     6.948668    8.691083
                  republican |   4.637681   .4212993    11.01   0.000     3.784805    5.490557
                 independent |   6.849275   .4019355    17.04   0.000     6.035599    7.662951
                 noninvolved |   4.782609   .5159841     9.27   0.000     3.738053    5.827164
                             |
                      female |
                        male |   6.091787   .3016645    20.19   0.000       5.4811    6.702475
                      female |   6.247205   .3166041    19.73   0.000     5.606274    6.888136
                             |
                partyid#female |
                democrat#male |         7    .652058    10.74   0.000     5.679978    8.320022
              democrat#female |   8.714286   .5510896    15.81   0.000     7.598663    9.829908
              republican#male |         4   .5952448     6.72   0.000      2.79499     5.20501
            republican#female |   5.333333   .5952448     8.96   0.000     4.128323    6.538343
             independent#male |   7.444444   .4860153    15.32   0.000     6.460558    8.428331
                 independent # |
                       female |       6.2    .652058     9.51   0.000     4.879978    7.520022
             noninvolved#male |       5.5    .729023     7.54   0.000      4.02417     6.97583
                 noninvolved # |
                       female |         4    .729023     5.49   0.000      2.52417     5.47583
------------------------------------------------------------------------------
```

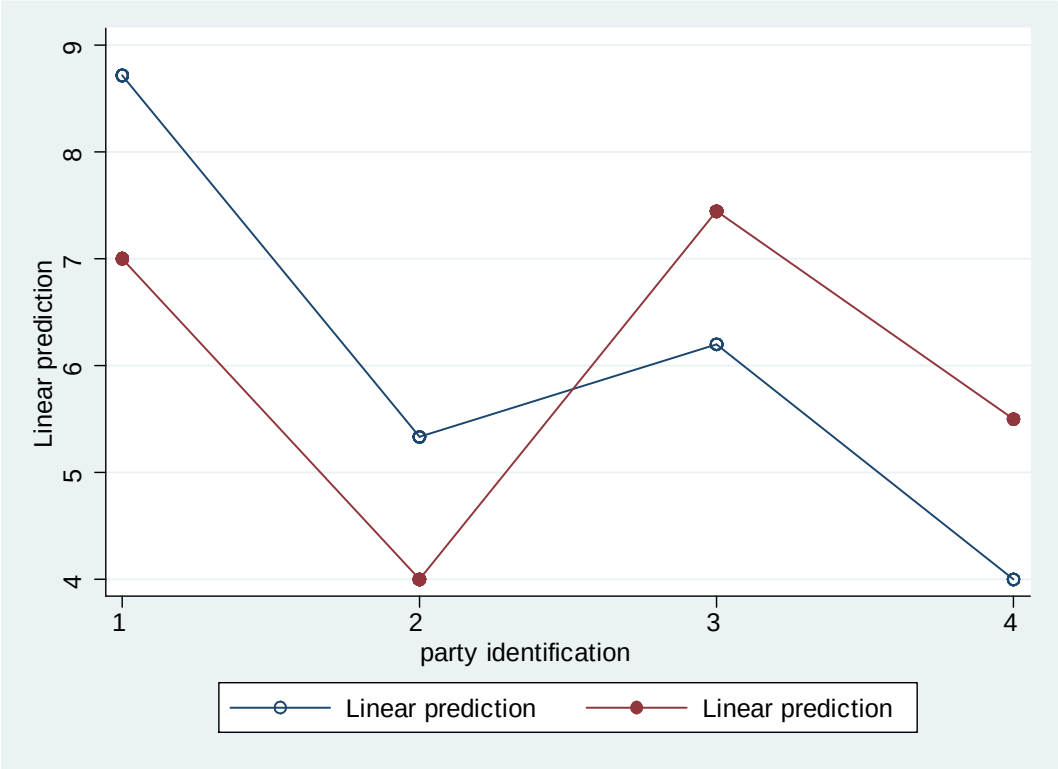
交互作用項の影響をグラフにするには，まず幹細胞研究の支持の度合いの予測値を出す必要があります．Statistics ▷ Postestimation ▷ Predictions, residuals, etc.（統計 (S) ▷ 推定後の分析 ▷ 予測，およびその標準誤差，レバレッジ統計量，距離統計量等）を選択します．*New variable name*（新しい変数名）に `predictedstem` と入力して新たな変数を作成します．その下で *Linear prediction (xb)*（線形予測 (xb)）が選択されていること確認します．発行されるコマンドは，

以下です．

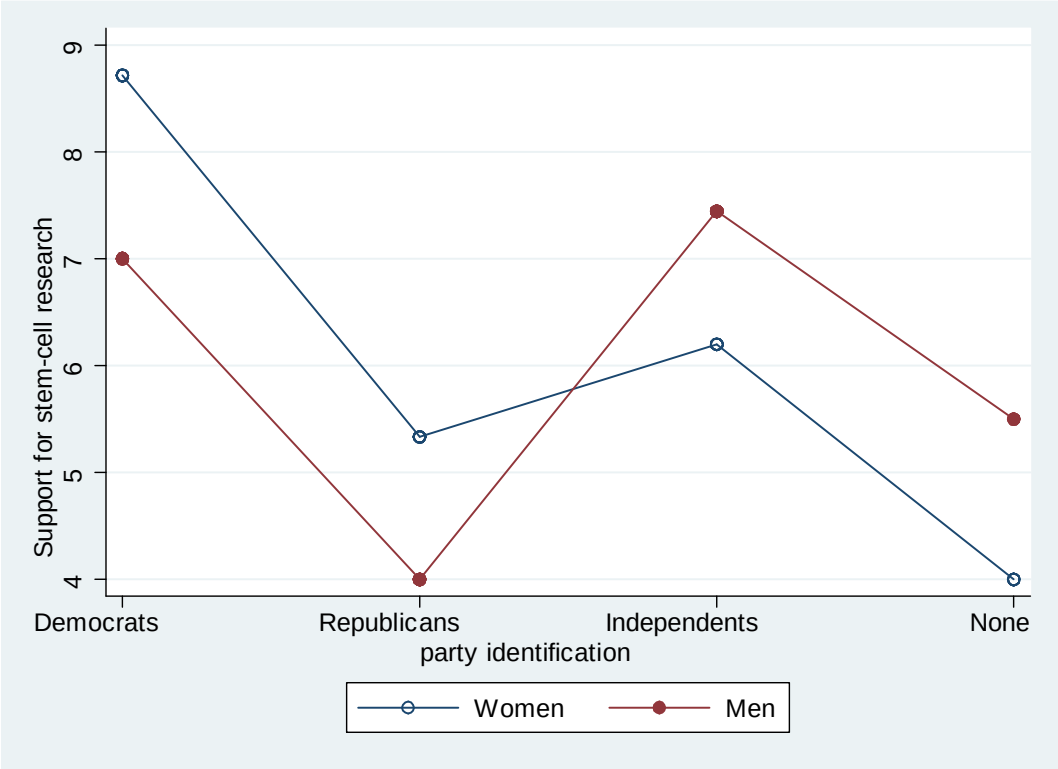
```
. predict predictedstem, xb
```

これでグラフを作成することができます．`partyid`$[F(3,38)=12.84, p<0.001]$，`female`（有意でない）による差，および両変数の交互作用項 $[F(3,38)=3.61, p<0.05]$ の差を可視化します．Graphics ▷ Twoway graph (scatter, line, etc.)（グラフィックス (G) ▷ 二元グラフ (散布図/折れ線など)）を選択します．男性と女性の二つのグラフを作成し，重ね合わせます．Create...（作成...）をクリックします．開いたウィンドウの左側で *Basic plots*（基本的なグラフ）が選択されていること確認し，右側で `Connected`（点を接続）を選択します．*Y variable*（y 変数）に `predictedstem` を指定し，*X variable*（x 変数）に `partyid` を指定します．Marker properties（マーカーのプロパティ）をクリックし，*Symbol*（シンボル）で `Hollow circle` を選択し，Accept（OK）をクリックします．次に if/in タブで，*if (expression)*（条件式）に `female==1` と入力します．Accept（OK）をクリックして，*Plot 1*（プロット 1）の作業を完了します．再び Create...（作成...）をクリックし，二つ目のグラフ，すなわち男性のグラフの作成を開始します．Marker properties（マーカーのプロパティ）の *Symbol*（シンボル）で `Circle` を選択すること，また if/in タブで `female==1` と入力すること以外，先ほどと同じステップを繰り返します．
コマンドとグラフは，以下です．

```
. twoway (connected  predictedstem partyid if female==1, msymbol(circle_hollow))
> (connected predictedstem partyid if female==0, msymbol(circle))
```



グラフエディタで諸々の編集を行い，次のようにすることもできます．



このグラフは，民主党派が共和党派よりも幹細胞研究を高く支持していることを示しています．無党派層もまた，共和党派よりも幹細胞研究を高く支持していますが，政治に関心がない人々は共和党派と支持の度合いが似ています．男女別で見ると，どちらかの党を支持する層では女性の支持度合いが高いのに対し，どちらの党も支持しない層では男性の支持度合いが高い点に，男女別の支持の度合いに有意差が見られなかった理由を見て取ることができます．このグラフには，交互作用項の効果も表れています．どちらか一方の党を支持する層では女性が男性よりも支持度合いが高いのに対し，どちらも支持しない層では男性が女性よりも支持度合いが高いという構図が表れています．

8. まず演習 1 のデータをワイド形式で再入力します．ただ，今回は追加要素として変数 `id` を作成し，ほかの変数の変数名を `groupa`，`groupb`，`groupc` とする代わりに `time1`，`time2`，`time3` とし，以下のようなデータにします．

```
. list
```

	id	time1	time2	time3
1.	1	9	5	2
2.	2	8	9	5
3.	3	6	6	3
4.	4	9	4	4
5.	5	5	8	6

上記のデータは，`reshape` コマンドを使用してロング形式に変換する必要があります．Data ▷ Create or change data ▷ Other variable-transformation commands ▷ Convert data between wide and long（データ (D) ▷ データの作成または変更 ▷ その他の変数変換コマンド ▷ データのワイド形式/ロング形式を変換）を選択します．*Long format from wide*（ワイド形式からロング形式へ）が選択していることを確認します．*ID variable(s) - the i() option*（ID 変数 - i() オプション）ボックスで `id` と入力し，*Subobservation identifier - the j() option*（サブオブザベーション識別子 - j() オプション）の *Variable*（変数）に `wave`，*Base(stub) names of X_ij variables*（X_ij 変数の基準 (接頭語) 名）に `time` と入力します．`wave` は，この時点ではまだ変数としてデータセットに存在していませんが，`reshape` コマンドの実行後に作成されることに注意してください．

コマンドと結果は，以下です．

```
. reshape long time, i(id) j(wave)
(note: j = 1 2 3)
Data                               wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                        5   ->      15
Number of variables                   4   ->       3
j variable (3 values)                     ->   wave
xij variables:
                      time1 time2 time3   ->   time
-----------------------------------------------------------------------------
```

これにより `id`，`wave`，`time` の 3 つの変数を得ました．

```
. list
```

	id	wave	time
1.	1	1	9
2.	1	2	5
3.	1	3	2
4.	2	1	8
5.	2	2	9
6.	2	3	5
7.	3	1	6

8.	3	2	6
9.	3	3	3
10.	4	1	9
11.	4	2	4
12.	4	3	4
13.	5	1	5
14.	5	2	8
15.	5	3	6

`id` が 1 である最初の人について，`wave` ごとに 3 行のデータがあります．5 人に対し 15 行のデータがあるロング形式です．

Statistics ▷ Linear models and related ▷ ANOVA/MANOVA ▷ Analysis of variance and covariance（統計 (S) ▷ 線形モデル他 ▷ ANOVA/MANOVA ▷ ANOVA/ANCOVA）を選択します．*Dependent variable*（従属変数）に `time` を指定します．*Model*（モデル）に `id wave` を指定します．*Repeated-measures variables*（反復測定変数）を選択して，ボックスに `wave` を指定します．これにより Stata に wave1, 2, 3 の間の変動を検定する繰り返し測定あり ANOVA を行うよう指示することができます．

コマンドと結果は，以下です．

```
. anova time id wave, repeated(wave)
                        Number of obs =         15    R-squared     =  0.5752
                        Root MSE      =    1.94079    Adj R-squared =  0.2566
                  Source | Partial SS         df         MS        F    Prob>F
                 --------+----------------------------------------------------
                   Model |       40.8          6        6.8     1.81  0.2152
                         |
                      id |  10.266667          4  2.5666667     0.68  0.6241
                    wave |  30.533333          2  15.266667     4.05  0.0609
                         |
                Residual |  30.133333          8  3.7666667
                 --------+----------------------------------------------------
                   Total |  70.933333         14  5.0666667

Between-subjects error term:  id
                     Levels:  5         (4 df)
     Lowest b.s.e. variable:  id
Repeated variable: wave
                                          Huynh-Feldt epsilon        =  0.8471
                                          Greenhouse-Geisser epsilon =  0.6556
                                          Box´s conservative epsilon =  0.5000
                                            ------------ Prob > F ------------
                  Source |     df      F    Regular    H-F      G-G      Box
                 --------+-------------------------------------------------
                    wave |      2   4.05    0.0609   0.0736   0.0937   0.1144
                Residual |      8
                 --------+-------------------------------------------------
```

結果には，モデル全体に対する $R^2 = 0.26$ とありますが，これは今回の関心事ではありません．この値は，測定回（1 回目，2 回目，3 回目）と被験者（1 人目，2 人目，...，5 人目）の両方により説明できる変動の割合を示しています．ここで注目しているのは測定回による R^2 であり，全体の変動に対して，測定回の違いで説明できる変動はどれほどなのかという点です．この R^2 は $30.53/70.93 = 0.43$ と計算されます．全体の F 値とその p 値は $F(2,8) = 4.05, p < 0.06$ であり，結果の最後の表の中に表示があります．この表にはほかに 3 種類の p 値も記載され，それらはいずれも 3 回の測定が同じ被験者に繰り返し行われたことによる独立の欠落を自由度に反映した調整後の値です．Box 調整は最も極端な調整で，自由度を半分に縮小し，p 値を割り出しています．おそらく 3 つとも報告書に記載するのが無難ですが，Huynh-Feldt 調整が最も採用されやすいでしょう．

最後に，`margins` コマンドで 3 回の測定の平均を推定します．このコマンドにより `wave` の各水準での従属変数 `time` の平均が与えられます．

```
. margins wave
Predictive margins                              Number of obs     =         15
Expression   : Linear prediction, predict()
```

	Margin	Delta-method Std. Err.	t	P>|t|	[95% Conf.	Interval]
wave						
1	7.4	.8679478	8.53	0.000	5.398509	9.401491
2	6.4	.8679478	7.37	0.000	4.398509	8.401491
3	4	.8679478	4.61	0.002	1.998509	6.001491

9. まずデータをワイド形式で入力し，`reshape` でロング形式に変換します．ロング形式のデータを直接入力もできますが，練習のために上記の方法で行いましょう．ワイド形式のデータは以下のようになります．

```
. list
```

	family	par1	par2	kid1	kid2
1.	1	9	3	9	8
2.	2	5	5	7	4
3.	3	4	2	8	10
4.	4	6	7	3	2
5.	5	8	10	8	8

上記のようなタイプのデータはときに distinguishable dyadic data（識別可能な二者間データ）

と呼ばれます．par1 が常に母親で par2 が常に父親，あるいはその逆となります．同様に，kid1 が常に年上で kid2 が常に年下となります．indistinguishable dyadic data（識別不可能な二者間データ）は，たとえば同姓カップルのように父親/母親というラベルを区別して付与できない場合のデータです．

次に，データを reshape で変換します．stub を二つ指定します．一つは親を意味する par，もう一つは子を意味する kid です．また，家族の識別する変数が family です．今は一行に複数の人が多重した状態なので，j() には individual と指定します．

```
. reshape long par kid, i(family) j(individual)
(note: j = 1 2)
Data                               wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                        5   ->      10
Number of variables                   5   ->       4
j variable (2 values)                     ->   individual
xij variables:
                               par1 par2   ->   par
                               kid1 kid2   ->   kid
-----------------------------------------------------------------------------
```

ロング形式のデータに list を実行した結果は，以下です．

```
. list
```

	family	indivi~l	par	kid
1.	1	1	9	9
2.	1	2	3	8
3.	2	1	5	7
4.	2	2	5	4
5.	3	1	4	8
6.	3	2	2	10
7.	4	1	6	3
8.	4	2	7	2
9.	5	1	8	8
10.	5	2	10	8

級内相関の推定には，loneway コマンドを使用します．ここでの関心事は，それぞれの親同士または子同士の組の中でスコアがどれほど近いかという点です．もしスコアの変動が家族間によるもの（ある家族ではスコアが高く，ある家族では低い）であるとき，家族内においては実質的な類似性がある，すなわち親同士内または子同士内で類似していることになります．Statistics ▷ Linear models and related ▷ ANOVA/MANOVA ▷ Large one-way ANOVA（統計 (S) ▷ 線形モデル他 ▷ ANOVA/MANOVA ▷ 大規模な一元配置分散分析）を選択して loneway ダイア

ログボックスを開きます．分析は親と子で別に行います．

まず親を対象とし，*Response variable*（応答変数）に par，*Group variable*（グループ変数）に family を指定します（父と母は家族でグループ化されています）．Submit（適用）をクリックして，コマンドを実行します．

```
. loneway par family
                    One-way Analysis of Variance for par:
                                                Number of obs =          10
                                                    R-squared =      0.6305
    Source                SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between family            38.4        4        9.6       2.13     0.2138
Within family             22.5        5        4.5
------------------------------------------------------------------------
Total                     60.9        9   6.7666667
         Intraclass       Asy.
         correlation      S.E.       [95% Conf. Interval]
         ---------------------------------------------------
            0.36170     0.41228       0.00000     1.16976
         Estimated SD of family effect           1.596872
         Estimated SD within family               2.12132
         Est. reliability of a family mean        0.53125
              (evaluated at n=2.00)

. loneway kid family
                    One-way Analysis of Variance for kid:
                                                Number of obs =          10
                                                    R-squared =      0.8865
    Source                SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between family            58.6        4       14.65      9.77     0.0139
Within family              7.5        5        1.5
------------------------------------------------------------------------
Total                     66.1        9   7.3444444
         Intraclass       Asy.
         correlation      S.E.       [95% Conf. Interval]
         ---------------------------------------------------
            0.81424     0.15986       0.50093     1.12756
         Estimated SD of family effect           2.564176
         Estimated SD within family              1.224745
         Est. reliability of a family mean        0.89761
              (evaluated at n=2.00)
```

級内相関は，親について 0.36 という実質的な類似性を示す値が出ました．また，子について，0.81 という極めて高い類似性が出ました．

第 9 章 (9.11 節, pp.278-279) の do-file

演習 9.1

```
/************** Begin do-file ************/
* chapter9.1.do
save "C:\data\agis9.dta"
/************** End do-file **************/
```

演習 9.2

```
/************** Begin do-file ************/
* chapter9.2.do
use "C:\data\agis9.dta"
oneway score group, tabulate
pwmean score, over(group) effects cimeans mcompare(bonferroni)
pwmean score, over(group) effects cimeans mcompare(bonferroni)
/************** End do-file **************/
```

演習 9.3

```
/************** Begin do-file ************/
* chapter9.3.do
use "C:\data\gss2002_chapter9.dta
oneway tvhours marital, bonferroni tabulate
graph bar (mean) tvhours, over(marital) ytitle(Mean Hours Watching TV per Day)
/************** End do-file **************/
```

演習 9.4

```
/************** Begin do-file ************/
* chapter9.4.do
use "C:\data\gss2002_chapter9.dta", clear
tabulate partyid
codebook partyid
fre partyid /*If installed*/
recode partyid (0/2=1 Democrat)(3=2 Independent)(4/6=3 Republican)(7=.), ///
generate(party)
tabulate party
oneway tvhours party, bonferroni tabulate
graph box tvhours, medtype(line) over(party) ///
ytitle(Median Hours Watching TV per Day)
graph box tvhours if tvhours < 15, medtype(line) ///
over(party) ytitle(Median Hours Watching TV per Day)
/************** End do-file **************/
```

演習 9.5

```
/************** Begin do-file ************/
* chapter9.5.do
use "C:\data\gss2002_chapter9.dta
oneway tvhours marital, bonferroni tabulate
kwallis tvhours, by(marital)
graph box tvhours, over(marital) ///
ytitle(Median hours watching TV per day)
graph box tvhours if tvhours < 15, over(marital) ///
ytitle(Median hours watching TV per day)
/************** End do-file **************/
```

演習 9.6

```
/************** Begin do-file ************/
* chapter9.6.do
use "C:\data\nlsy97_selected_variables.dta", clear
anova pvol97 fun97 c.age97, partial
margins fun97
table fun97, contents(mean pvol97)
/************** End do-file **************/
```

演習 9.7

```
/************** Begin do-file ************/
* chapter 9.7.do
* c9_exercise_8.do
use "C:\data\partyid.dta", clear
anova stemcell partyid female partyid#female, partial
margins partyid female partyid#female
predict predictedstem, xb
twoway (connected predictedstem partyid if female==1, msymbol(circle_hollow)) ///
(connected predictedstem partyid if female==0, msymbol(circle))
/************** End do-file **************/
```

演習 9.8

```
/************** Begin do-file ************/
* chapter9.8.do
* Reenter the data, adding the id variable and renaming the variable
* labels as indicated
list
reshape long time, i(id) j(wave)
* Use a list to verify that the transformation was done as you intended.
list
anova time id wave, repeated(time)
margins wave
```

```
/************** End do-file **************/
```

演習 9.9

```
/************** Begin do-file ************/
* chapter9.9.do
* Re-enter the data
list
reshape long par kid, i(family) j(individual)
* Use a list to verify that the transformation was done as you intended.
list
loneway par family
loneway kid family
/************** End do-file **************/
```

第 10 章 (10.15 節, pp.342-344) の解答

1. recode コマンドを用いて sex と wrkslf を再コード化します．sex の再コード化後の値は新規変数 male に収め，wrkslf については新規変数 selfemp に収めることにします．codebook で元データのコードの意味を確認し，変換後に新旧変数のクロス表で間違いがないか確認します．コマンドと結果は，以下です．

```
. codebook sex wrkslf
-------------------------------------------------------------------------------
sex                                                              respondents sex
-------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  sex

                 range:  [1,2]                        units:  1
         unique values:  2                        missing .:  0/2,765

            tabulation:  Freq.   Numeric  Label
                         1,228         1  male
                         1,537         2  female
-------------------------------------------------------------------------------
wrkslf                                              r self-emp or works for somebody
-------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  wrkslf

                 range:  [1,2]                        units:  1
         unique values:  2                        missing .:  96/2,765

            tabulation:  Freq.   Numeric  Label
                           307         1  self-employed
                         2,362         2  someone else
                            96         .

. recode sex (1=1 male) (2=0 female), generate(male)
(1537 differences between sex and male)

. tabulate sex male, missing
             |      RECODE of sex
respondent   |    (respondents sex)
     s sex   |     female       male |     Total
-------------+----------------------+----------
      male   |          0      1,228 |     1,228
    female   |      1,537          0 |     1,537
-------------+----------------------+----------
     Total   |      1,537      1,228 |     2,765

. recode wrkslf (1=1 self-employed) (2=0 "someone else"), generate(selfemp)
(2362 differences between wrkslf and selfemp)

. tabulate wrkslf selfemp, missing
r self-emp or | RECODE of wrkslf (r self-emp or
```

works for somebody	works for somebody) someone e	self-empl	.	Total
self-employed	0	307	0	307
someone else	2,362	0	0	2,362
.	0	0	96	96
Total	2,362	307	96	2,765

次に，重回帰を行い，ベータ重みも求めます．

```
. regress hrs1 male age selfemp, beta
      Source |       SS           df       MS      Number of obs   =     1,721
-------------+----------------------------------   F(3, 1717)      =     31.61
       Model |  19171.738         3  6390.57933   Prob > F        =    0.0000
    Residual |  347146.361     1,717  202.181922   R-squared       =    0.0523
-------------+----------------------------------   Adj R-squared   =    0.0507
       Total |  366318.099     1,720  212.975639   Root MSE        =    14.219

------------------------------------------------------------------------------
        hrs1 |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
        male |   6.424665   .6864235     9.36   0.000                 .2201195
         age |  -.0632973   .0270648    -2.34   0.019                -.0559048
     selfemp |  -1.052756   1.037442    -1.01   0.310                -.0242797
       _cons |   41.36069    1.19508    34.61   0.000                        .
------------------------------------------------------------------------------
```

線形式に表すと，次のように書けます．

$$\texttt{hrs1}\ \text{推定値} = 41.36 + 6.42(\texttt{male}) - 0.06(\texttt{age}) - 1.05(\texttt{selfemp})$$

次に，20 歳の自営業の女性の労働時間を推定します．方法はいくつかありますが，最も正確なやり方は，保存されている B の値を使用するものです．それ以外にも以下のように `display` コマンドで `regress` の結果を使って計算することもできます．

```
. display "estimated hours = "6.42*0 - .06*20 - 1.05*1 + 41.36
estimated hours = 39.11
```

性別，年齢，自営業かいわゆる雇われかを取り入れたこのモデルは，労働時間の変動の 5% しか説明できないため，あまり良いモデルとは言えません．結果自体は，$F(3, 1717) = 31.61, p < 0.001$ と有意です．ただ，この有意性は回帰結果が強力だったというよりは，標本の大きさが大きかったことに起因します（$R^2 = 0.05$）．統計的な有意性すなわち R^2 がゼロとの有意差があることと，実質的な有意性すなわち R^2 がゼロとの有意差があり且つその差が実質的であることは，違うことに留意してください．

2. ((訂正) 本演習については本体書籍の記述に誤植がありました．以下のように訂正するととも

に，心からお詫び申し上げます．)

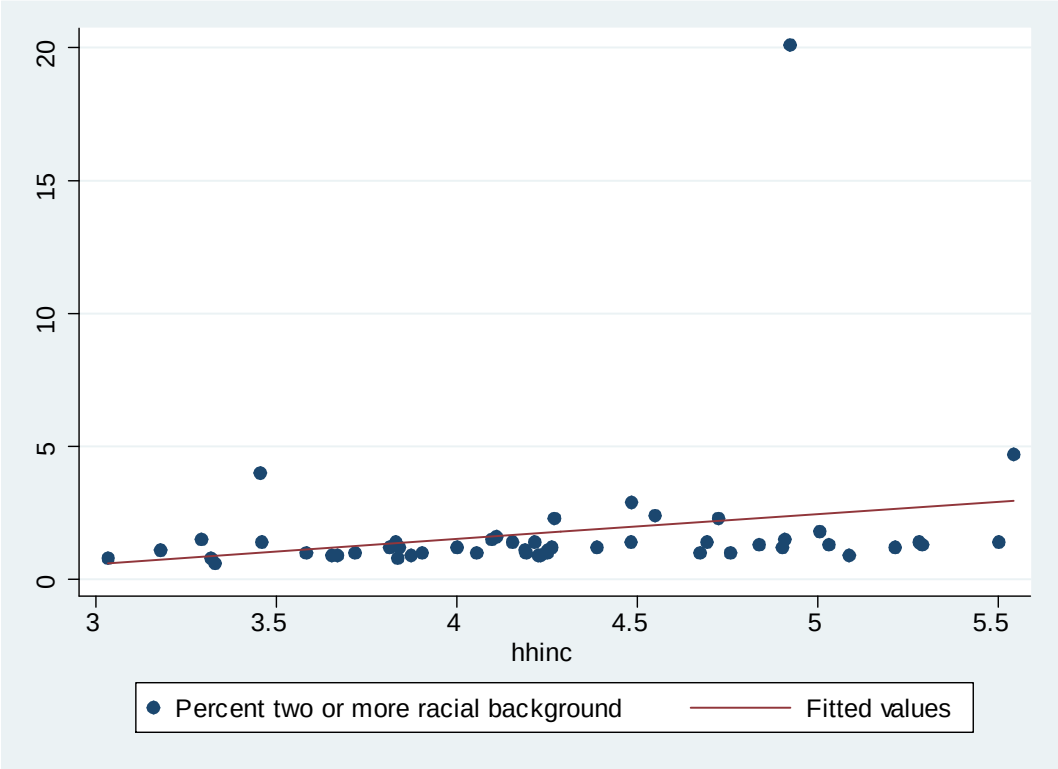
(誤) tworase

(正) tworace

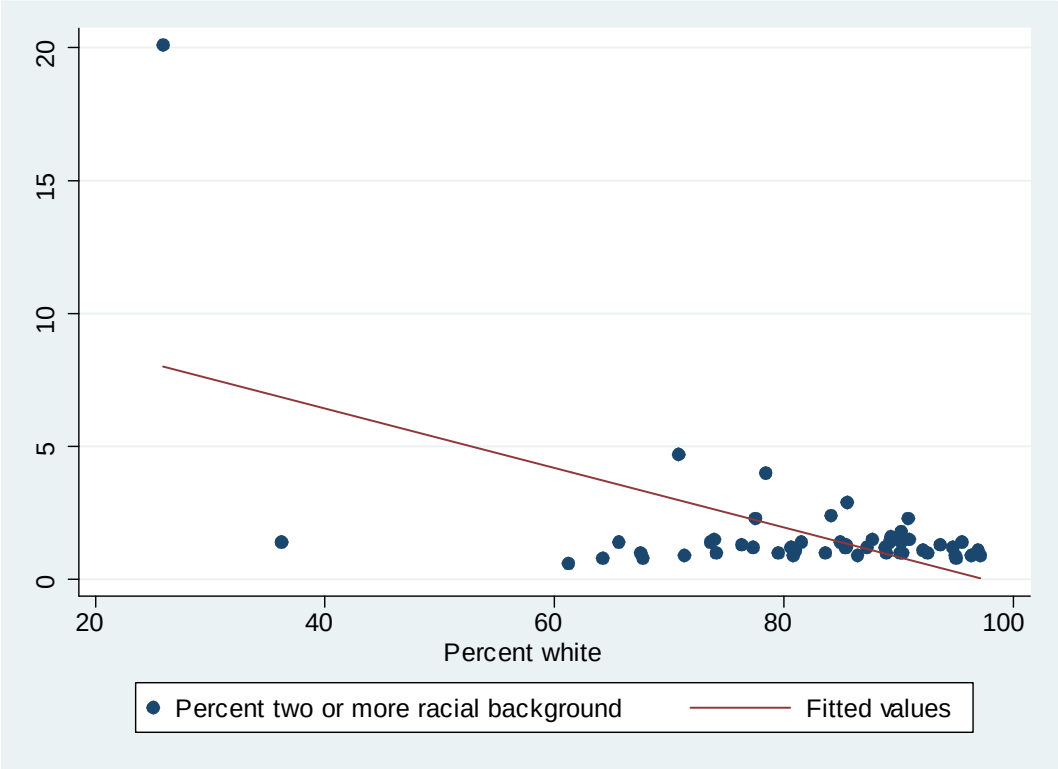
まず，重回帰を行います．結果，以下を得ます．

```
. regress tworace white hhinc ba
      Source |       SS           df       MS      Number of obs   =        51
-------------+----------------------------------   F(3, 47)        =     11.45
       Model |  156.575467         3  52.1918222   Prob > F        =    0.0000
    Residual |  214.306115        47  4.55970458   R-squared       =    0.4222
-------------+----------------------------------   Adj R-squared   =    0.3853
       Total |  370.881582        50  7.41763163   Root MSE        =    2.1353

------------------------------------------------------------------------------
     tworace |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       white |  -.1204406   .0223738    -5.38   0.000    -.1654508   -.0754304
       hhinc |   1.623711   .6146224     2.64   0.011       .38725    2.860172
          ba |  -.1680948    .070885    -2.37   0.022     -.310697   -.0254926
       _cons |   9.165264   2.931213     3.13   0.003     3.268424     15.0621
------------------------------------------------------------------------------
```

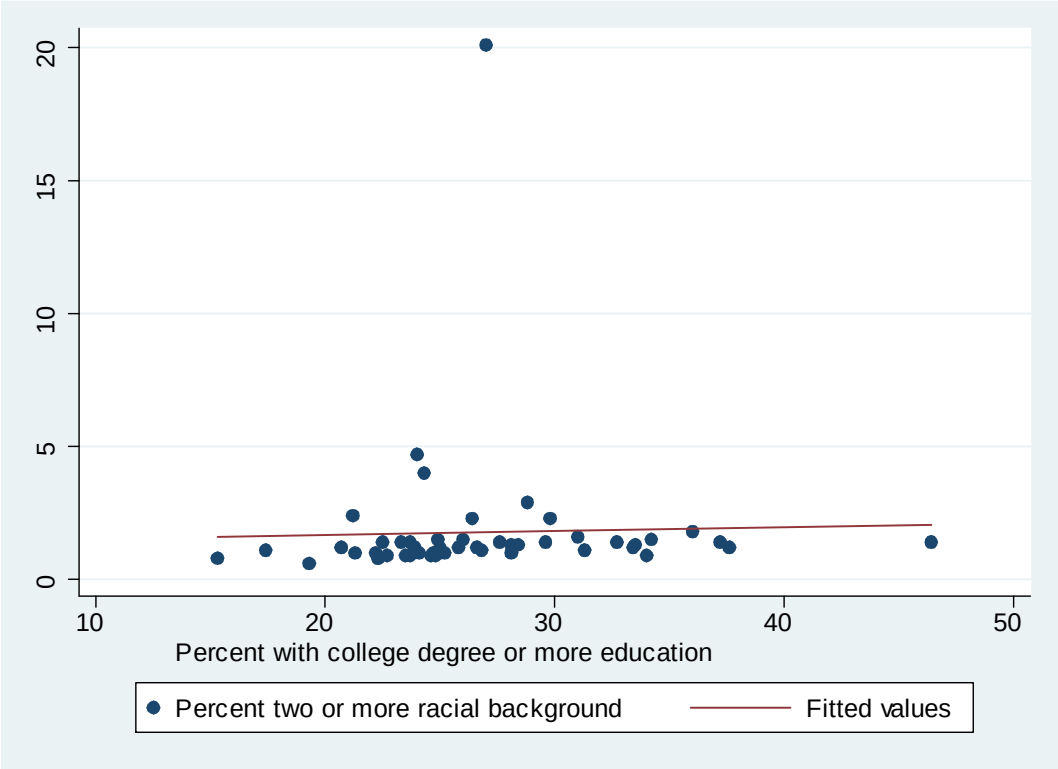
図 10.4 は，独立変数 1 つに対する従属変数の散布図です．今回は散布図に，hhinc に対する tworace，white に対する tworace，ba に対する tworace の 3 つが必要です．また，回帰予測を重ねた重複グラフにする必要があります．二元の重複グラフを用い，散布図をプロット 1，線形予測をプロット 2 として描きます．一つ目のグラフは，以下です．



上記グラフには，外れ値が 1 つあるのが分かります．これが巨大な残差を生み出しています．またこの外れ値を除くと，`hhinc` と `tworace` の間には有意な関係が存在しません．



上記のグラフはさらに問題です．`white` に関し，白人の割合が低い州 1 つが外れ値となっています．また，白人の割合が高くなるにつれ，予測値からのばらつきは小さくなっていきます．



上記のグラフには，大きな外れ値が 1 つあります．この外れ値は，独立変数の分布の中央付近にあるため，傾きの推定にそれほど大きなバイアスを生みません．外れ値が独立変数の最小値または最大値にある場合，傾きの推定に非常に大きなバイアスを生みます．

演習では，線形予測値と標準化残差も求める指示があります．コマンドは，以下です．

```
. predict yhat
. predict rstandard, rstandard
```

最後に，混血者の割合をよく予測できなかったものを特定します．標準化残差 $|1.96|$ を超える州をリストアップすることにします．リストには，州名，各変数値，`tworace` の予測値，標準化残差を表示します．演習の記述では，標準化残差が 1.96 を超えるものを一覧にするよう指示がありますが，-1.96 未満のものも同様に推定が悪いケースであるため，こちらも一覧に含めます．1.96 という数字を用いる理由は，これが z スコアにおける 95% 信頼区間の上限であるからです．結果は，以下です．

```
. list state white hhinc ba tworace yhat rstandard if rstandard > 1.96 | rstandard <
> -1.96
```

	state	white	hhinc	ba	tworace	yhat	rstandard
12.	Hawaii	25.9	4.9232	27	20.1	9.5011468	6.4458849

結果はさほど驚くものではありません．独立変数対従属変数の散布図には，それぞれに外れ値が 1 つずつありました．また，この州では白人の割合が低いことが分かります．ハワイ州は人種の多様な地域です．夫婦全体のなかで異民族同士の組は 9.5% と予測されましたが，実際の値は 20.1% と，はるかに高いものでした．ハワイ州についての標準化残差は 6.45 で，z スコアと同様に解釈可能なものです．

3. DFbeta は，問題のある観測値の特定方法の一つです．演習 2 では，ハワイ州を除いたとき，分析結果が劇的に変化しそうなことが分かりました．本来ならば，どの観測値を除いても分析結果は劇的に変化すべきではありません．DFbeta は，各独立変数と各観測値に対して別々に計算されます．この解答では，演習で求められている以上の解答を示したいと思います．具体的には，`tworace` の予測値について，推定後の分析を行います．問題のある観測値を一覧にする際，DFbeta の値に加え，予測値に対するスコア，従属変数の値，従属変数の予測値も表示します．過度に影響が高い観測値にどんな問題があるかを確認するため，各独立変数値の平均を `summarize` で表示します．結果は，以下です．

```
. regress tworace white hhinc ba
```

(output omitted)

```
. predict yhat
(option xb assumed; fitted values)

. dfbeta
                      _dfbeta_1: dfbeta(white)
                      _dfbeta_2: dfbeta(hhinc)
                      _dfbeta_3: dfbeta(ba)

. list state white hhinc ba tworace yhat _dfbeta_1 _dfbeta_2 _dfbeta_3 if abs(_dfbeta
> _1) > 2/sqrt(51) | abs(_dfbeta_3) > 2/sqrt(51) | abs(_dfbeta_3) > 2/sqrt(51)
```

	state	white	hhinc	ba	tworace	yhat
9.	District of Columbia	36.2	4.1539	46.4	1.4	3.7504487

_dfbeta_1	_dfbet~2	_dfbeta_3
.8512678	1.052215	-1.462431

```
12.  +-----------------------------------------------------------------------+
     |                state | white |  hhinc |   ba | tworace |      yhat |
     |               Hawaii |  25.9 | 4.9232 |   27 |    20.1 | 9.5011468 |
     |-----------------------------------------------------------------------|
     |     _dfbeta_1        |      _dfbet~2        |      _dfbeta_3        |
     |     -14.46868        |       5.37189        |     -5.755651         |
     +-----------------------------------------------------------------------+

25.  +-----------------------------------------------------------------------+
     |                state | white |  hhinc |   ba | tworace |      yhat |
     |          Mississippi |  61.2 | 3.3305 | 19.3 |      .6 | 3.9578385 |
     |-----------------------------------------------------------------------|
     |     _dfbeta_1        |      _dfbet~2        |      _dfbeta_3        |
     |      .4650196        |      .1884046        |       .2309257        |
     +-----------------------------------------------------------------------+

. summarize white hhinc ba
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       white |         51    81.71961    13.89722       25.9       97.1
       hhinc |         51    4.266153    .6318089     3.0342     5.5426
          ba |         51    26.70588    5.603371       15.3       46.4
```

演習 2 と 3 では，`list` の条件設定を異なる方法で行っています．上記にある 2 つの `list` コマンドを見比べてください．演習 3 では，$2/\sqrt{N} = 2/\sqrt{51}$ を超える大きい DFbeta を選び出す条件文を，`if` 文の中に数式を用いて記述しています．代わりに，事前に値を計算して値だけ用いても構いません．また演習 3 では，絶対値を返す関数も用いています．

結果では，コロンビア特別区 (D.C.) が問題のある観測値であり，3 つの DFbeta がすべて $2/\sqrt{N}$ 基準値を大きく上回っていることが示されています．D.C. は，その経済が連邦政府に依存する異質なものであることを論拠に，分析の対象から外しても良いかもしれません．D.C. における白人の割合は他よりかなり低く，学士取得率が他よりかなり高くなっています．個人所得は全体の平均とさほど変わりませんが，白人の割合の低さと教育水準の突出した高さを基にすると，実際より低い値が予測されます．

ハワイ州は，DFbeta の値が 3 つとも極めて高い値です．演習 2 のグラフでも，ハワイ州が外れ値であることを示していました．ミシシッピ州は，3 つの DFbeta の値が基準値よりわずかに高い値です．同州で異民族の夫婦の割合は 4% と予測されましたが，実際は 0.6% でかなり稀な存在です．同州では全米平均と比べて，白人の割合が低く，貧困率が高く，教育水準が低くなっています．白人の割合が全米平均より低いことから，異民族カップルが多いことが予測されますが，実際にはそうなっていない原因としては，その土地の文化規範がモデルに含めた要因を超えて影響していると考えられます．

こうした問題のある観測値はどう扱うべきでしょうか．D.C. については，州でないこと，また農業生産や郊外の欠如など，他の州と様々な面で異なることを理由に，分析の対象からの除外を

主張することも可能です．ハワイ州については，その独自の歴史を持つ特別なケースとして，分析の対象からの除外を主張することもできます．ミシシッピ州については，類似の州が存在するため，分析の対象からの除外を正当化する根拠はほとんどありません．もしかすると，D.C. とハワイ州を除くと，ミシシッピ州がもはや問題のある州ではなくなっているかもしれません．

4. ((訂正) 本演習については本体書籍の記述に誤植がありました．以下のように訂正するとともに，心からお詫び申し上げます．)

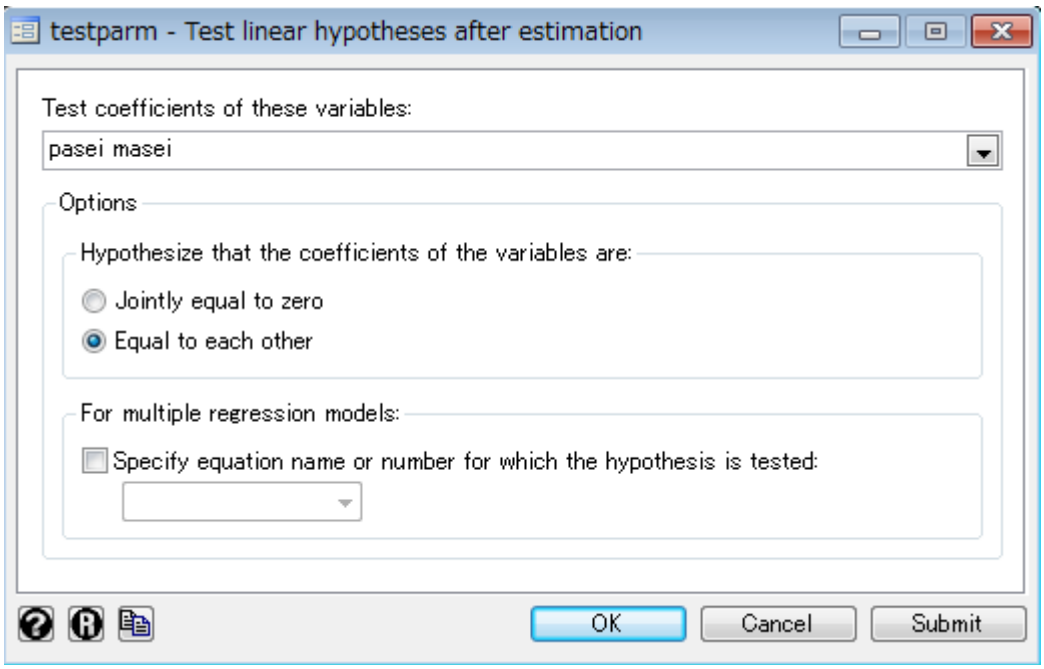
   (訂正文) 4. `gss2002_chapter10.dta` を開いてください．ある男性の社会経済的状況は，父の社会経済的状況のほうに母のそれより強く依存すると考えたとします．`sei`，`pasei`，`masei` を使用して，回帰分析を行ってください．その際，週労働時間 `hrs1` をコントロールしてください．`pasei` と `masei` の係数が等しいか検定してください．図 10.7 のようなグラフを作成し，残差の分布を慎重に解釈してください．

回帰分析は，`regress` コマンドで行えます．今回は男性についてのみ回帰を実施するので，コンマの前に `if` 文を含めます．

```
. regress sei pasei masei hrs1 if sex==1, beta

      Source |       SS           df       MS      Number of obs   =       393
-------------+----------------------------------   F(3, 389)       =     16.28
       Model |  16385.6873         3  5461.89575   Prob > F        =    0.0000
    Residual |  130503.416       389   335.48436   R-squared       =    0.1116
-------------+----------------------------------   Adj R-squared   =    0.1047
       Total |  146889.103       392  374.717101   Root MSE        =    18.316

------------------------------------------------------------------------------
         sei |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
       pasei |   .1752678   .0518092     3.38   0.001                 .1736297
       masei |   .2276482   .0519268     4.38   0.000                  .224405
        hrs1 |   .0534024   .0632797     0.84   0.399                 .0404509
       _cons |   31.08142   3.947154     7.87   0.000                        .
------------------------------------------------------------------------------
```

結果は，父の社会経済的状況と母の社会経済的状況が共に，息子の社会経済的状況に対する有意な効果を示しています．父についての標準化後回帰係数は $\beta = 0.18, t = 3.38, p < 0.01$ です．母についての標準化後回帰係数は $\beta = 0.23, t = 4.38, p < 0.001$ です．モデルは，全体として息子の社会経済的状況の変動の 11% を説明し，$F(3, 389) = 16.28, p < 0.001$ です．これを見ると，母は父よりもわずかに影響が大きいように見えます．両者の差異が統計的に有意かを検定するには，回帰後に `testparm` コマンドを実施します．`testparm` ダイアログボックスを開くには，Statistics ▷ Postestimation ▷ Test parameters.（Stata 14 では 2015 年 11 月 17 日現在，メ

ニューからは選択できません）を選択します．開いたダイアログボックスで，比較したい 2 つの変数を指定し，それらの係数が等しいかを検定する選択肢を選択します．指定後のダイアログボックスは，以下です．



上記によるコマンドとその結果は，以下です．

```
. testparm pasei masei, equal
 ( 1)  - pasei + masei = 0
       F(  1,   389) =    0.38
            Prob > F =    0.5403
```

上記の代わりに，以下のようなさらに直感的な方法で検定することもできます．

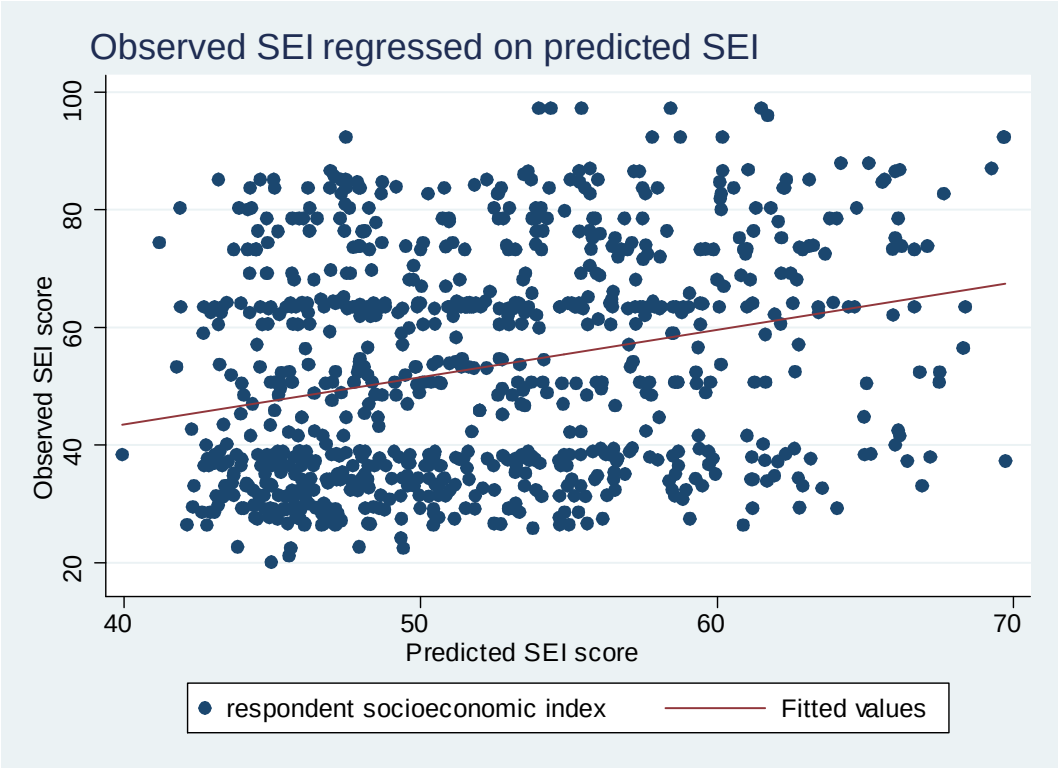
```
. test pasei = masei
 ( 1)  pasei - masei = 0
       F(  1,   389) =    0.38
            Prob > F =    0.5403
```

この方法では `pasei = masei` の検定と明示的に入力し，さらに `pasei - masei = 0` と書き換えられているため，内容がより明確かもしれません．上記の 2 つは，どちらも $F(1, 389) = 0.38, p$ not significant で有意性は示されませんでした．

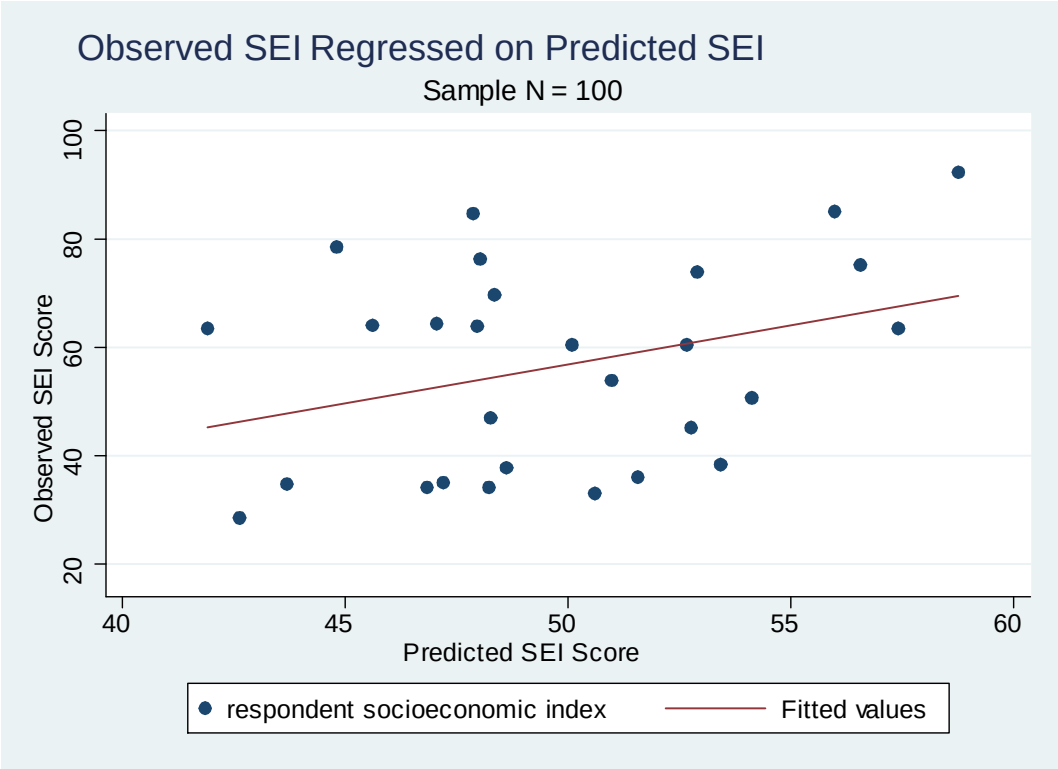
β 重み（標準化後偏回帰係数）は多くの論文に掲載され，値同士は有意性に差があるものとされます．たとえば，t 検定で有意性があった係数もあれば，そうでない係数もあります．これらの t 検定は，係数がゼロであるという仮説検定について行われます．ゼロとは有意差のある係数が，必ずしもほかの係数の有意差があるとは限りません．一方の係数がもう一方より強力かどうかを示すには，同等性の検定を両係数に対して行います．標準化後の β の同等性は，標準化前の B の同等性検定で暗示的に示すことができます．

最後に，息子の社会経済的状況に関し，予測値に対する観測値のプロットを行います．予測値を独立変数として扱うと聞くと，奇異に聞こえるかもしれませんが，残差の様子を見るには良い方法です．

一つは観測したデータ全体をもとに作成されるグラフで，もう一つは抽出した標本 100 個をもとに作成されるグラフです．こうした抽出後のデータでのグラフ作成は，データの数が膨大でグラフ化すると解釈が難しい場合にのみ用います．



上記のグラフでは，ドットがあまりにも多く，重なっているかどうかも分からないため，読むのが少し大変です．次は，標本を 100 個抽出して描いたグラフです．

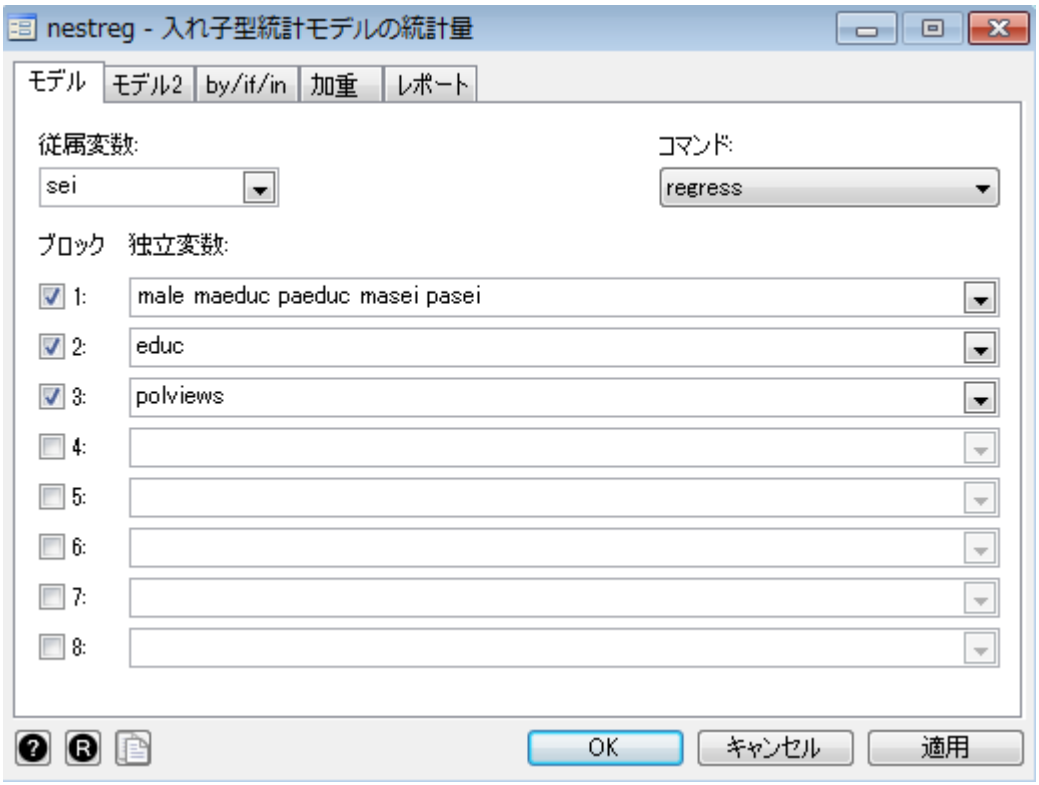


残差は予測値全体にわたり，きわめて均一に分散している様子が見られます．

5. まず，`recode` コマンドで変数 `sex` を再コード化し，`male` というダミー変数を作成します．

```
. recode sex (1=1 male) (2=0 female), generate(male)
(1537 differences between sex and male)
```

入れ子型回帰分析のコマンドは，本体書籍にもあるようにタイプもできますが，ダイアログボックスでの作成もできます．本体書籍の内容を逸脱するので，ここでこのダイアログボックスを説明します．Statistics ▷ Other ▷ Nested model statistics（統計 (S) ▷ その他 ▷ 入れ子型統計モデルの統計量）を選択します．Model（モデル）タブでは，以下のように入力します．



Dependent variable（従属変数）に `sei` を指定します．演習の指示に基づいて 3 つのブロックを選択します．1 つ目のブロックに，*Independent variables*（独立変数）に `male maeduc paeduc masei pasei` とを指定します．2 つ目のブロックには `educ`，3 つ目のブロックには `polviews` をそれぞれ指定します．右上にある *Command*（コマンド）で，`regress` を選択します．ボックスの右端にある矢印をクリックし，プルダウンメニューを表示すると，様々な種類の OLS 回帰が選択できるのが分かります．それぞれで適切なタイプの回帰がブロックに対して実施されます．このコマンドはとても強力なコマンドです．`regress` を実施するときには，通常 `beta` オプションが用いられますが，Model2（モデル 2）タブでは，こうした各コマンドに有効なオプションを入力できるボックスがあります．今回はここに `beta` と入力します．
適用をクリックすると，以下を得ます．

```
. nestreg : regress sei (male maeduc paeduc masei pasei) (educ) (polviews), beta
Block  1: male maeduc paeduc masei pasei
```

Source	SS	df	MS			
				Number of obs	=	531
				F(5, 525)	=	14.68
Model	23042.7443	5	4608.54887	Prob > F	=	0.0000
Residual	164860.423	525	314.019853	R-squared	=	0.1226
				Adj R-squared	=	0.1143
Total	187903.167	530	354.534278	Root MSE	=	17.721

sei	Coef.	Std. Err.	t	P>|t|	Beta
male	-1.189079	1.547602	-0.77	0.443	-.0314593
maeduc	.527248	.3373834	1.56	0.119	.0904839
paeduc	.2962253	.292221	1.01	0.311	.0589205
masei	.1162678	.051822	2.24	0.025	.1162073
pasei	.1786331	.0504431	3.54	0.000	.1796234
_cons	28.27565	3.299471	8.57	0.000	.

Block 2: educ

Source	SS	df	MS			
				Number of obs	=	531
				F(6, 524)	=	43.00
Model	61990.7508	6	10331.7918	Prob > F	=	0.0000
Residual	125912.416	524	240.290871	R-squared	=	0.3299
				Adj R-squared	=	0.3222
Total	187903.167	530	354.534278	Root MSE	=	15.501

sei	Coef.	Std. Err.	t	P>|t|	Beta
male	.2879349	1.358746	0.21	0.832	.0076178
maeduc	-.0411352	.2984878	-0.14	0.890	-.0070594
paeduc	-.2523983	.2592306	-0.97	0.331	-.0502031
masei	.1005997	.0453486	2.22	0.027	.1005474
pasei	.1378762	.0442417	3.12	0.002	.1386404
educ	3.354689	.2634985	12.73	0.000	.5066337
_cons	-2.827332	3.78138	-0.75	0.455	.

Block 3: polviews

Source	SS	df	MS			
				Number of obs	=	531
				F(7, 523)	=	37.39
Model	62674.7443	7	8953.5349	Prob > F	=	0.0000
Residual	125228.423	523	239.442491	R-squared	=	0.3335
				Adj R-squared	=	0.3246
Total	187903.167	530	354.534278	Root MSE	=	15.474

sei	Coef.	Std. Err.	t	P>|t|	Beta
male	.2822825	1.356349	0.21	0.835	.0074683
maeduc	-.0827567	.2989764	-0.28	0.782	-.0142023
paeduc	-.2743854	.2590994	-1.06	0.290	-.0545764
masei	.1027444	.0452863	2.27	0.024	.102691
pasei	.1403687	.0441882	3.18	0.002	.1411468
educ	3.344203	.2631061	12.71	0.000	.5050501
polviews	-.8334485	.4931209	-1.69	0.092	-.0610382
_cons	1.227449	4.472568	0.27	0.784	.

Block	F	Block df	Residual df	Pr > F	R2	Change in R2
1	14.68	5	525	0.0000	0.1226	
2	162.09	1	524	0.0000	0.3299	0.2073
3	2.86	1	523	0.0916	0.3335	0.0036

結果から，政治思想をモデルに追加しても，社会経済状況に対しては統計的に有意な効果は持たないことが読み取れます．1 つ目と 2 つ目のブロックにより息子の社会経済状況の変動が 33% 説明され，そこに政治思想を追加しても，説明できる変動は 0.4 パーセントポイントしか増えず，$F(1, 523) = 2.86, p < 0.10$ と有意ではありません．

表 1：政治信条の保守性と背景要因変数による社会経済状況の予測を行った回帰分析の要約表（$N = 531$）

	Model 1			Model 2			Model 3		
変数	B	SE B	β	B	SE B	β	B	SE B	β
性別	−1.19	1.55	−0.03	0.29	1.36	0.01	0.28	1.35	0.01
母の教育水準	0.53	0.34	0.09	−0.04	0.30	−0.01	−0.08	0.30	−0.01
父の教育水準	0.30	0.29	0.06	−0.25	−0.97	−0.05	−0.27	0.26	−0.05
母の社会経済状況	0.12*	0.12	0.12	0.10*	2.22	0.10	0.10*	0.04	0.10
父の社会経済状況	0.18*	0.05	0.18	0.14**	3.12	0.14	0.14**	0.04	0.14
教育水準				3.35*	0.26	0.51	3.34***	0.26	0.50
政治信条の保守性							−0.83	0.49	−0.06
切片	28.28***	3.30		−2.83	−0.75		1.23	4.47	
R^2	0.12			0.33			0.33		
R^2 の増分の F	14.68***			162.09***			2.86		

注：年齢と鬱傾向は各平均でセンタリング化した．

*$p < 0.05$　**$p < 0.01$　***$p < 0.001$

表 1 のような表は有用です．先ほどは，背景要因変数をコントロールすると政治信条の保守性が社会経済状況をより高い精度で予測できないと述べました．表からは，主要因は自身の教育水準であることが分かります．この要因は，R^2 を有意に押し上げ，モデル 2 とモデル 3 において極めて強力な β 重みを示しています．

6. ((訂正) 本演習については本体書籍の記述に意図しない改行が入るという誤植がありました．以下のように訂正するとともに，心からお詫び申し上げます．)

(訂正文) 6. c10interaction.dta を開いてください．このデータセットについて，本章で行っている回帰分析を再び実行してください．その際，性別と教育の交互作用項を収める変数を新たに作成してください．summarize などを用いて平均を調べるか，center コマンドをインストールなどして，educ を平均の値でセンタリングしてください．センタリングした値は，新たな変数 c_educ に収めてください．その後，generate educc_male = c_educ*male で交互作用項を作成してください．regress inc male c_educ educc_male で回帰分析を実施し，2 つのグラフを重ねた 2 軸グラフを作成してください．作成したグラフを，本章のものと比較してください．両グラフで切片が異なるのはなぜですか．切片が，予測変数の値がゼロのときの結果変数の値であること，および，センタリングした後の予測変数のゼロの意味を踏まえて考えてください．

交互作用項に用いられる連続変数のセンタリングは，常に実施されたり全く実施されなかったりと，分析者によって様々です．センタリングを実施することにより，切片に一定の意味が生まれます．センタリングは center コマンドで実施できます．同コマンドは findit center により見つけことができます．center を実行すると，変数名の先頭に c_が付いた新たな変数が作成されます．すなわち，center educ とすると c_educ が作成されます．
センタリング後，以下により交互作用項を作成します．

```
. generate educc_male = c_educ*male
```

続いて回帰を実施します．演習 5 のように入れ子型回帰を行うか，!missing の指定で標本の大きさを揃えた regress を行うかの 2 通りの方法が取れます．regress による方法は，以下です．

```
. regress inc c_educ male if !missing(inc, c_educ, male, educc_male)
      Source |       SS           df       MS      Number of obs   =       120
-------------+----------------------------------   F(2, 117)       =     37.19
       Model |  100464.105         2  50232.0527   Prob > F        =    0.0000
    Residual |  158015.895       117   1350.5632   R-squared       =    0.3887
-------------+----------------------------------   Adj R-squared   =    0.3782
       Total |     258480       119  2172.10084   Root MSE        =     36.75

------------------------------------------------------------------------------
         inc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      c_educ |   8.045694   1.008586     7.98   0.000     6.048243    10.04315
        male |   19.04991   6.719787     2.83   0.005     5.741726     32.3581
       _cons |   66.47504   4.748008    14.00   0.000     57.07186    75.87822
------------------------------------------------------------------------------
```

```
. regress inc c_educ male educc_male if !missing(inc, c_educ, male, educc_male)
      Source |       SS           df       MS      Number of obs   =       120
-------------+----------------------------------   F(3, 116)       =     34.89
       Model |  122604.719         3  40868.2397   Prob > F        =    0.0000
    Residual |  135875.281       116  1171.33863   R-squared       =    0.4743
-------------+----------------------------------   Adj R-squared   =    0.4607
       Total |      258480       119  2172.10084   Root MSE        =    34.225

------------------------------------------------------------------------------
         inc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      c_educ |   3.602369   1.388076     2.60   0.011     .8531092    6.351628
        male |   19.17645   6.258121     3.06   0.003     6.781452    31.57145
  educc_male |   8.196446   1.885263     4.35   0.000     4.462445    11.93045
       _cons |   65.66043   4.425728    14.84   0.000     56.89472    74.42615
------------------------------------------------------------------------------
```

`!missing` は便利な関数です．冒頭の`!`（感嘆符）は“not（否定）”を意味します．カッコ内では変数をコンマで区切る必要があります．交互作用項を含む全変数をカッコに指定して，入れ子型回帰のときと同じ標本に対する分析ができます．`!missing` は，今回の回帰がもし単独の回帰であれば必要ではありませんでした．しかし，同関数は必要なときのために覚えておくと便利な関数です．

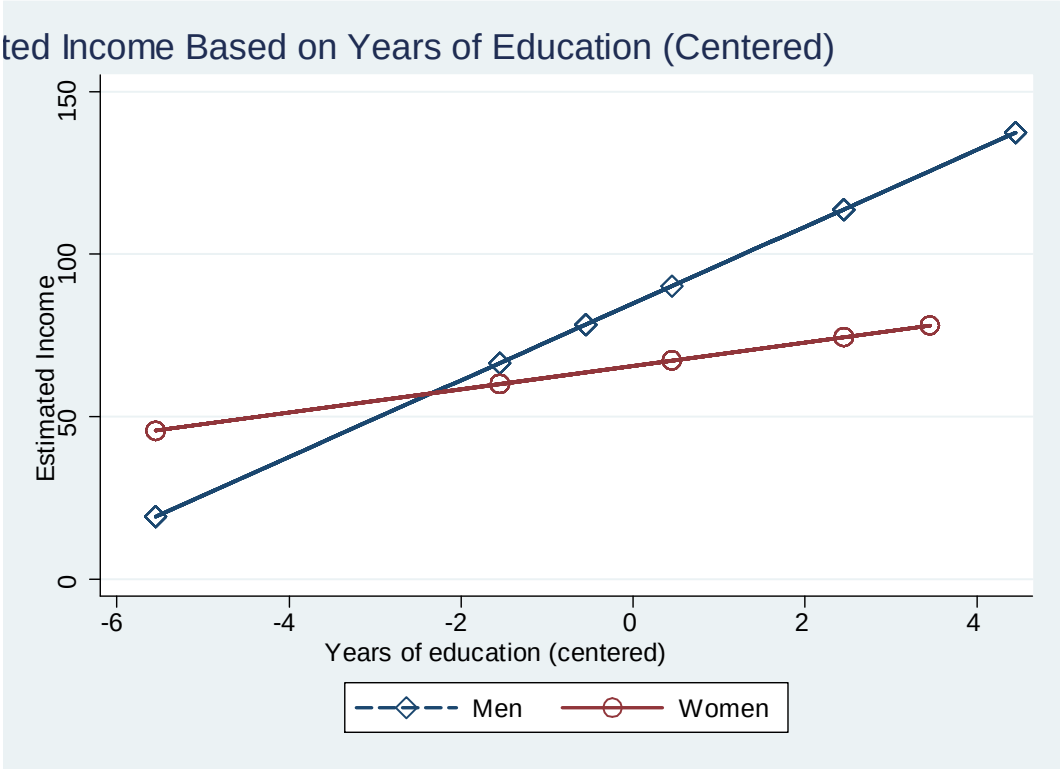
結果では予測変数が 3 つとも統計的に有意でした．その効果に有意性が示されるときは，交互作用項もモデルに含める必要があります．交互作用項の主効果への影響を見るため，グラフにして確認しなければいけません．

グラフをプロットする前に，事後分析コマンド `predict` を実行します．同コマンドで，女性の所得に関するスコア `incf`，および男性の所得に関するスコア `incm` を予測します．各コマンドは，以下です．

```
. predict incf if male==1
. predict incm if male==0
```

これで本体書籍の 9.6 節のようなグラフを作成する準備が整いました．コマンドとグラフは，以下です．

```
. twoway (connected incf c_educ if male==1, lpattern(dash) lwidth(medthick) msymbol(d
> iamond_hollow) msize(medlarge)) (connected incm c_educ if male==0, lwidth(medthick)
>  msymbol(circle_hollow) msize(large)), ytitle(Estimated Income) title(Estimated Inc
> ome Based on Years of Education (Centered)) legend(order(1 "Men" 2 "Women"))
```



グラフからは，予想通りの結果が表れているといえます．教育水準を積み上げることによる所得への対価が，男性で女性より高くなっています．回帰式は，以下です．

$$\texttt{inc}\ \text{の予測値} = 65.66 + 3.60(\texttt{educ}) + 19.18(\texttt{male}) + 8.20(\texttt{educ} \times \texttt{male})$$

女性では，回帰式が以下になります（`male` と `ed_male` が女性ではゼロのため，回帰式からなくなります）．

$$\texttt{inc}\ \text{の予測値} = 65.66 + 3.60(\texttt{educ})$$

男性では，回帰式が以下になります．

$$\texttt{inc}\ \text{の予測値} = (65.66 + 19.18) + (3.60 + 8.20)(\texttt{educ})$$

$$\texttt{inc}\ \text{の予測値} = 84.84 + 11.80(\texttt{educ})$$

切片の解釈：educ がセンタリングされているとき，その平均はゼロです．従って，女性についての回帰式の切片である 65.66 は，教育水準が女性の平均である女性の所得平均です．同様に，男性の 84.84 は，教育水準が男性の平均である男性の所得平均です．分析者によっては，センタリングを行わず，切片が教育水準がゼロのときの所得を表すようにする方法を好む場合もあります．しかし，実際に教育を全く受けなかった人は稀であるため，切片の値自体の有用性は低くなります．グラフを見ると，男性の所得は教育水準の低いケースでは女性より低い一方で，教育水準の高いケースでは女性よりも高くなり，同じ教育を受けたときの恩恵が男性のほうが女性より高いと言えます．

7. ((訂正) 本演習については本体書籍において，先にある演習 6 の記述に意図せず行われた改行により演習番号が一つ繰り上がる誤植がありました．このため，本体書籍で演習 7 とされた内容の解答は，上記にある演習 6 についての解答に含まれています．以下は，本体書籍で演習 8 とされた内容の解答です．何卒，ご容赦ください．)

   演習 6 では，以下のように male と c_educ の交互作用項を作成し，変数に収めました．

```
. generate educc_male = c_educ*male
```

   上記の変数を含め，回帰分析を実施します．

```
. regress inc c_educ male educc_male
      Source |       SS           df       MS      Number of obs   =       120
-------------+----------------------------------   F(3, 116)       =     34.89
       Model |  122604.719         3  40868.2397   Prob > F        =    0.0000
    Residual |  135875.281       116  1171.33863   R-squared       =    0.4743
-------------+----------------------------------   Adj R-squared   =    0.4607
       Total |      258480       119  2172.10084   Root MSE        =    34.225

------------------------------------------------------------------------------
         inc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      c_educ |   3.602369   1.388076     2.60   0.011     .8531092    6.351628
        male |   19.17645   6.258121     3.06   0.003     6.781452    31.57145
  educc_male |   8.196446   1.885263     4.35   0.000     4.462445    11.93045
       _cons |   65.66043   4.425728    14.84   0.000     56.89472    74.42615
------------------------------------------------------------------------------
```

   同様のことを，i.，c.，#記号を使用して 1 行の regress で実施することもできます．

```
. regress inc c_educ male i.male#c.c_educ, beta
      Source |       SS           df       MS      Number of obs   =       120
-------------+----------------------------------   F(3, 116)       =     34.89
       Model |  122604.719         3  40868.2397   Prob > F        =    0.0000
```

```
    Residual |  135875.281       116  1171.33863   R-squared       =    0.4743
-------------+------------------------------------   Adj R-squared   =    0.4607
       Total |     258480       119  2172.10084   Root MSE        =    34.225

------------------------------------------------------------------------------
         inc |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
      c_educ |   3.602369   1.388076     2.60   0.011                 .2585699
        male |   19.17645   6.258121     3.06   0.003                 .2065929
             |
male#c.c_educ |
           1 |   8.196446   1.885263     4.35   0.000                 .4328146
             |
       _cons |   65.66043   4.425728    14.84   0.000                        .
------------------------------------------------------------------------------
```

結果自体は変わりません．コマンドに交互作用項を数式で指定すると Stata が計算を実施してくれることが，ここでのポイントです．数式は `i.male#c.c_educ` です．`i.male` により `male` がカテゴリ変数であること，`c.c_educ` により `c_educ` が連続変数であることを Stata に伝えています．`#`は前後の変数を交互作用させることを伝えています．今回は状況が簡単で，`i.`，`c.`，`#`記号を使用したことによる特典は少ないと認めざるを得ません．もし 4 つのカテゴリで分類した政治志向（`pol`）と 5 つのカテゴリで分類した宗教（`rel`）があるとすると，`i.pol#i.rel` は可能な組み合わせすべてについての交互作用を作成します．たとえば，共和党派とカトリックの交互作用，共和党派とプロテスタントの交互作用，共和党派とイスラム教の交互作用などです．

8. ((訂正) 本演習については本体書籍において，先にある演習 6 の記述に意図せず行われた改行により演習番号が一つ繰り上がる誤植がありました．このため，本体書籍で演習 8 とされた内容の解答は，上記にある演習 7 についての解答に含まれています．以下は，本体書籍で演習 9 とされた内容の解答です．何卒，ご容赦ください．)
分析前に把握しておくことは何でしょう．

   `r2f` は大きいほうの（full model の）R^2 です= 表を参照すると中程度の効果を示す値は 0.13
   `r2r` は小さいほうの（reduced model または 0 の）R^2 です $= 0$
   `nvar` は予測変数の数です $= 3$
   `ntest` は検出力検定で吟味する予測変数の数です $= 3$
   `alpha` は有意水準です $= 0.05$
   `power` は検出力です $= 0.90$
   `n` は標本数の大きさです = ここで予測します

```
. powerreg, r2f(0.13) r2r(0) nvar(3) ntest(3) alpha(0.05) power(0.90)
 Linear regression power analysis
 alpha=.05  nvar=3  ntest=3
 R2-full=.13  R2-reduced=0  R2-change=0.1300
 nominal     actual
  power       power          n
  0.9000     0.9038        100
```

9. ((訂正) 本演習については本体書籍において，先にある演習 6 の記述に意図せず行われた改行により演習番号が一つ繰り上がる誤植がありました．このため，本体書籍で演習 9 とされた内容の解答は，上記にある演習 8 についての解答に含まれています．以下は，本体書籍で演習 10 とされた内容の解答です．何卒，ご容赦ください．)
分析前に把握しておくことは何でしょう．

`r2f` は大きいほうの（full model の）R^2 です$= 0.20 + 0.05 = 0.25$
`r2r` は小さいほうの（reduced model または 0 の）R^2 です $= 0.20$
`nvar` は予測変数の数です $= 4$
`ntest` は検出力検定で吟味する予測変数の数です $= 1$
`alpha` は有意水準です $= 0.05$
`power` は検出力です $= 0.90$
`n` は標本数の大きさです = ここで予測します

```
. powerreg, r2f(0.25) r2r(0.20) nvar(4) ntest(1) alpha(0.05) power(0.90)
 Linear regression power analysis
 alpha=.05  nvar=4  ntest=1
 R2-full=.25  R2-reduced=.2  R2-change=0.0500
 nominal     actual
  power       power          n
  0.9000     0.9043        162
```

10. ((訂正) 本演習については本体書籍において，先にある演習 6 の記述に意図せず行われた改行により演習番号が一つ繰り上がる誤植がありました．このため，本体書籍で演習 10 とされた内容の解答は，上記にある演習 9 についての解答に含まれています．以下は，本体書籍で演習 11 とされた内容の解答です．何卒，ご容赦ください．)
分析前に把握しておくことは何でしょう．

`r2f` は大きいほうの（full model の）R^2 です$= 0.20 + 0.06 = 0.26$
`r2r` は小さいほうの（reduced model または 0 の）R^2 です $= 0.20$

nvar は予測変数の数です = 4

ntest は検出力検定で吟味する予測変数の数です = 1

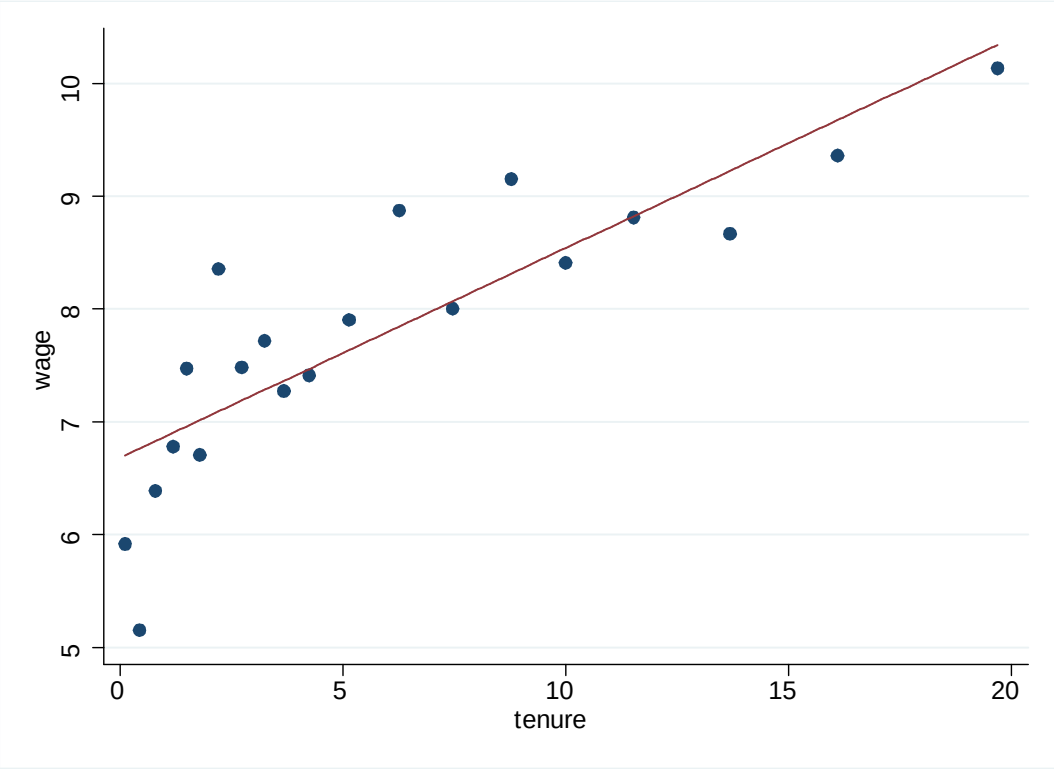
alpha は有意水準です = 0.05

power は検出力です = ここで予測します

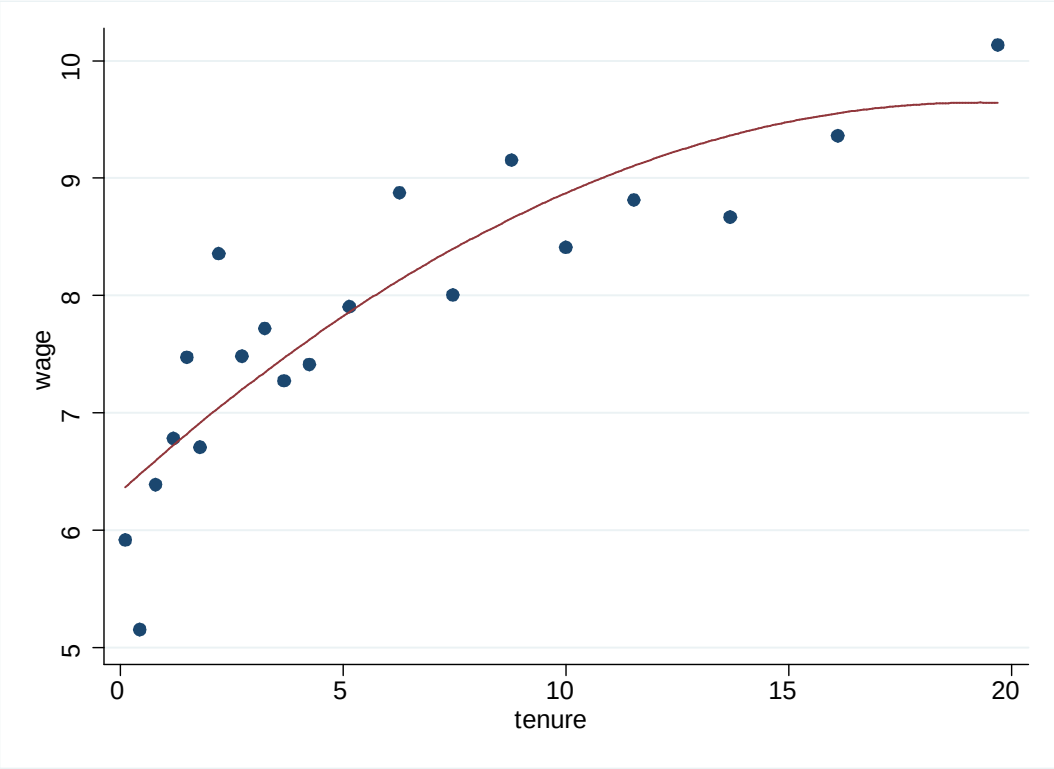
n は標本数の大きさです = 75

```
. powerreg, r2f(0.26) r2r(0.20) nvar(4) ntest(1) alpha(0.05) n(75)
 Linear regression power analysis
 alpha=.05  nvar=4  ntest=1
 R2-full=.26  R2-reduced=.2  R2-change=.06
 n = 75    power = 0.6816
```

11. ((訂正) 本演習については本体書籍において，先にある演習 6 の記述に意図せず行われた改行により演習番号が一つ繰り上がる誤植がありました．このため，本体書籍で演習 11 とされた内容の解答は，上記にある演習 10 についての解答に含まれています．以下は，本体書籍で演習 12 とされた内容の解答です．何卒，ご容赦ください．)
まず，`binscatter wage tenure` を実行して，以下を得ます．



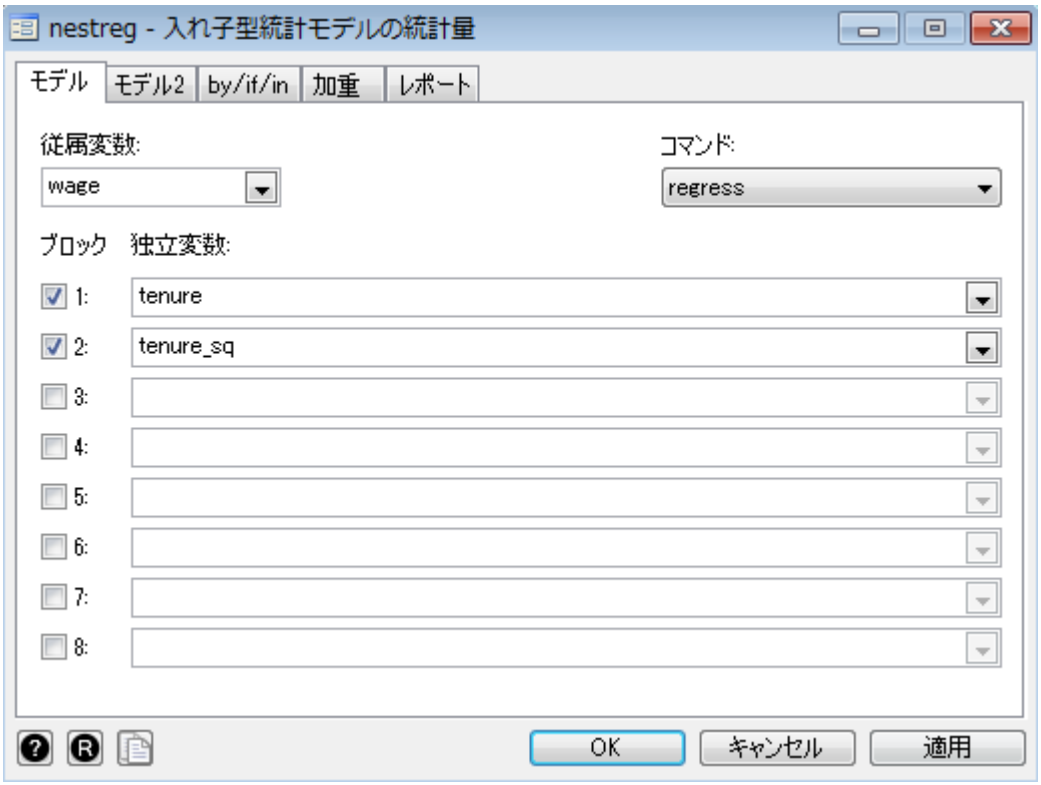
続いて，`binscatter wage tenure, line(qfit)` を実行して，以下を得ます．



2 乗項は必要に見えるでしょうか．そのように見えるかもしれませんが，検定して確かめます．まず，勤続年数の 2 乗項を作成します．

```
. gen tenure_sq = tenure*tenure
```

続いて，入れ子型回帰を行い，1 つ目のブロックに `tenure`，2 つ目のブロックに `tenure_sq` を入力します．Statistics ▷ Other ▷ Nested model statistics（統計 (S) ▷ その他 ▷ 入れ子型統計モデルの統計量）を選択します．



結果は，以下です．

```
. nestreg : regress wage (tenure) (tenure_sq)
Block  1: tenure
      Source |       SS           df       MS      Number of obs   =     2,231
-------------+----------------------------------   F(1, 2229)      =     72.66
       Model |  2339.38077         1  2339.38077   Prob > F        =    0.0000
    Residual |  71762.4469     2,229  32.1949066   R-squared       =    0.0316
-------------+----------------------------------   Adj R-squared   =    0.0311
       Total |  74101.8276     2,230  33.2295191   Root MSE        =    5.6741

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tenure |   .1858747   .0218054     8.52   0.000     .1431138    .2286357
       _cons |   6.681316   .1772615    37.69   0.000     6.333702    7.028931
------------------------------------------------------------------------------

Block  2: tenure_sq
      Source |       SS           df       MS      Number of obs   =     2,231
-------------+----------------------------------   F(2, 2228)      =     39.14
       Model |  2514.91563         2  1257.45782   Prob > F        =    0.0000
```

Residual	71586.912	2,228	32.1305709	R-squared	=	0.0339
				Adj R-squared	=	0.0331
Total	74101.8276	2,230	33.2295191	Root MSE	=	5.6684

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
tenure	.3436887	.0709456	4.84	0.000	.2045623	.482815
tenure_sq	-.0089112	.0038125	-2.34	0.020	-.0163877	-.0014347
_cons	6.326825	.2331543	27.14	0.000	5.869602	6.784047

Block	F	Block df	Residual df	Pr > F	R2	Change in R2
1	72.66	1	2229	0.0000	0.0316	
2	5.46	1	2228	0.0195	0.0339	0.0024

2 つ目のブロックの予測変数は tenure_sq しかなく，$B = -0.009, t = -2.34, p < 0.05$ であることから，2 乗項に有意性があるとすぐに判断できます．その効果の程は最下部の表で明らかになっています．$F(1, 2228) = 5.46, p < 0.05$ であり，（2 乗項だけを含む）ブロックの R^2 の増分は 0.0024 です．回帰線を曲線にすることで，R^2 が 0.24% 増加し，$R^2 = 0.0316$ から $R^2 = 0.0339$ へと変化しました．数字からは曲線にする重要性は確かに見られますが，その統計的有意性が見られた 2 乗項の追加による説明可能な変動の増分はとても小さいものでした．

12. ((訂正) 本演習については本体書籍において，先にある演習 6 の記述に意図せず行われた改行により演習番号が一つ繰り上がる誤植がありました．このため，本体書籍で演習 12 とされた内容の解答は，上記にある演習 11 についての解答に含まれています．以下は，本体書籍で演習 13 とされた内容の解答です．何卒，ご容赦ください．)

変数をセンタリングする方法は 3 つあります．1 つ目は，ユーザ作成コマンド center（未インストールの場合は ssc install center を実行してください）を用いる方法です．2 つ目は，summarize tenure を実行して平均値を確認し，その平均値（5.97785）を tenure から引いた値を持つ新たな変数を作成する方法です．そして 3 つ目は，summarize tenure を実行し，tenure から r(mean) を得る方法で，おそらく 2 つ目のものより良い方法です．r(mean) は summarize コマンドにより内部で保存される値で，数値精度が高い値です．3 つの方法のコマンドは，以下です．

```
. sysuse nlsw88.dta, clear
. center tenure
. label variable c_tenure "Center tenure by using center command"
. sum tenure
. gen ctenure = tenure - 5.97785
. label variable ctenure "Center tenure by subtracting 5.97785"
```

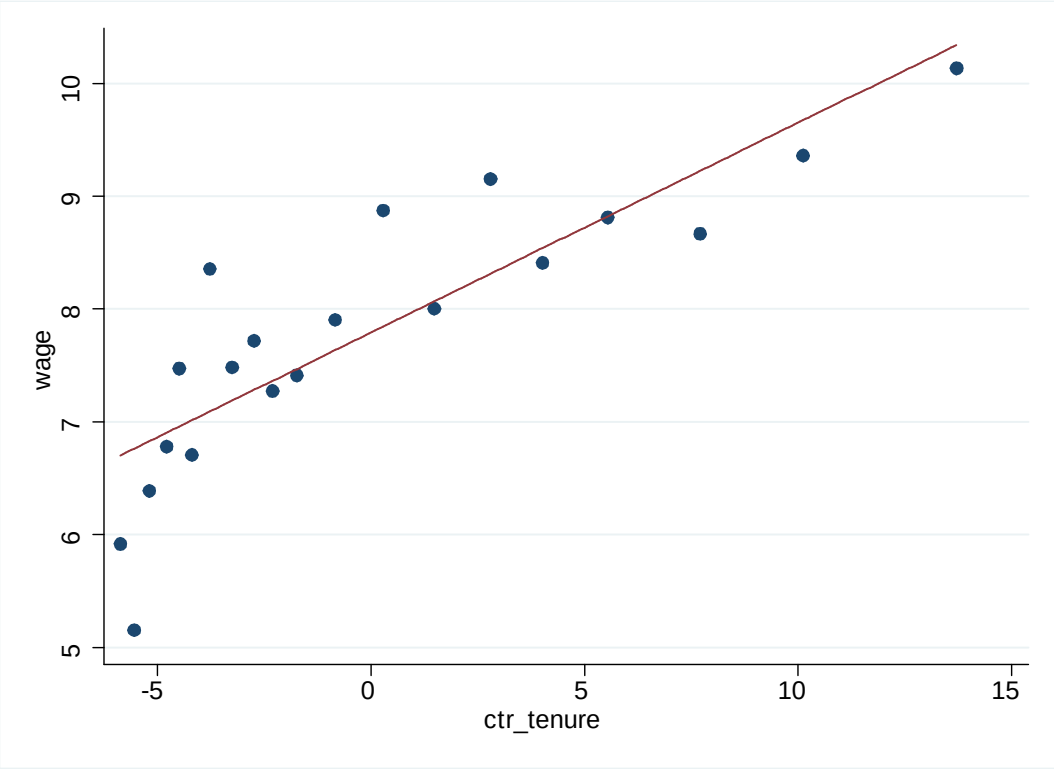
```
. sum tenure
. gen ctr_tenure = tenure - r(mean)
. label variable ctr_tenure "Center tenure by subtracting r(mean)"
. sum c_tenure ctenure ctr_tenure
```

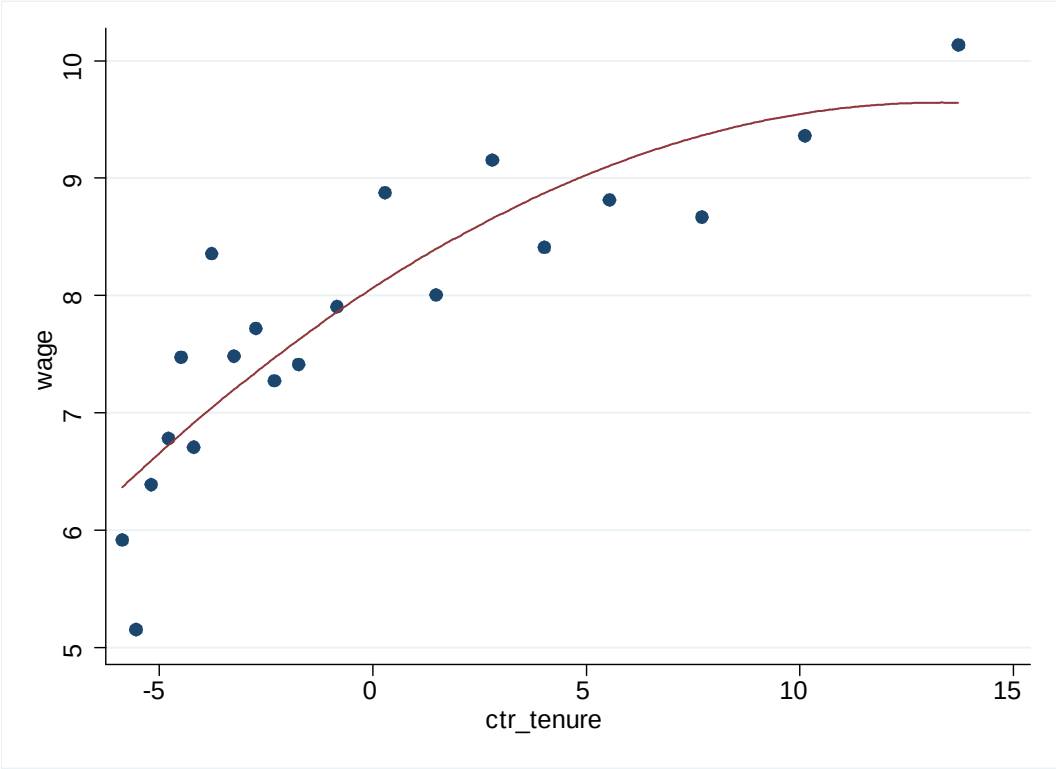
最後のコマンドでは，3 つの方法による結果をまとめています．

```
. sum c_tenure ctenure ctr_tenure
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     c_tenure |      2,231    3.87e-08    5.510331   -5.97785   19.93882
      ctenure |      2,231    3.44e-08    5.510331   -5.97785   19.93882
   ctr_tenure |      2,231    3.44e-08    5.510331   -5.97785   19.93882
```

e-08 は，小数点を左に 8 桁移動させることを意味し，実際の値は 0.0000000344，すなわち実質的にゼロであることを示しています．上記のように，3 つの方法はいずれも同様の結果を生み出しました．これで，得点がゼロのデータは勤続年数が平均値の人を代表するようになりました．

binscatter wage ctr_tenure および binscatter wage ctr_tenure, line(qfit) によるグラフは，演習 11 のものと同じ形ですが，今回は勤続年数に関してセンタリングしたグラフです．


wage
5
6
7
8
9
10
-5
0
5
10
15
ctr_tenure



nestreg コマンドにより，2 乗項 ctr_tenure_sq が重要かどうかを検定することができます．結果の最下部に表れる要約表は演習 11 のものと変わりません．2 乗項の追加により，説明可能な変動の割合 R^2 が 0.0024，あるいは 0.24% 増加します．この項は $F(1, 2228) = 5.46, p < 0.05$ と有意ですが，実践的に意味のある貢献はありません．実はこれについては，グラフ上を見ると 5 年以上の相対勤続年数を持つ点が比較的少ないことからヒントが得られます．

演習 11 と大きく異なるのは，回帰係数の値です．切片の値を見ると，1 つ目のブロックが 7.79 です．これは勤続年数が平均の人に対する wage の値です．演習 11 では，切片が 6.68 という勤続年数がゼロの人に対する wage の値でした．

さらに，ctr_tenure_sq を追加した 2 つ目のブロックには，大きな違いが表れています．演習 11 では tenure について $B = 0.344$ でした．この値は，勤続年数がゼロの時点における wage の変化率を示し，tenure= 0.0 での曲線の傾きを示します．今回はこの値が $B = 0.237$ であり，勤続年数が平均の時点での変化率，すなわち ctr_tenure_sq= 0.0 での曲線の傾きを示します．

注目したいケースがどちらになるかは時と場合によりますが，両者には明らかな違いがあります．両者により，勤め始めは賃金の上昇が早いのに対し，年数を経るにつれて上昇率が低下していく事実が示されています．

```
. nestreg : regress wage (ctr_tenure) (ctr_tenure_sq)
Block  1: ctr_tenure
      Source |       SS           df       MS      Number of obs   =     2,231
-------------+----------------------------------   F(1, 2229)      =     72.66
       Model |  2339.38077         1  2339.38077   Prob > F        =    0.0000
    Residual |  71762.4469     2,229  32.1949066   R-squared       =    0.0316
-------------+----------------------------------   Adj R-squared   =    0.0311
       Total |  74101.8276     2,230  33.2295191   Root MSE        =    5.6741

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  ctr_tenure |   .1858747   .0218054     8.52   0.000     .1431138    .2286357
       _cons |   7.792448   .1201279    64.87   0.000     7.556873    8.028022
------------------------------------------------------------------------------

Block  2: ctr_tenure_sq
      Source |       SS           df       MS      Number of obs   =     2,231
-------------+----------------------------------   F(2, 2228)      =     39.14
       Model |  2514.91564         2  1257.45782   Prob > F        =    0.0000
    Residual |   71586.912     2,228  32.1305709   R-squared       =    0.0339
-------------+----------------------------------   Adj R-squared   =    0.0331
       Total |  74101.8276     2,230  33.2295191   Root MSE        =    5.6684

------------------------------------------------------------------------------
         wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+----------------------------------------------------------------
   ctr_tenure |   .2371488   .0309152     7.67   0.000     .1765231    .2977745
ctr_tenure_sq |  -.0089112   .0038125    -2.34   0.020    -.0163877   -.0014347
        _cons |   8.062904   .1667062    48.37   0.000     7.735989     8.38982
------------------------------------------------------------------------------
```

Block	F	Block df	Residual df	Pr > F	R2	Change in R2
1	72.66	1	2229	0.0000	0.0316	
2	5.46	1	2228	0.0195	0.0339	0.0024

第 10 章 (10.15 節, pp.342-344) の do-file

演習 10.1

```
/************** Begin do-file ************/
* chapter10.1.do
use "C:\data\gss2002_chapter10.dta", clear
codebook sex wrkslf
recode sex (1=1 male) (2=0 female), generate(male)
tabulate sex male, missing
recode wrkslf (1=1 self-employed) (2=0 "someone else"), generate(selfemp)
tabulate wrkslf selfemp, missing
regress hrs1 male age selfemp, beta
display "estimated hours = "6.42*0 -.06*20 -1.05*1 + 41.36
/************** End do-file **************/
```

演習 10.2

```
/************** Begin do-file ************/
* chapter10.2.do
use "C:\data\census.dta", clear
regress tworace white hhinc ba
predict yhat, xb
predict rstandard, rstandard
twoway (scatter tworace hhinc) (lfit tworace hhinc), ///
ytitle(Estimated Hours Worked) ///
title(Estimated Hours Worked on Household Income)
twoway (scatter tworace white) (lfit tworace white), ///
ytitle(Estimated Hours Worked) ///
title(Estimated Hours Worked on Percentage of State that is White)
twoway (scatter tworace ba) (lfit tworace ba), ///
ytitle(Estimated Hours Worked) ///
title(Estimated Hours Worked on Percent College Graduate)
list state white hhinc ba tworace yhat rstandard ///
if rstandard>1.96 | rstandard < -1.96
/************** End do-file **************/
```

演習 10.3

```
/************** Begin do-file ************/
* chapter10.3.do
use "C:\data\census.dta", clear
regress tworace white hhinc ba
predict yhat
dfbeta
list state white hhinc ba hhinc tworace yhat _dfbeta_1 _dfbeta_2 _dfbeta_3 ///
if abs(_dfbeta_1) > 2/sqrt(51) | abs(_dfbeta_2) > 2/sqrt(51) ///
| abs(_dfbeta_3) > 2/sqrt(51)
sum white hhinc ba
/************** End do-file **************/
```

演習 10.4

```
/************** Begin do-file ************/
* chaper10.4.do
use "C:\data\gss2002_chapter10", clear
regress sei pasei masei hrs1 if sex==1, beta
test pasei = masei
predict seihat
twoway (scatter sei seihat) (lfit sei seihat), ytitle(Observed SEI Score) ///
xtitle(Predicted SEI Score) title(Observed SEI Regressed on Predicted SEI)
preserve
set seed 222
sample 100, count
twoway (scatter sei seihat) (lfit sei seihat), ytitle(Observed SEI Score) ///
xtitle(Predicted SEI Score) title(Observed SEI Regressed on Predicted SEI) ///
subtitle(Sample N = 100)
restore
/************** End do-file **************/
```

演習 10.5

```
/************** Begin do-file ************/
* chapter10.5.do
use "C:\data\gss2002_chapter10", clear
recode sex (1=1 male) (2=0 female), generate(male)
nestreg: regress sei (male maeduc paeduc masei pasei) (educ) (polviews), beta
/************** End do-file **************/
```

演習 10.6

```
/************** Begin do-file ************/
* chapter10.6.do
use "C:\data\c10interaction.dta", clear
set more off
center educ
generate educc_male = c_educ*male
regress inc c_educ male if !missing(inc, c_educ, male, educc_male)
regress inc c_educ male educc_male if !missing(inc, c_educ, male, educc_male)
* nestreg: regress inc (c_educ male) (educc_male)
predict incm if male==1
predict incf if male==0
twoway (connected incm c_educ if male==1, lpattern(dash) ///
lwidth(medthick) msymbol(diamond_hollow) msize(medlarge)) ///
(connected incf c_educ if male==0, lwidth(medthick) ///
msymbol(circle_hollow) msize(large)), ytitle(Estimated Income) ///
title(Estimated Income Based on Years of Education (Centered)) ///
legend(order(1 "Men" 2 "Women"))
/************** End do-file **************/
```

演習 10.7

```
/************** Begin do-file ************/
* chapter10.7.do
use "C:\data\c10interaction.dta", clear
set more off
center educ
regress inc c_educ male i.male#c.c_educ
generate educc_male = c_educ*male
regress inc c_educ male educc_male
/************** End do-file **************/
```

演習 10.8

```
/************** Begin do-file ************/
* chapter10.8.do
powerreg, r2f(.13) r2r(0) nvar(3) ntest(3) alpha(.05) power(.90)
/************** End do-file **************/
```

演習 10.9

```
/************** Begin do-file ************/
* chapter10.9.do
powerreg, r2f(.25) r2r(0) nvar(4) ntest(1) alpha(.05) power(.90)
/************** End do-file **************/
```

演習 10.10

```
/************** Begin do-file ************/
* chapter10.10.do
powerreg, r2f(.26) r2r(.20) nvar(4) ntest(1) alpha(.05) n(75)
/************** End do-file **************/
```

演習 10.11

```
/************** Begin do-file ************/
* chapter10.11.do
sysuse nlsw88.dta, clear
binscatter wage tenure
binscatter wage tenure, line(qfit)
gen tenure_sq = tenure*tenure
nestreg: regress wage (tenure) (tenure_sq)
/************** End do-file **************/
```

演習 10.12

```
/************** Begin do-file ************/
* chapter10.12.do
sysuse nlsw88.dta, clear
center tenure
label variable c_tenure "Center tenure by using center command"
sum tenure
gen ctenure = tenure - 5.97785
label variable ctenure "Center tenure by subtracting 5.97785"
sum tenure
gen ctr_tenure = tenure - r(mean)
label variable ctr_tenure "Center tenure by subtracting r(mean)"
sum c_tenure ctenure ctr_tenure
binscatter wage ctr_tenure
binscatter wage ctr_tenure, line(qfit)
nestreg : regress wage (ctr_tenure) (ctr_tenure_sq)
/************** End do-file **************/
```

第 11 章 (11.11 節, pp.375-376) の解答

1. 使用できるコマンドは 2 つあります．一つは，デフォルトで係数 (B) を表示する `logit` で，もう一つはデフォルトでオッズ比を表示する `logistic` です．両コマンドともに Statistics ▷ Binary outcomes（統計 (S) ▷ アウトカム (二値)）から選択できます．それぞれ，コマンドと結果は以下です．

```
. logit severity liberal female
Iteration 0:   log likelihood = -331.35938
Iteration 1:   log likelihood =  -217.1336
Iteration 2:   log likelihood = -216.94677
Iteration 3:   log likelihood = -216.94661
Iteration 4:   log likelihood = -216.94661
Logistic regression                               Number of obs     =        480
                                                  LR chi2(2)        =     228.83
                                                  Prob > chi2       =     0.0000
Log likelihood = -216.94661                       Pseudo R2         =     0.3453

------------------------------------------------------------------------------
    severity |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     liberal |   1.055704    .090975    11.60   0.000     .8773958    1.234011
      female |   .6526588   .2423695     2.69   0.007     .1776233    1.127694
       _cons |  -3.547764   .3393913   -10.45   0.000    -4.212959   -2.882569
------------------------------------------------------------------------------

. logistic severity liberal female
Logistic regression                               Number of obs     =        480
                                                  LR chi2(2)        =     228.83
                                                  Prob > chi2       =     0.0000
Log likelihood = -216.94661                       Pseudo R2         =     0.3453

------------------------------------------------------------------------------
    severity | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     liberal |   2.873997   .2614619    11.60   0.000     2.404629    3.434981
      female |   1.920641   .4655047     2.69   0.007     1.194375    3.088527
       _cons |   .0287889   .0097707   -10.45   0.000     .0148025    .0559907
------------------------------------------------------------------------------
```

ここではオッズ比を表示する `logistic`（ロジスティック回帰）に焦点を当てます．懲役刑を厳しすぎると考えるかどうか，リベラル度と性別を要因としたロジスティック回帰の結果は，$\chi^2(2) = 228.83, p < .001$ で有意でした．リベラル度については $z = 11.60, p < .001$，性別については $z = 2.69, p < .01$ となり，共に有意でした．

5 つの水準で区分けしたリベラル度が 1 だけ増加すると，懲役刑を厳しすぎると考えるオッズが 2.87 倍に増加します．最も保守的な人，つまりリベラル度が 1 の人と，最もリベラルな人，つまりリベラル度が 5 の人とを比較すると，リベラル度が 4 だけ離れているので，懲役刑を厳しすぎると考えるオッズは，$2.87^4 = 194.72$ 倍になります．

性別についてもその効果に有意性が示されています．女性が懲役刑を厳しすぎると考えるオッズは，男性と比較して 1.92 倍です．

演習の発展：Stata は，リベラル度や性別といった各予測変数の有意性を検定するのに Wald 検定を用いています．Stata は，Wald 検定の検定量を z スコアとして報告します．Wald 検定量は χ^2 検定量の近似値です．これに対して，Wald 検定より良いとされる尤度比を用いて検定することもできます．尤度比を用いる方法では，まずはじめにリベラル度だけを予測変数としてフィッティングを行い，その後性別を加えてフィッティングを行います．さらに性別だけを予測変数としてフィッティングを行い，その後リベラル度を加えてフィッティングを行います．こうした手続きは，予測変数が多いと煩雑になります．そこで，本体書籍ではユーザ作成コマンド `lrdrop1`（1 回ごとに変数を 1 つ除外する likelihood test）に言及しています．演習の発展問題として，`logistic` と `logit` の後に，このコマンドを実施してみてください．結果は，以下です．

```
. lrdrop1
Likelihood Ratio Tests: drop 1 term
logistic regression
number of obs = 480
----------------------------------------------------------------------
severity     Df      Chi2      P>Chi2    -2*log ll   Res. Df   AIC
----------------------------------------------------------------------
Original Model                            433.89      477     439.89
-liberal      1    198.78      0.0000     632.68      476     636.68
 -female      1      7.24      0.0071     441.13      476     445.13
----------------------------------------------------------------------
Terms dropped one at a time in turn.
```

リベラル度については $\chi^2(1) = 198.78, p < .001$，性別については $\chi^2(1) = 7.24, p < .01$ です．リベラル度の χ^2 値の平方根は 14.10 であり，Wald 検定による z スコアの結果よりも幾分大きな値です．

2. ((訂正) 本演習については本体書籍の記述に誤植がありました．以下のように訂正するとともに，心からお詫び申し上げます．)

   (誤) `2002_chapter11.dta`

   (正) `gss2002_chapter11.dta`

ロジスティック回帰を実行する前に再コード化をする必要があります．次の方法は，そのやり方の一例です．`abort12` の再コード化の際，`recode` コマンドでは新たに作成する変数の変数名 `abort` を指定することができます．`tabulate abort 12 abort, missing` で間違いがないか確

認することができます．以下では，reliten を新たな変数 religious へと再コード化する際，スペースを含むためラベルを二重引用符で囲っています．

準備が整ったので，ロジスティック回帰を実施し，その後 listcoef コマンドを実行します．結果は，以下です．

```
. logistic abort religious conservative premarsxok sei
Logistic regression                               Number of obs     =        409
                                                  LR chi2(4)        =      98.34
                                                  Prob > chi2       =     0.0000
Log likelihood = -226.25386                       Pseudo R2         =     0.1785

------------------------------------------------------------------------------
       abort | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   religious |   .6763732   .0826416    -3.20   0.001     .5323332    .8593877
conservative |    .759082   .0666057    -3.14   0.002     .6391458    .9015242
  premarsxok |     1.8045   .1957157     5.44   0.000     1.458934    2.231917
         sei |  1.020954   .0062494     3.39   0.001     1.008779    1.033277
       _cons |   .3922666   .2986108    -1.23   0.219     .0882285    1.744029
------------------------------------------------------------------------------

. listcoef
logistic (N=409): Factor Change in Odds
  Odds of: yes vs no

------------------------------------------------------------------------------
       abort |      b         z     P>|z|     e^b    e^bStdX      SDofX
-------------+----------------------------------------------------------------
   religious |  -0.39101    -3.200   0.001   0.6764   0.6807     0.9836
conservative |  -0.27565    -3.141   0.002   0.7591   0.6758     1.4216
  premarsxok |   0.59028     5.442   0.000   1.8045   2.1012     1.2579
         sei |   0.02074     3.388   0.001   1.0210   1.4835    19.0180
------------------------------------------------------------------------------
```

人工妊娠中絶の支持は，religious，conservative，premarsxok，sei の 4 つの変数の組に対し $\chi^2(4) = 98.34, p < 0.001$ で，有意に関係しています．Pseudo R2 は 0.18 であり，弱い関係から中程度の関係のあたりを示しています．信仰の強さが 1 単位増加すると，理由を問わず人工中絶を支持するオッズが 0.68 倍に減少します．別の表現を用いると，人工中絶を支持するオッズは $(1.00 - 0.68) \times 100 =$32% 減少します．この減少の効果は $z = -3.20, p < 0.001$ であり，有意です．保守派が強くなるほど，理由を問わず人工中絶を支持する確率が低くなります．保守派の度合いが 1 単位増加したときのオッズ比は，0.76 つまりオッズは 24% 減少します．保守派の度合いが 1 標準偏差分だけ増加したときのオッズは 0.68 倍になります．婚前交渉を問題なしと考える人は，人工中絶を高い確率で支持します．婚前交渉を問題なしとする度合いが 1 単位増加すると，人工中絶を支持するオッズが 1.80 倍になります．$(1.80 - 1.00) \times 100 = 80\%$ より，80% の増加です．社会経済的における高いステータスは，人工中絶に対する高い支持率

に連関しています．社会経済的状況は 0 から 100 までの数値で表されます．社会経済的状況が 1 標準偏差分だけ増加したときの人工中絶を賛成するオッズは 1.48 倍になります．従って，同数値が 1 標準偏差分だけ増加したときの人工中絶を賛成するオッズは 48% 増加します．

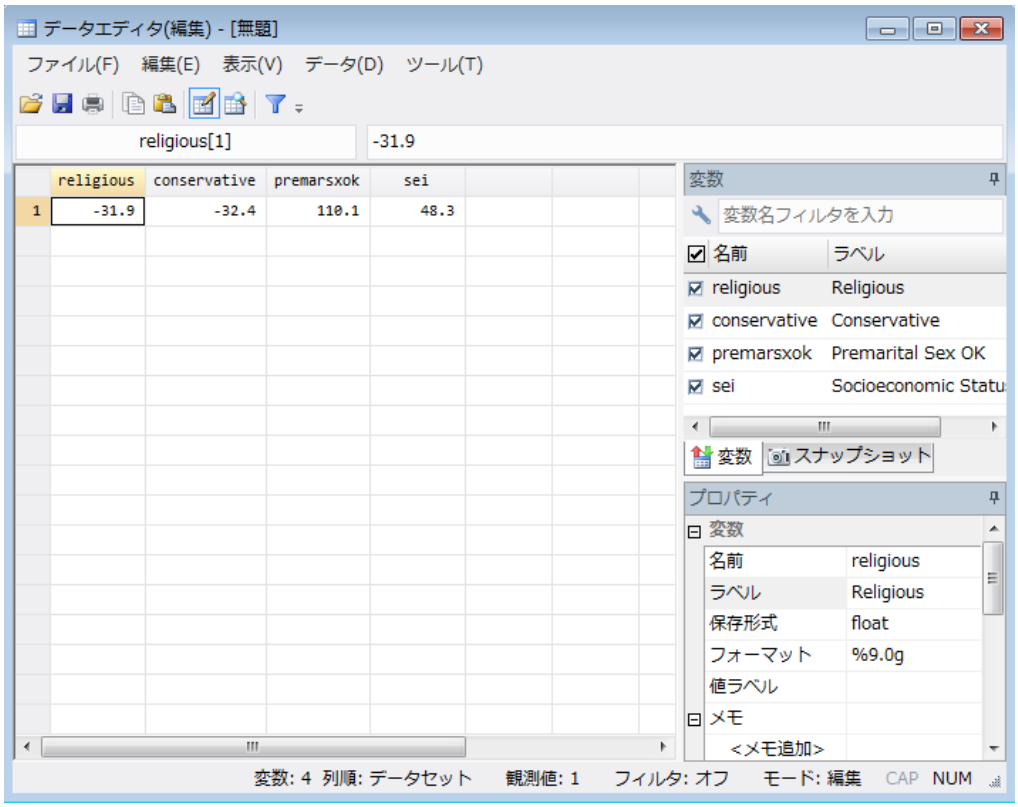
3. 演習 2 と同じコマンドを実行し，その後以下のコマンドを実行します．

```
. listcoef, help percent
logistic (N=409): Percentage Change in Odds
  Odds of: yes vs no
```

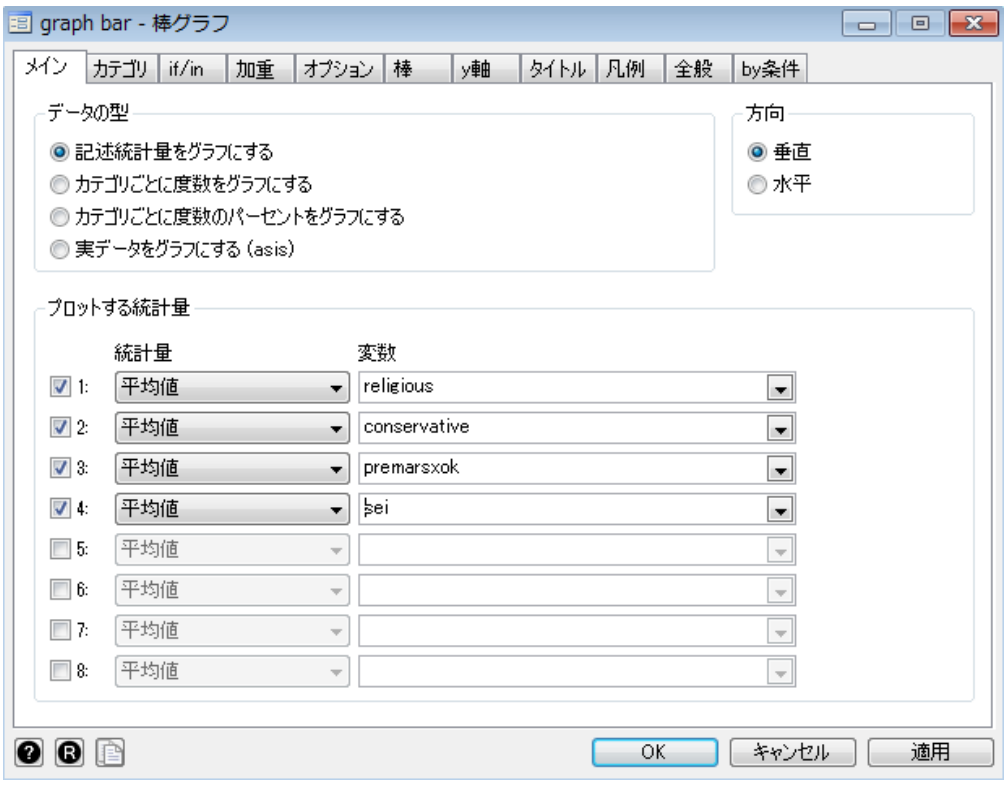
abort	b	z	P>|z|	%	%StdX	SDofX
religious	-0.39101	-3.200	0.001	-32.4	-31.9	0.9836
conservative	-0.27565	-3.141	0.002	-24.1	-32.4	1.4216
premarsxok	0.59028	5.442	0.000	80.5	110.1	1.2579
sei	0.02074	3.388	0.001	2.1	48.3	19.0180

```
       b = raw coefficient
       z = z-score for test of b=0
   P>|z| = p-value for z-test
       % = percent change in odds for unit increase in X
   %StdX = percent change in odds for SD increase in X
   SDofX = standard deviation of X
```

次に，データセットを一新し，各予測変数が 1 標準偏差分だけ増加したときのオッズの増減（パーセントポイント）（%StdX）を収めた 4 つの変数を作成します．今回はありませんが，もし性別を示す gender などの二値変数があった場合，1 単位分だけ増加したときのオッズ増減を用います．データセットは，以下のようになります．

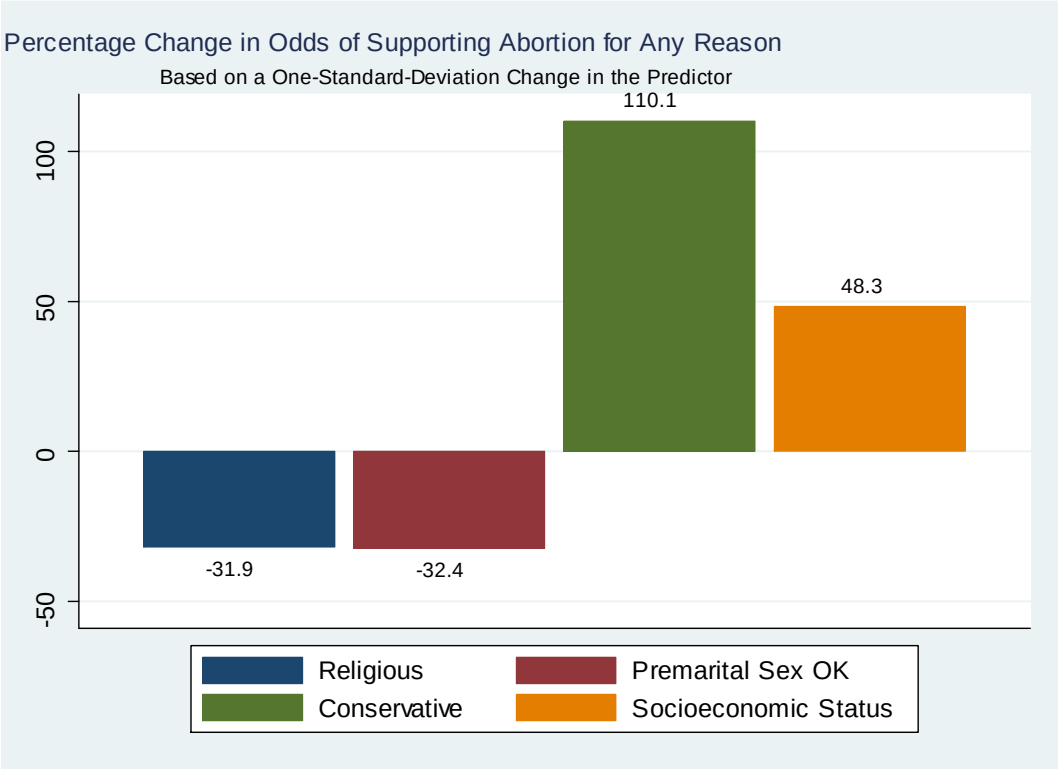


棒グラフのダイアログボックスは多くのソフトウェアでシンプルなものですが，Stata の graph bar はそれに当てはまりません．Stata では数多くのオプションが提供されており，ここで紹介するのはその一部です．Graphics ▷ Bar chart（グラフィックス (G) ▷ 棒グラフ）を選択して，ダイアログボックスを開きます．Main（メイン）タブで，*Graph by calculating summary statistics*（記述統計量をグラフにする）を選択します．各変数は単一の値しか持たないため，平均を用いてグラフが作成できます．値が 1 つしかない場合は，その単一の値が平均になります．左側にある 1 から 4 までのチェックボックスを選択し，*Variables*（変数）ボックスで変数を 1 つずつ指定します．指定後のダイアログボックスは以下のようになります．



グラフの見映えをよくするため，Titles（タイトル）タブでタイトルとサブタイトルを追加しても構いません．Bars（棒）タブで，*Gap between bars*（棒の間隔）に 10（% ）を指定したり，*Bar labels*（棒のラベル）の *Label with bar height*（棒の高さ）を選択したりしても構いません．凡例も工夫すると良いと思います．コマンドとグラフは以下になります．

```
. graph bar (mean) religious (mean) conservative (mean) premarsxok (mean) sei, bargap
> (10) blabel(bar) title("Percentage Change in Odds of Supporting Abortion for Any Re
> ason", size(medsmall)) subtitle("Based on a One-Standard-Deviation Change in the Pr
> edictor", size(small)) legend(order(1 "Religious" 2 "Premarital Sex OK" 3 "Conserva
> tive" 4 "Socioeconomic Status"))
```



上記のグラフにさらに強調を加えても良いかもしれませんが，少なくとも上記からは，政治的に保守派が強いときと宗教的な信仰が深いときに，人工中絶を支持するオッズに同じ程度の低下が見られることが分かります．これに対し，婚前交渉を良しとするときと社会経済的なステータスが高いときに，人工中絶を支持するオッズに上昇が見られます．

もし性別のような二値の予測変数があった場合，1 単位分の増加によるオッズ変化（パーセントポイント）を示したほうが良いでしょう．二値の予測変数では，1 標準偏差分の増加に意味がないためです．グラフには，1 単位分の増加によるオッズ変化と共に，注記として `gender` では 1 単位分の増加，それ以外では 1 標準偏差分の増加によるオッズ変化が示されていることを明記します．

4. 演習 2 でも簡単に触れましたが，もし `lrdrop1` をインストールしていない場合，`findit lrdrop1` を実行し，指示に従ってインストールを行ってください．分析ではロジスティック回帰を行った後，`lrdrop1` を実行します．結果は，以下です．

```
. logistic abort religious conservative premarsxok sei
Logistic regression                               Number of obs     =        409
                                                  LR chi2(4)        =      98.34
                                                  Prob > chi2       =     0.0000
Log likelihood = -226.25386                       Pseudo R2         =     0.1785

------------------------------------------------------------------------------
       abort | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   religious |   .6763732   .0826416    -3.20   0.001     .5323332    .8593877
conservative |    .759082   .0666057    -3.14   0.002     .6391458    .9015242
  premarsxok |     1.8045   .1957157     5.44   0.000     1.458934    2.231917
         sei |  1.020954   .0062494     3.39   0.001     1.008779    1.033277
       _cons |   .3922666   .2986108    -1.23   0.219     .0882285    1.744029
------------------------------------------------------------------------------

. lrdrop1
Likelihood Ratio Tests: drop 1 term
logistic regression
number of obs = 409
---------------------------------------------------------------------------
   abort     Df      Chi2      P>Chi2    -2*log ll   Res. Df   AIC
---------------------------------------------------------------------------
Original Model                            452.51      404     462.51
-religious     1    10.57      0.0012     463.08      403     471.08
-conservative     1    10.15   0.0014     462.66      403     470.66
-premarsxok     1    33.75     0.0000     486.26      403     494.26
    -sei      1     11.81      0.0006     464.32      403     472.32
---------------------------------------------------------------------------
Terms dropped one at a time in turn.
```

（得られた結果は，フォーマットに少し問題があります.）変数名が `lrdrop1` での想定より長かったため，縦の配列が崩れています．表をフォーマットし直してもよいのですが，今の状態でもまだ読めるのでそのままにします.

二つの表を比較するには，上のロジスティック回帰の結果にある z スコアを二乗して Wald カイ二乗値を得るか，下の尤度比検定にあるカイ二乗値の平方根を計算します．この作業は後で行うことにして，二つの表の差異をまとめた表を以下に記載します.

	ロジスティック回帰		尤度比検定からの χ^2 の平方根	
予測変数	z 検定値（Wald）	z 検定値による p 値（Wald）	尤度比検定からの $z = \sqrt{\chi^2}$	尤度比検定からの p 値
信仰の深さ	-3.20	0.001	-3.25	0.001
保守性	-3.14	0.002	-3.19	0.001
婚前交渉	5.44	0.000	5.81	0.000
社会経済的ステータス	3.39	0.001	3/44	0.001

二つの結果は微妙に異なりますが，結論には違いがないといえます．ときには両者に大きな違いが出ますが，その場合には `lrdrop1` による尤度比 χ^2 検定値のほうを採用します．今回は比

較のために χ^2 値の平方根を計算しましたが，通常は必要ありません．通常は，自由度が 1 の χ^2 値として報告します．

5. 懲役刑を厳しすぎると考える人の割合は何パーセントでしょうか．tabulate で表を表示すると答えを得ることができます．

```
. tabulate severity
   Prisons
 sentences
too severe |      Freq.     Percent        Cum.
------------+-----------------------------------
         0 |       222       46.25       46.25
         1 |       258       53.75      100.00
------------+-----------------------------------
     Total |       480      100.00
```

変数 severity には値ラベルがありませんが，一般に高い値ほど程度が甚だしくなるようコード化します．従って，1 が「厳しすぎる」，0 が「そうは思わない」を示すと解釈します．すなわち，53.75% の人が，懲役刑を厳しすぎると考えていることが分かります．

演習での 2 つ目の課題は，liberal と female の 2 つの予測変数に対し，education という 3 つ目を追加するときの検出力分析についての質問です．既に予測変数が 2 つあるので，その 2 つの変数 female と liberal に対して，education が交絡するかを考慮しなければなりません．以前の研究結果から，education が説明する変動のうち 5% が性別とリベラル度によって説明できると予想されます．つまり，education を liberal と female で回帰分析すると，R^2 の値としては約 0.05 が見込まれます．逆に言えば，教育水準の変動の 95% は 2 つの予測変数に対して独立であり，検定力分析ではこの点を調整する必要があります．

分析前に把握しておくことは何でしょう．

p1 は新たな変数なしでの割合です $= 0.5375$

p2 は新たな変数（教育水準）ありでの割合です $= 0.6375$（p1 の値 + 教育水準による増分として重要と考える量）

alpha は有意水準です $= 0.05$

rsq は共線性 R^2 です $= 0.05$（この値は低めの見積もりです）

コマンドと結果は，以下です．

```
. powerlog, p1(0.5375) p2(0.6375) alpha(0.05) rsq(0.05)
 Logistic regression power analysis
 One-tailed test: alpha=.05  p1=.5375  p2=.6375  rsq=.05  odds ratio=1.51323175621491
> 6
```

```
power          n
0.60          95
0.65         109
0.70         123
0.75         141
0.80         161
0.85         186
0.90         221
```

80% の検出力には 161 の標本の大きさが，90% の検出力には 221 の標本の大きさがそれぞれ必要と分かります．教育水準と他の 2 つの予測変数の間の関係はより強いという予想もあるでしょう．教育水準で説明される変動の 20% が，それら 2 つの変数であるリベラル度と性別で説明されるとします．こうするとどの程度の変化があるでしょうか．試してみてください．結果は，80% の検出力で 191 人，90% の検出力で 262 人となります．ここで，検出力分析は分析者の予想により，良くも悪くもなることを記しておきます．予想を変化させると，検出力分析の結果も大きく変わる場合があります．

第 11 章 (11.11 節, pp.375-376) の do-file

演習 11.1

```
/************** Begin do-file ************/
* chapter11.1.do
use "C:\data\severity.dta"
logit severity liberal female
logistic severity liberal female
lrdrop1
/************** End do-file **************/
```

演習 11.2

```
/************** Begin do-file ************/
* chapter11.2.do
use "C:\data\gss2002_chapter11.dta", clear
codebook abort12
recode abort12 (1=1 yes) (2=0 no), generate(abort)
label variable abort "Abortion OK for any reason, 1 yes, 0 no"
tabulate abort12 abort, missing
recode reliten (1=4 strong) (2=3 "not very strong") ///
(3=2 "somewhat strong") (4=1 "no religion"), generate(religious)
clonevar conservative = polviews
clonevar premarsxok = premarsx
logistic abort religious conservative premarsxok sei
listcoef
/************** End do-file **************/
```

演習 11.3

```
/************** Begin do-file ************/
* chapter11.3.do
use "C:\data\gss2002_chapter11.dta", clear
codebook abort12
recode abort12 (1=1 yes) (2=0 no), generate(abort)
label variable abort "Abortion OK for any reason, 1 yes, 0 no"
tabulate abort12 abort, missing
recode reliten (1=4 strong) (2=3 "not very strong") ///
(3=2 "somewhat strong") (4=1 "no religion"), generate(religious)
clonevar conservative = polviews
clonevar premarsxok = premarsx
logistic abort religious conservative premarsxok sei
listcoef, help percent
clear
set obs 1
generate var1 = -31.9 in 1
generate var2 = -32.4 in 1
generate var3 = 110.1 in 1
generate var4 = 48-3 in 1
```

```
rename var1 religious
rename var2 conservative
rename var3 premarsxok
rename var4 sei
label variable religious ‘ "Religious" ’
label variable conservative ‘ "Conservative" ’
label variable premarsxok ‘ "Premarital Sex OK" ’
label variable sei ‘ "Socieconomic Status" ’
graph bar (mean) religious (mean) conservative (mean) premarsxok (mean) sei, ///
bargap(10) blabel(total) title(Percentage Change in Odds of Supporting ///
Abortion for Any Reason, size(medsmall)) subtitle(Based on a ///
One-Standard-Deviation Change in the Predictor, size(small)) ///
legend(label(1 "Religious") label(2 "Premarital Sex OK") ///
label(3 "Conservative") label(4 "Socioeconomic Status"))
/************** End do-file **************/
```

演習 11.4

```
/************** Begin do-file ************/
* chapter11.4.do
use "C:\data\gss2002_chapter11.dta", clear
codebook abort12
recode abort12 (1=1 yes) (2=0 no), generate(abort)
label variable abort "Abortion OK for any reason, 1 yes, 0 no"
tabulate abort12 abort, missing
recode reliten (1=4 strong) (2=3 "not very strong") ///
(3=2 "somewhat strong") (4=1 "no religion"), generate(religious)
clonevar conservative = polviews
clonevar premarsxok = premarsx
logistic abort religious conservative premarsxok sei
lrdrop1
/************** End do-file **************/
```

演習 11.5

```
/************** Begin do-file ************/
* chapter11.5.do
use "C:\data\severity.dta"
tabulate severity
powerlog, p1(.5375) p2(.6375) alpha(.05) rsq(.05)
powerlog, p1(.5375) p2(.6375) alpha(.05) rsq(.20)
/************** End do-file **************/
```

第 12 章 (12.8 節, pp.409-410) の解答

1. ((訂正) 本演習については本体書籍の記述に誤植がありました．以下のように訂正するとともに，心からお詫び申し上げます．)

   (誤) `gss2002 and 2006 chapter12.dta`

   (正) `gss2006_chapter12.dta`

   因子分析のコマンドと結果は，以下です．

```
. factor natspac-natsci, pcf
(obs=2,115)
Factor analysis/correlation                      Number of obs    =      2,115
    Method: principal-component factors          Retained factors =          4
    Rotation: (unrotated)                        Number of params =         50
```

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	2.90147	1.53239	0.2072	0.2072
Factor2	1.36908	0.11888	0.0978	0.3050
Factor3	1.25020	0.20128	0.0893	0.3943
Factor4	1.04892	0.09827	0.0749	0.4693
Factor5	0.95065	0.02654	0.0679	0.5372
Factor6	0.92411	0.11213	0.0660	0.6032
Factor7	0.81198	0.04207	0.0580	0.6612
Factor8	0.76991	0.01165	0.0550	0.7162
Factor9	0.75826	0.03099	0.0542	0.7703
Factor10	0.72727	0.04134	0.0519	0.8223
Factor11	0.68593	0.05432	0.0490	0.8713
Factor12	0.63161	0.03410	0.0451	0.9164
Factor13	0.59751	0.02445	0.0427	0.9591
Factor14	0.57307	.	0.0409	1.0000

```
    LR test: independent vs. saturated:  chi2(91) = 3063.46 Prob>chi2 = 0.0000
Factor loadings (pattern matrix) and unique variances
```

Variable	Factor1	Factor2	Factor3	Factor4	Uniqueness
natspac	-0.0703	0.6774	-0.2377	0.2335	0.4251
natenvir	0.5082	0.1493	-0.2578	-0.2182	0.6053
natheal	0.6136	0.0747	0.0945	-0.3839	0.4616
natcity	0.5299	-0.0667	-0.0065	0.3018	0.6236
natcrime	0.4503	-0.0903	0.5223	0.3675	0.3812
natdrug	0.5493	-0.0363	0.3812	0.3862	0.4025
nateduc	0.5724	0.0687	0.0347	-0.3965	0.5092
natrace	0.5812	-0.1463	-0.3046	0.2398	0.4905
natarms	-0.1305	0.4077	0.5617	-0.0436	0.4994
natfare	0.4617	-0.0522	-0.3989	0.2183	0.5773
natroad	0.1334	0.4402	0.1731	-0.0198	0.7581
natsoc	0.4482	-0.0774	0.2478	-0.3927	0.5775
natchld	0.5591	-0.0566	-0.1242	-0.0483	0.6664

```
      natsci |   0.2224    0.6859   -0.1571    0.0546 |    0.4524
```

上記の結果から，natspac，natarms，natroad，natsci を取り除きます．第 1 主因子に対する負荷量が高いものから 10 項目が残った状態になります．

次に，egen コマンドで前後半の 5 項目ずつの平均を計算します．変数名はそれぞれ firsthalf と secondhalf とします．最後に，作成した両変数の相関を計算します．結果は，以下です．

```
. * exclude natspac natarms natroad natsci as not loading on first Windows
. egen firsthalf = rowmean(natenvir-natdrug)
(5904 missing values generated)
. egen secondhalf = rowmean(nateduc natrace natfare natsoc natchld)
(1717 missing values generated)
. correlate firsthalf secondhalf
(obs=4,268)
             | firsth~f second~f
-------------+------------------
   firsthalf |   1.0000
  secondhalf |   0.4868   1.0000
```

折半法による相関は $r = 0.49$ でした．この相関は 10 項目でなく 5 項目で計算されたため，信頼性が低くなっています．5 項目と 10 項目では，10 項目のほうが計算される相関の信頼性がより高くなります．

2. 信頼性係数 α を計算するには，Statistics ▷ Multivariate analysis ▷ Cronbach's alpha（統計 (S) ▷ 多変量解析 ▷ クロンバックのアルファ）を選択します．Main（メイン）タブで，10 項目を指定します．Options（オプション）タブで，*Take sign of each item as is*（各項目の符号をそのまま利用する），*Delete cases with missing values*（欠損値のある場合は行ごと削除する），*List individual interitem correlations and covariances*（項目間の相関と共分散を表示する），*Display item-test and item-rest correlations*（I-T 相関及び I-R 相関を表示する），および *Standardized items in the scale to mean 0, variance 1*（項目を平均 0，分散 1 に標準化する）を選択します．相関行列を得るため，得点は標準化しましたが，報告時には，標準化のチェックを外して，標準化前の α を掲載します．結果は，以下です．

```
. alpha natenvir-natdrug nateduc natrace natfare natsoc natchld, asis casewise detail
> item std
Test scale = mean(standardized items)
                                                          average
                                  item-test     item-rest   interitem
Item          |  Obs  Sign   correlation   correlation   correlation     alpha
--------------+--------------------------------------------------------------
natenvir      | 2217    +       0.5031        0.3427        0.2083       0.7031
natheal       | 2217    +       0.5928        0.4493        0.1951       0.6857
```

natcity	2217	+	0.5345	0.3796	0.2037	0.6972
natcrime	2217	+	0.4831	0.3195	0.2113	0.7069
natdrug	2217	+	0.5624	0.4127	0.1996	0.6917
nateduc	2217	+	0.5638	0.4143	0.1994	0.6915
natrace	2217	+	0.5734	0.4258	0.1980	0.6896
natfare	2217	+	0.4851	0.3219	0.2110	0.7065
natsoc	2217	+	0.4735	0.3086	0.2127	0.7086
natchld	2217	+	0.5538	0.4024	0.2009	0.6934
Test scale					0.2040	0.7193

(output omitted)

信頼性係数 α は 0.72 です．理想的には 0.80 を超えるほしいところですが，少なくとも適格に達しているといえます．$\alpha = 0.72$ は，折半法による相関すべてについて，項目の除去による減衰の影響を調整して平均化した値と考えることができます．別の言い方をすれば，ここでの尺度が有効であるならば，その尺度での変動の 72% は真の変動を表し，28% は残差を表すと考えられます．item-test correlation はそれぞれの項目と全 10 項目との相関を示します．ここでの項目はすべて，全項目に対し中程度の相関を示しています．item-test correlation では，得点に自身との相関が含まれ，それが相関係数 r を擬似的に大きくしてしまう点に問題があります．一方，item-rest correlation はそれぞれの項目と自身以外の 9 項目との相関を示します．通常，item-test correlation より小さくなります．ここでの結果は，共に中程度でした．

結果で最終列にある alpha は，各項目を取り除いたときの α を示します．ここではどの項目についても取り除くと α が下がるため，尺度の信頼性の観点からどの項目も取り除く理由が見当たりません．分析はこの後，コマンドから `std` を除き，今度は標準化なしの分析結果を表示させます．標準化なしでも，α は 0.72 から変わりません．

3. 結果は，以下です．

```
. alpha natspac-natsci, asis casewise detail item std
Test scale = mean(standardized items)
```

Item	Obs	Sign	item-test correlation	item-rest correlation	average interitem correlation	alpha
natspac	2115	+	0.1776	0.0070	0.1296	0.6594
natenvir	2115	+	0.4781	0.3303	0.1071	0.6092
natheal	2115	+	0.5600	0.4253	0.1009	0.5934
natcity	2115	+	0.4833	0.3363	0.1067	0.6082
natcrime	2115	+	0.4406	0.2879	0.1099	0.6161
natdrug	2115	+	0.5224	0.3813	0.1037	0.6008
nateduc	2115	+	0.5261	0.3855	0.1035	0.6000
natrace	2115	+	0.5013	0.3569	0.1053	0.6048
natarms	2115	+	0.1253	-0.0457	0.1336	0.6671
natfare	2115	+	0.4358	0.2825	0.1102	0.6170

natroad	2115	+	0.2952	0.1291	0.1208	0.6411
natsoc	2115	+	0.4337	0.2801	0.1104	0.6174
natchld	2115	+	0.4977	0.3528	0.1056	0.6055
natsci	2115	+	0.3798	0.2205	0.1144	0.6269
Test scale					0.1116	0.6374

(output omitted)

```
. * drop natarms
. alpha natspac-natrace natfare-natsci, asis casewise detail item std
Test scale = mean(standardized items)
```

Item	Obs	Sign	item-test correlation	item-rest correlation	average interitem correlation	alpha
natspac	2122	+	0.1652	-0.0070	0.1581	0.6926
natenvir	2122	+	0.5029	0.3575	0.1284	0.6386
natheal	2122	+	0.5661	0.4313	0.1228	0.6268
natcity	2122	+	0.5013	0.3556	0.1285	0.6389
natcrime	2122	+	0.4350	0.2803	0.1343	0.6506
natdrug	2122	+	0.5227	0.3804	0.1266	0.6350
nateduc	2122	+	0.5351	0.3949	0.1255	0.6327
natrace	2122	+	0.5270	0.3854	0.1262	0.6342
natfare	2122	+	0.4598	0.3083	0.1322	0.6463
natroad	2122	+	0.2829	0.1149	0.1477	0.6753
natsoc	2122	+	0.4345	0.2797	0.1344	0.6507
natchld	2122	+	0.5133	0.3695	0.1274	0.6367
natsci	2122	+	0.3672	0.2054	0.1403	0.6620
Test scale					0.1333	0.6665

(output omitted)

```
. * drop natspac
. alpha natenvir-natrace natfare-natsci, asis casewise detail item std
Test scale = mean(standardized items)
```

Item	Obs	Sign	item-test correlation	item-rest correlation	average interitem correlation	alpha
natenvir	2163	+	0.5008	0.3534	0.1574	0.6727
natheal	2163	+	0.5841	0.4510	0.1487	0.6578
natcity	2163	+	0.5177	0.3729	0.1557	0.6698
natcrime	2163	+	0.4599	0.3066	0.1617	0.6797
natdrug	2163	+	0.5379	0.3964	0.1536	0.6662
nateduc	2163	+	0.5497	0.4103	0.1523	0.6641
natrace	2163	+	0.5446	0.4043	0.1529	0.6650
natfare	2163	+	0.4675	0.3152	0.1609	0.6784
natroad	2163	+	0.2742	0.1038	0.1811	0.7087
natsoc	2163	+	0.4523	0.2980	0.1625	0.6810
natchld	2163	+	0.5286	0.3856	0.1545	0.6678
natsci	2163	+	0.3342	0.1677	0.1749	0.6998

Test scale	0.1597	0.6952

(output omitted)

```
. * drop natroad
. alpha natenvir-natrace natfare natsoc-natsci, asis casewise detail item std
Test scale = mean(standardized items)
```

Item	Obs	Sign	item-test correlation	item-rest correlation	average interitem correlation	alpha
natenvir	2179	+	0.5096	0.3580	0.1811	0.6886
natheal	2179	+	0.5896	0.4525	0.1712	0.6738
natcity	2179	+	0.5237	0.3745	0.1793	0.6860
natcrime	2179	+	0.4670	0.3090	0.1863	0.6961
natdrug	2179	+	0.5489	0.4040	0.1762	0.6814
nateduc	2179	+	0.5582	0.4151	0.1751	0.6797
natrace	2179	+	0.5620	0.4196	0.1746	0.6790
natfare	2179	+	0.4750	0.3182	0.1853	0.6947
natsoc	2179	+	0.4569	0.2975	0.1876	0.6978
natchld	2179	+	0.5502	0.4056	0.1760	0.6812
natsci	2179	+	0.3238	0.1505	0.2041	0.7194
Test scale					0.1815	0.7093

(output omitted)

```
. * drop natsci
. alpha natenvir-natrace natfare natsoc-natchld, asis casewise detail item std
Test scale = mean(standardized items)
```

Item	Obs	Sign	item-test correlation	item-rest correlation	average interitem correlation	alpha
natenvir	2217	+	0.5031	0.3427	0.2083	0.7031
natheal	2217	+	0.5928	0.4493	0.1951	0.6857
natcity	2217	+	0.5345	0.3796	0.2037	0.6972
natcrime	2217	+	0.4831	0.3195	0.2113	0.7069
natdrug	2217	+	0.5624	0.4127	0.1996	0.6917
nateduc	2217	+	0.5638	0.4143	0.1994	0.6915
natrace	2217	+	0.5734	0.4258	0.1980	0.6896
natfare	2217	+	0.4851	0.3219	0.2110	0.7065
natsoc	2217	+	0.4735	0.3086	0.2127	0.7086
natchld	2217	+	0.5538	0.4024	0.2009	0.6934
Test scale					0.2040	0.7193

(output omitted)

はじめに，14 項目について尺度を計算し，$\alpha = 0.64$ を得ます．natarms を取り除いたときの

α は 0.67 になるので，これを分析から取り除きます．実際に，natarms を取り除いたときの α は，先ほどとわずかに異なっていますが，これは（natarms にある欠損値の影響により）標本の大きさが異なったためです．結果を辿ると，遂には先ほどの演習と同様に項目数 10,（natarms にある欠損値の影響のため）α が 0.7193 という結果に行き着きます．今回のようにすんなり進むケースは稀であり，通常は，どの項目を取り除いていくと α が最大になるかは未知数です．

4. ((訂正) 本演習については本体書籍の記述に誤植がありました．以下のように訂正するとともに，心からお詫び申し上げます．)

   (誤) gss2002 and 2006 chapter12.dta

   (正) gss2006_chapter12.dta

結果は，以下です．

```
. factor natenvir-natdrug nateduc natrace natfare natsoc natchld, pcf
(obs=2,217)
Factor analysis/correlation                      Number of obs    =      2,217
    Method: principal-component factors          Retained factors =          3
    Rotation: (unrotated)                        Number of params =         27
```

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	2.85830	1.75449	0.2858	0.2858
Factor2	1.10382	0.06585	0.1104	0.3962
Factor3	1.03796	0.13622	0.1038	0.5000
Factor4	0.90174	0.11751	0.0902	0.5902
Factor5	0.78423	0.00606	0.0784	0.6686
Factor6	0.77817	0.07783	0.0778	0.7464
Factor7	0.70034	0.05415	0.0700	0.8165
Factor8	0.64618	0.03879	0.0646	0.8811
Factor9	0.60739	0.02552	0.0607	0.9418
Factor10	0.58188	.	0.0582	1.0000

```
    LR test: independent vs. saturated:  chi2(45) = 2752.74 Prob>chi2 = 0.0000
Factor loadings (pattern matrix) and unique variances
```

Variable	Factor1	Factor2	Factor3	Uniqueness
natenvir	0.4950	-0.2354	-0.2739	0.6246
natheal	0.6111	-0.0394	-0.3951	0.4690
natcity	0.5368	0.0511	0.2959	0.6217
natcrime	0.4639	0.6690	0.1591	0.3120
natdrug	0.5665	0.5137	0.2187	0.3673
nateduc	0.5770	-0.0629	-0.4082	0.4965
natrace	0.5855	-0.3161	0.3948	0.4014
natfare	0.4678	-0.4367	0.4080	0.4240
natsoc	0.4510	0.0829	-0.3959	0.6330
natchld	0.5635	-0.1772	0.0221	0.6506

第 1 因子に対して，負荷量が少ないものを取り除き，負荷量の大きな項目のみで再分析すると，通常得られる結果は，第 1 因子のみが必要となります．このことは，第 1 因子の固有値のみが極めて大きくなり，その他の因子は 1.00 未満あるいはその付近に抑えられるという形で表れます．今回は，第 1 因子の固有値が 2.86 と極めて大きな値であり，全 10 項目の変動の 28% を説明できることが示される一方，固有値が 1.00 をわずかに超えた因子が他に 2 つありました．報告では，10 項目から固有値 2.86 の第 1 因子が構成され，第 2 因子は固有値 1.00 をわずかに超えるのみというところでしょう．全項目は第 1 因子に対して 0.43 を超える負荷量を持ちます．

5. 因子得点の計算は `alpha` で行えます．Statistics ▷ Multivariate analysis ▷ Cronbach's alpha（統計 (S) ▷ 多変量解析 ▷ クロンバックのアルファ）を選択します．Main（メイン）タブで，10 項目を指定します．Options（オプション）タブで，*Take sign of each item as is*（各項目の符号をそのまま利用する），*Delete cases with missing values*（欠損値のある場合は行ごと削除する），*List individual interitem correlations and covariances*（項目間の相関と共分散を表示する），*Save the generated scale in variable*（算出した尺度を変数として保存する），*Display item-test and item-rest correlations*（I-T 相関及び I-R 相関を表示する），および *Standardized items in the scale to mean 0, variance 1*（項目を平均 0，分散 1 に標準化する）を選択します．*Save the generated scale in variable*（算出した尺度を変数として保存する）の下で，新たな変数の変数名 `factorscore` を指定します．

10 項目の平均得点の計算は，以下の `egen` コマンドで行えます．結果は，以下です．

```
. alpha natenvir-natdrug nateduc natrace natfare natsoc natchld, asis casewise detail
>  generate(factorscore) item std
Test scale = mean(standardized items)
```

Item	Obs	Sign	item-test correlation	item-rest correlation	average interitem correlation	alpha
natenvir	2217	+	0.5031	0.3427	0.2083	0.7031
natheal	2217	+	0.5928	0.4493	0.1951	0.6857
natcity	2217	+	0.5345	0.3796	0.2037	0.6972
natcrime	2217	+	0.4831	0.3195	0.2113	0.7069
natdrug	2217	+	0.5624	0.4127	0.1996	0.6917
nateduc	2217	+	0.5638	0.4143	0.1994	0.6915
natrace	2217	+	0.5734	0.4258	0.1980	0.6896
natfare	2217	+	0.4851	0.3219	0.2110	0.7065
natsoc	2217	+	0.4735	0.3086	0.2127	0.7086
natchld	2217	+	0.5538	0.4024	0.2009	0.6934
Test scale					0.2040	0.7193

(output omitted)

```
. egen meanscore = rowmean(natenvir-natdrug nateduc natrace natfare natsoc natchld)
(1710 missing values generated)
. summarize factorscore meanscore
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
 factorscore |      2,217    6.39e-17    .5325336  -.8350667   2.328268
   meanscore |      8,469    1.532534     .463135          1          3
. correlate factorscore meanscore
(obs=2,217)
             | factor~e meansc~e
-------------+------------------
 factorscore |   1.0000
   meanscore |   0.9979   1.0000
```

`meanscore` に収められた平均得点は平均 1.42，標準偏差 0.46 です．要約表には最小値 1，最大値 3 とあります．`factorscore` に収められた因子得点は平均 −1.98e-09，覚えていると思いますが，標準偏差 0.53 です．Stata では小数表示ができないとき，指数表示が用いられます．ここでも，小数表示は，小数点を左に 9 桁ずらした 0.00000000198 となります．因子平均では，尺度の設定の仕方から，常に平均がほぼゼロになります．

因子得点と平均得点は相関が高いので，項目間の負荷量が類似している限り，わざわざ相関係数を計算することは稀です．全項目が同じ負荷量を持つとき，相関係数は 1.00 になります．今回は 0.9979 です．

6. ((訂正) 本演習については本体書籍の記述に誤植がありました．以下のように訂正するとともに，心からお詫び申し上げます．)

   (誤) `gss2002 and 2006 chapter12.dta`

   (正) `gss2006_chapter12.dta`

結果は，以下です．

```
. alpha natenvir-natdrug nateduc natrace natfare natsoc natchld, asis casewise detail
>  generate(factorscore10) item std
Test scale = mean(standardized items)
                                                          average
                                  item-test     item-rest   interitem
Item          |  Obs  Sign   correlation   correlation   correlation     alpha
--------------+--------------------------------------------------------------
natenvir      | 2217    +       0.5031        0.3427        0.2083       0.7031
natheal       | 2217    +       0.5928        0.4493        0.1951       0.6857
natcity       | 2217    +       0.5345        0.3796        0.2037       0.6972
natcrime      | 2217    +       0.4831        0.3195        0.2113       0.7069
```

```
natdrug     | 2217   +     0.5624      0.4127        0.1996     0.6917
nateduc     | 2217   +     0.5638      0.4143        0.1994     0.6915
natrace     | 2217   +     0.5734      0.4258        0.1980     0.6896
natfare     | 2217   +     0.4851      0.3219        0.2110     0.7065
natsoc      | 2217   +     0.4735      0.3086        0.2127     0.7086
natchld     | 2217   +     0.5538      0.4024        0.2009     0.6934
------------+-------------------------------------------------------
Test scale  |                                        0.2040     0.7193
------------------------------------------------------------------------
```

(output omitted)

```
. alpha natspac-natsci, asis casewise detail generate(factorscore14) item std
Test scale = mean(standardized items)
                                                     average
                                  item-test     item-rest     interitem
Item          |  Obs  Sign   correlation   correlation   correlation     alpha
--------------+----------------------------------------------------------------
natspac       | 2115    +       0.1776        0.0070        0.1296      0.6594
natenvir      | 2115    +       0.4781        0.3303        0.1071      0.6092
natheal       | 2115    +       0.5600        0.4253        0.1009      0.5934
natcity       | 2115    +       0.4833        0.3363        0.1067      0.6082
natcrime      | 2115    +       0.4406        0.2879        0.1099      0.6161
natdrug       | 2115    +       0.5224        0.3813        0.1037      0.6008
nateduc       | 2115    +       0.5261        0.3855        0.1035      0.6000
natrace       | 2115    +       0.5013        0.3569        0.1053      0.6048
natarms       | 2115    +       0.1253       -0.0457        0.1336      0.6671
natfare       | 2115    +       0.4358        0.2825        0.1102      0.6170
natroad       | 2115    +       0.2952        0.1291        0.1208      0.6411
natsoc        | 2115    +       0.4337        0.2801        0.1104      0.6174
natchld       | 2115    +       0.4977        0.3528        0.1056      0.6055
natsci        | 2115    +       0.3798        0.2205        0.1144      0.6269
--------------+----------------------------------------------------------------
Test scale    |                                            0.1116      0.6374
------------------------------------------------------------------------------
```

(output omitted)

```
. correlate factorscore10 factorscore14
(obs=2,115)
             | facto~10 facto~14
-------------+------------------
factorsco~10 |   1.0000
factorsco~14 |   0.9193   1.0000
```

ひとつ前の演習と同じ方法を用いて，因子得点を収める 2 つの変数 `factorscore10`, `factorscore14` を作成します．14 項目による尺度の信頼度は低いものの，2 変数の相関は 0.92 もあります．追加の項目を持ち信頼度が低い尺度とそうでない尺度で，相関が高くなるのはなぜでしょうか．主な原因は，追加的要素である 4 つの項目が，尺度にとっては必ずしも重要でないことです．同 4 項目は，第 1 因子への負荷量がいずれも低く，`factorscore14` の値への寄与がほとんどありません．14 項目で平均得点を計算すると，10 項目での平均得点との相関は 0.81 に抑えられ

ます．相関が，先ほどの 0.92 より低くなるのは，尺度にフィットするしないに関わらず，各項目の得点を等しく勘定するためです．

第 12 章 (12.8 節, pp.409-410) の do-file

演習 12.1

```
/************** Begin do-file ************/
* chapter12.1.do
use "C:\data\gss2006_chapter12.dta"
factor natspac-natsci, pcf
* exclude natspac natarms natroad natsci as not loading on first Principal
* Component
egen firsthalf = rowmean(natenvir-natdrug)
egen secondhalf = rowmean(nateduc natrace natfare natsoc natchld)
correlate firsthalf secondhalf
/************** End do-file **************/
```

演習 12.2

```
/************** Begin do-file ************/
* chapter12.2.do
use "C:\data\gss2006_chapter12.dta"
alpha natenvir-natdrug nateduc natrace natfare natsoc natchld, ///
asis casewise detail item std
alpha natenvir-natdrug nateduc natrace natfare natsoc natchld, ///
asis casewise detail item label
/************** End do-file **************/
```

演習 12.3

```
/************** Begin do-file ************/
* chapter12.3.do
use "C:\data\gss2006_chapter12.dta"
alpha natspac-natsci, asis casewise detail item std
* drop natarms
alpha natspac-natrace natfare-natsci, asis casewise detail item std
* drop natspac
alpha natenvir-natrace natfare-natsci, asis casewise detail item std
* drop natroad
alpha natenvir-natrace natfare natsoc-natsci, asis casewise detail item std
* drop natsci
alpha natenvir-natrace natfare natsoc-natchld, asis casewise detail item std
/************** End do-file **************/
```

演習 12.4

```
/************** Begin do-file ************/
* chapter12.4.do
use "C:\data\gss2006_chapter12.dta"
```

```
factor natenvir-natdrug nateduc natrace natfare natsoc natchld, pcf
/************** End do-file **************/
```

演習 12.5

```
/************** Begin do-file ************/
*chapter12.5.do
use "C:\data\gss2006_chapter12.dta"
alpha natenvir-natdrug nateduc natrace natfare natsoc natchld, asis ///
casewise detail generate(factorscore) item std
egen meanscore = rowmean(natenvir-natdrug nateduc natrace ///
natfare natsoc natchld)
sum factorscore meanscore
correlate factorscore meanscore
/************** End do-file **************/
```

演習 12.6

```
/************** Begin do-file ************/
* chapter12.6
use "C:\data\gss2006_chapter12.dta"
alpha natenvir-natdrug nateduc natrace natfare natsoc natchld, asis ///
casewise detail generate(factorscore10) item std
alpha natspac- natsci, asis ///
casewise detail generate(factorscore14) item std
correlate factorscore10 factorscore14
egen meanscore10 = rowmean(natenvir-natdrug nateduc natrace ///
natfare natsoc natchld)
egen meanscore14 = rowmean(natspac - natsci)
correlate meanscore10 meanscore14
/************** End do-file **************/
```

第 13 章 (13.7 節, pp.430-431) の解答

1. 本演習のポイントは, 欠損値の生成メカニズムがランダムである点です．一例としては, 回答者ごとに質問がランダムに抽出されるコンピュータ利用面接（computer-assisted interview(CAI)）が挙げられます．ただ, そこでも回答者が回答を拒否して欠損値となったものについては MCAR ではありません.
2. 一例としては，適切な補助変数が存在するような調査が挙げられます．学部生への専攻に対する意識調査で，ある質問に対する未回答があった（10% ）場合でも，もしその質問に関する補助変数があれば，その調査は MAR と考えられます．そうした補助変数は過去の研究で明らかにされているものもあります．たとえば，男性は女性より回答をよくスキップすることが分かっています．なかには，補助変数として明らかでないものもあります．たとえば，有色人種の学部生が回答をスキップしやすいかどうかなどです．質問に未回答だった学部生を 1，そうでない学部生を 0 とするダミー変数を作成すると，評価分析をすることができます．同ダミー変数に対し，人種を説明変数とするロジスティック回帰を実施します．同ダミー変数と人種が有意に関連していれば，人種を補助変数に含めます．質問の得点を良く予測する予測変数も，補助変数に含めると良いです．たとえば，GPA スコアの低い回答者が，全体的に自身の専攻をあまり良く思っていないときには，GPA スコアを補助変数に含めることが推奨されます.
3. 補助変数としては，回答者の得点を良く予測できる変数のほかに，得点を予測できなくとも欠損値を生む回答者を良く予測できる変数も含まれます．研究者のなかには，前者の変数のほうが後者よりも重要とする人もいます．しかし，後者の変数も，多重代入の前提となる MAR の根拠として重要です．未回答者を予測する変数に対してコントロールを行えば，未回答者のばらつきはランダムとみなすことができます.
4. `dftable` オプションを指定すると，予測変数ごとの自由度を表示できます．標本の大きさと比べて値が小さい自由度に対しては，多重回数を増やし自由度を増やすことにより，パラメータ推定の有意性をより多くの情報量で検定することができます．事後推定コマンド `mibeta` を `miopts(vartable)` オプションとともに実行すると，20 回多重の相対効率を表示できます．相対効率が 1.0 に近い値でない場合，多重回数を増やす恩恵がある状態といえます.
5. 多重代入はごまかしではありません．多重代入で得られた値は，補助変数を含む観測値と一貫性を保ったデータです．このため，新たな情報は全く追加されません．唯一の例外として，欠損値を予測する補助変数の追加が，新たな情報の追加に当たるのみです．多重代入を行うと，そのプロセスの不確定さに対して自由度の調整が行われます．実は，多重代入による結果の変

化は，リストワイズ除去による結果の変化が有意なときでさえも有意にならないほど微量です．

6. 多重代入を行うと，内在する不確実性を考慮して自由度が減少します．欠損値が多いほど，自由度の減少幅も大きくなります．また，多重回数が増えるほど，不確実性さが小さくなり，自由度の減少幅は小さくなります．初期の研究では，多重回数が 3〜5 回と少なめであったため，自由度が相当減少しています．
7. 参考のため，多重代入なしでの回帰分析の結果を以下に示します．リストワイズ除去により N が $3,769$ であることに注意してください．

```
. regress env_con educat inc com3 hlthprob epht3, beta

      Source |       SS           df       MS      Number of obs   =     3,769
-------------+----------------------------------   F(5, 3763)      =    320.15
       Model |  647.67794         5  129.535588   Prob > F        =    0.0000
    Residual |  1522.55872     3,763  .404613001   R-squared       =    0.2984
-------------+----------------------------------   Adj R-squared   =    0.2975
       Total |  2170.23666     3,768  .575965144   Root MSE        =    .63609

------------------------------------------------------------------------------
     env_con |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
      educat |  -.0011841    .004077    -0.29   0.772               -.0044584
         inc |  -5.51e-08   3.62e-07    -0.15   0.879               -.0023317
        com3 |   .0503162   .0092717     5.43   0.000                .074352
    hlthprob |  -.2974035   .0248129   -11.99   0.000                -.172927
       epht3 |  -.4020741    .012687   -31.69   0.000               -.4575999
       _cons |   3.726345   .0651735    57.18   0.000                      .
------------------------------------------------------------------------------
```

(a) 答えは `misstable pattern` コマンド，あるいは `misschk` コマンドで得られます．実行するコマンドは以下です．

```
. misstable patterns env_con educat inc com3 hlthprob epht3
. misschk env_con educat inc com3 hlthprob epht3, gen(d_) dummy
```

`misstable pattern` では，回答者の 84% が欠損値なし，すなわち 16% が欠損値ありであることが示されます．`misschk` では，いずれの変数にも欠損値がないケースが 3,769 あり，全体の 83.61% を占めることが示されます．また，欠損値のあるケースのなかで，欠損値が 1 つのものが 637 ケースで最大であることが示されています．

(b) 最初の作業は，未回答を `1`，それ以外を `0` としたダミー変数の作成です．先ほどの演習の `misschk` の結果をよく見ると，`gen(d_)` および `dummy` オプションをしていたのが分かります．このコマンドでは，欠損値を示すダミー変数を作成し，各変数 `env con`，`educat`，`inc`，`com3`，`hlthprob`，`epht3` に対してそれぞれ `d env con`，`d educat`，`d inc`，`d com3`，`d hlthprob`，`d epht3` という変数名を付けています．ロジスティック回帰分析の前に，ダミー変数に対して `summarize`（または `tabulate`）コマンドを実行してみます．もし欠損値が全くない，または

ないに等しい場合，ロジスティック回帰をする必要はありません．こうした場合，そもそも実行自体できません．`d env con` では 1 を持つケースが 2 つしかないので，敢えてこれに対するロジスティック回帰分析を行う必要はありません．実行するロジスティック回帰コマンドは，以下です．

```
. logit d_educat male
. logit d_inc male
. logit d_com3 male
. logit d_hlthprob male
. logit d_epht3 male
```

結果では, `male` で `d_inc` を予測するロジスティック回帰のみが有意になりました（$z = 4.96, p < 0.001$）．これは，性別を補助変数として含めるべきことを示しています．性別は，所得の回答における欠損値の説明に役立つ変数になっています．

(c) 多重代入を行って 20 個のデータセットを作成するコマンドは，以下です．

```
. mi set flong
. mi register imputed env_con educat inc com3 hlthprob epht3 male
. mi impute mvn env_con educat inc com3 hlthprob epht3 male, add(20)
> rseed(222)
```

1 つ目のコマンドは，作成する 20 個のデータセットを，現在のデータセットにロング形式で追加するという指示です．得られるデータセットには，元のデータに続き多重代入で作成された欠損値なしのデータセットのデータが追加されています．2 つ目のコマンドは，多重代入する変数の指定です．今回は補助変数も含めて全ての変数を指定しました．3 つ目のコマンドは，実際に多重代入を行うコマンドで，`mvn` を指定して多変量正規分布を前提とする多重代入を行います．コンマの後にある `add(20)` オプションは，多重代入でデータセットをいくつ作成するかを制御しています．`rseed(222)` オプションは，ランダムプロセスの開始値を指定しています．同オプションは，もし繰り返し同じ結果を得たいときは重要な指定です．このオプションの指定なしでは，Stata は実行のたびに時刻を基にしてランダムプロセスの開始値を算出し，このため実行のたびに異なる結果を生成します．

(d) コマンドと結果は，以下です．

```
. mi estimate, dftable: regress env_con educat inc com3 hlthprob epht3
Multiple-imputation estimates                   Imputations       =         20
Linear regression                               Number of obs     =      4,508
                                                Average RVI       =     0.0344
                                                Largest FMI       =     0.1493
                                                Complete DF       =       4502
DF adjustment:   Small sample                   DF:     min       =     712.71
                                                        avg       =   3,604.21
                                                        max       =   4,455.37
Model F test:       Equal FMI                   F(   5, 4168.8)   =     357.85
```

```
Within VCE type:          OLS                    Prob > F          =     0.0000
```

env_con	Coef.	Std. Err.	t	P>|t|	DF	% Increase Std. Err.
educat	-.0019951	.0038223	-0.52	0.602	3376.4	1.82
inc	-1.02e-07	3.65e-07	-0.28	0.780	712.7	8.30
com3	.0490917	.0085662	5.73	0.000	4302.3	0.60
hlthprob	-.3006827	.0234605	-12.82	0.000	4394.5	0.41
epht3	-.4037722	.011838	-34.11	0.000	4455.4	0.24
_cons	3.74564	.0602982	62.12	0.000	4384.0	0.44

結果をみると，観測数は 4,508 であることに気づきます．つまり，欠損値は 1 つもないときと同様です．所得を示す変数 var の自由度は 712.71 ときわめて小さい値です．他の変数に比べて小さい値となった一つの理由は，多重代入された値が多いことです．

多重代入の後の回帰分析の結果は，リストワイズ除去の伴う回帰分析の結果とかなり似ています．今回のような類似が見られるのは稀です．理解が進んでいないうちは，多重代入をデータの創出により有意性を高めるものと勘違いする場合があります．実際は，リストワイズ除去のときと比較すると，t 値が類似の値，自由度が小さい値になります．

8. 演習 7 では，R^2 や β の表示はありませんでした．mibeta コマンドで表示すると，結果は以下です．

```
. mibeta env_con educat inc com3 hlthprob epht3, fisherz miopts(vartable)
Multiple-imputation estimates                   Imputations       =         20
Linear regression
Variance information
```

	Imputation variance Within	Between	Total	RVI	FMI	Relative efficiency
educat	.000014	4.9e-07	.000015	.036655	.035486	.998229
inc	1.1e-13	1.9e-14	1.3e-13	.172791	.149275	.992592
com3	.000073	8.4e-07	.000073	.0121	.01197	.999402
hlthprob	.000546	4.3e-06	.00055	.008231	.00817	.999592
epht3	.000139	6.3e-07	.00014	.004742	.004722	.999764
_cons	.003604	.00003	.003636	.008729	.008662	.999567

```
Note: FMIs are based on Rubin´s large-sample degrees of freedom.

Multiple-imputation estimates                   Imputations       =         20
Linear regression                               Number of obs     =      4,508
                                                Average RVI       =     0.0344
                                                Largest FMI       =     0.1493
                                                Complete DF       =       4502
DF adjustment:   Small sample                   DF:     min       =     712.71
                                                        avg       =   3,604.21
                                                        max       =   4,455.37
Model F test:       Equal FMI                   F(   5, 4168.8)   =     357.85
```

```
Within VCE type:          OLS                    Prob > F          =     0.0000
```

env_con	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educat	-.0019951	.0038223	-0.52	0.602	-.0094893	.005499
inc	-1.02e-07	3.65e-07	-0.28	0.780	-8.19e-07	6.15e-07
com3	.0490917	.0085662	5.73	0.000	.0322974	.0658859
hlthprob	-.3006827	.0234605	-12.82	0.000	-.3466771	-.2546883
epht3	-.4037722	.011838	-34.11	0.000	-.4269805	-.3805639
_cons	3.74564	.0602982	62.12	0.000	3.627425	3.863855

```
Standardized coefficients and R-squared
Summary statistics over 20 imputations
```

	mean*	min	p25	median	p75	max
educat	-.007505	-.011	-.0100364	-.0073385	-.005578	-.00224
inc	-.004264	-.0155	-.0077944	-.0046998	.0012209	.00617
com3	.0724949	.07	.0717025	.0723296	.0730453	.0753
hlthprob	-.1705718	-.172	-.1716305	-.1705947	-.1698727	-.168
epht3	-.453342	-.455	-.4539943	-.4532183	-.4527826	-.452
R-square	.2926484	.291	.292097	.2926992	.2931932	.294
Adj R-square	.2918628	.29	.2913108	.2919136	.2924082	.293

```
* based on Fisher´s z transformation
```

R^2 は 0.29 で，20 回の多重を通じて 0.291 から 0.294 の幅で推移し，安定しています．R^2 が安定すると，結果の信頼も高くなります．最下段の表で `mean*`とある列が，β 重みを示しています．教育と所得が環境への関心に弱い関連しかないことは興味深い発見です．一方で，健康に問題を抱えることは，環境への関心の度合いと関連しているようです．

第 13 章 (13.7 節, pp.430-431) の do-file

演習 13.7

```
/************** Begin do-file ************/
* chapter13.7.do
use "C:\data\ops2004.dta"
regress env_con educat inc com3 hlthprob epht3, beta
misstable summarize env_con educat inc com3 hlthprob epht3
misstable patterns env_con educat inc com3 hlthprob epht3
misschk env_con educat inc com3 hlthprob epht3, gen(d_) dummy
summarize d_env_con d_educat d_inc d_com3 d_hlthprob d_epht3
/* The summarize shows that there are virtually no missing values
on d_env_con (Mean is .0004). A tabulation shows that there are
only 2 missing observations on this variable in our sample of 4,508.
We will not run this logistic regression (if we try it, it will not
work). There are only a few missing observations on d_educat.
We will run the logistic regression but are reluctant to expect
a strong result for this.
Not run---logit d_env_con male
*/
logit d_educat male
logit d_inc male
logit d_com3 male
logit d_hlthprob male
logit d_epht3 male
/* Here is the multiple imputation analysis */
mi set flong
mi register imputed env_con educat inc com3 hlthprob epht3 male
mi impute mvn env_con educat inc com3 hlthprob epht3 male, add(20) rseed(222)
mi estimate, dftable: regress env_con educat inc com3 hlthprob epht3
mibeta env_con educat inc com3 hlthprob epht3, fisherz miopts(vartable)
/************** End do-file **************/
```

第 14 章 (14.6 節, pp.459-460) の解答

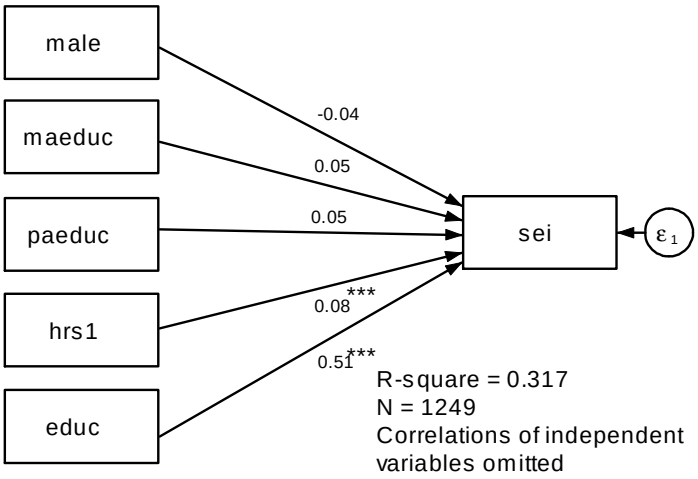
1. （(注）本演習については，解答の原文が本体書籍と大きく隔たり，演習の文章から導き出すのが困難であると判断できます．よって，以下にある本演習の解答は，必ずしも著者の意図通りのものでなく，差し替えミス等による内容の誤りである可能性があります.）
ここではいわゆるアダルトチルドレンの社会経済的ステータス（sei）が関心の対象です．予測変数は，性別（male），母親の教育期間（maeduc），父親の教育期間（paeduc），直近 1 週間の週労働時間（hrs1），自身の教育期間（educ）です．はじめに，gss2002_chapter10.dta を開き，その後 sex を再コード化して male（男性が 1，女性が 0）を作成します.

(a) 再コード化して regress コマンドを実行すると，以下を得ます.

```
. recode sex (1=1 Male) (2=0 Female), gen(male)
(1537 differences between sex and male)

. regress sei male maeduc paeduc hrs1 educ, beta

      Source |       SS           df       MS      Number of obs   =     1,249
-------------+----------------------------------   F(5, 1243)      =    115.55
       Model |  148516.146         5  29703.2293   Prob > F        =    0.0000
    Residual |   319533.62     1,243  257.066468   R-squared       =    0.3173
-------------+----------------------------------   Adj R-squared   =    0.3146
       Total |  468049.766     1,248  375.039877   Root MSE        =    16.033

------------------------------------------------------------------------------
         sei |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
        male |  -1.487118   .9289301    -1.60   0.110               -.0384085
      maeduc |   .2997065   .1897761     1.58   0.115                .0472928
      paeduc |   .2408598   .1619554     1.49   0.137                 .045601
        hrs1 |   .1148798   .0325929     3.52   0.000                .0847621
        educ |   3.752891   .1877156    19.99   0.000                 .511595
       _cons |  -10.71026   2.926604    -3.66   0.000                       .
------------------------------------------------------------------------------
```

結果から，個人の成果はその努力によるところが大きいことが示され，hrs1 は $\beta = 0.08$***，educ は $\beta = 0.51$***でそれぞれ社会経済的ステータス有意な予測変数であることが示されました．一方，父親および母親の教育期間，性別が社会経済的ステータスに直接及ぼす影響には，有意性がありませんでした.

(b) SEM ビルダを使用して作成できるパスモデルの一例は，以下です．図をシンプルにするため，以下では外生予測変数同士の共分散を含めていません.



(c) do-file は以下です.

```
recode sex (1=1 Male) (2=0 Female), gen(male)
sem (male maeduc paeduc hrs1 educ -> sei), standardized
estat eqgof
```

(d) 標準化係数を表示する回帰分析の結果は，以下のように得られます.

```
. use http://www.stata-press.com/data/r13/regsmpl.dta, clear
(NLS Women 14-26 in 1968)

. regress ln_wage age grade not_smsa south ttl_exp tenure, beta

      Source |       SS           df       MS      Number of obs   =    28,091
-------------+----------------------------------   F(6, 28084)     =   2626.73
       Model |  2305.54089         6  384.256816   Prob > F        =    0.0000
    Residual |  4108.32299    28,084   .14628696   R-squared       =    0.3595
-------------+----------------------------------   Adj R-squared   =    0.3593
       Total |  6413.86388    28,090  .228332641   Root MSE        =    .38247

------------------------------------------------------------------------------
     ln_wage |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
         age |  -.0038303   .0005265    -7.28   0.000               -.0536845
       grade |   .0670419   .0010237    65.49   0.000                .3266299
    not_smsa |  -.1637396   .0051791   -31.62   0.000               -.1542952
       south |  -.1135945   .0047533   -23.90   0.000               -.1168974
     ttl_exp |   .0287283   .0009252    31.05   0.000                .2798338
      tenure |   .0195421   .0008321    23.48   0.000                .1533823
       _cons |   .8004553   .0173735    46.07   0.000                       .
------------------------------------------------------------------------------
```

上記の回帰モデルは個人の所得のばらつきの 36% を説明します．予測変数はすべて有意な効

果がありました．米国南部または非都市部に居住していることは，低所得と連関しています．年齢は，他の予測変数をコントロールしたとき，弱いマイナスの効果を持ちます．最も強い効果が見られるのは，自身の教育期間およびこれまでの労働経験です．同一雇用者での勤続年数は，独立なプラスの効果を持ちます．上記の結果は 28,091 の標本から得られたものです．

(e) SEM ビルダを使用した結果は，以下です．

```
. sem (age -> ln_wage, ) (grade -> ln_wage, ) (not_smsa -> ln_wage, ) (south -> ln_wage,) (ttl_ex
> p -> ln_wage, ) (tenure -> ln_wage, ), standardized nocapslatent
(443 observations with missing values excluded)
Endogenous variables
Observed:  ln_wage
Exogenous variables
Observed:  age grade not_smsa south ttl_exp tenure
Fitting target model:
Iteration 0:   log likelihood = -344757.11
Iteration 1:   log likelihood = -344757.11
Structural equation model                       Number of obs     =     28,091
Estimation method  = ml
Log likelihood     = -344757.11
```

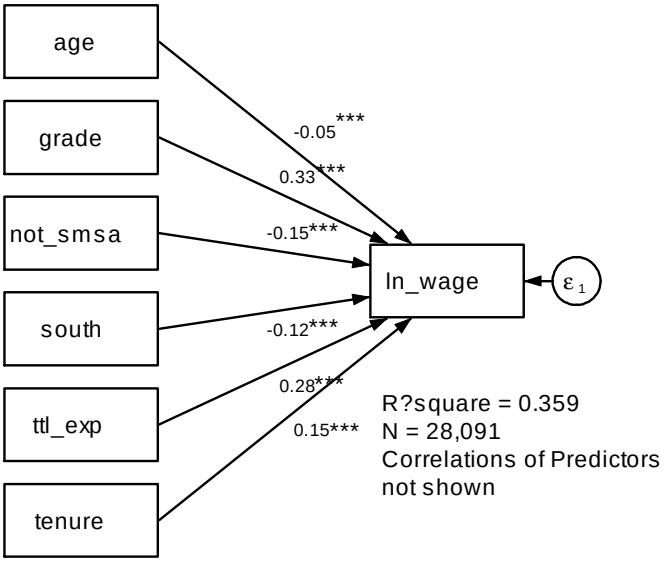
Standardized	Coef.	OIM Std. Err.	z	P>|z|	[95% Conf.	Interval]
Structural						
ln_wage <-						
age	-.0536845	.007372	-7.28	0.000	-.0681334	-.0392355
grade	.3266299	.0046535	70.19	0.000	.3175091	.3357507
not_smsa	-.1542952	.0048058	-32.11	0.000	-.1637143	-.144876
south	-.1168974	.0048487	-24.11	0.000	-.1264006	-.1073941
ttl_exp	.2798338	.0088795	31.51	0.000	.2624303	.2972372
tenure	.1533823	.0064761	23.68	0.000	.1406894	.1660752
_cons	1.675177	.0387981	43.18	0.000	1.599135	1.75122
var(e.ln_wage)	.6405379	.0041505			.6324546	.6487245

```
LR test of model vs. saturated: chi2(0)   =      0.00, Prob > chi2 =      .

. estat eqgof
Equation-level goodness of fit
```

depvars	fitted	Variance predicted	residual	R-squared	mc	mc2
observed						
ln_wage	.2283245	.082074	.1462505	.3594621	.5995516	.3594621
overall				.3594621		

```
mc  = correlation between depvar and its prediction
mc2 = mc^2 is the Bentler-Raykov squared multiple correlation coefficient
```

図は，以下です．



(a) では最小二乗法，(b) では最尤法という違いがあるものの，両者の結果は同じでした．

(f) (b) のモデルを欠損値ありの最尤法で推定しても，ほぼ同じ結果を得ます．両者の違いは，推定に用いた標本の大きさで，ここでは調査に参加した 28,534 人全員分のデータを用いています．ただし，欠損値が無作為に表れることを前提条件にしています．欠損値が無作為に表れるとは，データの欠損が完全に無作為かまたはモデル内の変数で説明できることを意味します．

2. データの読み込み，`abort12` と `reliten` の再コード化，`polviews` の変数名変更を行う do-file は，以下です．

```
clear
use http://www.stata-press.com/data/agis4/gss2002_chapter11
recode abort12 (1=1 Yes) (2=0 No), gen(abortion)
fre abortion reliten polviews premarsx sei
recode reliten (1=4 "Strong") (2=3 "Not Very Strong") (3=2 "Somewhat Strong") ///
(4=1 "No Religion"), gen(religious)
clonevar conservative=polviews
```

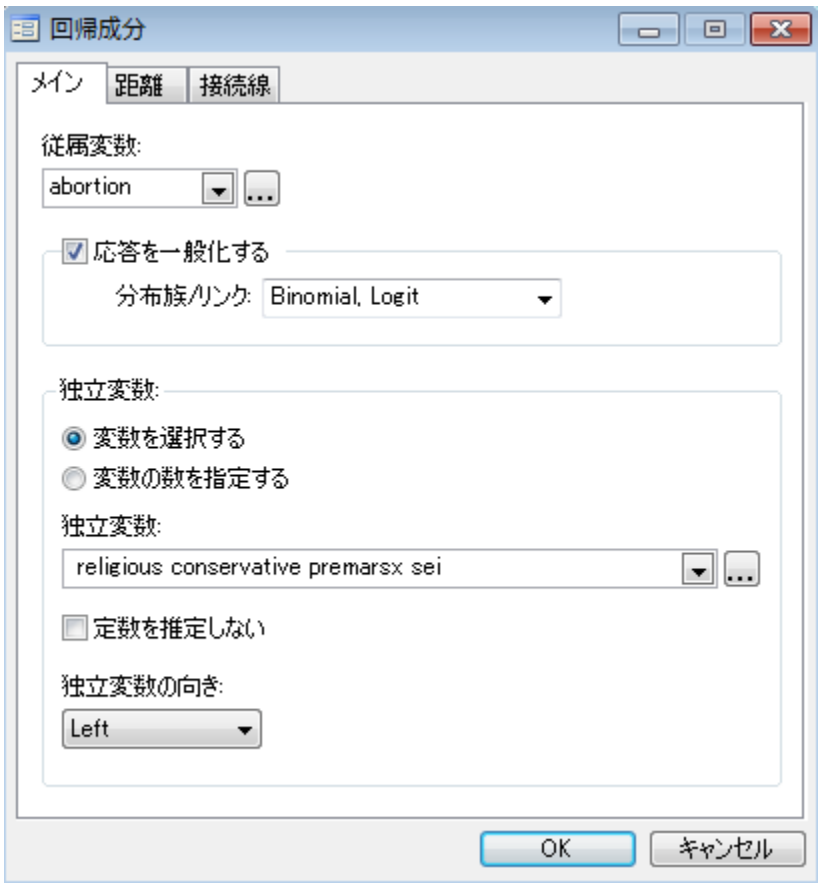
比較のために，以下のコマンドでロジスティック回帰分析を実施します．

```
logit abortion religious conservative premarsx sei
```

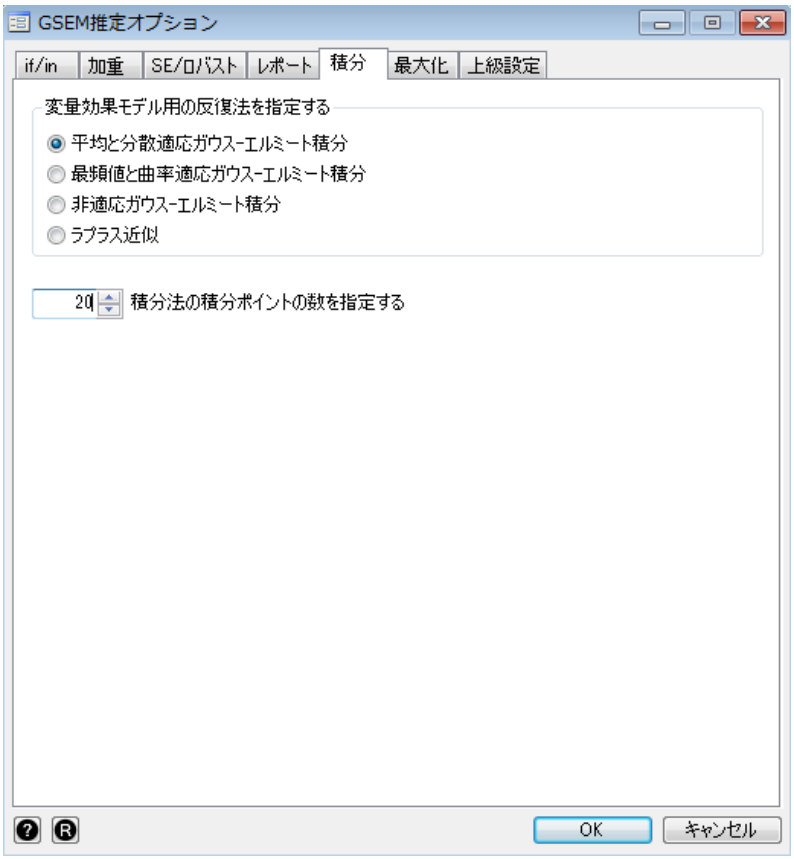
結果は，以下です．

```
. logit abortion religious conservative premarsx sei
Iteration 0:   log likelihood = -275.42314
Iteration 1:   log likelihood = -227.36054
Iteration 2:   log likelihood = -226.25548
Iteration 3:   log likelihood = -226.25386
Iteration 4:   log likelihood = -226.25386
Logistic regression                               Number of obs     =        409
                                                  LR chi2(4)        =      98.34
                                                  Prob > chi2       =     0.0000
Log likelihood = -226.25386                       Pseudo R2         =     0.1785

------------------------------------------------------------------------------
    abortion |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   religious |  -.3910103   .1221835    -3.20   0.001    -.6304856   -.1515351
conservative |  -.2756455    .087745    -3.14   0.002    -.4476226   -.1036684
    premarsx |   .5902836   .1084598     5.44   0.000     .3777063    .8028609
         sei |   .0207377   .0061212     3.39   0.001     .0087404     .032735
       _cons |  -.9358135   .7612444    -1.23   0.219    -2.427825    .5561982
------------------------------------------------------------------------------
```

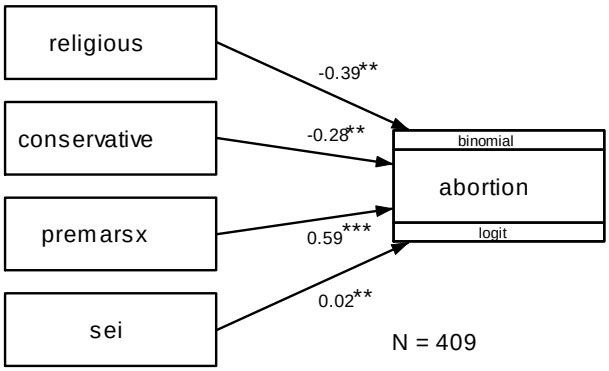
次に，SEM ビルダを開きます．ロジスティック回帰をするには一般化した SEM をする必要があるため `gsem` ボタンをクリックして，ボタンが押された状態にします．回帰要素を追加するボタンをクリックしてダイアログボックスを開き，以下のように情報を入力します．*Make response generalized*（応答を一般化する）を選択し，*Family/Link*（分布族/リンク）で `Binomial, Logit` を選択することに注意してください．



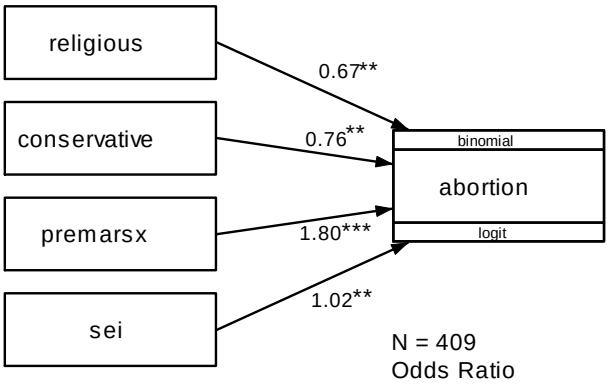
本体書籍では紹介はありませんでしたが，ここでモデルの推定の際には計算上のある自由度が残っています．Stata は特に指定がない場合，精度と速度の観点から推定プロセス内での積分を 7 点で実施しますが，敢えて 20 点の積分へと変更することもできます．ここでは点数を変更しなくても問題ありませんが，フィッティングに問題が出た場合には有用です．


GSEM推定オプション
if/in
加重
SE/ロバスト
レポート
積分
最大化
上級設定
変量効果モデル用の反復法を指定する
平均と分散適応ガウス-エルミート積分
最頻値と曲率適応ガウス-エルミート積分
非適応ガウス-エルミート積分
ラプラス近似
20
積分法の積分ポイントの数を指定する
OK
キャンセル

パス図は，以下です．



上記で接続線に表れている数字は係数（B）であり，場合によってはオッズ比が好まれます．オッズ比を表示するには，Estimation ▷ Other ▷ Report exponentiated estimates（推定 ▷ その他 ▷ 指数形式で係数を表示する）を選択します．`eform` を実行すると，結果で `exp(b)` 列にオッズ比が表れます．SEM ビルダの図にはオッズ比は出ません．図の中に組み込むには，推定結果をクリアして，オッズ比をテキストとして入力する必要があります．結果は，以下です．



ロジスティック回帰結果の解釈は，第 11 章と同様に行います．`sei` が 17.1 から 97.2 までの範囲に分布する連続変数であるため，1 単位分の増減は意味がありそうにありません．一方，`sei` の標準偏差が 19.242 であることから，`sei` のオッズ比を 1 標準偏差分の増加分で議論したいところです．これは，たとえば `display exp(.021*19.242)` を実行して，`1.498` を得て行います．

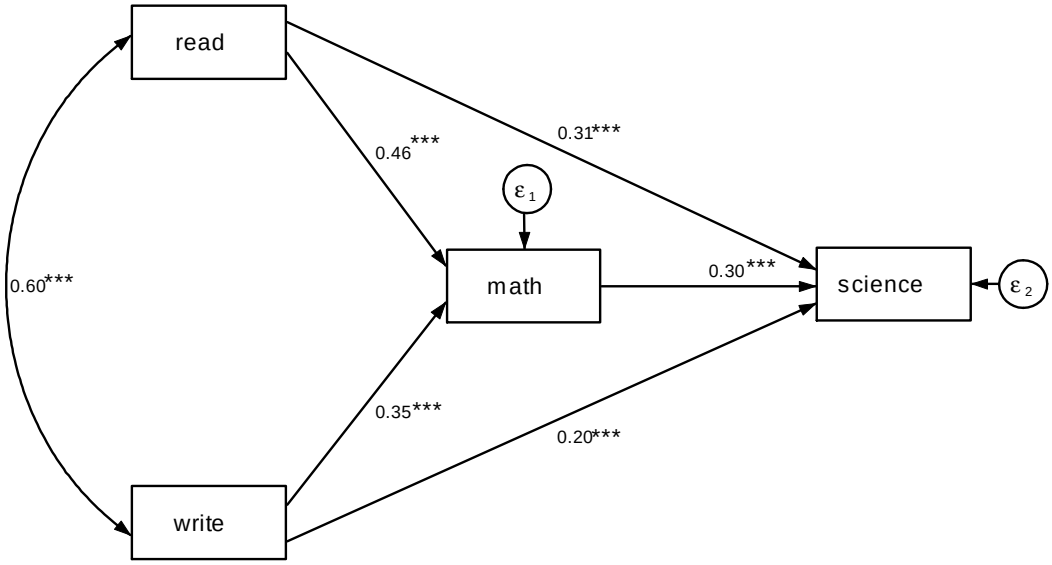
3. ((訂正) 本演習については本体書籍の記述に誤植がありました．以下のように訂正するとともに，心からお詫び申し上げます．

   (誤) http://www.ats.ucla.edu/stat/faq/pathreg.htm

   (正) http://www.ats.ucla.edu/stat/stata/faq/pathreg.htm

   (訂正終わり))

   SEM ビルダ上の図は，以下です．



ここでは，`math` が `read` および `write` の `science` への効果を介在するかどうかを調べるのが課題です．図から，`math` を介在するパスばかりでなく，`read`→`science` と `write`→`science` の両方の効果に有意性が見られることから，`math` は，両効果を完全には介在しなことが分かります．間接効果を得るには，Estimation ▷ Testing and CIs ▷ Direct and indirect effects（推定 ▷ 検定/信頼区間 ▷ 直接/間接効果）を選択し，*Decomposition of effects into total, direct, and indirect (teffects)*（合計，直接，間接へ効果を分解 (teffects)）を選択します．さらに，*Report standardized effects*（標準化済み効果を表示する）を選択し，OK をクリックします．結果は，以下です．

```
. estat teffects, standardized

Direct effects
------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.      z    P>|z|                Std. Coef.
-------------+----------------------------------------------------------------
```

Structural					
math <-					
read	.4169486	.0560586	7.44	0.000	.4563134
write	.3411219	.0606382	5.63	0.000	.3451322
science <-					
math	.3190094	.0759047	4.20	0.000	.301854
read	.3015317	.0679912	4.43	0.000	.3122533
write	.2065257	.0700532	2.95	0.003	.1977167

Indirect effects

	Coef.	OIM Std. Err.	z	P>|z|	Std. Coef.
Structural					
math <-					
read	0	(no path)			0
write	0	(no path)			0
science <-					
math	0	(no path)			0
read	.1330105	.0363514	3.66	0.000	.13774
write	.1088211	.0323207	3.37	0.001	.1041795

Total effects

	Coef.	OIM Std. Err.	z	P>|z|	Std. Coef.
Structural					
math <-					
read	.4169486	.0560586	7.44	0.000	.4563134
write	.3411219	.0606382	5.63	0.000	.3451322
science <-					
math	.3190094	.0759047	4.20	0.000	.301854
read	.4345423	.0627773	6.92	0.000	.4499933
write	.3153468	.0679059	4.64	0.000	.3018962

結果から，`write` から `math` を介在して `science` へ及ぶ効果は 0.104 であり，$z = 3.37, p < 0.01$ で有意であることが分かります．また，`read` から `math` を介在して `science` へ及ぶ効果は 0.104 であり，$z = 3.66, p < 0.001$ で有意であることが分かります．従って，このケースでは直接と間接の両方の効果が存在する部分媒介（partial mediation）が見られていることになります．以下は，結果をまとめた表の一例です．

関係	直接	間接	合計
read→science	0.312***	0.138***	0.450***
write→science	0.198**	0.104***	0.302***
write→math	0.345***		0.345***
read→math	0.456***		0.456***
math→science	0.302***		0.302***

第 14 章 (14.6 節, pp.459-460) の do-file

演習 14.1a

```
/************** Begin do-file ************/
* chapter14.1a.do
use "C:\data\gss2002_chapter10.dta", clear
recode sex (1=1 Male) (2=0 Female), gen(male)
regress sei male maeduc paeduc hrs1 educ, beta
/************** End do-file **************/
```

演習 14.1c

```
/************** Begin do-file ************/
* chapter14.1c.do
use "C:\data\gss2002_chapter10.dta", clear
recode sex (1=1 Male) (2=0 Female), gen(male)
sem (male maeduc paeduc hrs1 educ -> sei) standardized
estat eqgof
/************** End do-file **************/
```

演習 14.2a

```
/************** Begin do-file ************/
* chapter14.2a.do
use http://www.stata-press.com/data/r13/regsmpl.dta, clear
regress ln_wage age grade not_smsa south ttl_exp tenure, beta
/************** End do-file **************/
```

演習 14.2b

```
/************** Begin do-file ************/
* chapter14.2b.do
use http://www.stata-press.com/data/r13/regsmpl.dta, clear
regress ln_wage age grade not_smsa south ttl_exp tenure, beta
sem (age -> ln_wage, ) (grade -> ln_wage, ) (not_smsa -> ln_wage, ) ///
(south -> ln_wage,) (ttl_exp -> ln_wage, ) (tenure -> ln_wage, ), ///
standardized nocapslatent
/************** End do-file **************/
```

演習 14.3

```
/************** Begin do-file ************/
* chapter14.3.do
```

```
clear
use http://www.stata-press.com/data/agis4/gss2002_chapter11
recode abort12 (1=1 Yes) (2=0 No), gen(abortion)
fre abortion reliten polviews premarsx sei
recode reliten (1=4 "Strong") (2=3 "Not Very Strong") (3=2 "Somewhat Strong") ///
 (4=1 "No Religion"), gen(religious)
clonevar conservative=polviews
logit abortion religious conservative premarsx sei
gsem (abortion <- religious conservative premarsx sei), logit
estat eform
/************** End do-file **************/
```

演習 14.4

```
/************** Begin do-file ************/
* chapter14.4.do
use http://www.ats.ucla.edu/stat/data/hsb2, clear
sem (science <- read write math) (math <- read write), standardized
estat teffects, standardized
/************** End do-file **************/
```