

R Commander を用いた統計解析の基礎(2)

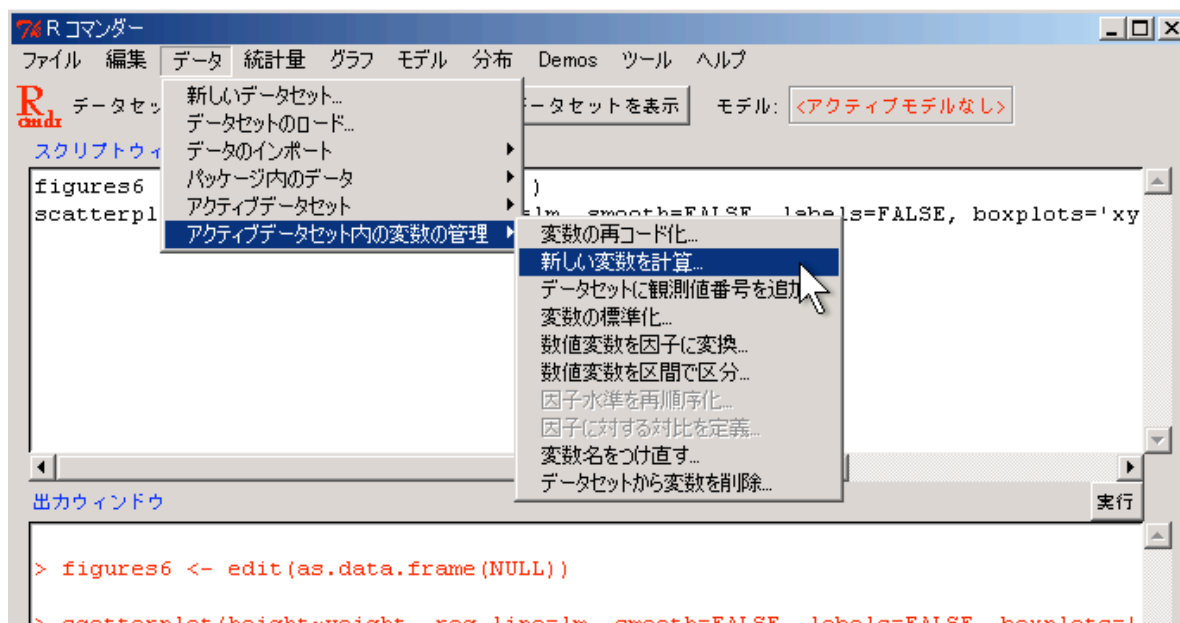
岡田 昌史

1. 変数の操作

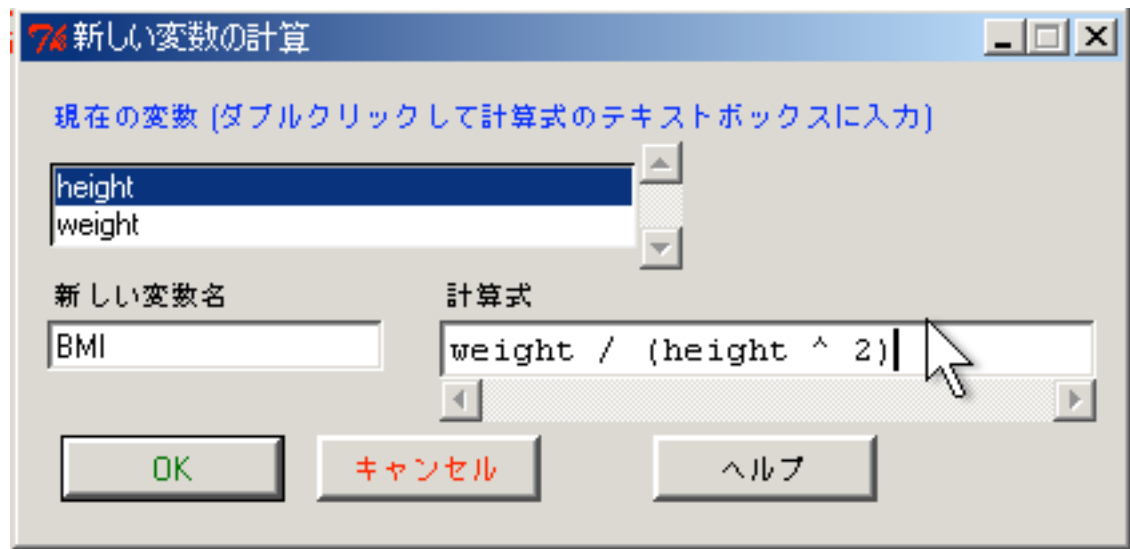
それでは、前回の復習からいってみましょう。まずは身長(height)と体重(weight)から、BMI を求めて、それをデータセットに新しい列として追加してみます。

BMI は、体重 / (身長)² ですから、height と weight を使って書くと、weight / (height ^2) あるいは、weight / (height * height) などと書けますね。R では、かけ算を " * "、割り算を " / "、べき乗を " ^ " であらわします。これは、Excel などと一緒にですね。

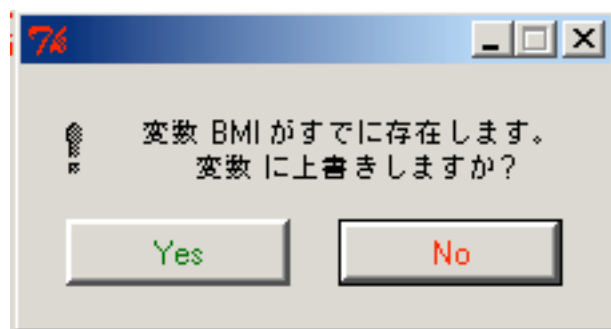
現在、データセット figures6 には、height と weight の2つの列（このように、名前のついた一連のデータを「変数」、正確には「ベクタ変数」と呼びます）しかありませんが、ここに BMI を新しい変数として追加してみましょう。「データ」メニューから、「アクティブデータセット内の変数の管理」を選び、「新しい変数を計算...」をクリックします。



「新しい変数の計算」ウィンドウが出ますから、「新しい変数名」のところに "BMI" , 「計算式」のところに "weight / (height ^ 2) " をいれて、「Ok」を押しましょう。

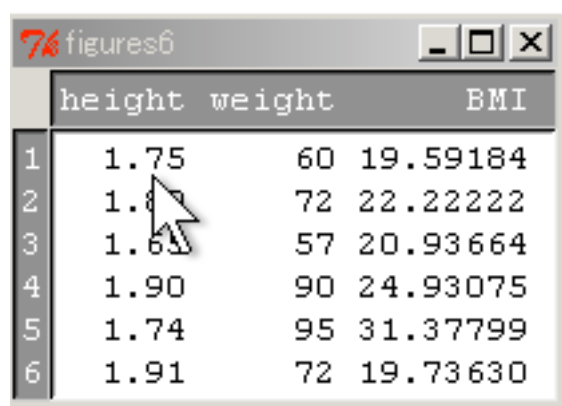


おっと、ここでもし、



というのが出てしまった方は、おそらく前回の演習がうまく進んでおり、すでに "BMI" 変数が作成されているということです。上書きしてもかまいませんので、"Yes" を選んでください。

それでは、うまく変数が追加されているか、「データセットを表示」で確認してみましょう。



	height	weight	BMI
1	1.75	60	19.59184
2	1.8	72	22.22222
3	1.65	57	20.93664
4	1.90	90	24.93075
5	1.74	95	31.37799
6	1.91	72	19.73630

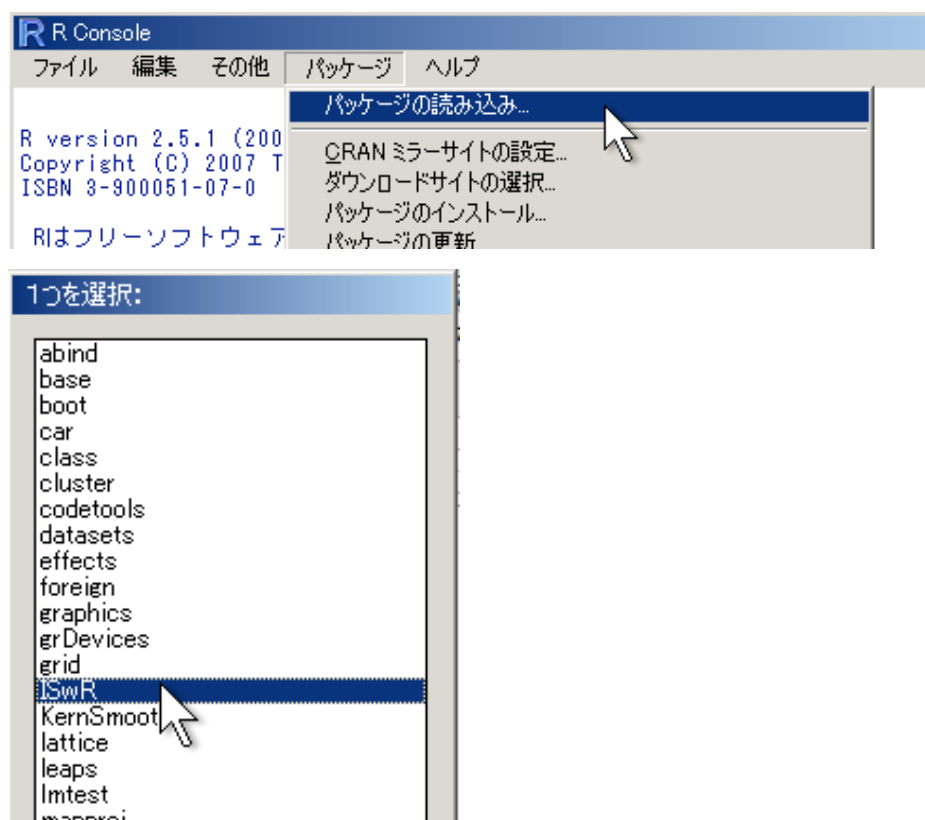
このように、3列目に「BMI」が追加されていれば成功です。

2. データの可視化

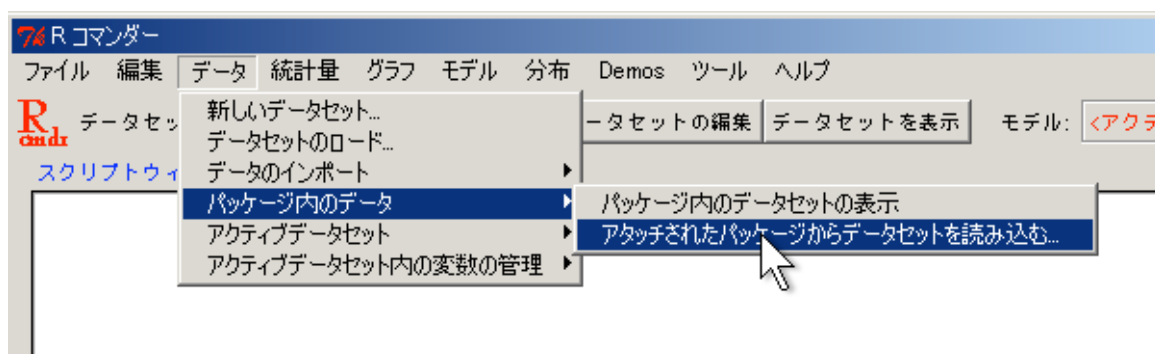
それでは、次に R コマンダーを用いて、データの特徴を示すさまざまな表やグラフを描いてみましょう。あらゆるデータ解析において、まず最初に必要になるのは、データを表やグラフにまとめて、その全体像を把握することです。R は「対話的」な統計解析環境なので、可視化の機能は比較的充実しています。

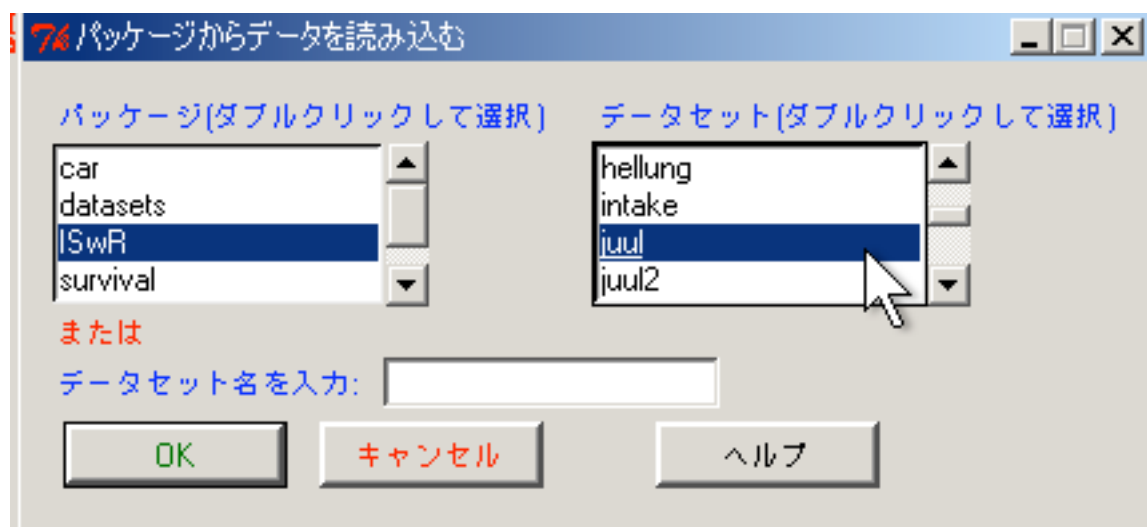
ここでは、さまざまなデータを扱いますが、その1つ1つを毎回 Excel などから読み込んでいては大変なので、書籍「R による医療統計学」(原書 "Introductory Statistics with R")に掲載されているサンプルデータがまとめて収録されているパッケージ、「ISwR」を読み込んで利用します。

まずは、ISwR パッケージを読み込んでみましょう。こんな感じですね？



ISwR パッケージを読み込むと、R コマンダーの「データ」メニュー、「パッケージ内のデータ」から、「アタッチされたパッケージからデータセットを読み込む...」でいろいろなデータを読み込めるようになります。



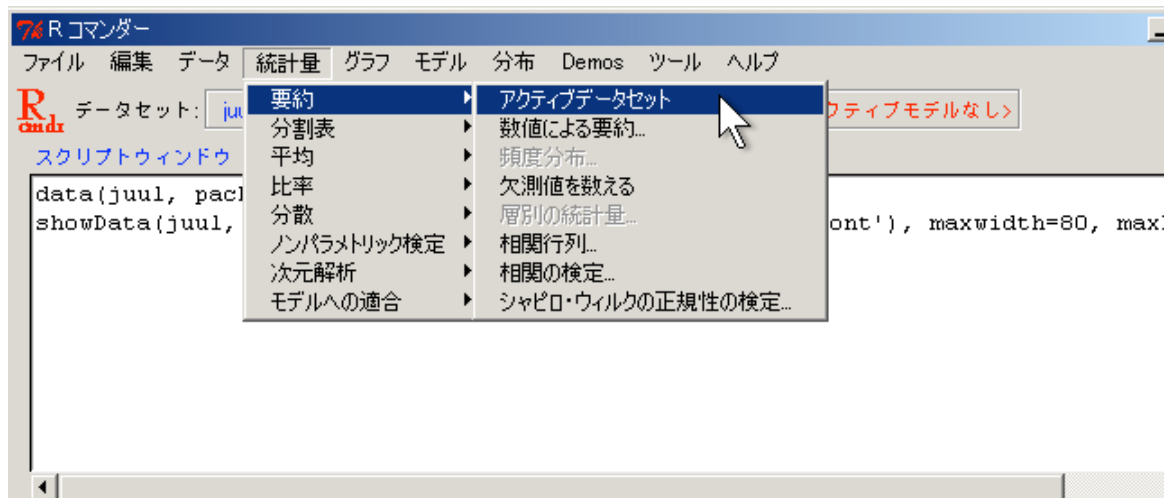


「パッケージからデータを読み込む」ウィンドウが出たら、「パッケージ」のところで "ISwR", 「データセット」では "juul" を選択し、Ok を押してください。これで、ISwR パッケージに内蔵されている juul という名前のデータセットが読み込まれ、アクティブデータセットとなります。

	age	menarche	sex	igf1	tanner	testvol
1	NA	NA	NA	90	NA	NA
2	NA	NA	NA	88	NA	NA
3	NA	NA	NA	164	NA	NA
4	NA	NA	NA	166	NA	NA
5	NA	NA	NA	131	NA	NA
6	0.17	NA	1	101	1	NA
7	0.17	NA	1	97	1	NA
8	0.17	NA	1	106	1	NA
9	0.17	NA	1	111	1	NA
10	0.17	NA	1	79	1	NA
11	0.17	NA	1	43	1	NA
12	0.17	NA	1	64	1	NA
13	0.25	NA	1	90	1	NA
14	0.25	NA	1	141	1	NA
15	0.42	NA	1	42	1	NA
16	0.50	NA	1	43	1	NA
17	0.67	NA	1	132	1	NA
18	0.75	NA	1	43	1	NA
19	0.75	NA	1	36	1	NA
20	1.00	NA	1	86	1	NA
21	1.16	NA	1	44	1	NA
22	1.50	NA	1	68	1	NA
23	1.50	NA	1	89	1	NA
24	1.58	NA	1	101	1	NA
25	1.67	NA	1	115	1	NA
26	1.67	NA	1	53	1	NA
27	1.75	NA	1	94	1	NA
28	1.83	NA	1	95	1	NA
29	1.92	NA	1	76	1	NA
30	2.00	NA	1	79	1	NA

「データセットを表示」をさせてみると、ウィンドウの右側にスクロールバーがでていますね。このデータは 1339 行もある大きなもので、age,menarche,sex,igf1,tanner,testvol という 6 つの変数が格納されています。これは、学校での健康診断の際に、思春期の子供のインスリン様成長因子(IGF-I)の分布を調べたデータです。「NA」という値が目立ちますね。これは、「Not Available」の略、つまり、いわゆる欠損値です。R では欠損値のことを一貫して「NA」と呼びます。

それではまず、このデータの特性をあらわす「記述統計量」を表にまとめて出力してみましょう。「統計量」メニューから「要約」を選び、さらに「アクティブデータセット」を選択してください。





出力ウィンドウのほうに、青文字で要約結果が出力されたことと思います。

```

出力ウィンドウ
> showData(juul, placement='-20+200', font=getRcmdr('logFont'), maxwic

> summary(juul)

      age      menarche      sex      igf1
Min.   : 0.170   Min.   : 1.000   Min.   :1.000   Min.   : 25.0
1st Qu.: 9.053   1st Qu.: 1.000   1st Qu.:1.000   1st Qu.:202.2
Median :12.560   Median : 1.000   Median :2.000   Median :313.5
Mean   :15.095   Mean   : 1.476   Mean   :1.534   Mean   :340.2
3rd Qu.:16.855   3rd Qu.: 2.000   3rd Qu.:2.000   3rd Qu.:462.8
Max.   :83.000   Max.   : 2.000   Max.   :2.000   Max.   :915.0
NA's   : 5.000   NA's   :635.000   NA's   :5.000   NA's   :321.0

      tanner      testvol
Min.   : 1.000   Min.   : 1.000
1st Qu.: 1.000   1st Qu.: 1.000
Median : 2.000   Median : 3.000
Mean   : 2.640   Mean   : 7.896
3rd Qu.: 5.000   3rd Qu.:15.000
Max.   : 5.000   Max.   :30.000
NA's   :240.000   NA's   :859.000

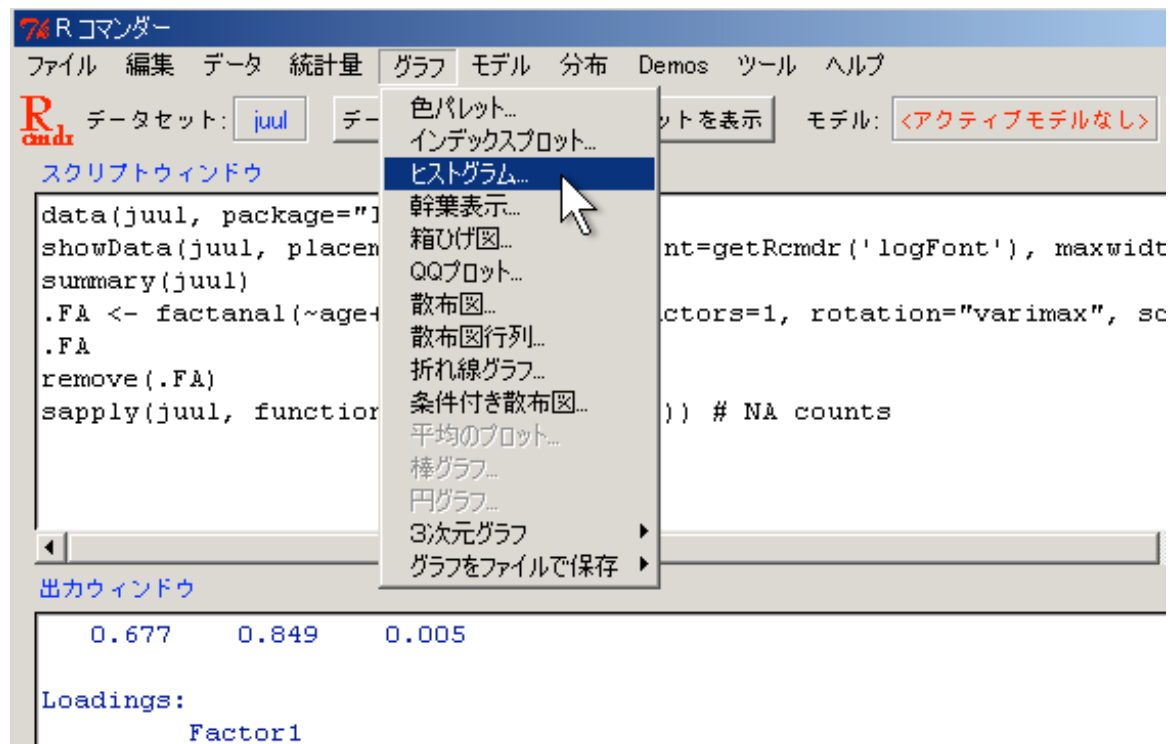
```

これは、アクティブデータセットを構成するそれぞれの変数について、その値の最小値、4分位値、平均値、最大値、欠損値の数をもとめたものをまとめて表示しています。これらの値をみることで、そのデータがどのように「分布」しているのか、すなわち、どちらかという小さい値が多いのか、大きい値が多いのか、また平均値と中央値にどのくらいずれがあるのか、欠損値はどのくらいあるのか、といったことがすぐにわかります。

ただ、sex, menarche, tanner の各変数については、4分位値や最大、最小値がちょっと「きりのいい」値になりすぎている感じがします。また、testvol に関しては、欠損値が多すぎるのがわかるでしょう。これは、sex, menarche, tanner はそれぞれ性別、初経の有無、ターナーの成熟度の5段階分類を示しているためです。これらはカテゴリカルデータなのです。

そして、testvol に関しては、精巣体積です。つまり、男子にしか測定していないため、女子では欠損値になっているのです。

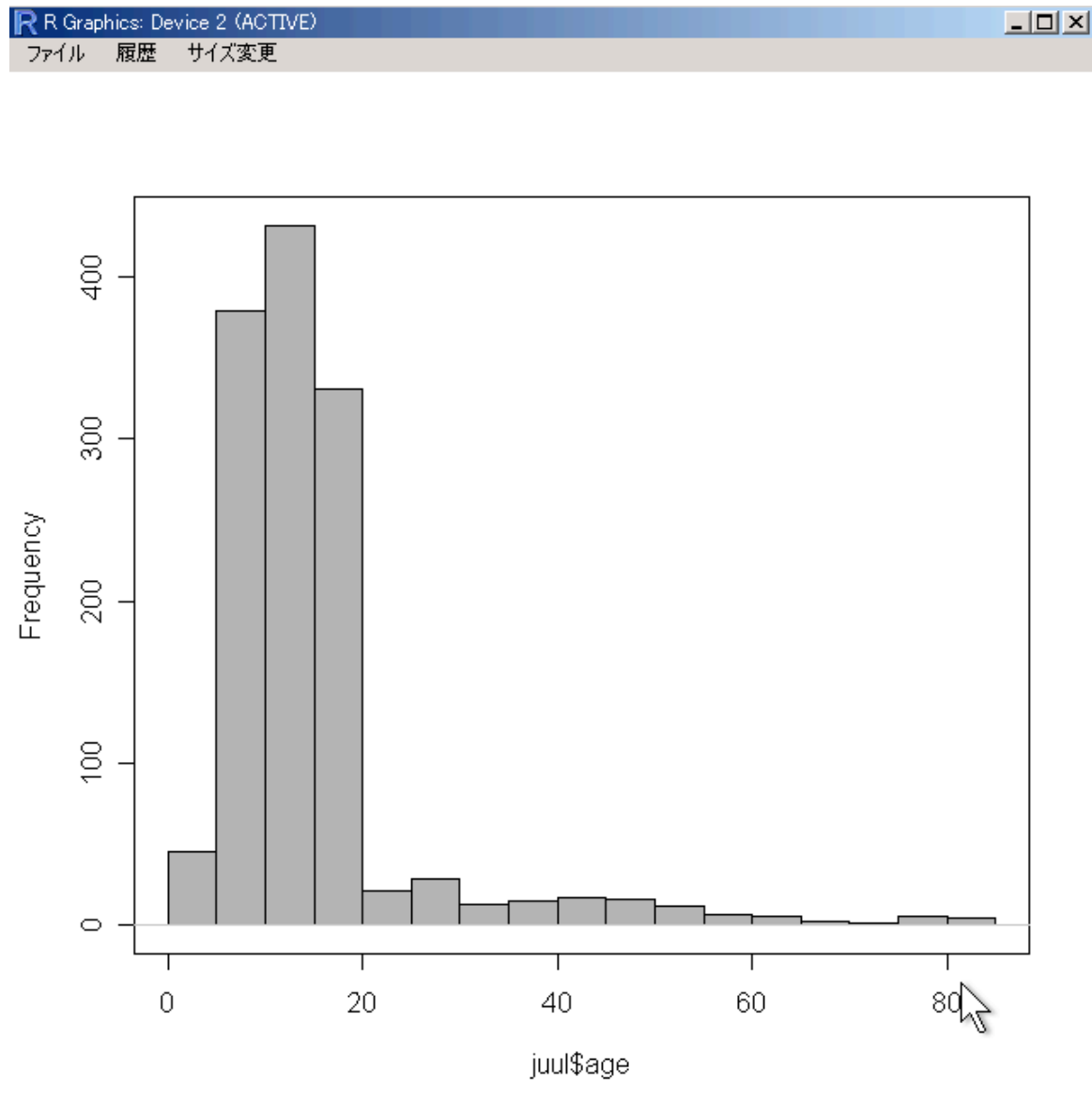
それでは次に、データの分布を知るために用いるもっとも基本的な図であるヒストグラムを描いてみましょう。「グラフ」メニューから「ヒストグラム...」を選んでください。



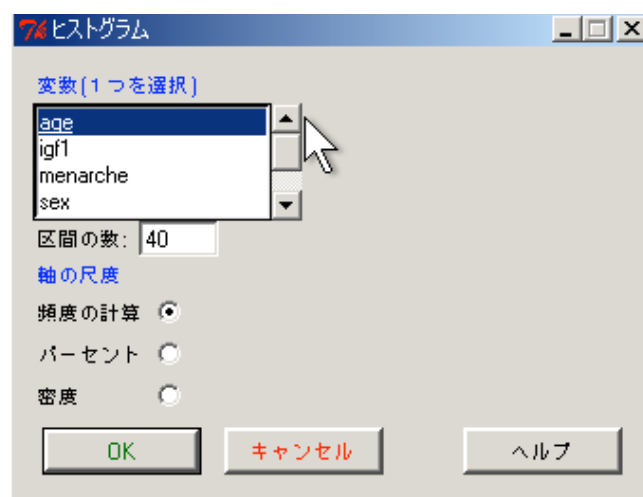
「ヒストグラム」ウィンドウがでたら、変数として「age」を選択、あとはそのまま「Ok」を押してみましょう。

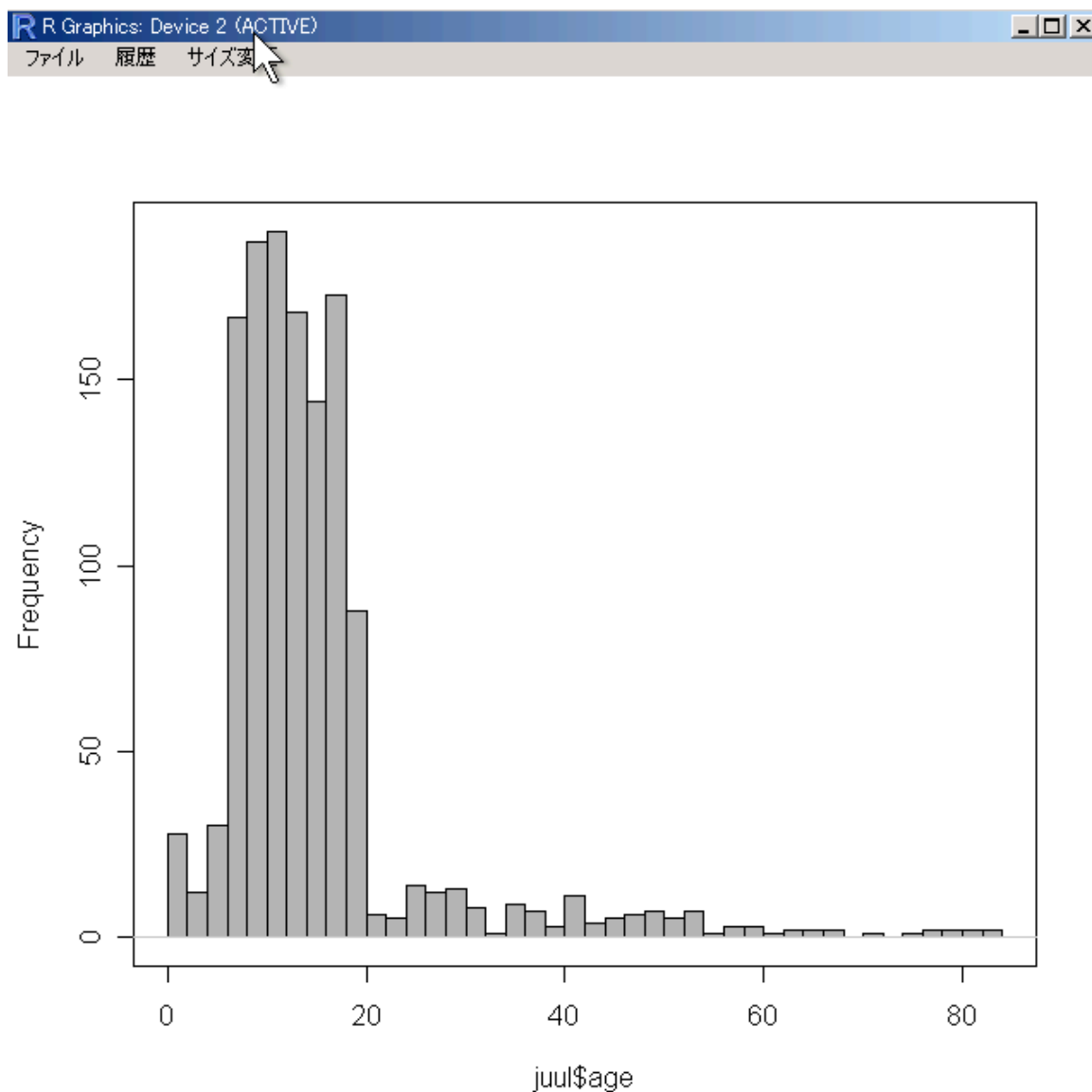


ヒストグラムが無事出力されたでしょうか？年齢分布が圧倒的に 20 歳以下にかたよっていることがわかるでしょう。



デフォルトでは、ヒストグラムの棒1本あたりに含まれる年齢の範囲は、自動的に決まっています。このデータでは少数の高い年齢層の影響を受けて、年齢の範囲が広めになっているようです。もう少し細かいヒストグラムにしてみましょう。「ヒストグラム」ウィンドウで、「区間の数」を40ぐらいにします。



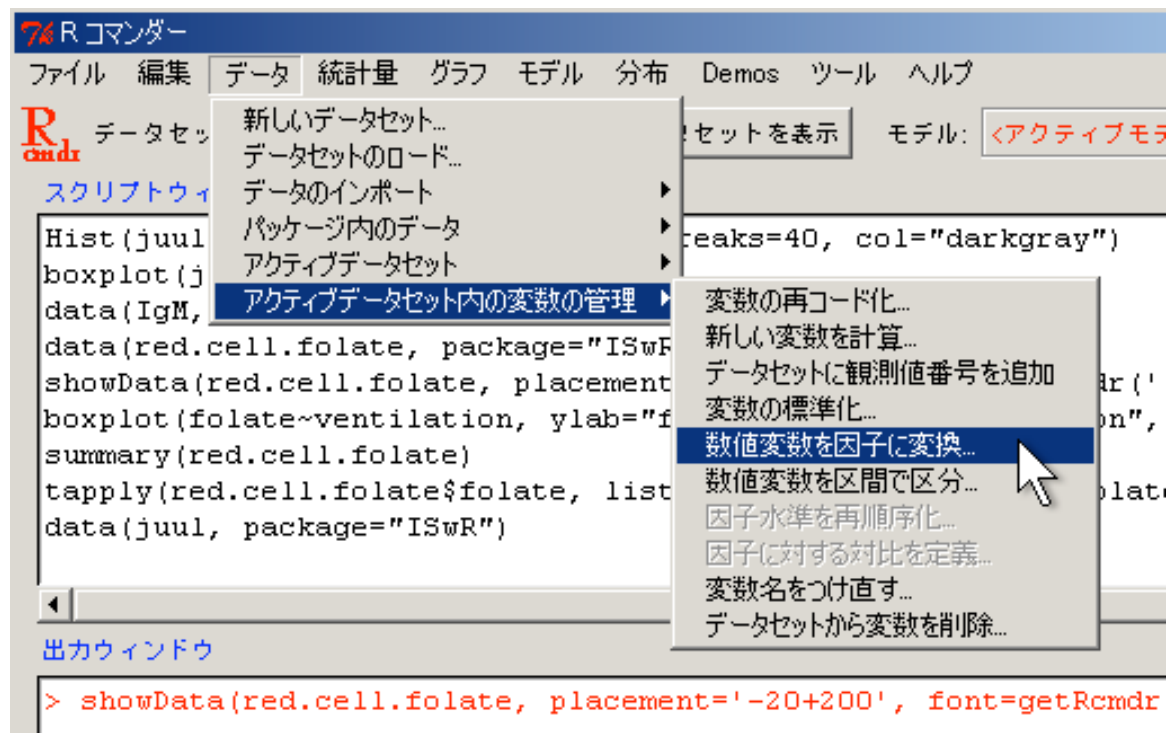


区間分けが細くなったグラフが出力されました。14-16 歳あたりや、2-4 歳あたりに落ち込んでいる部分があることが検出できましたね。

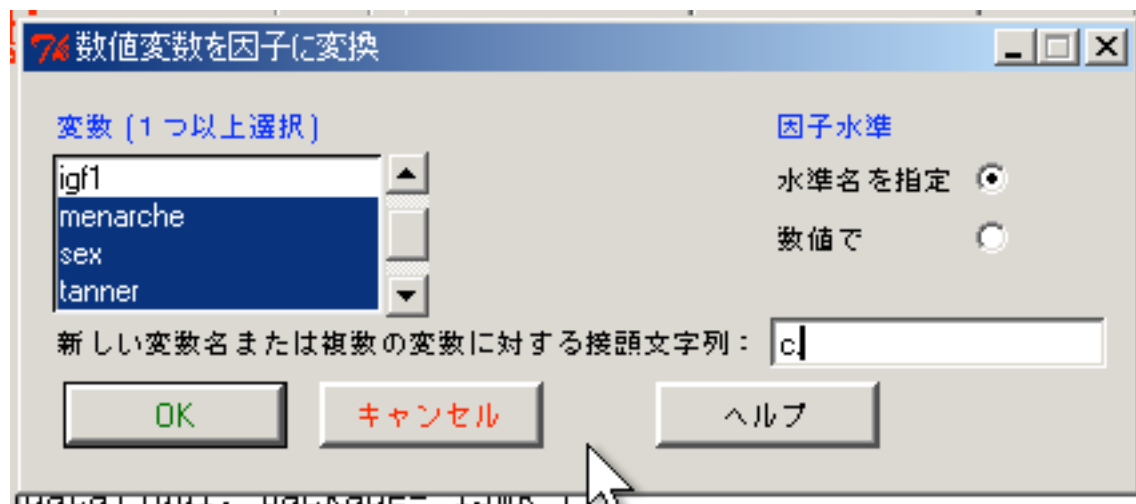
次に箱ひげ図を見ていきたいと思いますが、これは層別にしたほうが見やすくなりますから、まずはこのデータに含まれる、数値で入力されているカテゴリカルデータ(sex, menarche, tanner)がカテゴリカルデータであることを R に教えてあげましょう。

R では、カテゴリカルデータのことを「因子(Factor)」と呼びます。因子分析などの因子とは若干ニュアンスが違い、データの層別に使われる要因、といった意味合いのようです。


「データ」メニューから「アクティブデータセット内の変数の管理...」を選び、さらに「数値変数を因子に変換...」を選んでください。



「数値変数を因子に変換」ウィンドウが出たら、「変数」で「menarche」「sex」「tanner」を選んで、「新しい変数名または複数の変数に対する接頭文字列」に「c.」と入力しましょう。こうすることで、これらの変数のコピーに、「これはカテゴリカルデータですよ」という印をつけた新しい変数、つまり因子化した変数は、もとの変数名の頭に「c.」をつけた、「c.menarche」とか「c.sex」とか「c.tanner」という名前で作成されます。



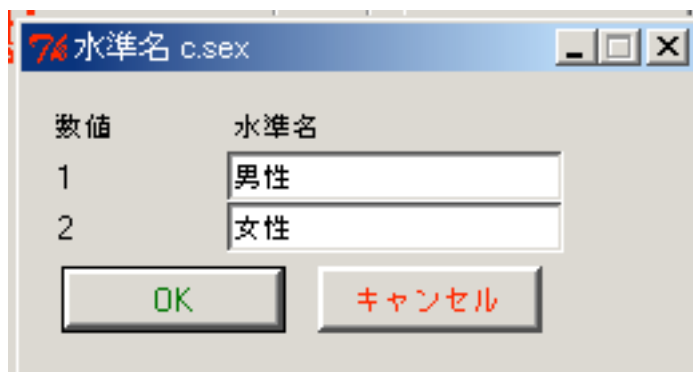
「Ok」を押すと、まずは menarche について聞かれます。



数値	水準名
1	なし
2	あり

OK キャンセル

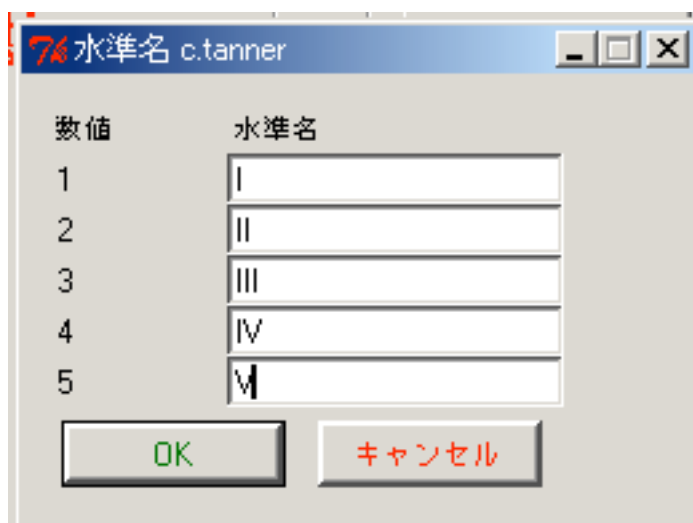
juul データにおける menarche の値は、1: なし, 2: あり ですから、そのように水準名を入力しましょう。続けて c.sex です。こちらは 1: 男性, 2: 女性です。



数値	水準名
1	男性
2	女性

OK キャンセル

最後に c.tanner です。これはローマ数字の I-V にしましょうか。



数値	水準名
1	I
2	II
3	III
4	IV
5	V

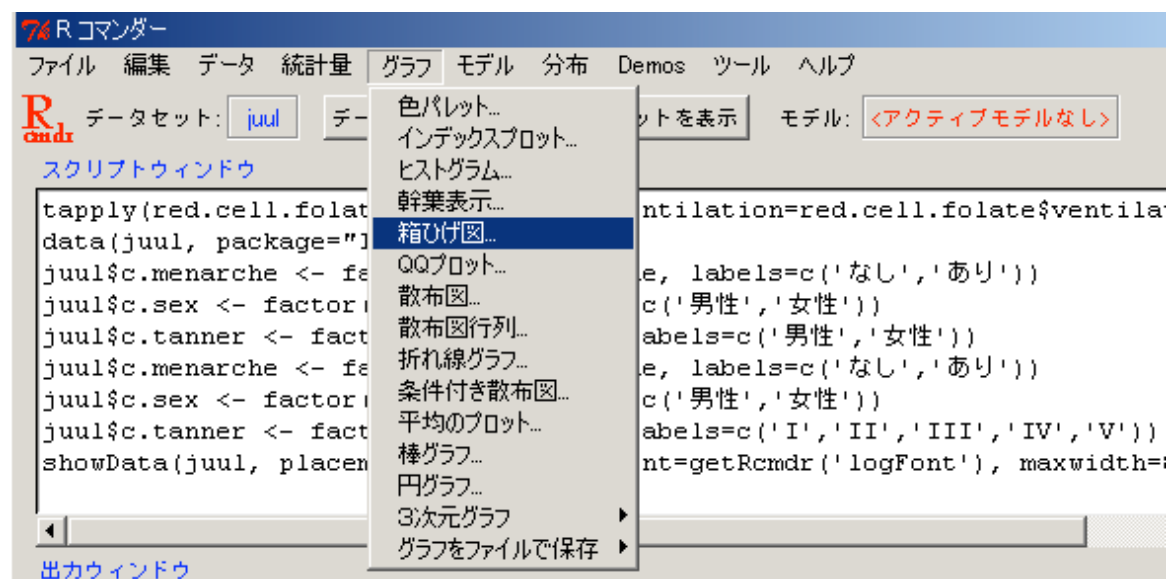
OK キャンセル

Ok を押せば、無事 3 つの因子変数が新しく作成されたはずですが、「データセットを表示」でみてみましょう。

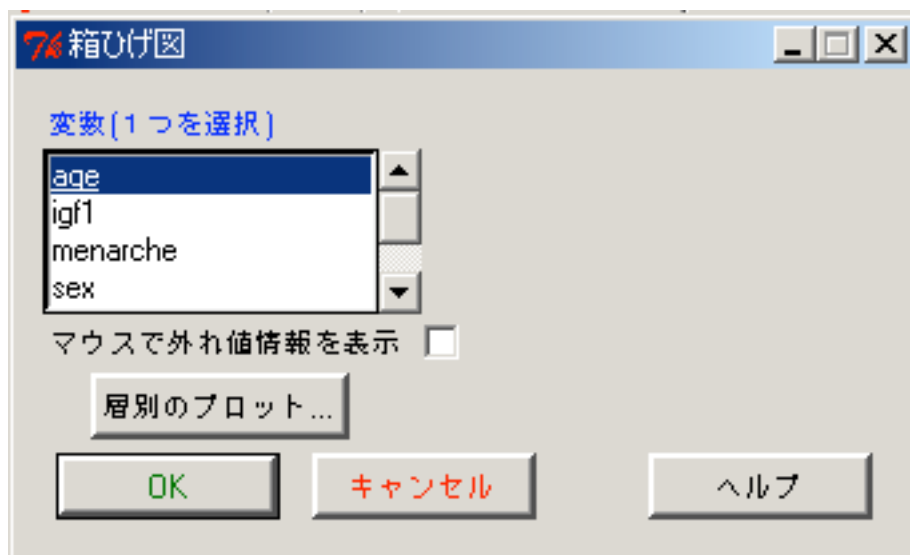


	age	menarche	sex	igf1	tanner	testvol	c.menarche	c.sex	c.tanner
614	59.00	NA	1	134	NA	NA	<NA>	男性	<NA>
615	60.00	NA	1	105	NA	NA	<NA>	男性	<NA>
616	63.00	NA	1	160	NA	NA	<NA>	男性	<NA>
617	65.00	NA	1	92	NA	NA	<NA>	男性	<NA>
618	67.00	NA	1	222	NA	NA	<NA>	男性	<NA>
619	70.05	NA	1	NA	NA	NA	<NA>	男性	<NA>
620	77.00	NA	1	126	NA	NA	<NA>	男性	<NA>
621	78.00	NA	1	90	NA	NA	<NA>	男性	<NA>
622	79.00	NA	1	119	NA	NA	<NA>	男性	<NA>
623	80.00	NA	1	122	NA	NA	<NA>	男性	<NA>
624	81.00	NA	1	112	NA	NA	<NA>	男性	<NA>
625	81.00	NA	1	87	NA	NA	<NA>	男性	<NA>
626	83.00	NA	1	149	NA	NA	<NA>	男性	<NA>
627	83.00	NA	1	104	NA	NA	<NA>	男性	<NA>
628	0.25	NA	2	51	1	NA	<NA>	女性	I
629	0.91	NA	2	25	NA	NA	<NA>	女性	<NA>
630	2.64	1	2	NA	1	NA	なし	女性	I
631	3.25	NA	2	250	NA	NA	<NA>	女性	<NA>
632	5.12	1	2	NA	1	NA	なし	女性	I
633	5.68	1	2	NA	1	NA	なし	女性	I

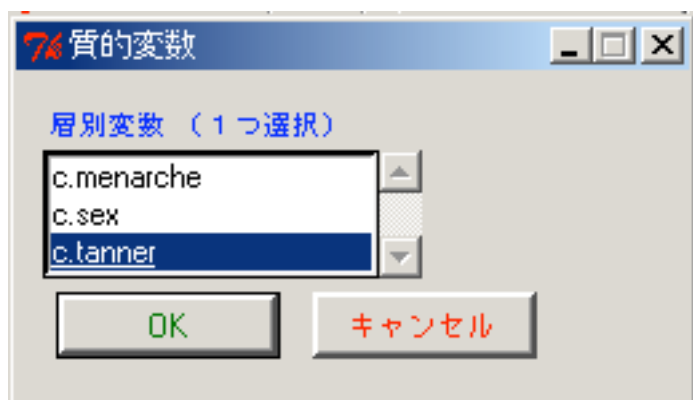
それでは、因子によって層別した箱ひげ図を作成しましょう。「グラフ」メニューから「箱ひげ図...」を選んでください。



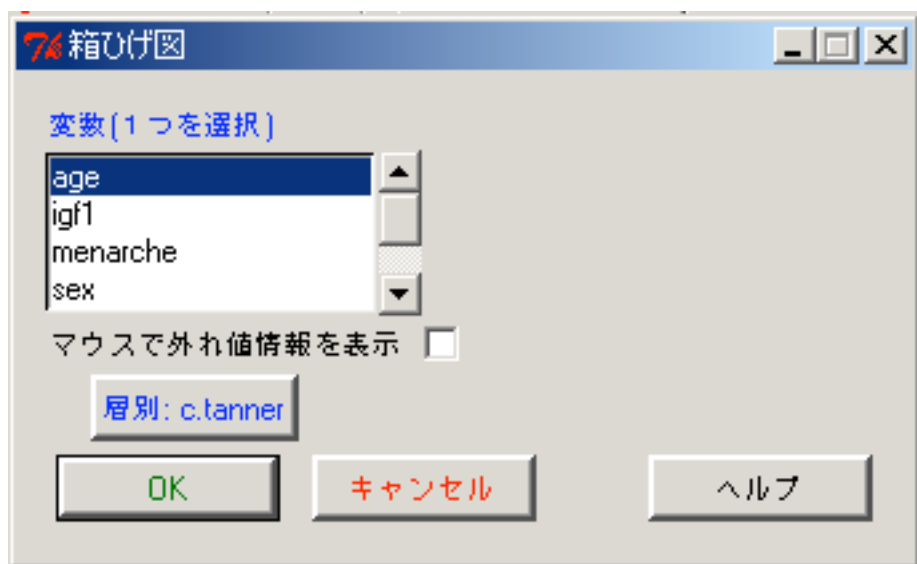
「箱ひげ図」のウィンドウが出たら、変数には「age」を選び、「層別のプロット」ボタンを押しましょう。

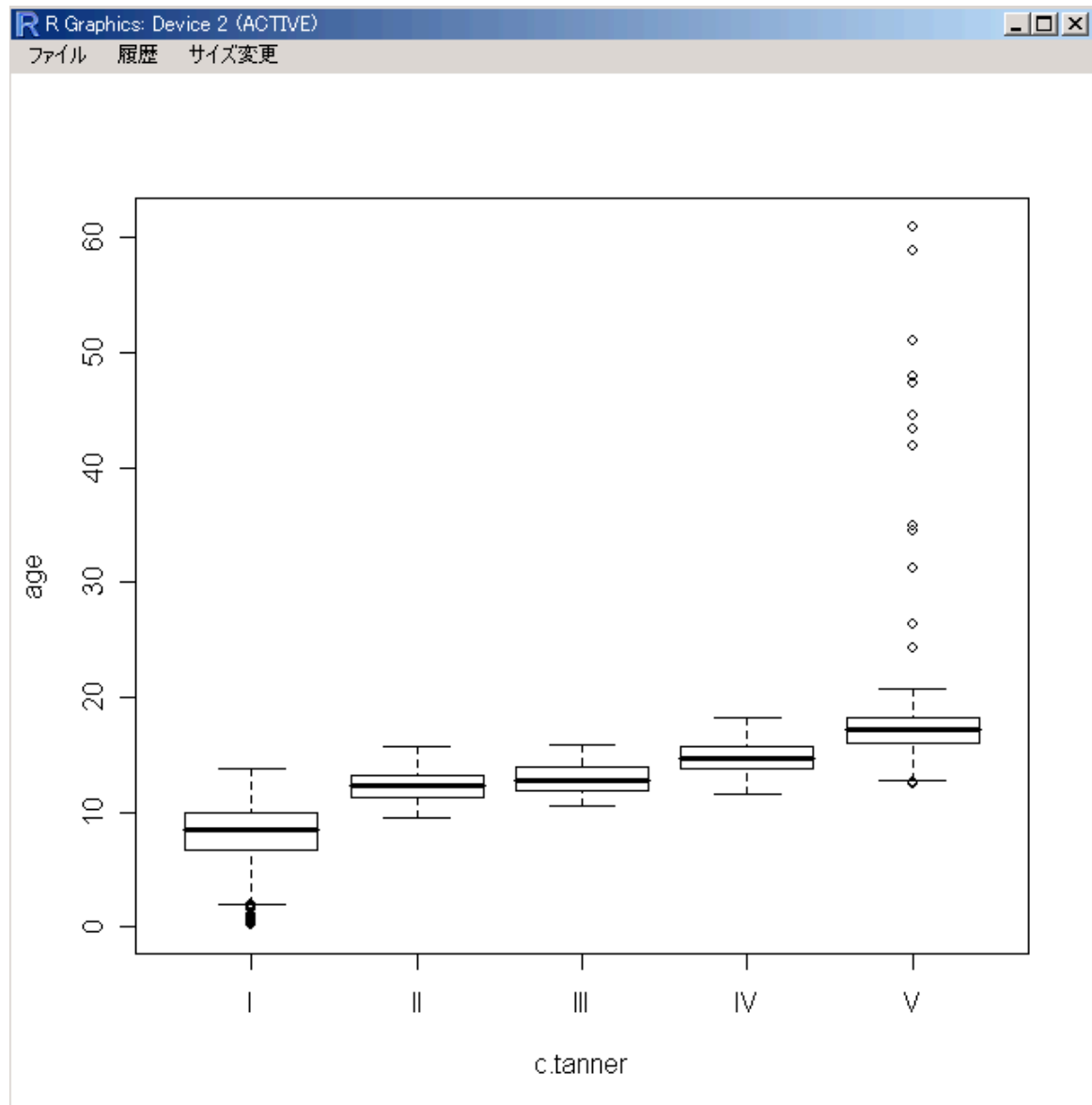


「質的変数」ウィンドウが開いたら、因子化した変数を1つ選ぶことができます。c.tanner を選んでみましょう。



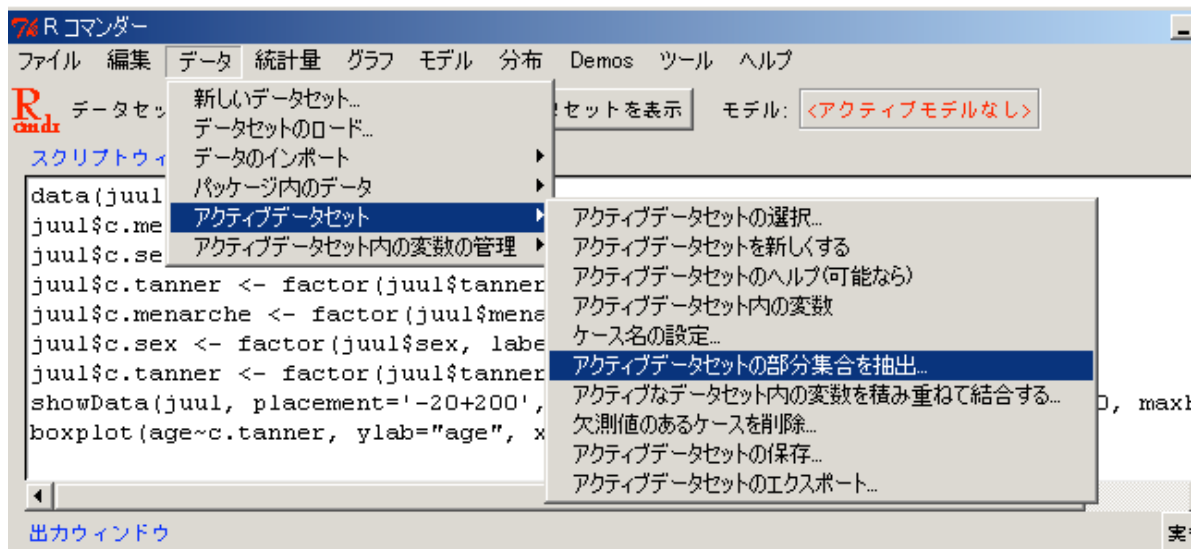
「層別: c.tanner」となっているのがわかりますね。ここで "Ok" を押せばグラフがかかれます。



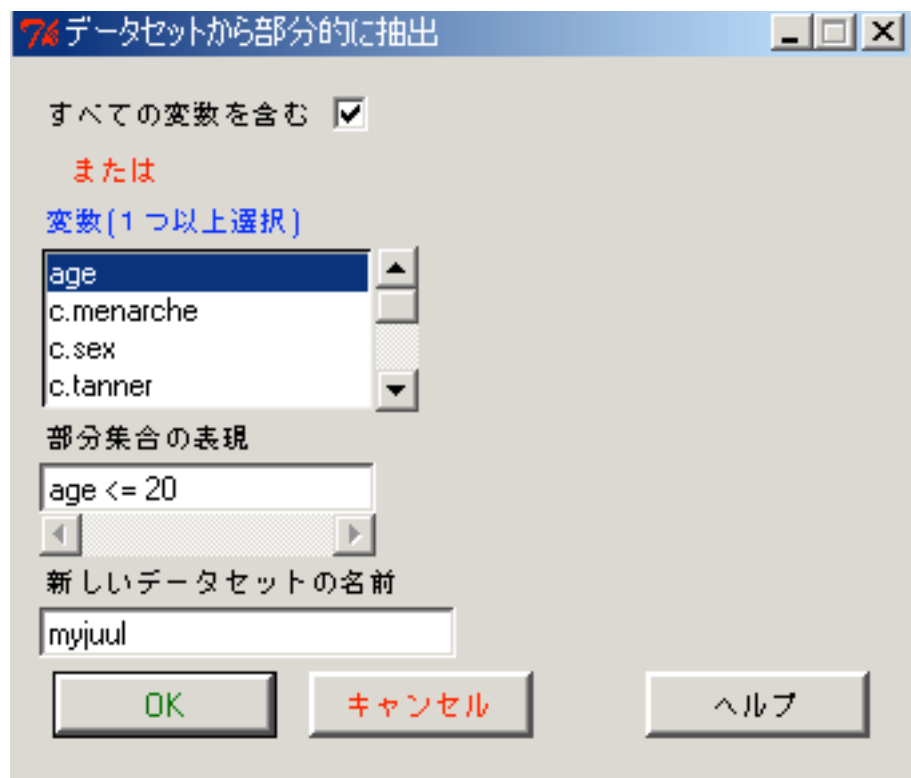


c.tanner は成熟度の指標ですから、年齢が上がるにつれて進んでいることがわかります。でも、V のところはかなり年齢の高い人が含まれていますから、その影響で I-IV がわかりにくくなっていますね。では、年齢が 20 歳以下の人のデータだけをみてみましょう。

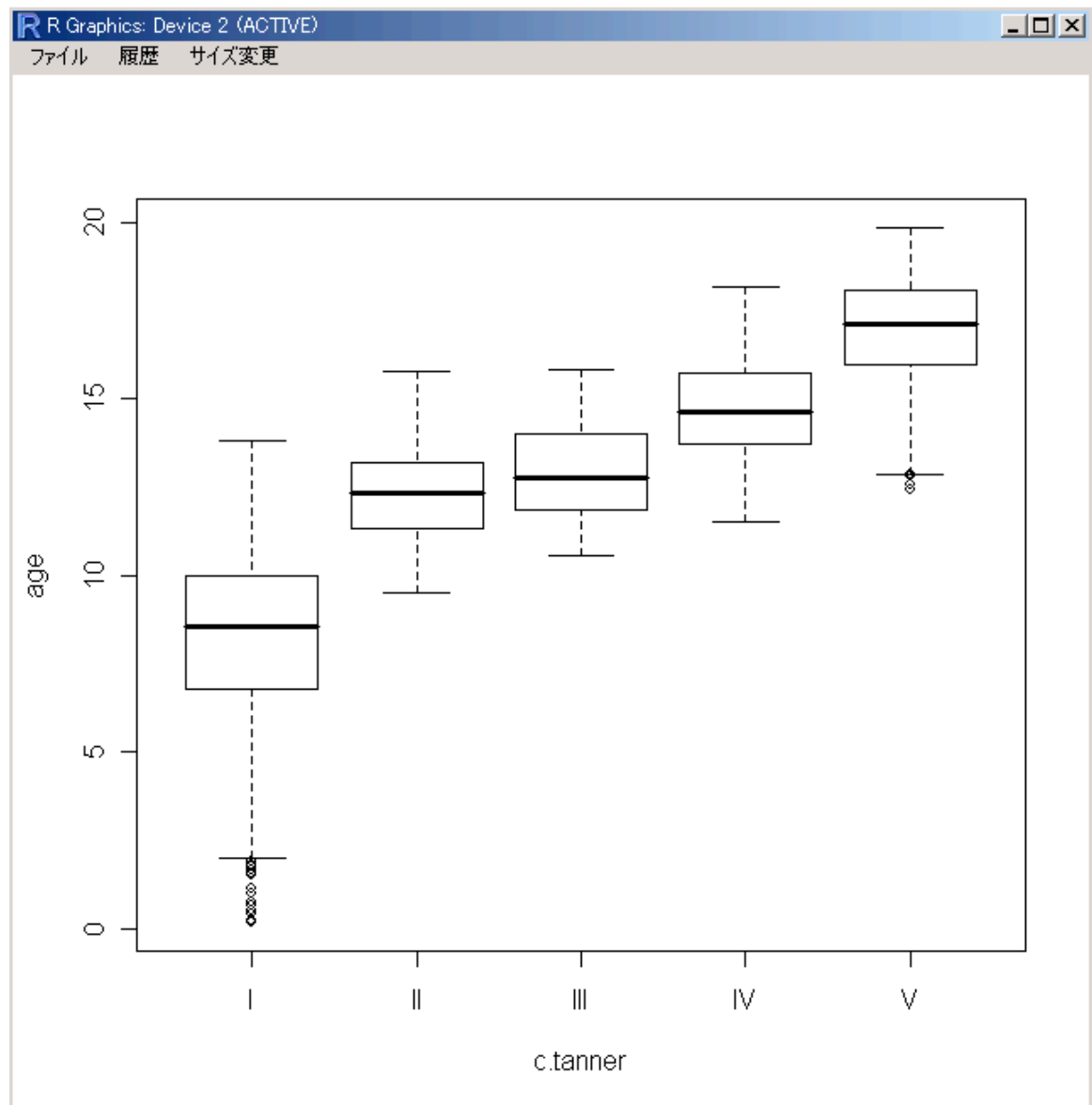
まず、データセットからその部分集合を抽出します。「データ」メニューから「アクティブデータセット」を選び、さらに「アクティブデータセットの部分集合を抽出」としましょう。



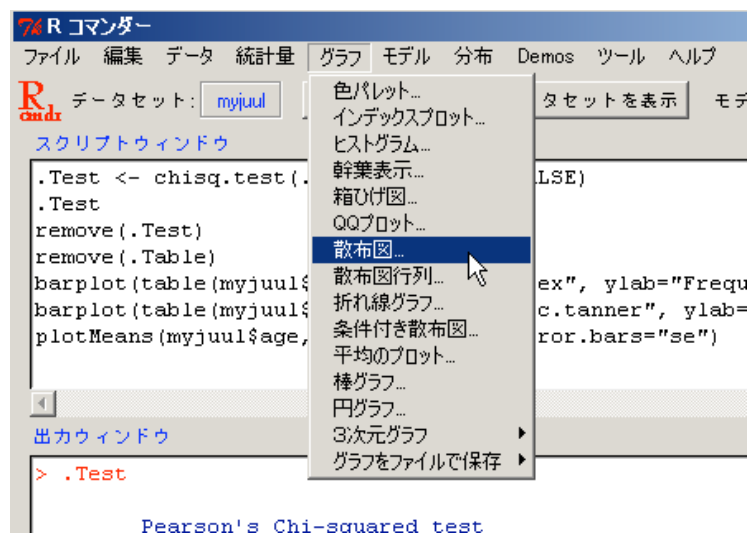
「データセットから部分的に抽出」ウィンドウが開いたら、「すべての変数を含む」がチェックされていることを確認し、「部分集合の表現」として、20 歳以下ということで "age <= 20" を入力。そして、抽出結果のあたらしいデータセットの名前として「myjuul」などとして "Ok" を押します。



それでは、myjuul で、さきほどと同様の箱ひげ図をかいてみましょう。以下のようにになりましたか？



では、次に連続（量的）変数同士のプロットをやってみましょう。最初にためしたときと同様、「グラフ」メニューから、「散布図...」を選択してください。



74 散布図

x変数 [1 つ選択] y変数 [1 つ選択]

age
igf1
menarche
sex

age
igf1
menarche
sex

点を確認する ☐

x の値にゆらぎを与えて表示 ☐

y の値にゆらぎを与えて表示 ☐

周辺箱ひげ図 ☒

最小 2 乗直線 ☒

平滑線 ☒

スムージングの幅

部分集合の表現

<全ての有効なケース>

層別のプロット...

x 軸ラベル

y 軸ラベル

プロットするパラメータ

プロットするパラメータ

点の大きさ

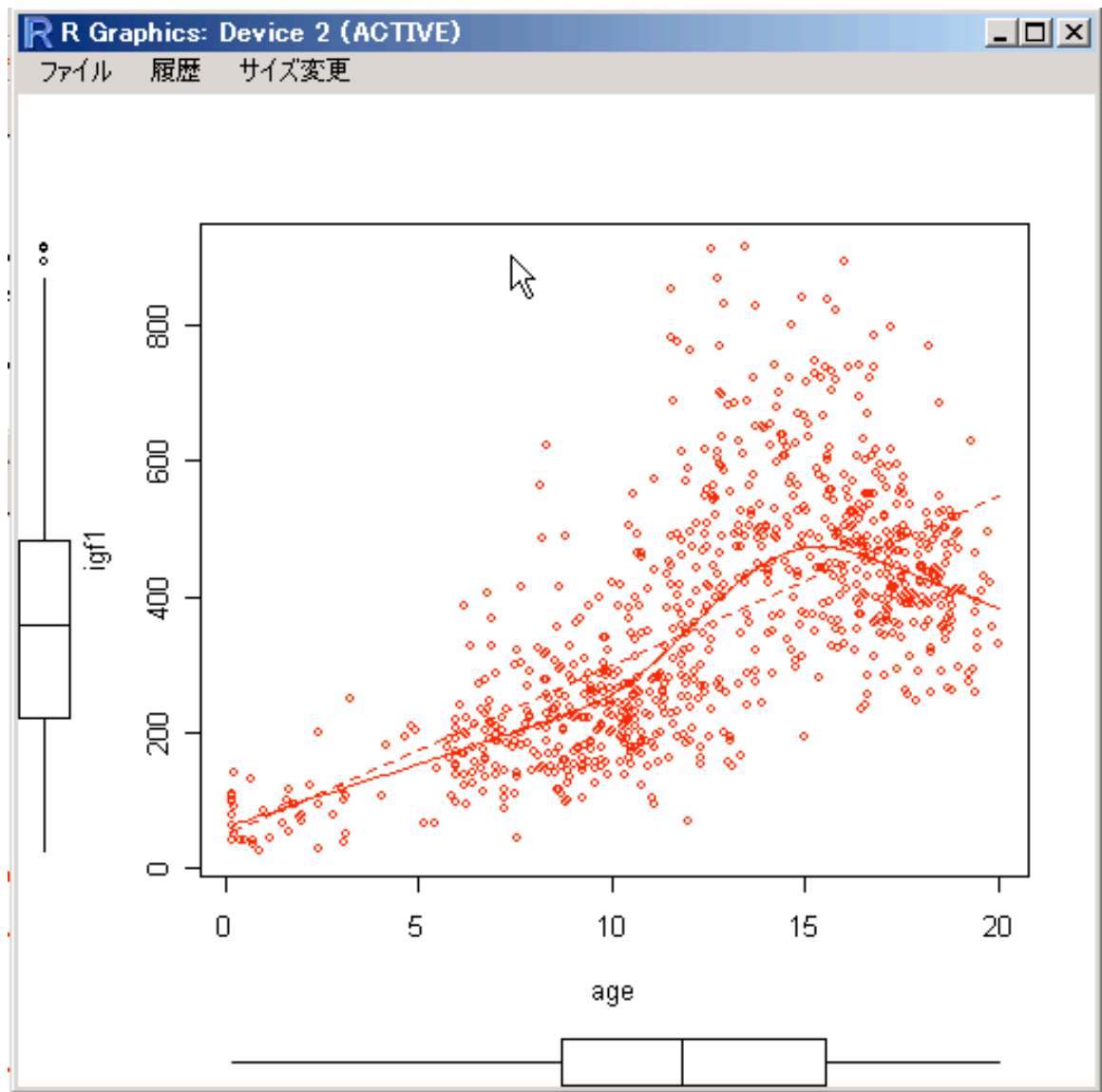
軸テキストの大きさ

軸ラベルのテキストの大きさ

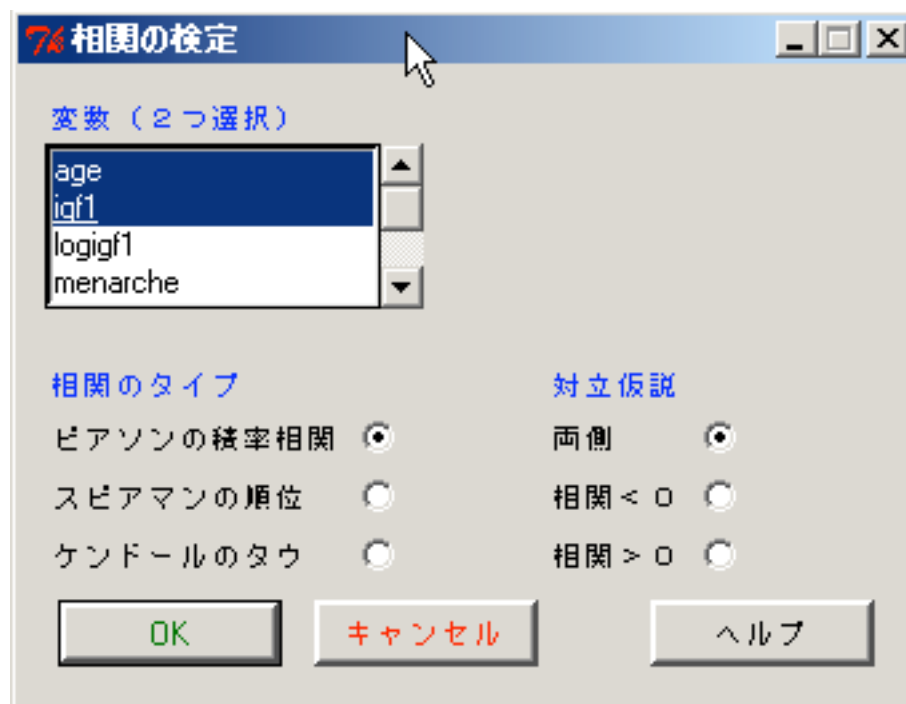
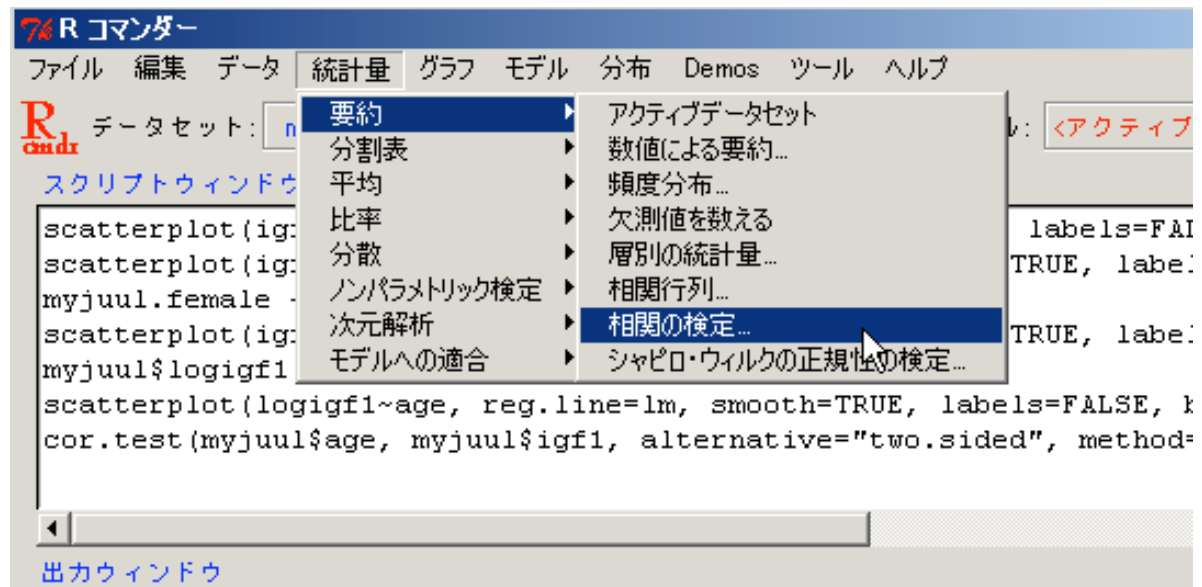
OK キャンセル ヘルプ

「散布図」ウィンドウが出たら、x 変数として "age", y 変数として "igf1" を選択して "Ok" とします。

年齢と IGF-I の値の関係を示すプロットが出力されました。軸の周辺に、それぞれの変数の箱ひげ図も表示されているので、x 軸, y 軸変数の分布も同時に把握することができます（「周辺箱ひげ図」チェック）。それと、ノンパラメトリック回帰による平滑化曲線（「平滑線」チェック）、回帰直線（「最小 2 乗直線」チェック）も同時にプロットされています。



プロットに対応する、age と igf1 の相関係数も計算してみましょう。「統計量」メニューの「要約」、「相関の検定...」です。



変数として age と igf1 を選んで、ピアソンの積率相関係数を選ぶと、出力ウィンドウに計算結果が出ます。



```
> cor.test(myjuul$age, myjuul$igf1, alternative="two.sided", method="pearson")

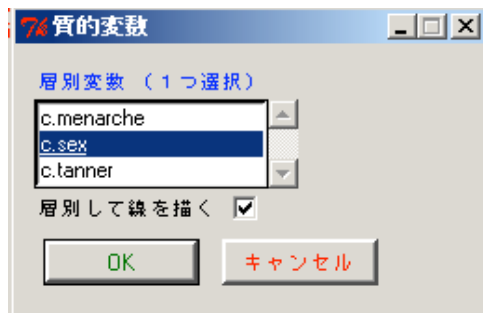
Pearson's product-moment correlation

data: myjuul$age and myjuul$igf1
t = 24.6737, df = 869, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6010404 0.6793067
sample estimates:
      cor 
0.641842
```

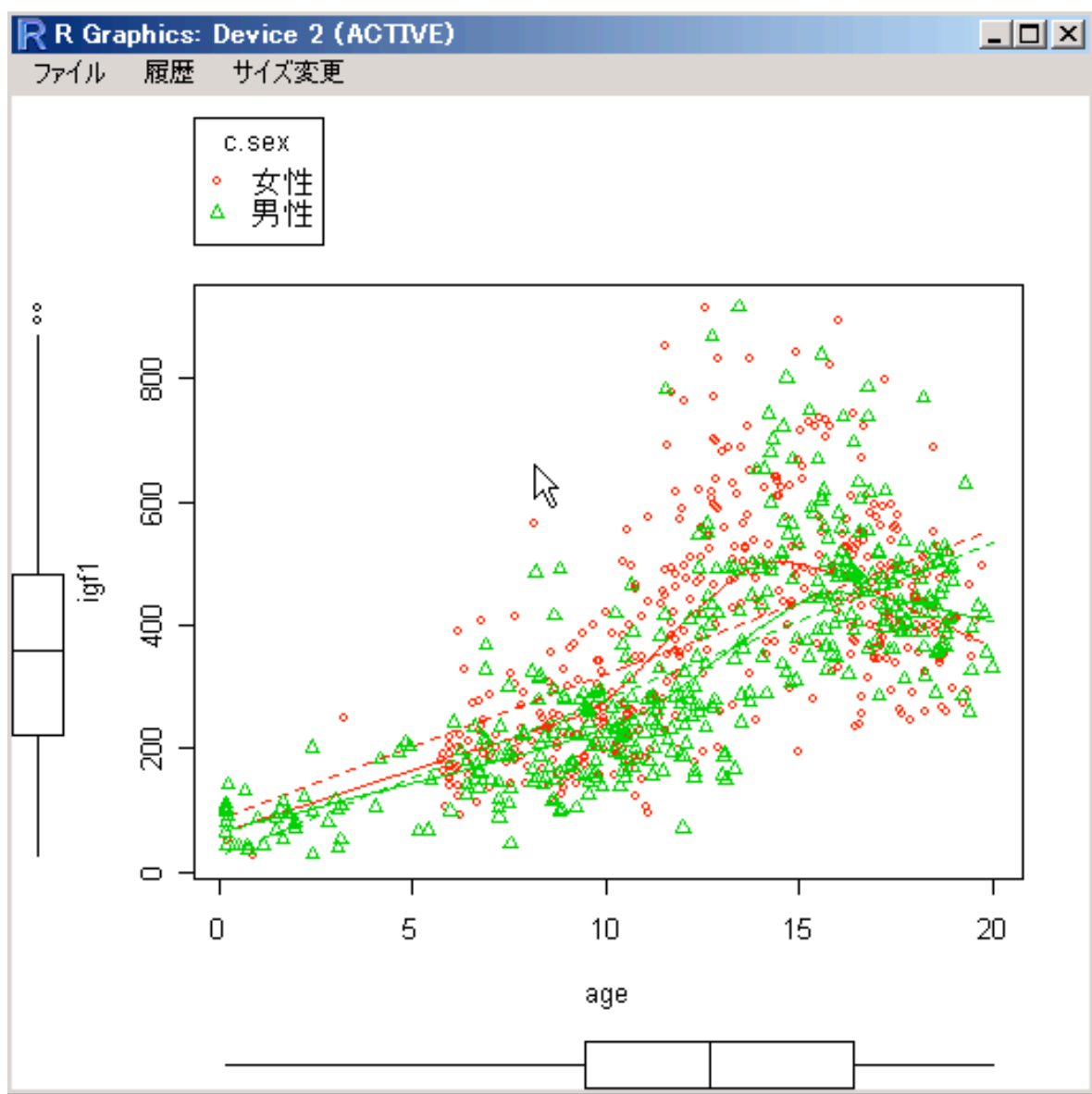
標本相関係数 0.642, 母相関係数 =0 の帰無仮説は非常に小さい p 値で棄却されていますね。

しかし、この図をみると、年齢が7歳ぐらいまでは直線へのあてはまりがよさそうですが、それ以降は IGF-I のばらつきがかなり大きくなっていますね。成長ホルモンに関連するものですから、年齢とは関係がありそうな気がします...他にも関係する要因があるのかもしれませんが。そこで、性別で層別してみましょう。

「散布図」ウィンドウにも、「層別のプロット」というボタンがありましたね。早速、c.sexで層別をしてみましょう。

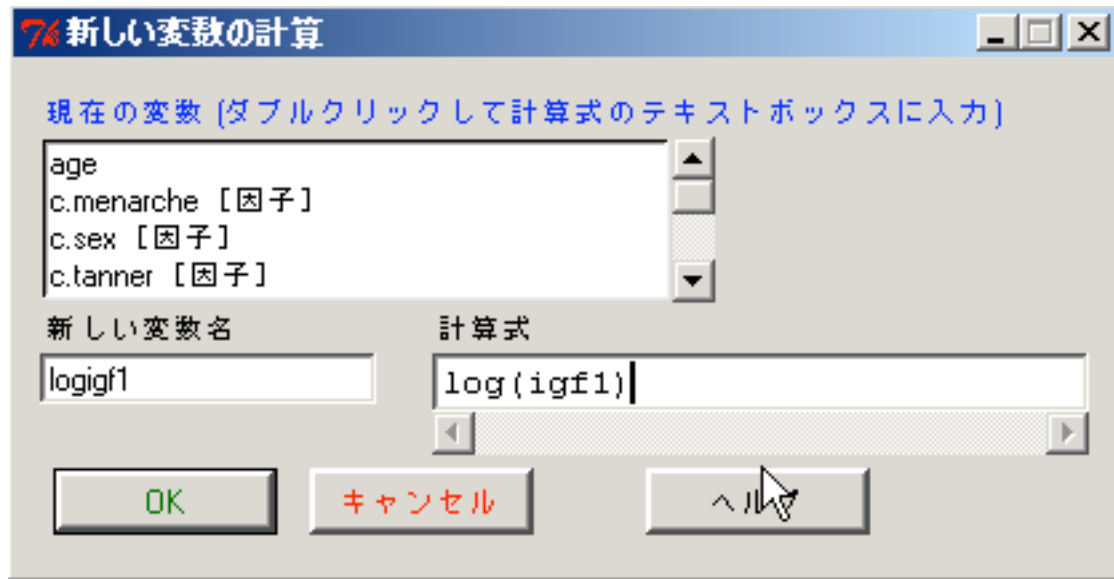


そうすると、プロットの点が男性と女性で色分けされて、下図のようなグラフが出ます。

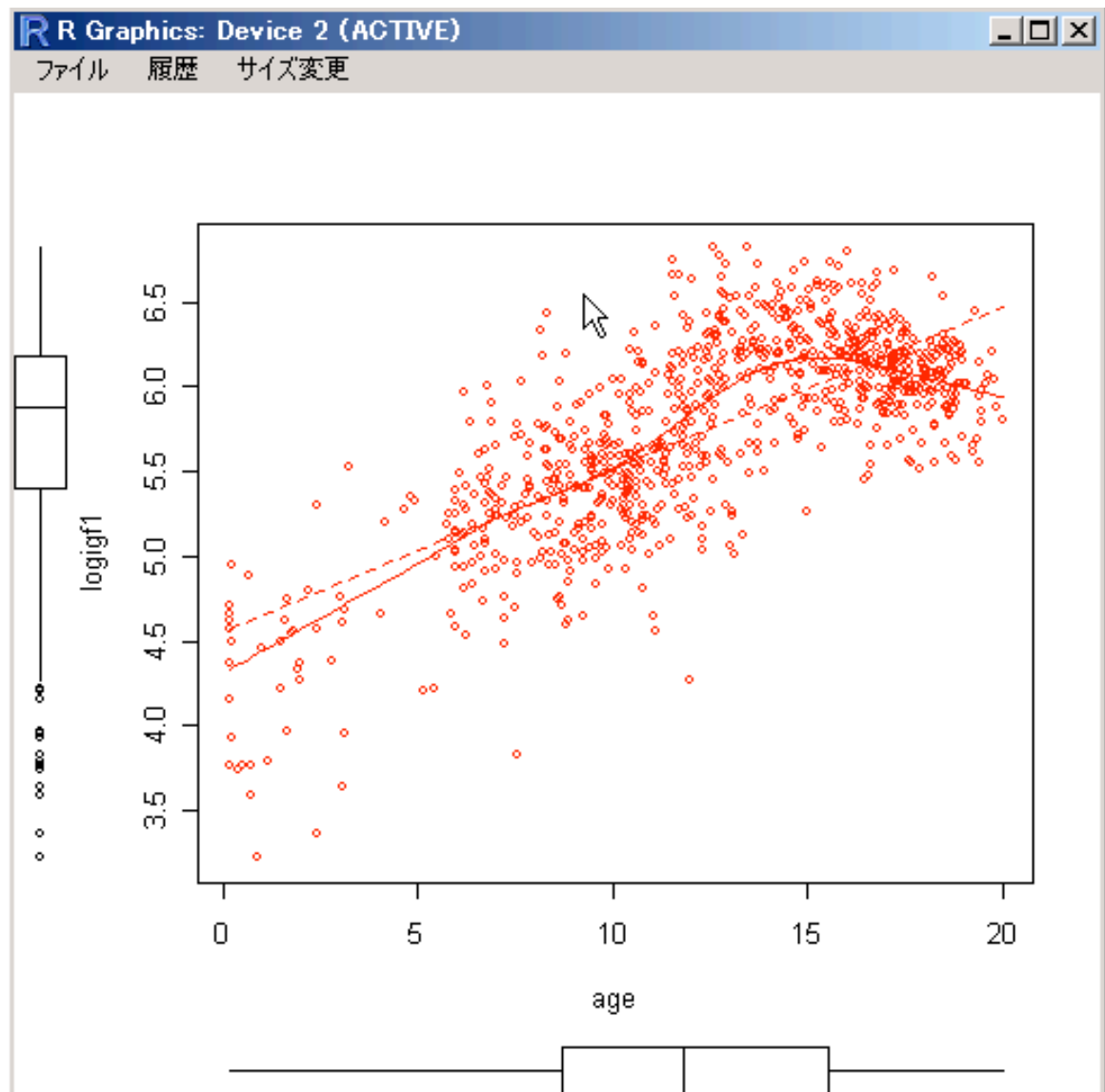


...うーん、あんまりきれいにわかれた感じではありませんね。

どうも、igf1 が年齢がすすむにつれて、直線的ではなく指数的に増えている感じを受けます。では、igf1 を「対数変換」して、もう一度プロットしてみましょう。weight と height から BMI をつくったときの思い出して、igf1 を対数変換した変数、logigf1 を作成してください。「データ」メニューの「アクティブデータセット内の変数の管理」、その先の「新しい変数を計算...」です。



そして、age と logigf1 でプロットしてみると...



前よりは直線によくのってきましたね。相関係数も計算してみましょう。

```
> cor.test(myjuul$age, myjuul$logigf1, alternative="two.sided", method="pearson")

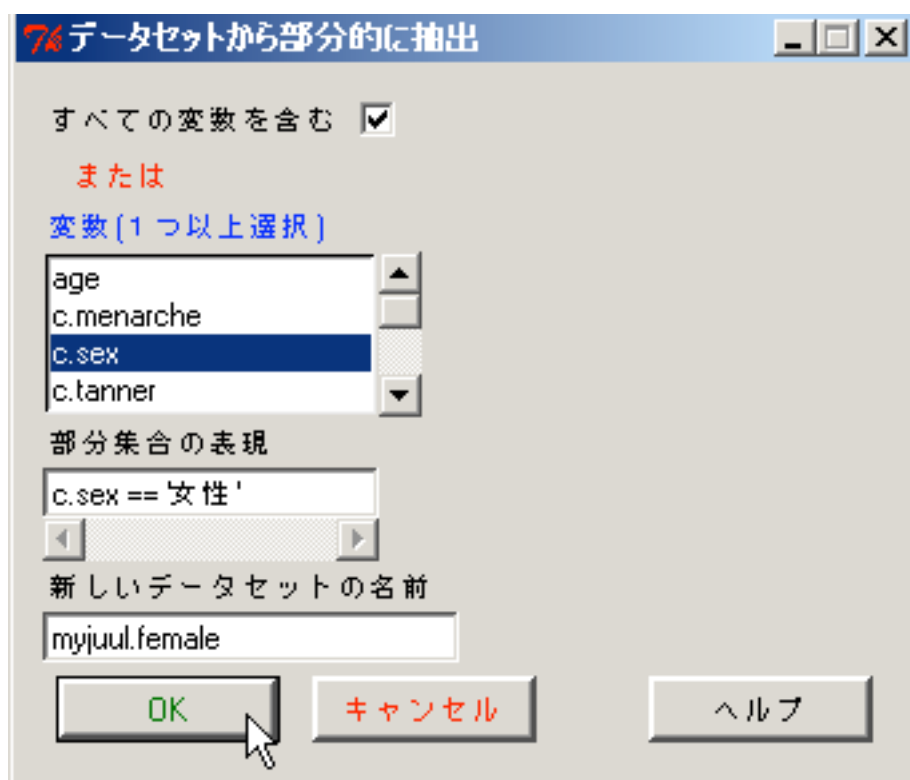
Pearson's product-moment correlation

data: myjuul$age and myjuul$logigf1
t = 32.0789, df = 869, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7043406 0.7653123
sample estimates:
      cor 
0.7363176
```

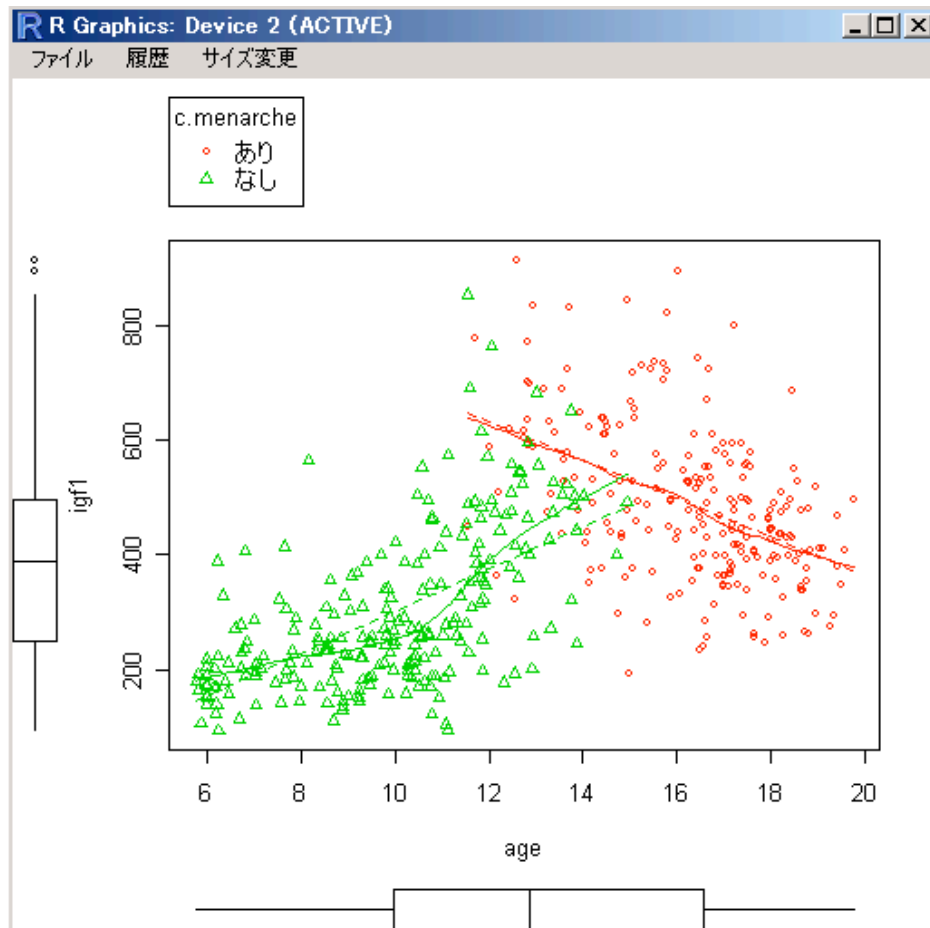
標本相関係数 0.736, その 95%信頼区間 0.704-0.765 ということで、igf1 そのままよりはよく相関しています。

ただ、15 歳以降でちょっと減少傾向に転じているような...

では、今度は初経の有無で層別してみましょう。c.menarche は女性にしかないので、まずはデータセットからの部分抽出で、女性だけのデータセット、myjuul.female をつくります。やり方はわかりますね。



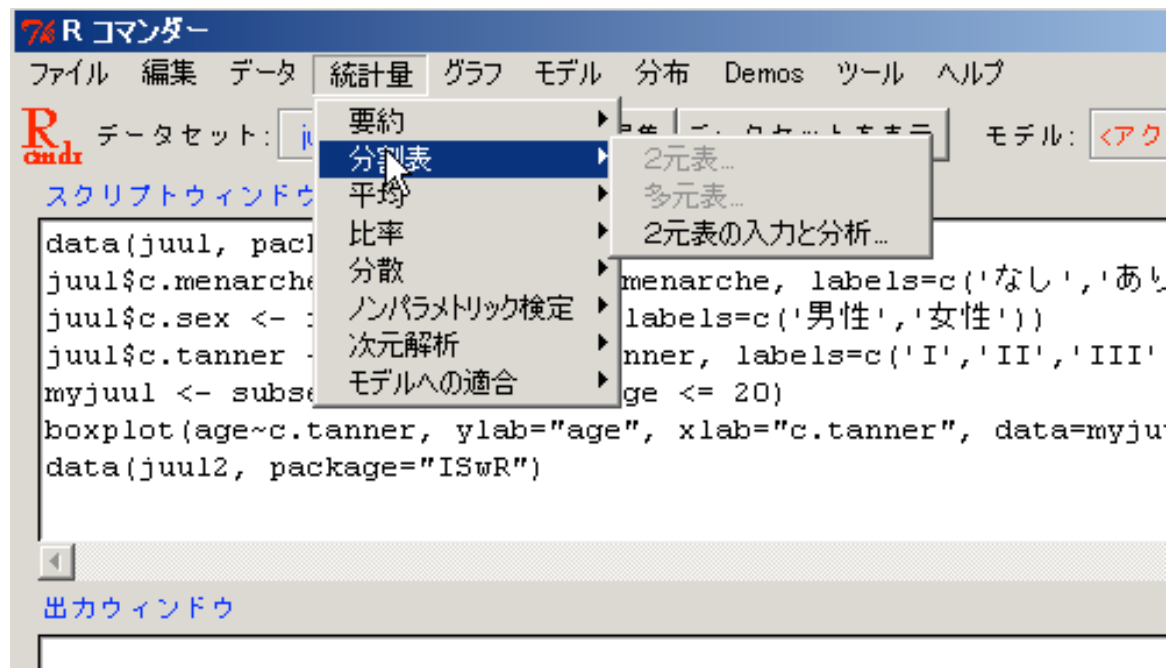
そして、myjuul.female における age と igf1 の散布図を、c.menarche で層別して描いてみましょう。



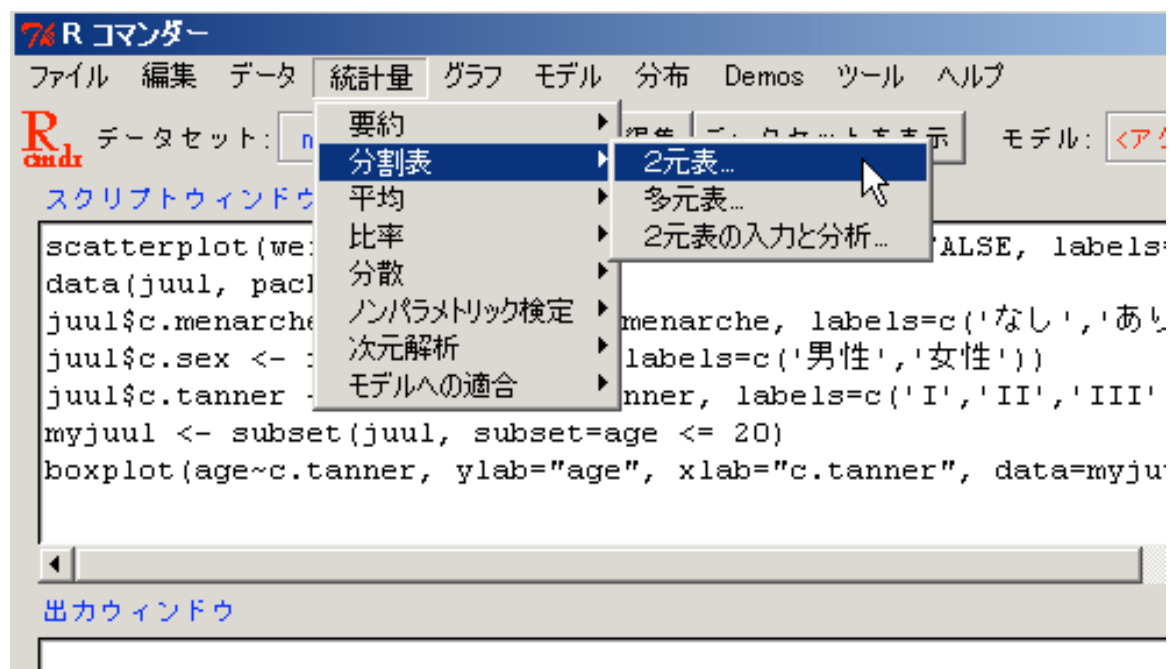
初経の前後で増加傾向から、減少傾向に転じていることがわかりますね。

3. カテゴリカルデータの集計

myjuul データセットを用いて、カテゴリカルデータの処理についてもう少しみていきましょう。「統計量」メニューから「分割表」とたどると、変数の因子化をする前は一部のメニューがかすれた文字になっていて無効化されていたと思います。



それが、因子化したあとの現在のデータセットでは、「2 元表...」などが選択できるようになっています。これは、因子化をすることによって、このデータセットのなかのいくつかの変数はカテゴリカルデータである、ということが R に伝わったためです。R はデータが数量データなのか、カテゴリカルデータなのかをいつも見えています。それに応じて処理を変更してくれるのです。それでは、「分割表」から「2 元表...」を選んでみましょう。



この「2 元表」とは、いわゆる 2×2 の分割表です。行、列それぞれに指定したカテゴリカル変数の値に基づいて、それぞれの条件を満たすデータが「いくつあるのか」を集計して表にしてくれます。では、まず c.sex と c.menarche でみてみましょうか。

2元表

行の変数 (1つ選択) 列の変数 (1つ選択)

c.menarche
c.sex
c.tanner

c.menarche
c.sex
c.tanner

パーセントの計算

行のパーセント ☐
列のパーセント ☐
総計のパーセント ☐
パーセント表示無し ☒

仮説検定

独立性のカイ2乗検定 ☒
カイ2乗統計量の要素 ☐
期待度数の表示 ☐
フィッシャーの正確検定 ☐

部分集合の表現

<全ての有効なケース>

OK キャンセル ヘルプ

上のように指定して「Ok」を押すと...

```
出力ウィンドウ
> .Table
      c.menarche
c.sex なし あり
  男性    0    0
  女性  368  263

> .Test <- chisq.test(.Table, correct=FALSE)

> .Test

      Pearson's Chi-squared test

data:  .Table
X-squared = NaN, df = 1, p-value = NA

> remove(.Test)
```

出力ウィンドウの上のほうに、c.sex と c.menarche でクロス集計をした結果が表示されます。c.sex は性別、c.menarche は初経の有無ですから、当然男性ではすべてゼロになりますね。何らかのデータ入力ミスがあれば、ここで気がつくことができます。では、今度は c.menarche と、c.tanner で集計してみましょう。

```

出力ウィンドウ

> .Table <- xtabs(~c.menarche+c.tanner, data=myjuul)

> .Table

      c.tanner
c.menarche  I   II  III  IV   V
   なし  221  43   32   14   2
   あり    1   1    5   26  190

> .Test <- chisq.test(.Table, correct=FALSE)

> .Test

      Pearson's Chi-squared test

data:  .Table |
X-squared = 463.5167, df = 4, p-value < 2.2e-16

```

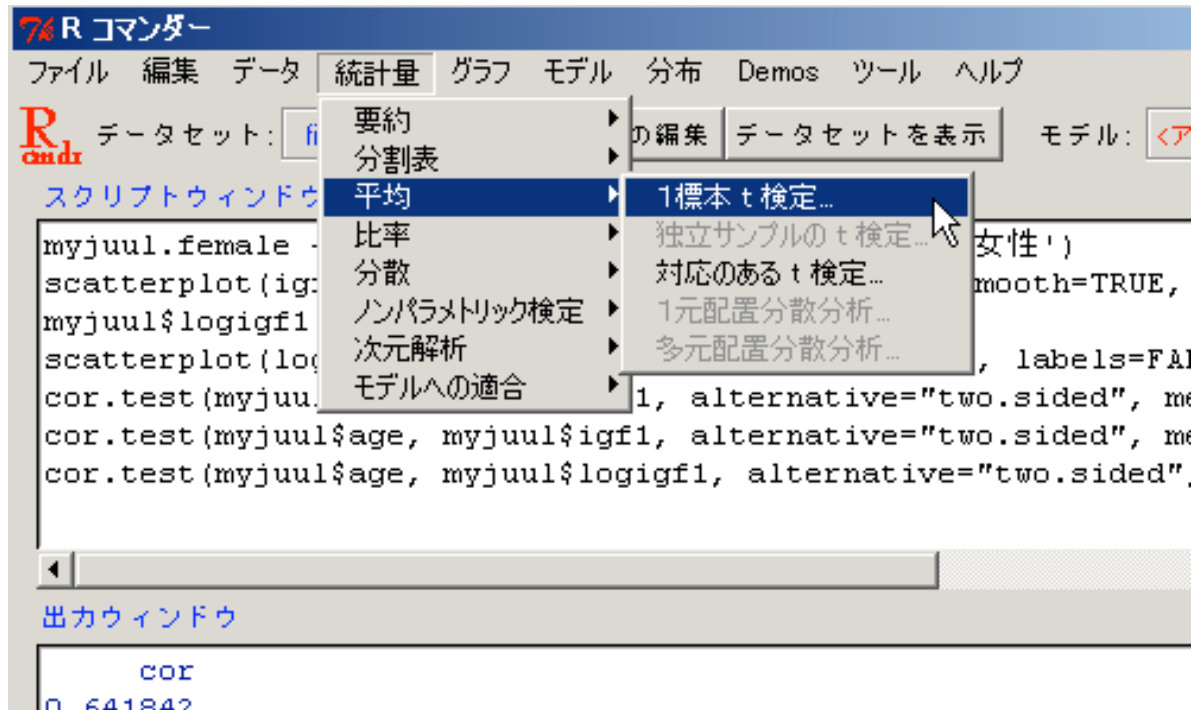
このように集計されます。初経の有無と、ターナーの成熟度分類の値には何らかの関係があるように見えますね。出力ウィンドウの下の方には、この分割表（今回は成熟度分類が5段階なので、2x2 ではないですが）に対してカイ 2 乗検定を行った結果が出ています。p 値が非常に小さく、この2つのカテゴリ変数は独立しているという仮説が棄却できることがわかります。

なお、ここではやりませんが、分割表が 2x2 になったときには、「フィッシャーの正確検定」にチェックをいれるべきだったことも忘れずに。

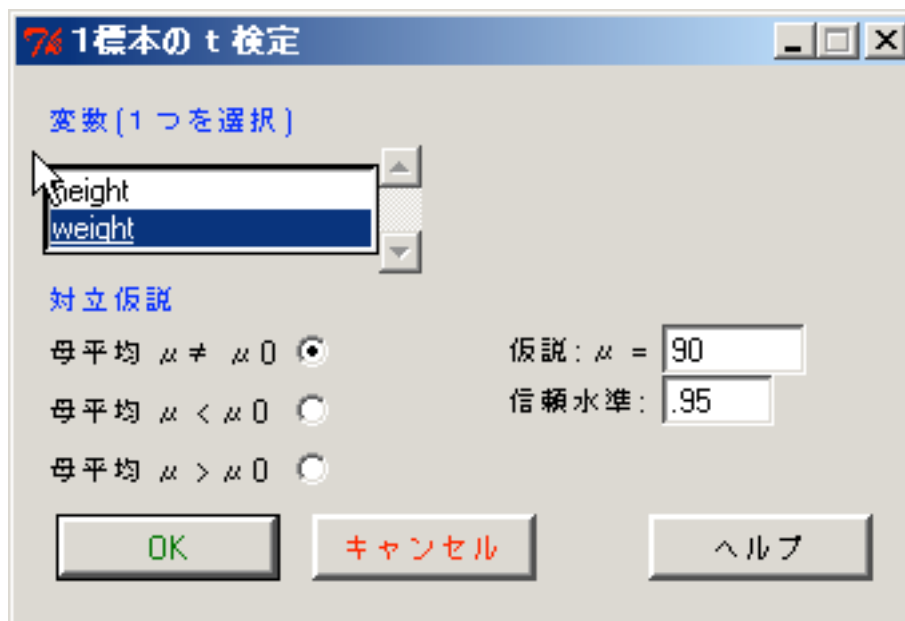
4. 1 標本または 2 標本の検定

ここまでで、カイ 2 乗検定と相関係数の検定はやりましたが、おそらくもっともよく知られている検定、t 検定についてはまだでしたね。figures6 に戻って、まずは 1 標本の検定からやってみましょう。データセットを figures6 に指定しなおしたら、「統計量」メニューから「平均」、「1 標本

t 検定」にいきます。



weight は平均 74.33, 最大値 95 でしたから、この母集団は体重がちょっと重そうです。そこで、体重の母平均値が 90kg である、ということを帰無仮説、90kg とはいえないということを対立仮説として、両側検定をしてみましょう。



母平均値 μ に 90 を入力、対立仮説に母平均 $\mu \neq \mu_0$ を指定して、Ok を押してみます。

```
> t.test(figures6$weight, alternative='two.sided', mu=90, conf.level=.95)

One Sample t-test

data:  figures6$weight |
t = -2.4882, df = 5, p-value = 0.05528
alternative hypothesis: true mean is not equal to 90
95 percent confidence interval:
 58.14796 90.51870
sample estimates:
mean of x
 74.33333
```

ということで、標本平均は 74.33 ですが、母平均値が 90kg である、という仮説は $p=0.055$ でぎりぎりですが棄却することができませんでした。そのため、母平均値が 90kg ということも否定できませんが、 p 値がぎりぎりであることに対応して、母平均値の 95%信頼区間も 58.1-90.5 ですから、たとえば母平均値が 91kg である、という仮説ですと棄却されることがわかります。そこまでは重くない集団のようです。

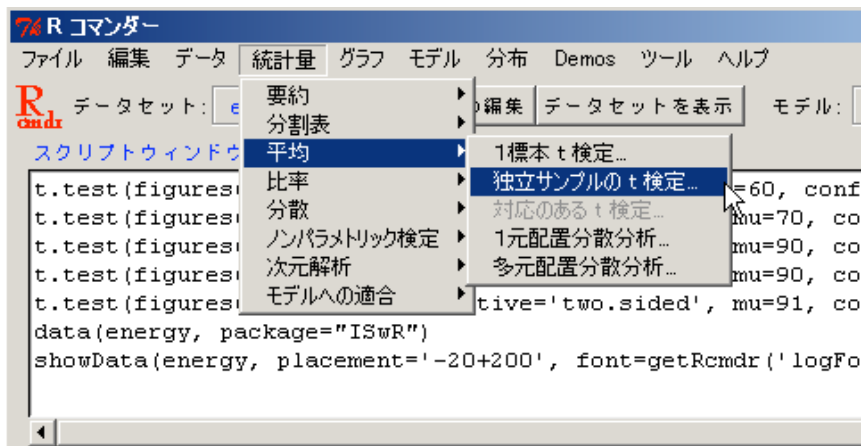
それでは、2 標本の場合、すなわち、平均値の差の検定に移りましょう。使うデータは ISwR パッケージ内の "energy" データセットです。「データ」→「パッケージ内のデータ」→「アタッチされたパッケージからデータセットを読み込む...」で、アクティブデータセットに指定しましょう。

energy		
	expend	stature
1	9.21	obese
2	7.53	lean
3	7.48	lean
4	8.08	lean
5	8.09	lean
6	10.15	lean
7	8.40	lean
8	10.88	lean
9	6.13	lean
10	7.90	lean
11	11.51	obese
12	12.79	obese
13	7.05	lean
14	11.85	obese
15	9.97	obese
16	7.48	lean
17	8.79	obese
18	9.69	obese
19	9.68	obese
20	7.58	lean
21	9.19	obese
22	8.11	lean

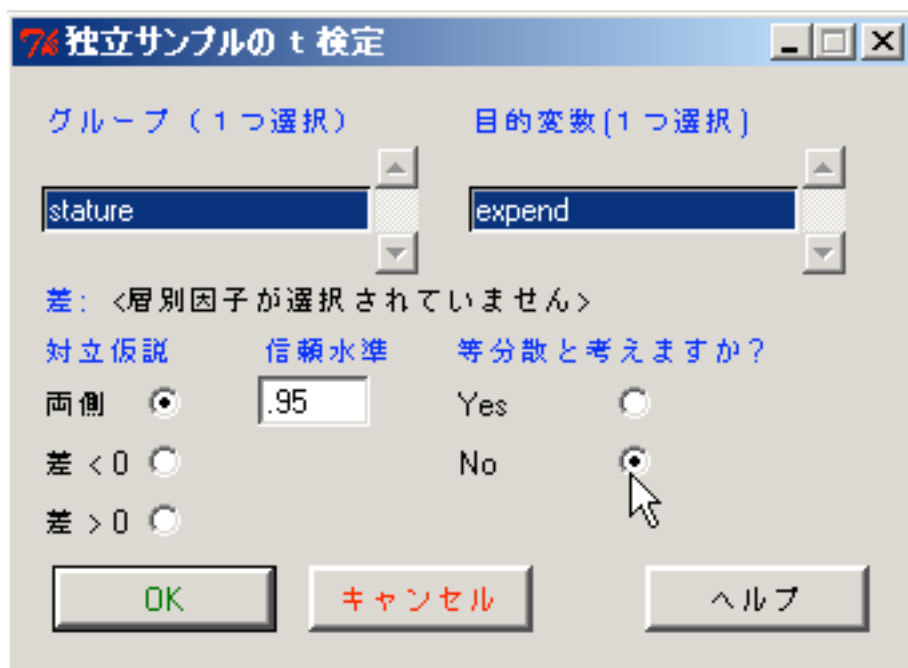
energy データセットは、やせている女性と太っている女性で、24 時間でのエネルギー消費量を測定したものです。変数 `expend` がエネルギー消費量（単位はメガジュール）、`stature` は体型で、すでに因子化されています。

やせている場合と太っている場合で、エネルギー消費量の平均値に差があるかどうかに興味深いところです。

2 標本の t 検定は、「統計量」→「平均」→「独立サンプルの t 検定...」です。



「独立サンプルの t 検定」ウィンドウがでますから、差をとりたい「グループ」をあらわす因子として、体型データの stature, 実際のエネルギー消費量をあらわす連続変数として、expend を指定します。グループ間で、測定値の母分散が等しいと考えられる場合は「等分散と考えますか？」に「Yes」をつけますが、ここでは必ずしもそうとは考えられないかなと思うので、「No」にしておきましょう。「Yes」の場合は Student の t 検定、「No」の場合は Welch の t 検定となります。



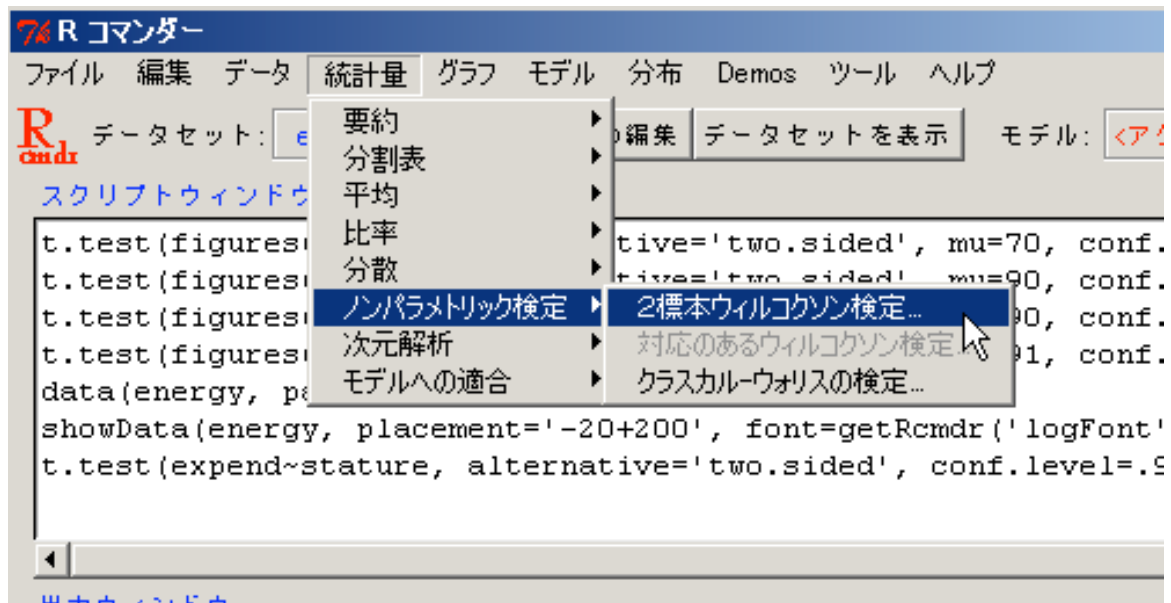
検定結果は出力ウィンドウです。

```
Welch Two Sample t-test

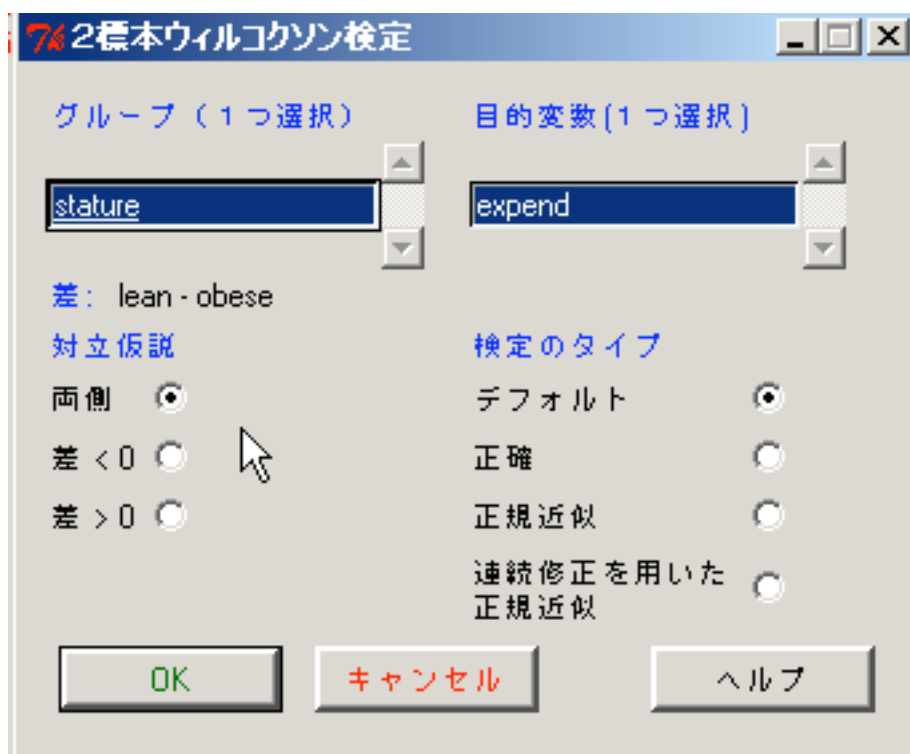
data:  expend by stature
t = -3.8555, df = 15.919, p-value = 0.001411
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.459167 -1.004081
sample estimates:
mean in group lean mean in group obese
      8.066154      10.297778
```

帰無仮説、「2群でエネルギー消費量の平均値には差はない」は、p 値 0.001 ということで棄却されました。すなわち、やせている場合と太っている場合でエネルギー消費量には差が認められたことになります。この結果には、各群のエネルギー消費量の標本平均値もでていきますね。それぞれ 8.07 と 10.3 です。

エネルギー消費量はおそらく正規分布すると思われますから、t 検定で問題ありませんが、母集団の年齢構成にそもそも偏りがあることなどがわかっており、エネルギー消費量が正規分布しないことが明らかな場合などには、t 検定を使うことは適切ではありません。このような場合、ノンパラメトリックな手法を用います。2 標本の平均値の差の検定におけるノンパラメトリック法は、Wilcoxon の順位和検定です。「統計量」→「ノンパラメトリック検定」→「2 標本ウィルコクソン検定...」にあります。



変数の指定方法は、t 検定の場合とほとんど同じですね。





結果は p 値が 0.002 と、t 検定のときに比べて若干大きくなっていますが、2 群が同じ母集団から抽出されたという帰無仮説は棄却され、やはりエネルギー消費量には 2 群で差があるという結論となります。ただ、t 検定のほうが正規分布を仮定しているため、データから利用できる情報量が多くなり、差の検出力があがります。したがって、p 値は t 検定を用いた場合のほうが小さくなるのです。

```
> wilcox.test(expend ~ stature, alternative="two.sided", data=energy)
```

```
Wilcoxon rank sum test with continuity correction
```

```
data: expend by stature
```

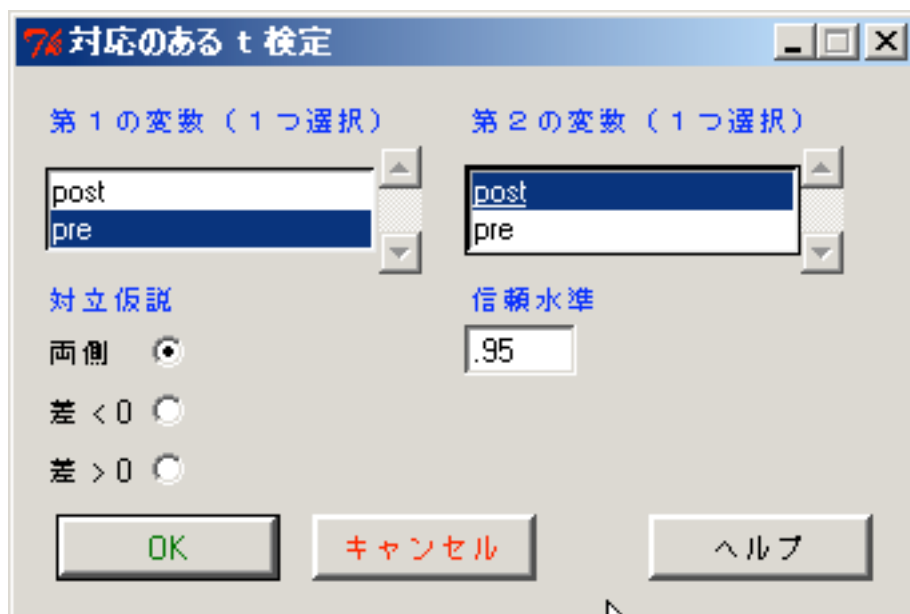
```
W = 12, p-value = 0.002122
```

```
alternative hypothesis: true location shift is not equal to 0
```

ではでは、次に「対応がある」データの場合に進みましょう。データは同じく ISwR パッケージにある "intake" です。

	pre	post
1	5260	3910
2	5470	4220
3	5640	3885
4	6180	5160
5	6390	5645
6	6515	4680
7	6805	5265
8	7515	5975
9	7515	6790
10	8230	6900
11	8770	7335

intake データは小さいものですが、11名の女性の月経前10日間と月経後10日間で、エネルギー換算した食事摂取量を比較したものです。月経前10日間の食事摂取量が変数 pre, 月経後10日間が post に格納されています。1人の女性について2回の測定が行われているので、pre 群と post 群は「対応がある」2群ということになります。対応のある2群の平均値の差の検定には、やはり t 検定という名前がついています。「統計量」→「平均」→「対応のある t 検定...」で呼び出してみましょう。



第1の変数、第2の変数として、それぞれの群のデータを示す変数を指定します。Ok ボタンを押せば、結果が出力ウィンドウに表示されます。

```

Paired t-test

data:  intake$pre and intake$post
t = 11.9414, df = 10, p-value = 3.059e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1074.072 1566.838
sample estimates:
mean of the differences
      1320.455

```

結果は通常の t 検定と似ていますが、"Paired" t-test となっていることに注意しましょう。

p 値は非常に小さく、月経前と月経後で摂取エネルギー量に差があったことが示されました。

一方、この検定もノンパラメトリック手法で行うこともできます。「統計量」→「ノンパラメトリック検定」→「対応のあるウィルコクソン検定...」です。

```

> wilcox.test(intake$pre, intake$post, alternative='two.sided', paired=TRUE)

Wilcoxon signed rank test with continuity correction

data:  intake$pre and intake$post
V = 66, p-value = 0.00384
alternative hypothesis: true location shift is not equal to 0

```

結果はやはり大きくは変化しませんが、p 値は 0.00384 と、t 検定に比べて大きくなっています。

5. 回帰分析

続いて、データを数理モデルにあてはめることで、変数間の直線的な関係の有無を明らかにする解析手法、回帰分析に入っていきます。散布図のところで若干、これに近いことをしましたが、回帰分析とは、

$$y_i = \alpha + \beta x_i + \epsilon_i$$

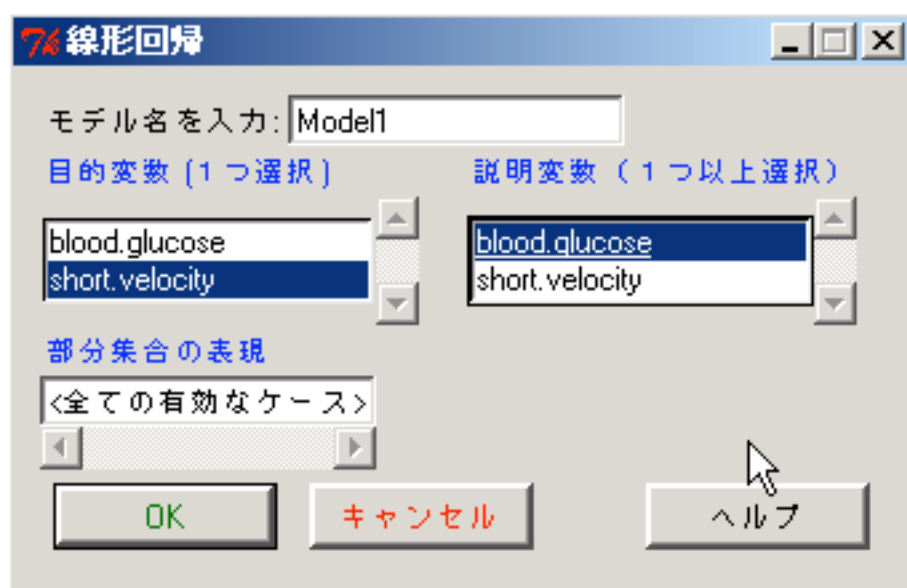
というように、x 変数で y を説明することが適切かどうかを検討することです。

それでは、R コマンダーで実際に回帰分析を行ってみます。今度のデータはやはり ISwR パッケージ、thuesen データセットです。

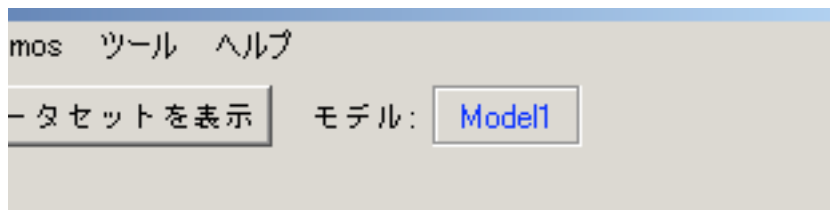
	blood.glucose	short.velocity
1	15.3	1.76
2	10.8	1.34
3	8.1	1.27
4	19.5	1.47
5	7.2	1.27
6	5.3	1.49
7	9.3	1.31
8	11.1	1.09
9	7.5	1.18
10	12.2	1.22
11	6.7	1.25
12	5.2	1.19
13	19.0	1.95
14	15.1	1.28
15	6.7	1.52
16	8.6	NA
17	4.2	1.12
18	10.3	1.37
19	12.5	1.19
20	16.1	1.05
21	13.3	1.32
22	4.9	1.03
23	8.8	1.12
24	9.5	1.70

データセットに含まれるのは blood.glucose と short.velocity という変数。これは、I 型糖尿病の患者 24 人における、空腹時血糖値と、心エコーの結果から求められた心室内周短縮速度の平均値を示しています。心室内周短縮速度は、心臓の収縮機能を示す指標の 1 つで、この変動が血糖値の変化で説明できるかが興味あるところです。このように「ある変数でほかの変数を説明できるかを検討する」際には、変数間の関係を数式で示した「数理モデル」を構築するのですが、R コマンダーではここでいよいよ、今まで赤いままだった「アクティブモデル」の出番となります。

thuesen をデータセットに指定後、「統計量」→「モデルへの適合」→「線形回帰...」を選びましょう。



数理モデルは、変数と同様に名前がつけられ、作業スペースに保存されます。「線形回帰」ウィンドウが出たら、モデル名を「Model1」として名付け、目的変数（従属変数）に short.velocity, 説明変数（独立変数）に blood.glucose を指定して Ok を押してください。



すると、出力ウィンドウに構築されたモデルの詳細が表示されるとともに、コマンドーの「モデル」欄にも、いま作成した「Model1」が入ります。

```
> summary(Model1)

Call:
lm(formula = short.velocity ~ blood.glucose, data = thuesen)

Residuals:
    Min       1Q   Median       3Q      Max
-0.40141 -0.14760 -0.02202  0.03001  0.43490

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.09781    0.11748   9.345 6.26e-09 ***
blood.glucose  0.02196    0.01045   2.101  0.0479 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2167 on 21 degrees of freedom
(1 observation deleted due to missingness)
Multiple R-Squared:  0.1737,    Adjusted R-squared:  0.1343
F-statistic: 4.414 on 1 and 21 DF,  p-value: 0.0479
```

Model1 は、このような数式であらわされます。

$$\text{short.velocity} = 1.10 + 0.0220 \times \text{blood.glucose}$$

出力結果の「Coefficients」のところで、回帰係数の推定値(Estimate)をみると、1.10 と 0.020 がでてますね。(Intercept)が α の値、すなわち切片で、blood.glucose のところが β となります。

α 、 β には、Estimate のほかにも、Standard Error、つまり推定値のばらつきも表示されます。推定値があまり大きくばらついているようなら、あまり信用のできない推定値ということになるでしょう。そして、t value と Pr(>|t|) は、それぞれの回帰係数が 0 である、という帰無仮説を t 検定で検定した結果の t 統計量と p 値を示しています。Pr(>|t|)のところが 0.05 未満であれば、帰無仮説は棄却され、回帰係数がゼロ（すなわち、その変数はまったく説明に寄与していない）とはいえないことになります。なお、切片のところでも一応 p 値がでていますが、切片がゼロかどうかはあまりモデルの解釈に影響しないので、それほど注目はされません。

このモデルがデータにどの程度よくあてはまっているのか？ということをお知らせするのが、Residuals、すなわち「残差」です。Min, 1Q, Median, 3Q, Max とありますが、これはデータの要約のところでもでてきた、最小、第 1 四分位点、中央値、第 3 四分位点、最大値です。回帰モデルは、この残渣の平均値がゼロになるように計算されているので、平均値は必ずゼロですからここには出てきません。モデルがよくあてはまるならば、残差はなるべく小さくなるべきであり、そして



正規分布すべきですが、そうなっているかをここで簡単ではありますが確認することができます。

Residual Standard error のところは、残差のばらつきですから、実際の観測値が回帰直線のまわりにどの程度ばらついているか、ということを示しています。これはあてはめた測定値(short. velocity)の単位に依存する値なので、一般的なモデル適合度の指標値ではありませんが、測定値に比べてあまり大きいとあてはまりはよくないことがわかります。

Multiple R-squared は、説明変数と目的変数の間のピアソンの相関係数を 2 乗したものですが、これは回帰によって説明される偏差平方和が、総平方和の何%であるか、という値と同じことになります。分散分析のところでもう少し詳しく説明しますが、実は分散分析の考え方と回帰分析の考え方は統一的にとらえることができるのです。ここでは、この Multiple R-squared(R^2)は全てのデータの変動のうちモデルによって説明できる割合を示していると理解してください。