11
Phân tích phương sai (Analysis of variance)

Phân tích phương sai, như tên gọi, là một số phương pháp phân tích thống kê mà trọng điểm là phương sai (thay vì số trung bình). Phương pháp phân tích phương sai nằm trong "đại gia đình" các phương pháp có tên là mô hình tuyến tính (hay general linear models), bao gồm cả hồi qui tuyến tính mà chúng ta đã gặp trong chương trước. Trong chương này, chúng ta sẽ làm quen với cách sử dụng R trong phân tích phương sai. Chúng ta sẽ bắt đầu bằng một phân tích đơn giản, sau đó sẽ xem đến phân tích phương sai hai chiều, và các phương pháp phi tham số thông dụng.

11.1 Phân tích phương sai đơn giản (one-way analysis of variance - ANOVA)

Ví dụ 1. Bảng **thống kê 11.1** dưới đây so sánh độ galactose trong 3 nhóm bệnh nhân: nhóm 1 gồm 9 bệnh nhân với bệnh Crohn; nhóm 2 gồm 11 bệnh nhân với bệnh viêm ruột kết (colitis); và nhóm 3 gồm 20 đối tượng không có bệnh (gọi là nhóm đối chứng). Câu hỏi đặt ra là độ galactose giữa 3 nhóm bệnh nhân có khác nhau hay không? Gọi giá trị trung bình của ba nhóm là μ_1, μ_2, và μ_3, và nói theo ngôn ngữ của kiểm định giả thiết thì giả thiết đảo là:

$$H_o: \mu_1 = \mu_2 = \mu_3$$

Và giả thiết chính là: H_A: có một khác biệt giữa 3 μ_j (j=1,2,3)

Bảng 11.1. Độ galactose cho 3 nhóm bệnh nhân Crohn, viêm ruột kết và đối chứng

Nhóm 1: bệnh Crohn	Nhóm 2: bệnh viêm ruột kết	Nhóm 3: đối chứng (control)
1343 1393 1420 1641 1897 2160 2169 2279 2890	1264 1314 1399 1605 2385 2511 2514 2767 2827 2895 3011	1809 2850 1926 2964 2283 2973 2384 3171 2447 3257 2479 3271 2495 3288 2525 3358 2541 3643 2769 3657
n=9 Trung bình: 1910 SD: 516	*n*=11 Trung bình: 2226 SD: 727	*n*=20 Trung bình: 2804 SD: 527

Chú thích: SD là độ lệch chuẩn (standard deviation).

Mới xem qua vấn đề, có lẽ bạn đọc sẽ nghĩ rằng chúng ta cần làm 3 so sánh (bằng phương pháp kiểm định t): giữa nhóm 1 và 2, nhóm 2 và 3, và nhóm 1 và 3. Nhưng cách làm này không hợp lí, vì có ba phương sai khác nhau. Cách thích hợp nhất để so sánh này là phân tích phương sai. Phân tích phương sai có thể ứng dụng để so sánh nhiều nhóm cùng một lúc (simultaneous comparisons).

11.1.1 Mô hình phân tích phương sai

Để minh họa cho phương pháp phân tích phương sai, chúng ta phải dùng kí hiệu. Gọi độ galactose của bệnh nhân *i* thuộc nhóm *j* (*j* = 1, 2, 3) là x_{ij}. Mô hình phân tích phương sai phát biểu rằng:

$$x_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad [1]$$

Hay cụ thể hơn:

$$x_{i1} = \mu + \alpha_1 + \varepsilon_{i1}$$
$$x_{i2} = \mu + \alpha_2 + \varepsilon_{i2}$$
$$x_{i3} = \mu + \alpha_3 + \varepsilon_{i3}$$

Tức là, giá trị galactose của bất cứ bệnh nhân nào bằng giá trị trung bình của toàn quần thể (μ) cộng/trừ cho ảnh hưởng của nhóm *j* được đo bằng hệ số ảnh hưởng α_i, và sai số ε_{ij}. Một giả định khác là ε_{ij} phải tuân theo luật phân phối chuẩn với trung bình 0 và phương sai σ^2. Hai thông số cần ước tính là μ và α_i. Cũng như phân tích hồi qui tuyến tính, hai thông số này được ước tính bằng phương pháp bình phương nhỏ nhất; tức là tìm ước số $\hat{\mu}$ và $\hat{\alpha}_j$ sao cho $\sum\left(x_{ij} - \hat{\mu} - \hat{\alpha}_j\right)^2$ nhỏ nhất.

Quay lại với số liệu nghiên cứu trên, chúng ta có những tóm tắt thống kê như sau:

Nhóm	Số đối tượng (n_i)	Trung bình	Phương sai
1 – Crohn	$n_1 = 9$	$\bar{x}_1 = 1910$	$s_1^2 = 265944$
2 – Viêm ruột kết	$n_2 = 11$	$\bar{x}_2 = 2226$	$s_2^2 = 473387$
3 – Đối chứng	$n_3 = 20$	$\bar{x}_3 = 2804$	$s_3^2 = 277500$
Toàn bô mẫu	$n = 40$	$\bar{x} = 2444$	

Chú ý: $x_{ij} = \bar{x} + \left(\bar{x}_j - \bar{x}\right) + \left(x_{ij} - \bar{x}_j\right)$ [2]

Trong đó, $\bar{x}$ là số trung bình của toàn mẫu, và $\bar{x}_j$ là số trung bình của nhóm *j*. Nói cách khác, phần $\left(\bar{x}_j - \bar{x}\right)$ phản ánh độ khác biệt (hay cũng có thể gọi là hiệu số) giữa trung bình từng nhóm và trung bình toàn mẫu, và phần $\left(x_{ij} - \bar{x}_j\right)$ phản ánh hiệu số giữa một galactose của một đối tượng và số trung bình của từng nhóm. Theo đó, chúng ta có các nguồn dao động như sau:

- Tổng bình phương cho toàn bộ mẫu là:

$$SST = \sum_i \sum_j \left(x_{ij} - \bar{x}\right)^2$$

$= (1343–2444)^2 + (1393–2444)^2 + (1343 – 2444)^2 + \ldots + (3657–2444)^2$

$= 12133923$

- Tổng bình phương phản ánh độ khác nhau *giữa* các nhóm:

$$SSB = \sum_i \sum_j \left(\bar{x}_i - \bar{x}\right)^2 = \sum_j n_j \left(\bar{x}_j - \bar{x}\right)^2$$

$= 9(1910 – 2444)^2 + 11(2226 – 2444)^2 + 20(2804 – 2444)^2$

$= 5681168$

- Tổng bình phương phản ánh độ dao động *trong* mỗi nhóm:

$$SSW = \sum_i \sum_j \left(x_{ij} - \bar{x}_j\right)^2 = \sum_j \left(n_j - 1\right) s_j^2$$

$= (9-1)(265944) + (11-1)(473387) + (20-1)(277500)$

$= 12133922$

Có thể chứng minh rằng: $SST = SSB + SSW$.

SSW được tính từ mỗi bệnh nhân cho 3 nhóm, cho nên trung bình bình phương cho từng nhóm (mean square – *MSW*) là:

$$MSW = SSW / (N – k) = 12133922 / (40-3) = 327944$$

và trung bình bình phương giữa các nhóm là:

$$MSB = SSB / (k – 1) = 5681168 / (3-1) = 2841810$$

Trong đó *N* là tổng số bệnh nhân ($N = 40$) của ba nhóm, và $k = 3$ là số nhóm bệnh nhân. Nếu có sự khác biệt giữa các nhóm, thì chúng ta kì vọng rằng *MSB* sẽ lớn hơn *MSW*. Thành ra, để kiểm tra giả thiết, chúng ta có thể dựa vào kiểm định *F*:

$$F = MSB / MSW = 8.67 \qquad [3]$$

Với bậc tự do *k*-1 và *N-k*. Các số liệu tính toán trên đây có thể trình bày trong một bảng phân tích phương sai (ANOVA table) như sau:

Nguồn biến thiên (source of variation)	Bậc tự do (degrees of freedom)	Tổng bình phương (sum of squares)	Trung bình bình phương (mean square)	Kiểm định F
Khác biệt giữa các nhóm (between-group)	2	5681168	2841810	8.6655
Khác biệt trong từng nhóm (with-group)	37	12133923	327944	
Tổng số	39	12133923		

11.1.2 Phân tích phương sai đơn giản với R

Tất cả các tính toán trên tương đối phức tạp, và tốn khá nhiều thời gian. Tuy nhiên với R, các tính toán đó có thể làm trong vòng 1 giây, sau khi dữ liệu đã được chuẩn bị đúng cách.

(a) Nhập dữ liệu. Trước hết, chúng ta cần phải nhập dữ liệu vào R. Bước thứ nhất là báo cho R biết rằng chúng ta có ba nhóm bệnh nhân (1, 2 và 3), nhóm 1 gồm 9 người, nhóm 2 có 11 người, và nhóm 3 có 20 người:

```
> group <- c(1,1,1,1,1,1,1,1,1, 2,2,2,2,2,2,2,2,2,2,2,
             3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3)
```

Để phân tích phương sai, chúng ta phải định nghĩa biến `group` là một yếu tố - factor.

```
> group <- as.factor(group)
```

Bước kế tiếp, chúng ta nạp số liệu galactose cho từng nhóm như định nghĩa trên (gọi object là `galactose`):

```
> galactose <- c(1343,1393,1420,1641,1897,2160,2169,2279,
                 2890,1264,1314,1399,1605,2385,2511,2514,
                 2767,2827,2895,3011,1809,2850,1926,2964,
                 2283,2973,2384,3171,2447,3257,2479,3271,
                 2495,3288,2525,3358,2541,3643,2769,3657)
```

Đưa hai biến `group` và `galactose` vào một dataframe và gọi là `data`:

```
> data <- data.frame(group, galactose)
```

```
> attach(data)
```

Sau khi đã có dữ liệu sẵn sàng, chúng ta dùng hàm `lm()` để phân tích phương sai như sau:

```
> analysis <- lm(galactose ~ group)
```

Trong hàm trên chúng ta cho R biết biến `galactose` là một hàm số của `group`. Gọi kết quả phân tích là `analysis`.

(b) Kết quả phân tích phương sai. Bây giờ chúng ta dùng lệnh `anova` để biết kết quả phân tích:

```
> anova(analysis)
Analysis of Variance Table

Response: galactose
          Df  Sum Sq  Mean Sq F value  Pr(>F)
group      2 5683620  2841810 8.6655 0.0008191 ***
Residuals 37 12133923  327944
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Trong kết quả trên, có ba cột: `Df` (degrees of freedom) là bậc tự do; `Sum Sq` là tổng bình phương (sum of squares), `Mean Sq` là trung bình bình phương (mean square); `F value` là giá trị *F* như định nghĩa [3] vừa đề cập phần trên; và `Pr(>F)` là trị số P liên quan đến kiểm định *F*.

Dòng `group` trong kết quả trên có nghĩa là bình phương giữa các nhóm (between-groups) và `residual` là bình phương trong mỗi nhóm (within-group). Ở đây, chúng ta có:

$$SSB = 5683620 \text{ và } MSB = 2841810$$

và:

$$MSB = 2841810 \text{ và } MSB = 327944$$

Như vậy, $F = 2841810 / 327944 = 8.6655$.

Trị số p = 0.00082 có nghĩa là tín hiệu cho thấy có sự khác biệt về độ galactose giữa ba nhóm.

(c) Ước số. Để biết thêm chi tiết kết quả phân tích, chúng ta dùng lệnh `summary` như sau:

```
> summary(analysis)

Call:
```

```
lm(formula = galactose ~ group)

Residuals:
  Min   1Q Median   3Q  Max
-995.5 -437.9 102.0 456.0 979.8

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  1910.2   190.9    10.007   4.5e-12 ***
group2        316.3   257.4     1.229   0.226850
group3        894.3   229.9     3.891   0.000402 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 572.7 on 37 degrees of freedom
Multiple R-Squared: 0.319,   Adjusted R-squared: 0.2822
F-statistic: 8.666 on 2 and 37 DF, p-value: 0.0008191
```

Theo kết quả trên đây, `intercept` chính là $\hat{\mu}$ trong mô hình [1]. Nói cách khác, $\hat{\mu}$ = 1910 và sai số chuẩn là 190.9.

Để ước tính thông số $\hat{\alpha}_j$, R đặt $\hat{\alpha}_1$=0, và $\hat{\alpha}_2 = \hat{\alpha}_2 - \hat{\alpha}_1$ = 316.3, với sai số chuẩn là 257, và kiểm định t = 316.3 / 257 = 1.229 với trị số p = 0.2268. Nói cách khác, so với nhóm 1 (bệnh nhân Crohn), bệnh nhân viêm ruột kết có độ galactose trung bình cao hơn 257, nhưng độ khác biệt này không có ý nghĩa thống kê.

Tương tự, $\hat{\alpha}_3 = \hat{\alpha}_3 - \hat{\alpha}_1$ = 894.3, với sai số chuẩn là 229.9, kiểm định t=894.3/229.9=3.89, và trị số p = 0.00040. So với bệnh nhân Crohn, nhóm đối chứng có độ galactose cao hơn 894, và mức độ khác biệt này có ý nghĩa thống kê.

11.2 So sánh nhiều nhóm (multiple comparisons) và điều chỉnh trị số p

Cho *k* nhóm, chúng ta có ít nhất là *k(k-1)*/2 so sánh. Ví dụ trên có 3 nhóm, cho nên tổng số so sánh khả dĩ là 3 (giữa nhóm 1 và 2, nhóm 1 và 3, và nhóm 2 và 3). Khi *k*=10, số lần so sánh có thể lên rất cao. Như đã đề cập trong chương 7, khi có nhiều so sánh, trị số p tính toán từ các kiểm định thống kê không còn ý nghĩa ban đầu nữa, bởi vì các kiểm định này có thể cho ra kết quả dương tính giả (tức kết quả với p<0.05 nhưng trong thực tế không có khác nhau hay ảnh hưởng). Do đó, trong trường hợp có nhiều so sánh, chúng ta cần phải điều chỉnh trị số p sao cho hợp lí.

Có khá nhiều phương pháp điều chỉnh trị số p, và 4 phương pháp thông dụng nhất là: Bonferroni, Scheffé, Holm và Tukey (tên của 4 nhà thống kê học). Phương pháp nào thích hợp nhất? Không có câu trả lời dứt khoát cho câu hỏi này, nhưng hai điểm sau đây có thể giúp bạn đọc quyết định tốt hơn:

(a) Nếu $k < 10$, chúng ta có thể áp dụng bất cứ phương pháp nào để điều chỉnh trị số p. Riêng cá nhân tôi thì thấy phương pháp Tukey thường rất hữu ích trong so sánh.

(b) Nếu $k > 10$, phương pháp Bonferroni có thể trở nên rất "bảo thủ". Bảo thủ ở đây có nghĩa là phương pháp này rất ít khi nào tuyên bố một so sánh có ý nghĩa thống kê, dù trong thực tế là có thật! Trong trường hợp này, hai phương pháp Tukey, Holm và Scheffé có thể áp dụng.

Ở đây, chúng ta sẽ không bàn đến lí thuyết đằng sau các phương pháp này (vì bạn đọc có thể tham khảo trong các sách giáo khoa về thống kê), nhưng sẽ chỉ cách sử dụng R để tiến hành các so sánh theo phương pháp của Tukey.

Quay lại ví dụ trên, các trị số p trên đây là những trị số chưa được điều chỉnh cho so sánh nhiều lần. Trong chương về trị số p, đã nói các trị số này phóng đại ý nghĩa thống kê, không phản ánh trị số p lúc ban đầu (tức 0.05). Để điều chỉnh cho nhiều so sánh, chúng ta phải sử dụng đến phương pháp điều chỉnh Bonferroni.

Chúng ta có thể dùng lệnh `pairwise.t.test` để có được tất cả các trị số p so sánh giữa ba nhóm như sau:

```
> pairwise.t.test(galactose, group, p.adj="bonferroni")

    Pairwise comparisons using t tests with pooled SD

data: galactose and group

   1      2
 2 0.6805 -
 3 0.0012 0.0321

P value adjustment method: bonferroni
```

Kết quả trên cho thấy trị số p giữa nhóm 1 (Crohn) và viêm ruột kết là 0.6805 (tức không có ý nghĩa thống kê); giữa nhóm Crohn và đối chứng là 0.0012 (có ý nghĩa thống kê), và giữa nhóm viêm ruột kết và đối chứng là 0.0321 (tức cũng có ý nghĩa thống kê).

Một phương pháp điều chỉnh trị số p khác có tên là phương pháp Holm:

```
> pairwise.t.test(galactose, group)
    Pairwise comparisons using t tests with pooled SD

data: galactose and group

  1      2
2 0.2268 -
3 0.0012 0.0214

P value adjustment method: holm
```

Kết quả này cũng không khác so với phương pháp Bonferroni.

Tất cả các phương pháp so sánh trên sử dụng một sai số chuẩn chung cho cả ba nhóm. Nếu chúng ta muốn sử dụng cho từng nhóm thì lệnh sau đây (`pool.sd=F`) sẽ đáp ứng yêu cầu đó:

```
> pairwise.t.test(galactose, group, pool.sd=FALSE)

    Pairwise comparisons using t tests with non-pooled SD

data: galactose and group

   1     2
2 0.2557 -
3 0.0017 0.0544

P value adjustment method: holm
```

Một lần nữa, kết quả này cũng không làm thay đổi kết luận.

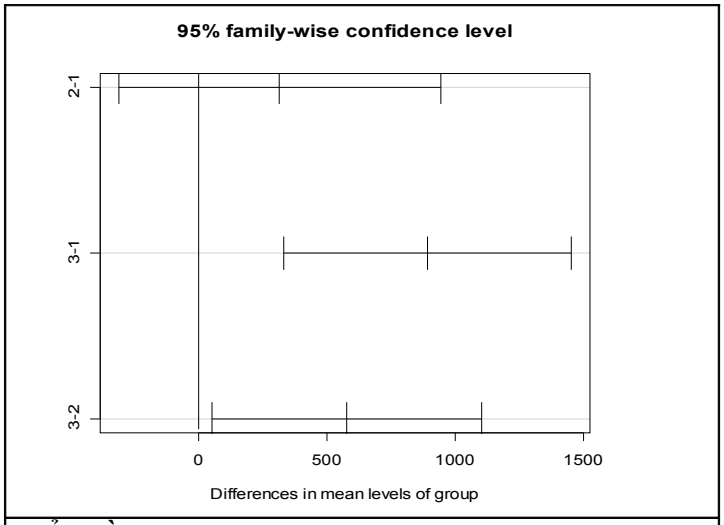
11.2.1 So sánh nhiều nhóm bằng phương pháp Tukey

Trong các phương pháp trên, chúng ta chỉ biết trị số p so sánh giữa các nhóm, nhưng không biết mức độ khác biệt cũng như khoảng tin cậy 95% giữa các nhóm. Để có những ước số này, chúng ta cần đến một hàm khác có tên là `aov` (viết tắt từ analysis of variance) và hàm `TukeyHSD` (HSD là viết tắt từ Honest Significant Difference, tạm dịch là "Khác biệt có ý nghĩa thành thật") như sau:

```
> res <- aov(galactose ~ group)
> TukeyHSD (res)
 Tukey multiple comparisons of means
  95% family-wise confidence level
```

```
Fit: aov(formula = galactose ~ group)

$group
    diff     lwr        upr     p adj
2-1 316.3232 -312.09857 944.745 0.4439821
3-1 894.2778 333.07916 1455.476 0.0011445
3-2 577.9545  53.11886 1102.790 0.0281768
```

Kết quả trên cho chúng ta thấy nhóm 3 và 1 khác nhau khoảng 894 đơn vị, và khoảng tin cậy 95% từ 333 đến 1455 đơn vị. Tương tự, galactose trong nhóm bệnh nhân viêm ruột kết thấp hơn nhóm đối chứng (nhóm 3) khoảng 578 đơn vị, và khoảng tin cậy 95% từ 53 đến 1103.



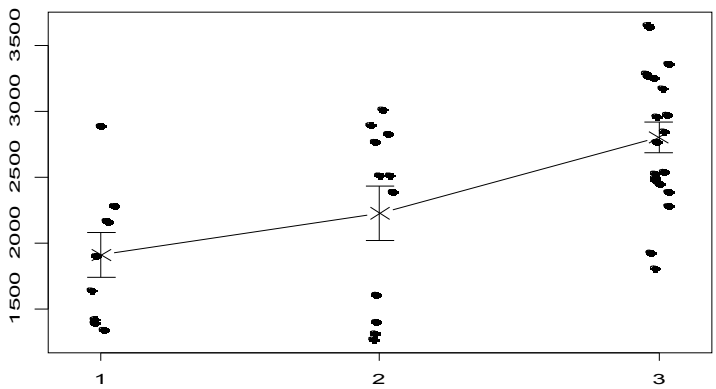
Biểu đồ 11.1. Trung bình hiệu và khoảng tin cậy 95% giữa nhóm 1 và 2, 1 và 3, và 3 và 2. Trục hoành là độ galactose, trục tung là ba so sánh.

11.2.2 Phân tích bằng biểu đồ

Một phân tích thống kê không thể nào hoàn tất nếu không có một đồ thị minh họa cho kết quả. Các lệnh sau đây vẽ đồ thị thể hiện độ galactose trung bình và sai số chuẩn cho từng nhóm bệnh nhân. Biểu đồ này cho thấy, nhóm bệnh nhân Crohn có độ galactose thấp nhất (nhưng không thấp hơn nhóm viêm ruột kết), và cả hai nhóm thấp hơn nhóm đối chứng và sứ khác biệt này có ý nghĩa thống kê.

```
> xbar <- tapply(galactose, group, mean)
> s <- tapply(galactose, group, sd)
> n <- tapply(galactose, group, length)
```

```
> sem <- s/sqrt(n)
> stripchart(galactose ~ group, "jitter", jit=0.05,
             pch=16, vert=TRUE)
> arrows(1:3, xbar+sem, 1:3, xbar-sem, angle=90, code=3,
         length=0.1)
> lines(1:3, xbar, pch=4, type="b", cex=2)
```



Biểu đồ 11.2. Độ galactose của nhóm 1 (bệnh nhân Crohn), nhóm 2 (bệnh nhân viêm ruột kết), và nhóm 3 (đối chứng).

11.3 Phân tích bằng phương pháp phi tham số

Phương pháp so sánh nhiều nhóm phi tham số (non-parametric statistics) tương đương với phương pháp phân tích phương sai là Kruskal-Wallis. Cũng như phương pháp Wilcoxon so sánh hai nhóm theo phương pháp phi tham số, phương pháp Kruskal-Wallis cũng biến đổi số liệu thành thứ bậc (ranks) và phân tích độ khác biệt thứ bậc này giữa các nhóm. Hàm `kruskal.test` trong R có thể giúp chúng ta trong kiểm định này:

```
> kruskal.test(galactose ~ group)

    Kruskal-Wallis rank sum test

data: galactose by group

Kruskal-Wallis chi-squared = 12.1381, df = 2, p-value =
0.002313
```

Trị số p từ kiểm định này khá thấp (p = 0.002313) cho thấy có sự khác biệt giữa ba nhóm như phân tích phương sai qua hàm `lm` trên đây. Tuy nhiên,

một bất tiện của kiểm định phi tham số Kruskal-Wallis là phương pháp này không cho chúng ta biết hai nhóm nào khác nhau, mà chỉ cho một trị số p chung. Trong nhiều trường hợp, phân tích phi tham số như kiểm định Kruskal-Wallis thường không có hiệu quả như các phương pháp thống kê tham số (parametric statistics).

11.4 Phân tích phương sai hai chiều (two-way analysis of variance - ANOVA)

Phân tích phương sai đơn giản hay một chiều chỉ có một yếu tố (factor). Nhưng phân tích phương sai hai chiều (two-way ANOVA), như tên gọi, có hai yếu tố. Phương pháp phân tích phương sai hai chiều chỉ đơn giản khai triển từ phương pháp phân tích phương sai đơn giản. Thay vì ước tính phương sai của một yếu tố, phương pháp phương sai hai chiều ước tính phương sai của hai yếu tố.

Ví dụ 2. Trong ví dụ sau đây, để đánh giá hiệu quả của một kĩ thuật sơn mới, các nhà nghiên cứu áp dụng sơn trên 3 loại vật liệu (1, 2 và 3) trong hai điều kiện (1, 2). Mỗi điều kiện và loại vật liệu, nghiên cứu được lặp lại 3 lần. Độ bền được đo là chỉ số bền bỉ (tạm gọi là `score`). Tổng cộng, có 18 số liệu như sau:

Bảng 11.2. Độ bền bæ của sơn cho 2 điều kiện và 3 vật liệu

Điều kiện (*i*)	Vật liệu (*j*)		
	1	2	3
1	4.1, 3.9, 4.3	3.1, 2.8, 3.3	3.5, 3.2, 3.6
2	2.7, 3.1, 2.6	1.9, 2.2, 2.3	2.7, 2.3, 2.5

Số liệu này có thể tóm lược bằng số trung bình cho từng điều kiện và vật liệu trong bảng thống kê sau đây:

Bảng 11.3. Tóm lược số liệu từ thí nghiệm độ bền bĩ của nước sơn

Điều kiện (*i*)	Vật liệu (*j*)			Trung bình cho 3 vật liệu
	1	2	3	
Trung bình				
1	4.10	3.07	3.43	**3.533**
2	2.80	2.13	2.50	**2.478**
Trung bình 2 nhóm	**3.450**	**2.600**	**2.967**	**3.00**

Phương sai				
1	0.040	0.063	0.043	
2	0.070	0.043	0.040	

Những tính toán sơ khởi trên đây cho thấy có thể có sự khác nhau (hay ảnh hưởng) của điều kiện và vật liệu thí nghiệm.

Gọi x_{ij} là `score` của điều kiện i (i = 1, 2) cho vật liệu j (j = 1, 2, 3). (Để đơn giản hóa vấn đề, chúng ta tạm thời bỏ qua k đối tượng). Mô hình phân tích phương sai hai chiều phát biểu rằng:

$$x_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \qquad [4]$$

Hay cụ thể hơn:

$$x_{11} = \mu + \alpha_1 + \beta_1 + \varepsilon_{11}$$

$$x_{12} = \mu + \alpha_1 + \beta_2 + \varepsilon_{12}$$

$$x_{13} = \mu + \alpha_1 + \beta_3 + \varepsilon_{11}$$

$$x_{21} = \mu + \alpha_2 + \beta_1 + \varepsilon_{21}$$

$$x_{22} = \mu + \alpha_2 + \beta_2 + \varepsilon_{22}$$

$$x_{23} = \mu + \alpha_2 + \beta_3 + \varepsilon_{21}$$

μ là số trung bình cho toàn quần thể, các hệ số α_i (ảnh hưởng của điều kiện i)và β_j (ảnh hưởng của vật liệu j) cần phải ước tính từ số liệu thực tế. ε_{ij} được giả định tuân theo luật phân phối chuẩn với trung bình 0 và phương sai σ^2.

Trong phân tích phương sai hai chiều, chúng ta cần chia tổng bình phương ra thành 3 nguồn:

- Nguồn thứ nhất là tổng bình phương do khác biệt giữa 2 điều kiện:

$$SSc = \sum_i n_i \left(\bar{x}_i - \bar{x} \right)^2$$

$$= 9(3.533 - 3.00)^2 + 9(2.478 - 3.00)^2$$

$$= 5.01$$

- Nguồn thứ hai là tổng bình phương do khác biệt giữa 3 vật liệu:

$$SSm = \sum_j n_j \left(\bar{x}_j - \bar{x} \right)^2$$

$= 6(3.45 - 3.00)^2 + 6(2.60 - 3.00)^2 + 6(2.967 - 3.00)^2$

$= 2.18$

- Nguồn thứ ba là tổng bình phương phần dư (residual sum of squares):

$$SSe = \sum_i \sum_j \left(x_{ij} - \bar{x}_i - \bar{x}_j + \bar{x} \right)^2 = \sum \left(n_{ij} - 1 \right) s_{ij}^2$$

$= 2(0.040) + 2(0.063) + 2(0.043) + 2(0.070) + 2(0.043) + 2(0.040)$

$= 0.73$

Trong các phương trình trên, $n = 3$ (lặp lại 3 lần cho mỗi điều kiện và vật liệu), $m = 3$ vật liệu, $\bar{x}$ là số trung bình cho toàn mẫu, $\bar{x}_i$ là số trung bình cho từng điều kiện, $\bar{x}_j$ là số trung bình cho từng vật liệu. Vì *SSc* có *m*-1 bậc tự do, *SSm* có (*n* -1) bậc tự do, và *SSe* có *N–nm*+2 bậc tự do, trong đó *N* là tổng số mẫu (tức 18). Do đó, các trung bình bình phương

- Giữa hai điều kiện: $MSc = SSc / (m-1) = 5.01 / 1 = 5.01$
- Giữa ba vật liệu: $MSm = SSc / (n-1) = 2.18 /2 = 1.09$
- Phần dư: $MSe = SSe / (N-nm+2) = 0.73 / 14 = 0.052$

Do đó, so sánh độ khác biệt giữa hai điều kiện dựa vào kiểm định $F = MSc/Mse$ với bậc tự do 1 và 14. Tương tự, so sánh độ khác biệt giữa ba vật liệu có thể dựa vào kiểm định $F = MSm/Mse$ với bậc tự do 2 và 14. Các phân tích trên có thể trình bày trong một bảng phân tích phương sai như sau:

Nguồn biến thiên (source of variation)	Bậc tự do (degrees of freedom)	Tổng bình phương (sum of squares)	Trung bình bình phương (mean square)	Kiểm định F
Khác biệt giữa 2 điều kiện	1	5.01	5.01	95.6
Khác biệt giữa 3 vật liệu	2	2.18	1.09	20.8
Phần dư (residual)	14	0.73	0.052	
Tổng số	17	7.92		

11.4.1 Phân tích phương sai hai chiều với R

(a) Bước đầu tiên là nhập số liệu từ bảng 11.2 vào R. Chúng ta cần phải tổ chức dữ liệu sao cho có 4 biến như sau:

Condition (điều kiện)	Material (vật liệu)	Đối tượng	Score
1	1	1	4.1
1	1	2	3.9
1	1	3	4.3
1	2	4	3.1
1	2	5	2.8
1	2	6	3.3
1	3	7	3.5
1	3	8	3.2
1	3	9	3.6
2	1	10	2.7
2	1	11	3.1
2	1	12	2.6
2	2	13	1.9
2	2	14	2.2
2	2	15	2.3
2	3	16	2.7
2	3	17	2.3
2	3	18	2.5

Chúng ta có thể tạo ra một dãy số bằng cách sử dụng hàm `gl` (generating levels). Cách sử dụng hàm này có thể minh họa như sau:

```
> gl(9, 1, 18)
[1] 1 2 3 4 5 6 7 8 9 1 2 3 4 5 6 7 8 9
Levels: 1 2 3 4 5 6 7 8 9
```

Trong lệnh trên, chúng ta tạo ra một dãy số 1,2,3, … 9 hai lần (với tổng số 18 số). Mỗi một lần là một nhóm. Trong khi lệnh:

```
> gl(4, 9, 36)
[1] 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 4 4 4 4
4 4 4 4 4
Levels: 1 2 3 4
```

Trong lệnh trên, chúng ta tạo ra một dãy số với 4 bậc (1,2,3, 4) 9 lần (với tổng số 36 số).

Do đó, để tạo ra các bậc cho điều kiện và vật liệu, chúng ta lệnh như sau:

```
> condition <- gl(2, 9, 18)
```

```
> material <- gl(3, 3, 18)
```

Và tạo nên 18 mã số (từ 1 đến 18):

```
> id <- 1:18
```

Sau cùng là số liệu cho `score`:

```
> score <- c(4.1,3.9,4.3, 3.1,2.8,3.3, 3.5,3.2,3.6,
             2.7,3.1,2.6, 1.9,2.2,2.3, 2.7,2.3,2.5)
```

Tất cả cho vào một dataframe tên là `data`:

```
> data <- data.frame(condition, material, id, score)
> attach(data)
```

(b) Phân tích và kết quả sơ khởi. Bây giờ số liệu đã sẵn sàng cho phân tích. Để phân tích phương sai hai chiều, chúng ta vẫn sử dụng lệnh `lm` với các thông số như sau:

```
> twoway <- lm(score ~ condition + material)
> anova(twoway)
Analysis of Variance Table

Response: score
          Df Sum Sq Mean Sq  F value  Pr(>F)
condition 1  5.0139  5.0139  95.575  1.235e-07 ***
material  2  2.1811  1.0906  20.788  6.437e-05 ***
Residuals 14 0.7344  0.0525
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ba nguồn dao động (variation) của `score` được phân tích trong bảng trên. Qua trung bình bình phương (mean square), chúng ta thấy ảnh hưởng của điều kiện có vẻ quan trọng hơn là ảnh hưởng của vật liệu thí nghiệm. Tuy nhiên, cả hai ảnh hưởng đều có ý nghĩa thống kê, vì trị số p rất thấp cho hai yếu tố.

(c) Ước số. Chúng ta yêu cầu R tóm lược các ước số phân tích bằng lệnh `summary`:

```
> summary(twoway)
Call:
lm(formula = score ~ condition + material)

Residuals:
   Min    1Q  Median    3Q   Max
-0.32778 -0.16389 0.03333 0.16111 0.32222
```

```
Coefficients:
          Estimate Std. Error  t value Pr(>|t|)
(Intercept)  3.9778   0.1080  36.841   2.43e-15 ***
condition2  -1.0556   0.1080   -9.776  1.24e-07 ***
material2   -0.8500   0.1322   -6.428  1.58e-05 ***
material3   -0.4833   0.1322   -3.655  0.0026 **
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.229 on 14 degrees of freedom
Multiple R-Squared: 0.9074,   Adjusted R-squared: 0.8875
F-statistic: 45.72 on 3 and 14 DF, p-value: 1.761e-07
```

Kết quả trên cho thấy so với điều kiện 1, điều kiện 2 có `score` thấp hơn khoảng 1.056 và sai số chuẩn là 0.108, với trị số p = 1.24e-07, tức có ý nghĩa thống kê. Ngoài ra, so với vật liệu 1, `score` cho vật liệu 2 và 3 cũng thấp hơn đáng kể với độ thấp nhất ghi nhận ở vật liệu 2, và ảnh hưởng của vật liệu thí nghiệm cũng có ý nghĩa thống kê.

Giá trị có tên là "`Residual standard error`" được ước tính từ trung bình bình phương phần dư trong phần (a), tức là $\sqrt{0.0525}$ = 0.229, tức là ước số của $\hat{\sigma}$.

Hệ số xác định bội (R^2) cho biết hai yếu tố điều kiện và vật liệu giải thích khoảng 91% độ dao động của toàn bộ mẫu. Hệ số này được tính từ tổng bình phương trong kết quả phần (a) như sau:

$$R^2 = \frac{5.0139 + 2.1811}{5.0139 + 2.1811 + 0.7344} = 0.9074$$

Và sau cùng, hệ số R^2 điều chỉnh phản ánh độ "cải tiến" của mô hình. Để hiểu hệ số này tốt hơn, chúng ta thấy phương sai của toàn bộ mẫu là s^2 = (5.0139 + 2.1811 + 0.7344) / 17 = 0.4644. Sau khi điều chỉnh cho ảnh hưởng của điều kiện và vật liệu, phương sai này còn 0.0525 (tức là residual mean square). Như vậy hai yếu tố này làm giảm phương sai khoảng 0.4644 – 0.0525 = 0.4119. Và hệ số R^2 điều chỉnh là:

$$\text{Adj } R^2 = 0.4119 / 0.4644 = 0.88$$

Tức là sau khi điều chỉnh cho hai yếu tố điều kiện và vật liệu phương sai của score giảm khoảng 88%.

(d) Hiệu ứng tương tác (interaction effects)

Để cho phân tích hoàn tất, chúng ta còn phải xem xét đến khả năng ảnh hưởng của hai yếu tố này có thể tương tác nhau (interactive effects). Tức là mô hình `score` trở thành:

$$x_{ij} = \mu + \alpha_i + \beta_j + \left(\alpha_i \beta_j\right)_{ij} + \varepsilon_{ij}$$

Chú ý phương trình trên có phần $\left(\alpha_i \beta_j\right)_{ij}$ phản ánh sự tương tác giữa hai yếu tố. Và chúng ta chỉ đơn giản lệnh R như sau:

```
> anova(twoway <- lm(score ~ condition+
material+condition*material))
Analysis of Variance Table

Response: score
              Df Sum Sq   Mean Sq F value   Pr(>F)
condition      1  5.0139  5.0139  100.2778  3.528e-07 ***
material       2  2.1811  1.0906   21.8111  0.0001008 ***
condition:material
               2  0.1344  0.0672    1.3444  0.2972719
Residuals     12  0.6000  0.0500
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Kết quả phân tích trên (p = 0.297 cho ảnh hưởng tương tác). Chúng ta có bằng chứng để kết luận rằng ảnh hưởng tương tác giữa vật liệu và điều kiện không có ý nghĩa thống kê, và chúng ta chấp nhận mô hình [4], tức không có tương tác.

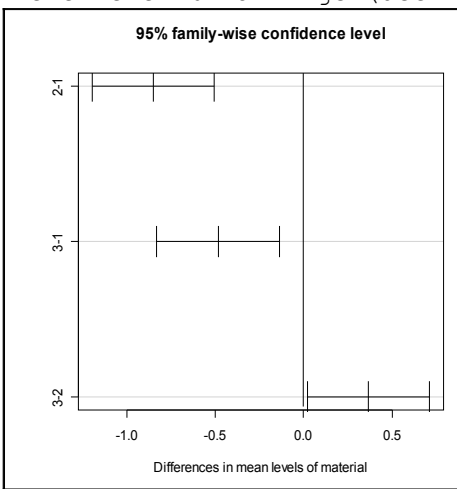
(e) So sánh giữa các nhóm. Chúng ta sẽ ước tính độ khác biệt giữa hai điều kiện và ba vật liệu bằng hàm `TukeyHSD` với `aov`:

```
> res <- aov(score ~ condition+ material+condition)
> TukeyHSD(res)
 Tukey multiple comparisons of means
  95% family-wise confidence level

Fit: aov(formula = score ~ condition + material +
condition)

$condition
     diff      lwr       upr       p adj
2-1 -1.055556 -1.287131 -0.8239797 1e-07

$material
     diff       lwr         upr       p adj
2-1 -0.8500000 -1.19610279 -0.5038972 0.0000442
```

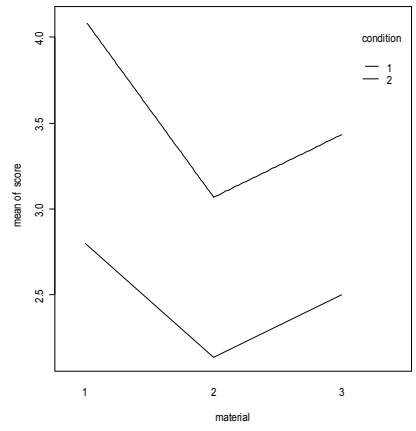
```
3-1 -0.4833333 -0.82943612 -0.1372305 0.0068648
3-2  0.3666667  0.02056388  0.7127695 0.0374069
```

Biểu đồ 11.3 sau đây sẽ minh họa cho các kết quả trên:

```
> plot(TukeyHSD(res), ordered=TRUE)
There were 16 warnings (use warnings() to see them)
```



Biểu đồ 11.3. So sánh giữa 3 loại vật liệu bằng phương pháp Tukey.



Biểu đồ 11.4. Trung bình score cho từng điều kiện 1 (đường đứt đoạn) và điều kiện 2 cho 3 loại vật liệu.

(f) Biểu đồ. Để xem qua độ ảnh hưởng của hai yếu tố điều kiện và vật liệu, chúng ta cần phải có một đồ thị, mà trong phân tích phương sai gọi là đồ thị tương tác. Hàm `interaction.plot` cung cấp phương tiện để vẽ biểu đồ này (xem **biểu đồ 11.4**):

```
> interaction.plot(score, condition, material)
```

11.5 Phân tích hiệp biến (analysis of covariance - ANCOVA)

Phân tích hiệp biến (sẽ viết tắt là ANCOVA) là phương pháp phân tích sử dụng cả hai mô hình hồi qui tuyến tính và phân tích phương sai. Trong phân tích hồi qui tuyến tính, cả hai biến phụ thuộc (dependent variable, cũng có thể gọi là "biến ứng" – response variable) và biến độc lập (independent variable hay predictor variable) phần lớn là ở dạng liên tục (continuous variable), như độ cholesterol và độ tuổi chẳng hạn. Trong phân tích phương sai, biến phụ thuộc là biến liên tục, còn biến độc lập thì ở dạng thứ bậc và thể loại (categorical variable), như độ galactose và nhóm bệnh nhân trong ví dụ 1 chẳng hạn. Trong

phân tích hiệp biến, biến phụ thuộc là liên tục, nhưng biến độc lập có thể là liên tục và thể loại.

Ví dụ 3. Trong nghiên cứu mà kết qủa được trình bày dưới đây, các nhà nghiên cứu đo chiều cao và độ tuổi của 18 học sinh thuộc vùng thành thị (urban) và 14 học trò thuộc vùng nông thôn (rural).

Bảng 11.4. Chiều cao của học trò vùng thành thị và nông thôn

Area	ID	Age (months)	Height (cm)
urban	1	109	137.6
urban	2	113	147.8
urban	3	115	136.8
urban	4	116	140.7
urban	5	119	132.7
urban	6	120	145.4
urban	7	121	135.0
urban	8	124	133.0
urban	9	126	148.5
urban	10	129	148.3
urban	11	130	147.5
urban	12	133	148.8
urban	13	134	133.2
urban	14	135	148.7
urban	15	137	152.0
urban	16	139	150.6
urban	17	141	165.3
urban	18	142	149.9
rural	1	121	139.0
urban	2	121	140.9
urban	3	128	134.9
urban	4	129	149.5
urban	5	131	148.7
urban	6	132	131.0
urban	7	133	142.3
urban	8	134	139.9
urban	9	138	142.9
urban	10	138	147.7
urban	11	138	147.7
urban	12	140	134.6
urban	13	140	135.8
urban	14	140	148.5

Câu hỏi đặt ra là có sự khác biệt nào về chiều cao giữa trẻ em ở thành thị và nông thôn hay không. Nói cách khác, môi trường cư trú có ảnh hưởng đến chiều cao hay không, và nếu có thì mức độ ảnh hưởng là bao nhiêu?

Một yếu tố có ảnh hưởng lớn đến chiều cao là độ tuổi. Trong độ tuổi trưởng thành, chiều cao tăng theo độ tuổi. Do đó, so sánh chiều cao giữa hai nhóm chỉ có thể khách quan nếu độ tuổi giữa hai nhóm phải tương đương nhau. Để đảm bảo tính khách quan của so sánh, chúng ta cần phải phân tích số liệu bằng mô hình hiệp biến.

Việc đầu tiên là chúng ta phải nhập số liệu vào R với những lệnh sau đây:

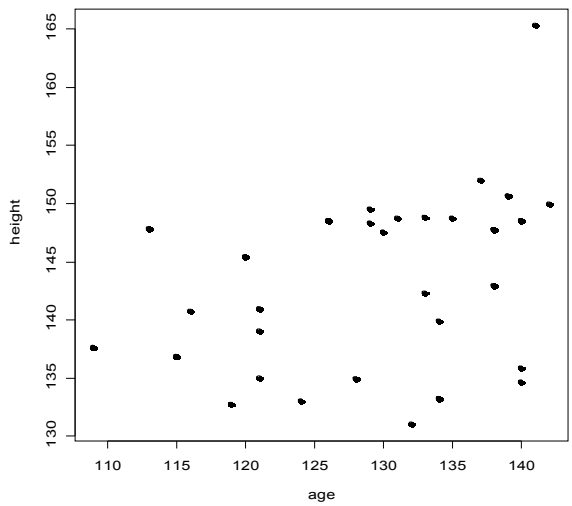
```
# tạo ra dãy số id
> id <- c(1:18, 1:14)
# group 1=urban 2=rural và cần phải xác định group là một
"factor"

> group <- c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
             2,2,2,2,2,2,2,2,2,2,2,2,2,2)
> group <- as.factor(group)

> # nhập dữ liệu
> age <- c(109,113,115,116,119,120,121,124,126,129,130,
           133,134,135, 137,139,141,142, 121,121,128,
           129,131,132,133,134,138,138,138,140,140,140)

> height <- c(137.6,147.8,136.8,140.7,132.7,145.4,135.0,
              133.0,148.5, 148.3,147.5,148.8,133.2,
              148.7,152.0,150.6,165.3,149.9,
              139.0,140.9,134.9,149.5,148.7,131.0,142.3,
              139.9,142.9,147.7,147.7,134.6,135.8,148.5)

> # tạo một data frame
> data <- data.frame(id, group, age, height)
> attach(data)
```

Chúng ta thử xem qua vài chỉ số thống kê mô tả bằng cách ước tính độ tuổi và chiều cao trung bình cho từng nhóm học sinh:

```
> tapply(age, group, mean)
    1    2
126.8333 133.0714

> tapply(height, group, mean)
    1    2
144.5444 141.6714
```

Kết quả trên cho thấy nhóm học sinh thành thị có độ tuổi thấp hơn học sinh nông thôn khoảng 6.3 tháng (126.8 – 133.1). Tuy nhiên, chiều cao của học sinh thành thị cao hơn học sinh nông thôn khoảng 2.8 cm (144.5 – 141.7). Bạn đọc có thể dùng kiểm định t để thấy rằng sự khác biệt về độ tuổi giữa hai nhóm có ý nghĩa thống kê (p = 0.045).

Ngoài ra, biểu đồ sau đây còn cho thấy có một mối liên hệ tương quan giữa tuổi và chiều cao:



Biểu đồ 11.5. Chiều cao (cm) và độ tuổi (tháng tuổi) của hai nhóm học sinh thành thị và nông thôn.

Vì hai nhóm khác nhau về độ tuổi, và tuổi có liên hệ với chiều cao, cho nên chúng ta không thể phát biểu hay so sánh chiều cao giữa 2 nhóm học sinh mà không điều chỉnh cho độ tuổi. Để điều chỉnh độ tuổi, chúng ta sử dụng phương pháp phân tích hiệp biến.

11.5.1 Mô hình phân tích hiệp biến

Gọi *y* là chiều cao, *x* là độ tuổi, và *g* là nhóm. Mô hình căn bản của ANCOVA giả định rằng mối liên hệ giữa *y* và *x* là một đường thẳng, và độ dốc (gradient hay slope) của hai nhóm trong mối liên hệ này không khác nhau. Nói cách khác, viết theo kí hiệu của hồi qui tuyến tính, chúng ta có:

$$y_1 = \alpha_1 + \beta x + e_1 \qquad \text{in group 1}$$

$$y_2 = \alpha_2 + \beta x + e_2 \qquad \text{in group 2.} \qquad [5]$$

Trong đó:

α_1 : là giá trị trung bình của *y* khi *x*= 0 của nhóm 1;

α_2 : là giá trị trung bình của *y* khi *x*= 0 của nhóm 2;

β: độ dốc của mối liên hệ giữa *y* và *x*;

e_1 *và* e_2: biến số ngẫu nhiên với trung bình 0 và phương sai σ^2.

Gọi $\bar{x}$ là số trung bình của độ tuổi cho cả 2 nhóm, $\bar{x}_1$ và $\bar{x}_2$ là tuổi trung bình của nhóm 1 và nhóm 2. Như nói trên, nếu $\bar{x}_1 \neq \bar{x}_2$, thì so sánh chiều cao trung bình của nhóm 1 và 2 ($\bar{y}_1$ và $\bar{y}_2$) sẽ thiếu khách quan, vì

$$\bar{y}_1 = \alpha_1 + \beta\bar{x}_1 + e_1$$
$$\bar{y}_2 = \alpha_2 + \beta\bar{x}_2 + e_2$$

và mức độ khác biệt giữa hai nhóm bây giờ tùy thuộc vào hệ số β:

$$\bar{y}_1 - \bar{y}_2 = \alpha_1 - \alpha_2 + \beta\left(\bar{x}_1 - \bar{x}_2\right)$$

Chú ý rằng trong mô hình [5], chúng ta có thể diễn dịch $\alpha_1 - \alpha_2$ là độ khác biệt chiều cao trung bình giữa hai nhóm *nếu* cả hai nhóm có cùng tuổi trung bình. Mức khác biệt này thể hiện ảnh hưởng của hai nhóm nếu không có một yếu tố nào liên hệ đến *y*. Để ước tính $\alpha_1 - \alpha_2$, chúng ta không thể đơn giản trừ hai số trung bình $\bar{y}_1$ - $\bar{y}_2$, nhưng phải điều chỉnh cho *x*. Gọi *x** là một giá trị chung cho cả hai nhóm, chúng ta có thể ước tính giá trị điều chỉnh *y* cho nhóm 1 (kí hiệu $\bar{y}_{1a}$) như sau:

$$\bar{y}_{1a} = \bar{y}_1 - \beta\left(\bar{x}_1 - x^*\right)$$

$\bar{y}_{1a}$ có thể xem là một ước số cho chiều cao trung bình của nhóm 1 (thành thị) cho giá trị *x* là *x** . Tương tự:

$$\bar{y}_{2a} = \bar{y}_2 - \beta\left(\bar{x}_2 - x^*\right)$$

là số cho chiều cao trung bình của nhóm 1 (nông thôn) với cùng giá trị *x**. Từ đây, chúng ta có thể ước tính ảnh hưởng của thành thị và nông thôn bằng công thức sau đây:

Do đó, vấn đề là $\bar{y}_{1a} - \bar{y}_{2a} = \bar{y}_2 - \bar{y}_1 - \beta\left(\bar{x}_1 - \bar{x}_2\right)$
chúng ta phải ước tính β. Có thể chứng minh rằng ước số β từ phương pháp bình phương nhỏ nhất cũng là ước tính khách quan cho $\alpha_1 - \alpha_2$. Khi viết bằng mô hình tuyến tính, mô hình hiệp biến có thể mô tả như sau:

$$y = \alpha + \beta x + \gamma g + \delta\left(xg\right) + e \qquad [6]$$

Nói cách khác, mô hình trên phát biểu rằng chiều cao của một học sinh bị ảnh hưởng bởi 3 yếu tố: độ tuổi (β), thành thị hay nông thôn (γ), và tương tác giữa hai yếu tố đó (δ). Nếu δ = 0 (tức ảnh hưởng tương tác không có ý nghĩa thống kê), mô hình trên giảm xuống thành:

$$y = \alpha + \beta x + \gamma g + e \qquad [7]$$

Nếu γ = 0 (tức ảnh hưởng của thành thị không có ý nghĩa thống kê), mô hình trên giảm xuống thành:

$$y = \alpha + \beta x + e \qquad [8]$$

11.5.2 Phân tích bằng R

Các thảo luận vừa trình bày trên xem ra khá phức tạp, nhưng trong thực tế, với R, cách ước tính rất đơn giản bằng hàm `lm`. Chúng ta sẽ phân tích ba mô hình [6], [7] và [8]:

```
> # model 6
> model6 <- lm(height ~ group + age + group:age)

> # model 7
> model7 <- lm(height ~ group + age)

> # model 8
> model8 <- lm(height ~ age)
```

Chúng ta cũng có thể so sánh cả ba mô hình cùng một lúc bằng lệnh `anova` như sau:

```
> anova(model6, model7, model8)
Analysis of Variance Table

Model 1: height ~ group + age + group:age
Model 2: height ~ group + age
```

```
Model 3: height ~ age

  Res.Df   RSS   Df Sum of Sq  F      Pr(>F)
1   28   1270.44
2   29   1338.02 -1  -67.57   1.4893  0.23251
3   30   1545.95 -1  -207.93  4.5827  0.04114 *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Chú ý "`model 1`" chính là mô hình [6], "`model 2`" là mô hình [7], và "`model 3`" là mô hình [8]. `RSS` là `residual sum of squares`, tức tổng bình phương phần dư cho mỗi mô hình. Kết quả phân tích trên cho thấy:

- Toàn bộ mẫu có 18+14=32 học sinh, mô hình [6] có 4 thông số (α, β, γ và δ), cho nên mô hình này có 32-4 = 28 bậc tự do. Tổng bình phương của mô hình là 1270.44.

- Mô hình [7] có 3 thông số (tức còn 29 bậc tự do), cho nên tổng bình phương phần dư cao hơn mô hình [7]. Tuy nhiên, đứng trên phương diện xác suất thì trung bình bình phương phần dư của mô hình này 1338.02 / 29 = 46.13, không khác mấy so với mô hình [6] (trung bình bình phương là: 1270.44 / 28 = 45.36), vì trị số p=0.2325, tức không có ý nghĩa thống kê. Nói cách khác, bỏ hệ số tương tác δ không làm thay đổi khả năng tiên đoán của mô hình một cách đáng kể.

- Mô hình [8] chỉ có 2 thông số (và do đó có 30 bậc tự do), với tổng bình phương là 1545.95. Trung bình bình phương phần dư của mô hình này là 51.53 (1545.95 / 30), tức cao hơn hai mô hình [6] một cách đáng kể, vì trị số p = 0.0411.

Qua phân tích trên, chúng ta thấy mô hình [7] là tối ưu hơn cả, vì chỉ cần 3 thông số mà có thể "giải thích" được dữ liệu một cách đầy đủ. Bây giờ chúng ta sẽ chú tâm vào phân tích kết quả của mô hình này.

```
> summary(model7)

Call:
lm(formula = height ~ group + age)

Residuals:
  Min    1Q Median    3Q   Max
-14.324 -3.285  0.879  3.956 14.866

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 91.8171  17.9294  5.121  1.81e-05 ***
group2      -5.4663   2.5749 -2.123  0.04242 *
age          0.4157   0.1408  2.953  0.00619 **
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.793 on 29 degrees of freedom
Multiple R-Squared: 0.2588,   Adjusted R-squared: 0.2077
F-statistic: 5.063 on 2 and 29 DF, p-value: 0.01300
```

Qua phần ước tính thông số trình bày trên đây, chúng ta thấy tính trung bình chiều cao học sinh tăng khoảng 0.41 cm cho mỗi tháng tuổi. Chú ý trong kết quả trên, phần "`group2`" có nghĩa là hệ số hồi qui (regression coefficient) cho nhóm 2 (tức là nông thôn), vì R phải đặt hệ số cho nhóm 1 bằng 0 để tiện việc tính toán. Vì thế, chúng ta có hai phương trình (hay hai đường biểu diễn) cho hai nhóm học sinh như sau:

Đối với học sinh thành thị:

```
Height = 91.817 + 0.4157(age)
```

Và đối với học sinh nông thôn:

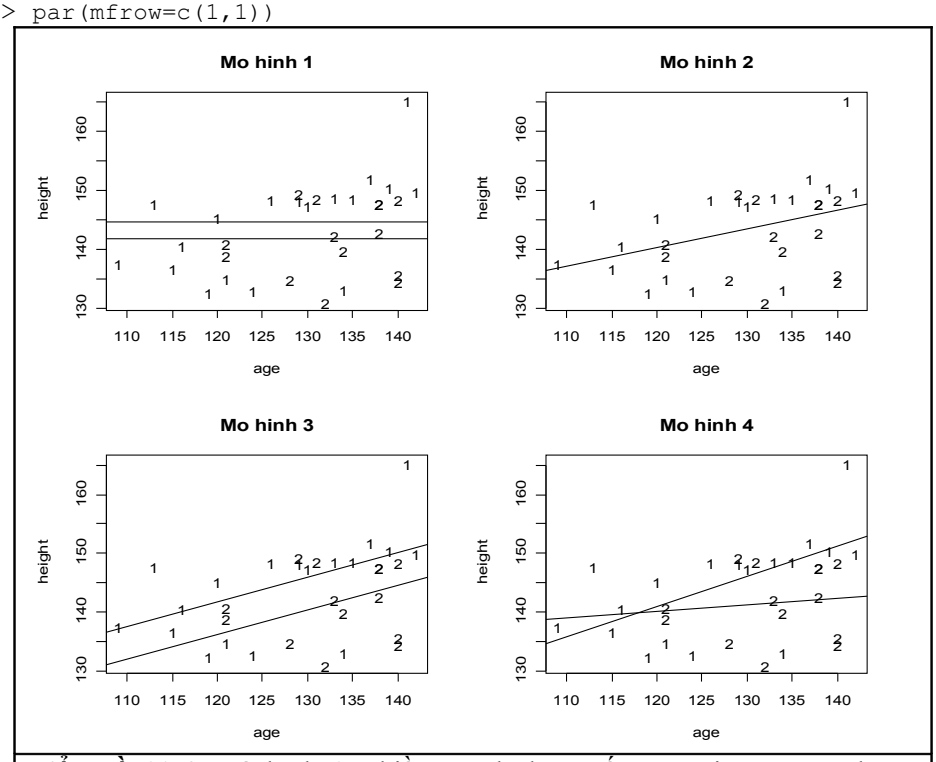
```
Height = 91.817 - 5.4663(rural) + 0.4157(age)
```

Nói cách khác, sau khi điều chỉnh cho độ tuổi, nhóm học sinh nông thôn (rural) có chiều cao thấp hơn nhóm thành thị khoảng 5.5 cm và mức độ khác biệt này có ý nghĩa thống kê vì trị số p = 0.0424. (Chú ý là trước khi điều chỉnh cho độ tuổi, mức độ khác biệt là 2.8 cm).

Các biểu đồ sau đây sẽ minh họa cho các mô hình trên:

```
> par(mfrow=c(2,2))

> plot(age, height, pch=as.character(group),
       main="Mo hinh 1")
> abline(144.54, 0) #mean value for urban
> abline(141.67, 0) #mean value for rural

> plot(age, height, pch=as.character(group),
       main="Mo hinh 2")
> abline(102.63, 0.3138) #single line for dependence on age

> plot(age, height, pch=as.character(group),
       main="Mo hinh 3")
> abline(91.8, 0.416) #line for males
```

```
> abline(91.8-5.46,0.416) #line for females parallel

> plot(age, height, pch=as.character(group),
      main="Mo hinh 4")
> abline(79.7, 0.511) #line for males
> abline(79.7+47.08, 0.511-0.399) #line for females parallel
> par(mfrow=c(1,1))
```



Biểu đồ 11.6. Mô hình 1: chiều cao là hàm số của nơi trú ngụ, nhưng không có liên hệ với độ tuổi; Mô hình 2 giả thiết rằng chiều cao phụ thuộc vào độ tuổi, nhưng không có khác nhau giữa hai nhóm thành thị và nông thôn; Mô hình 3 giả thiết rằng mối liên hệ giữa chiều cao và tuổi của nhóm thành thị tương đương với với nhóm nông thôn (hai đường song song), nhưng học sinh thành thị có chiều cao cao hơn nông thôn; và mô hình 4 giả thiết rằng mức độ khác biệt về chiều cao giữa hai nhóm tùy thuộc vào độ tuổi (tức có tương tác giữa độ tuổi và nơi trú ngụ): ở tuổi <120 tháng, chiều cao hai nhóm không khác nhau mấy, nhưng khi tuổi >120 tháng tuổi thì nhóm học sinh thành thị có chiều cao cao hơn nhóm nông thôn. Phân tích trên cho thấy mô hình 3 là tốt nhất.

11.6 Phân tích phương sai cho thí nghiệm giai thừa (factorial experiment)

Ví dụ 4. Để khảo sát ảnh hưởng của 4 loại thuốc trừ sâu (1, 2, 3 và 4) và ba loại giống (B1, B2 và B3) đến sản lượng của cam, các nhà nghiên cứu tiến hành một thí nghiệm loại giai thừa. Trong thí nghiệm này, mỗi giống cam có 4 cây cam được chọn một cách ngẫu nhiên, và 4 loại thuốc trừ sâu áp dụng (cũng ngẫu nhiên) cho mỗi cây cam. Kết quả nghiên cứu (sản lượng cam) cho từng giống và thuốc trừ sâu như sau:

Bảng 11.5. Sản lượng cam cho 3 loại giống và 4 loại thuốc trừ sâu

Giống cam (variety)	Thuốc trừ sâu (pesticide)				Tổng số
	1	2	3	4	
B1	29	50	43	53	**175**
B2	41	58	42	73	**214**
B3	66	85	63	85	**305**
Tổng số	**136**	**193**	**154**	**211**	**694**

Mô hình phân tích thí nghiệm giai thừa cũng không khác gì so với phân tích phương sai hai chiều như trình bày trong phần trên. Cụ thể hơn, mô hình mà chúng ta xem xét là:

$$\texttt{product} = \alpha + \beta(\texttt{variety}) + \gamma(\texttt{pesticide}) + \varepsilon$$

Trong đó, α là hằng số biểu hiện trung bình toàn mẫu, α là hệ số ảnh hưởng của ba giống cam, và γ là hệ số ảnh hưởng của 4 loại thuốc trừ sâu, và ε là phần dư (residual) của mô hình.

Chúng ta có thể sử dụng hàm `aov` của R để ước tính các thông số trên như sau:

```
# trước hết chúng ta nhập số liệu
> variety <- c(1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3)
> pesticide <- c(1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4)
> product <- c(29,50,43,53,41,58,42,73,66,85,69,85)

# định nghĩa variety và pesticide là hai yếu tố (factors)
> variety <- as.factor(variety)
> pesticide <- as.factor(pesticide)

# cho vào một data frame tên là data
> data <- data.frame(variety, pesticide, product)

# phân tích phương sai bằng aov và cho vào object analysis
> analysis <- aov(product ~ variety + pesticide)
> anova(analysis)
```

```
Analysis of Variance Table

Response: product
          Df Sum Sq   Mean Sq   F value  Pr(>F)
variety    2  2225.17  1112.58  44.063  0.000259 ***
pesticide  3  1191.00   397.00  15.723  0.003008 **
Residuals  6   151.50    25.25
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

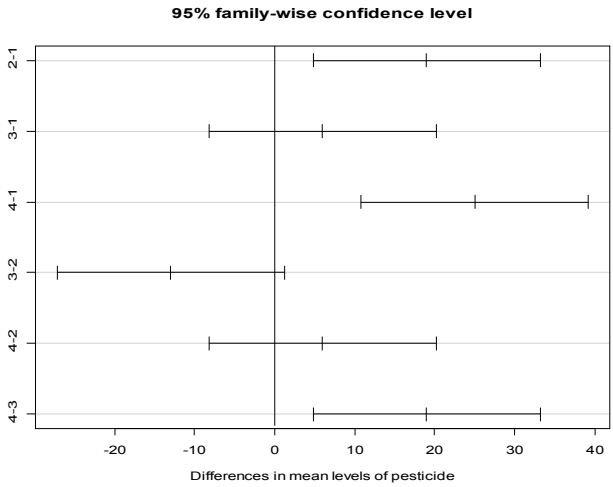
Kết quả trên cho thấy cả hai yếu tố giống cây (variety) và thuốc trừ sâu (pesticide) đều có ảnh hưởng đến sản lượng cam, vì trị số $p < 0.05$. Để so sánh cụ thể cho từng hai nhóm, chúng ta sử dụng hàm `TukeyHSD` như sau:

```
> TukeyHSD(analysis)

Tukey multiple comparisons of means
  95% family-wise confidence level

Fit: aov(formula = product ~ variety + pesticide)

$variety
    diff    lwr      upr      p adj
2-1  9.75 -1.152093  20.65209  0.0749103
3-1 32.50 21.597907  43.40209  0.0002363
3-2 22.75 11.847907  33.65209  0.0016627

$pesticide
    diff    lwr     upr        p adj
2-1  19   4.797136  33.202864  0.0140509
3-1  6   -8.202864  20.202864  0.5106152
4-1  25  10.797136  39.202864  0.0036109
3-2 -13 -27.202864   1.202864  0.0704233
4-2  6   -8.202864  20.202864  0.5106152
4-3  19   4.797136  33.202864  0.0140509
```

Kết quả phân tích giữa các loại giống cho thấy giống B3 có sản lượng cao hơn giống B1 khoảng 32 đơn vị với khoảng tin cậy 95% từ 21 đến 43 ($p = 0.0002$). Giống cam B3 cũng tốt hơn giống B2, với độ khác biệt trung bình khoảng 22 đơn vị ($p = 0.0017$). Nhưng không có khác biệt đáng kể giữa giống B2 và B1.

So sánh giữa các loại thuốc trừ sâu, kết quả trên cho chúng ta biết các thuốc trừ sâu 4 có hiệu quả cao hơn thuốc 1 và 3. Ngoài ra, thuốc 2 cũng có hiệu quả cao hơn thuốc 1. Còn các so sánh khác không có ý nghĩa thống kê. Biểu đồ Tukey sau đây minh họa cho kết luận trên.

```
> plot(TukeyHSD(analysis), ordered=TRUE)
```



11.7 Phân tích phương sai cho thí nghiệm hình vuông Latin (Latin square experiment)

Ví dụ 5. Để so sánh hiệu quả của 2 loại phân bón (A và B) cùng 2 phương pháp canh tác (a và b), các nhà nghiên cứu tiến hành một thí nghiệm hình vuông Latin. Theo đó, có 4 nhóm can thiệp tổng hợp từ hai loại phân bón và phương pháp canh tác: Aa, Ab, Ba, và Bb (sẽ cho mã số, lần lược, là 1=Aa, 2=Ab, 3=Ba, 4=Bb). Bốn phương pháp (treatment) đó được áp dụng trong 4 mẫu ruộng (sample = 1, 2, 3, 4) và 4 loại cây trồng (variety = 1, 2, 3, 4). Tổng cộng, thí nghiệm có 4x4 = 16 mẫu. Tiêu chí để đánh giá là sản lượng, và kết quả sản lượng được tóm tắt trong bảng sau đây:

Bảng 11.6. Sản lượng cho 2 loại phân bón và 2 phương pháp canh tác:

Mẫu ruộng (sample)	Giống (variety)			
	1	2	3	4
1	175 Aa	143 Ba	128 Bb	166 Ab
2	170 Ab	178 Aa	140 Ba	131 Bb
3	135 Bb	173 Ab	169 Aa	141 Ba

4	145 Ba	136 Bb	165 Ab	173 Aa

Câu hỏi đặt ra là các phương pháp canh tác và phân bón có ảnh hưởng đến sản lượng hay không. Để trả lời câu hỏi đó, chúng ta phải xem xét đến các nguồn làm cho sản lượng thay đổi hay biến thiên. Nhìn qua thí nghiệm và bảng số liệu trên, rất dễ dàng hình dung ra 3 nguồn biến thiên chính:

- Nguồn thứ nhất là khác biệt giữa các phương pháp canh tác và phân bón;
- Nguồn thứ hai là khác biệt giữa các loại giống cây;
- Nguồn thứ ba là khác biệt giữa các mẫu ruộng;

Và phần còn lại là khác biệt trong mỗi mẫu ruộng và loại giống. Để có một cái nhìn chung về số liệu, chúng ta hãy tính trung bình cho từng nhóm qua bảng số sau đây:

Trung bình cho từng loại giống	Trung bình cho từng mẫu	Trung bình cho từng phương pháp
1: 156.25 **2**: 157.50 **3**: 150.50 **4**: 152.75 Tổng trung bình: 154.25	**1**: 153.00 **2**: 154.75 **3**: 154.50 **4**: 154.75 Tổng trung bình: 154.25	**1**: 173.75 **2**: 168.50 **3**: 142.25 **4**: 132.50 Tổng trung bình: 154.25

Bảng tóm lược trên cho phép chúng ta tính tổng bình phương cho từng nguồn biến thiên. Khởi đầu là tổng bình phương cho toàn bộ thí nghiệm (sẽ tạm gọi là SStotal):

- Tổng bình phương chung cho toàn thí nghiệm:

$$SStotal = (175 - 154.25)^2 + (143 - 154.25)^2 + \ldots (165 - 154.25)^2 + (173 - 154.25)^2 = 4941$$

- Tổng bình phương do khác biệt giữa các loại giống (SSvariety). Chú ý là vì trung bình mỗi giống được tính từ 4 số, cho nên chúng ta phải nhân cho 4 khi tính tổng bình phương:

$$SSvariety = 4(156.25 - 154.25)^2 + 4(157.50 - 154.25)^2 + 4(150.50 - 154.25)^2 + 4(152.75 - 154.25)^2 = 123.5$$

Vì có 4 loại giống và một thông số, cho nên bậc tự do là 4-1=3. Theo đó, trung bình bình phương (mean square) là:
123.5 / 3 = 41.2.

- Tổng bình phương do khác biệt giữa giống (SSsample). Chú ý là vì trung bình mỗi mẫu được tính từ 4 số, cho nên khi tính tổng bình phương, cần phải nhân cho 4:

  $$SSsample= 4(153.00 - 154.25)^2 + 4(154.75 - 154.25)^2 + 4(154.50 - 154.25)^2 + 4(154.75 - 154.25)^2 = 8.5$$

  Vì có 4 mẫu và một thông số, cho nên bậc tự do là 4-1=3, và theo đó trung bình bình phương là: 8.5 / 3 = 2.8.

- Tổng bình phương do khác biệt giữa các phương pháp (SSmethod). Chú ý là vì trung bình mỗi phương pháp được tính từ 4 số, cho nên khi tính tổng bình phương, cần phải nhân cho 4:

  $$SSsample= 4(173.75 - 154.25)^2 + 4(168.50 - 154.25)^2 + 4(142.25 - 154.25)^2 + 4(132.50 - 154.25)^2 = 4801.50$$

  Vì có 4 phương pháp và một thông số, cho nên bậc tự do là 4-1=3, và theo đó trung bình bình phương là: 4801.5 / 3 = 1600.5.

- Tổng bình phương phần dư (residual sum of squares):

  $$\begin{aligned} SSresidual &= SStotal - SSmethod - SSsample - SSvariety \\ &= 4941.0 - 4801.5 - 8.5 - 123.5 \\ &= 7.5 \end{aligned}$$

Những ước tính trên đây có thể trình bày trong một bảng phân tích phương sai như sau:

Nguồn biến thiên	Bậc tự do (degrees of freedom)	Tổng bình phương (Sum of squares)	Trung bình bình phương (Mean square)	Kiểm định F
Giữa 4 mẫu ruộng	3	8.5	2.8	2.3
Giữa 4 loại giống	3	123.5	41.2	32.9
Giữa 4 phương pháp	3	4801.5	1600.5	1280.4
Phần dư (residual)	6	7.5		

Tổng số	16	4941.0		

Qua phân tích thủ công và đơn giản trên, chúng ta thấy phương pháp canh tác và loại giống có ảnh hưởng lớn đến sản lượng. Để tính toán chính xác trị số p, chúng ta có thể sử dụng R để tiến hành phân tích phương sai cho thí nghiệm hình vuông Latin.

Vấn đề tổ chức số liệu sao cho thích hợp để R có thể tính toán là rất quan trọng. Nói một cách ngắn gọn, mỗi số liệu phải là một số đặc thù (unique). Trong thí nghiệm trên, chúng ta có 4 loại giống, 4 mẫu, cho nên tổng số là 16 số liệu. Và, 16 số liệu này phải được định nghĩa cho từng loại giống, từng mẫu, và quan trọng hơn là cho từng phương pháp canh tác. Chẳng hạn như, trong ví dụ bảng số liệu 10.6 trên, 175 là sản lượng của phương pháp canh tác 1 (tức Aa), loại giống 1, và mẫu 1; nhưng 173 (số ở góc mặc cuối bảng) là sản lượng của phương pháp canh tác 1, nhưng từ loại giống 4, và mẫu 4; v.v...

- Trước hết, chúng ta nhập số liệu sản lượng, và gọi đó là `y`:

```
> y <- c(175, 143, 128, 166,
         170, 178, 140, 131,
         135, 173, 169, 141,
         145, 136, 165, 173)
```

- Kế đến, gọi `variety` là giống gồm 4 bậc (1,2,3,4) cho từng số liệu trong `y` (và cũng định nghĩa rằng `variety` là một factor, tức biến thứ bậc):

```
> variety <- c(1,2,3,4,
               1,2,3,4,
               1,2,3,4,
               1,2,3,4,)
> variety <- as.factor(variety)
```

- Gọi `sample` là mẫu gồm 4 bậc (1,2,3,4) cho từng số liệu trong `y` (và cũng định nghĩa rằng `sample` là một factor, tức biến thứ bậc):

```
> sample  <- c(1,1,1,1,
               2,2,2,2,
               3,3,3,3,
               4,4,4,4)
> sample <- as.factor(sample)
```

- Nhập số liệu cho phương pháp, `method`, cũng gồm 4 bậc (1,2,3,4) cho từng số liệu trong `y` (và cũng định nghĩa rằng `method` là một factor, tức biến thứ bậc):

```
> method <- c(1, 3, 4, 2,
              2, 1, 3, 4,
```

```
              4, 2, 1, 3,
              3, 4, 2, 1)
> method <- as.factor(method)
```

- Tổng hợp tất cả các số liệu trên vào một data frame và gọi là `data`:

```
> data <- data.frame(sample, variety, method, y)
```

- In `data` để kiểm tra xem số liệu có đúng và thích hợp hay chưa:

```
> data
      sample variety method y
1        1      1     1   175
2        1      2     3   143
3        1      3     4   128
4        1      4     2   166
5        2      1     2   170
6        2      2     1   178
7        2      3     3   140
8        2      4     4   131
9        3      1     4   135
10       3      2     2   173
11       3      3     1   169
12       3      4     3   141
13       4      1     3   145
14       4      2     4   136
15       4      3     2   165
16       4      4     1   173
```

Bây giờ chúng ta đã sẵn sàng dùng hàm `lm` hay `aov` để phân tích số liệu. Ở đây chúng ta sẽ sử dụng hàm `aov` để tính các nguồn biến thiên trên (kết quả tính toán sẽ chứa trong đối tượng `latin`):

```
> latin <- aov(y ~ sample + variety + method)
> summary(latin)
          Df  Sum Sq  Mean Sq  F value  Pr(>F)
sample     3    8.5     2.8    2.2667 0.1810039
variety    3  123.5    41.2   32.9333 0.0004016 ***
method     3 4801.5  1600.5 1280.4000 8.293e-09 ***
Residuals  6    7.5     1.3
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tất cả các kết quả này (dĩ nhiên) là những kết quả mà chúng ta đã tóm tắt trong bảng phân tích phương sai một cách "thủ công" trên đây. Tuy nhiên, ở đây R cung cấp cho chúng ta trị số p (trong `Pr > F`) để có thể suy luận thống kê. Và, qua trị số p, chúng ta có thể phát biểu rằng mẫu ruộng không có ảnh

hưởng đến sản lượng, nhưng loại giống và phương pháp canh tác thì có ảnh hưởng đến sản lượng.

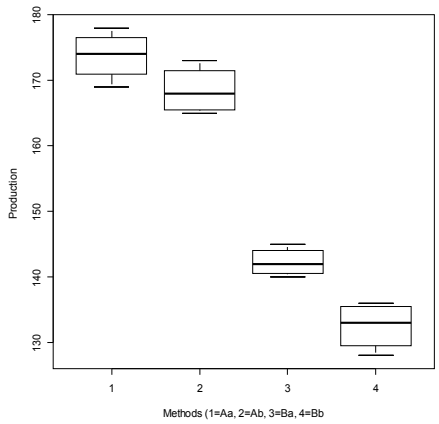
Để biết mức độ khác biệt giữa các phương pháp canh tác và giữa các loại giống, chúng ta dùng hàm `TukeyHSD` như sau:

```
> TukeyHSD(latin)
$variety
     diff    lwr     upr        p adj
2-1  1.25 -1.4867231 3.9867231  0.4528549
3-1 -5.75 -8.4867231 -3.0132769 0.0014152
4-1 -3.50 -6.2367231 -0.7632769 0.0173206
3-2 -7.00 -9.7367231 -4.2632769 0.0004803
4-2 -4.75 -7.4867231 -2.0132769 0.0038827
4-3  2.25 -0.4867231  4.9867231 0.1034761

$method
     diff    lwr      upr        p adj
2-1  -5.25  -7.986723  -2.513277 0.0023016
3-1 -31.50 -34.236723 -28.763277 0.0000001
4-1 -41.25 -43.986723 -38.513277 0.0000000
3-2 -26.25 -28.986723 -23.513277 0.0000004
4-2 -36.00 -38.736723 -33.263277 0.0000000
4-3  -9.75 -12.486723  -7.013277 0.0000730
```

So sánh giữa các loại giống cho thấy có sự khác biệt giữa giống 3 và 1, 4 và 1, 3 và 2, 4 và 2.

Tất cả các so sánh giữa các phương pháp canh tác đều có ý nghĩa thống kê. Nhưng loại nào có sản lượng cao nhất? Để trả lời câu hỏi này, chúng ta sẽ sử dụng biểu đồ hộp:

```
> boxplot(y ~ method, xlab="Methods (1=Aa, 2=Ab, 3=Ba,
          4=Bb", ylab="Production")
```



Biểu đồ so sánh sản lượng của bốn phương pháp canh tác.

11.8 Phân tích phương sai cho thí nghiệm giao chéo (cross-over experiment)

Ví dụ 6. Để thử nghiệm hiệu ứng của một thuốc mới đối với chứng ra mồ hôi (thuốc này được bào chế để chữa trị bệnh tim, nhưng ra mồ hồi là một ảnh hưởng phụ), các nhà nghiên cứu tiến hành một nghiên cứu trên 16 bệnh nhân. Số bệnh nhân này được chia thành 2 nhóm (tạm gọi là nhóm AB và BA) một cách ngẫu nhiên. Mỗi nhóm gồm 8 bệnh nhân. Bệnh nhân được theo dõi hai lần: tháng thứ nhất và tháng thứ 2. Đối với bệnh nhân nhóm AB, tháng thứ nhất họ được điều trị bằng thuốc, tháng thứ hai họ được cho sử dụng giả dược (placebo). Ngược lại, với bệnh nhân nhóm BA, tháng thứ nhất sử dụng giả dược, và tháng thứ hai được điều trị bằng thuốc. Tiêu chí để đánh giá là thời gian ra mồ hôi trên trán (tính từ lúc uống thuốc đến khi ra mồ hôi) sau khi sử dụng thuốc hay giả dược. Kết quả nghiên cứu được trình bày trong bảng số liệu sau đây:

Bảng 11.7. Kết quả nghiên cứu hiệu ứng ra mồ hôi của thuốc điều trị bệnh tim

Nhóm	Mã số bệnh nhân số (id)	Thời gian (phút) ra mồ hôi trên trán	
		Tháng 1	Tháng 2
AB		**A**	**Placebo**
	1	6	4
	3	8	7
	5	12	6
	6	7	8
	9	9	10
	10	6	4
	13	11	6

	15	8	8
BA		**Placebo**	**A**
	2	5	7
	4	9	6
	7	7	11
	8	4	7
	11	9	8
	12	5	4
	14	8	9
	16	9	13

Câu hỏi chính là có sự khác biệt về thời gian ra mồ hôi giữa hai nhóm điều trị bằng thuốc và giả dược hay không?

Để trả lời câu hỏi trên, chúng ta cần tiến hành phân tích phương sai. Nhưng vì cách thiết kế nghiên cứu khá đặc biệt (hai nhóm bệnh nhân với cách sắp xếp can thiệp theo hai thứ tự khác nhau), nên các phương pháp phân tích trên không thể áp dụng được. Có một phương pháp thông dụng là phân tích phương sai trong từng nhóm, rồi sau đó so sánh giữa hai nhóm. Một trong những vấn đề chúng ta cần phải lưu ý là khả năng hiệu ứng kéo dài (còn gọi là carry-over effect), tức là trong nhóm AB, hiệu quả của tháng thứ 2 có thể chịu ảnh hưởng kéo dài từ tháng thứ nhất khi bệnh được được điều trị bằng thuốc thật. Trước hết, chúng ta thử tóm lược dữ liệu bằng bảng sau đây:

Bảng 11.8. Tóm lược kết quả thí nghiệm hiệu ứng ra mồ hôi của thuốc điều trị bệnh tim

Nhóm	Mã số bệnh nhân số (id)	Thời gian (phút) ra mồ hôi trên trán		Trung bình cho từng bệnh nhân
		Tháng 1	Tháng 2	
AB		**A**	**Placebo**	
	1	6	4	**5.0**
	3	8	7	**7.5**
	5	12	6	**9.0**
	6	7	8	**7.5**
	9	9	10	**9.5**
	10	6	4	**5.0**
	13	11	6	**8.5**
	15	8	8	**8.0**
	Trung bình	**8.375**	**6.625**	**7.50**
BA		**Placebo**	**A**	
	2	5	7	**6.0**
	4	9	6	**7.5**
	7	7	11	**9.0**

	8	4	7	**5.5**
	11	9	8	**8.5**
	12	5	4	**4.5**
	14	8	9	**8.5**
	16	9	13	**11.0**
	Trung bình	**7.000**	**8.125**	**7.5625**
Trung bình cho 2 nhóm		**7.6875**	**7.3750**	**7.5312**

Trung bình cho nhóm A = (8.375 + 8.125) / 2 = 8.25
Trung bình cho nhóm P (giả dược) = (6.625 + 7.000) / 2 = 6.8125

Qua bảng tóm lược trên, chúng ta có thể tính toán một số tổng bình phương:

- Tổng bình phương do khác biệt giữa hai nhóm điều trị bằng thuốc và giả dược:

  $SSTreat = 16(8.25 - 7.5312)^2 + 16(8.8125 - 7.5312)^2 = 16.53$

- Tổng bình phương do khác biệt giữa tháng 1 và tháng 2:

  $SSPeriod = 16(7.6875 - 7.5312)^2 + 16(7.3750 - 7.5312)^2 = 0.781$

- Tổng bình phương do khác biệt giữa hai nhóm AB và BA (thứ tự):

  $SSseq = 16(7.50 - 7.5312)^2 + 16(7.5625 - 7.5312)^2 = 0.031$

- Tổng bình phương do khác biệt giữa các bệnh nhân trong cùng nhóm AB hay BA:

  $$SSw = (5.0 - 7.50)^2 + (7.5 - 7.50)^2 + (9.0 - 7.50)^2 + \ldots + (8.0 - 7.50)^2 + (6.0 - 7.5625)^2 + (7.5 - 7.5625)^2 + (9.0 - 7.5625)^2 + \ldots + (11.0 - 7.5625)^2 = 103.44$$

- Tổng bình phương cho toàn bộ mẫu:

  $$SStotal = (6 - 7.5312)^2 + (9 - 7.5312)^2 + \ldots + (13 - 7.5312)^2 + (9 - 7.5312)^2 = 167.97$$

- Tổng bình phương còn lại (tức phần dư):

  $SSres = 167.97 - 16.53 - 0.781 - 0.031 - 103.44 = 47.19$

Đến đây, chúng ta có thể lập bảng phân tích phương sai như sau:

Bảng 11.9. Kết quả phân tích phương sai số liệu trong bảng 11.7

Nguồn biến thiên	Bậc tự do (degrees of freedom)	Tổng bình phương (Sum of squares)	Trung bình bình phương (Mean square)	Kiểm định F
Giữa hai nhóm điều trị	1	16.53	16.53	4.90
Giữa hai tháng	1	0.781	0.781	0.23
Giữa AB và BA	1	0.031	0.031	0.004
Trong mỗi nhóm	14	103.44	7.39	
Phần dư (residual)	14	47.19	3.37	
Tổng số	31	167.97		

Qua phân tích trên, chúng ta thấy độ khác biệt giữa thuốc và giả dược lớn hơn là độ khác biệt giữa hai tháng hay hai nhóm AB và BA. Kiểm định F để thử nghiệm giả thiết thuốc và giả dược có hiệu quả như nhau là kiểm định F = 16.53 / 3.37 = 4.90 với bậc tự do 1 và 14. Dựa trên lí thuyết xác suất, trị số F với bậc tự do 1 và 14 là 4.60. Do đó, chúng ta có thể kết luận rằng thuốc này có hiệu ứng làm ra mồ hôi lâu hơn nhóm giả dược.

Tất cả các tính toán "thủ công" trên chỉ là minh họa cho cách phân tích phương sai trong thí nghiệm giao chéo. Trong thực tế, chúng ta có thể sử dụng R để tiến hành các tính toán đó như cách tính phương sai cho các thí nghiệm đơn giản. Vấn đề chính là tổ chức số liệu cho phân tích. R (cũng như nhiều phần mềm khác) yêu cầu người sử dụng phải nhập từng số liệu một, và **mỗi số liệu phải gắn liền với một bệnh nhân, một nhóm điều trị, một tháng (hay giai đoạn), và một nhóm thứ tự.** Đó là một yêu cầu rất quan trọng, vì nếu tổ chức số liệu không đúng, kết quả phân tích có thể sai.

Phần sau đây sẽ mô tả từng bước một:

```
# bước 1: nhập dữ liệu và đặt tên object là y

> y <- c(6,8,12,7,9,6,11,8,
         4,7,6,8,10,4,6,8,
         5,9,7,4,9,5,8,9
         7,6,11,7,8,4,9,13)

# bước 2: cứ mỗi số liệu trong bước 1, chỉ ra nhóm AB
hay BA (mã số 1 và 2)

> seq <- c(1,1,1,1,1,1,1,1,
```

```
           1,1,1,1,1,1,1,1,
           2,2,2,2,2,2,2,2,
           2,2,2,2,2,2,2,2)
> seq <- as.factor(seq)
```

 # bước 3: cứ mỗi số liệu trong bước 1, chỉ ra tháng 1 hay tháng 2

```
> period <- c(1,1,1,1,1,1,1,1,
              2,2,2,2,2,2,2,2,
              2,2,2,2,2,2,2,2,
              1,1,1,1,1,1,1,1)
> period <- as.factor(period)
```

 # bước 4: cứ mỗi số liệu trong bước 1, chỉ ra nhóm A hay placebo bằng mã số 1 và 2:

```
> treat <- c(1,1,1,1,1,1,1,1,
             2,2,2,2,2,2,2,2,
             1,1,1,1,1,1,1,1,
             2,2,2,2,2,2,2,2)
> treat <- as.factor(treat)
```

 # bước 5: cứ mỗi số liệu trong bước 1, chỉ ra mã số cho từng bệnh nhân

```
> id <- c(1,3,5,6,9,10,13,15,
          1,3,5,6,9,10,13,15,
          2,4,7,8,11,12,14,16,
          2,4,7,8,11,12,14,16)
> id <- as.factor(id)
```

 # bước 6: lập thành một data frame tên là data và in ra để kiểm tra một lần nữa.

```
> data <- data.frame(seq, period, treat, id, y)
> data
  seq period treat id    y
 1  1   1     1     1    6
 2  1   1     1     3    8
 3  1   1     1     5   12
 4  1   1     1     6    7
 5  1   1     1     9    9
 6  1   1     1     10   6
 7  1   1     1     13  11
 8  1   1     1     15   8
```

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

Bây giờ chúng ta đã sẵn sàng dùng hàm `lm` của R để phân tích số liệu. Chú ý rằng cách dùng hàm `lm` cho phân tích phương sai áp dụng cho thí nghiệm giao chéo hoàn toàn không khác gì với cách dùng cho các thí nghiệm khác. Khía cạnh khác biệt duy nhất là cách tổ chức dữ liệu cho phân tích như trình bày trên.

```
> xover <- lm(y ~ treat+seq+period)
> anova(xover)

Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value Pr(>F)
treat      1 16.531 16.531 4.9046 0.04388 *
seq        1  0.031  0.031 0.0093 0.92466
period     1  0.781  0.781 0.2318 0.63764
id        14 103.438 7.388 2.1921 0.07711 .
Residuals 14 47.187  3.371
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Kết quả phân tích trên đây dĩ nhiên giống với cách tính thủ công mà chúng ta đã tiến hành ở phần trên. Nói tóm lại, mức độ khác biệt giữa thuốc và giả dược có ý nghĩa thống kê, với trị số F là 0.044.

Chúng ta cũng có thể yêu cầu khoảng tin cậy 95% cho độ khác biệt giữa hai nhóm (bằng cách lệnh `TukeyHSD`) như sau (chú ý là với `TukeyHSD` chúng ta chỉ sử dụng hàm `aov` chứ không phải `lm`):

```
> TukeyHSD(aov(y ~ treat+seq+period+id))
 Tukey multiple comparisons of means
  95% family-wise confidence level

Fit: aov(formula = y ~ treat + seq + period + id)

$treat
    diff     lwr      upr         p adj
2-1 -1.4375 -2.829658 -0.04534186 0.0438783

$seq
    diff     lwr      upr         p adj
2-1 0.0625 -1.329658 1.454658 0.924656

$period
    diff     lwr      upr         p adj
2-1 -0.3125 -1.704658 1.079658 0.6376395
```

Chú ý kết quả:

```
$treat
    diff     lwr      upr         p adj
2-1 -1.4375 -2.829658 -0.04534186 0.0438783
```

cho biết tính trung bình thời gian ra mồ hôi của nhóm được điều trị cao hơn nhóm giả dược khoảng 1.44 phút, và khoảng tin cậy 95% là từ 0.05 phút đến 2.8 phút. Còn các kết quả so sánh giữa hai nhóm AB và BA (`seq`) hay giữa tháng 1 và tháng 2 (`period`) không có ý nghĩa thống kê.

11.9 Phân tích phương sai cho thí nghiệm tái đo lường (repeated measure experiment)

Ví dụ 7. Một nghiên cứu sơ khởi (pilot study) được tiến hành để đánh giá hiệu nghiệm của một vắc-xin mới chống bệnh thấp khớp. Nghiên cứu gồm 8 bệnh nhân, được chia thành 2 nhóm một cách ngẫu nhiên. Nhóm 1 gồm 4 bệnh nhân được điều trị bằng vắc-xin; nhóm 2 cũng gồm 4 bệnh nhân nhưng được nhận giả dược (placebo, hay đối chứng). Bệnh nhân được theo dõi trong 3

tháng, và cứ mỗi tháng, bệnh nhân được hỏi về tình trạng của bệnh ra sao. Tình trạng bệnh được "đo lường" bằng một chỉ số có giá trị từ 0 (không có hiệu nghiệm, bệnh vẫn như trước) đến 10 (có hiệu nghiệm tuyệt đối, hết bệnh). Kết quả nghiên cứu có thể tóm tắt trong bảng số liệu sau đây:

Bảng 11.10. Kết quả nghiên cứu vắc-xin chống đau thấp khớp

Nhóm	Mã số bệnh nhân (id)	Chỉ số bệnh qua từng tháng		
		Tháng 1	Tháng 2	Tháng 3
Vắc-xin				
	1	6	3	0
	2	7	3	1
	3	4	1	2
	4	8	4	3
Placebo				
	5	6	5	5
	6	9	4	6
	7	5	3	4
	8	6	2	3

Câu hỏi chính là có sự khác biệt nào giữa hai nhóm vắc-xin và giả dược hay không?

Để đơn giản hóa cách phân tích phương sai cho thí nghiệm tái đo lường, chúng ta sẽ tránh dùng kí hiệu toán, mà chỉ minh họa bằng vài phép tính "thủ công" để bạn đọc có thể theo dõi. Trước hết, chúng ta cần phải tóm lược số liệu bằng cách tính trung bình cho mỗi bệnh nhân, mỗi nhóm điều trị, và mỗi tháng như sau:

Bảng 11.11. Tóm lược số liệu nghiên cứu vắc-xin chống đau thấp khớp

Nhóm điều trị	id	Chỉ số bệnh qua từng tháng			Trung bình
		1	2	3	
Vắc-xin	1	6	3	0	**3.000**
	2	7	3	1	**3.667**
	3	4	1	2	**2.333**
	4	8	4	3	**5.000**
	Trung bình	**6.25**	**2.75**	**1.50**	**3.500**
	SD	**1.71**	**1.26**	**1.29**	
Placebo	5	6	5	5	**5.333**
	6	9	4	6	**6.333**
	7	5	3	4	**4.000**
	8	6	2	3	**3.667**
	Trung bình	**6.50**	**3.50**	**4.50**	**4.833**

	SD	**1.73**	**1.29**	**1.29**	
Trung bình cho hai nhóm		**6.375**	**3.125**	**3.000**	**4.167**

Qua bảng trên, chúng ta có thể thấy rằng có 5 nguồn làm cho kết quả thí nghiệm khác nhau:

(a) Giữa vắc-xin và giả dược (có lẽ là nguồn mà chúng ta cần biết!);
(b) Giữa 3 tháng theo dõi;
(c) Giữa ba tháng trong mỗi nhóm điều trị, mà giới thống kê thường đề cập đến là "interaction" (tương tác), và trong trường hợp này, tương tác giữa nhóm điều trị và thời gian;
(d) Giữa các bệnh nhân trong cùng một nhóm điều trị;
(e) Và sau cùng là phần dư, tức phần mà chúng ta không thể "giải thích" sau khi xem xét các nguồn (a) đến (d) trên.

- Trước hết là tổng bình phương giữa hai nhóm điều trị (vắc-xin và giả dược), tôi sẽ gọi là SS*treat*:

$$SS\textit{treat} = 12(3.500 - 4.167)^2 + 12(4.833 - 4.167)^2 = 10.667$$

- Kế đến là tổng bình phương giữa 3 tháng điều trị, gọi là SS*time*:

$$SS\textit{time} = 8(6.375 - 4.167)^2 + 8(3.125 - 4.167)^2 + 8(3.000 - 4.167)^2 = 58.583$$

- Nguồn thứ ba là tổng bình phương do tương tác giữa điều trị và thời gian, gọi là SS*int*

$$\begin{aligned} SS\textit{int} = \; & 4(6.25 - 4.167)^2 + \\ & 4(2.75 - 4.167)^2 + \\ & 4(1.50 - 4.167)^2 + \\ & 4(6.50 - 4.167)^2 + \\ & 4(3.50 - 4.167)^2 + \\ & 4(4.50 - 4.167)^2 - \\ & SS\text{vắcxin} - SS\text{time} \\ = \; & 77.833 - 10.667 - 58.583 \\ = \; & 8.583 \end{aligned}$$

- Nguồn thứ tư là tổng bình phương do tương tác giữa bệnh nhân trong mỗi nhóm điều trị, gọi là SS*patient(treat)*:

$$SS\textit{patient(treat)} = 3(3.000-3.350)^2 + 3(3.667-3.350)^2 + 3(2.333-3.350)^2 + 3(5.000-3.350)^2 + 3(5.333-4.833)^2 + 3(6.333-4.833)^2$$

$+3(4.000–4.833)^2 +3(3.667–4.833)^2$
$= 25.333$

- Ngoài ra, tổng bình phương cho toàn mẫu là:

$$SS\textit{total} = (6\text{-}4.167)^2 +(3\text{-}4.167)^2 +(0\text{-}4.167)^2 + \ldots +(3\text{-}4.167)^2 = 115.333$$

- Từ đó, chúng ta có thể ước tính tổng bình phương cho phần dư:

$$\begin{aligned} SSE &= SStotal – SSvắcxin – SStime – SSpatient(vắcxin)– SSvắcxin\text{-}time \\ &= 115.333 – 10.667 – 58.583 – 25.333 – 8.583 \\ &= 12.167 \end{aligned}$$

Tất cả các tính toán thủ công trên, như bạn đọc có thể thấy, khá phức tạp, và rất dễ sai sót. Nhưng trong R, chúng ta có thể có kết quả nhanh chóng . Sau đây, sẽ trình bày cách phân tích phương sai tái đo lường bằng R:

Chúng ta có thể lập bảng phân tích phương sai như sau:

Nguồn biến thiên	Bậc tự do (degrees of freedom)	Tổng bình phương (Sum of squares)	Trung bình bình phương (Mean square)	Kiểm định F
Giữa vắcxin và placebo	1	10.667	10.667	2.53
Bệnh nhân (nhóm điều trị)	6	25.333	4.222	-
Giữa 3 tháng	2	58.583	29.292	28.89
Thời gian và nhóm điều trị	2	8.583	4.292	4.23
Phần dư (residual)	12	12.167	1.014	-
Tổng số	23	115.333		

- Trước hết, chúng ta nhập dữ liệu cho từng bệnh nhân. Cũng như bất cứ phần mềm thống kê nào, mỗi giá trị phải được kèm theo những biến số đặc trưng như cho mỗi bệnh nhân, mỗi nhóm, và mỗi thời gian:

```
y <- c(6,7,4,8,
       3,3,1,4,
       0,1,2,3,
       6,9,5,6,
       5,4,3,2,
       5,6,4,3)
```

- Trong mỗi số liệu trên, cho R biết thuộc nhóm điều trị (mã số 1) hay giả dược (mã số 2). Cũng nên cho R biết `treat` là một biến thứ bậc (categorical variable) chứ không phải biến số (numerical variable):

```
treat <- c(1,1,1,1,
           1,1,1,1,
           1,1,1,1,
           2,2,2,2,
           2,2,2,2,
           2,2,2,2)
treat <- as.factor(treat)
```

- Trong mỗi số liệu trên, cho R biết thuộc tháng nào (mã số 1, 2, 3), và định nghĩa `time` là một biến thứ bậc.

```
time <- c(1,1,1,1,
          2,2,2,2,
          3,3,3,3,
          1,1,1,1,
          2,2,2,2,
          3,3,3,3)
time <- as.factor(time)
```

- Trong mỗi số liệu trên, cho R biết thuộc bệnh nhân nào (mã số 1, 2, 3, …,8), và định nghĩa `id` là một biến thứ bậc.

```
id <- c(1,2,3,4, 1,2,3,4, 1,2,3,4, 5,6,7,8, 5,6,7,8,
        5,6,7,8)
id <- as.factor(id)
```

- Nhập tất cả biến vào một data frame và đặt tên là `data`. Kiểm tra một lần nữa xem số liệu đã đúng với ý định sắp xếp hay chưa. Xin nhắc lại, trước khi phân tích số liệu, việc quan trọng là phải kiểm tra lại cho thật kĩ số liệu để đảm bảo số liệu đã được tổ chức đúng và thích hợp.

```
data <- data.frame(id, time, treat, y)
data
  id time treat y
1  1    1    1    6
2  2    1    1    7
3  3    1    1    4
4  4    1    1    8
5  1    2    1    3
6  2    2    1    3
7  3    2    1    1
8  4    2    1    4
9  1    3    1    0
```

```
10 2   3   1    1
11 3   3   1    2
12 4   3   1    3
13 5   1   2    6
14 6   1   2    9
15 7   1   2    5
16 8   1   2    6
17 5   2   2    5
18 6   2   2    4
19 7   2   2    3
20 8   2   2    2
21 5   3   2    5
22 6   3   2    6
23 7   3   2    4
24 8   3   2    3
```

Bây giờ, chúng ta đã sẵn sàng sử dụng R để phân tích. Hàm chính để phân tích phương sai là `aov` (analysis of variance). Trong hàm này, chú ý cách cung cấp thông số bằng cách dùng một hàm khác có tên là `Error`. Trong hàm `Error`, chúng ta cho R biết rằng mỗi bệnh nhân (`id`) "thuộc" vào một nhóm điều trị và do đó thuộc vào biến `time`. Cách để cho R biết là: `Error(id/time)`. Cụ thể hơn:

```
> repeated <- aov(y ~ treat*time + Error(id/time))
```

Lệnh trên đây yêu cầu R phân tích theo mô hình: `y = treat + time + treat*time` (chú ý `treat*time` tương đương với `treat+time+treat*time`), và trung bình bình phương phần dư phải được tách thành hai phần: một phần trong các bệnh nhân, và một phần giữa các tháng điều trị (viết tắt bằng kí hiệu `id/time`). Tất cả kết quả cho vào đối tượng có tên là `repeated`. Chúng ta yêu cầu một bảng tóm lược kết quả từ đối tượng `repeated`:

```
> summary(repeated)

Error: id
         Df Sum Sq  Mean Sq F value Pr(>F)
treat     1 10.6667 10.6667 2.5263 0.1631
Residuals 6 25.3333  4.2222

Error: id:time
          Df Sum Sq  Mean Sq  F value  Pr(>F)
time       2 58.583  29.292  28.8904  2.586e-05 ***
treat:time 2 8.583    4.292   4.2329  0.04064 *
Residuals 12 12.167   1.014
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Kết quả phân tích trong phần đầu của bảng trên cho thấy sự khác biệt giữa nhóm điều trị bằng thuốc và giả dược không có ý nghĩa thống kê (p = 0.16). Như vậy chúng ta có thể kết luận thuốc không có hiệu nghiệm giảm đau thấp khớp?

Câu trả lời là "không", bởi vì phần thứ hai của bảng phân tích phương sai cho thấy mối tương tác giữa `treat` và `time` (trị số p = 0.041). Điều này có nghĩa là độ khác biệt giữa thuốc và giả dược tùy thuộc vào tháng điều trị. Thật vậy, nếu chúng ta xem lại bảng 10.11 sẽ thấy trong tháng 1, trung bình của nhóm vắc-xin và giả dược không mấy khác nhau (6.25 và 6.50), nhưng đến tháng thứ 2 và nhất là tháng thứ 3 thì độ khác biệt giữa hai nhóm rất cao (như tháng thứ ba: 1.50 cho vắc-xin và 4.50 cho nhóm giả dược). Như vậy, độ hiệu nghiệm trong nhóm được điều trị tăng dần theo thời gian, còn trong nhóm giả dược thì hầu như không có khác biệt giữa 3 tháng. Tóm lại, qua thí nghiệm sơ khởi này chúng ta có thể nói vắc-xin có vẻ có hiệu quả giảm đau trong các bệnh nhân thấp khớp.

***

Trên đây là vài cách sử dụng cho việc phân tích phương sai với các thí nghiệm thông dụng. Thiết kế và phân tích thí nghiệm (experimental design) là một lĩnh vực nghiên cứu tương đối chuyên sâu, những chỉ dẫn trên đây không thể mô tả tất cả các phép tính cũng như phương pháp cho tất cả thí nghiệm. Tuy nhiên, trong thực tế, các phương pháp và thí nghiệm rất thường được áp dụng trong khoa học thực nghiệm. R có một package tên là `nlme` (non-linear mixed-effects) cũng có thể sử dụng cho các phân tích trên và các mô hình phức tạp hơn với đa biến và đa thứ bậc. Package này cũng có thể tải về máy miễn phí tại website của R: http://cran.R-project.org.