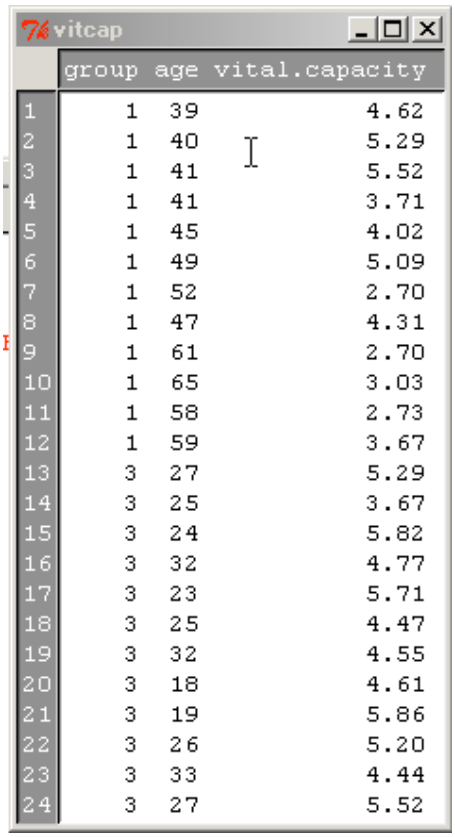
R Commander を用いた統計解析の基礎(ex) 問題解答

岡田　昌史

1.　まとめ(演習問題)解答例

1．1　ISwR パッケージのデータセット vitcap に関して、２つのグループ間で肺活量に差があるかどうかを t 検定を用いて分析してみましょう。

vitcap データセットはカドミウム産業の労働者の肺活量を、10 年以上曝露があった群と曝露がなかった群について測定したものです。



vitcap

	group	age	vital.capacity
1	1	39	4.62
2	1	40	5.29
3	1	41	5.52
4	1	41	3.71
5	1	45	4.02
6	1	49	5.09
7	1	52	2.70
8	1	47	4.31
9	1	61	2.70
10	1	65	3.03
11	1	58	2.73
12	1	59	3.67
13	3	27	5.29
14	3	25	3.67
15	3	24	5.82
16	3	32	4.77
17	3	23	5.71
18	3	25	4.47
19	3	32	4.55
20	3	18	4.61
21	3	19	5.86
22	3	26	5.20
23	3	33	4.44
24	3	27	5.52

変数 group が群（1: 曝露あり, 3:曝露なし）、age が年齢、vital.capacity が肺活量ですね。

ただし、**group は因子化されていない**ことに注意が必要です。このままでは R は group を連続量として解釈するので、このデータが２群に分かれていると判断してくれません。

そこで、まずは group を因子に変換します。「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...」ですね。

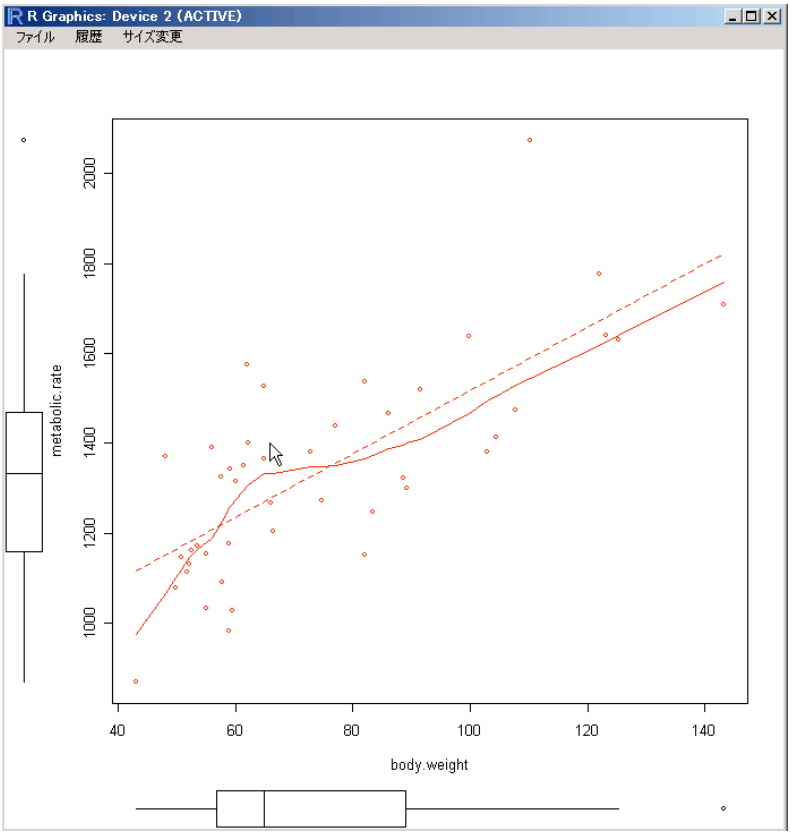
これで、「統計量」→「平均」→「独立サンプルの t 検定...」が実行できるようになります。

```
> t.test(vital.capacity~c.group, alternative='two.sided', conf.level=.

        Welch Two Sample t-test

data:  vital.capacity by c.group
t = -2.9228, df = 19.019, p-value = 0.008724
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.7904211 -0.2962456
sample estimates:
mean in group 曝露あり mean in group 曝露なし
             3.949167              4.992500
```

結果、p=0.008724 と帰無仮説は棄却され、曝露の有無によって肺活量の平均値に差があることが示されました。ただし、年齢によっても肺活量には差がでてくる可能性があるので、群間の年齢分布の差異についても検討が必要でしょう。

1．2　ISwR パッケージのデータセット rmr に関して、代謝率と体重との関係をグラフに示してください。また、この関係を回帰分析してみましょう。そのモデルに基づくと、体重が 70kg の場合、代謝率はどのくらいになるでしょうか？

rmr データセットは 44 名の女性について、体重(body.weight)と安静時のエネルギー代謝率(metabolic.rate)を示しているデータです。どちらも連続変数ですから、グラフは散布図ということになりますね。このあと、体重から代謝率を予測することになるので、体重を x 変数, 代謝率を y 変数としてプロットすると、そこに描かれる回帰直線が回帰分析の結果そのものになります。



では、回帰分析をしてみましょう。「統計量」→「モデルへの適合」→「線形回帰...」ですね。

```
Call:
lm(formula = metabolic.rate ~ body.weight, data = rmr)

Residuals:
    Min      1Q  Median      3Q     Max
-245.74 -113.99  -32.05  104.96  484.81

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 811.2267    76.9755  10.539 2.29e-13 ***
body.weight   7.0595     0.9776   7.221 7.03e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 157.9 on 42 degrees of freedom
Multiple R-Squared: 0.5539,     Adjusted R-squared: 0.5433
F-statistic: 52.15 on 1 and 42 DF,  p-value: 7.025e-09
```

このようなモデルになります。モデル式は、

$$metabolic.rate = 7.0595 \times body.weight + 811.2267$$

です。このモデルに基づけば、体重が 70kg の場合、代謝率は

$$7.0595 \times 70 + 811.2267 = 1305.392$$

となります。グラフの回帰直線から読み取ることでもだいたいの値を知ることができますね。

1．3　データセット juul のうち 25 歳を超える被験者のグループにおいて、IGF-I の値の平方根と年齢の関係を回帰分析してみましょう。

まず、「データ」→「アクティブデータセット」→「アクティブデータセットの部分集合を抽出...」を用いて、age > 25 という条件で juul データセットの部分集合をつくります。

さらに、「データ」→「アクティブデータセット内の変数の管理」→「新しい変数を計算...」で、igf1 の平方根を示す変数をつくります。平方根を求めるには、sqrt(igf1) を「計算式」にいれましょう。

あとは、この IGF-I の平方根を目的変数、age を説明変数とした回帰分析をすれば Ok です。

```
Call:
lm(formula = sqigf1 ~ age, data = juul.over25)

Residuals:
    Min      1Q  Median      3Q     Max
-4.8642 -1.1661  0.1018  0.9450  4.1136

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.71025    0.49462  37.828   <2e-16 ***
age         -0.10533    0.01072  -9.829   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.741 on 120 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-Squared: 0.446,      Adjusted R-squared: 0.4414
F-statistic:  96.6 on 1 and 120 DF,  p-value: < 2.2e-16
```

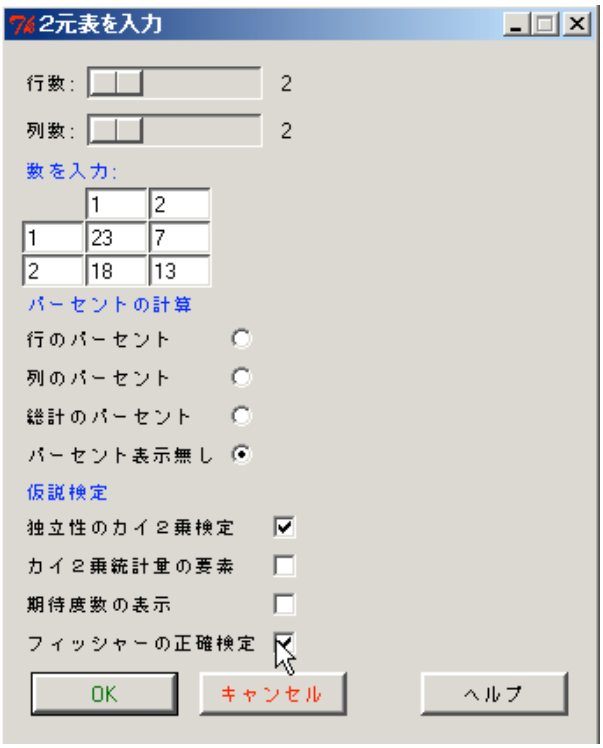
なかなか良いモデルになっていますね。

1．4　２種類の消化性潰瘍治療薬の比較を行ったところ、下表のような結果が得られました。この２つの薬の効果には差がみられたでしょうか？

薬剤名	治癒	非治癒	合計
薬 A	23	7	30
薬 B	18	13	31
合計	41	20	61

これは 2×2 の分割表で、治癒 - 非治癒という成績の違いと、A, B という薬剤の違いが独立しているかどうかを、Fisher の正確検定で検討すればよいです。

データはすでにクロス集計され、コンパクトな分割表にまとめられているので、データセットを読み込む必要はありません。「統計量」→「分割表」→「２元表の入力と分析...」を選択します。



「数を入力：」のところで分割表の値を直接入力し、2x2 ですから「フィッシャーの正確検定」をチェックします。

```
> fisher.test(.Table)

        Fisher's Exact Test for Count Data

data:  .Table
p-value = 0.1737
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.6936416 8.4948588
sample estimates:
odds ratio
  2.339104


> remove(.Table)
```

フィッシャーの正確検定の結果、p=0.1737 ということで、２つの要因は独立であるという帰無仮説を棄却できませんでしたので、成績の違いと薬の違いとの間に有意な関係は認められませんでした。

1．5　ISwR パッケージの lung データセットにおいて、３つの測定方法は異なる結果をもたらしているでしょうか？もしそうなら、有意に他と異なる結果をもたらしている方法はどれでしょうか？

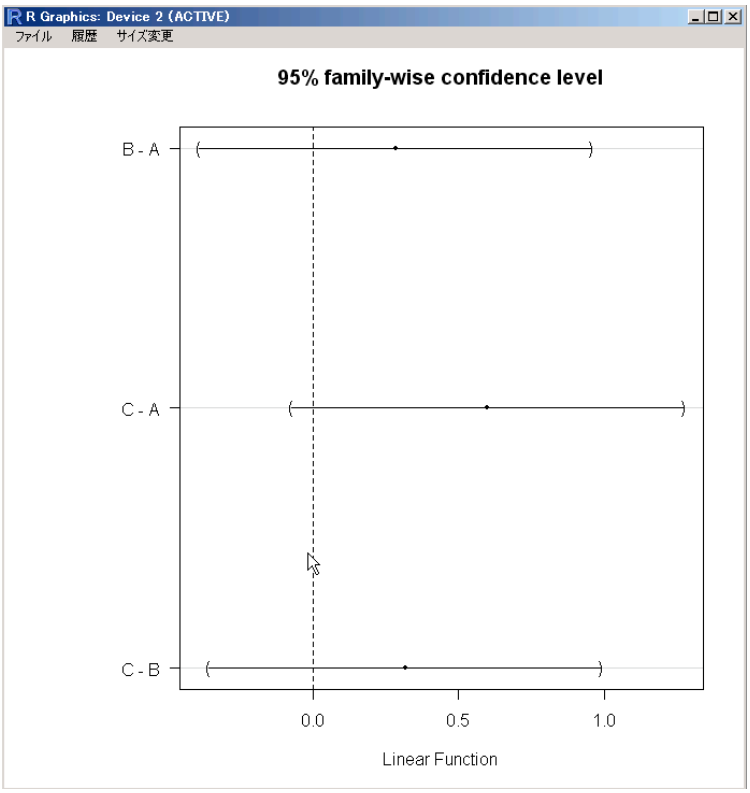
これは、肺容積の測定方法として３つの異なる方法を用いたときの測定結果のデータで、分散分析の例としてあげたつもりだったのですが、データをみてみると、測定方法のほかに対象者のデー

タも記録されていました。したがって、実際には測定法の違いだけを考慮した一元配置分散分析よりも、対象者の違いも考慮して二元配置分散分析を行うべき事例でしたね。講義では扱っていない部分を出題してしまいすみません。

対象者のデータがないものとして、一元配置分散分析をしてみると、method 変数を「グループ」、volume 変数を「目的変数」として分散分析をすることで、以下のような分散分析表が得られます。

```
> anova(.Anova)
Analysis of Variance Table

Response: volume
           Df  Sum Sq Mean Sq F value Pr(>F)
method      2 1.08111 0.54056  2.6893 0.1004
Residuals 15 3.01500 0.20100
```

F 検定の結果は p = 0.1004 ということで、測定法による肺容積の差は有意ではありませんでした。多重比較も行うことが出来ますが、



このようなグラフとなり、すべての差の信頼区間が 0 をはさんでいますから、やはり有意な差はどの測定法間にも存在しないようです。

しかし、対象者のデータを使って二元配置分散分析相当のことをすると話が変わってきます。

多元配置分散分析はコマンダーのメニューにもありますが、lung データに対して使用するとエラーが出てうまく結果が出ない(最新版では改善されているかもしれません)ようですから、ここは線形モデルでやってみましょう。Console からになります。

```
> attach(lung)
> model <- lm(volume ~ method + subject)
> anova(model)
Analysis of Variance Table

Response: volume
          Df  Sum Sq Mean Sq F value  Pr(>F)
method     2 1.08111 0.54056  6.4953 0.01557 *
subject    5 2.18278 0.43656  5.2457 0.01271 *
Residuals 10 0.83222 0.08322
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(model)

Call:
lm(formula = volume ~ method + subject)

Residuals:
     Min       1Q   Median       3Q      Max
-0.35556 -0.16389 -0.03889  0.17361  0.32778

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.17222    0.19232  16.494  1.4e-08 ***
method[T.B]  0.28333    0.16656   1.701  0.11975
method[T.C]  0.60000    0.16656   3.602  0.00483 **
subject[T.2] -0.83333    0.23555  -3.538  0.00538 **
subject[T.3]  0.10000    0.23555   0.425  0.68016
subject[T.4] -0.06667    0.23555  -0.283  0.78293
subject[T.5] -0.03333    0.23555  -0.142  0.89027
subject[T.6] -0.60000    0.23555  -2.547  0.02900 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2885 on 10 degrees of freedom
Multiple R-Squared: 0.7968,     Adjusted R-squared: 0.6546
F-statistic: 5.603 on 7 and 10 DF,  p-value: 0.00768

>
```

volume を method と subject で説明するモデルを作成して、その分散分析表を出力させたところ、F 検定で method, subject ともに有意になりました。(p=0.01557, p=0.01271)。したがって、測定法、対象者とも要因としての効果は有意にあることになります。

また、モデル要約で method[T.C] 係数が有意になっています(p=0.00483)。これは、測定法を A から C にかえるということが、肺容積測定値の変化を有意に説明できることを示しています。つまり、対象者(subject)の効果を調整して考えたときには、測定法 A と C の間には有意な差がみられるということです。

なお、B と C の比較をするには、測定法をダミー変数化するときの方式を(A を基準とするものから、B あるいは C を基準とするものへ)変えなければなりませんので、上記の解析ではそこまではやってみていません。

1．6　ISwR パッケージの secher データセットでは、腹部直径および児頭大横径、出生時体重をすべて対数変換すると、変数間の関係をいいモデルにすることができます。出生時体重を腹部直径と児頭大横径で予測するモデルをつくってみましょう。このモデルと、腹部直径のみ、あるいは児頭大横径のみで予測を行うモデルとはどちらがよいモデルになっているでしょうか。

まず、secher データセットを読み込んだのち、腹部直径(ad), 児頭大横径(bpd), 出生時体重(bwt)をそれぞれ対数変換して、logad, logbpd, logbwt などの新しい変数をつくります。これをもちいて、「統計量」→「モデルへの適合」→「線形モデル」から、

このようにモデルをつくります。モデルを要約してみると、

出力ウィンドウ

```
Call:
lm(formula = logbwt ~ logad + logbpd, data = secher)

Residuals:
      Min         1Q     Median         3Q        Max
-0.350742 -0.067409 -0.007916  0.057502  0.363595

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -5.8615     0.6617  -8.859 2.36e-14 ***
logad         1.4667     0.1467   9.998  < 2e-16 ***
logbpd        1.5519     0.2294   6.764 8.09e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1068 on 104 degrees of freedom
Multiple R-Squared: 0.8583,     Adjusted R-squared: 0.8556
F-statistic: 314.9 on 2 and 104 DF,  p-value: < 2.2e-16
```

となります。R^2が 0.8583 と、かなりあてはまりが良いモデルになっていることがわかりますね。一方、たとえば logbwt を logbpd だけで説明しようとするモデルをつくると、

出力ウィンドウ

```
Call:
lm(formula = logbwt ~ logbpd, data = secher)

Residuals:
     Min       1Q   Median       3Q      Max
-0.36478 -0.09725  0.01251  0.07703  0.51154

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -7.0862     0.9062  -7.819 4.35e-12 ***
logbpd        3.3320     0.2017  16.516  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1488 on 105 degrees of freedom
Multiple R-Squared: 0.7221,     Adjusted R-squared: 0.7194
F-statistic: 272.8 on 1 and 105 DF,  p-value: < 2.2e-16
```

このようになります。さきほどのモデルにくらべて、Residual Standard Error が大きくなっており、R^2は減っています。変数を減らすとむしろあてはまりがよくなるような場合もあるのですが、この場合は bpd と ad どちらもあったほうがいいようですね。