統計分析ソフト R によるデータ分析マニュアル

Data Analysis Manual Using R

Version 1.1

国際開発センター

International Development Center of Japan (IDCJ)

佐々木亮／Ryo SASAKI, Ph.D.

このマニュアル作成は、一般財団法人国際開発センター（IDCJ）の公益目的計画の自主研究事業として実施されました。

© R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form.

はしがき

本書は、無料で使える統計分析ソフトRの解説書です。Rのインストールから始めて、基本的な統計データ分析の操作の仕方を解説しました。初心者でも必ずわかるように、操作の仕方を丁寧にステップ・バイ・ステップで説明しています。だれでも、この解説書に沿って操作すれば、ひととおりのRの操作方法が身につくように工夫されています。ぜひトライしてみてください。

佐々木亮／Ryo SASAKI, Ph.D.

なお、さらに深く統計分析について学んでみたい方は、以下の講座を受講することをお勧め致します。「だれでも必ずわかる」をキャッチフレーズにしており、3日間で統計分析の技術がひととおり身につくように工夫されています。
＜『プロフェッショナル統計分析ワークショップ』: https://www.idcj.jp/seminar＞

＜謝辞 1＞ この研修教材は、一般財団法人国際開発センターの公益目的計画の自主研究事業の一環で作成されました。この場を借りて御礼申し上げます。
＜謝辞 2＞ R および R コマンダーに関しては、三好崇弘様（有限会社エムエム・サービス代表）より有益なご助言をいただきました。この場を借りて御礼申し上げます。

著者紹介

佐々木亮／Ryo SASAKI, Ph.D.

<現在の仕事>
国際開発センター評価部主任研究員
International Development Center of Japan (IDCJ)

<学位>
米国ウェスタンミシガン大学評価研究所 評価学博士（2008）
米国ニューヨーク大学ワグナー公共行政大学院 公共行政学修士（公共政策分析）（1996）
立教大学法学部 法学士(1991)

<連絡先>
Email: sasaki.ryo(a)idcj.or.jp

目　次

はしがき・著者紹介　i

第1章　統計分析ソフト R のインストール　1

1.　統計分析ソフト R の紹介 ＞
2.　R 本体のインストール
3.　RStudio のインストール
4.　R コマンダーのインストール

第 2 章　平均値と標準偏差　18

第 1 セッション：平均値と標準偏差：データの特徴を把握する

第 3 章　t 検定　34

第 2 セッション::　事前・事後の t 検定（対応のあるt検定）：保護者の満足度は向上したか？1:過去の満足度と比較してみよう。
第 3 セッション：　2群の t 検定（対応のない t 検定）：保護者の満足度は向上したか？2: :隣の学校の保護者満足度と比較してみよう。

第 4 章　比率の検定　56

第 4 セッション：　カイ二乗検定：適合度検定（Goodness-of-fit test）：住民サーベイの回答「はい」「いいえ」の比率に差があったと言えるのか？
第 5 セッション：　カイ二乗検定：独立性の検定（Chi-Square Test of Association (or Independence)）：教育法の違いで卒業試験の結果に差があったと言えるのか？

第 5 章　回帰分析　67

第 6 セッション：　回帰分析1：　クロスセクションデータを用いる
第 7 セッション：　回帰分析2：　時系列データを用いる
第 8 セッション：　プロビット回帰分析：「はい」「いいえ」の回帰分析

参考文献　103
あとがき　104

第1章　統計分析ソフト R のインストール

詳細目次

１．統計分析ソフト R の紹介　1

2. R 本体のインストール　2

最初に R 本体をインストールします。

3. RStudio のインストール　9

RStudio は R を見やすくするための補助ソフトです。

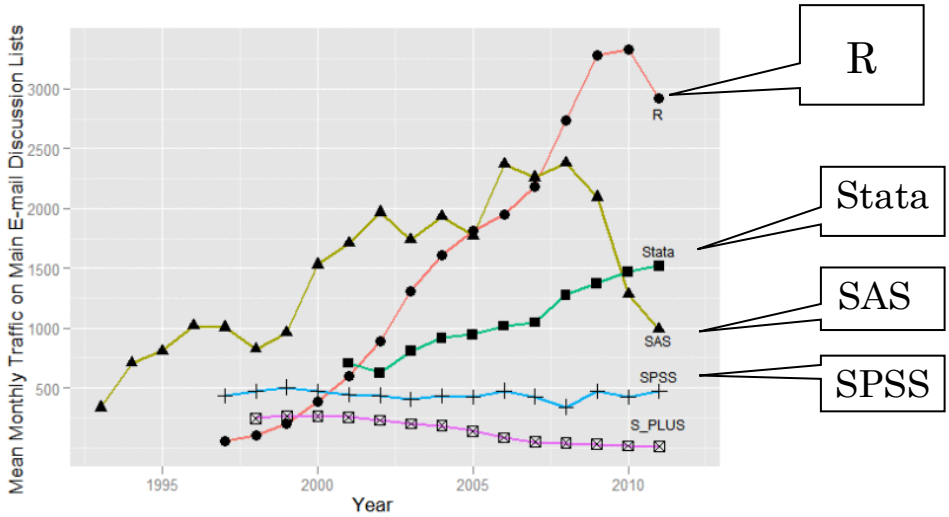
4. R コマンダーのインストール　15

R コマンダーは、エクセルのような分かりやすいメニューを提供する補助ソフトです。

1. 統計分析ソフト R の紹介

R は、無料かつオープンソースの統計分析ソフトです。以前（私の大学院時代）は、統計分析ソフトと言えば、SAS あるいは SPSS の二者択一でした。近年になって 1985 年に開発された STATA がシェアを伸ばしており、経済学や国際開発などの分野で広く使わるようになっています。しかしいずれも 15 万円～30万円の購入価格となっており、個人や零細企業では手が出しづらい価格設定となっています。これに対して、近年、無料の統計ソフトである R が普及しているという話をよく耳にします。統計分析ソフトの市場占有率を示すデータの一つとして以下のグラフがウェブで紹介されていました。現在は、R の命令コードを SAS の命令コードに変換できる機能を SAS が実装しました。SPSS も同様の機能を実装したと言われています。これからますます R が普及していくでしょう。

図 1 ： データ分析ソフトウェアのポピュラリティ

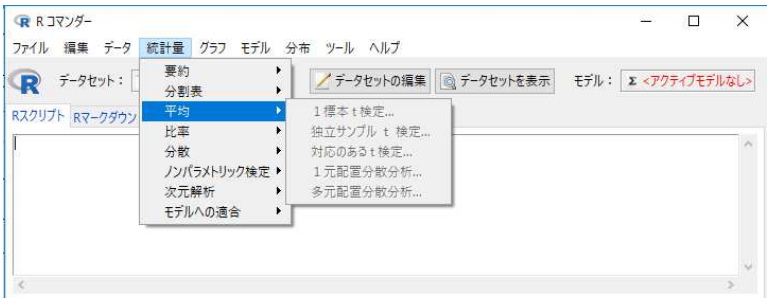


（出所）プリンストン大学 https://blogs.princeton.edu/onpopdata/2012/10/05/popularity-of-data-analysis-software/

R は 1996 年にニュージーランドのオークランド大学の Ross Ihaka と Robert Clifford Gentleman により作られました（その Ross の頭文字をとって R と呼称していると理解されます）。オープンソースで誰でも無料で使えますが、命令をテキストベースで打ち込むスタイルなので、長らく普及しないと見ていました。

しかし近年、R Commander なる追加モジュールが公開されて、エクセル風に操作することが可能となったことを知りました。これにより、自動的に別窓に命令コードが生成されていきます。これからますます R が普及してくでしょう。

図 2 ： R によるドロップダウンメニュー（t 検定の例）

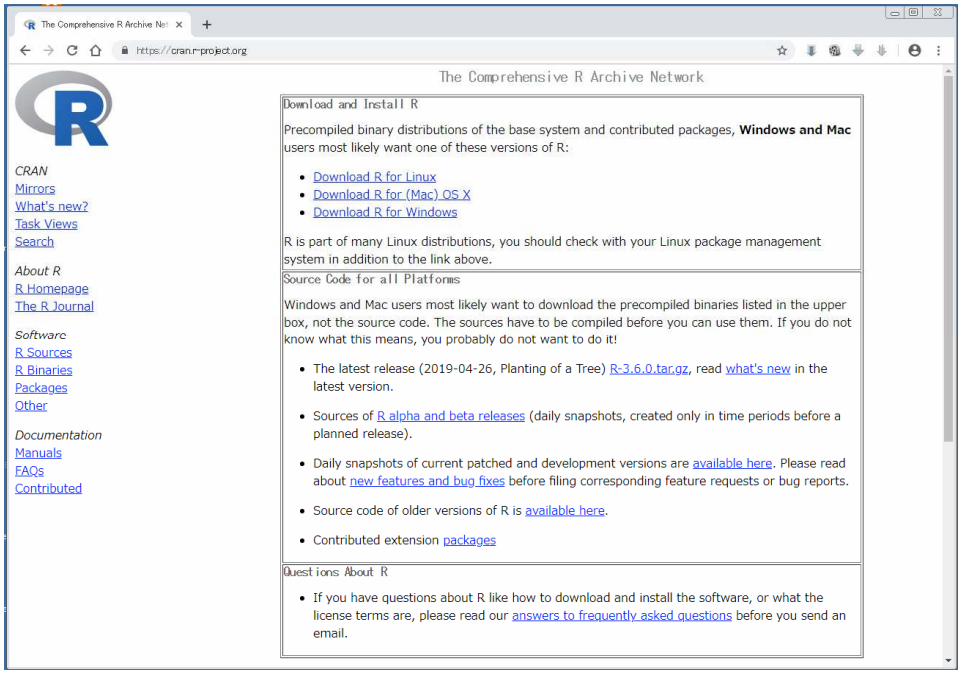


2. R 本体のインストール

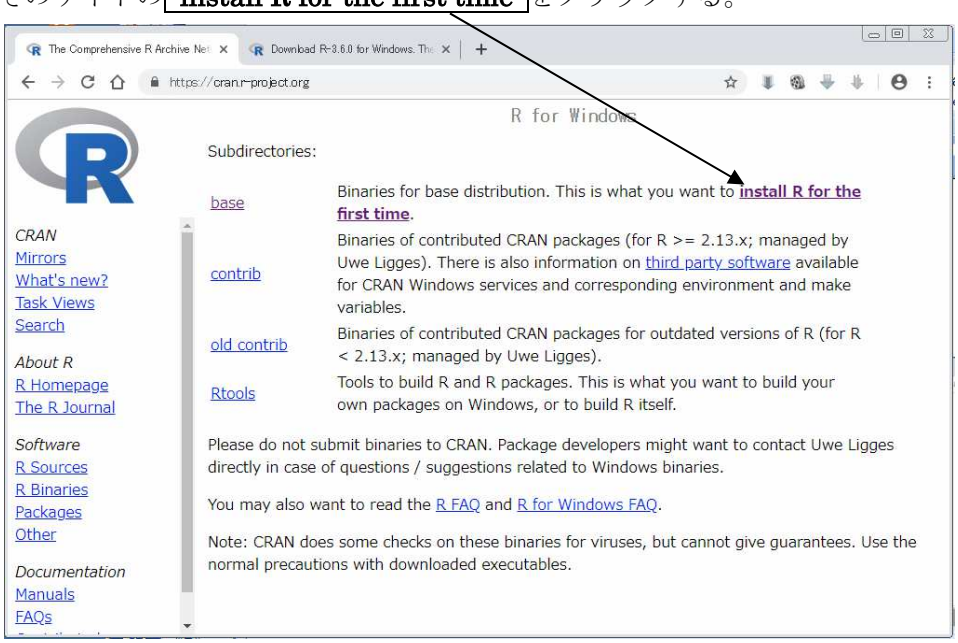
最初に R 本体をインストールします。

1. 以下のサイトにアクセスする。すると以下のようなサイトが表示される。

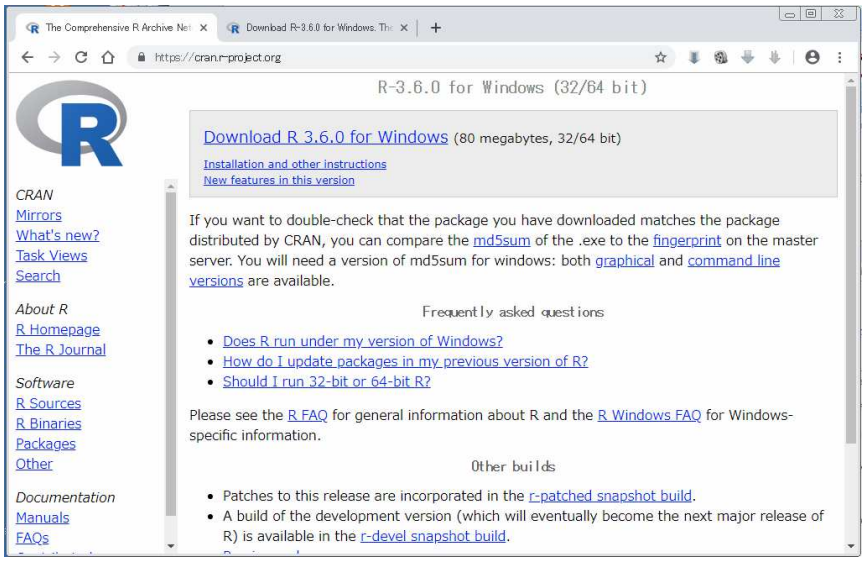
   https://cran.r-project.org/



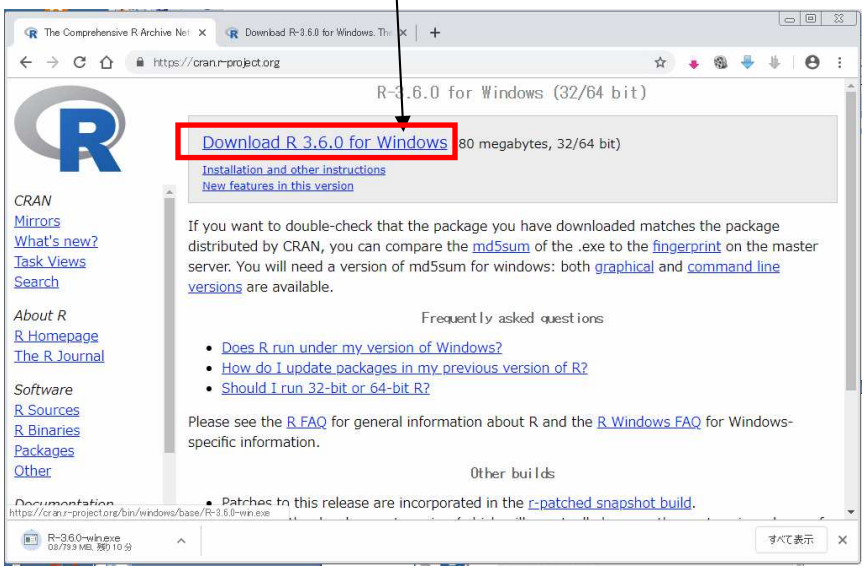
2. そのサイトの **install R for the first time** をクリックする。



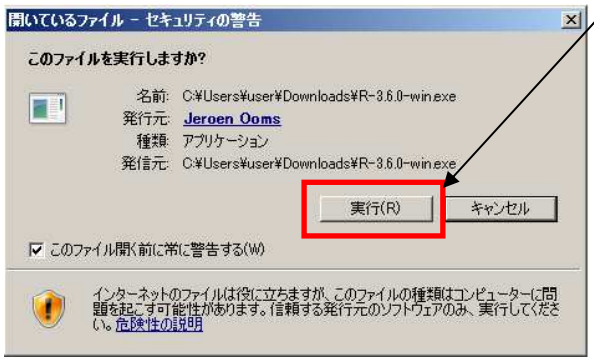
3. すると下のように表示される。



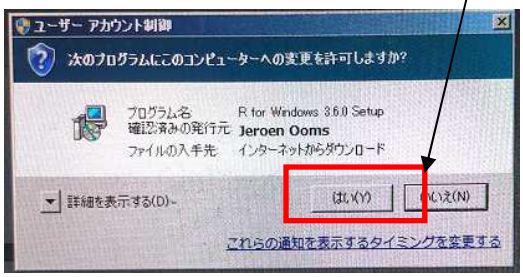
4. 一番上の Download R.3.6.0 for Windows をクリックする。（3.6.0 より新しいバージョンになっているかも知れません）



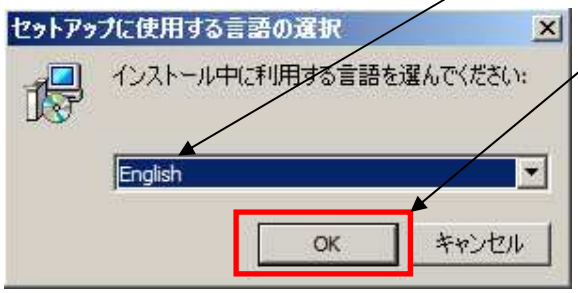
5. すると下のようなボックスが現れる。 実行(R) をクリックする。



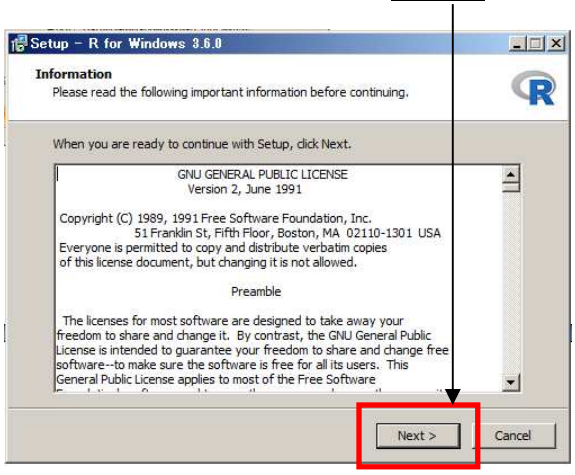
6. 以下のようなボックスが現れる。 はい をクリックする。



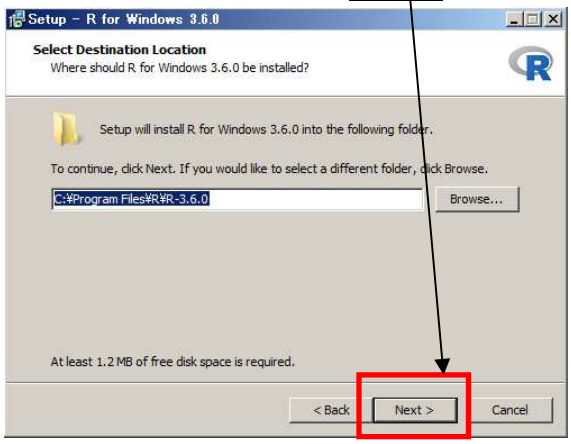
7. 以下のようなボックスが現れる。 English を選ぶ。 OK をクリックする。



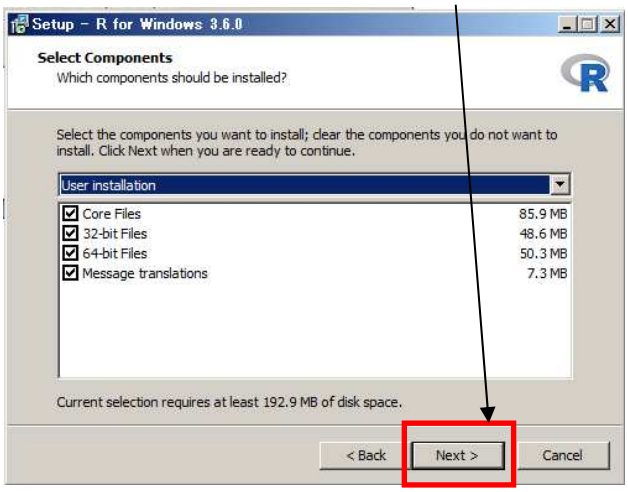
8. 以下のようなボックスが現れる。 Next ＞をクリックする。



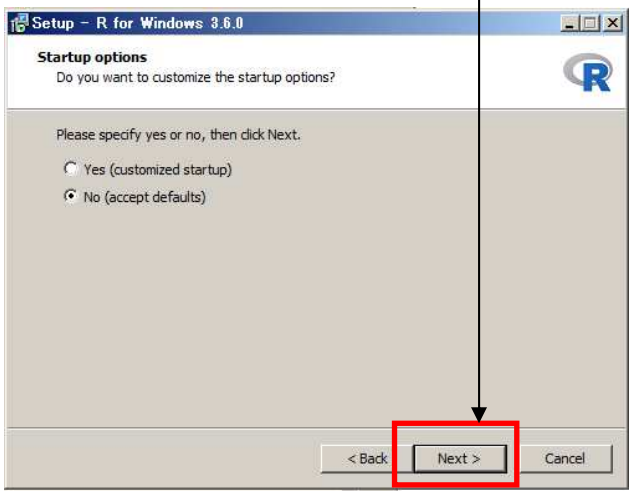
9. 以下のようなボックスが現れる。 Next ＞をクリックする。



10. 以下のようなボックスが現れる。 Next ＞をクリックする。

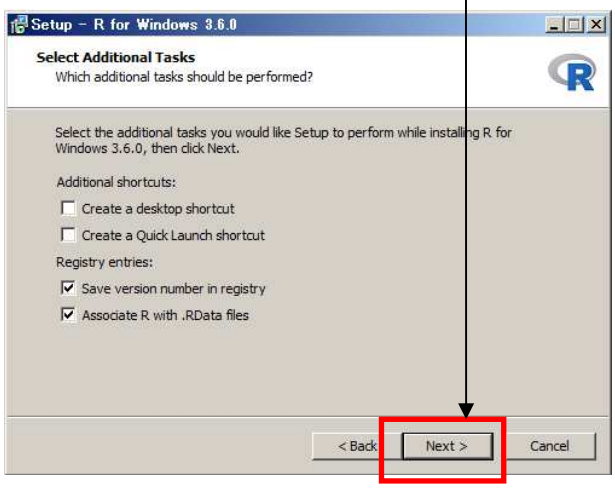


11. 以下のようなボックスが現れる。 Next ＞をクリックする。

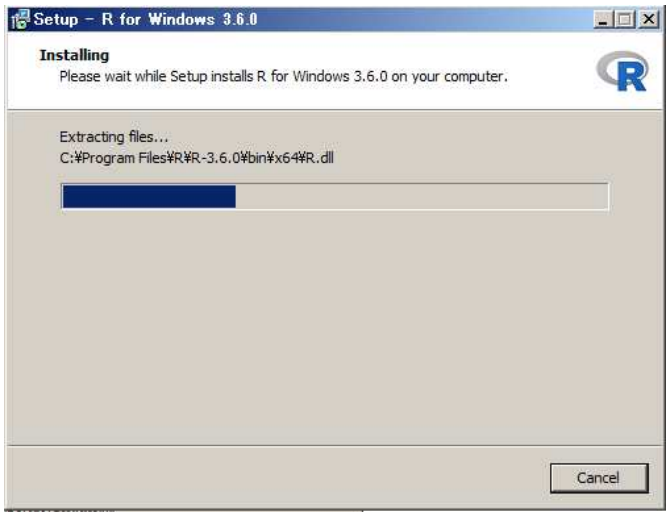


12. 以下のようなボックスが現れる。 Next ＞をクリックする。

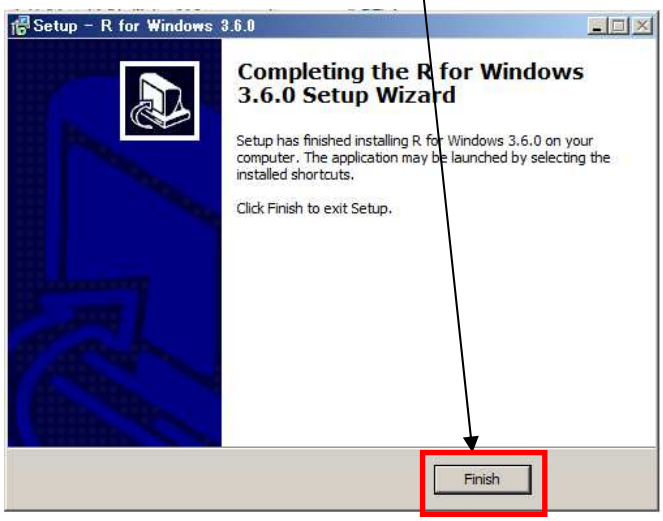
13. 以下のようなボックスが現れる。 Next＞ をクリックする。



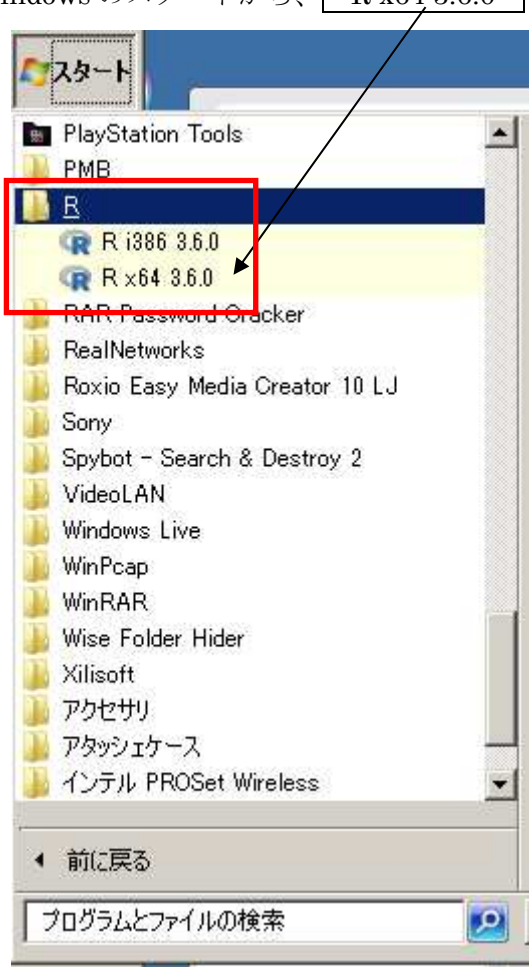
14. インストールが始まる。インストールが終わるまで待つ。



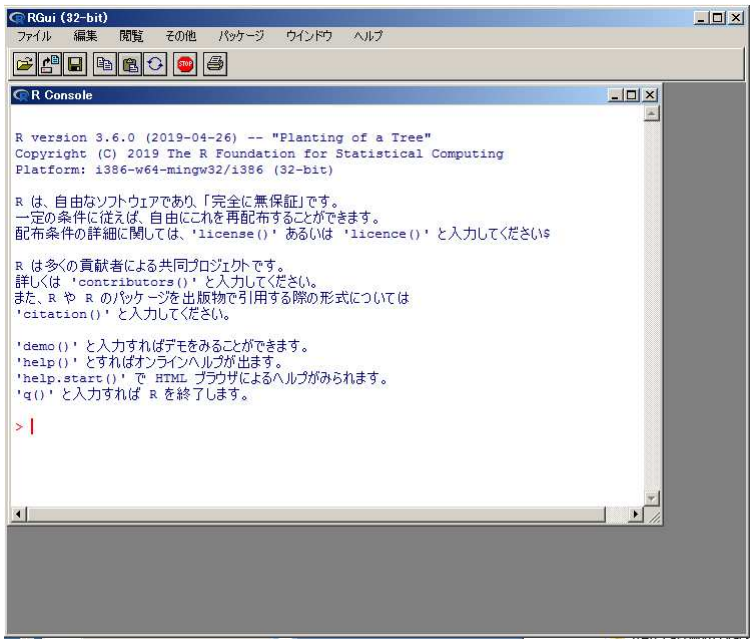
15.インストールが終わったら、 Finish をクリックする



16. Windows のスタートから、 R x64 3.6.0 をクリックする。



17.　以下のようなウィンドウが開く。R 本体の起動に成功した。

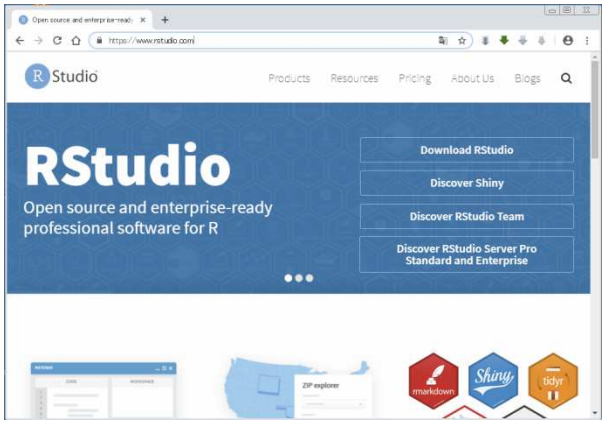


3. RStudio のインストール

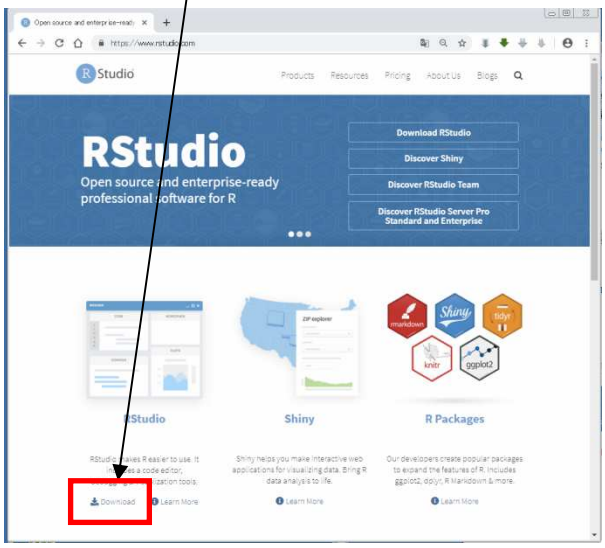
続いて、「Rstudio」をインストールします。Rstudio は R を見やすくするための補助ソフトです。

1. 以下のサイトにアクセスする。すると以下のようなサイトが表示される。

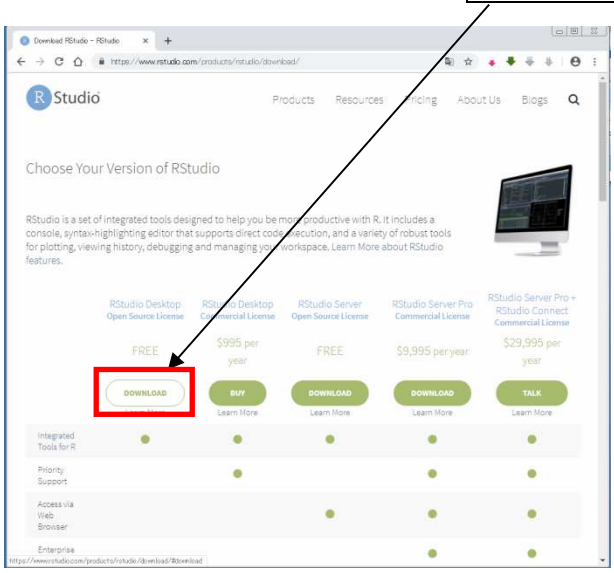
https://www.rstudio.com/



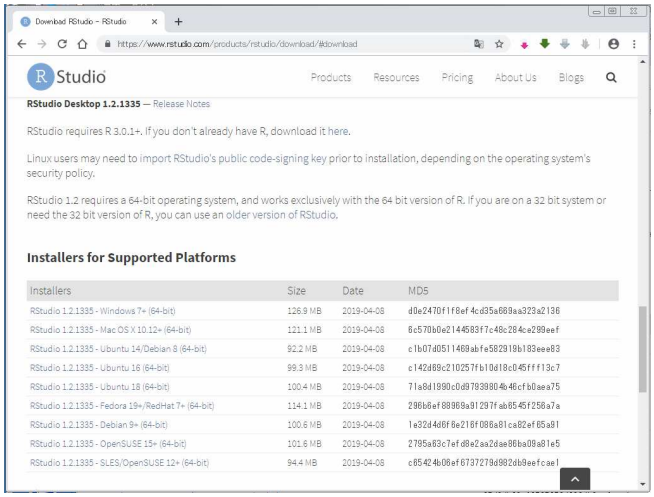
2. RStudio の Download をクリックする。



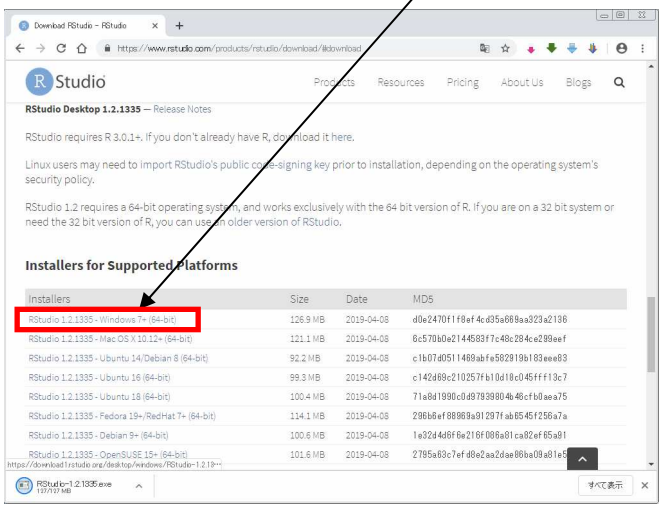
3. すると以下のサイトが表示される。 DOWNLOAD をクリックする。



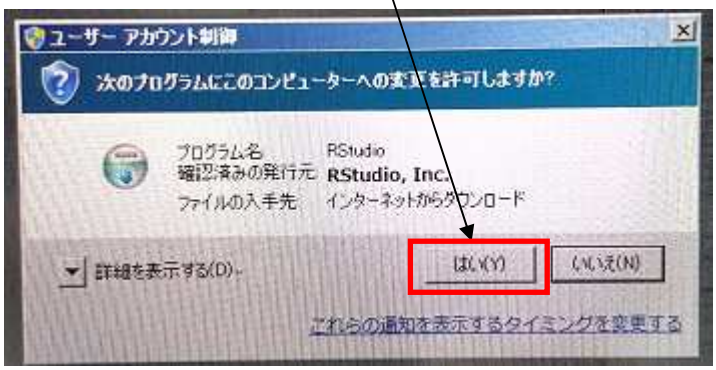
4. すると次のようなサイトが表示される。



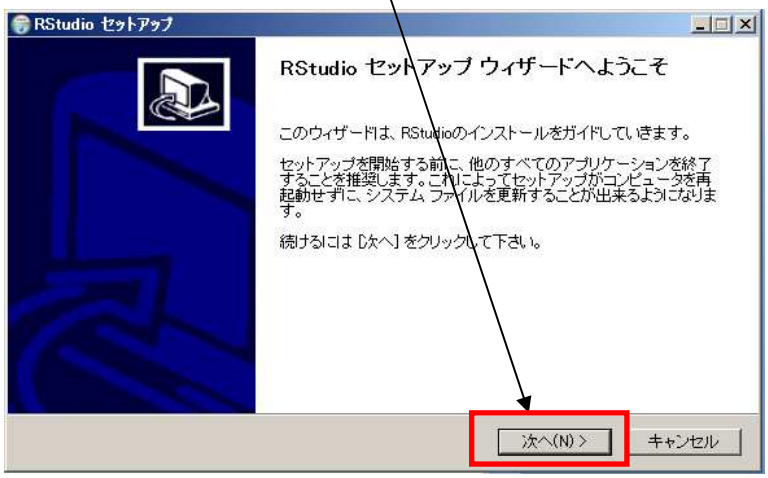
5. 一番上の RStudio 1.2.1335 Windows7+ (64-bit) をクリックする。（あなたがダウンロードするときは、これよりバージョンナンバーが上がっているかも知れない。）



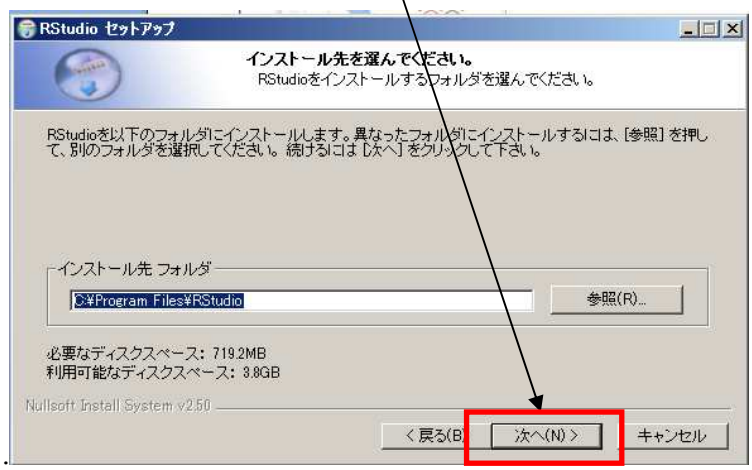
6. 以下のようなボックスが現れる。 はい をクリックする。



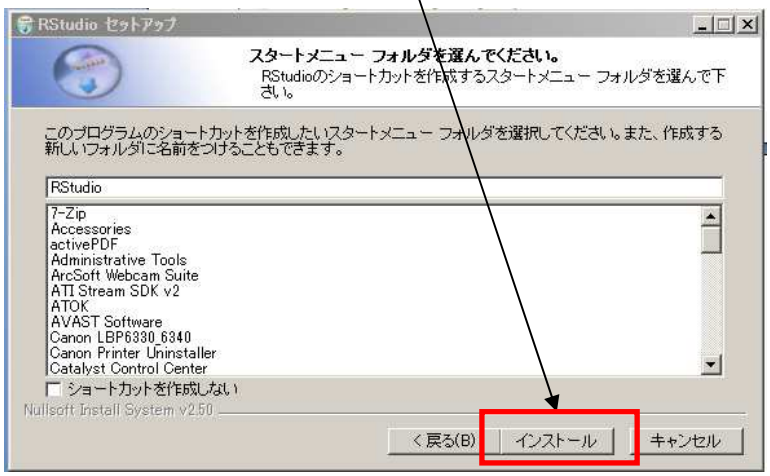
7. 以下のようなボックスが現れる。 次へ> あるいは Next> をクリックする。



8. 以下のようなボックスが現れる。 次へ> あるいは Next> をクリックする。

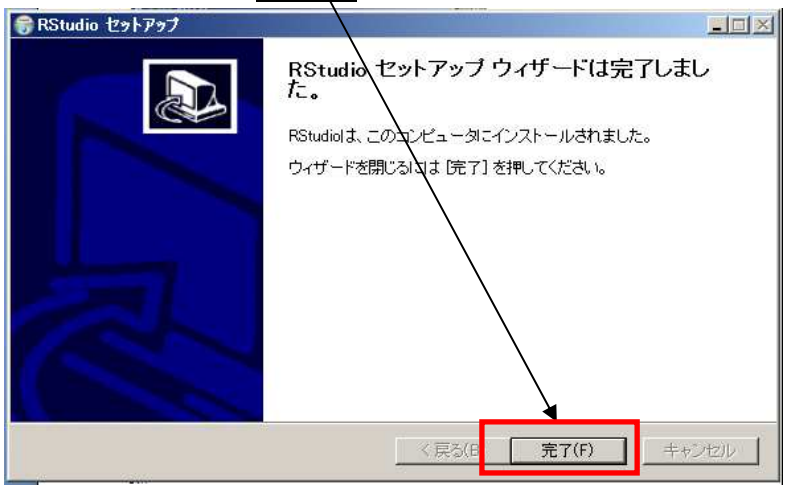


9. 以下のようなボックスが現れる。 インストール をクリックする。

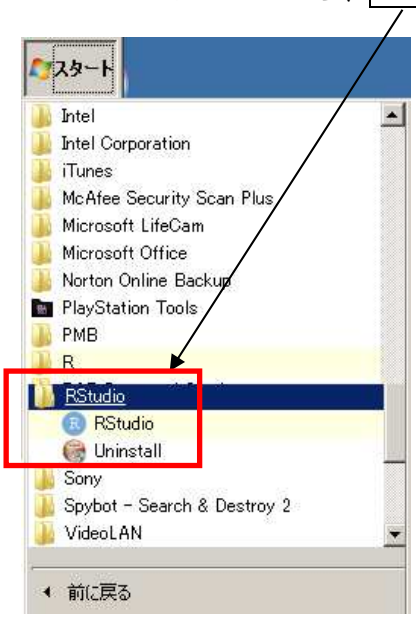


10. インストールが始まる。 30-40 秒 ほど待つ。

11. インストールが終わる。 終了 を押す。

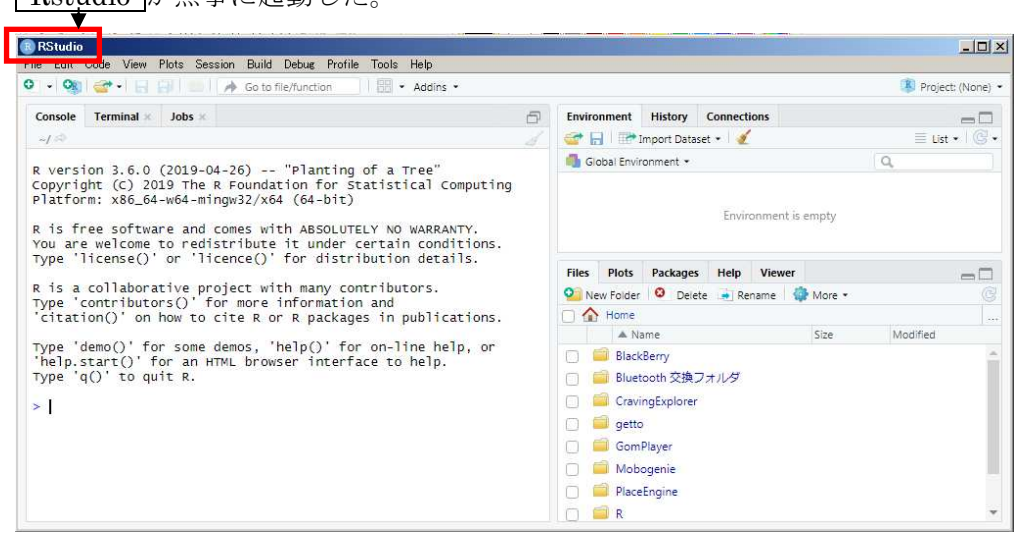


12. Windows のメニューから、Rstudio をクリックする。



13. 下のようなウィンドウが開く。

Rstudio が無事に起動した。

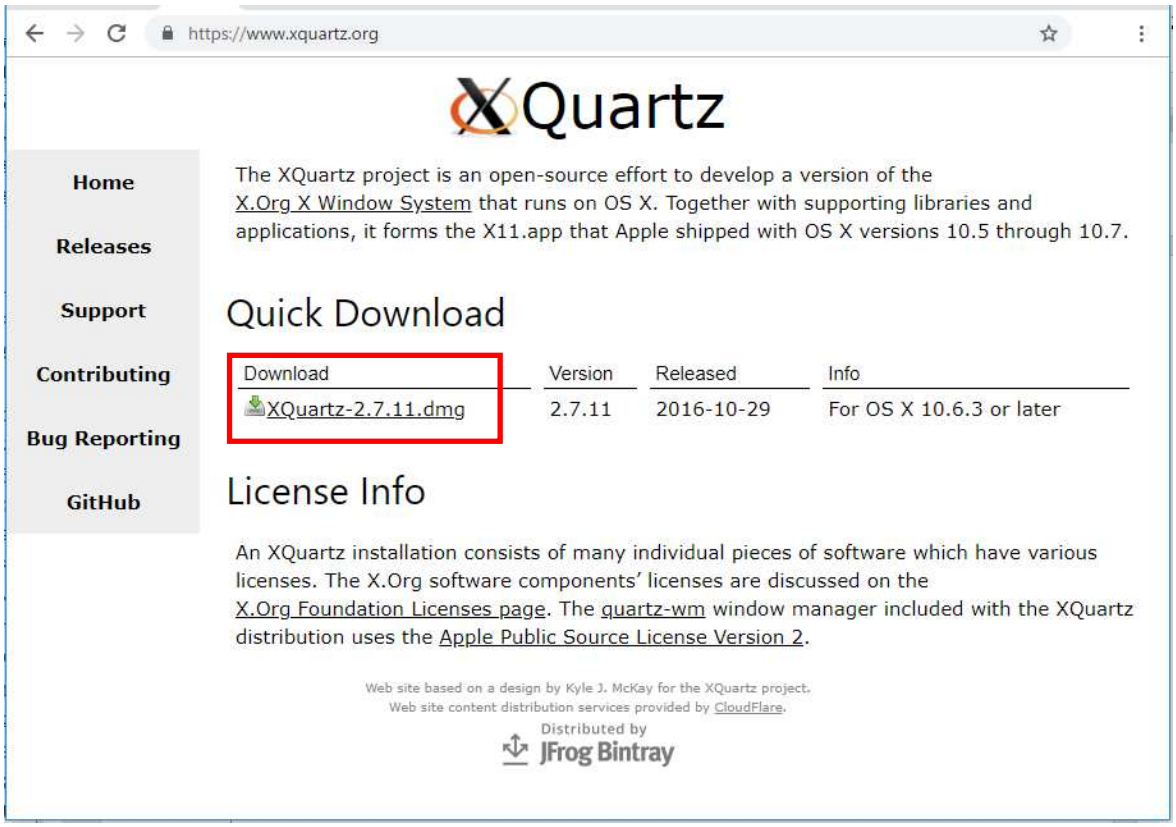


マックをお使いの方はここで以下の操作をしてください。
（Windows の方は必要ありません。）

① 以下のサイトにアクセスします。

https://www.xquartz.org/

② そのサイトから XQuarts-2.7.11.dmg ファイルをダウンロードします。（バージョン名はこのインストールマニュアル作成時の 2.7.11 より最新になっているかも知れません。）



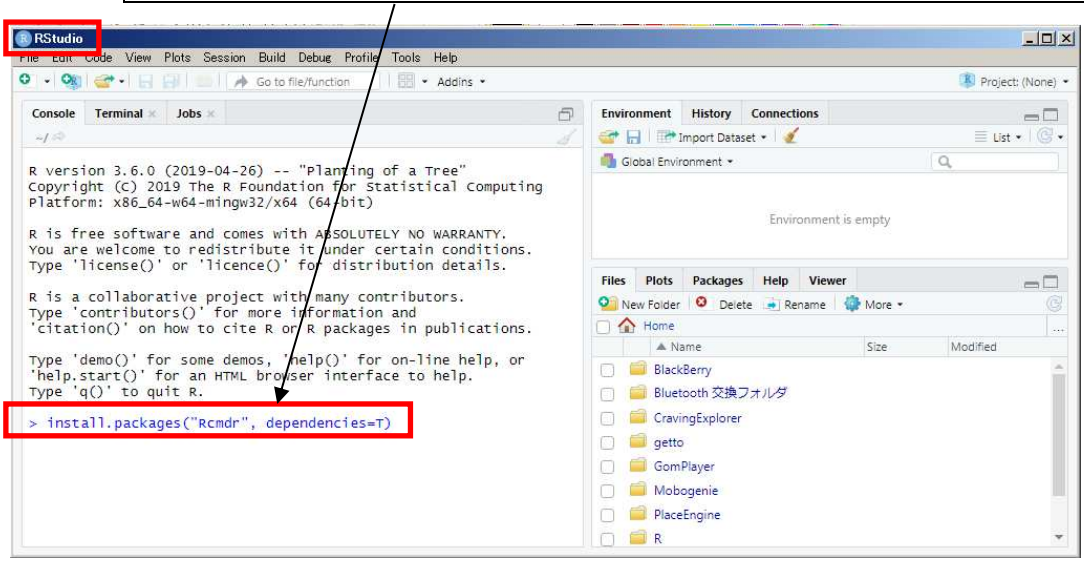
③ ダウンロードしたファイルをインストールします。
④ マックを再起動します。
その後、次のページからの R コマンダーのインストールを実施してください。

4. R コマンダーのインストール

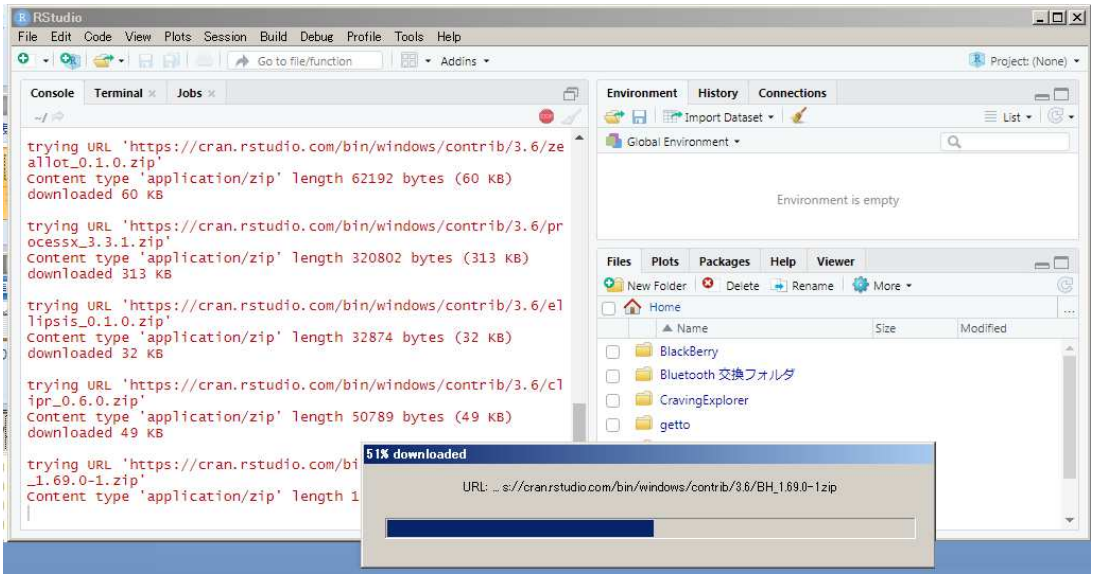
最後に、「R コマンダー」をインストールします。R コマンダーは、エクセルのようなプルダウンメニュー（GUI）を提供する補助ソフトです。コマンドを手でタイプする必要がなくなります。

1. 前ページ起動させた Rstudio で、以下のようにタイプする。

```
> install.packages("Rcmdr", dependencies=T)
```

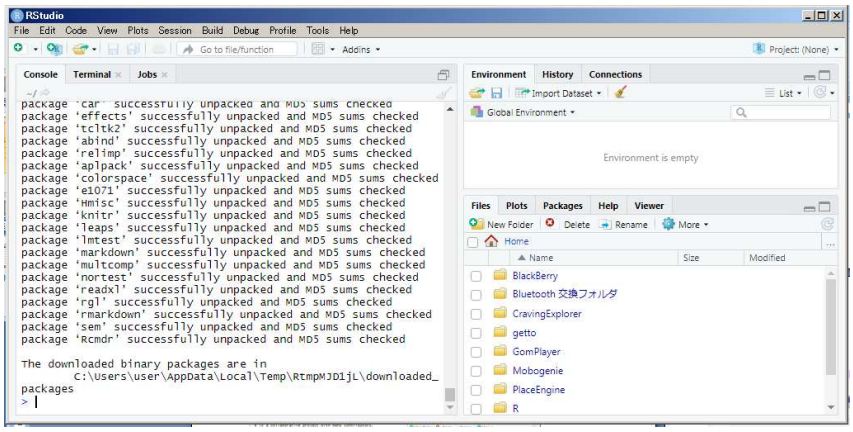


2. 赤い字でインストールが始まる。 **5-10 分** ほどインストールし続けるので、我慢強く待つ。

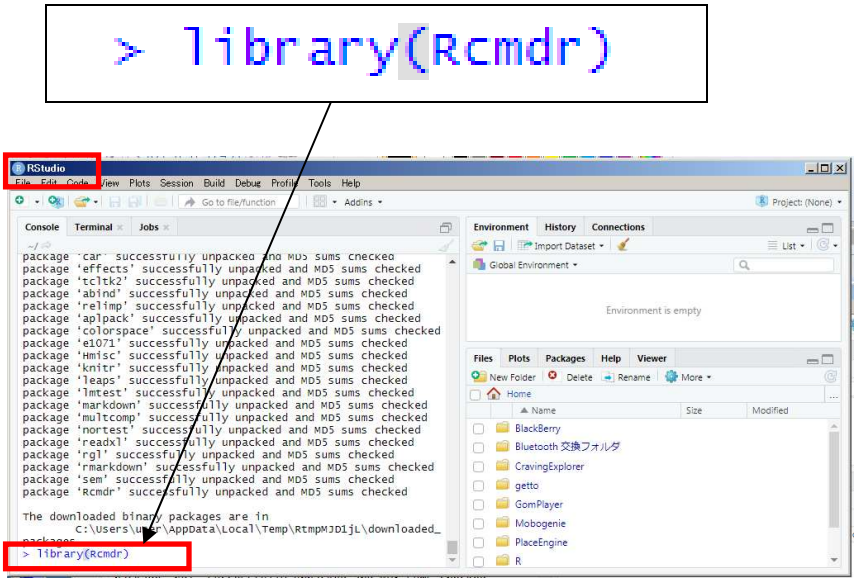


.

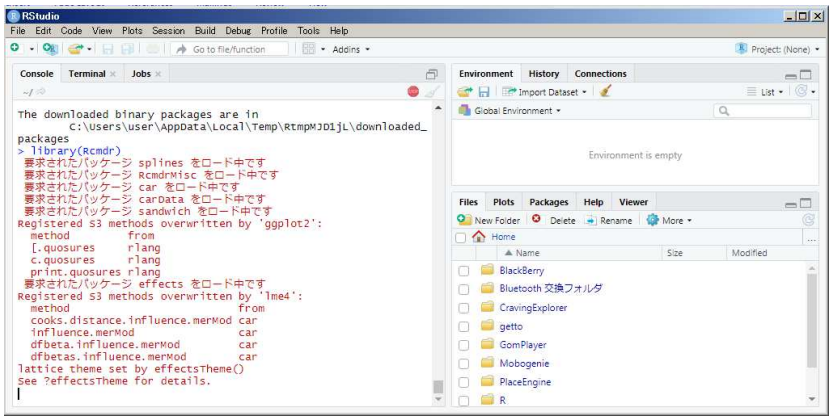
3. インストールが終わる。



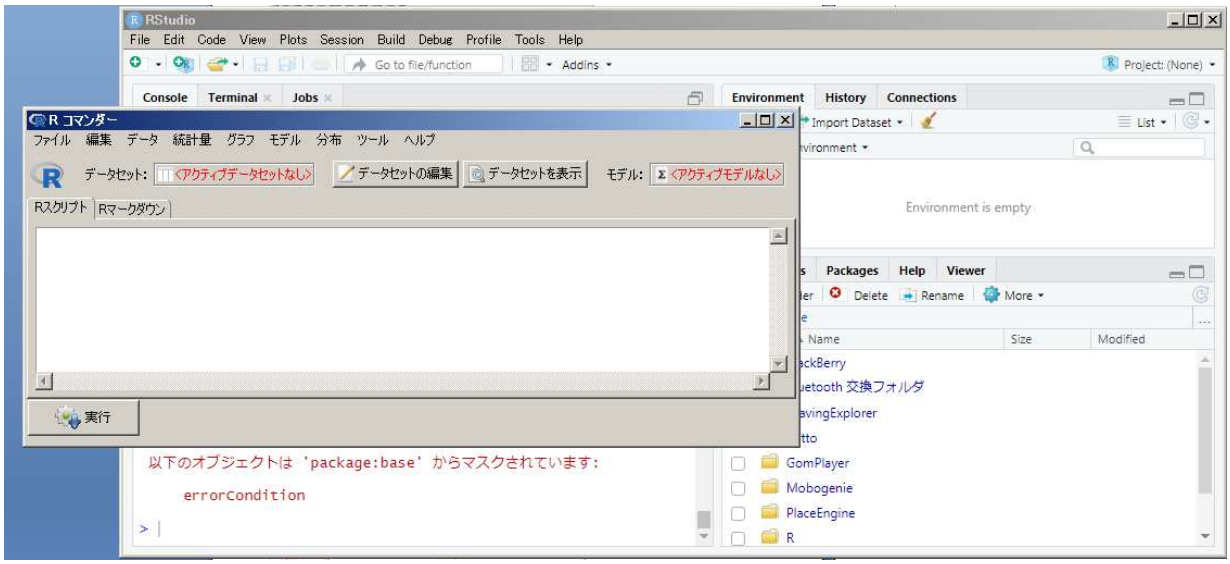
4. Rstudio で以下のようにタイプする。



5. 自動的に赤い字で 15 行ほど走って、R コマンダー が立ち上がる。



6. R コマンダー が立ち上がった。Rstudio の上にかぶさるようにして、R コマンダー のウィンドウが立ち上がっているのが分かる。



おめでとうございます！！

これで各種の統計分析ができるようになりました。

第 2 章　平均値と標準偏差

詳細目次

第 1 セッション　19

平均値と標準偏差：データの特徴を把握する

第１セッション:

平均値と標準偏差:

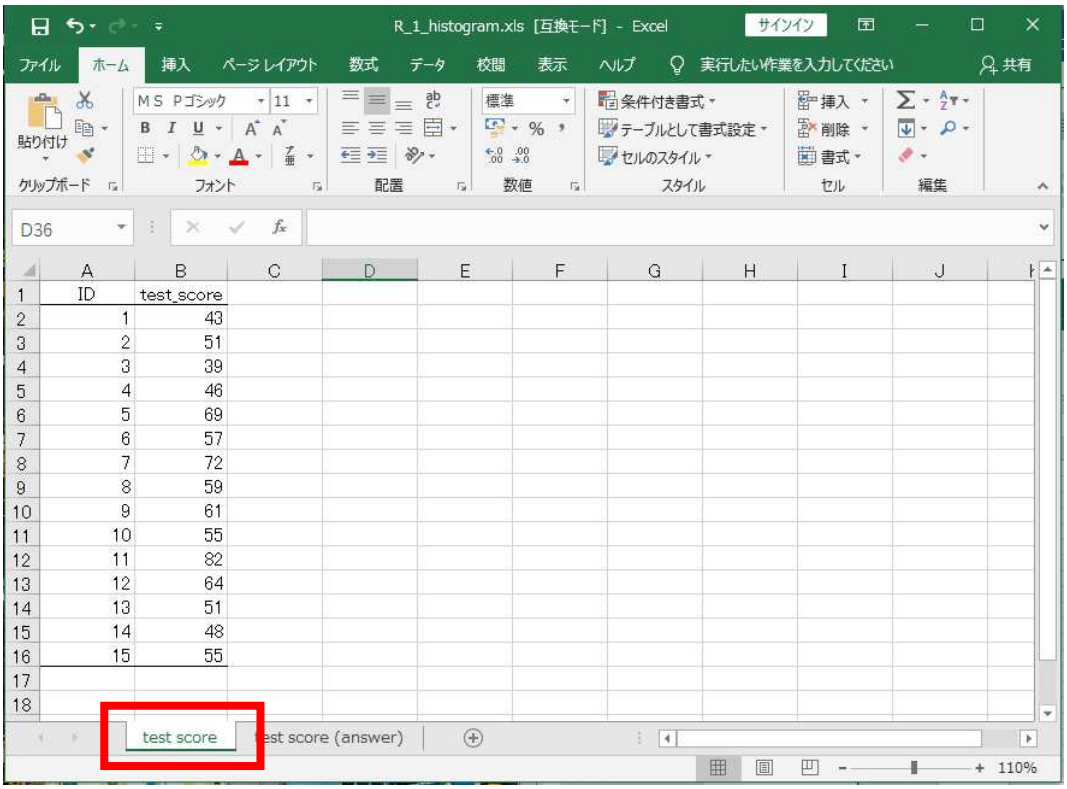
データの特徴を把握する

Q. 日本のある NGO は「革新的な学校運営プログラム」の普及を支援している。この「革新的な学校運営プログラム」を導入している A 小学校において、あるクラスの学力テストの点数が集計された。 満点は 100 点満点である。A 小学校の校長先生は、集計されたデータの基本的な特徴を知らねばならない。とくに以下の特徴である。

平均値と中央値はいくらか？
大半のデータが入る範囲はどこからどこまでか？

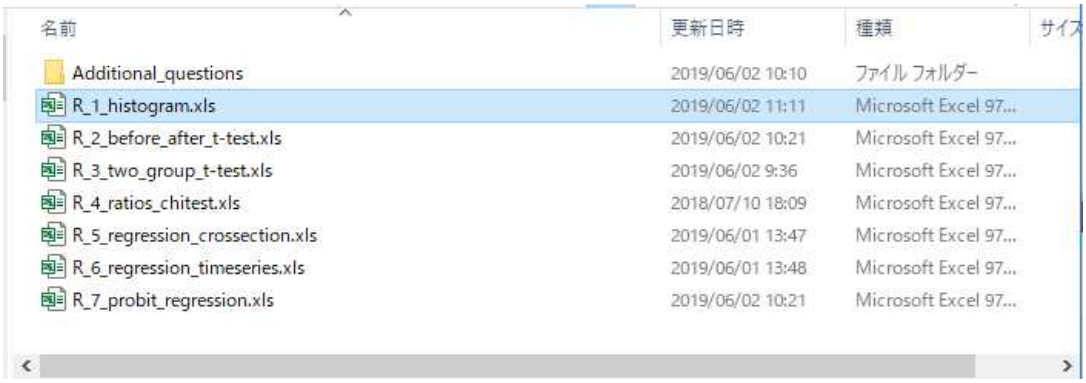
［1］ヒストグラムを作る

●以下のようにデータを入力する。タブに「test_score」などの名前をつける。



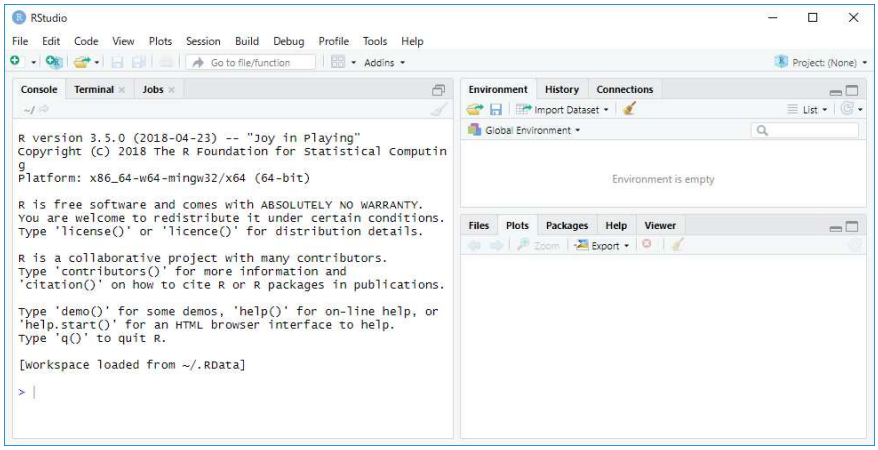
ID	test_score
1	43
2	51
3	39
4	46
5	69
6	57
7	72
8	59
9	61
10	55
11	82
12	64
13	51
14	48
15	55

●デスクトップに、R_1_histogram.xls などと適当な名前をつけて保存する。

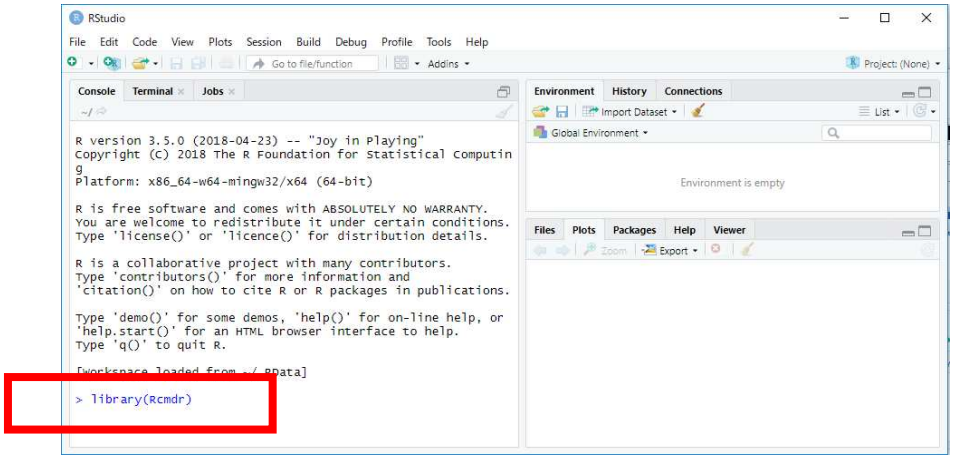


● Windows のメニュー＞“RStudio”を選ぶ。

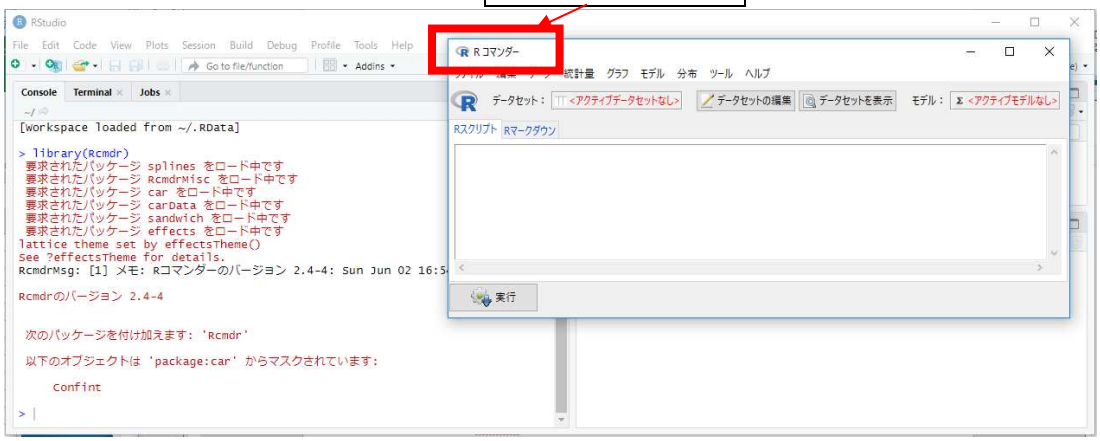
● 下のようなウィンドウが開く。



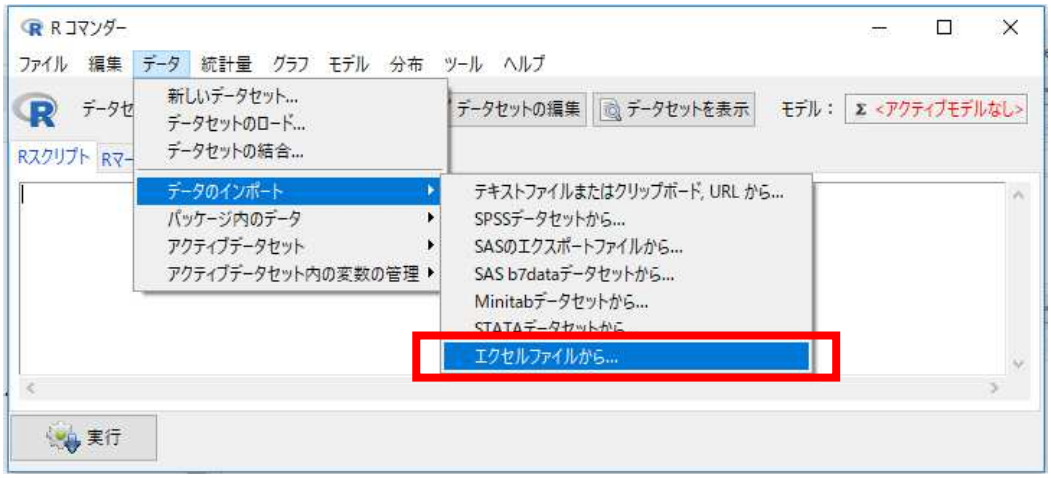
● ” library(Rcmdr) “と入力する。＞Enter を押す。



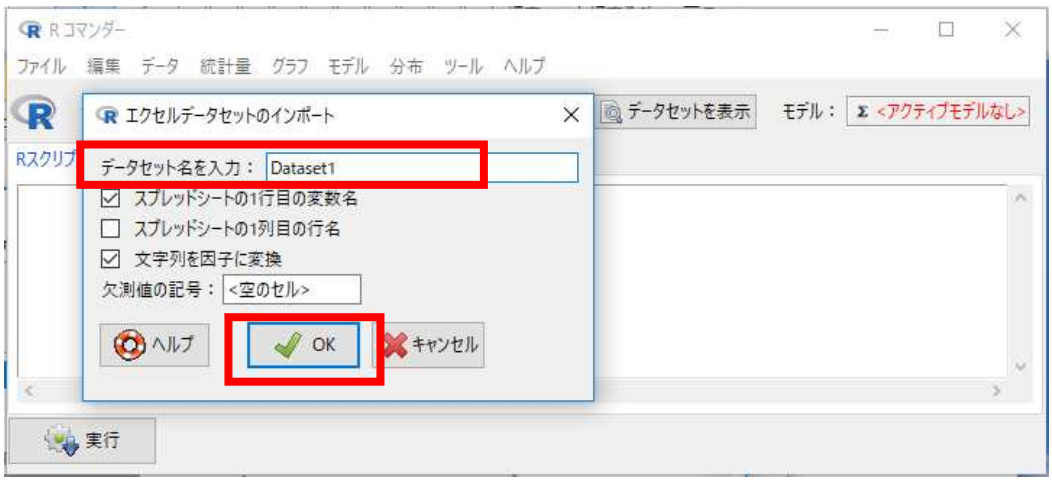
●下のようなメッセージが流れて、 R コマンダー のウィンドウが開く。



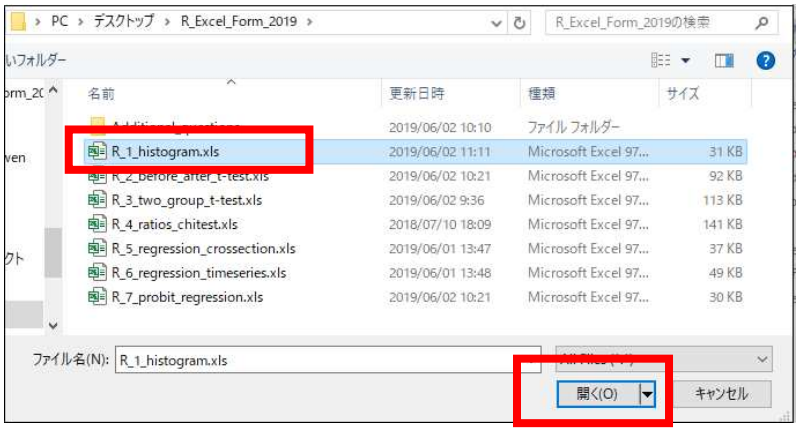
● R コマンダー のウィンドウで、データ ＞ データのインポート ＞ エクセルファイルから を選ぶ。



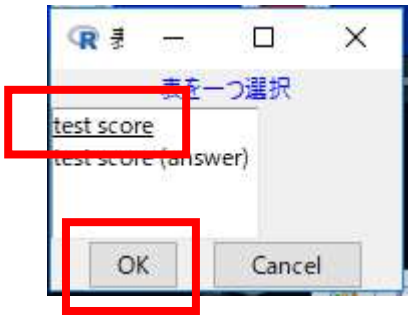
●現れたウィンウィンドウで、“データセットを入力” に “Dataset1” と入力する。＞ OK を押す。



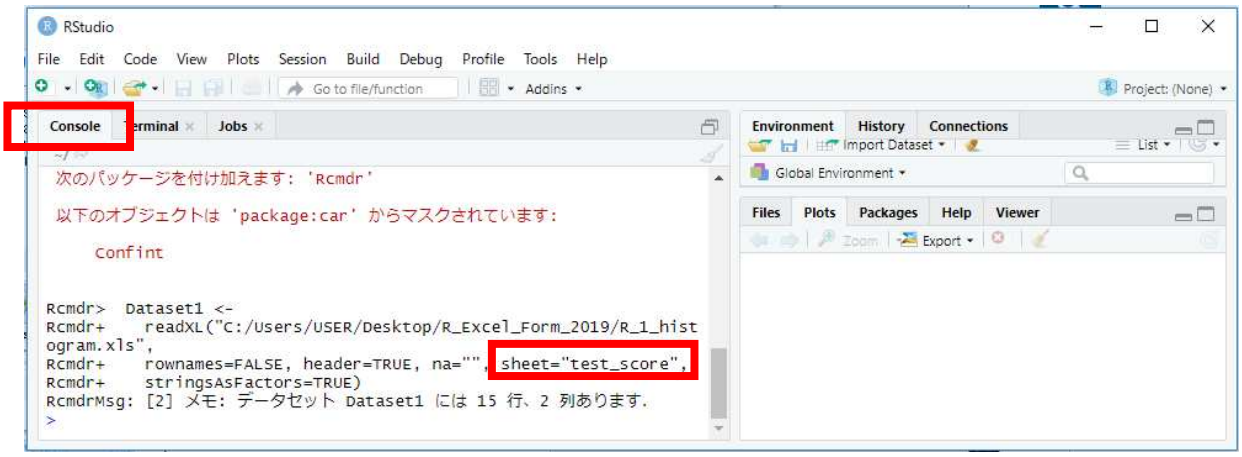
●さきほど保存した Excel ファイルを選択する。＞ 開く を押す。



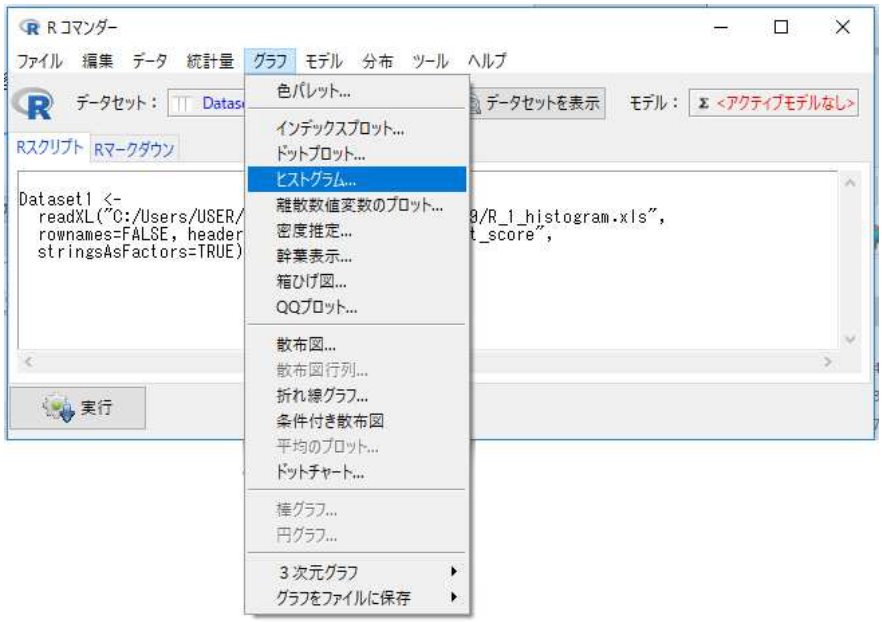
●もし Excel ファイルに２枚のシートが格納されていると下のような小さなウィンドウが開く。＞　開きたいシートを選ぶ。この場合は、test_score 。＞ OK を押す。



●すると RStudio の Console で、次のような表示が現れる。Excel ファイルの当該シート（ test_score ）の読み込みが成功した。

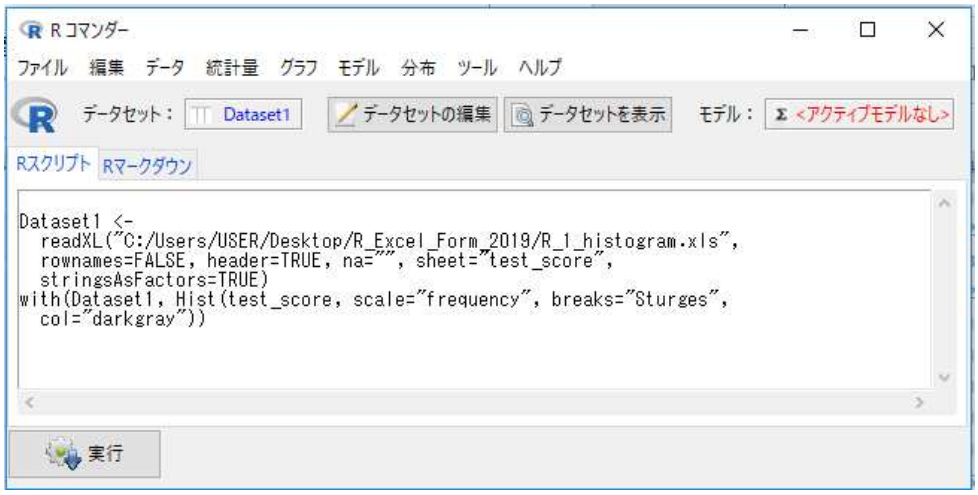


● R コマンダー のメニュー　＞グラフ　＞　ヒストグラム　と選択する。

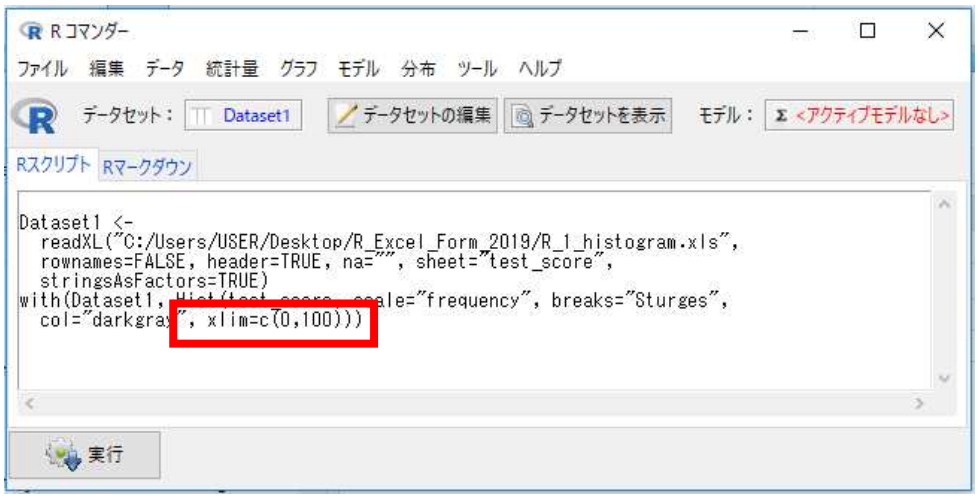


●開いたウィンドウで、 test_score を選択する。＞ OK を押す。

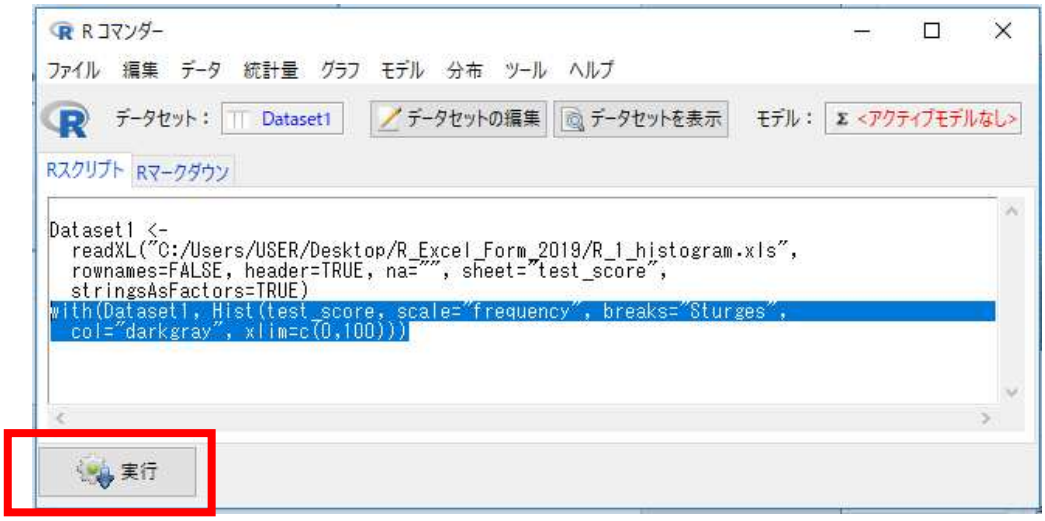
● R コマンダー で、以下のようなコマンドが記述されている。



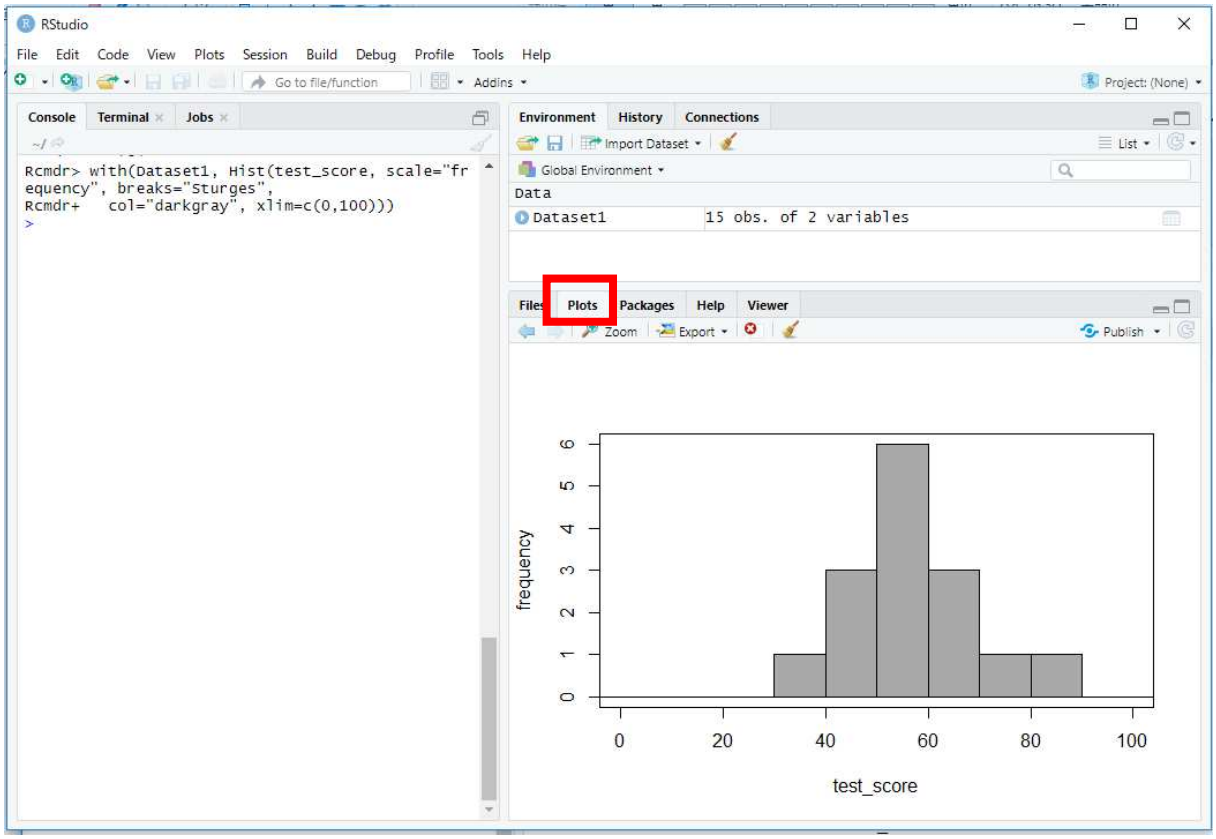
● R コマンダー で、カッコの中に下のように書き加える。 xlim=c(0,100)



● R コマンダー で、その２行をハイライトする。＞ 実行 を押す。

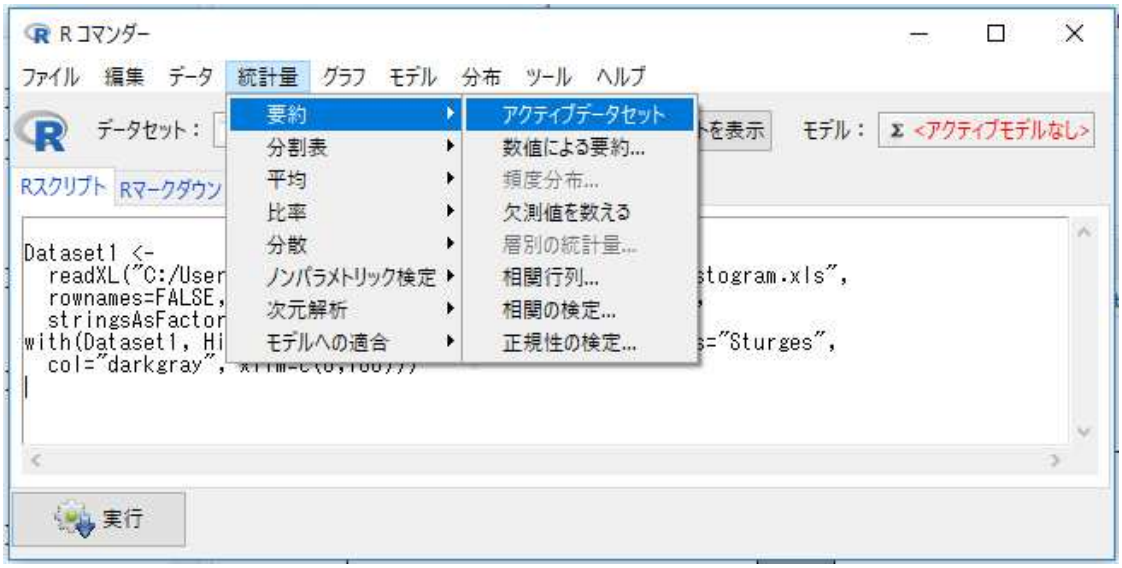


● RStudio の Plots のタブの中に、ヒストグラムが作成された。

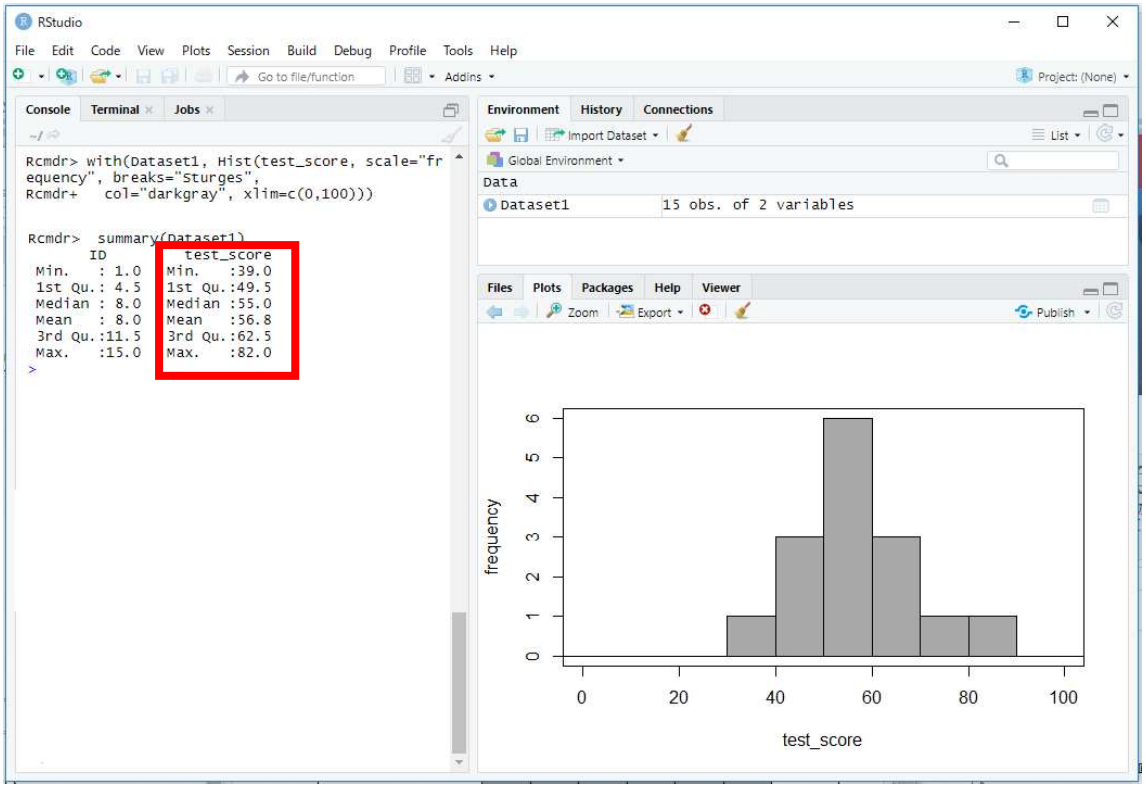


［2］平均値、中央値、標準偏差の計算

●R コマンダー のメニューから 統計量 ＞ 要約 ＞ アクティブデータセット を選ぶ。

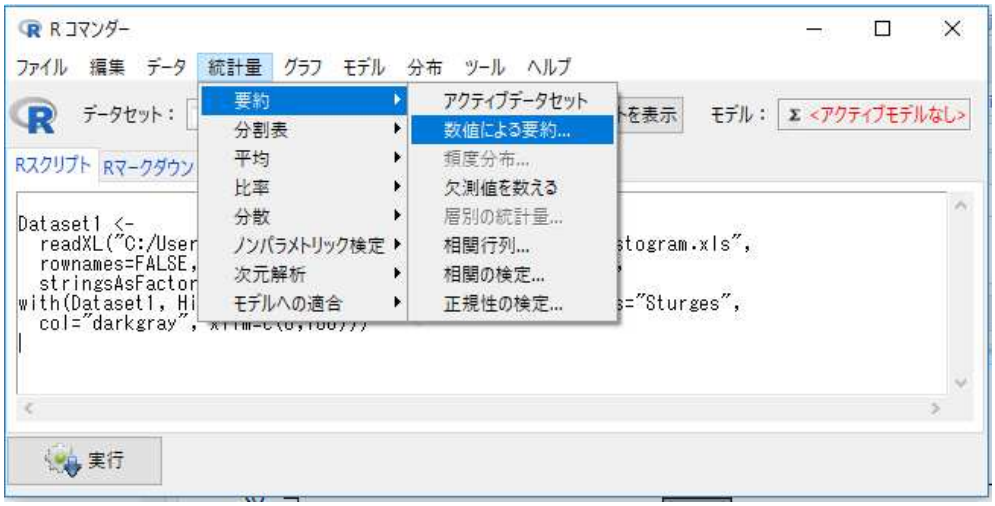


●RStudio の Console に以下にような出力される。min（最小値）、median（中央値）、mean（平均値）、max（最大値）が表示されている。

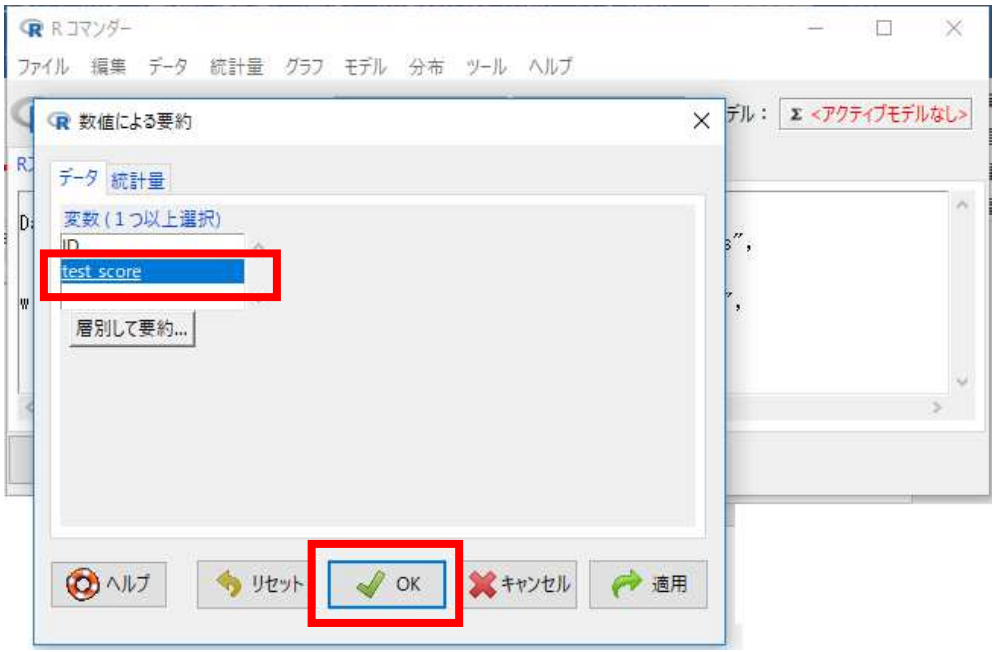


＝＞ median（中央値）は 55.0、mean（平均値）は 56.8。

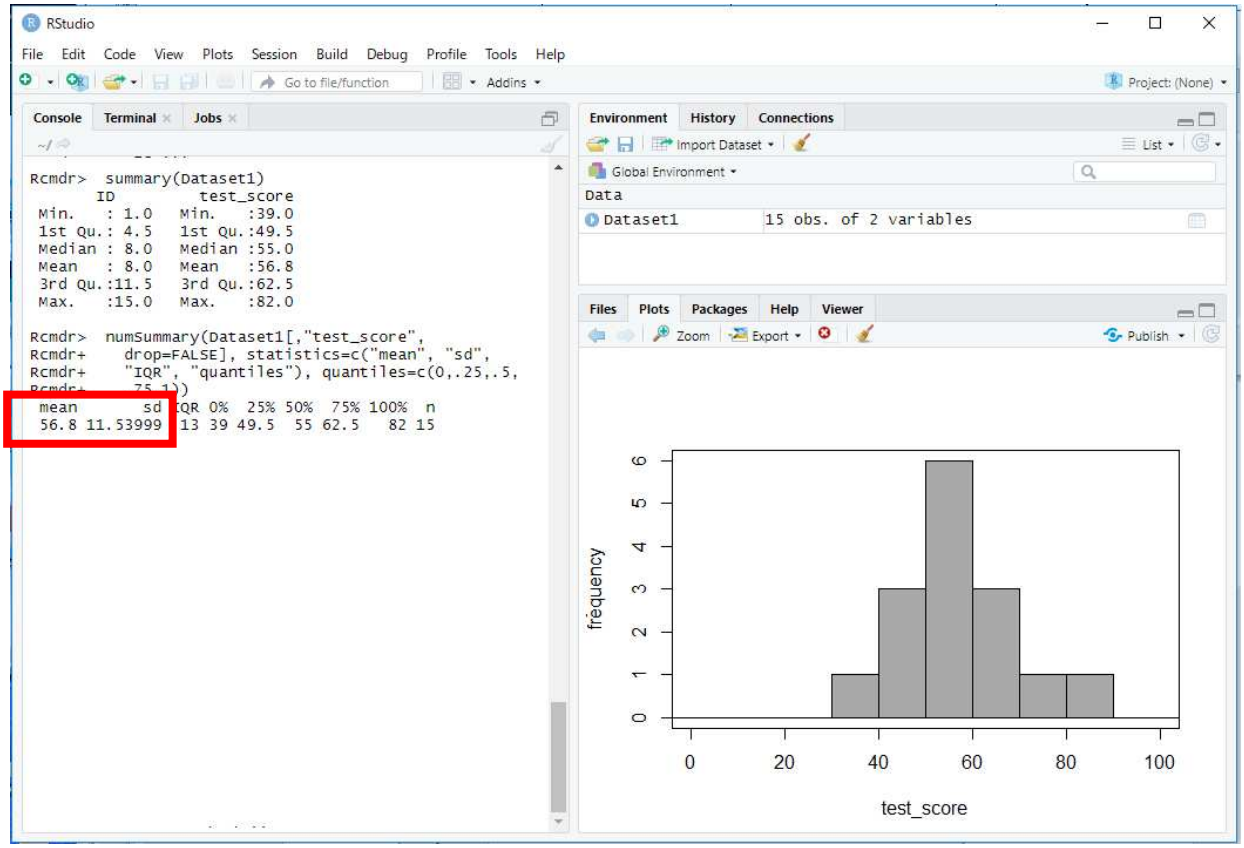
●R コマンダー のメニューから 統計量 ＞ 要約 ＞ 数値による集約 を選ぶ。



●開いたウィンドウで、test_score を選択する。＞ OK を押す。



●RStudio の Console　に以下にような出力される。
mean （平均値）、sd （標準偏差）が表示されている。



＝＞　mean （平均値）＝ 56.8、sd （標準偏差）＝ 11.54。

評価結果

あなたは以下の情報を得た。

● クラスの 15 人の平均点は 56.8 点、中央値は 55 点であった。

● 標準偏差は 11.54 点。

● ヒストグラムを見ると、ばらつき具合はわりと普通（標準的）である（relatively normal）。

表１：担当クラスの記述統計（Descriptive Statistics）

	担当クラス
n	15
平均値	56.8
標準偏差	11.54

発展学習１：平均値の手計算

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

ID	X
1	43
2	51
3	39
4	46
5	69
6	57
7	72
8	59
9	61
10	55
11	82
12	64
13	51
14	48
15	55
Total＝ $\left(\sum_{i=1}^{n} X_i\right)$	852
Average＝ $(\overline{X})$	56.80

発展学習 2：標準偏差の手計算

$$s_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

ID	X (X_i)	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
1	43	−13.80	190.44
2	51	−5.80	33.64
3	39	−17.80	316.84
4	46	−10.80	116.64
5	69	12.20	148.84
6	57	0.20	0.04
7	72	15.20	231.04
8	59	2.20	4.84
9	61	4.20	17.64
10	55	−1.80	3.24
11	82	25.20	635.04
12	64	7.20	51.84
13	51	−5.80	33.64
14	48	−8.80	77.44
15	55	−1.80	3.24
		Total (①)＝	1864.40
		Variance (②)＝	133.17
		Standard deviation (s_X) ＝	11.54

^2

=sum(H4:H18)

= ①/(15-1)

= sqrt(②)

	下限		上限
3分の2の範囲	45.26	~	68.34
95%の範囲	33.72	~	79.88

発展学習 3：正規分布（Normal Distribution）

［ 標準偏差の応用］

(1)　データの３分の２（68%）の個数が入る範囲を求めよ

答え：平均値 - 1 * 標準偏差 ～ 平均値 + 1 x 標準偏差
56.80 - 1 * 11.54 ～ 56.80 + 1 * 11.54 = **45.26 ～ 68.34**
理論値：15 * 2/3 = 10　（このケースの場合、10 個が入るはず）
実測値：確かに **10 個入っている**。確かめてみよ。

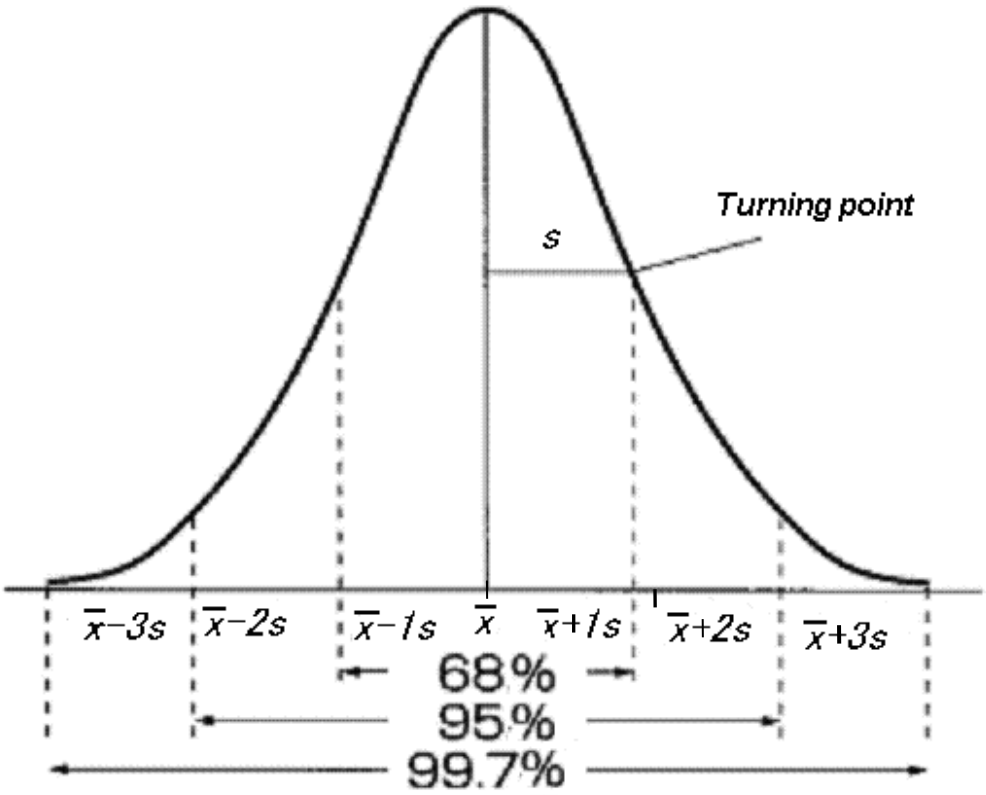
	A	B
1	ID	test_score
2	1	43
3	2	51
4	3	39
5	4	46
6	5	69
7	6	57
8	7	72
9	8	59
10	9	61
11	10	55
12	11	82
13	12	64
14	13	51
15	14	48
16	15	55

	A	B
1	ID	test_score
2	1	43
3	2	51
4	3	39
5	4	46
6	5	69
7	6	57
8	7	72
9	8	59
10	9	61
11	10	55
12	11	82
13	12	64
14	13	51
15	14	48
16	15	55

(2) データの 95%の個数が入る範囲

答え：平均値 - 2 * 標準偏差 ～ 平均値 + 12 x 標準偏差
56.80 - 2 * 11.54 ～ 56.80 + 2 * 11.54 = **33.72 ～ 79.88**
理論値：5 * 0.95 = 14.25　（このケースの場合、約 14 個が入る。）
実測値：確かに **14 個入っている**。確かめてみよ。

正規分布の正式なグラフ



かつで流通していた 10 マルク紙幣には、正規分布のグラフと式が書き込んであった。この、お札を教育に使うという発想に脱帽である。

Carl Friedrich Gauss
(1777-1855)

第 3 章　t 検定

詳細目次

第 2 セッション　35

事前・事後の t 検定（対応のあるt検定）:

保護者の満足度は向上したか？1:過去の満足度と比較してみよう。

第 3 セッション　45

２群の t 検定（対応のない t 検定）:

保護者の満足度は向上したか？2: :隣の学校の保護者満足度と比較してみよう。

第 2 セッション

事前・事後の t 検定（対応のあるt検定）[1][2]

保護者の満足度は向上したか？1: 過去の満足度と比較してみよう

日本の NGO の支援を受けて、カンボジアの某小学校は「革新的な学校運営プログラム」を導入しました。この教育プログラムによって、保護者の満足度は向上したと言えるでしょうか？導入前と導入後の満足度を比較して判断してみます。

<必要な条件>

- 同一のサーベイ用紙（アンケート用紙）が、事前サーベイと事後サーベイで使用されねばならない。
- 事前と事後で、同一の顧客がサーベイ用紙に記入せねばならない。
- 定量的な選択肢が用いられねばならない。
  (e.g.)　5= たいへん満足（Very satisfactory）
  4= 満足（Satisfactory）
  3= どちらとも言えない（Hard to say）
  2= 不満足（Unsatisfactory）
  1= たいへん不満足（Very unsatisfactory）

[1] Likert type scale のデータに t-test などのパラメトリック検定が適用可能かどうかに関しては議論があるが、現実には調査研究において広く用いられている。また、Winder and Dodou (2012) の最新の研究では次のように結論されている。 “In conclusion, the t-test (parametric test) and MWW test (nonparametric test) generally have similar power, and researchers do not have to worry about finding a difference”. (Source) Joost C. F. de Winter and Dimitra Dodou.(2012). Five-Point Likert Items: t test versus Mann-Whitney-Wilcoxon, Practical Assessment, Research and Evaluation. http://pareonline.net/pdf/v15n11.pdf

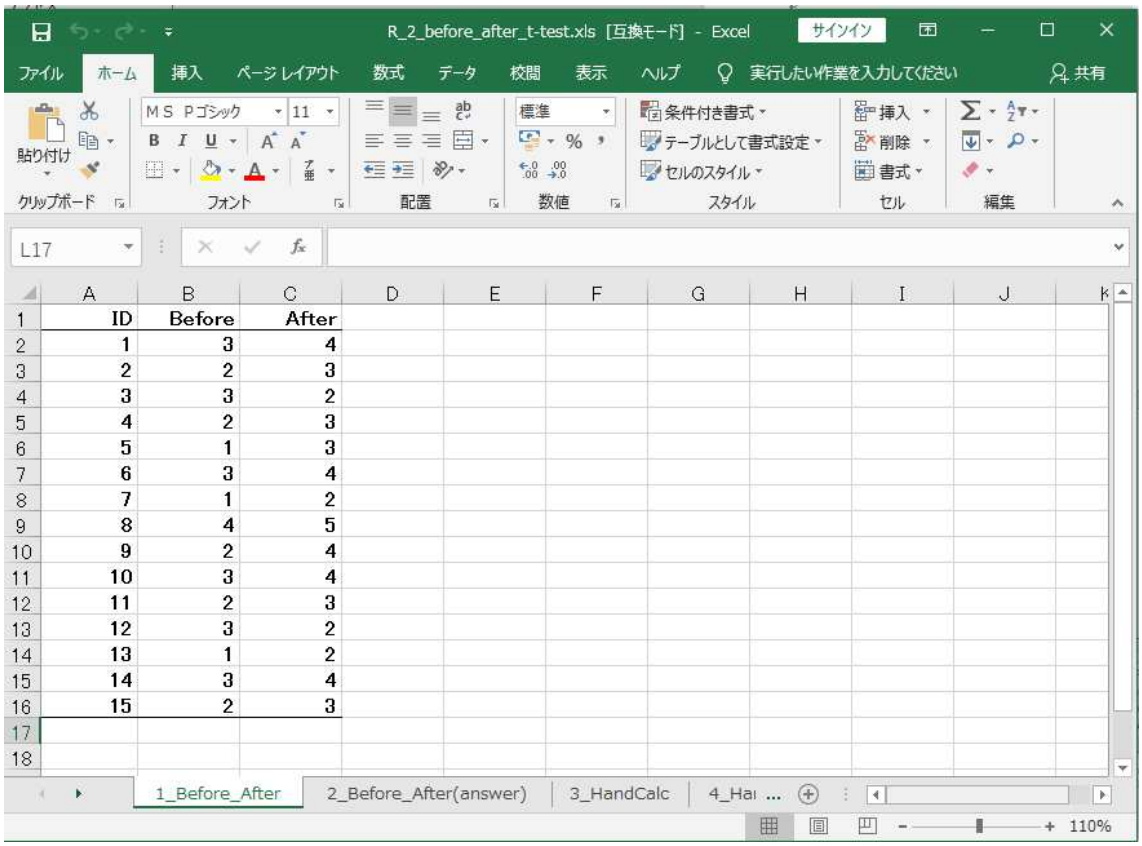
[2] JICA 研究所の「カイゼン」に関するインパクト評価報告書では次の 5-point scale を用いている。Five-point scale: 5.Very good, 4. Good, 3. Moderate, 2. Bad, 1. Very bad. (Source) Shimada, G. & Sonobe, T. (2018). p.31. https://www.jica.go.jp/jica-ri/ja/publication/workingpaper/wp_173.html

保護者 15 人に対して実施されたサーベイ結果は以下のとおりであった。

ID	Before	After
1	3	4
2	2	3
3	3	2
4	2	3
5	1	3
6	3	4
7	1	2
8	4	5
9	2	4
10	3	4
11	2	3
12	3	2
13	1	2
14	3	4
15	2	3

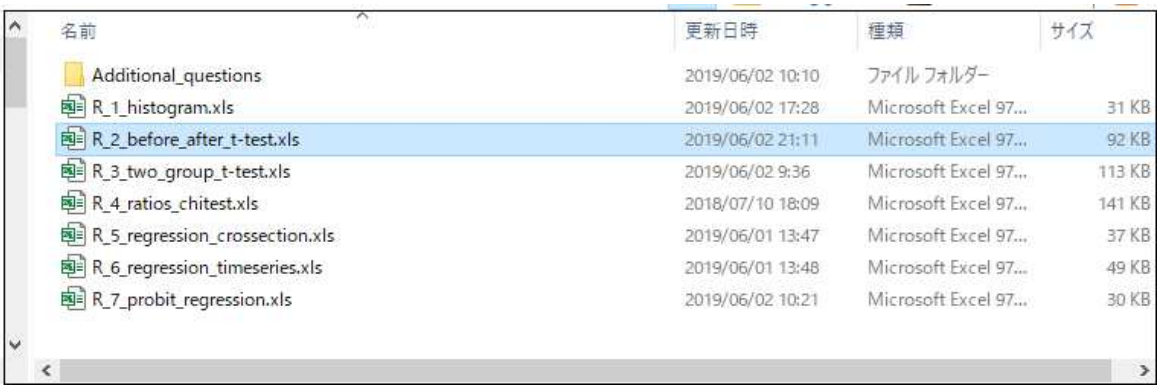
データの入力と読み込み

●以下のようにデータを入力する。タブに「1_before_after」などの名前をつける。



ID	Before	After
1	3	4
2	2	3
3	3	2
4	2	3
5	1	3
6	3	4
7	1	2
8	4	5
9	2	4
10	3	4
11	2	3
12	3	2
13	1	2
14	3	4
15	2	3

●デスクトップに、R_2_before_after_t-test.xls などと適当な名前をつけて保存する。

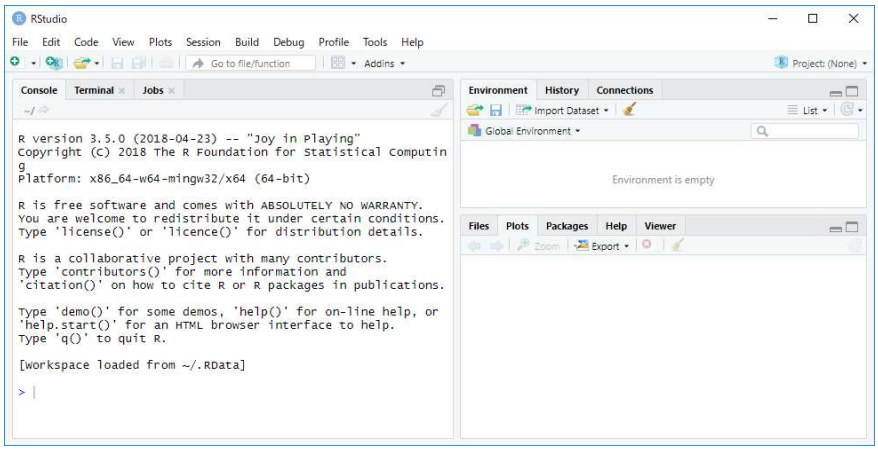


名前	更新日時	種類	サイズ
Additional_questions	2019/06/02 10:10	ファイル フォルダー	
R_1_histogram.xls	2019/06/02 17:28	Microsoft Excel 97...	31 KB
R_2_before_after_t-test.xls	2019/06/02 21:11	Microsoft Excel 97...	92 KB
R_3_two_group_t-test.xls	2019/06/02 9:36	Microsoft Excel 97...	113 KB
R_4_ratios_chitest.xls	2018/07/10 18:09	Microsoft Excel 97...	141 KB
R_5_regression_crossection.xls	2019/06/01 13:47	Microsoft Excel 97...	37 KB
R_6_regression_timeseries.xls	2019/06/01 13:48	Microsoft Excel 97...	49 KB
R_7_probit_regression.xls	2019/06/02 10:21	Microsoft Excel 97...	30 KB

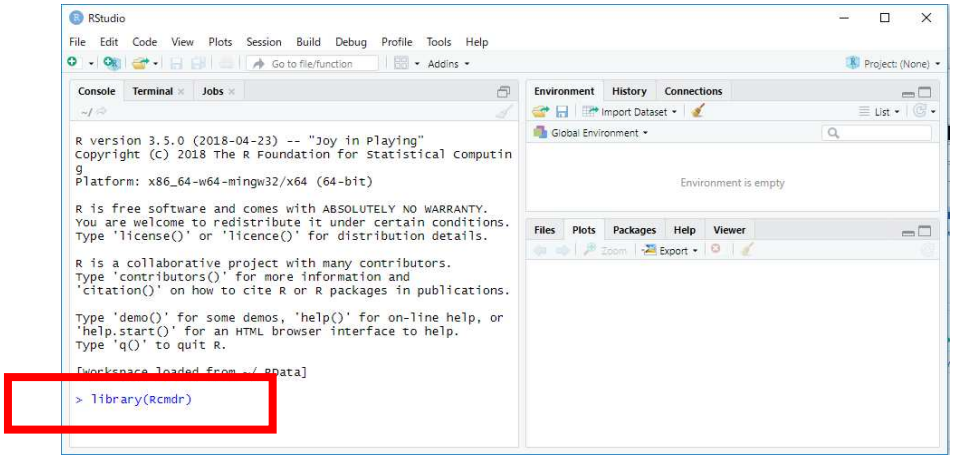
（すでに RStudio と R コマンダーを立ち上げている場合はこのページを省略してください。）

● Windows のメニュー＞“RStudio”を選ぶ。

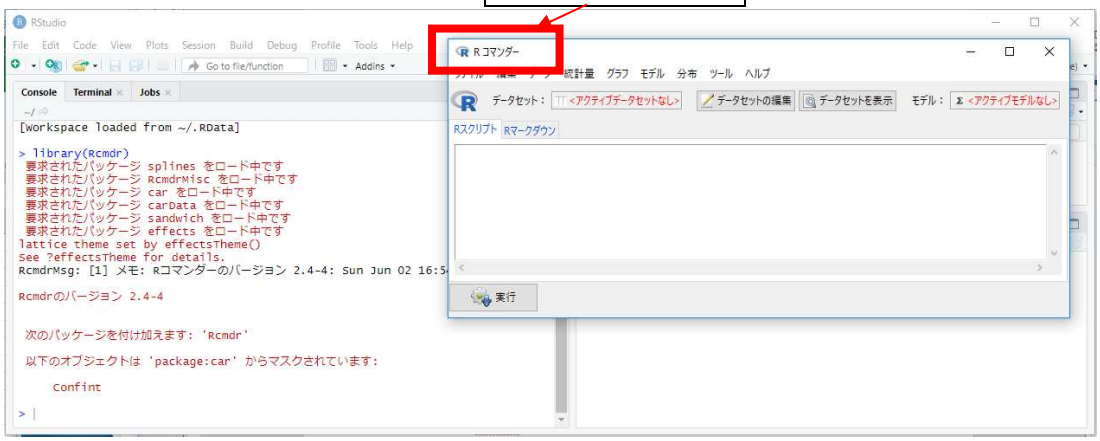
● 下のようなウィンドウが開く。



● ” library(Rcmdr) “と入力する。＞Enter を押す。



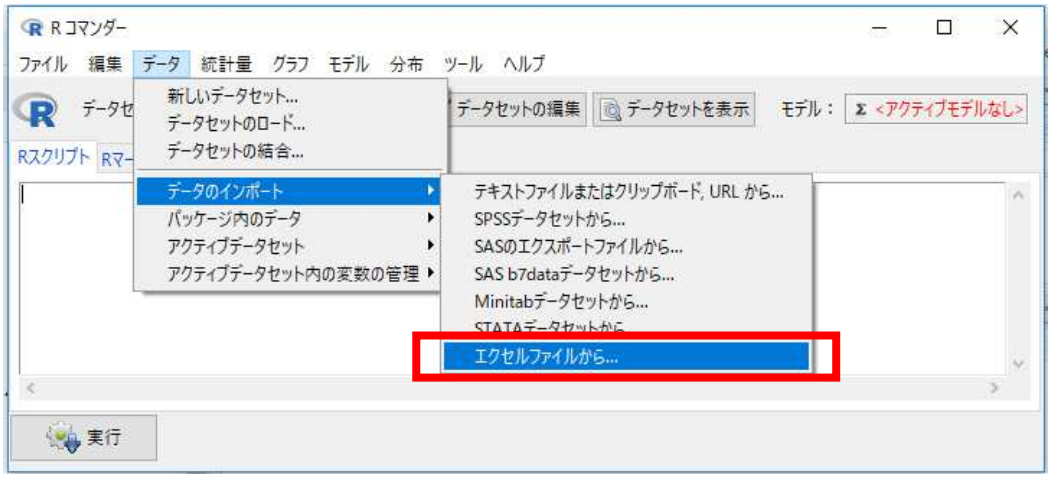
●下のようなメッセージが流れて、 R コマンダー のウィンドウが開く。



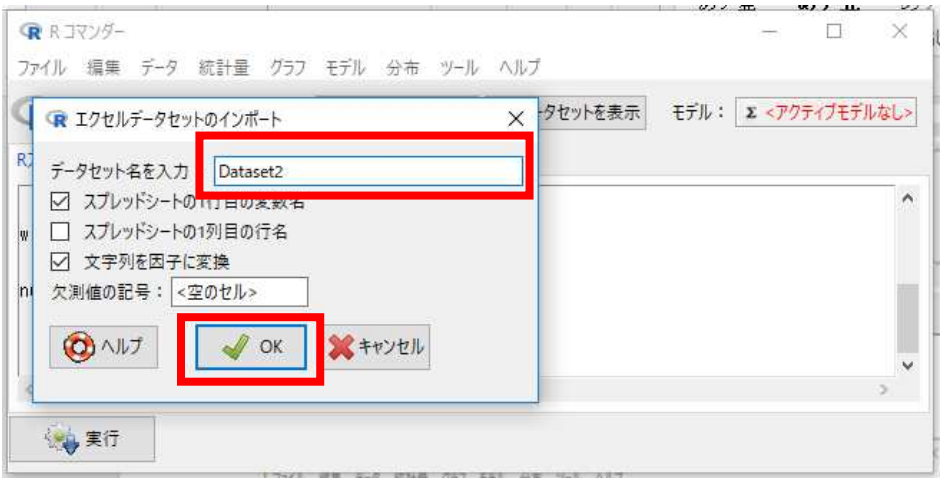
（前ページを飛ばした方はここから再開してください。）

ル

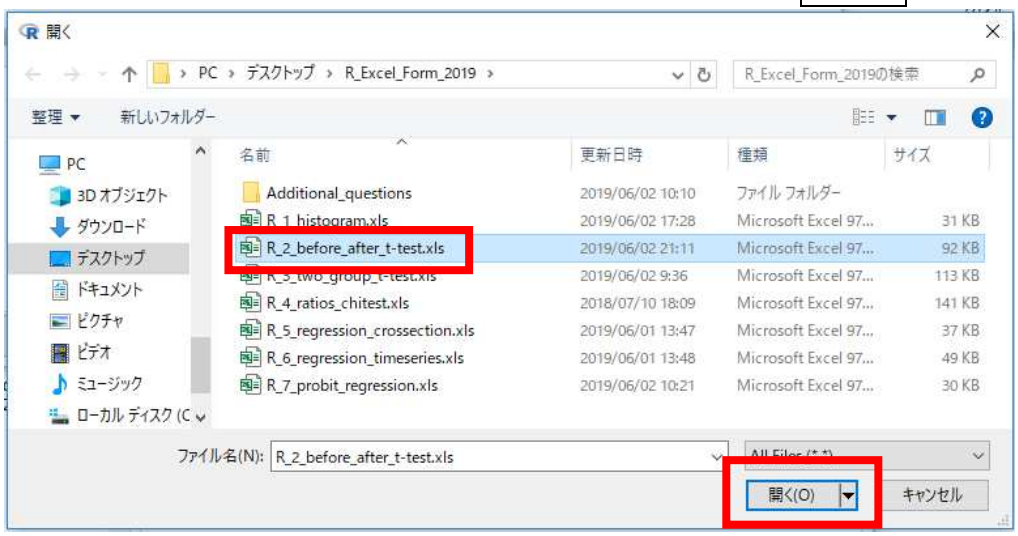
● R コマンダー のウィンドウで、データ ＞ データのインポート ＞ エクセルファイルから を選ぶ。



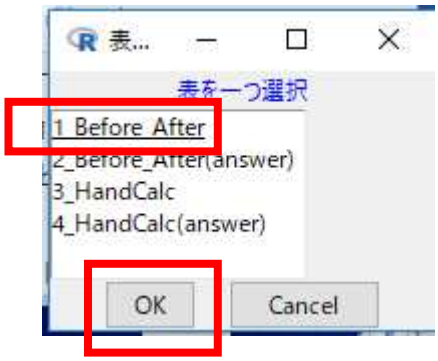
●現れたウィンウィンドウで、“データセットを入力” に “Dataset2” と入力する。＞ OK を押す。



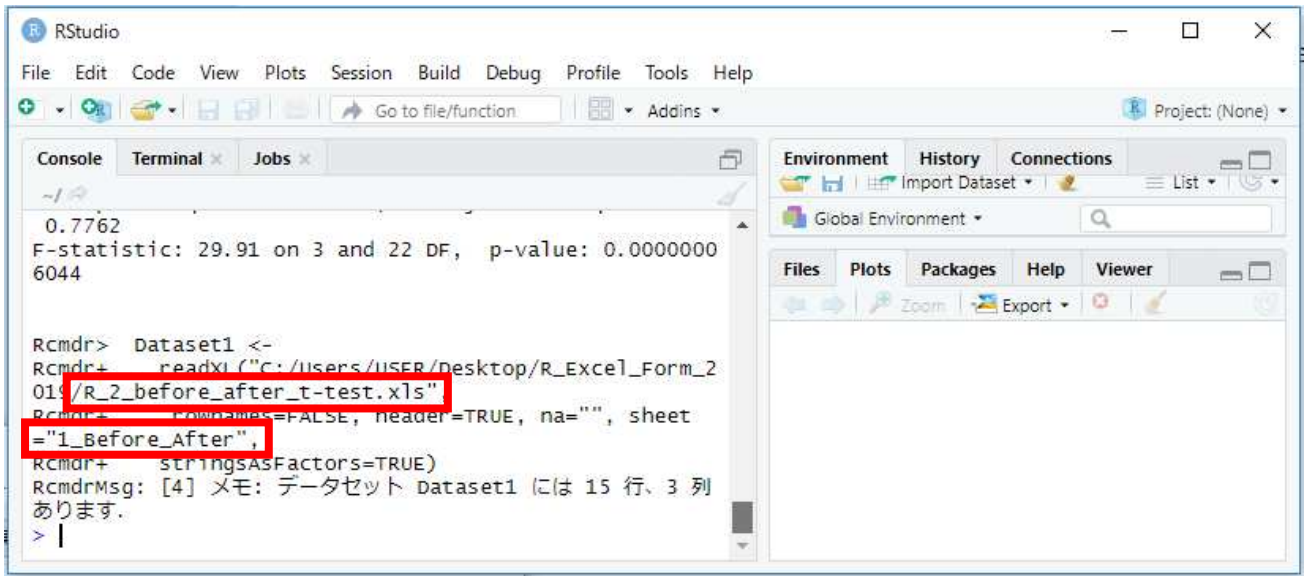
●さきほど保存した Excel ファイルを選択する。＞ 開く を押す。



●もし Excel ファイルに２枚のシートが格納されていると下のような小さなウィンドウが開く。＞　開きたいシートを選ぶ。この場合は、1_before_after 。＞ OK を押す。

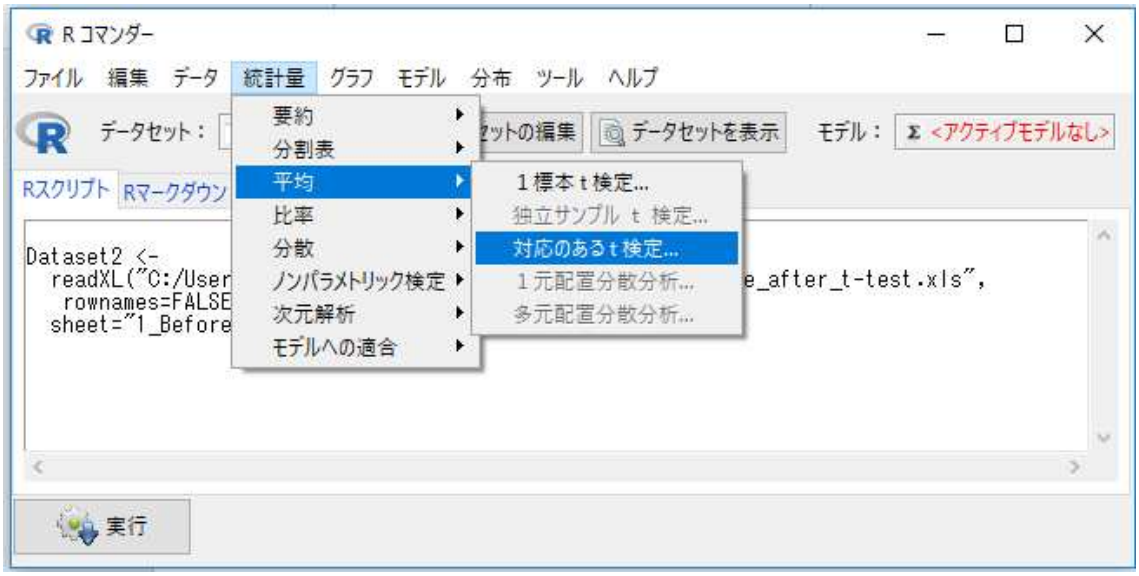


●すると RStudio の Cosole で、次のような表示が現れる。Excel ファイルの当該シート（ 1_Before_After ）の読み込みが成功した。



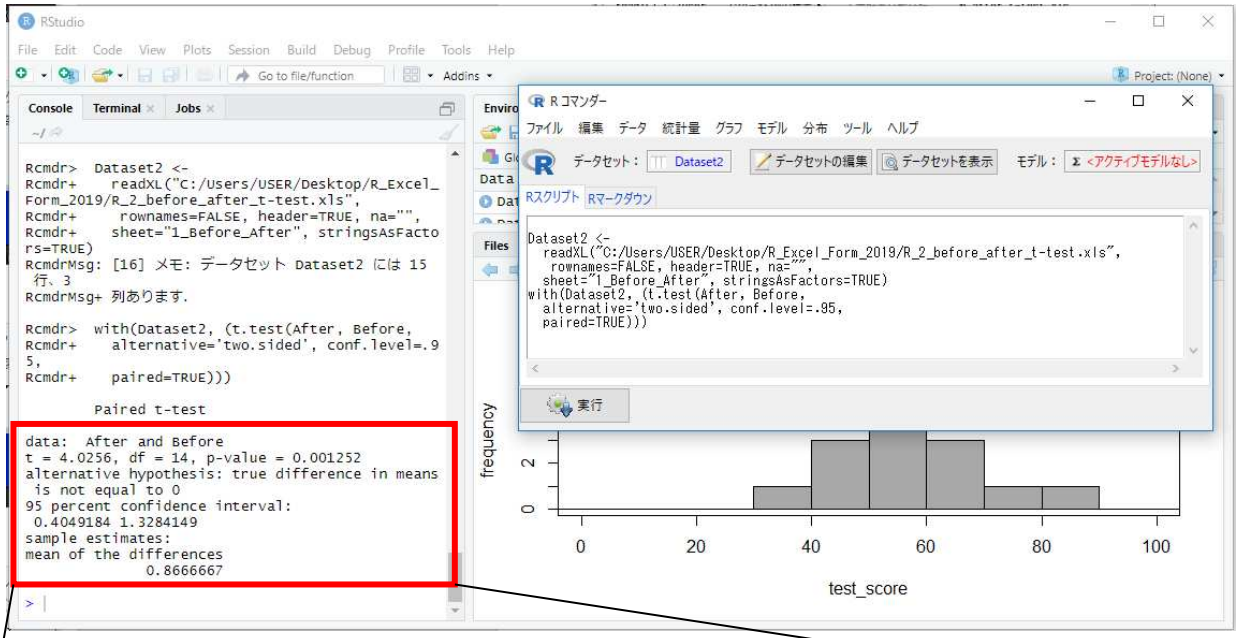
「対応のある t 検定」の選択

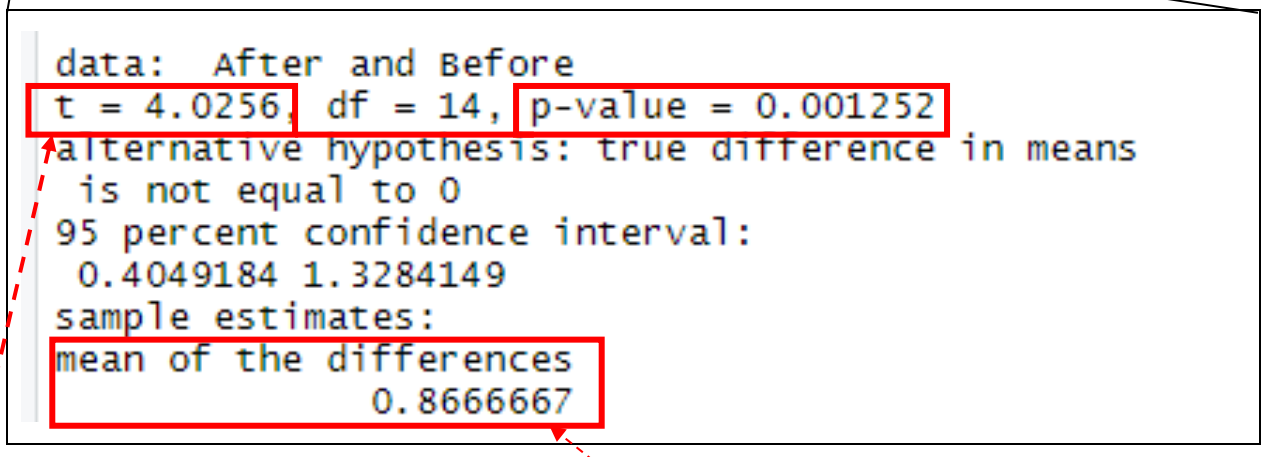
● R コマンダー のメニューから 統計量 ＞ 平均 ＞ 対応のある t 検定 を選ぶ。



●現れたウィンドウで、「第 1 の変数（1 つ選択）」で After を選ぶ。
「第 2 の変数（1 つ選択）」で Before を選ぶ。 OK を押す。

●RStudio の Console　に以下にような出力される。 t 値 、 p 値 、 mean of the difference （二つの平均値の差）が表示されている。





平均値：事後（After）は 3.2、事前（Before）は 2.33 なので、その差は 0.87 である。

t 値 ：その差(0.87)が、偶然かそれも偶然では起こりえないほど大きな差なのかを判断する計算結果。（なお、t 値は絶対値で見る。）

t ＞2 なら、統計学的に有意 。

「偶然では起こりえないほど大きな差である。」

t ≦ 2 なら、統計学的に有意とは言えない。

「偶然でも起こり得る程度の差である。」

p 値 ：確率値（probability）。t 値の別の形の表現。

p ＜0.05 (5%)なら、統計学的に有意 。

「偶然では起こりえないほど大きな差である。」

p ≧0.05(5%)なら、統計学的に有意とは言えない。

「偶然でも起こり得る程度の差である。」

結論

「事前」の平均値と「事後」の平均値の差は、統計学的に有意である（$t = 4.03$, $p = 0.00125 < 5\%$）。したがって、「革新的な学校運営プログラム」によって保護者の満足度は改善したと判断できる。

表１：「革新的な学校運営プログラム」の満足度の事前・事後の t 検定

	事前	事後	差	*t*	*p*
n	15	15			
平均値	2.33	3.20	0.87	4.025	0.00125
標準偏差	0.90	0.94			

発展学習１:「対応のあるｔ検定（事前・事後のｔ検定）」の手計算

(Before-After Comparison)

ID	Before	After	$d = (X_{After} - X_{Before})$	$(d_n - \bar{d})$	$(d_n - \bar{d})^2$
1	3	4	1	0.133	0.018
2	2	3	1	0.133	0.018
3	3	2	-1	-1.867	3.484
4	2	3	1	0.133	0.018
5	1	3	2	1.133	1.284
6	3	4	1	0.133	0.018
7	1	2	1	0.133	0.018
8	4	5	1	0.133	0.018
9	2	4	2	1.133	1.284
10	3	4	1	0.133	0.018
11	2	3	1	0.133	0.018
12	3	2	-1	-1.867	3.484
13	1	2	1	0.133	0.018
14	3	4	1	0.133	0.018
15	2	3	1	0.133	0.018
Average	2.3333	3.2			
		Total_d=	13	*Total=*	9.733
		Average $(\bar{d})$ =	0.8667		

$$s_d = \sqrt{\frac{Total}{(n-1)}} = 0.8338$$

*n….Sample size.
In this case, n=15.

$$SE = \frac{s_d}{\sqrt{n}} = 0.2153$$

$$t = \frac{\overline{X}_{after} - \overline{X}_{before}}{SE} = 4.0256$$

p-value = 0.00125

*p-value is obtained by EXCEL Command: =tdist (t, n-1,2)
In this case, type [=tdist(4.0256,14,2)] at any cell. You can get 0.0012

<Advance: Effect size (Cohen's *d*)>

$$d = \frac{\left|\overline{X}_{after} - \overline{X}_{before}\right|}{s_d} = 1.039$$

Criteria: Small: d =0.2, Medium: d=0.5, Large: d=0.8 (Source) Cohen (1988)

第 3 セッション

対応のないt検定（2群の t 検定）)[3][4]:

保護者の満足度は向上したか？2: :
隣の学校の保護者満足度と比較してみよう。

Q. 日本の NGO の支援を受けて、カンボジアの某小学校は「革新的な学校運営プログラム」を導入しました。この教育プログラムによって、保護者の満足度は向上したと言えるでしょうか？

よく似た近隣の学校の保護者の満足度と比較して判断してみます。

<必要な条件>

- 同一のサーベイ用紙（アンケート用紙）が、2 つの地域で使用されねばならない。
- 定量的な選択肢が用いられねばならない。
  (e.g.)　5= たいへん満足（Very satisfactory）
  4= 満足（Satisfactory）
  3= どちらとも言えない（Hard to say）
  2= 不満足（Unsatisfactory）
  1= たいへん不満足（Very unsatisfactory）

---

[3] Likert type scale のデータに t-test などのパラメトリック検定が適用可能かどうかに関しては議論があるが、現実には調査研究において広く用いられている。また、Winder and Dodou (2012) の最新の研究では次のように結論されている。 “In conclusion, the t-test (parametric test) and MWW test (nonparametric test) generally have similar power, and researchers do not have to worry about finding a difference”. (Source) Joost C. F. de Winter and Dimitra Dodou.(2012). Five-Point Likert Items: t test versus Mann-Whitney-Wilcoxon, Practical Assessment, Research and Evaluation. http://pareonline.net/pdf/v15n11.pdf

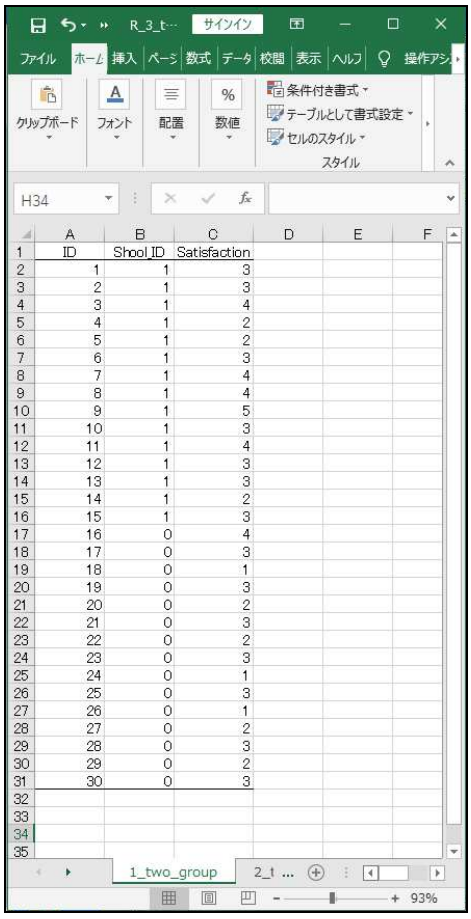
[4] JICA 研究所の「カイゼン」に関するインパクト評価報告書では次の 5-point scale を用いている。Five-point scale: 5.Very good, 4. Good, 3. Moderate, 2. Bad, 1. Very bad. (Source) Shimada, G. & Sonobe, T. (2018). p.31. https://www.jica.go.jp/jica-ri/ja/publication/workingpaper/wp_173.html

Shool_ID	Satisfaction
1	3
1	3
1	4
1	2
1	2
1	3
1	4
1	4
1	5
1	3
1	4
1	3
1	3
1	2
1	3
0	4
0	3
0	1
0	3
0	2
0	3
0	2
0	3
0	1
0	3
0	1
0	2
0	3
0	2
0	3

データの入力

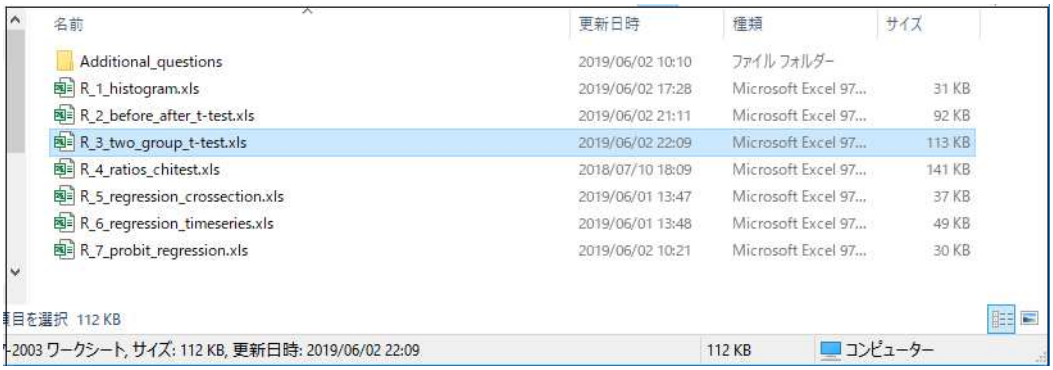
●以下のようにデータを入力する。タブに「1_two_group」などの名前をつける。

School_ID = 1（わが校)、School_ID = 0（近隣校）



ID	Shool_ID	Satisfaction
1	1	3
2	1	3
3	1	4
4	1	2
5	1	2
6	1	3
7	1	4
8	1	4
9	1	5
10	1	3
11	1	4
12	1	3
13	1	3
14	1	2
15	1	3
16	0	4
17	0	3
18	0	1
19	0	3
20	0	2
21	0	3
22	0	2
23	0	3
24	0	1
25	0	3
26	0	1
27	0	2
28	0	3
29	0	2
30	0	3

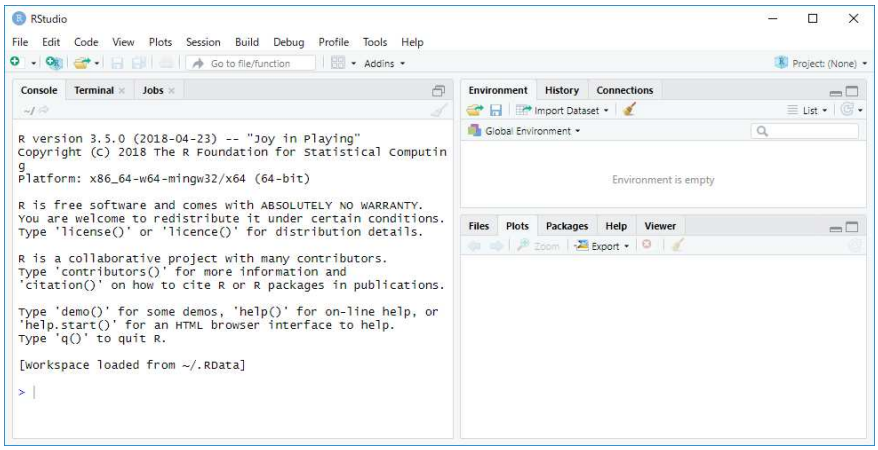
● デスクトップに、R_2_two_group_t-test.xls などと適当な名前をつけて保存する。



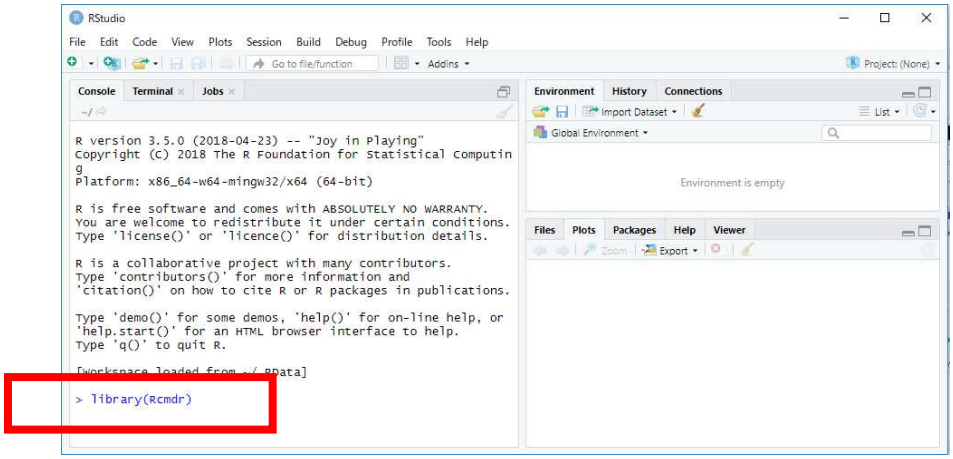
（すでに RStudio と R コマンダーを立ち上げている場合はこのページを省略してください。）

● Windows のメニュー＞“RStudio”を選ぶ。

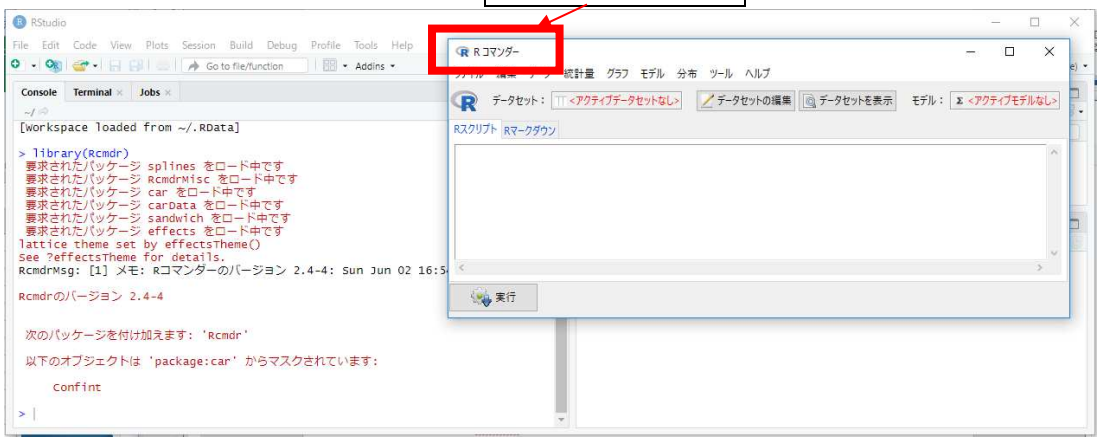
● 下のようなウィンドウが開く。



● ” library(Rcmdr) “と入力する。＞Enter を押す。

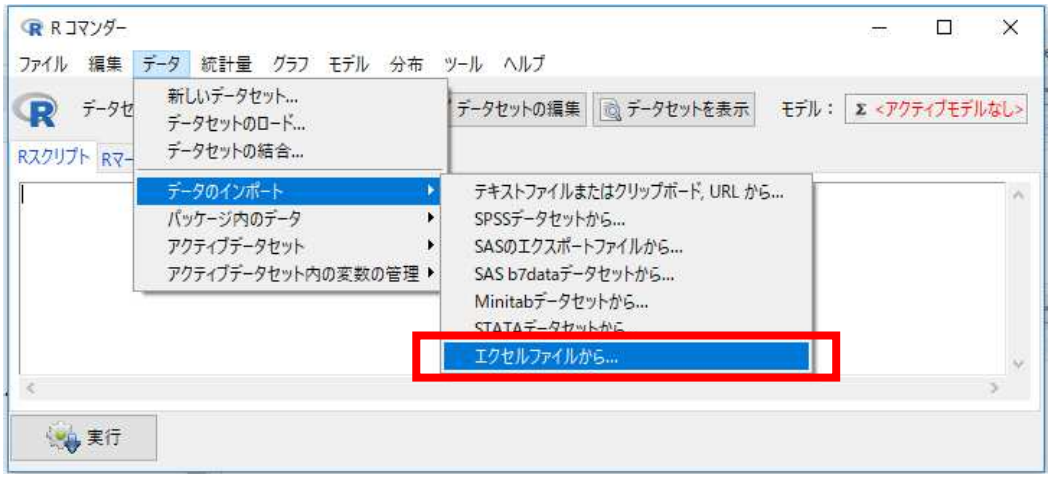


●下のようなメッセージが流れて、 R コマンダー のウィンドウが開く。

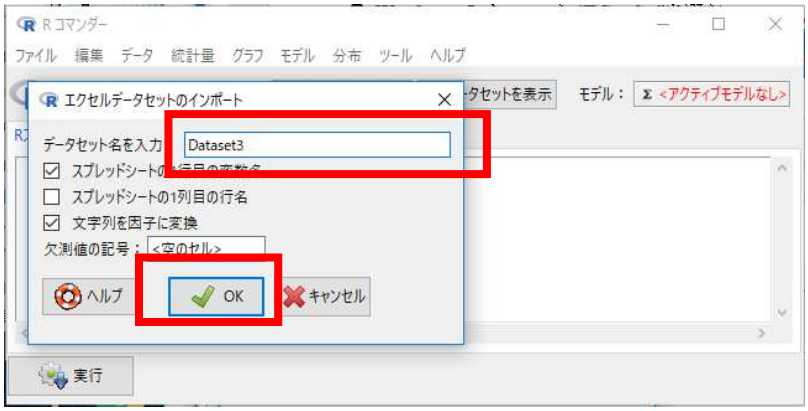


（前ページを飛ばした方はここから再開してください。）

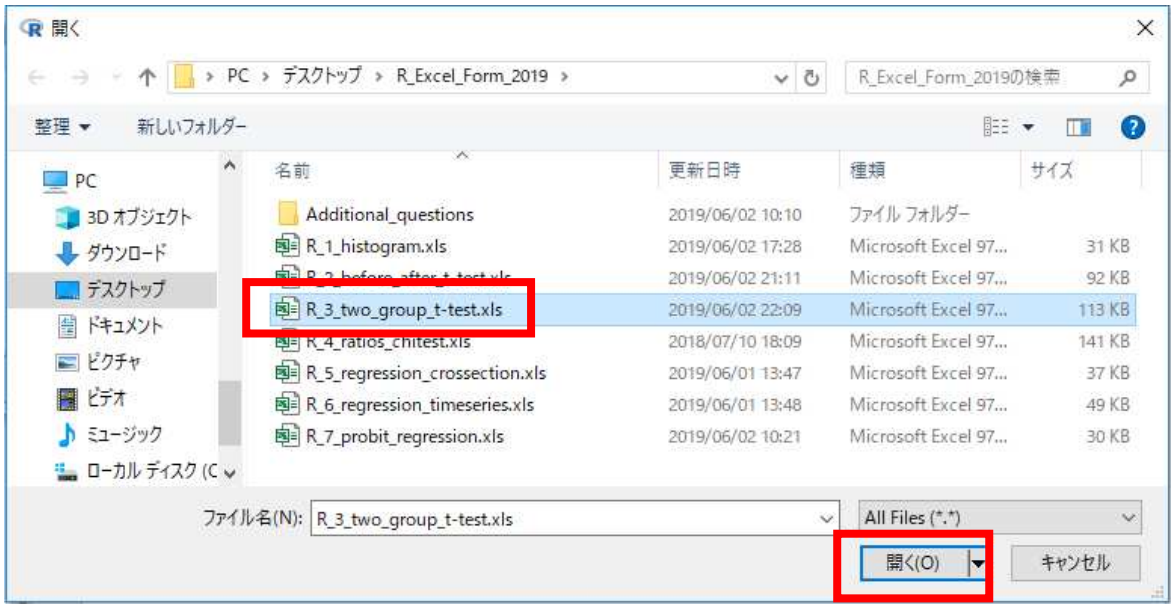
● R コマンダー のウィンドウで、データ ＞ データのインポート ＞ エクセルファイルから を選ぶ。



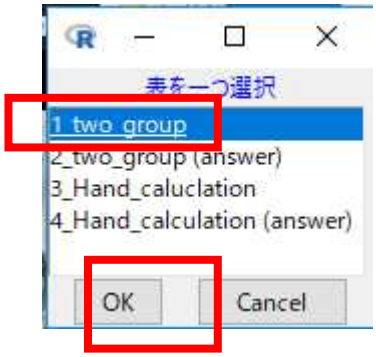
●現れたウィンウィンドウで、“データセットを入力”に“Dataset3”と入力する。＞ OK を押す。



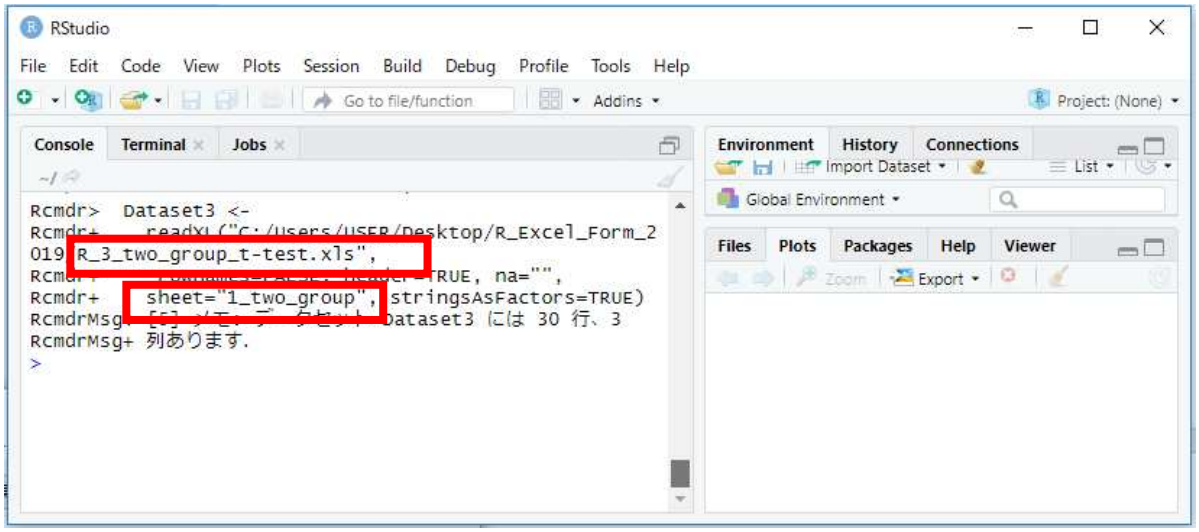
●さきほど保存した Excel ファイルを選択する。＞ 開く を押す。



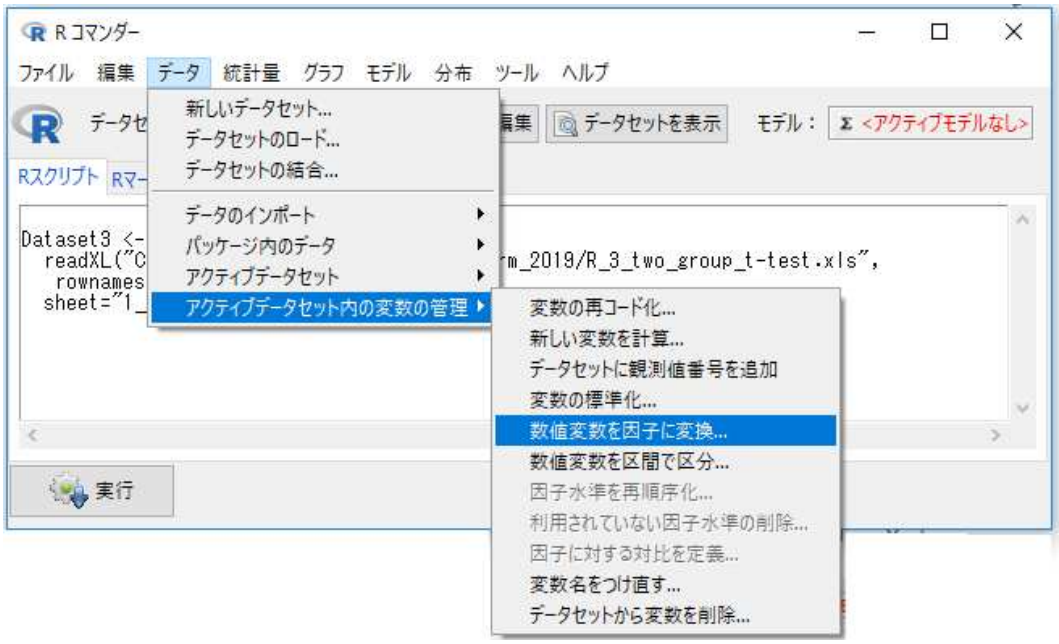
●もし Excel ファイルに２枚のシートが格納されていると下のような小さなウィンドウが開く。＞　開きたいシートを選ぶ。この場合は、1_two_group 。＞ OK を押す。



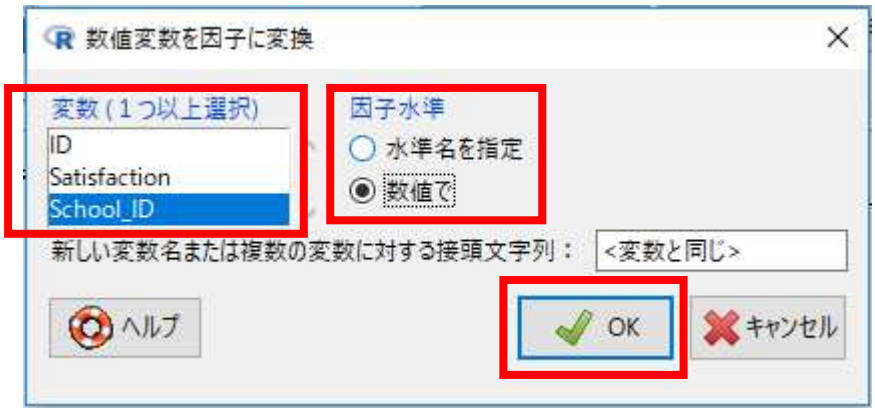
●すると R コマンダー で、次のような表示が現れる。Excel ファイルの当該シート 1_two_group の読み込みが成功した。



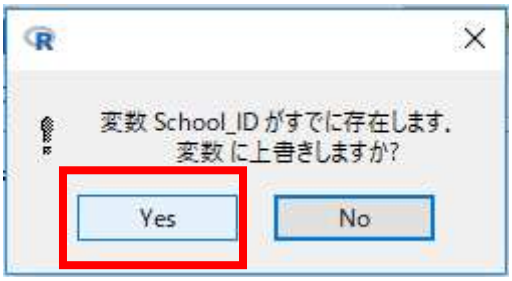
●R コマンダー で、データ ＞ アクティブデータセット内の変数の管理 ＞ 数値変数を因子に変換 を選ぶ。



●現れたウィンドウで、”変数（1 つ以上）”で School_ID を選ぶ。
続いて、”因子水準”で 数値で を選ぶ。 OK を押す。

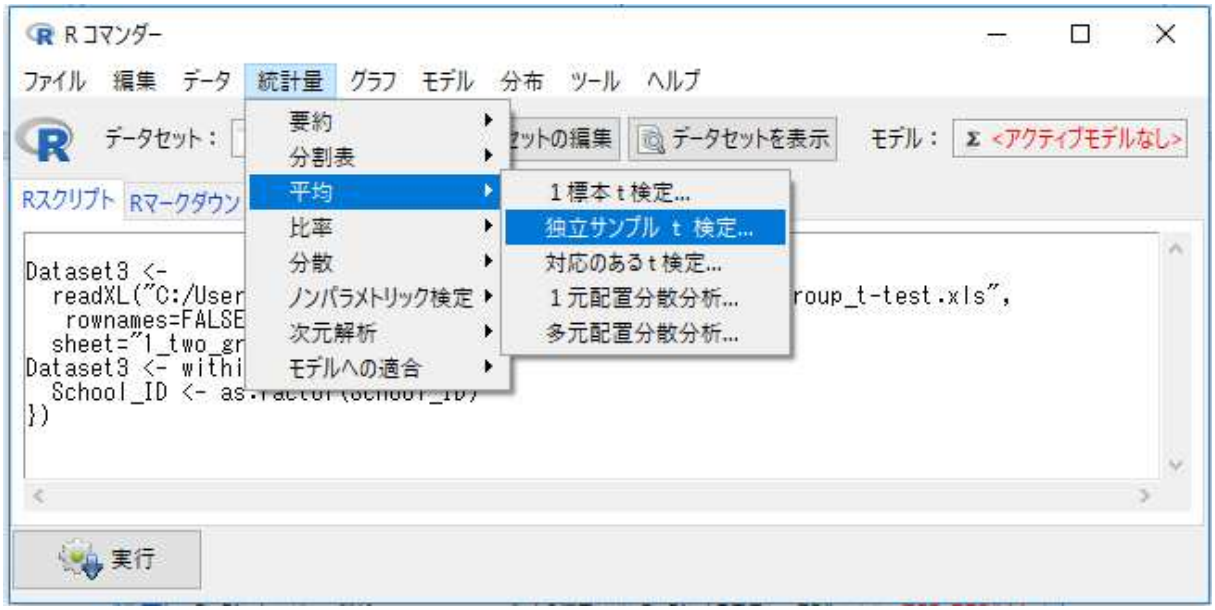


●現れたウィンドウで、 Yes を選ぶ。

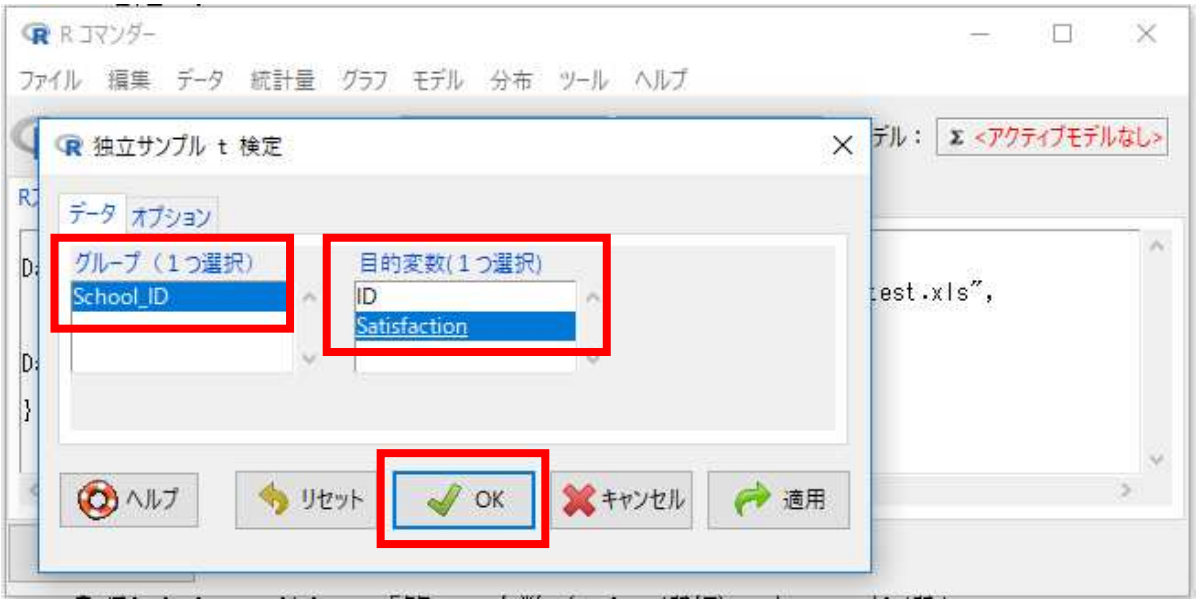


「独立サンプル t 検定」の選択

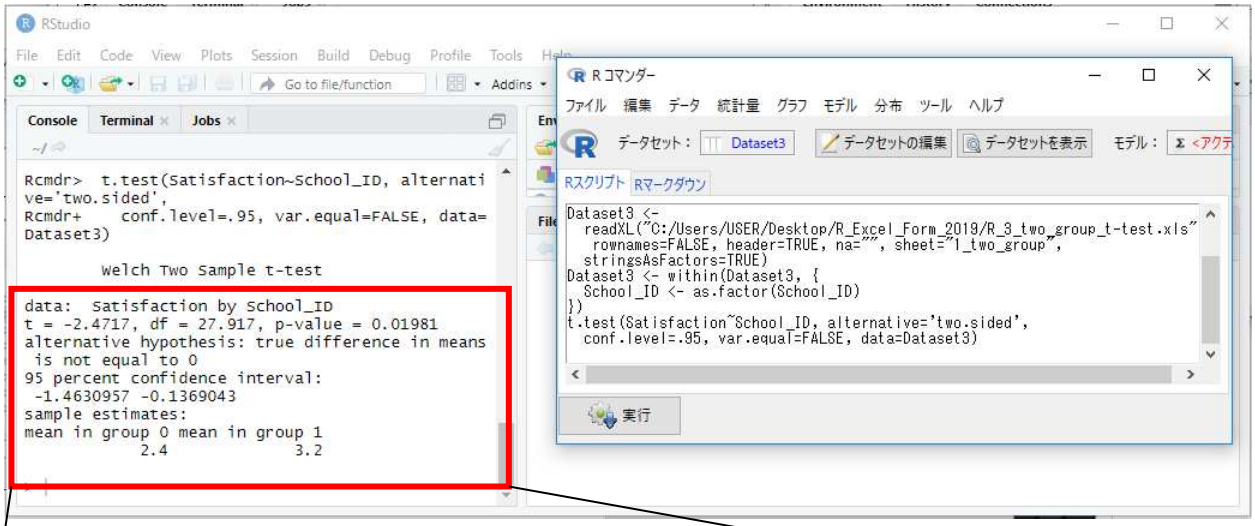
● R コマンダー のメニューから 統計量 ＞ 平均 ＞ 独立サンプル t 検定 を選ぶ。



●現れたウィンドウで、「グループ（1 つ選択）」で School_ID を選ぶ。
「目的変数（1 つ選択）で Satisfaction を選ぶ。 OK を押す。



●RStudio の Console　に以下にような出力される。t 値、p 値、mean of the difference（二つの平均値の差）が表示されている。




```
data:  Satisfaction by School_ID
t = -2.4717, df = 27.917, p-value = 0.01981
alternative hypothesis: true difference in means
 is not equal to 0
95 percent confidence interval:
 -1.4630957 -0.1369043
sample estimates:
mean in group 0 mean in group 1
            2.4             3.2
```

平均値：実施群（Group 1）は 3.2、比較群（Group 0）は 2.4 なので、その差は 0.8 である。

t 値：その差(0.87)が、偶然かそれも偶然では起こりえないほど大きな差なのかを判断する計算結果。（なお、t 値は絶対値で見る。）

t ＞2 なら、統計学的に有意 。

「偶然では起こりえないほど大きな差である。」

t ≦ 2 なら、統計学的に有意とは言えない。

「偶然でも起こり得る程度の差である。」

p 値：確率値（probability）。t 値の別の形の表現。

p ＜0.05 (5%)なら、統計学的に有意 。

「偶然では起こりえないほど大きな差である。」

p ≧0.05(5%)なら、統計学的に有意とは言えない。

「偶然でも起こり得る程度の差である。」

結論

“我校”の平均値と“近隣校”の平均値の差は統計学的に有意である（t = 2.47, p = 0.02 < 5%)。.

したがって、新しい学校運営プログラムの導入によって保護者の満足度は改善したと判断できる。

表 2 :「革新的な学校運営プログラム」の実施校・比較校の満足度の t 検定

	介入グループ	比較グループ	差	*t*	*p*
n	15	15			
平均値	3.2	2.40	0.80	2.47	0.0198
標準偏差	0.86	0.91			

発展学習１:「対応のないｔ検定（２群のｔ検定）」の手計算

Independent t-test
(Two-groups comparison)

School_ID	Satis-faction	$(X_{1i}-\overline{X}_1)$	$(X_{1i}-\overline{X}_1)^2$
1	3	-0.2	0.04
1	3	-0.2	0.04
1	4	0.8	0.64
1	2	-1.2	1.44
1	2	-1.2	1.44
1	3	-0.2	0.04
1	4	0.8	0.64
1	4	0.8	0.64
1	5	1.8	3.24
1	3	-0.2	0.04
1	4	0.8	0.64
1	3	-0.2	0.04
1	3	-0.2	0.04
1	2	-1.2	1.44
1	3	-0.2	0.04

Total = 48　　$Total_1$ = 10.4

Average $\overline{X}_1$ = 3.2

n_1 = 15

School_ID	Satis-faction	$(X_{2i}-\overline{X}_2)$	$(X_{2i}-\overline{X}_2)^2$
0	4	1.6	2.56
0	3	0.6	0.36
0	1	-1.4	1.96
0	3	0.6	0.36
0	2	-0.4	0.16
0	3	0.6	0.36
0	2	-0.4	0.16
0	3	0.6	0.36
0	1	-1.4	1.96
0	3	0.6	0.36
0	1	-1.4	1.96
0	2	-0.4	0.16
0	3	0.6	0.36
0	2	-0.4	0.16
0	3	0.6	0.36

Total = 36　　$Total_2$ = 11.6

Average $\overline{X}_2$ = 2.4

n_2 = 15

$$s_p^2 = \frac{Total_1 + Total_2}{(n_1)+(n_2) - 2} = \quad 0.7857$$

$$s_p = \sqrt{s_p^2} = \quad 0.886$$

$$SE = s_p \cdot \sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = \quad 0.3237$$

$$t = \frac{\overline{X}_1 - \overline{X}_2}{SE} = \quad 2.4717$$

p-value = 0.0198

*p-value is obtained by EXCEL Command: =tdist (t, n_1+n_2-2, 2)
In this case, type [=tdist(2.471658,28,2)] at any cell. You can get 0.019795.

<Advance: Effect size (Cohen's *d*)>

$$d = \frac{\left|\overline{X}_1 - \overline{X}_2\right|}{S_p} = \quad 0.90252$$

Criteria: Small: d =0.2, Medium: d=0.5, Large: d=0.8 (Source) Cohen (1988)

第 4 章　比率の検定

詳細目次

第 4 セッション 57

カイ二乗検定：適合度検定（Goodness-of-fit test）：

住民サーベイの回答「はい」「いいえ」の比率に差があったと言えるのか？

第 5 セッション 61

カイ二乗検定：独立性の検定（Chi-Square Test of Association (or Independence)）：

教育法の違いで卒業試験の結果に差があったと言えるのか？

第４セッション

カイ二乗検定：適合度検定（Goodness-of-fit test）：

住民サーベイの回答「はい」「いいえ」の比率に差があったと言えるのか？

Q. 研修受講生 100 人に対して「研修は有益でしたか？」と質問をしたところ、「はい」60 人、「いいえ」40 人という回答数を得た。この回答数に違いがあると言えるのだろうか。

<必要な条件>

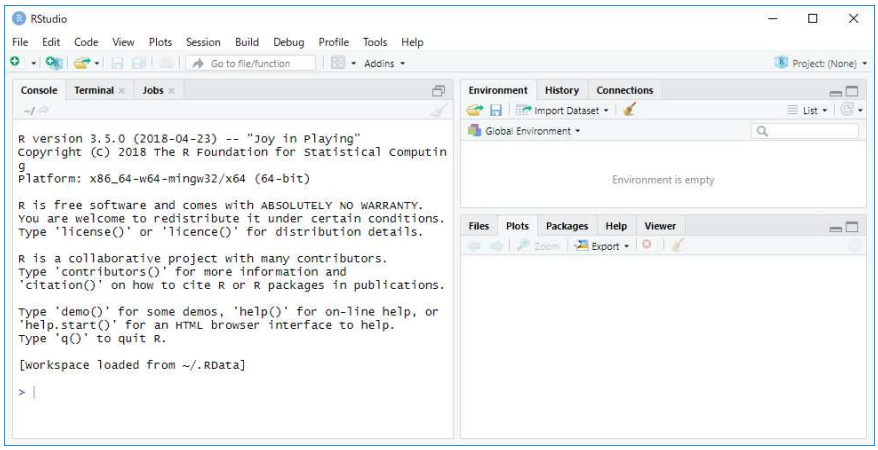
- 回答の選択肢が「はい」「いいえ」の２種類であること。

	観測度数(O)
はい	60
いいえ	40
合計	100

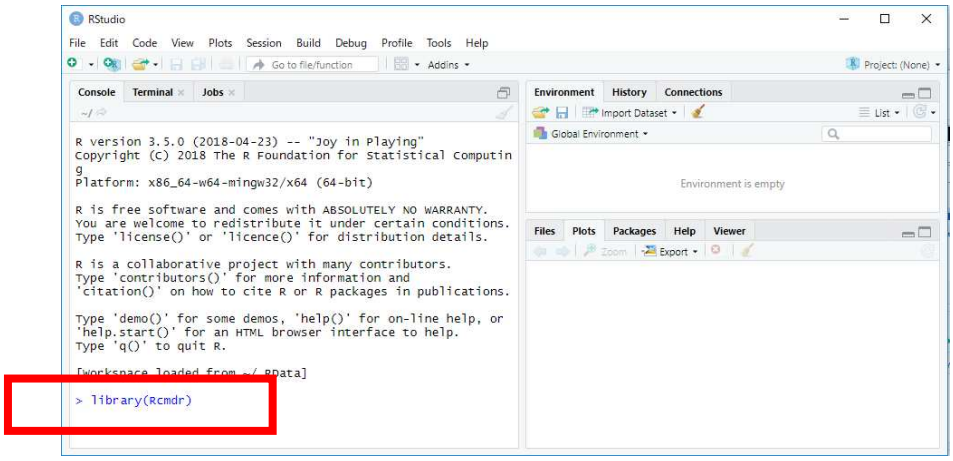
（すでに RStudio と R コマンダーを立ち上げている場合はこのページを省略してください。）

● Windows のメニュー＞“RStudio”を選ぶ。

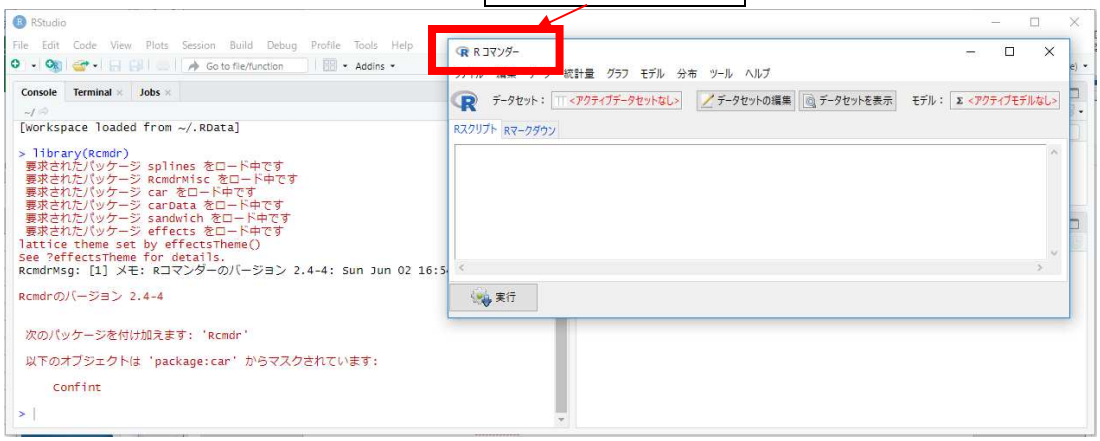
● 下のようなウィンドウが開く。



● library(Rcmdr)　と入力する。＞Enter を押す。



●下のようなメッセージが流れて、R コマンダー　のウィンドウが開く。

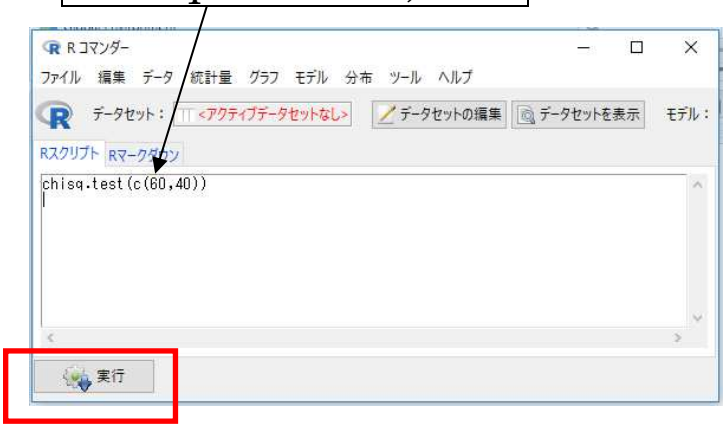


（前ページを飛ばした方はここから再開してください。）

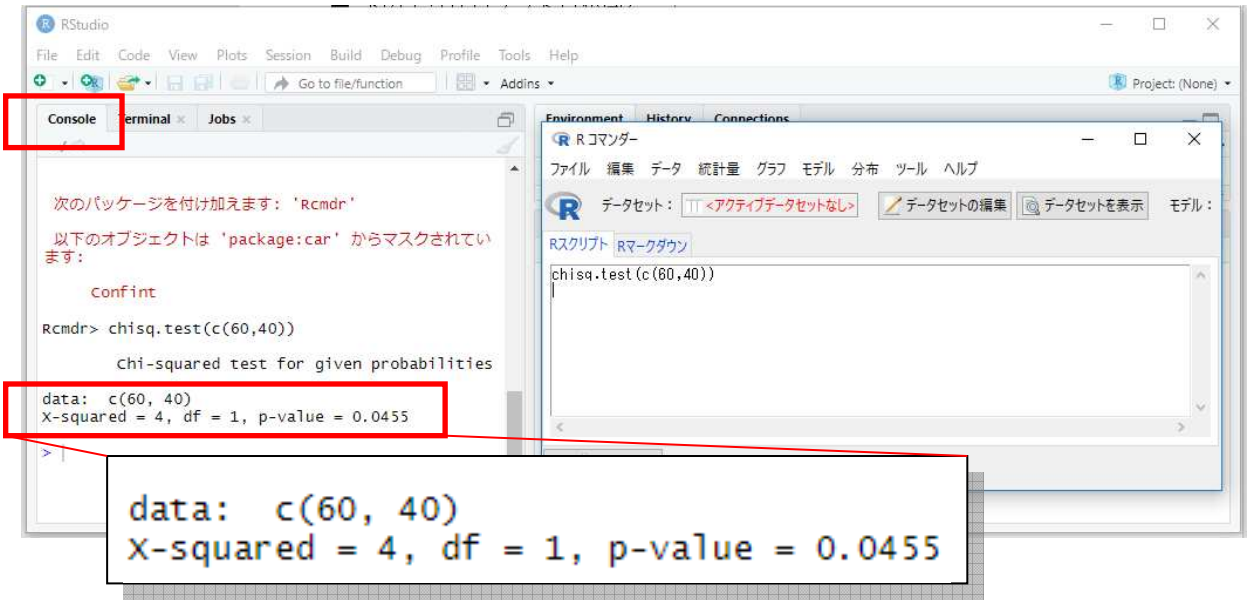
ル

● R コマンダー で、次の様に入力して、「実行」を押す。

chisq.test(c(60,40))



● RStudio の Console に、次の結果が出力される。。



data: c(60, 40)
X-squared = 4, df = 1, p-value = 0.0455

結　論

研修受講生 100 人に対して「研修は有益でしたか？」と質問をして「はい」「いいえ」のいずれかの回答を得た。この結果に関し、x^2 =4.0、p =0.046 < .05 で回答数には有意差が認められた。

表 1：研修の有益性に関する回答の比率の検定

	n	カイ二乗値	p
はい	60	4.0	0.0455
いいえ	40		
合計	100		

(The result indicated there was a statistically significant difference between the responses (yes/no), x^2 =4.0、p =0.045 < .05.)

発展学習 ： カイ二乗検定（適合度検定）の手計算

計算には以下の式を用いる。カイ二乗値（χ^2）を求めるとも言う。

$$\chi^2 = \sum \frac{(\text{観測度数} - \text{期待度数})^2}{(\text{期待度数})}$$

（１）最初に観測度数（O）を入力する。(O: Observed)

	観測度数(O)
はい	60
いいえ	40
合計	100

（２）次に期待度数（E）を入力する。期待値は１：１と考える。(E: Expected)

	観測度数(O)	期待度数(E)
はい	60	50
いいえ	40	50
合計	100	100

（３）次の３列のセルも手計算して合計値を計算する。合計値が χ^2 値である。

	観測度数(O)	期待度数(E)	(O−E)	$(O-E)^2$	$(O-E)^2/E$	
はい	60	50	10	100	2.0	
いいえ	40	50	−10	100	2.0	
合計	100	100			4.0	<= χ^2

（４）結果を解釈する。

χ^2値は 4.0 と計算された。（自由度１の場合には）χ^2 値が **3.8 以上**だと「有意差がある」と認められる。確認するために確率値(p 値)を計算してみる。

(i) まず自由度なるものを計算する。

自由度＝ 2-1 = 1

カテゴリー数から１引く。具体的には次のとおり。「はい」「いいえ」の 2−1=1

(ii)次に p 値を算出する。エクセルのコマンドは、=CHIDIST(χ^2値, 自由度)。

p =	0.0455003

=CHIDIST(4.0, 1)

得られた p 値（=0.0455）が意味するところは、今回得られたような大きな差は、偶然では 100 回中 4.55 回しか起こらないほど大きな差だということ。

第５セッション

カイ二乗検定：独立性の検定

（Chi-Square Test of Association (or Independence)：

教育法の違いで卒業試験の結果に差があったと言えるのか？

()

Q. 途上国への技術協力のひとつとして、日本型の理数科教育法導入が実施された。従来型の理数科教育法で教育を受けた 100 人のうち 60 人が合格した一方で、日本型の理数科教育法で教育を受けた 100 人のうち 75 人が合格した。日本型の理数科教育法と従来型の教育法で卒業試験の結果に差があったと言えるのか？

<必要な条件[5]>

- 横軸の選択肢が「日本型の教育法」「従来型の教育法」の２種類であること。
- 縦軸の選択肢が「合格」「不合格」の２種類であること。

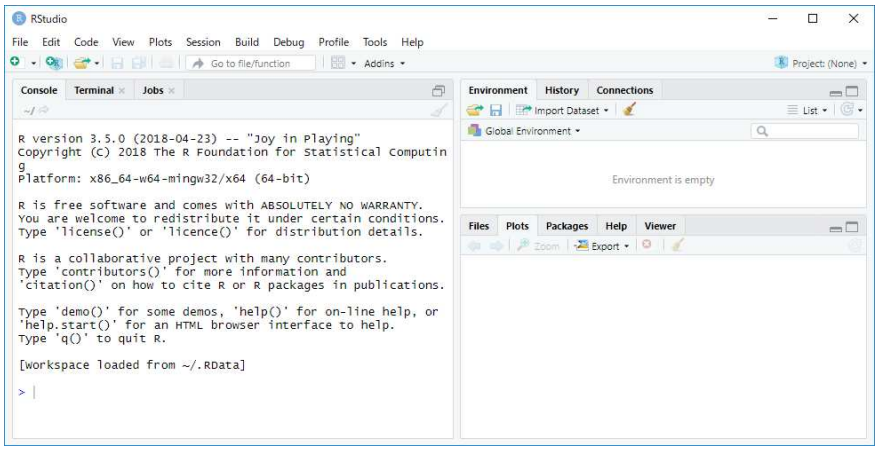
	日本型の教育法	従来型の教育法
合格	75	60
不合格	25	40
合計	100	100

[5] 選択肢のいずれか（あるいは両方）が 3 種類以上の場合にもこの章で説明する独立性検定（カイ二乗検定）が利用できるが、分散分析（Analysis of Variance (ANOVA)）および付随する post-hoc analysis という別の方法を用いることにより、より詳しく検定することができる。

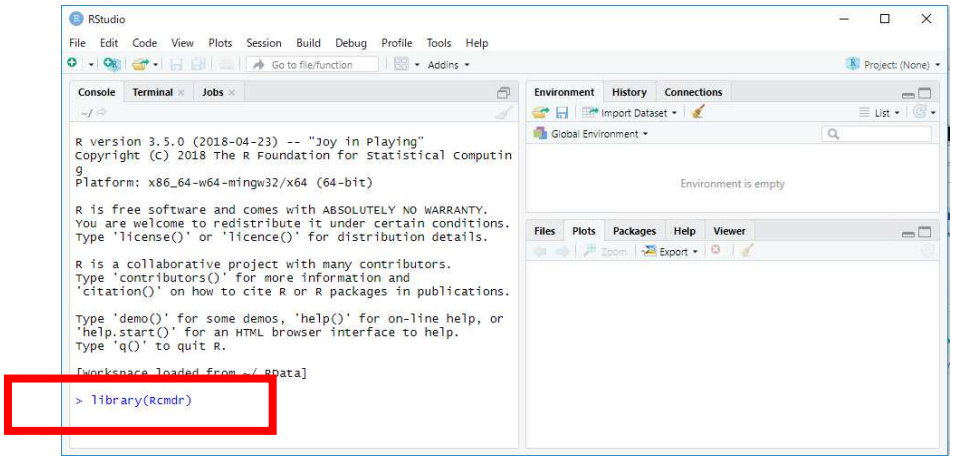
（すでに RStudio と R コマンダーを立ち上げている場合はこのページを省略してください。）

● Windows のメニュー＞“RStudio”を選ぶ。

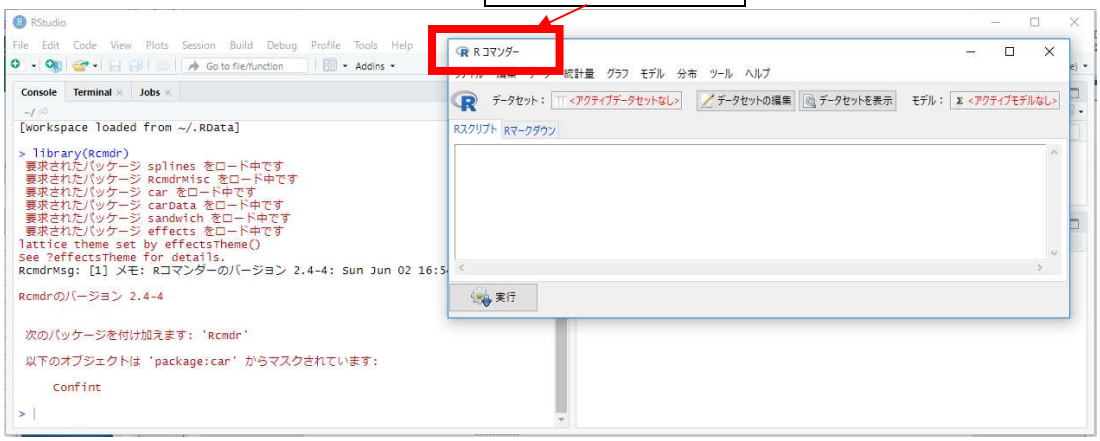
● 下のようなウィンドウが開く。



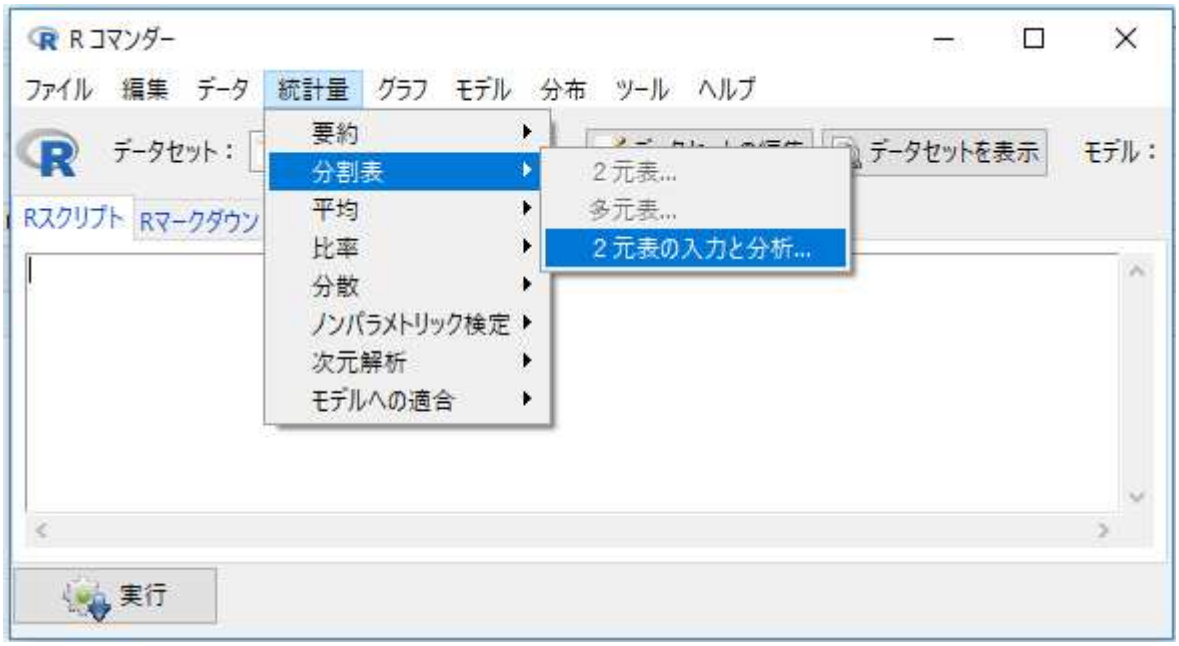
● library(Rcmdr)　と入力する。＞Enter を押す。



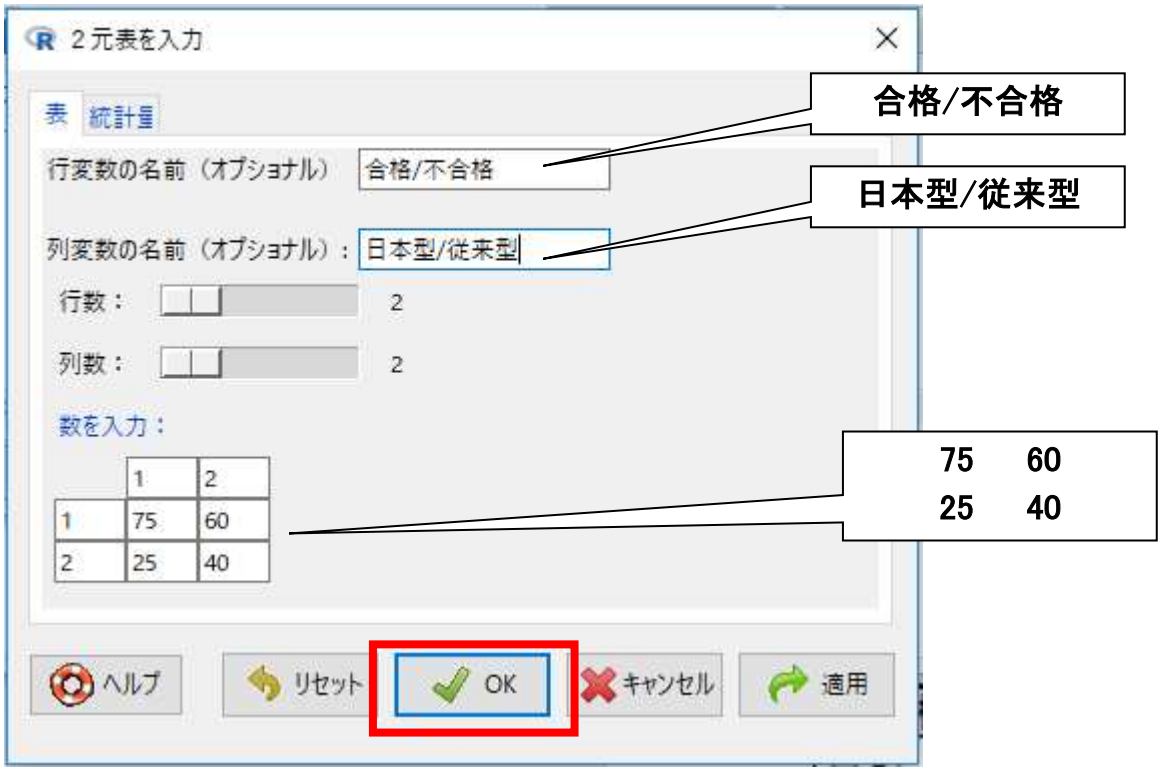
●下のようなメッセージが流れて、R コマンダー　のウィンドウが開く。



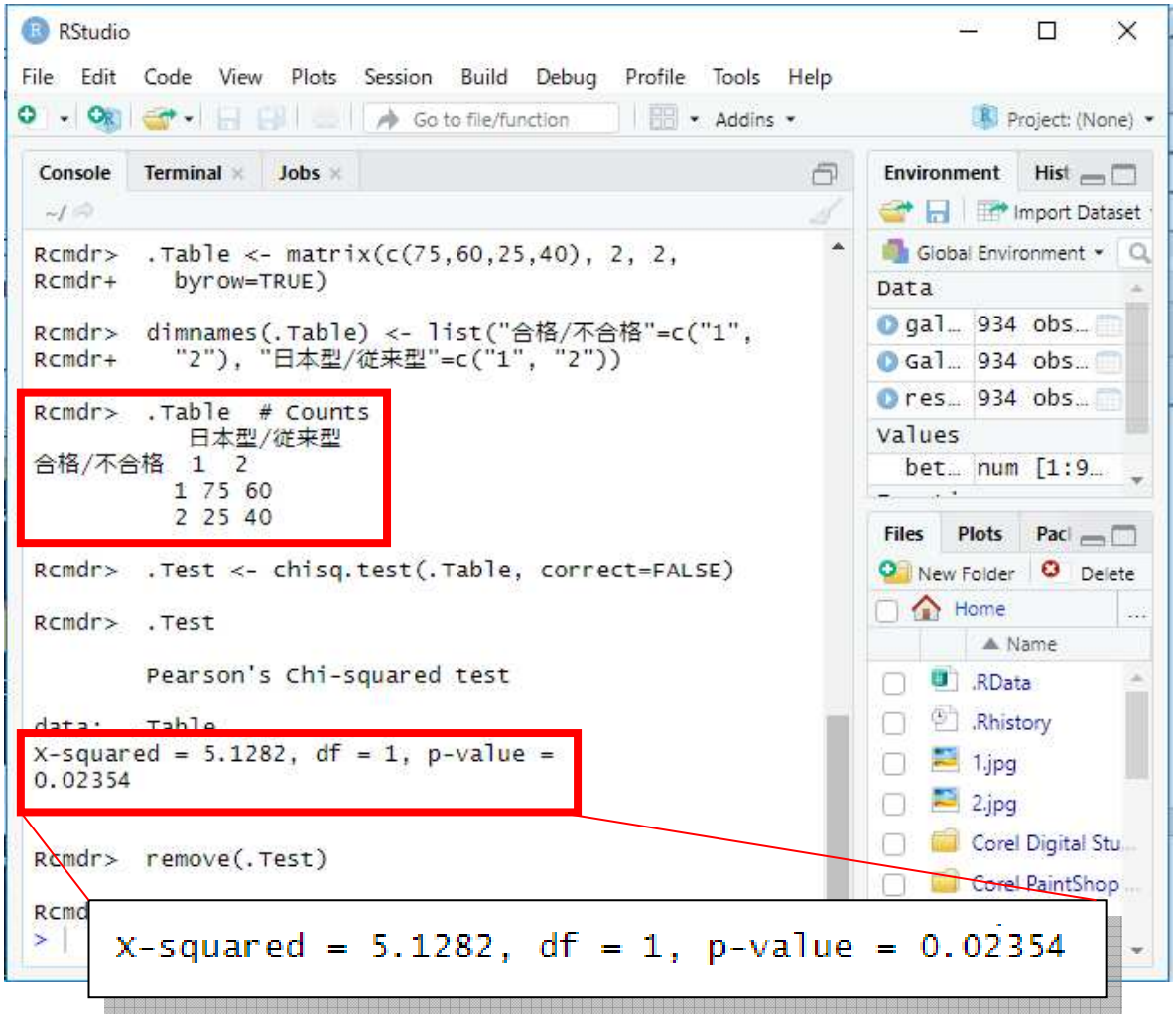
●R コマンダー で、 統計量 ＞ 分割表 ＞ 2 元素の入力と分割 を選ぶ。



●現れたウィンドウで以下のように 3 か所を入力する。OK を押す。



● すると、 RStudio の Console に以下のような出力がある。



結　論

日本型の理数科教育法の授業を受けた生徒の卒業試験の合格者数と従来型の理数科教育法を受けた生徒の卒業試験の合格者数を比べた。その結果、x^2=5.128、p=0.024 < .05 であったので、教授法と合格者数の間には統計学的に有意な関係があると結論できる。

表 1：教育法の違いによる合格者数の比率の検定

	日本型の教育法	従来型の教育法	カイ二乗値	p
合格	75	60	5.128	0.0235
不合格	25	40		
合計	100	100		

(The result indicated there was a statistically significant relationship between the methods of teaching (Japanese-style/traditional style) and the number of students passing the graduation examination, x^2=5.128、p=0.024 < .05.)

発展学習１：カイ二乗検定（独立性の検定）の手計算

計算には以下の式を用いる。カイ二乗値（X^2）を求めるとも言う。

$$\chi^2 = \sum \frac{(\text{観測度数} - \text{期待度数})^2}{(\text{期待度数})}$$

（１）最初に観測度数（O）を入力する(O: Observed)。

＜観測度数＞

	日本型の教育法	従来型の教育法	合計
合格	75	60	135
不合格	25	40	65
合計	100	100	200

（２）次に期待度数（E）を計算する（E: Expected)。たとえば、総合計数（$n_{合計}$=200）における総合格者数($n_{合格}$=135)の比率は、135÷200=0.675 となるので、その比率に「日本の教育法」の受講者数（$n_{日本型の教育法}$=100）を掛けて、67.5 人が期待度数となる。他のセルも同様に計算する。

＜期待度数＞

	日本型の教育法	従来型の教育法	合計	
合格	67.5	67.5	135	
不合格	32.5	32.5	65	
合計	100	100	200	<=総合計数

（３）上記（１）（２）の表から次のように書き写す。

教育法	合格_不合格	観測度数(O)	期待度数(E)
日本型の教育法	合格	75	67.5
日本型の教育法	不合格	25	32.5
従来型の教育法	合格	60	67.5
従来型の教育法	不合格	40	32.5
合計度数		200	200

（４）その次の 3 つの列も手計算して合計値を計算する。合計値が X^2値である。

教育法	合格_不合格	観測度数(O)	期待度数(E)	(O−E)	$(O-E)^2$	$(O-E)^2/E$	
日本型の教育法	合格	75	67.5	7.5	56.25	0.833	
日本型の教育法	不合格	25	32.5	−7.5	56.25	1.731	
従来型の教育法	合格	60	67.5	−7.5	56.25	0.833	
従来型の教育法	不合格	40	32.5	7.5	56.25	1.731	
合計度数		200	200			5.128	<= χ^2

（５）結果を解釈する。

χ^2値は 5.128 と計算された。（自由度 1 の場合には）χ^2 値が **3.8 以上**だと「有意差がある」と認められる。確認するために確率値(p 値)を計算してみる。

(i) まず自由度なるものを計算する。

自由度＝(2-1) * (2-1) = 1 * 1 = 1.

縦軸（この場合「教育法」）のカテゴリー数から１引く。具体的には次のとおり。
「日本～」「従来～」の２－１＝１

横軸≪この場合「合格_不合格」）のカテゴリー数から１引く。具体的には次のとおり。
「合格」「不合格」の２ －１＝１

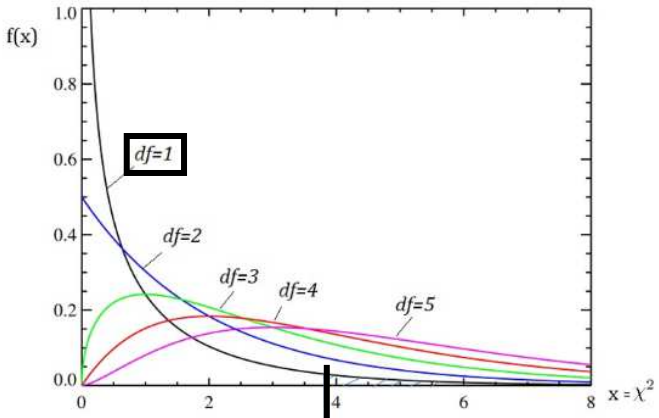
(ii)次に p 値を算出する。エクセルのコマンドは、=CHIDIST(X^2値, 自由度)。

p =	0.023540058

=CHIDIST(5.128, 1)

得られた p 値（=0.024）が意味するところは、今回得られたような大きな差は、偶然では 100 回中 2.4 回しか起こらないほど大きな差だということ。

カイ二乗分布（Chi-square distribution）



X^2=3.8（これより右側の面積はわずか 5%(p=0.05)

（注意）期待度数のセルのいずれかの度数が 5 より少ない（n<5）の場合には、フィッシャーの正確性検定（Fisher's exact test）という別の方法を用いる。第 1～３セッションいずれにも当てはまる前提条件である。なお、もちろんフィッシャーの正確性検定（Fisher's exact test）も手計算が可能である。

第 5 章　回帰分析

詳細目次

第 6 セッション: 68

回帰分析１:　クロスセクションデータを用いる

第 7 セッション: 79

回帰分析２:　時系列データを用いる

第 8 セッション: 89

プロビット回帰分析:「はい」「いいえ」の回帰分析

第 6 セッション

回帰分析１:クロスセクションデータを用いる

Q. E 市は、ゴミの総量を削減するために、有料ゴミ袋政策を導入した（“PaTrash”と命名)。この政策は、市のゴミ収集車が、事前に購入された有料ゴミ袋で出されたゴミしか収集しないとする政策である。周辺の他の市は、E 市が成功するかどうかを見ている。そして、主要な市のゴミ収集量が記録された。

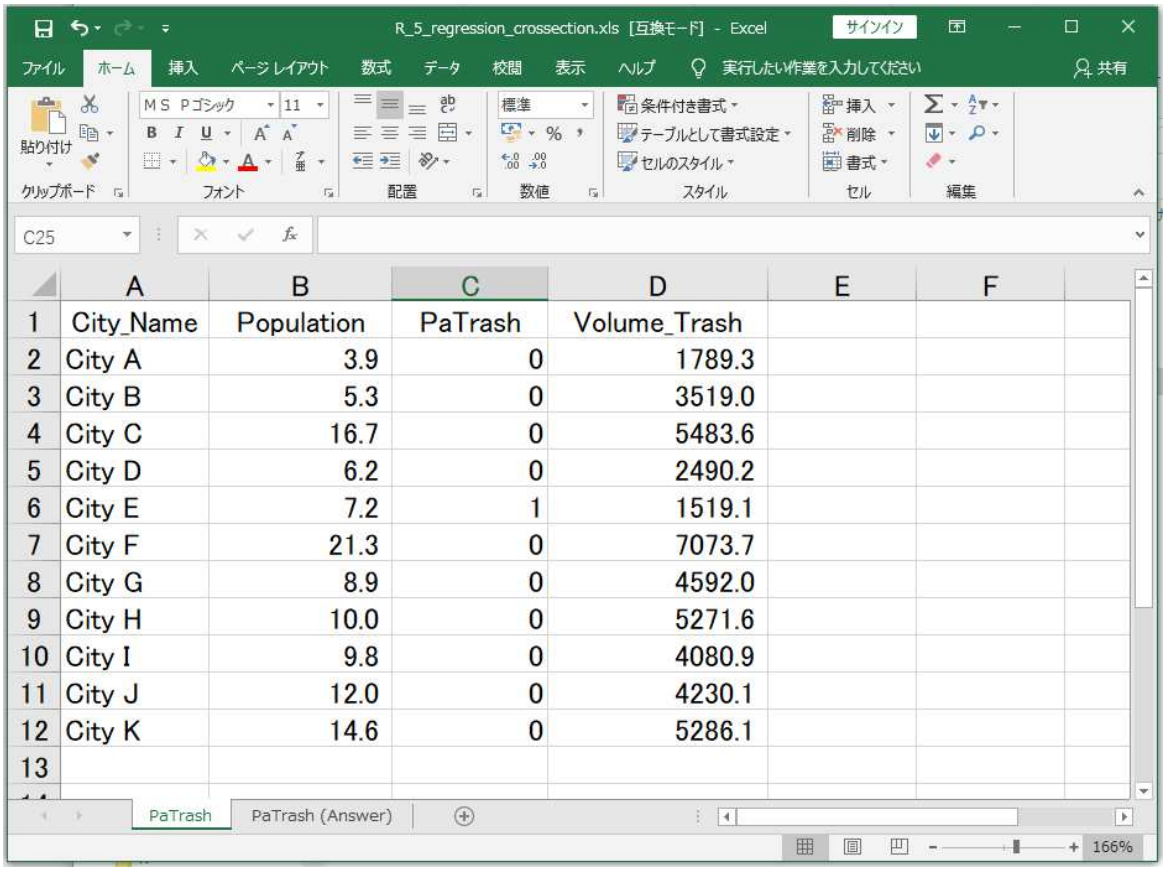
この新しい政策によってゴミ収集量は減ったのか？

<必要な条件>

- 成果に関する指標（Outcome indicators (“Volume of trash”))、介入行為に関する指標（Intervention indicators）(この場合は “PaTrash”)、 さらに、説明変数に関する指標（この場合は “人口”）のデータが必要である。
- 説明変数の間に強い相関関係がない。

Input of data

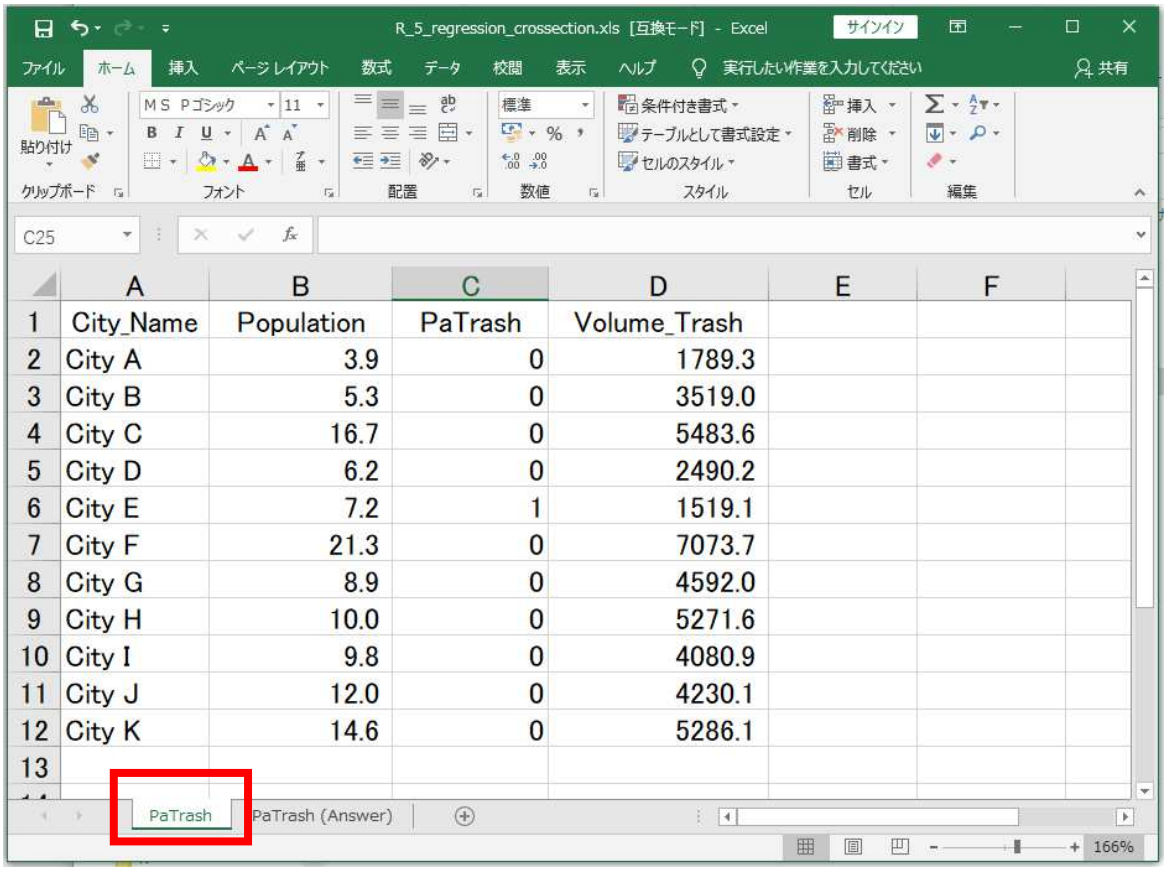
第 1 行 ....市名
第 2 行 ... 人口（,000 人）
第 3 行 ...“PaTrash” の導入の有無(1=Yes, 0=No)
第 4 行 ... ゴミ収集量（トン）



	A	B	C	D
1	City_Name	Population	PaTrash	Volume_Trash
2	City A	3.9	0	1789.3
3	City B	5.3	0	3519.0
4	City C	16.7	0	5483.6
5	City D	6.2	0	2490.2
6	City E	7.2	1	1519.1
7	City F	21.3	0	7073.7
8	City G	8.9	0	4592.0
9	City H	10.0	0	5271.6
10	City I	9.8	0	4080.9
11	City J	12.0	0	4230.1
12	City K	14.6	0	5286.1
13				

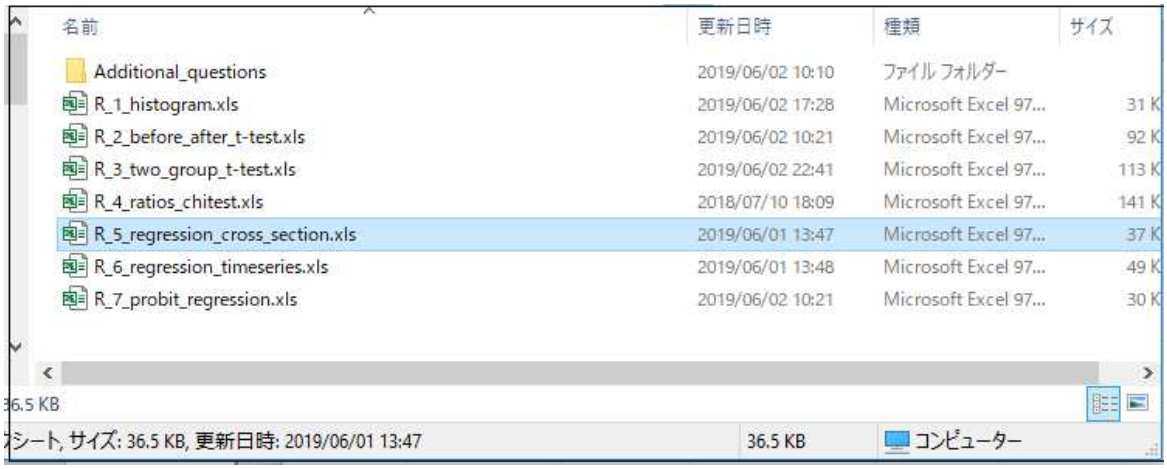
データの入力と読み込み

●以下のようにデータを入力する。タブに「PaTrash」などの名前をつける。



	A	B	C	D
1	City_Name	Population	PaTrash	Volume_Trash
2	City A	3.9	0	1789.3
3	City B	5.3	0	3519.0
4	City C	16.7	0	5483.6
5	City D	6.2	0	2490.2
6	City E	7.2	1	1519.1
7	City F	21.3	0	7073.7
8	City G	8.9	0	4592.0
9	City H	10.0	0	5271.6
10	City I	9.8	0	4080.9
11	City J	12.0	0	4230.1
12	City K	14.6	0	5286.1
13				

●デスクトップに、R_5_regression_cross_section.xls などと適当な名前をつけて保存する。

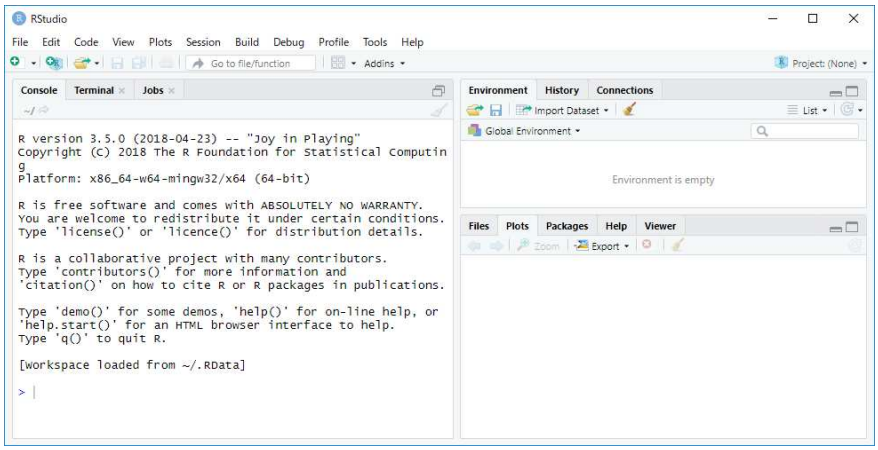


名前	更新日時	種類	サイズ
Additional_questions	2019/06/02 10:10	ファイル フォルダー	
R_1_histogram.xls	2019/06/02 17:28	Microsoft Excel 97...	31 K
R_2_before_after_t-test.xls	2019/06/02 10:21	Microsoft Excel 97...	92 K
R_3_two_group_t-test.xls	2019/06/02 22:41	Microsoft Excel 97...	113 K
R_4_ratios_chitest.xls	2018/07/10 18:09	Microsoft Excel 97...	141 K
R_5_regression_cross_section.xls	2019/06/01 13:47	Microsoft Excel 97...	37 K
R_6_regression_timeseries.xls	2019/06/01 13:48	Microsoft Excel 97...	49 K
R_7_probit_regression.xls	2019/06/02 10:21	Microsoft Excel 97...	30 K

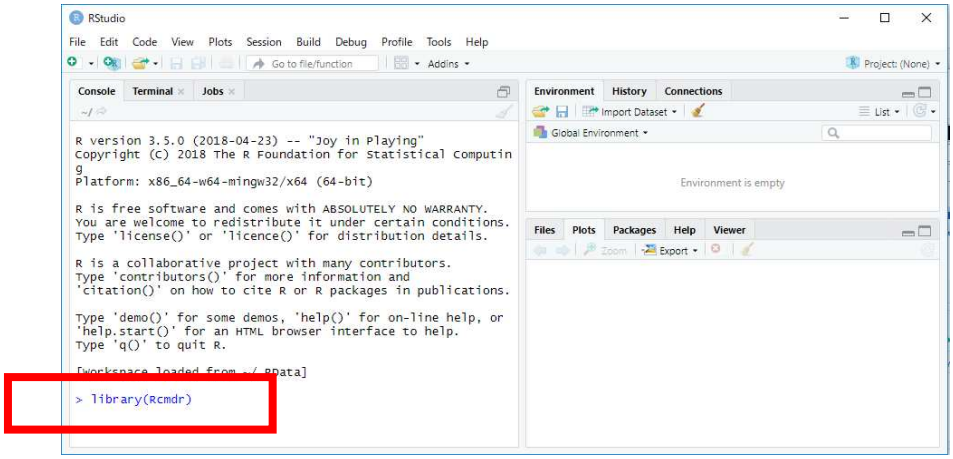
（すでに RStudio と R コマンダーを立ち上げている場合はこのページを省略してください。）

● Windows のメニュー＞“RStudio”を選ぶ。

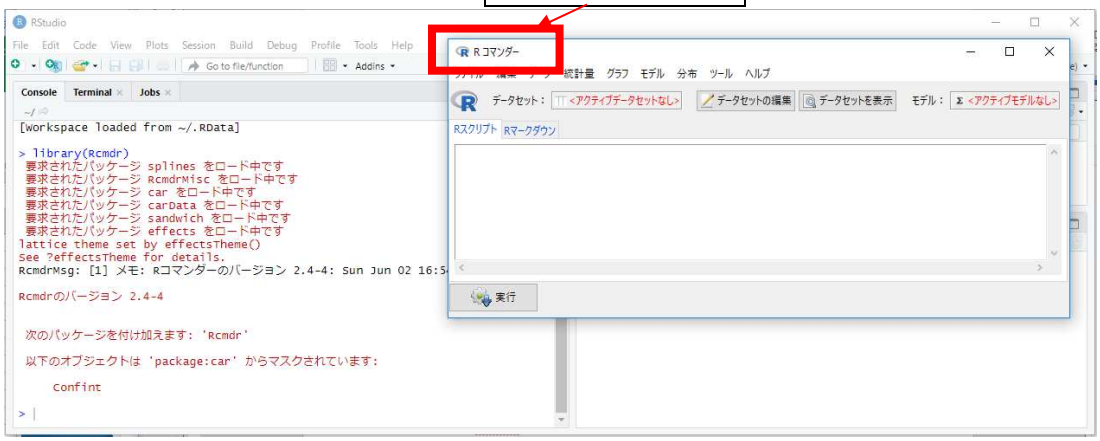
● 下のようなウィンドウが開く。



● ” library(Rcmdr) “と入力する。＞Enter を押す。

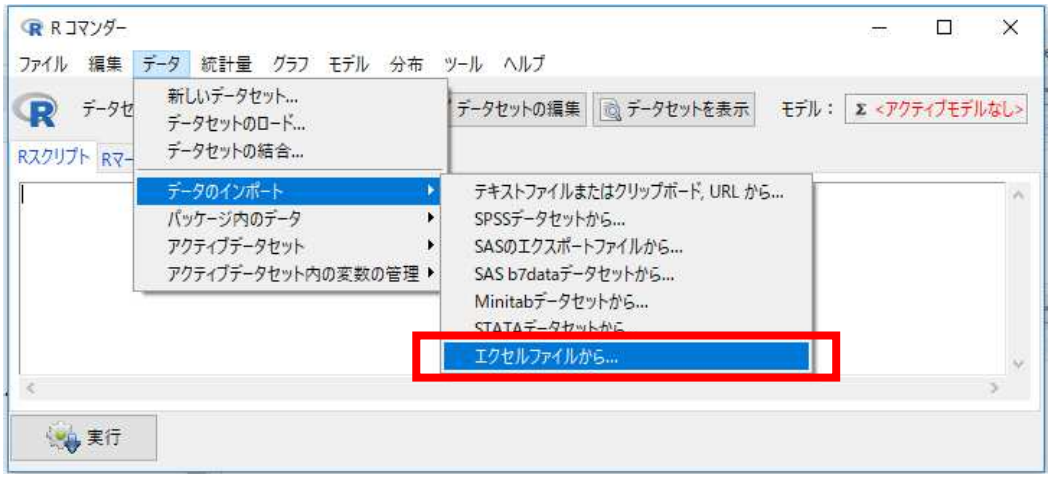


●下のようなメッセージが流れて、R コマンダー のウィンドウが開く。

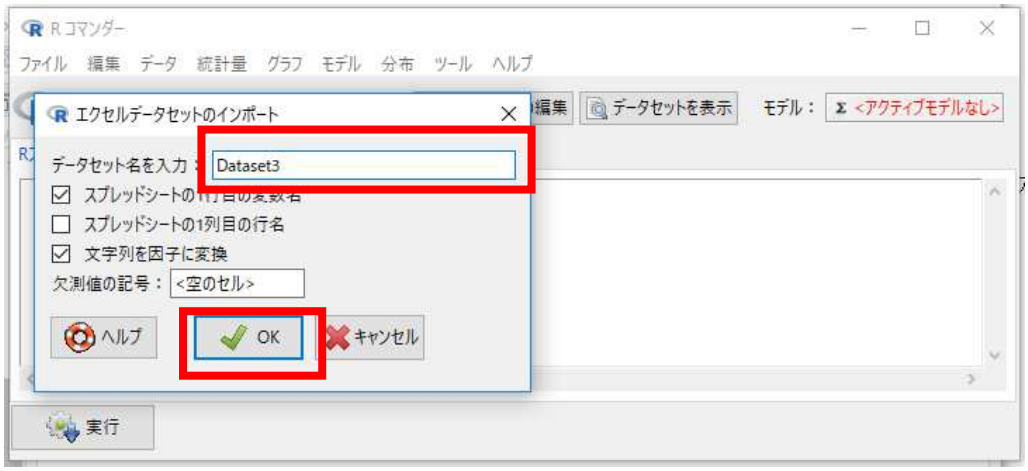


（前ページを飛ばした方はここから再開してください。）

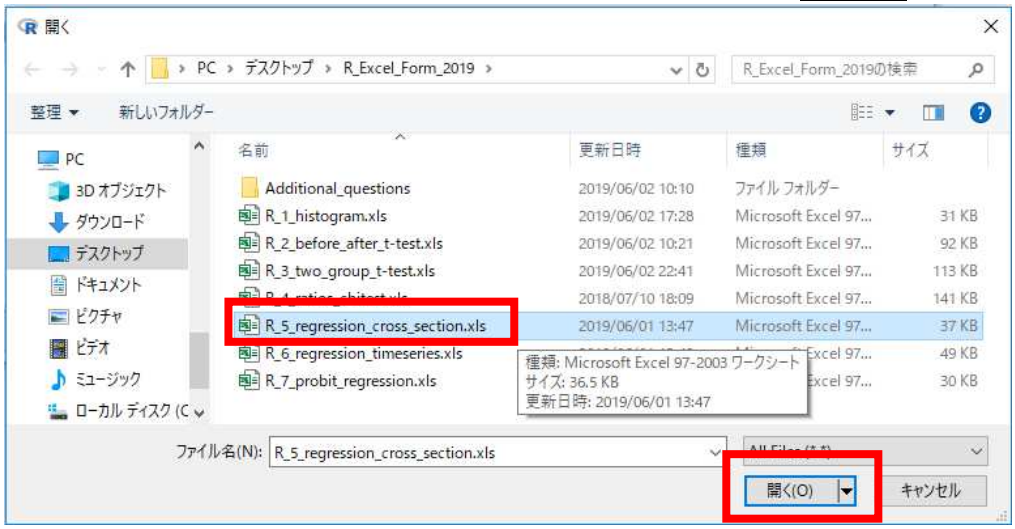
● R コマンダー のウィンドウで、データ ＞ データのインポート ＞ エクセルファイルから を選ぶ。



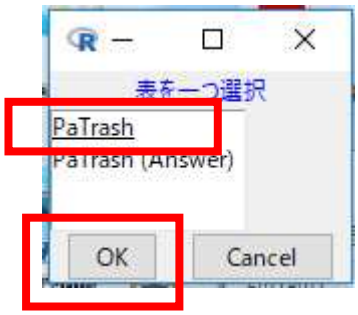
●現れたウィンウィンドウで、“データセットを入力” に “Dataset3” と入力する。＞ OK を押す。



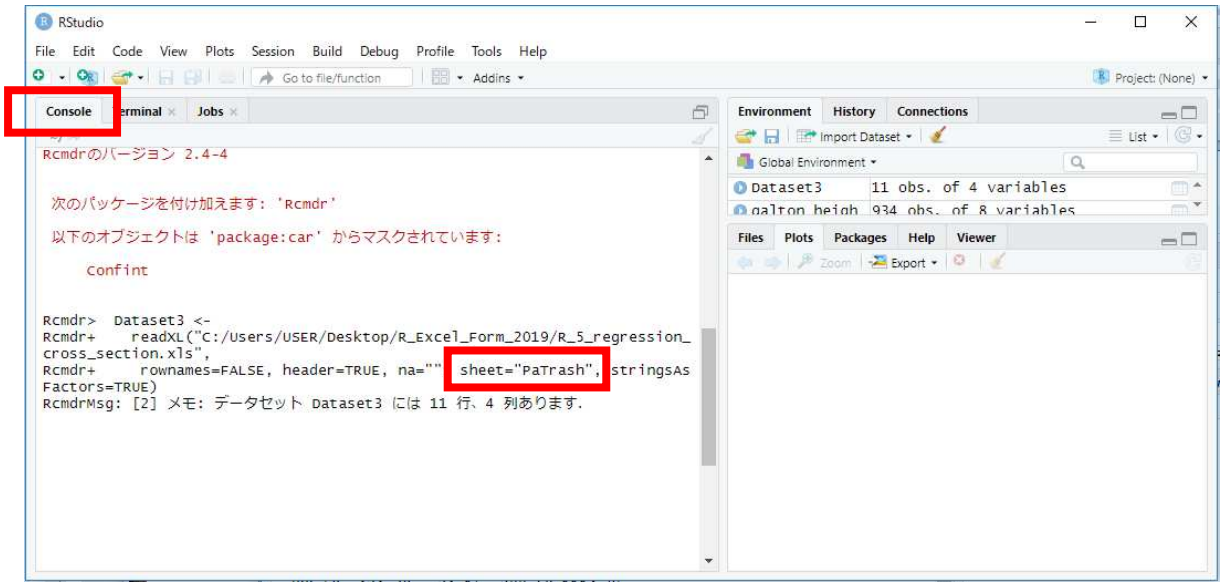
●さきほど保存した Excel ファイルを選択する。＞ 開く を押す。



●もし Excel ファイルに２枚のシートが格納されていると下のような小さなウィンドウが開く。＞　開きたいシートを選ぶ。この場合は、PaTrash 。＞ OK を押す。

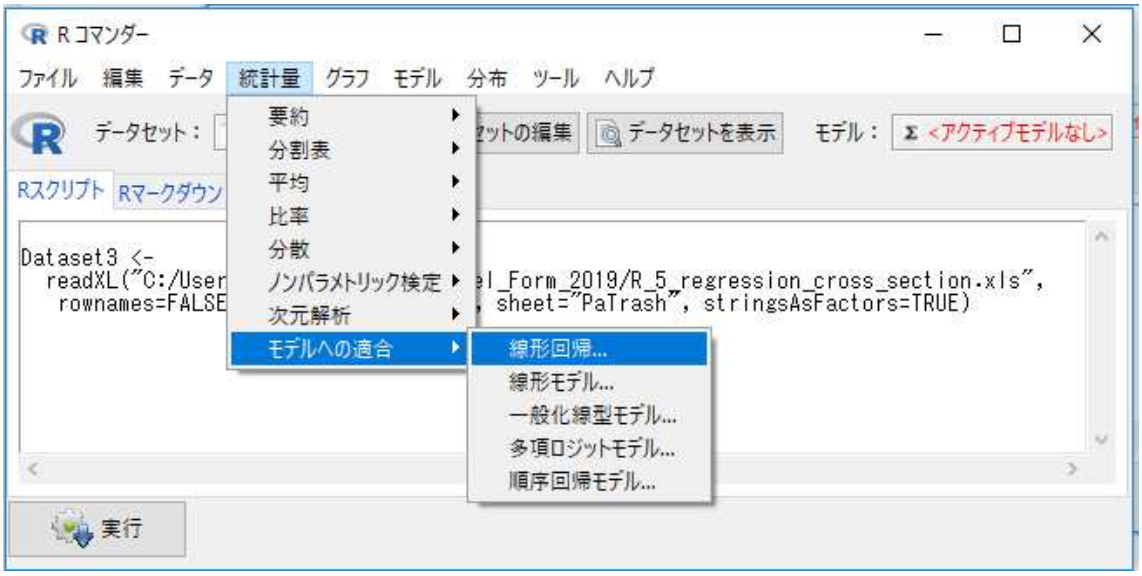


●すると RStudio の Console で、次のような表示が現れる。Excel ファイルの当該シート（ PaTrash ）の読み込みが成功した。

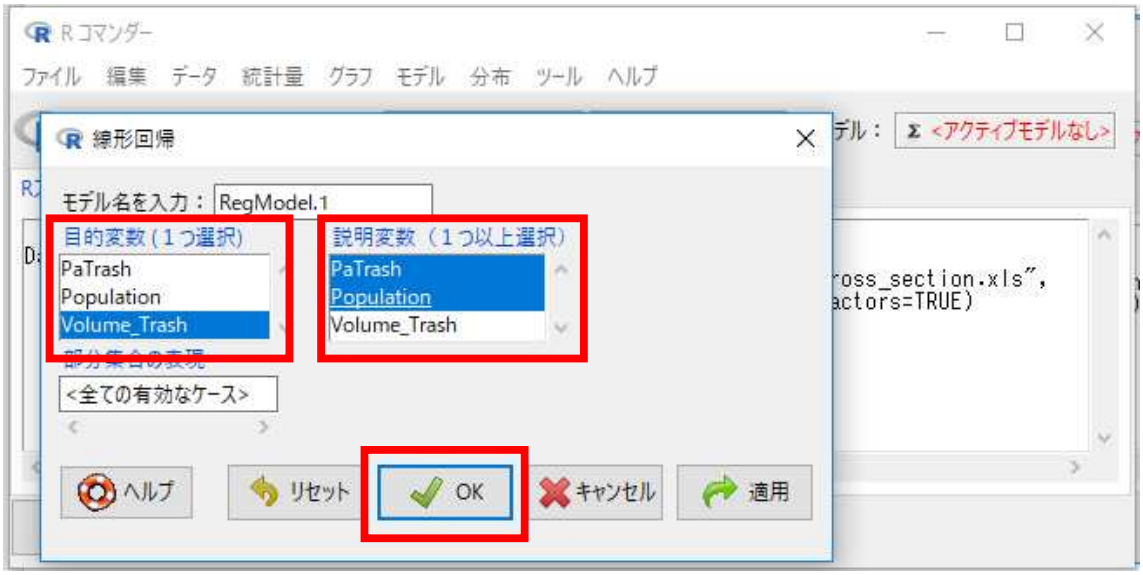


「線形回帰」の選択

● R コマンダー のメニューから 統計量 ＞ モデルへの適合 ＞ 線形回帰 を選ぶ。



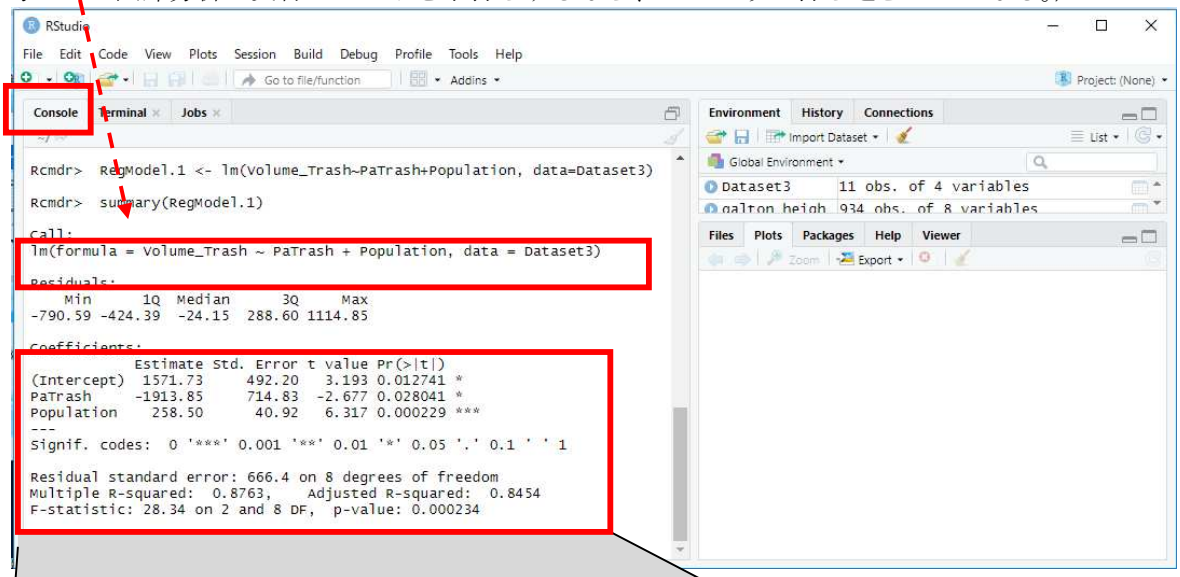
●現れたウィンドウで、「目的変数（1 つ選択）」で Volume_Trash を選ぶ。
「説明変数（1 つ以上選択）」で PaTrash と Population を選ぶ。 OK を押す。



分析結果を得る

● RStudio の Console に以下にように出力される。

（参考まで：回帰分析の実行コマンドを手打ちするなら、このように打ち込むことになる。）



③ 係数（傾き）：それぞれの変数が成果指標の変化をどれくらい説明するかを表す。このケースだと、“PaTrash”の導入によりゴミ収集量が 1,913.85 トン減少する。また、人口 1,000 人の増加により 258.50 トンのゴミ収集量が増加する。

④ 各係数の t 値と p 値：各変数の統計学的な有意度を表す。もし p 値 <0.05 なら、統計学的に有意である。この場合は、二つの変数とも統計学的に有意と判断される。

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1571.73     492.20   3.193 0.012741 *
PaTrash     -1913.85     714.83  -2.677 0.028041 *
Population    258.50      40.92   6.317 0.000229 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 666.4 on 8 degrees of freedom
Multiple R-squared:  0.8763,    Adjusted R-squared:  0.8454
F-statistic: 28.34 on 2 and 8 DF,  p-value: 0.000234
```

① R-Square (R2): 得られた等式でどの程度の変化量が説明されるか。この場合は、87.63%が等式によって説明される。

② F -statistic：等式全体の統計学的な有意度を表す。もし p-値<0.05 なら、等式は全体として有意である。このケースでは、F=28.34、p-value は 0.000234 であり、0.05 以下である。

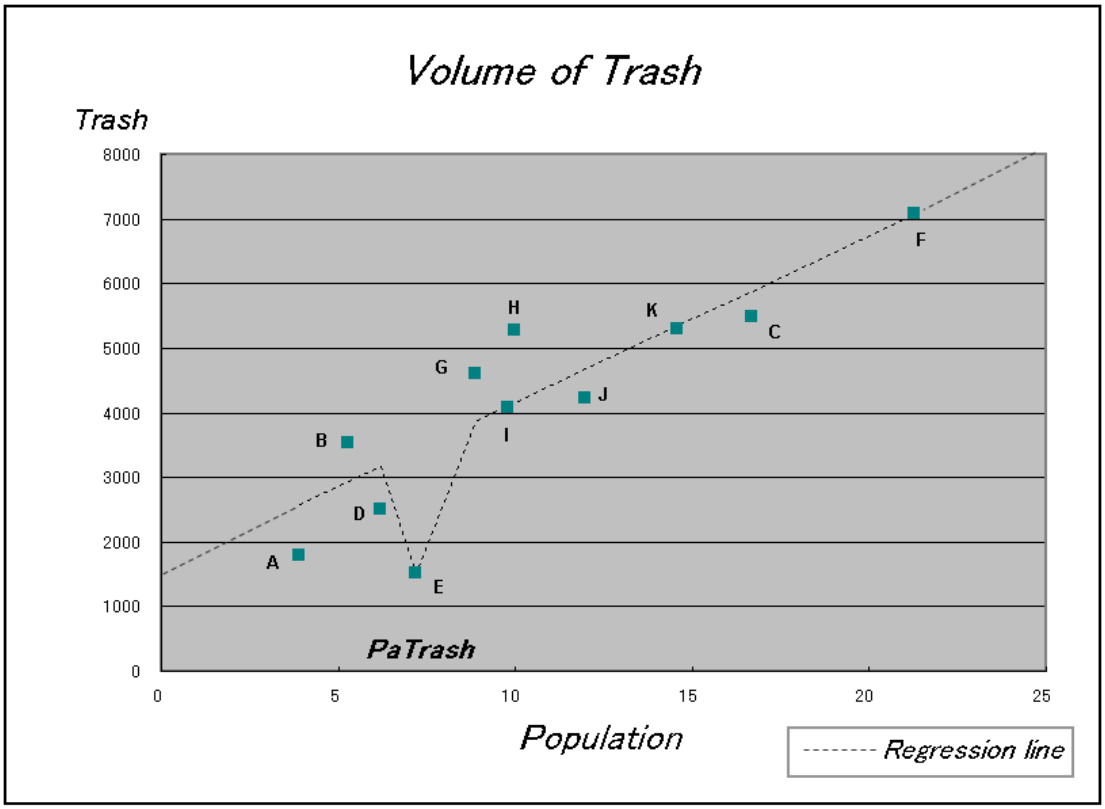
Volume_Trash＝1571.73+258.50×(population) -1913.85×（”PaTrash”）

分析結果

このケースでは、ゴミ収集量は、有料ゴミ袋政策によって減少したと結論された。当該政策によるゴミ収集量の減少は、1,914 トンと推定された。

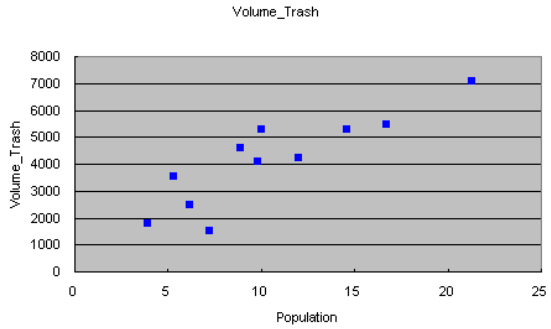
ところで、ゴミ収集量と人口の間に強い相関が観察され、それは統計学的に有意であることが確認された。つまり、人口増加によるゴミ収集量の増加が予想される。

(注意) 下のような散布図（Scatter plots）は読者の理解を高めるだろう。また、回帰分析の線を追加することも勧められる（ただし手書で加えるしかない)。

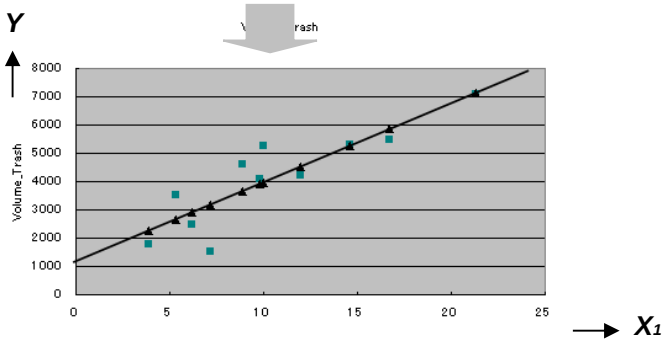


・・・新しい政策（“PaTrash”）は、ゴミ収集量を 1,914 トン減少させたと推定された。

発展学習 1：回帰分析を理解するための視覚的説明

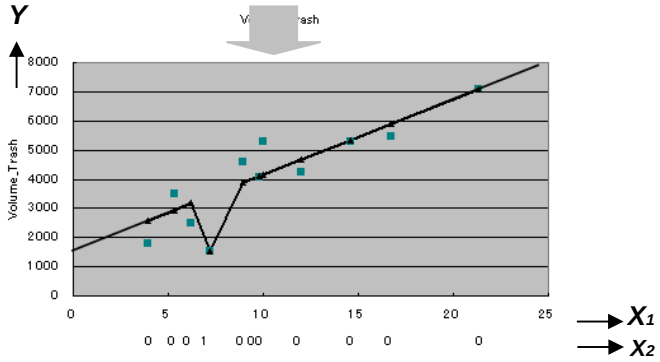


これはもとの散布図である。



***Y* は *X1* によって説明される**

ただ *X1*（人口のみ）を使う。



***Y* は *X1* と *X2* によって説明される。**

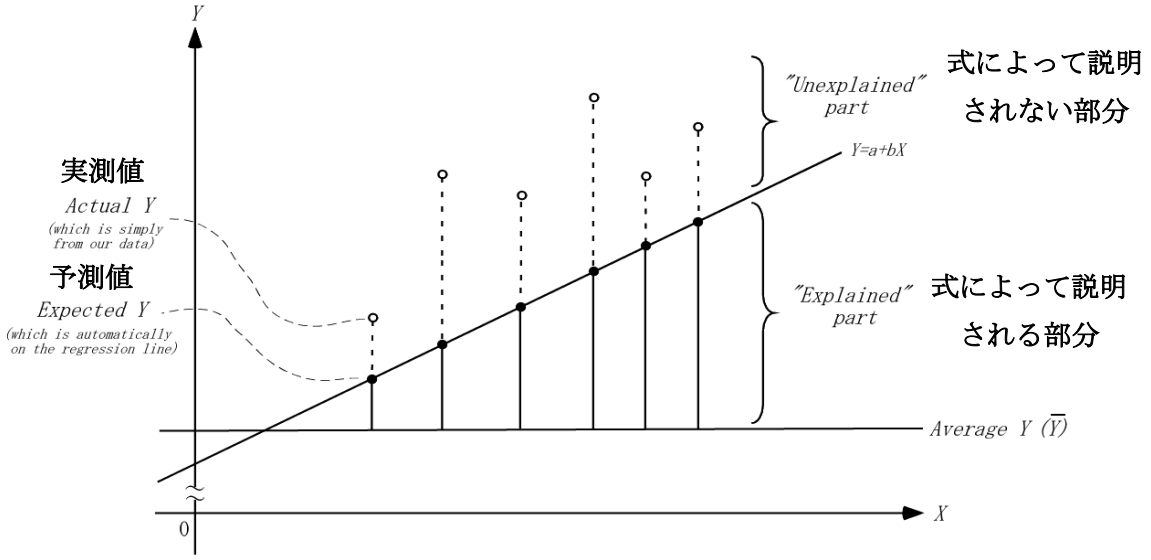
X2（PaTrash）を加える。.

X2：0（No）
　　1（Yes）=>線は下落する。.

……もし私が３時限の空間を使えるなら、グラフを３次元で正確に示すことができるだろう。なぜなら３軸を利用できるからだ（*Y, X1 & X2*）。しかし、紙は２次元なので、上記のように表した。

発展学習 2：R2 と F−値は何を意味しているのか?

以下が、R^2 と F における分子と分母のイメージである。



R^2 の公式は、

$$R^2 = \frac{\text{"Explained"}}{\text{"Explained"} + \text{"Unexplained"}}$$

簡単に言うと, R^2 とは、全体のうち "説明される部分"の比率はいくらかということである。(注：全体="説明される部分" + "説明されない" 部分)。

大まかに言って、もし比率が 0.60 (60%)以上なら, 等式は状況をよく説明していると言えるかも知れない。

F 値の公式は、

$$F = \frac{\dfrac{\text{"Explained"}}{(p-1)}}{\dfrac{\text{"Unexplained"}}{(n-p)}}$$

簡単に言うと, F-値は "説明される部分" は "説明されない部分" の何倍かということである。(注：ただしサンプルサイズ（n）と係数の数(p)で修正を施すので純粋のその比率というわけではない。)

大まかに言って、もし F-値が 3~4 (3.0~4.0)倍以上なら、対応する p-値は自動的に 0.05 (5%)以下になる。 そして我々は、式は全体として統計学的に有意だと結論する。

(注意) 実際の Y のドットは、回帰線の下にも存在するが、説明を単純にするため、回帰線の上の点だけを描いている。

第 7 セッション:

回帰分析2:時系列データを用いる

Q. あなたが経営する小売チェーン（“Shop Gold”と命名）は新しいセールスイベントを導入した(「どいち」“FirstSaturdy”と命名). それは、毎月第一土曜日の早朝に開催される。 そして、半年にわたって売上高が記録された。

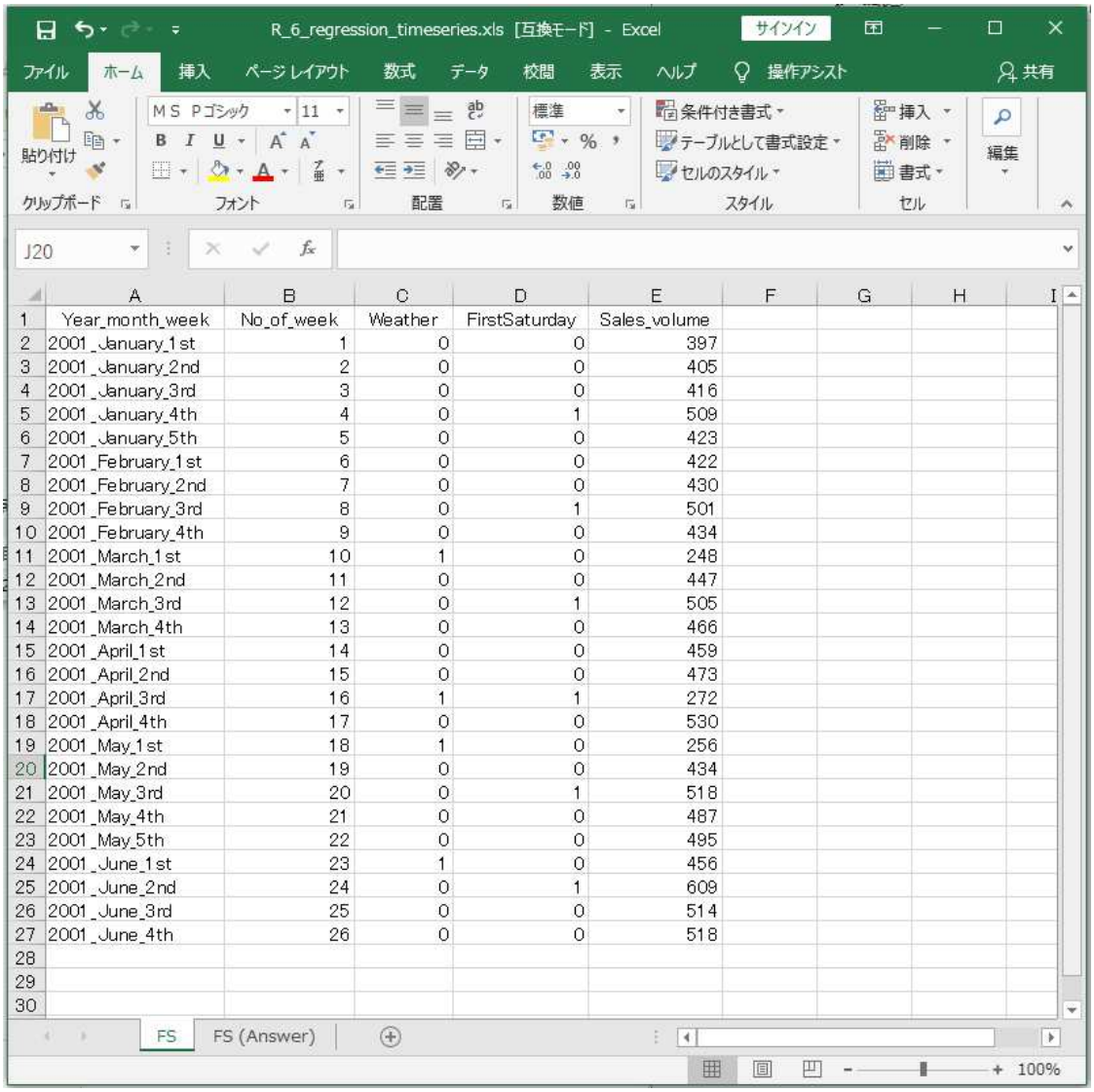
この新しいセールスキャンペーンの導入によって売上高は増加したと言えるだろうか。

<必要な条件>

- 成果に関する指標（Outcome indicators (“Sales”)）、介入行為に関する指標（Intervention indicators）（この場合は “FirstSaturday”）、 さらに、説明変数に関する指標（この場合は”週数”と “天候”）のデータが必要である。
- 説明変数の間に強い相関関係がない。

データの用意

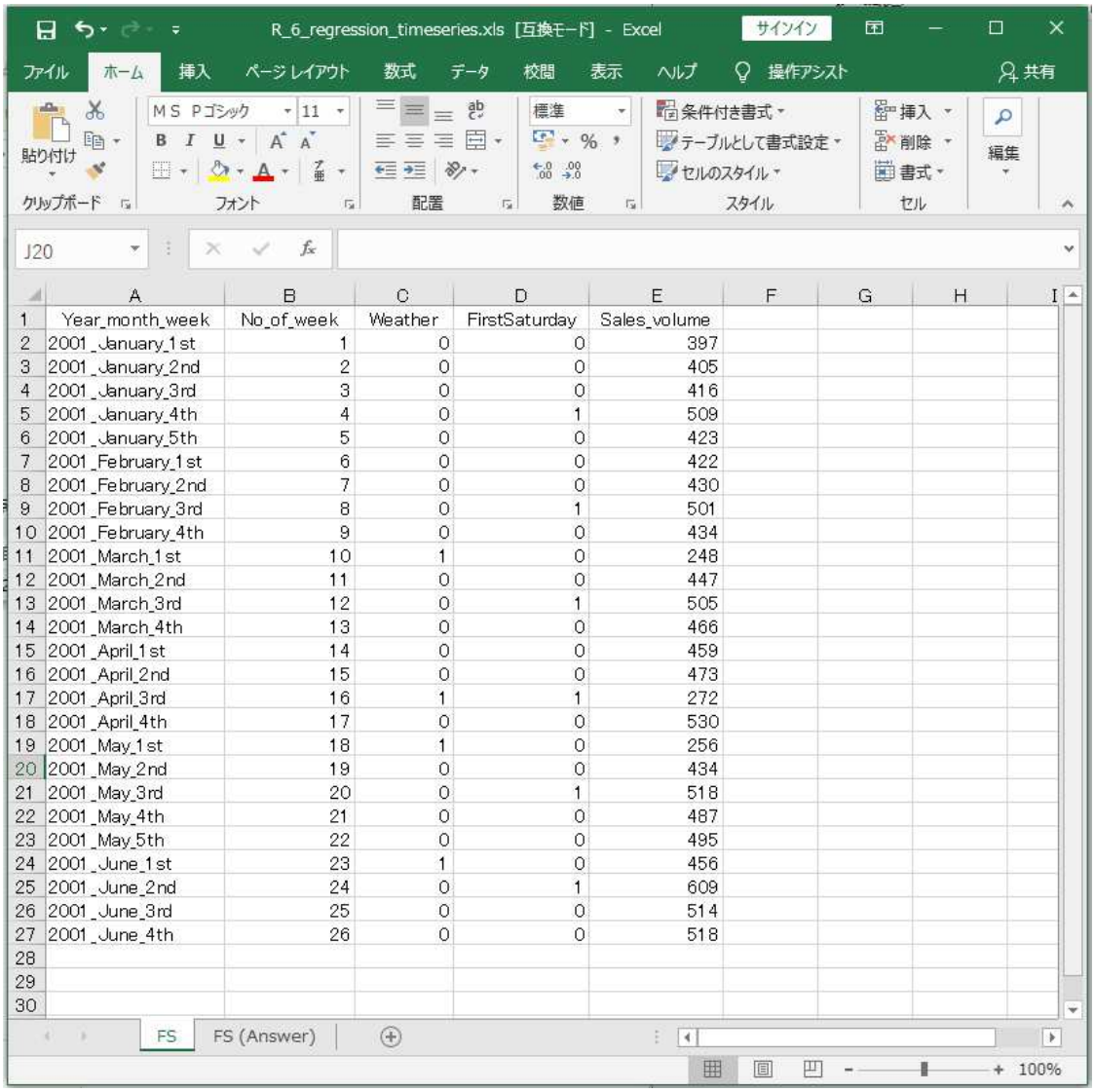
第 1 行 …年／月／週（Year/month/week）
第 2 行 …週数（No. of week）
第 3 行 …天候（Weather（１＝雨、0＝晴れ/曇り））
第 4 行 …「どいち」（“FirstSaturday” (1=Yes, 0=No)）
第 5 行 … 売上高（US$,000）（Sales volume (US$ thousands)）



Year_month_week	No_of_week	Weather	FirstSaturday	Sales_volume
2001_January_1st	1	0	0	397
2001_January_2nd	2	0	0	405
2001_January_3rd	3	0	0	416
2001_January_4th	4	0	1	509
2001_January_5th	5	0	0	423
2001_February_1st	6	0	0	422
2001_February_2nd	7	0	0	430
2001_February_3rd	8	0	1	501
2001_February_4th	9	0	0	434
2001_March_1st	10	1	0	248
2001_March_2nd	11	0	0	447
2001_March_3rd	12	0	1	505
2001_March_4th	13	0	0	466
2001_April_1st	14	0	0	459
2001_April_2nd	15	0	0	473
2001_April_3rd	16	1	1	272
2001_April_4th	17	0	0	530
2001_May_1st	18	1	0	256
2001_May_2nd	19	0	0	434
2001_May_3rd	20	0	1	518
2001_May_4th	21	0	0	487
2001_May_5th	22	0	0	495
2001_June_1st	23	1	0	456
2001_June_2nd	24	0	1	609
2001_June_3rd	25	0	0	514
2001_June_4th	26	0	0	518

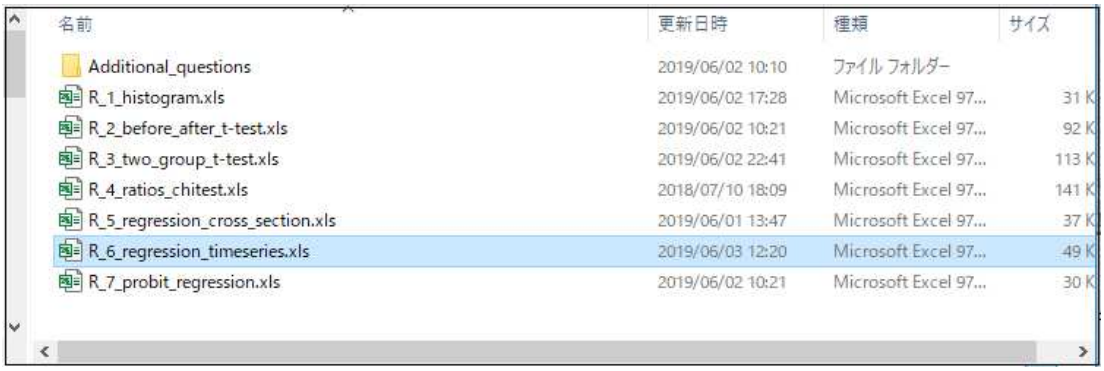
データの入力と読み込み

●以下のようにデータを入力する。タブに「FS」などの名前をつける。



	A	B	C	D	E
1	Year_month_week	No_of_week	Weather	FirstSaturday	Sales_volume
2	2001_January_1st	1	0	0	397
3	2001_January_2nd	2	0	0	405
4	2001_January_3rd	3	0	0	416
5	2001_January_4th	4	0	1	509
6	2001_January_5th	5	0	0	423
7	2001_February_1st	6	0	0	422
8	2001_February_2nd	7	0	0	430
9	2001_February_3rd	8	0	1	501
10	2001_February_4th	9	0	0	434
11	2001_March_1st	10	1	0	248
12	2001_March_2nd	11	0	0	447
13	2001_March_3rd	12	0	1	505
14	2001_March_4th	13	0	0	466
15	2001_April_1st	14	0	0	459
16	2001_April_2nd	15	0	0	473
17	2001_April_3rd	16	1	1	272
18	2001_April_4th	17	0	0	530
19	2001_May_1st	18	1	0	256
20	2001_May_2nd	19	0	0	434
21	2001_May_3rd	20	0	1	518
22	2001_May_4th	21	0	0	487
23	2001_May_5th	22	0	0	495
24	2001_June_1st	23	1	0	456
25	2001_June_2nd	24	0	1	609
26	2001_June_3rd	25	0	0	514
27	2001_June_4th	26	0	0	518

●デスクトップに、R_6_regression_timeseries.xls などと適当な名前をつけて保存する。

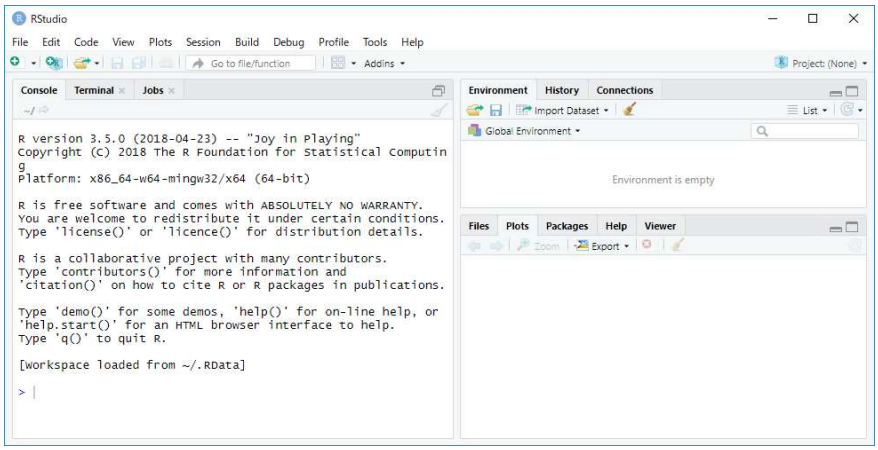


名前	更新日時	種類	サイズ
Additional_questions	2019/06/02 10:10	ファイル フォルダー	
R_1_histogram.xls	2019/06/02 17:28	Microsoft Excel 97...	31 K
R_2_before_after_t-test.xls	2019/06/02 10:21	Microsoft Excel 97...	92 K
R_3_two_group_t-test.xls	2019/06/02 22:41	Microsoft Excel 97...	113 K
R_4_ratios_chitest.xls	2018/07/10 18:09	Microsoft Excel 97...	141 K
R_5_regression_cross_section.xls	2019/06/01 13:47	Microsoft Excel 97...	37 K
R_6_regression_timeseries.xls	2019/06/03 12:20	Microsoft Excel 97...	49 K
R_7_probit_regression.xls	2019/06/02 10:21	Microsoft Excel 97...	30 K

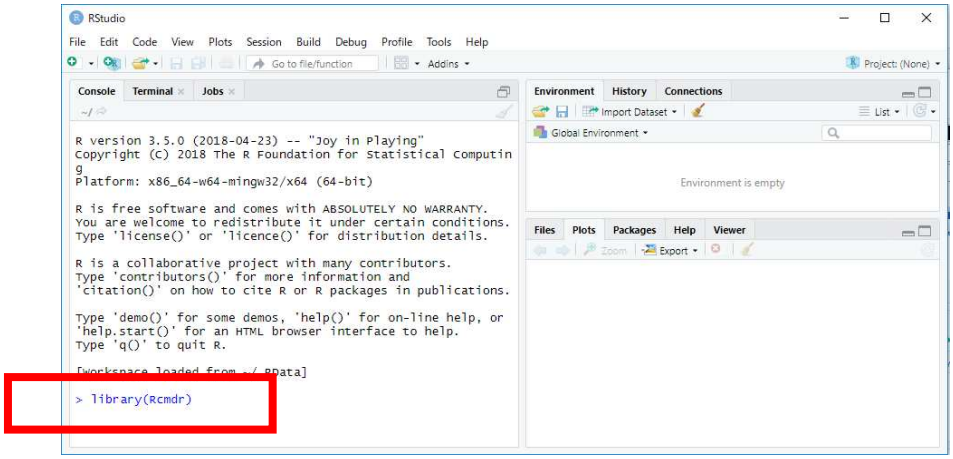
（すでに RStudio と R コマンダーを立ち上げている場合はこのページを省略してください。）

● Windows のメニュー＞“RStudio”を選ぶ。

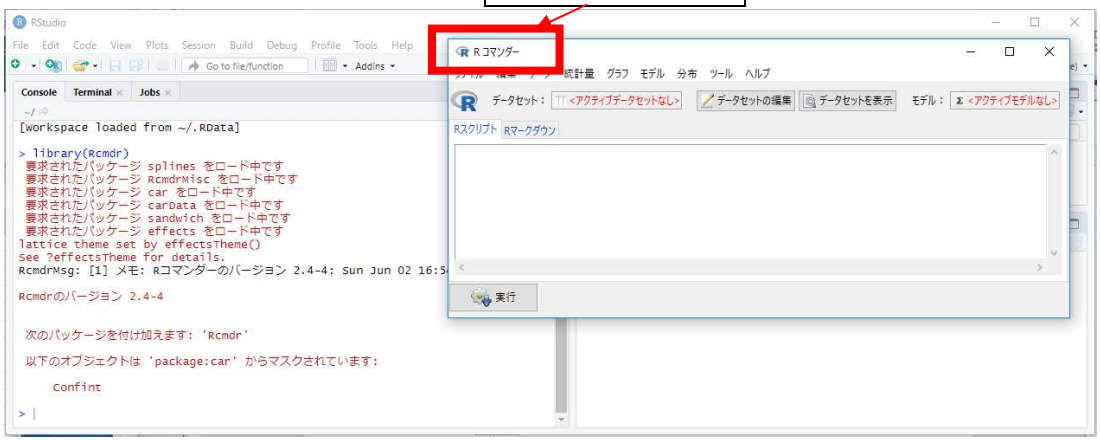
● 下のようなウィンドウが開く。



● ” library(Rcmdr) “と入力する。＞Enter を押す。

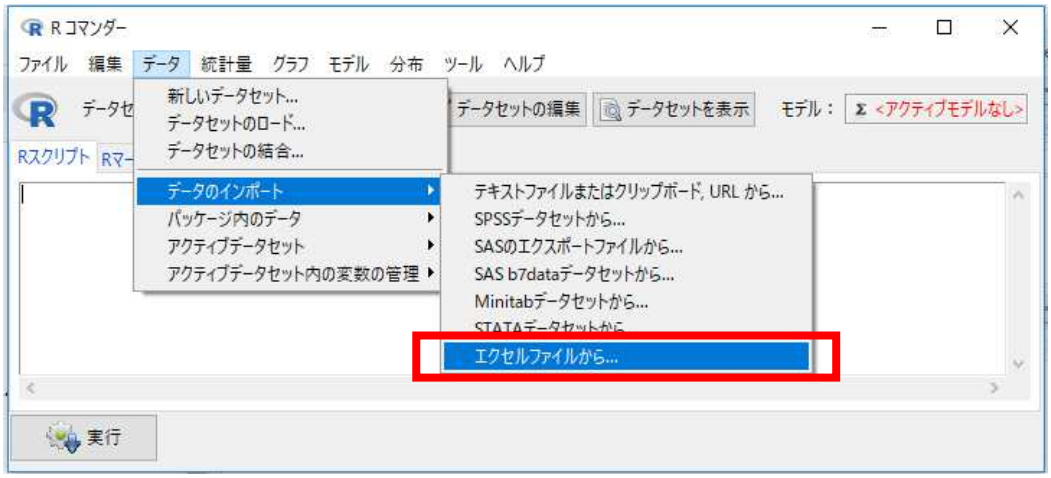


●下のようなメッセージが流れて、 R コマンダー のウィンドウが開く。

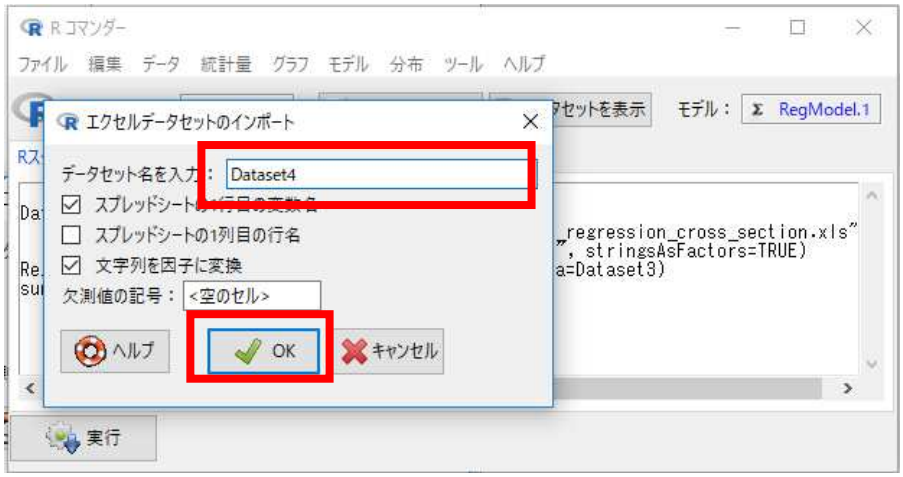


（前ページを飛ばした方はここから再開してください。）

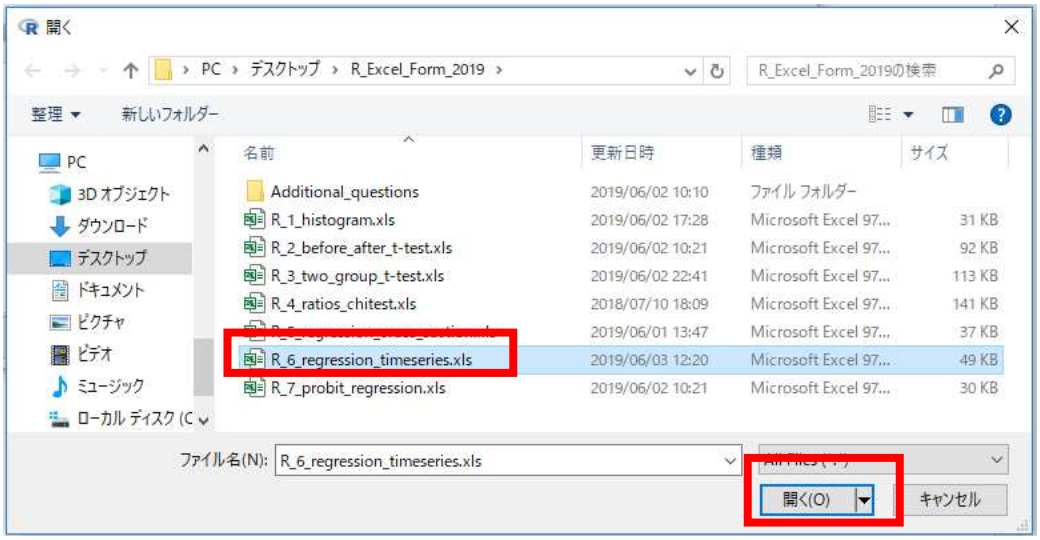
● R コマンダー のウィンドウで、データ ＞ データのインポート ＞ エクセルファイルから を選ぶ。



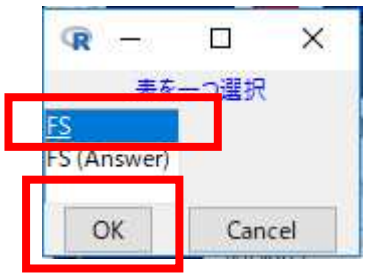
●現れたウィンウィンドウで、“データセットを入力” に “Dataset4” と入力する。＞ OK を押す。



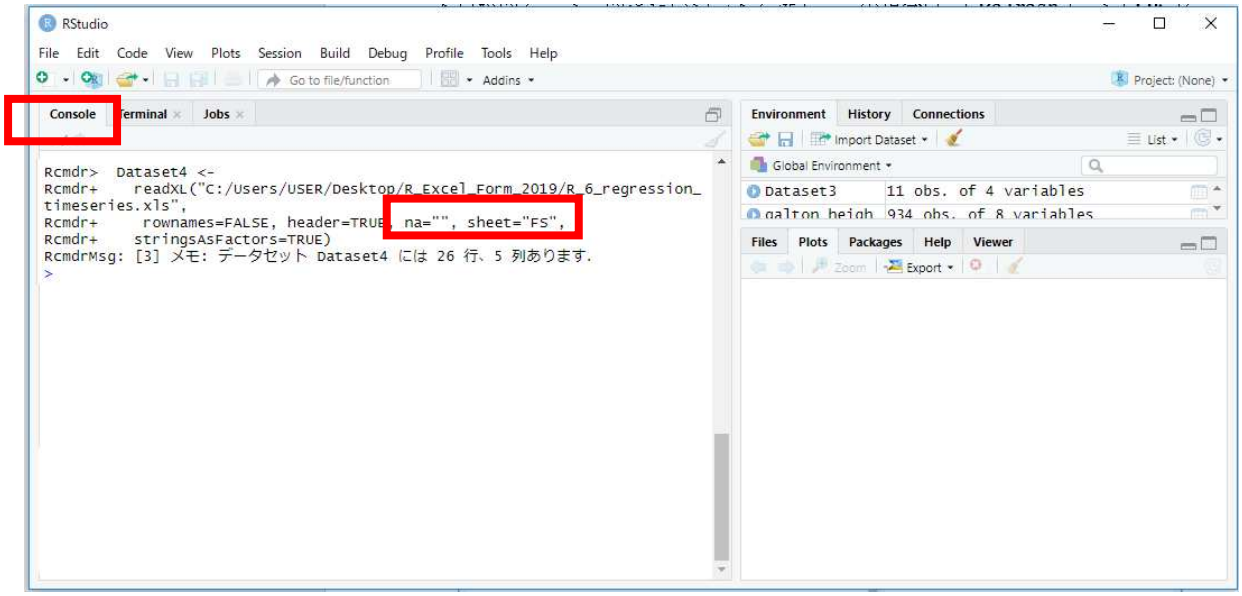
●さきほど保存した Excel ファイルを選択する。＞ 開く を押す。



●もし Excel ファイルに２枚のシートが格納されていると下のような小さなウィンドウが開く。＞　開きたいシートを選ぶ。この場合は、FS 。＞ OK を押す。

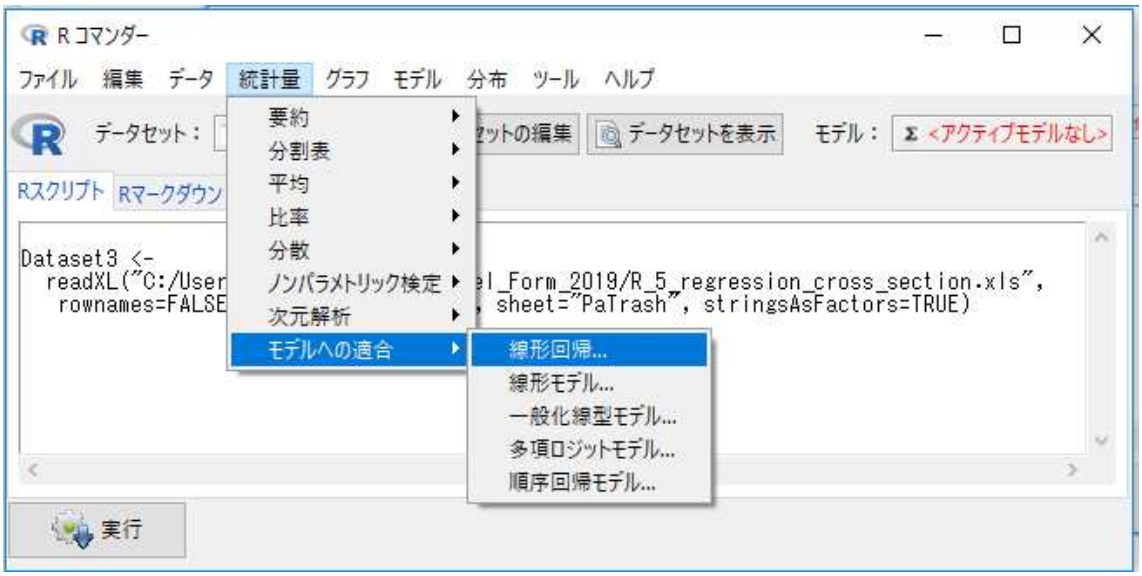


●すると RStudio の Console で、次のような表示が現れる。Excel ファイルの当該シート（ FS ）の読み込みが成功した。



「線形回帰」の選択

● R コマンダー のメニューから 統計量 ＞ モデルへの適合 ＞ 線形回帰 を選ぶ。

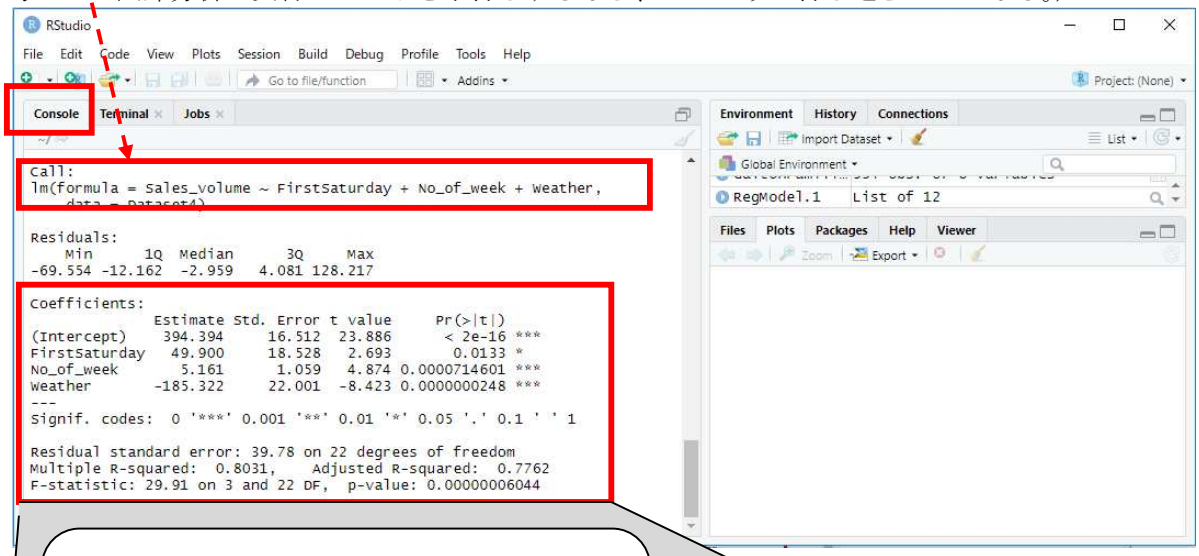


●現れたウィンドウで、「目的変数（1 つ選択）」で Sales_Volume を選ぶ。
「説明変数（1 つ以上選択）」で、Ctrl を押しながら
FirstSaturday と No_of_Week と Weather の 3 つを選ぶ。
OK を押す。

分析結果を得る

● RStudio の Console に以下にように出力される。

（参考まで：回帰分析の実行コマンドを手打ちするなら、このように打ち込むことになる。）



③ 係数（傾き）：それぞれの変数が成果指標の変化をどれくらい説明するかを表す。FirstSaturday が 0->1 になると、Sales_Volume が 49.90 ドル増加する。雨が降ると-185.322 等。

④ 各変数の t 値と p 値：各変数の統計学的な有意度を表す。もし p 値 <0.05 なら、統計学的に有意である。この場合は、三つの変数とも統計学的に有意と判断される。

```
Coefficients:
              Estimate Std. Error t value     Pr(>|t|)
(Intercept)    394.394     16.512  23.886      < 2e-16 ***
FirstSaturday   49.900     18.528   2.693       0.0133 *
No_of_week       5.161      1.059   4.874 0.0000714601 ***
Weather       -185.322     22.001  -8.423 0.0000000248 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 39.78 on 22 degrees of freedom
Multiple R-squared:  0.8031,    Adjusted R-squared:  0.7762
F-statistic: 29.91 on 3 and 22 DF,  p-value: 0.00000006044
```

① R-Square（R^2）：得られた等式でどの程度の変化量が説明されるか。この場合は、80.3%が等式によって説明される。

② F：等式全体の統計学的な有意度を表す。もし p-値（= Significance F）<0.05 なら、等式は全体として有意である。このケースでは、p-value は 6.0E-08（つまり 0.00000006）であり、0.05 以下である。

Sales_volume＝394.4 + 5.16 × (No_of_week) − 185.32 × (Weather) + 49.90 × (“FirstSaturday)

分析結果

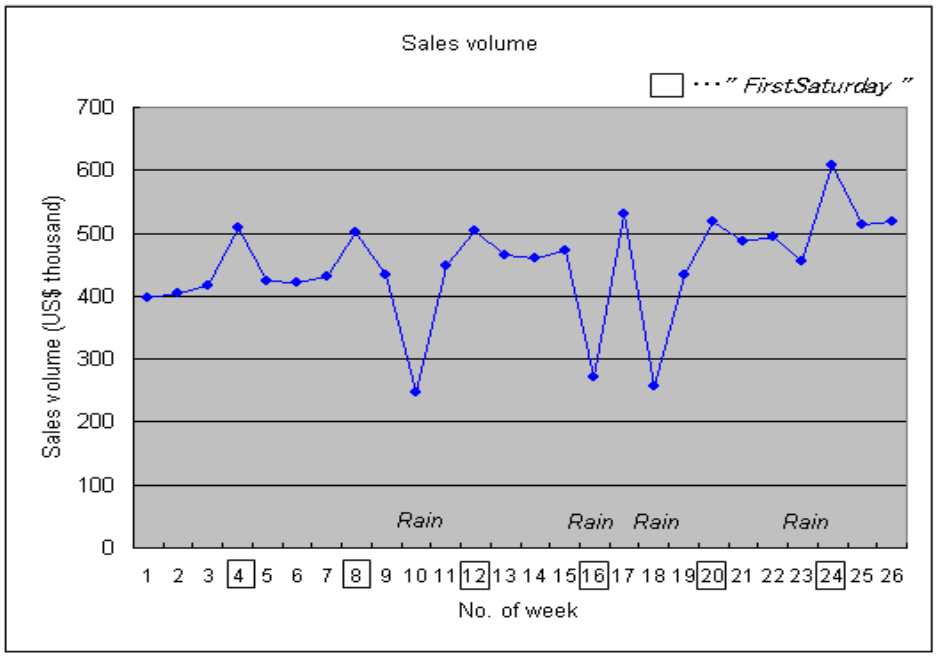
このケースでは、新しいセールスイベント(「どいち」（“FirstSaturdy”））が売上高を増加させていると結論された。「どいち」（“FirstSaturday”）の実施により、売上高は，US$49.9 thousands 増加すると推定された。

また、天候が売上高に影響していることも結論された。さらに、わずかな上昇傾向が観察された。確かな理由はわからないが、全国的な経済状況の改善が影響しているのかも知れない。

Sales_volume = 394.4+5.16 × (No_of_week) －185.32 × (Weather) + 49.90 × (“FirstSaturday)

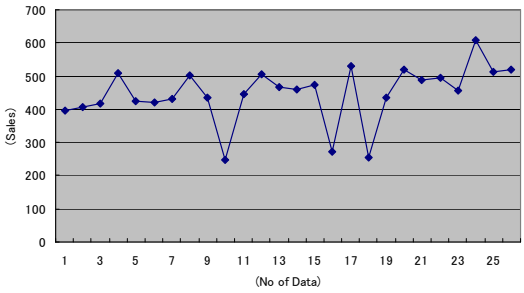
（Unit: US$ thousands）

以下のようなグラフを添付することが勧められる。(ただし、“FirstSaturday”（□）と雨（rain）は手書きしている。).

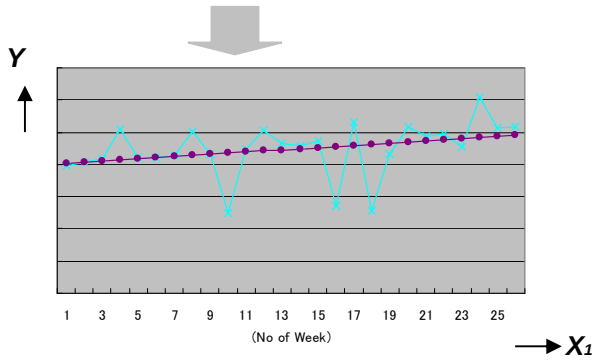


・・・売上高は新しいセールスイベント「どいち」（ “FirstSaturday”）によって増加していると推定された。

発展学習 1：回帰分析を理解するための視覚的説明



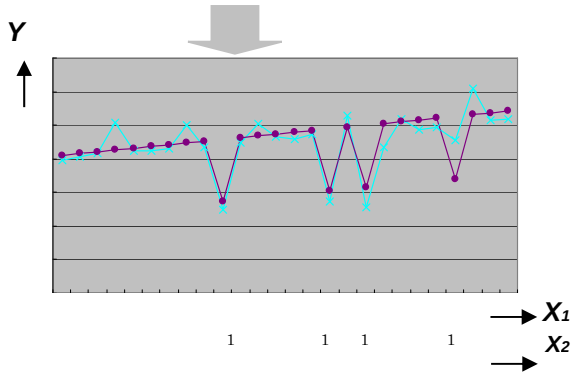
これはもとの散布図と折れ線グラフである。



***Y* は *X₁* によって説明される**

ただ *X₁* (週数(No.of Week)) を使う。

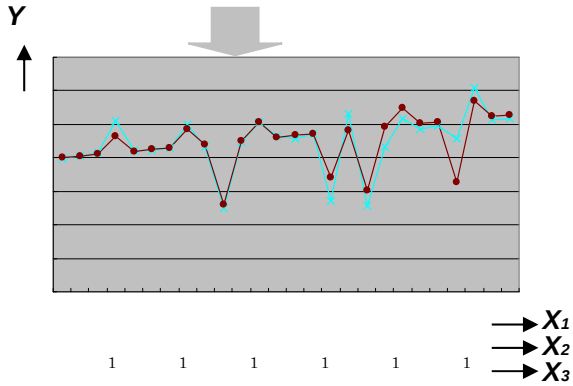
週数は、全国的な景気回復の傾向を反映している。



***Y* は *X₁* と *X₂* によって説明される。**

X₂ (天候（Weather）) を加える。

X₂：0 (晴れ/曇り)
1 (雨) =>線はところどころで下落する。.



***Y* は *X₁* と *X₂* と *X₃* によって説明される。**

X₃ (「どいち」（FirstSaturday）) を加える。

X₃ :0 (FirstSaturday is not)
1 (FirstSaturday is held)=>
線はところどころで上昇する。

……もし私が４時限の空間を使えるなら、グラフを４次元で正確に示すことができるだろう。なぜなら３軸を利用できるからだ(*Y, X₁, X₂ & X₃*).。しかし、紙は２次元なので、上記のように表した。

第 3 セッション:

プロビット回帰分析:

「はい」「いいえ」の回帰分析

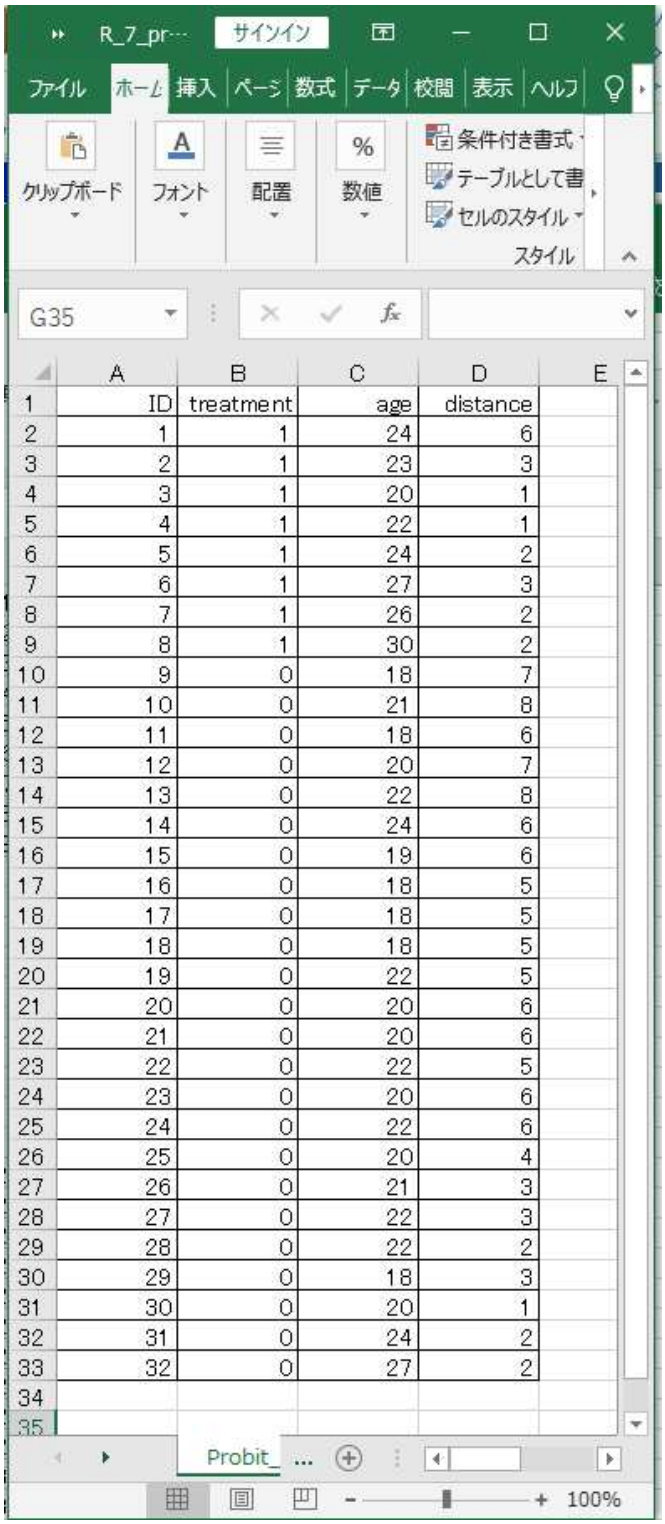
Q. マイクロファイナンスからの借り入れの有無は、①農村女性の住居からマイクロファンナンス事業者の支店までの距離（km）と、②年齢（歳）に大きく影響されているという主張がある。

（1） マイクロファイナンスの借り入れの有無に、①マイクロファイナンス事業者の支店までの距離（km）と②年齢（歳）は影響しただろうか。

（2） 年齢 30 歳、マイクロファイナンス業者の支店からの住居までの距離 4km の人がマイクロファイナンスから借り入れる確率はどの程度だろうか？

データの用意

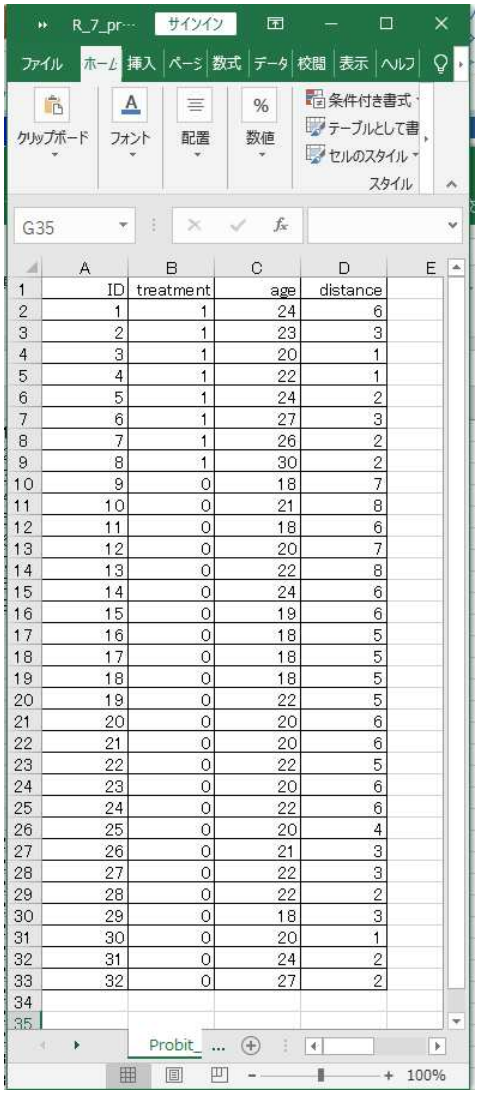
ID……個人番号
treatment……研修参加／不参加（参加=1，不参加=0）
age……年齢
distance……支店までの距離



	A	B	C	D
1	ID	treatment	age	distance
2	1	1	24	6
3	2	1	23	3
4	3	1	20	1
5	4	1	22	1
6	5	1	24	2
7	6	1	27	3
8	7	1	26	2
9	8	1	30	2
10	9	0	18	7
11	10	0	21	8
12	11	0	18	6
13	12	0	20	7
14	13	0	22	8
15	14	0	24	6
16	15	0	19	6
17	16	0	18	5
18	17	0	18	5
19	18	0	18	5
20	19	0	22	5
21	20	0	20	6
22	21	0	20	6
23	22	0	22	5
24	23	0	20	6
25	24	0	22	6
26	25	0	20	4
27	26	0	21	3
28	27	0	22	3
29	28	0	22	2
30	29	0	18	3
31	30	0	20	1
32	31	0	24	2
33	32	0	27	2

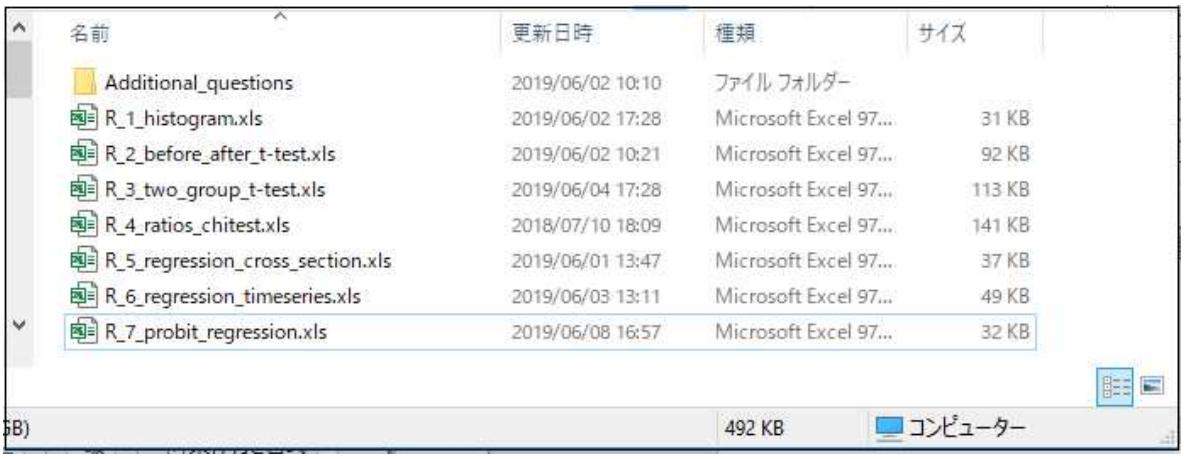
データの入力と読み込み

以下のようにデータを入力する。タブに「Probit」などの名前をつける。



ID	treatment	age	distance
1	1	24	6
2	1	23	3
3	1	20	1
4	1	22	1
5	1	24	2
6	1	27	3
7	1	26	2
8	1	30	2
9	0	18	7
10	0	21	8
11	0	18	6
12	0	20	7
13	0	22	8
14	0	24	6
15	0	19	6
16	0	18	5
17	0	18	5
18	0	18	5
19	0	22	5
20	0	20	6
21	0	20	6
22	0	22	5
23	0	20	6
24	0	22	6
25	0	20	4
26	0	21	3
27	0	22	3
28	0	22	2
29	0	18	3
30	0	20	1
31	0	24	2
32	0	27	2

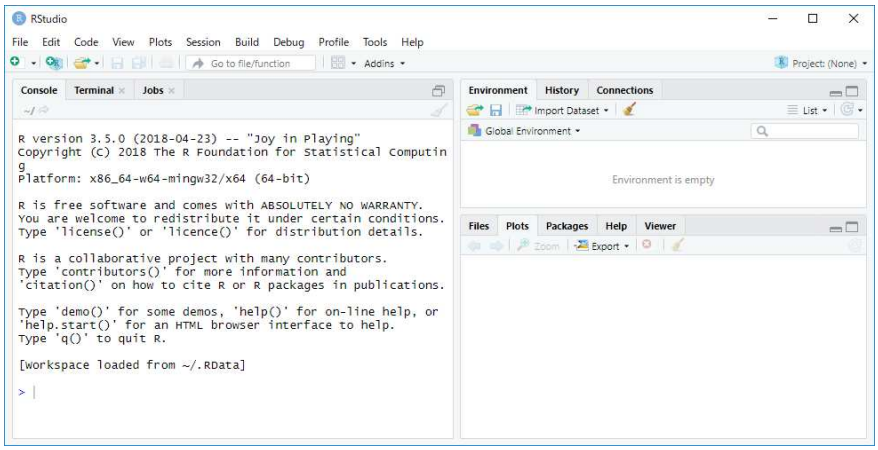
●デスクトップに、R_7_probit_regression.xls などと適当な名前をつけて保存する。



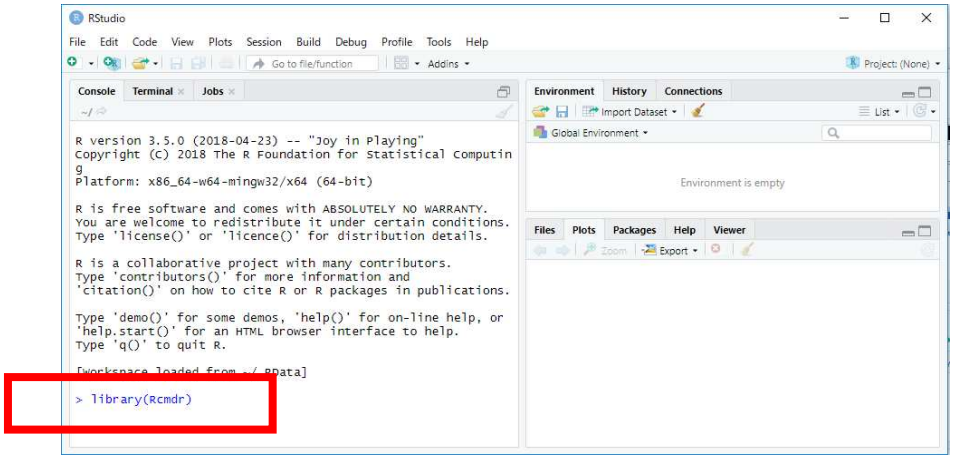
（すでに RStudio と R コマンダーを立ち上げている場合はこのページを省略してください。）

● Windows のメニュー＞“RStudio”を選ぶ。

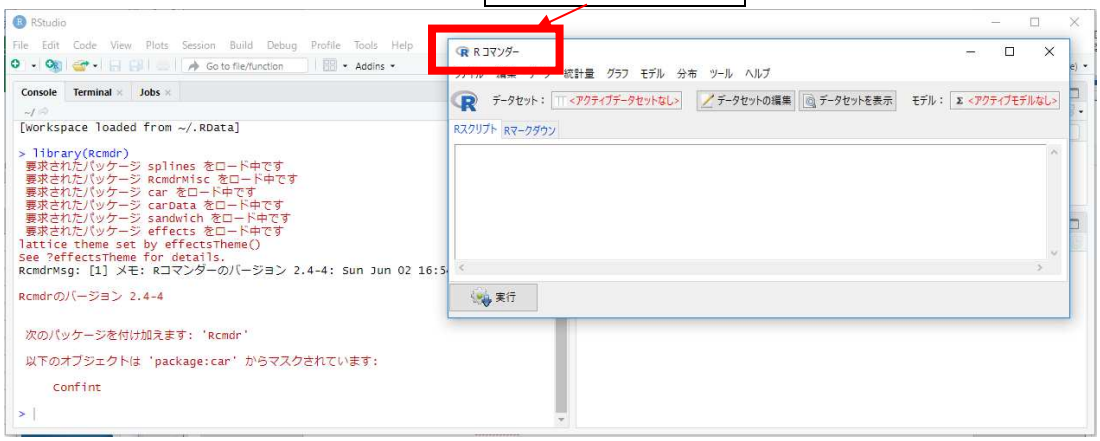
● 下のようなウィンドウが開く。



● ” library(Rcmdr) “と入力する。＞Enter を押す。

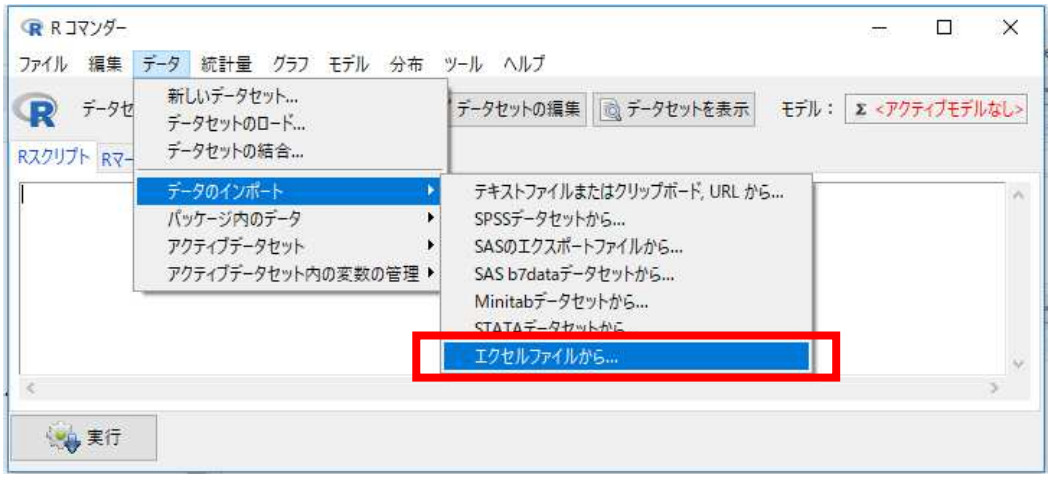


●下のようなメッセージが流れて、 R コマンダー のウィンドウが開く。

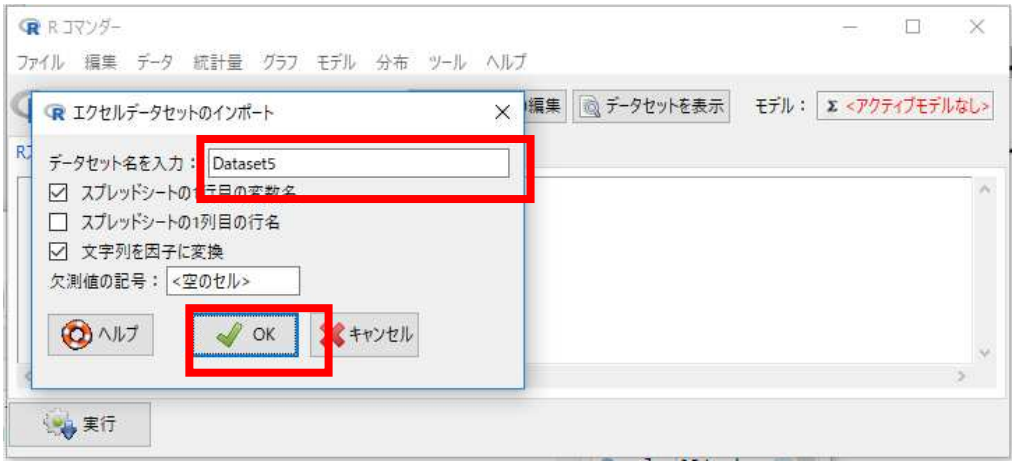


（前ページを飛ばした方はここから再開してください。）

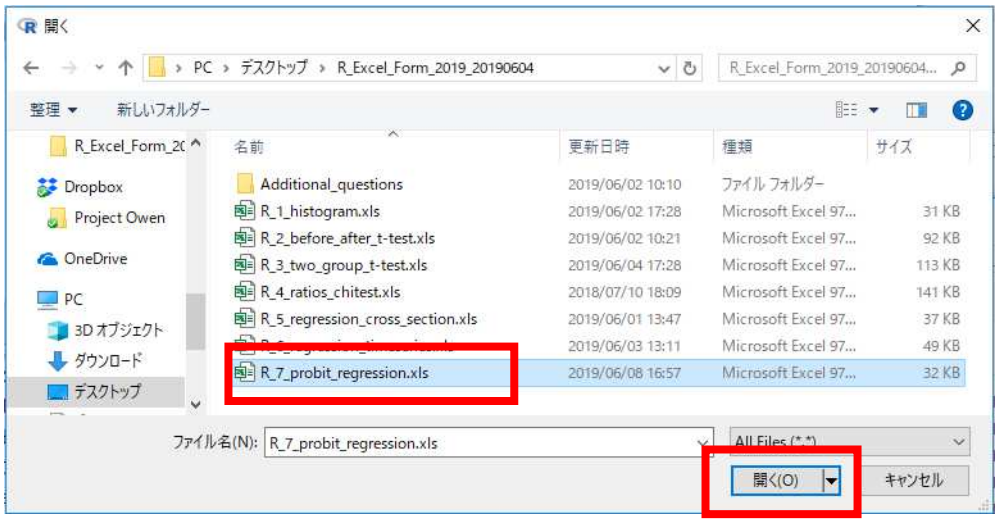
● R コマンダー のウィンドウで、データ ＞ データのインポート ＞ エクセルファイルから を選ぶ。



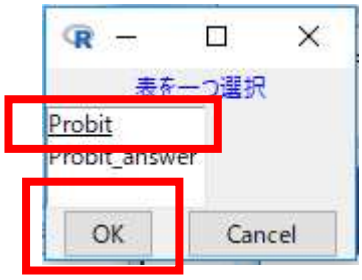
●現れたウィンウィンドウで、“データセットを入力”に“Dataset5”と入力する。＞ OK を押す。



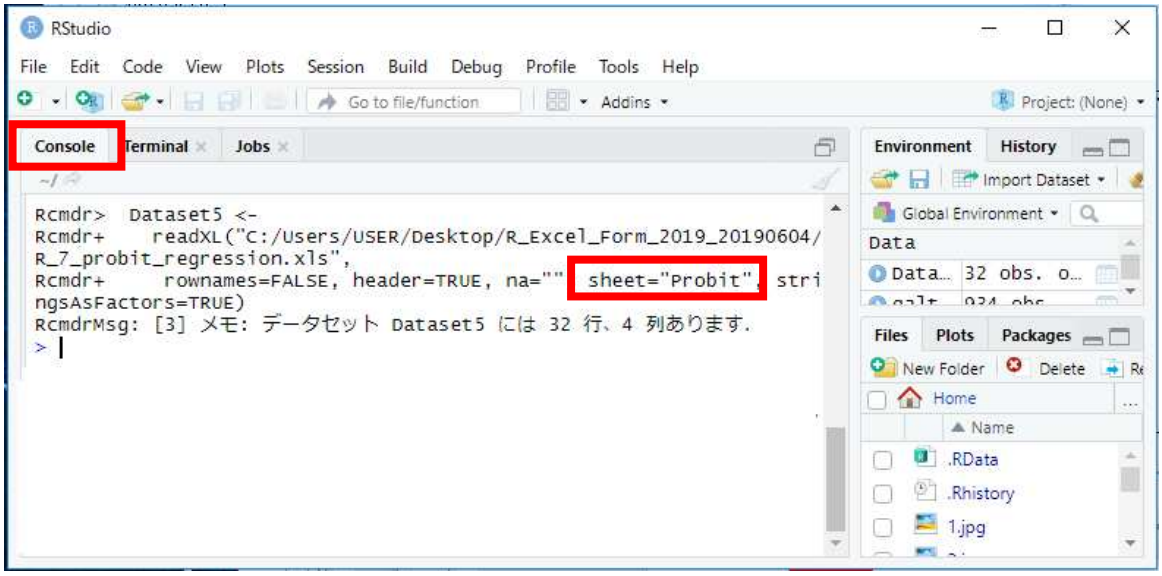
●さきほど保存した Excel ファイルを選択する。＞ 開く を押す。



●もし Excel ファイルに２枚のシートが格納されていると下のような小さなウィンドウが開く。＞　開きたいシートを選ぶ。この場合は、Probit 。＞ OK を押す。

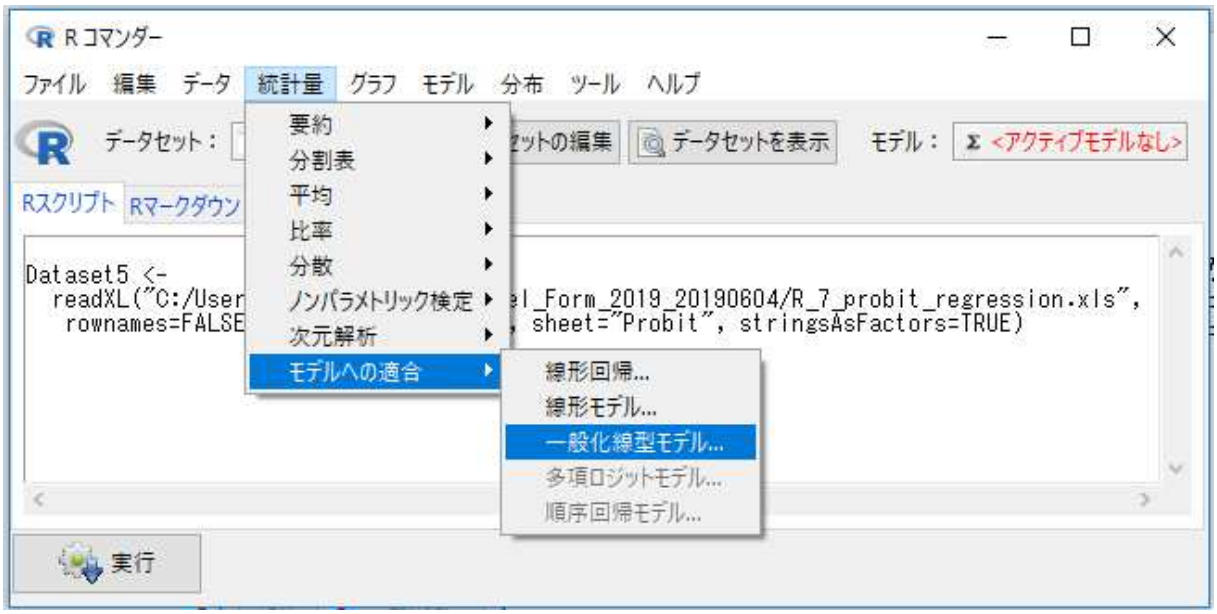


●すると RStudio の Console で、次のような表示が現れる。Excel ファイルの当該シート（ Probit ）の読み込みが成功した。



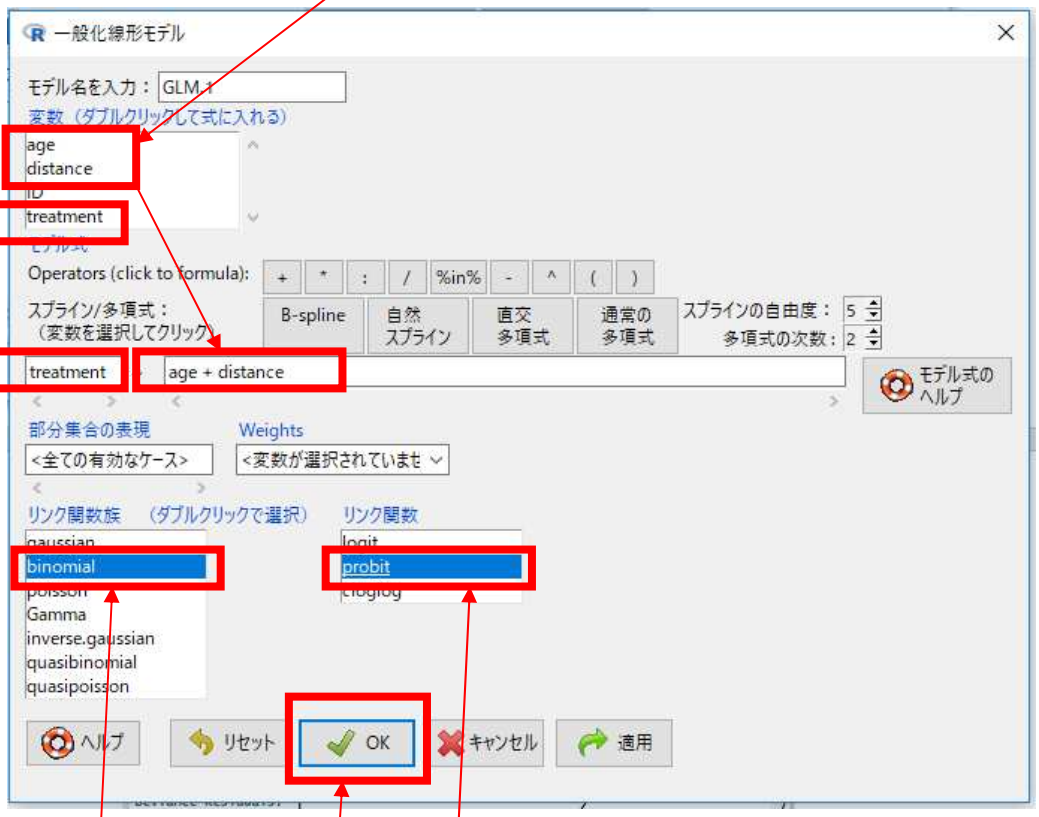
「線形回帰」の選択

● R コマンダー のメニューから 統計量 ＞ モデルへの適合 ＞ 一般化線型モデル を選ぶ。（英語では、General Linear Modeling (GLM)と呼ぶ。）



●開いたウィンドウで以下のように①～⑤を選択する。

②age をダブルクリックして式に入れる。続いて、distance をダブルクリックして、式に入れる。



① treatment をダブルクリックして、式に入れる。

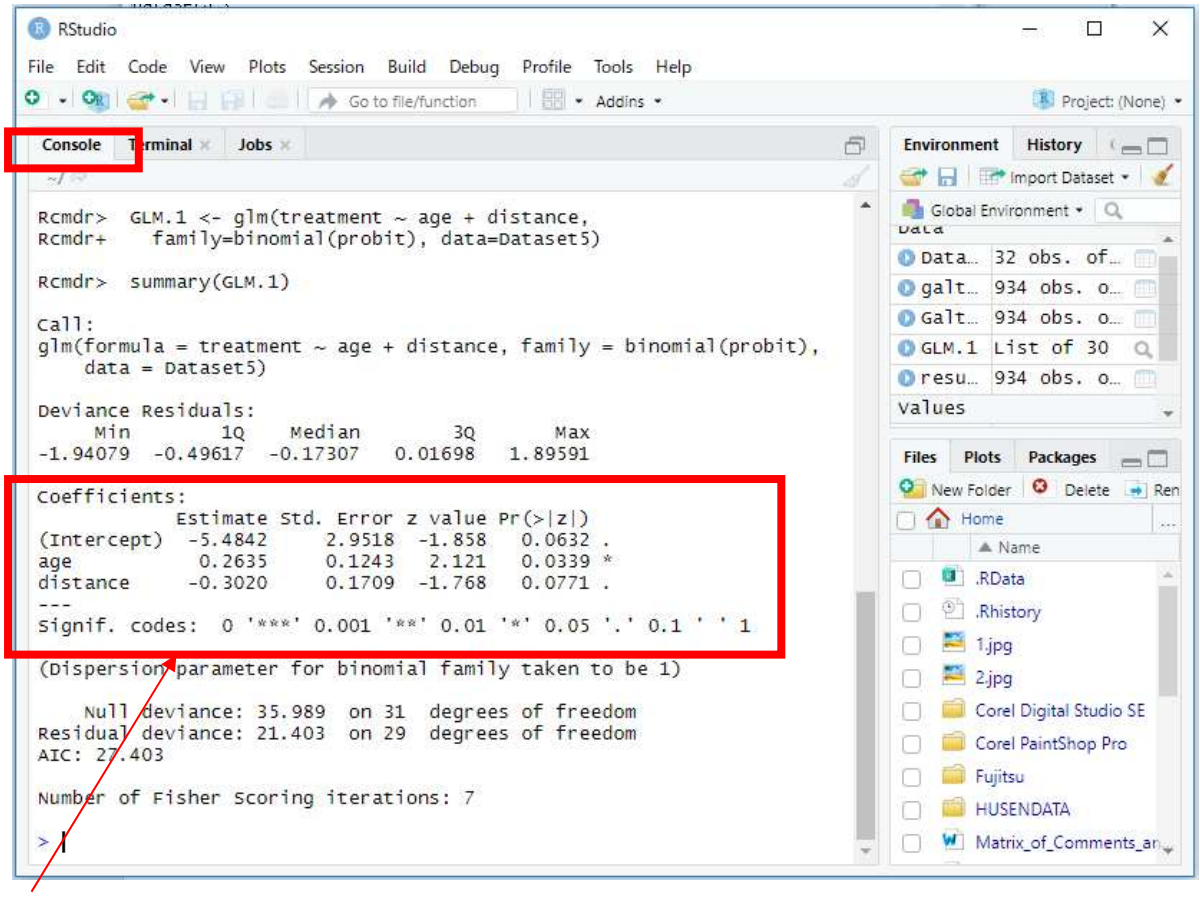
③「リンク関数族」で binomial を選ぶ。

④「リンク関数」で probit を選ぶ。

⑤ OK を押す。

分析結果を得る

● RStudio の Console に以下にように出力される。

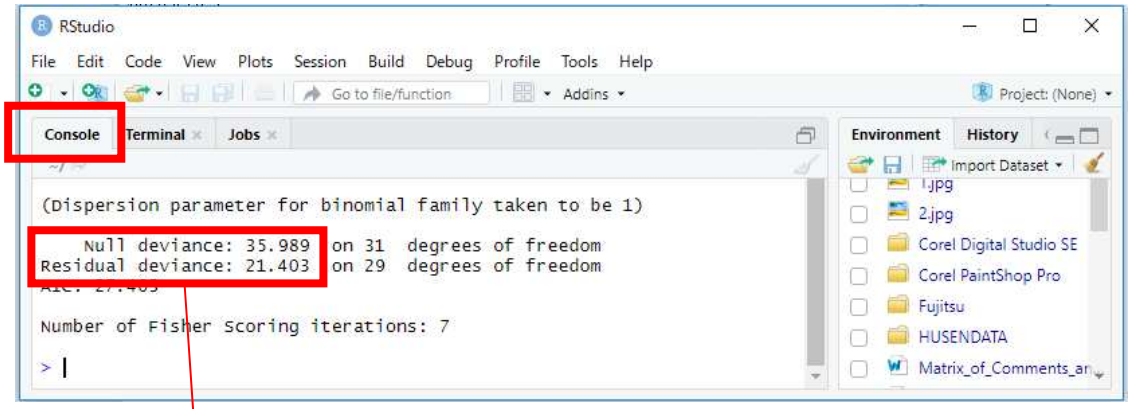


● 上記が、R が算出したプロビット回帰分析の結果である。全ての変数の係数は統計学的に有意である（p < 10%）。以下の式が得られた。

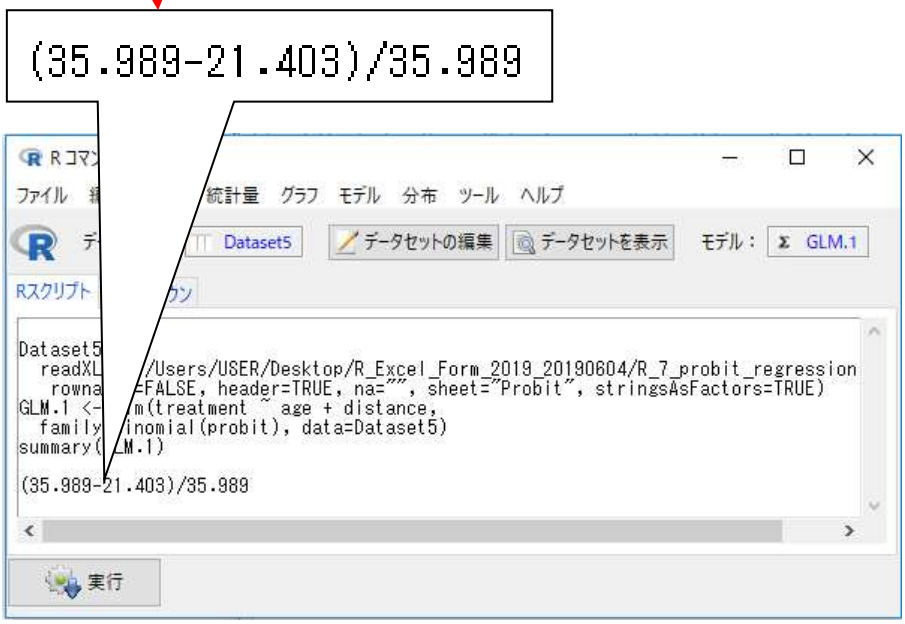
treatment = - 5.4842+ 0.2635 * age - 0.3020 * distance

● R^2 に相当する値をもとめる（Psuedo R^2 と言う）。

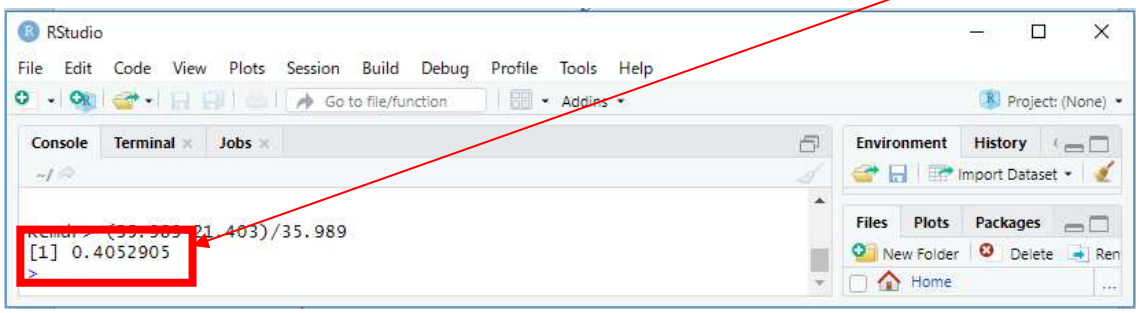
$$R^2 = \frac{\text{"Explained"}}{\text{"Explained" + "Unexplained"}} = \frac{\text{(Null deviance) – (Residual deviance)}}{\text{(Null deviance)}}$$



● したがって、　R コマンダーで次のように打ち込む。Enter を押す。



● すると RStudio の Console に次の値を得る。**Psuedo R^2= 0.405 (40.5%)**



z 値= -5.4842+ 0.2635 * age - 0.3020 * distance　　Psuedo R2=0.4053

BOX　プロビット回帰分析の説明

単純な回帰直線の式は、x → y を計算する（例 : $y=b_0+b_1x$）。プロビット分析では、(i)x → z を計算して(ii) z → y を計算する。つまり 2 段階の計算が必要になる。

今回の例では、まず、上記の回帰分析の結果から、(i) 二つの x 値（age と distance）から z 値を計算するための式が得られる（下の式）。

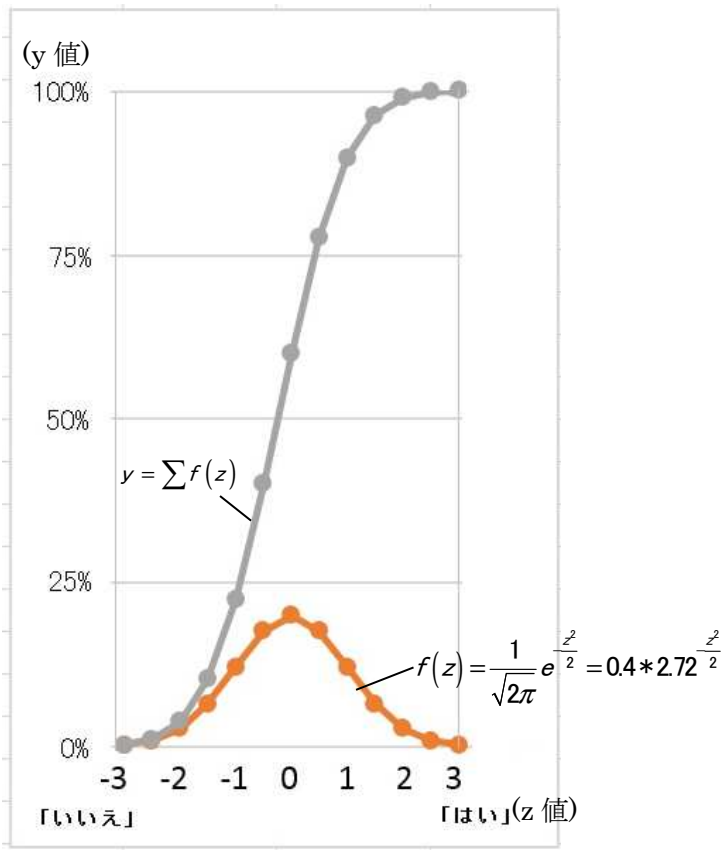
z 値= -5.4842 + 0.2635* age - 0.3020 * distance

x_1

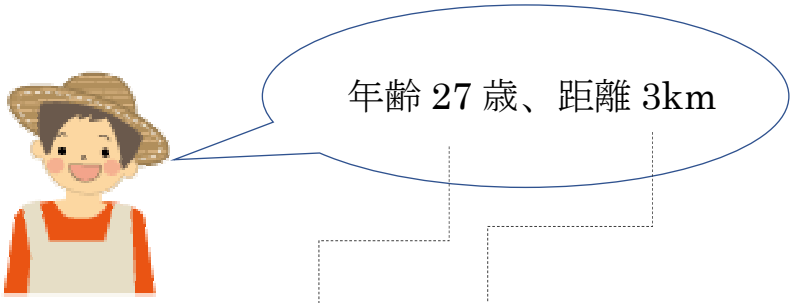
その式から得られた z 値を y 値を得るための式に代入すると y 値が得られる（R では下の式）。ちなみに y 値を「確率の累積値」と言い、0 から 1（0%から 100%）の間の値を取る。

y 値= pnorm (z 値, 0, 1)

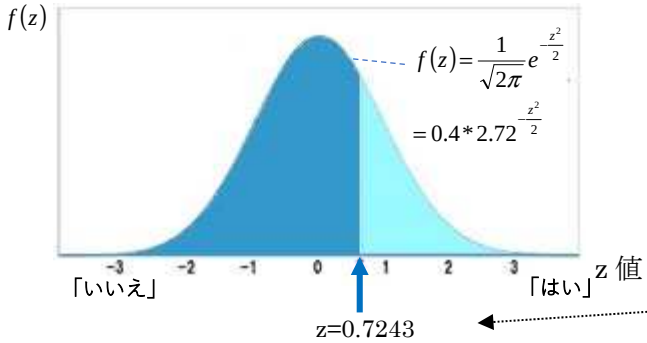
この手続きにより、ひとりひとりの素性から傾向スコアを計算できる。傾向スコアとは、**どれくらい「はい」なのかをパーセント表示した値** である。



（例 1）今回のデータの ID 6 の人の場合（ age = 27 歳、 distance = 3km ）

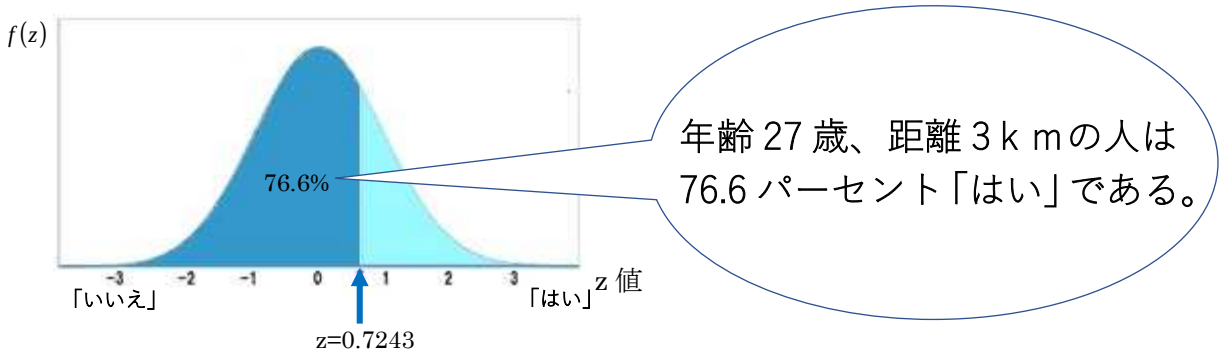


z 値 = -5.4842 + 0.2635 * 27 - 0.3020* 3 =0.7243



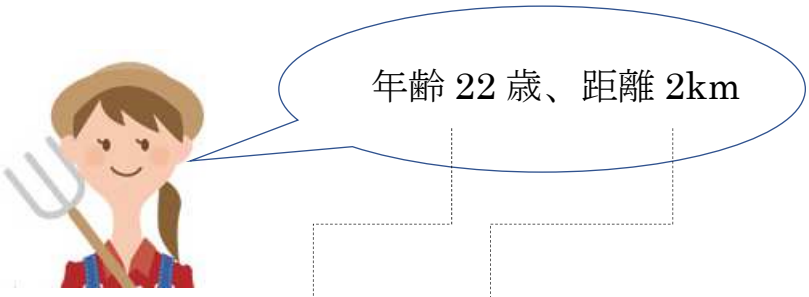
次に左側の青色の面積を求める。その面積を y 値（確率の累積値）と名付ける。

y 値（確率の累積値）= normdist (0.7243,0,1,1) = 0.7656 となる。

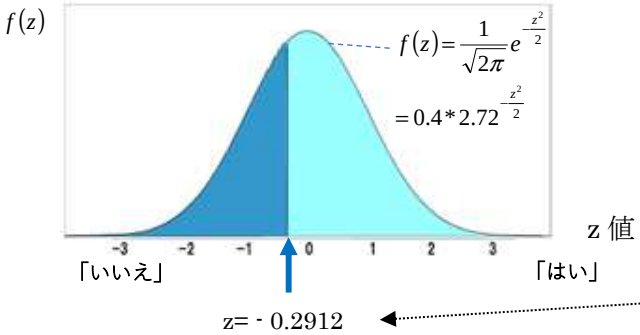


この左側の青色の面積は 76.6%である。つまり、彼女は 76.6%「はい」であり、彼女の傾向スコアは 76.6 点ということになる。

（例 2）今回のデータの ID28 の人の場合（age = 22 歳、 distance = 2km）

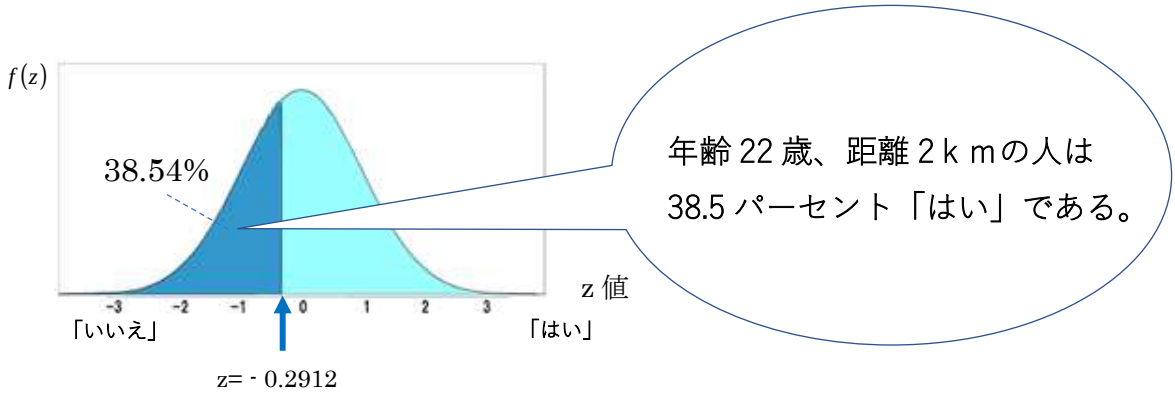


z 値 = -5.4842 + 0.2635 * 22 - 0.3020* 2 = - 0.2912



次に左側の青色の面積を求める。その面積を y 値（確率の累積値）と名付ける。

y 値（確率の累積値）= normdist (- 0.2912,0,1,1) = 0.3854 となる。



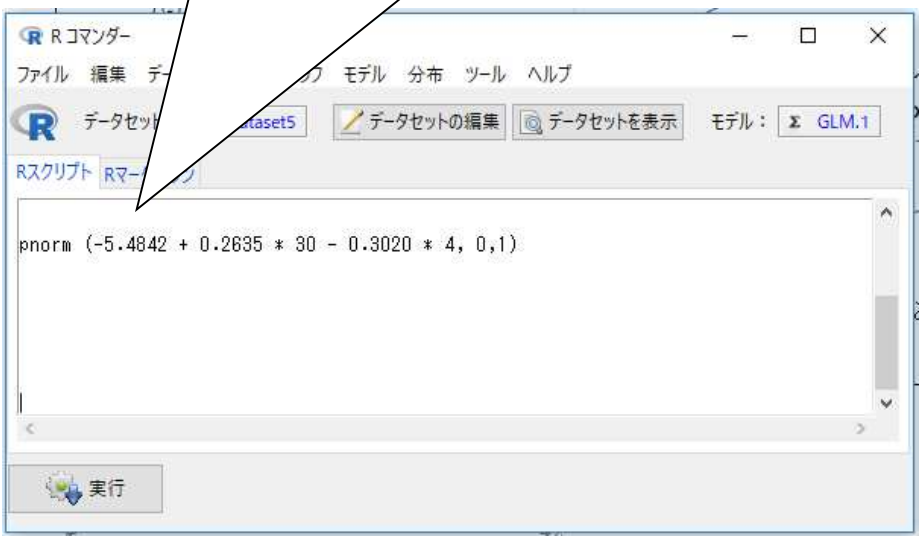
この左側の青色の面積は 38.5%である。つまり、彼女は 38.5%「はい」であり、彼女の傾向スコアは 38.5 点ということになる。

年齢 30 歳、距離 4km の人が借り入れる確率を計算する。

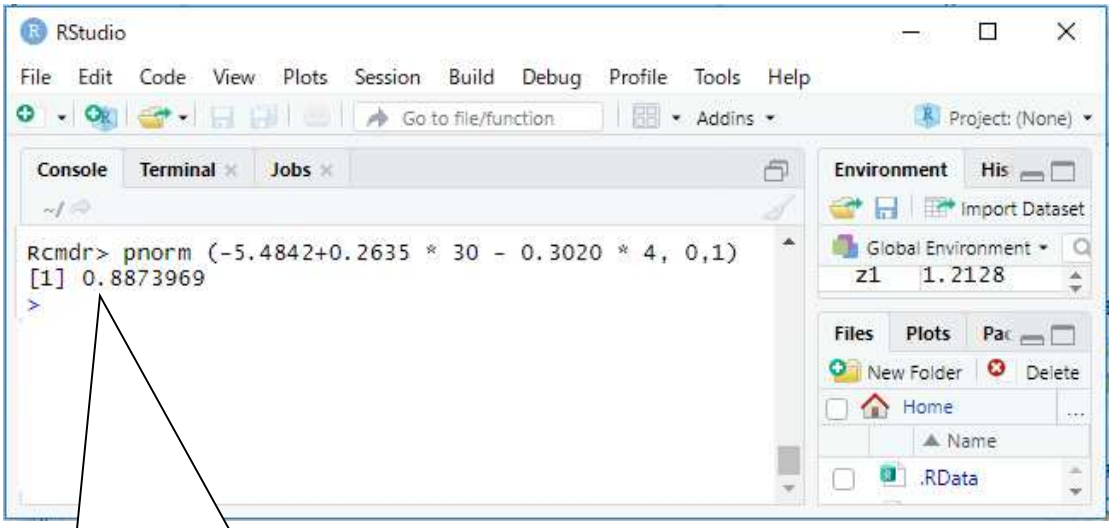
● R コマンダーで、pnorm（z の値、平均値、標準偏差）と打ち込む。

＜年齢 30 歳、距離 4km。この場合、平均値 0、標準偏差 1 になる。＞

```
pnorm (-5.4842 + 0.2635 * 30 - 0.3020 * 4, 0,1)
```



● RStudio の Console に以下にように出力される。



確率は 0.887 (88.7%)と計算された。

結論

以上の分析から以下の結論を得た。

（1）マイクロファイナンスの借り入れの有無に、①マイクロファイナンス事業者の支店までの距離（km）と②年齢（歳）は影響する。

（2）年齢 30 歳、マイクロファイナンス業者の支店からの住居までの距離 4km の人がマイクロファイナンスから借り入れる確率は 88.7%であった。

参考文献

佐々木亮(2019)『プロフェッショナル統計分析ワークショップ』の配布資料
（非売品・同ワークショップの参加者に配布しています）

西内啓（2013)『統計学が最強の学問である：データ社会を行きぬくための武器と教養』
（ダイヤモンド社）

あとがき

『統計学が最強の学問である：データ社会を行きぬくための武器と教養』（西内啓、2013）が 40 万部近い大ヒットを記録したことは、私にとって衝撃的でした。統計学を扱った本でそれほどの発行部数を記録した書籍を知らなかったからです。統計学の巨人である R.A.フィッシャーの名著である Statistical Methods for Research Workers (Oliver & Boyd, 1925) (http://psychclassics.yorku.ca/Fisher/Methods/) はつい最近まで重版されましたが、それ以来の大ヒットと言っても過言ではないでしょう。

その本には、「すべての学問は統計学のもとに」、「IT と統計学の素晴らしき結婚」、「ランダム化という最強の武器」などの刺激的な章や節のタイトルが並んでいます。また、その本で、「ランダム化比較実験（RCT）が最強の調査デザインである」という主旨の記載に触れたときには「よく言ってくれた！」と感激を憶えたものです。なお、ランダム化比較実験（RCT）をはじめとする調査デザインの概要と実例の解説は、私の著書をご覧ください。無料でダウンロードできます。
(https://www.idcj.jp/9evaluation/sub5_files/impact_eval_jirei_28july2011.pdf)

さて、それで実際に統計データ分析に取り組もうと思われた方が、はたと戸惑うのは、実際にどのソフトを使ってどのように操作すればいいのだろうかということです。できることなら、実際にそのソフトが行っている計算も、自分で手計算ができるレベルの簡単な解説で知りたいと思われるでしょう。そこで統計学を専門としている知人に尋ねると、SAS、SPSS、STATA などのソフトを紹介されるでしょう。そこでその価格をウェブで調べると 15 万円から 30 万円くらいすることを知り、目玉が飛び出る思いをするでしょう。「とてもそんなお金は用意できない。統計データ分析をやらなくても済むレポートの書き方を考えよう。何人かにインタビューするだけでもなんとかなるのではないか？そうだそうしよう。」という思考を経ることでしょう。これはたいへん残念なことでした。

そんな場面で、フリー（無料）の統計分析ソフトの R の出番です。R は以前は、コマンドを打ち込むソフトでした。真っ黒な画面にひたすら呪文を打ち込むソフトだったわけです。思い起こせば、その昔、MS-DOS というソフトがあって、真っ黒な画面にひたすら呪文を打ち込むことでパソコンを動かしていました。その後、1992 年に Windows 3.1 が発売されて、グラフィカルな分かりやすいインターフェースが提供されることになり、爆発的にパソコンが普及しました。今、R に関しても同じことが起こっています。R Commander により、グラフィカルな分かりやすいインターフェースが提供されたのです。これにより、R が爆発的に普及していくことでしょう。

皆様にとって、パソコンはもはや意識しなくてもなくてはならないものになっているでしょう。同様に、皆様が意識しなくても R がなくてはならないものになる日が来ると予想しており、その第一歩に、このマニュアルが貢献することを期待しています。

最後に、本を出版するときはいつもそうですが、執筆するときにリスニングしていた音楽を記載したいと思います。これらの音楽がなければ、本書の命とも言える直観的に理解できる図表ができることはなかったでしょう。

Avicii. *Without you.*
Avicii & Nicky Romero. *I Could Be the One (Live Video).*
Frankie Knuckles & Eric Kupper. *Baby I (Director's Cut mix).*
Satoshi Tomiie feat. Arnold Jarvis. *And I Loved you.*

佐々木亮／Ryo SASAKI, Ph.D.