The strategy to analyze 2 variables

Minato Nakazawa

18 Oct. 2010

1 Introduction

Here I give a brief summary to analyze 2 variables in R.

Variables	Summarizing method	Relationship to know	Methods of analysis	R functions
B1, B2	2 by 2 contingency table	Independence	Fisher's exact test	fisher.test(table(B1,B2))
(same)	(same)	(same)	Chi-square test	chisq.test(table(B1,B2))
(same)	(same)	Strength of relationship	Odds ratio	library(vcd) (OR <- oddsratio(table(B1,B2),log=FALSE) confint(OR)
(same)	(same)	(same)	Risk ratio	library(vcd) (RR <- riskratio(table(B1,B2),log=FALSE) confint(RR)
(same)	(same)	(same)	phi coefficient, contingency coefficient, and Cramer's V	library(vcd); assocstats(table(B1,B2))
(same)	(same)	Agreement of two measurement	Kappa coefficients	library(vcd); Kappa(table(B1,B2)) or library(fmsb); Kappa.test(table(B1,B2))
B1,C1	2 by n contingency table	Independence	Fisher's exact test	fisher.test(table(B1,C1))
(same)	(same)	(same)	Chi-square test	chisq.test(table(B1,C1))
(same)	(same)	Equal proportion among several groups or not	Test of equal proportions	prop.test(table(C1,B1)) or TAB <- table(C1,B1); prop.test(TAB[,1],rowSums(TAB))
C1, C2	m by n contingency table	Independence	Fisher's exact test	fisher.test(table(C1,C2))
(same)	(same)	(same)	Chi-square test	chisq.test(table(C1,C2))
(same)	(same)	Strength of relationship	phi coefficient, contingency coefficient, and Cramer's V	library(vcd); assocstats(table(C1,C2))
O1, O2 (similar to C1, C2)	Spearman's correlation coefficients	Strength of relationship	rank correlation	cor.test(C1,C2,method="spearman")
B1, X1	mean with sd for each B1	equal mean	t-test	t.test(X1 ~ B1)
(same)	median with semi-inter-quartile range for each B1	equal median	Wilcoxon's rank sum test	wilcox.test(X1 ~ B1)
(same)	logistic regression	how can X1 explain B1	Fitting logistic regression model	(res <- glm(as.factor(B1) ~ X1, family="binomial")); summary(res); library(fmsb); NagelkerkeR2(res)
C1, X1	mean with sd for each C1	equal mean	oneway ANOVA	oneway.test(X1 ~ C1) pairwise.t.test(X1,C1)
(same)	median with semi-inter-quartile range for each C1	equal order	Kruskal-Wallis test	kruskal.test(X1 ~ C1) pairwise.wilcox.test(X1,C1)
X1, X2	mean with sd	equal mean	t-test	t.test(X1,X2)
(same)	mean difference	no difference between each pair	paired-t-test	t.test(X1,X2,paired=TRUE) or t.test(X1-X2, mu=0)
(same)	median with semi-inter-quartile range	equal order	Wilcoxon's rank sum test	wilcox.test(X1,X2)
(same)	median difference	no order difference between each pair	Wilcoxon's signed rank test	wilcox.test(X1,X2,paired=TRUE)
(same)	Pearson' correlation coefficient	Strength of relationship	correlation	cor.test(X1,X2)
(same)	regression table	How can X2 explain X1	Fitting regression model	plot(X1 ~ X2); res <- lm(X1~X2); abline(res) summary(res)

Let's denote binary variables as `B1`, `B2`, ..., category variables as `C1`, `C2`, ... (ordered category variables as `O1`, `O2`, ...), and numeric variables as `X1`, `X2`, ...

The possible combinations of these variables and methods of analysis can be summarized as the above table.

For exercise, let's use the data below.

GENDER	SICKNESS	ETHNICITY	BMI	WELLBEING
M	HEALTHY	JAPANESE	25.0	GOOD
F	HEALTHY	JAPANESE	22.5	GOOD
M	DISEASE	JAPANESE	28.5	MODERATE
F	HEALTHY	INDONESIAN	17.5	BAD
M	DISEASE	INDONESIAN	30.0	BAD
M	HEALTHY	INDONESIAN	23.5	GOOD
F	DISEASE	KOREAN	20.0	MODERATE
F	HEALTHY	KOREAN	19.5	GOOD
M	DISEASE	KOREAN	24.5	BAD
F	HEALTHY	KOREAN	27.0	MODERATE

The easiest way to enter the data into R as a dataframe `x` is, first entering the data into the spreadsheet software like MS-Excel, selecting the range of the data, then activate R Console and enter `x <- read.delim("clipboard")`[*1].

However, it's also possible to directly enter the data into R as listed below.

```
x <- data.frame(
 GENDER=factor(c(1,2,1,2,1,1,2,2,1,2),labels=c("M","F")),
 SICKNESS=factor(c(1,1,2,1,2,1,2,1,2,1),labels=c("HEALTHY","DISEASE")),
 ETHNICITY=factor(c(1,1,1,2,2,2,3,3,3,3),labels=c("JAPANESE","INDONESIAN","KOREAN")),
 BMI=c(25.0,22.5,28.5,17.5,30.0,23.5,20.0,19.5,24.5,27.0),
 WELLBEING=ordered(c(3,3,2,1,1,3,2,3,1,2),labels=c("BAD","MODERATE","GOOD")))
```

2 Two binomial variables

```
table(x$GENDER,x$SICKNESS) # same as xtabs(~GENDER+SICKNESS, data=x)
fisher.test(table(x$GENDER,x$SICKNESS))
chisq.test(table(x$GENDER,x$SICKNESS))
require(vcd)
print(OR <- oddsratio(table(x$GENDER,x$SICKNESS),log=FALSE))
confint(OR)
assocstats(table(x$GENDER,x$SICKNESS))
```

3 Regression Analysis

単回帰分析とは，従属変数のばらつきを，1 つの独立変数のばらつきによって説明しようとする，モデルを使った分析法です。(Regression analysis is a model-based method to explain the variation of a dependent variable by the variation of a independent variable.)

作るグラフは散布図と回帰直線で，分析結果として求めるのは切片と回帰係数，それらの信頼区間，予測区間，相

[*1] For ordered variable, `x$WELLBEING <- ordered(x$WELLBEING, levels=c("BAD","MODERATE","GOOD"))` is needed after reading the data.

関係数の二乗（モデルが従属変数のばらつきのどれくらいの割合を説明するかを示します）といった値です。残差についても分析することを薦めます。(In the regression analysis, scattergram with regression line must be made, and the intercept and the regression coefficients with those confidence intervals and prediction intervals should be calculated. Adjusted R-squared is also to be calculated, which shows the proportion of the variance of the dependent variable explained by the model. Residuals are also to be analyzed.)

3.1 **分析の前提** (Prerequisite)

単回帰分析には，要素数が同じ 2 つのベクトルが必要です。通常は，同じ対象者について 2 つの量的変数（身長と体重など）があって，そのうち 1 つのばらつきを，他方のばらつきがどれくらい説明できるかというフレームで分析します。(Regression analysis requires 2 vectors with the same length. Usually, the frame of regression analysis is applyed to the data of 2 quantative variables (eg. height and weight) for each subject, in which how much variance of the dependent variable can be explained by the independent variable.)

本来, 独立変数は誤差を含んでいない既知の値であることが望ましいのですが, 実際には独立変数も測定値であることはよくあります。(As a principle, the independent variable should have accurate values without any measurement error, but in fact, both 2 values are measured values with some errors.)

実際のデータを使う場合，表計算ソフトで入力（ただし 1 行目は変数名とします）して `read.delim("clipboard")` あるいはファイルから入力して，`attach()` して分析するのが普通です。(To apply the regression analysis to actual data, reading the data is needed first, using `read.delim("clipboard")` after copying the area of data in MS-Excel table or using `read.delim("filename.txt")` after saving the data file "filename.txt" as tab-delimited text file in MS-Excel. The first line of the data should be the names of variables. After reading, the data frame should be attached by `attach()` function.)

3.2 **文法** (R statement rule)

従属変数が Y，独立変数が X だとすれば[*2]，分析の基本形は次の枠内のようになります。なお，どちらかの変数に欠損値が含まれている場合，その対象のデータは自動的に分析から除かれます。(Let Y as a dependent variable and X as an independent variable, the basic framework of regression analysis as following box, where the `runif` is a uniform random number generator in [0,1]. Here if missing value is included in either variable, its record is automatically omitted from the analysis.)

```
plot(Y ~ X)
res <- lm(Y ~ X)
abline(res)
summary(res)
pre <- data.frame(X=seq(1, 10, by=1))
pre.c <- predict(res, pre, interval="confidence", level=0.95)
pre.p <- predict(res, pre, interval="prediction", level=0.95)
matlines(pre, pre.c, lty=c(1,2,2), col=1)
matlines(pre, pre.p, lty=c(1,3,3), col=c(1,2,2))
plot(res)
```

1 行目が散布図を描く関数の基本形です。オプションとして `xlab="X` **軸の変数名**`"`とか，`main="`**グラフの表題**`"`とか `ylim=c(0,50)`（Y 軸の描画範囲を 0 から 50 までに設定する）とかいった設定が可能です。(The first line is the basic form to draw scattergram. Options of `xlab="the name of X axis variable"`, `main="title of the graph"`, `ylim=c(0,50)` (set graph limit of y axis in [0,50]) and so on are available.)

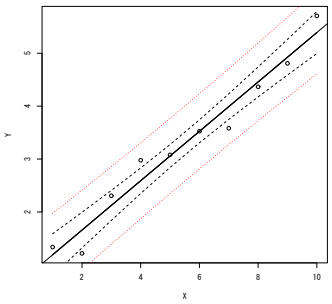
2 行目が回帰分析の本体です。(The second line is the regression analysis itself.)

[*2] `X <- 1:10; Y <- X/2+runif(10)` などとしてみてください。`runif(10)` は 0 以上 1 以下の一様乱数を 10 個発生させる関数です。

3 行目が回帰直線の描画です。`abline()` という関数に回帰分析の結果を付値したオブジェクトを渡せば，既に描かれている散布図に追加されます。(The third line is to add regression line to the scattergram.)

4 行目が回帰分析の結果の表示になります。`Coefficients` の `(intercept)` の行に切片が示され，`X` の行に回帰係数が示されます。右端の `p-value` はそれらがゼロと差がないという帰無仮説の検定結果です。(The 4th line shows the summary of regression analysis. The rightmost term of the line `X` is the result to test the null-hypothesis "the regression coefficient is 0".)

5 行目から 9 行目は回帰直線の信頼区間と予測区間を描いています。(The lines 5-9 are drawing the confidence intervals and prediction intervals of the regression line.)



10 行目では残差分析などがなされます。(The last line is to do residual analysis and so on.)

結果は，通常，散布図を示すとともに以下のような表の形でまとめます。複数の分析結果を 1 つの表にまとめて表示することも多いです。(The result of the regression analysis is usually shown as the following table. More than one results are often summarized into a single table.)

Table. The effect of X on Y.

Terms	Coefficients	p-value
Intercept	0.723	0.0069
X	0.467	5×10^{-7}

* Adjusted R^2=0.96

3.3 **実例** (practice)

組み込みデータ airquality を使って，`Ozone`（オゾン濃度）を従属変数，`Solar.R`（日照）を独立変数とする回帰分析をしてみましょう。(Let's try the regression analysis using the bulit-in data "airquality", where `Ozone` is a dependent variable and `Solar.R` is an independent variable.)

データフレーム内の変数を `lm()` 関数内で参照する場合，データフレーム名＄変数名とする方法の他に，変数名だけを指定しておき，data=データフレーム名とする方法もあります。この例では，下枠内のようになります。(The code listed below is an example to conduct this regression analysis.)

```
plot(Ozone ~ Solar.R, data=airquality)
res <- lm(Ozone ~ Solar.R, data=airquality)
abline(res)
summary(res)
pre <- data.frame(Solar.R=seq(min(airquality$Solar.R), max(airquality$Solar.R), len=100))
pre.c <- predict(res, pre, interval="confidence", level=0.95)
pre.p <- predict(res, pre, interval="prediction", level=0.95)
matlines(pre, pre.c, lty=c(1,2,2), col=1)
matlines(pre, pre.p, lty=c(1,3,3), col=c(1,2,2))
```