

Rで統計解析入門

(6) 分散分析と共分散分析



本日のメニュー

1. 分散分析

- ▶ イントロ

- ▶ データ「DEP」による例示

2. 共分散分析



分散分析

- ▶ 分散分析：要因が目的変数に影響を与えているかどうかを調べる手法
例：薬剤が QOL に影響を与えるか 薬剤は効果があるか
- ▶ 平方和：「平均からの距離の 2 乗」の和（= ばらつき, 変動）
例：薬剤の QOL への影響をみる場合, 以下の平方和が考えられる
 - ▶ 総平方和 (SST)：薬剤の平方和と誤差の平方和の和
 - ▶ 薬剤の平方和 (SSG)：意味のある変動, 興味のある変動
 - ▶ 誤差の平方和 (SSE)：意味のない変動
- ▶ 以下の方法で薬剤が QOL に影響を与えているかどうかを調べる
 - ▶ 薬剤の平方和が誤差の平方和（意味のない変動）と同じ
薬剤の影響は誤差と同程度 薬剤は QOL に影響を与えない 効果なし
 - ▶ 薬剤の平方和が誤差の平方和（意味のない変動）よりも大きい
薬剤の影響は誤差より大きい 薬剤は QOL に影響を与える 効果あり



分散分析の例

薬剤	A	A	A	B	B	B
QOL	1	2	3	4	5	6

- ▶ QOL に関するデータ（データ「DEP」とは別のもの）
 - ▶ GROUP：薬剤の種類（A, B）
 - ▶ QOL：QOL の点数（数値） 点数が大きい方が良い
- ▶ 以下の平方和を求める
 - ▶ 総平方和（Sum of Squares for Total；SST）
 - ▶ 薬剤の平方和（Sum of Squares for Group；SSG）
 - ▶ 誤差の平方和（Sum of Squares for Error；SSE）



分散分析の例

薬剤	A	A	A	B	B	B
QOL	1	2	3	4	5	6

- ▶ 総平方和 (SST) : (各データー全平均)² の総和
- ▶ 薬剤の平方和 (SSG) : (各薬剤の平均ー全平均)² の総和
- ▶ 誤差の平方和 (SSE) : (各データー各薬剤の平均)² の総和

- ▶ $F = \frac{SSG / \text{薬剤の種類} - 1}{SSE / \text{データ数} - \text{薬剤の種類}}$ が 1に近い : 薬剤の効果はない
1より大きい : 薬剤の効果はある



平方和の計算

- ▶ 関数 `lm()` と関数 `Anova()` の組み合わせで計算できる

```
> GROUP <- c("A", "A", "A", "B", "B", "B")
> QOL <- c(1, 2, 3, 4, 5, 6)
> result <- lm(QOL ~ GROUP)
> install.packages("car", dep=T) # パッケージ car のインストール
> library(car) # パッケージ car の呼び出し
> Anova(result, Type="II") # 分散分析表 (Type II 平方和)
```

Response: QOL

	Sum Sq	Df	F value	Pr(>F)
<u>GROUP</u>	<u>13.5</u>	1	<u>13.5</u>	<u>0.02131</u> *
<u>Residuals</u>	<u>4.0</u>	4		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



平方和の計算

```
> Anova(result, Type="II") # 分散分析表 (Type II 平方和)
Response: QOL
      Sum Sq Df F value Pr(>F)
GROUP  13.5  1  13.5 0.02131 *
Residuals  4.0  4
```

- ▶ 薬剤の平方和 (SSG) : 13.5
- ▶ 誤差の平方和 (SSE) : 4.0
- ▶ 総平方和 (SST) : $13.5 + 4.0 = 17.5$
- ▶ $F = \{ 13.5 / (2-1) \} \div \{ 4.0 / (6-2) \} = 13.5$
 - ▶ 検定の帰無仮説 $H_0 : F = 1$ (薬剤の効果はなし)
 - ▶ F が 1 かどうかの検定結果 $p = 0.02131$ 有意なので $F \neq 1$
 - ▶ よって薬剤の効果あり



平方和の計算

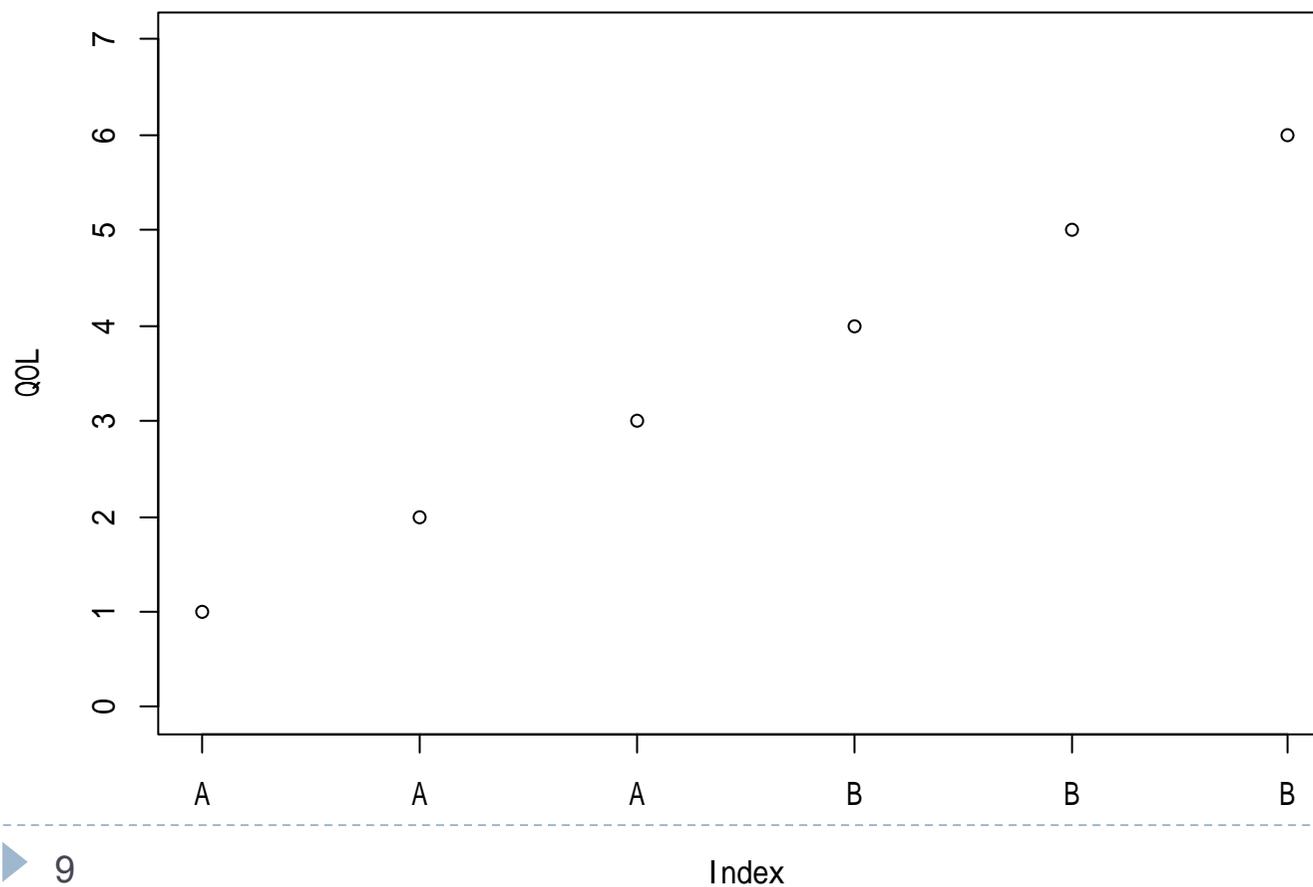
- ▶ 地道に計算することも可

```
> mean(QOL); mean(QOL[1:3]); mean(QOL[4:6]) # 全平均, Aの平均, Bの平均
[1] 3.5
[1] 2
[1] 5
> ( SST <- sum( (QOL-3.5)^2 ) )
[1] 17.5
> ( SSG <- 3*(2-3.5)^2 + 3*(5-3.5)^2 )
[1] 13.5
> ( SSE <- sum( (QOL[1:3]-2)^2 ) + sum( (QOL[4:6]-5)^2 ) )
[1] 4
> ( F <- ( SSG/(2-1) ) / ( SSE/(6-2) ) )
[1] 13.5
> 1-pf(F, 1, 4)
[1] 0.02131164
```



図による平方和の説明

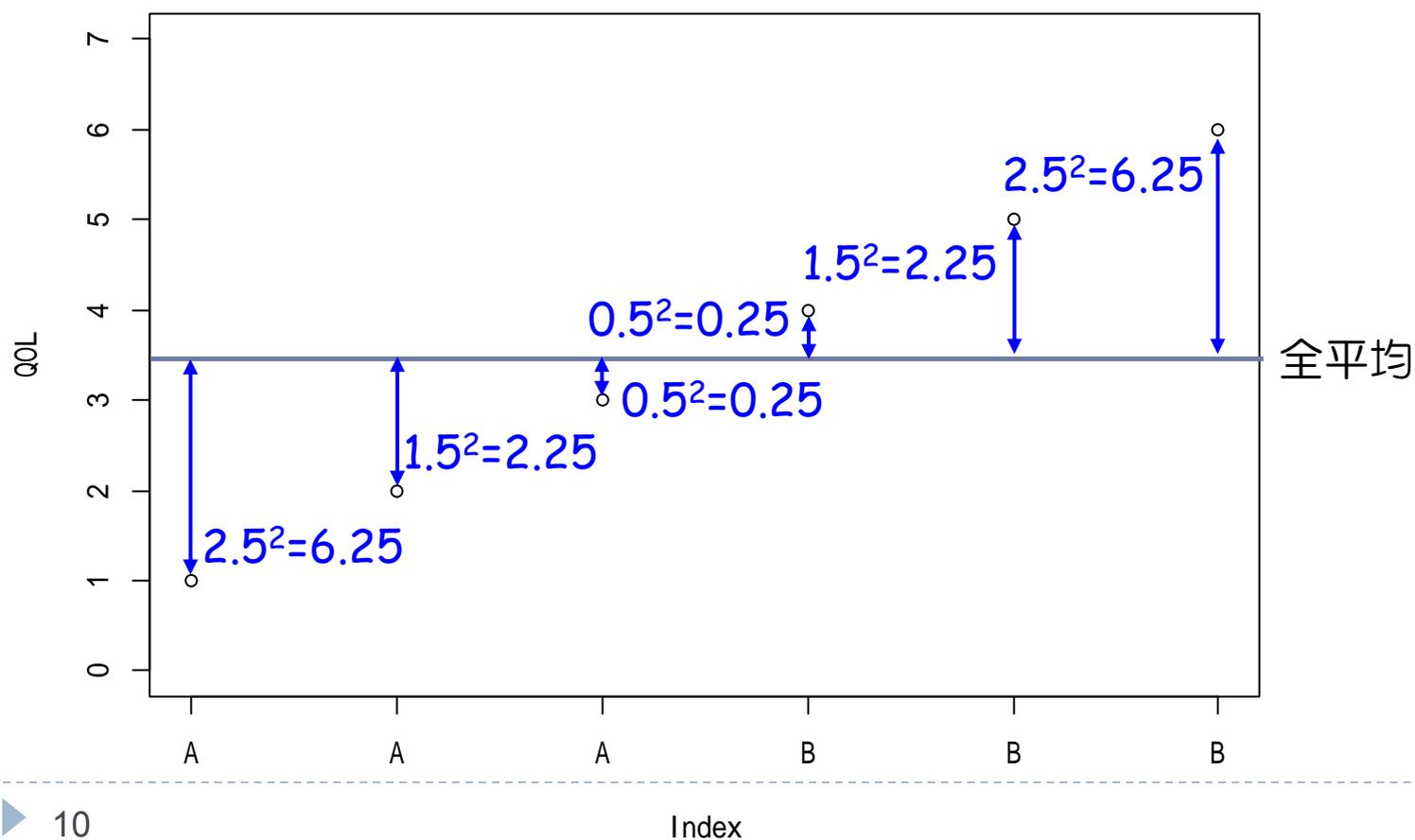
```
> plot(QOL, ylim=c(0,7), xaxt="n")  
> axis(1, 1:6, GROUP)
```





図による平方和の説明

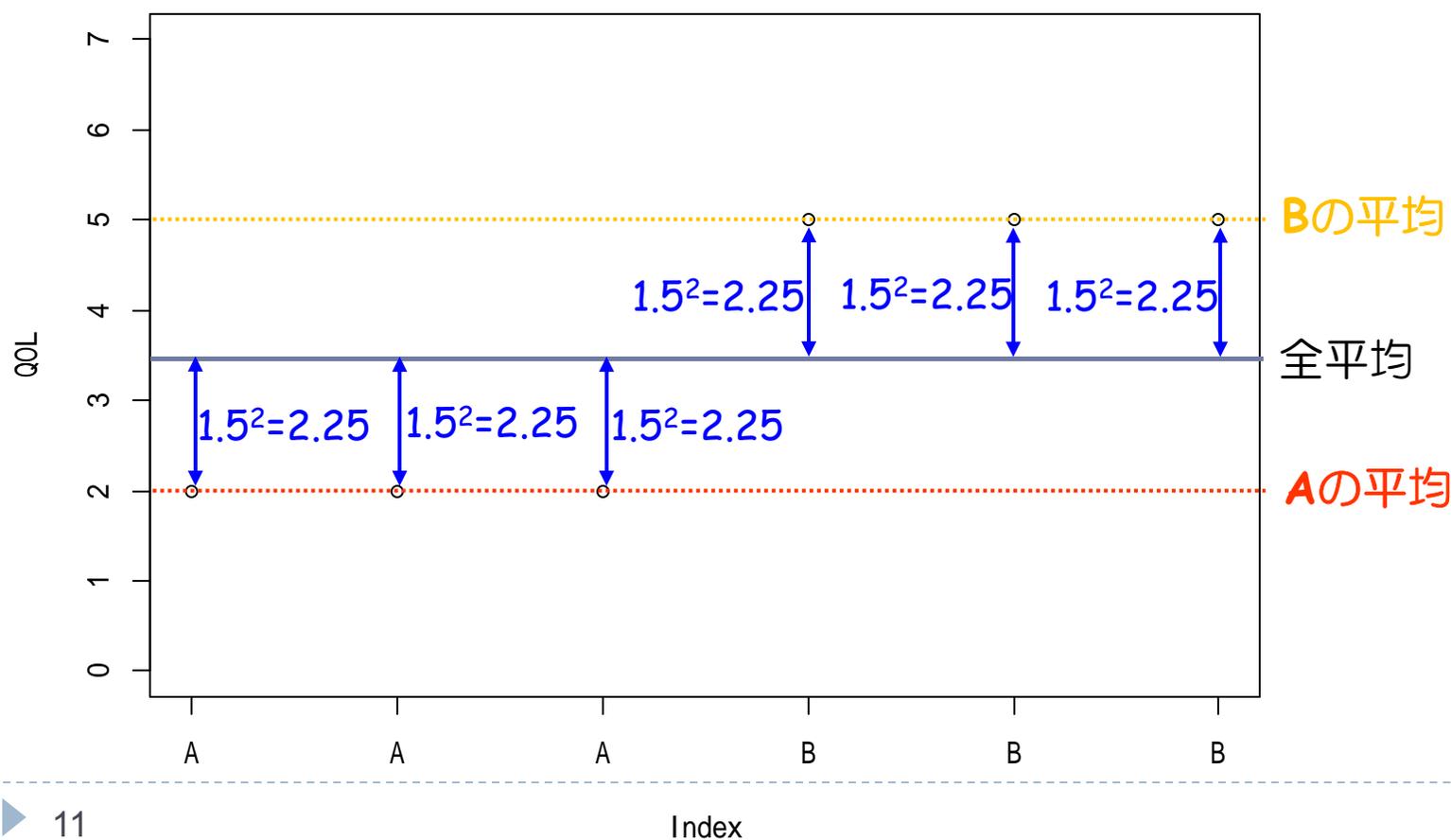
- ▶ 総平方和 (SST) : (各データ-全平均)² の総和
「各データ」と「全平均」の距離の 2 乗を足し算している





図による平方和の説明

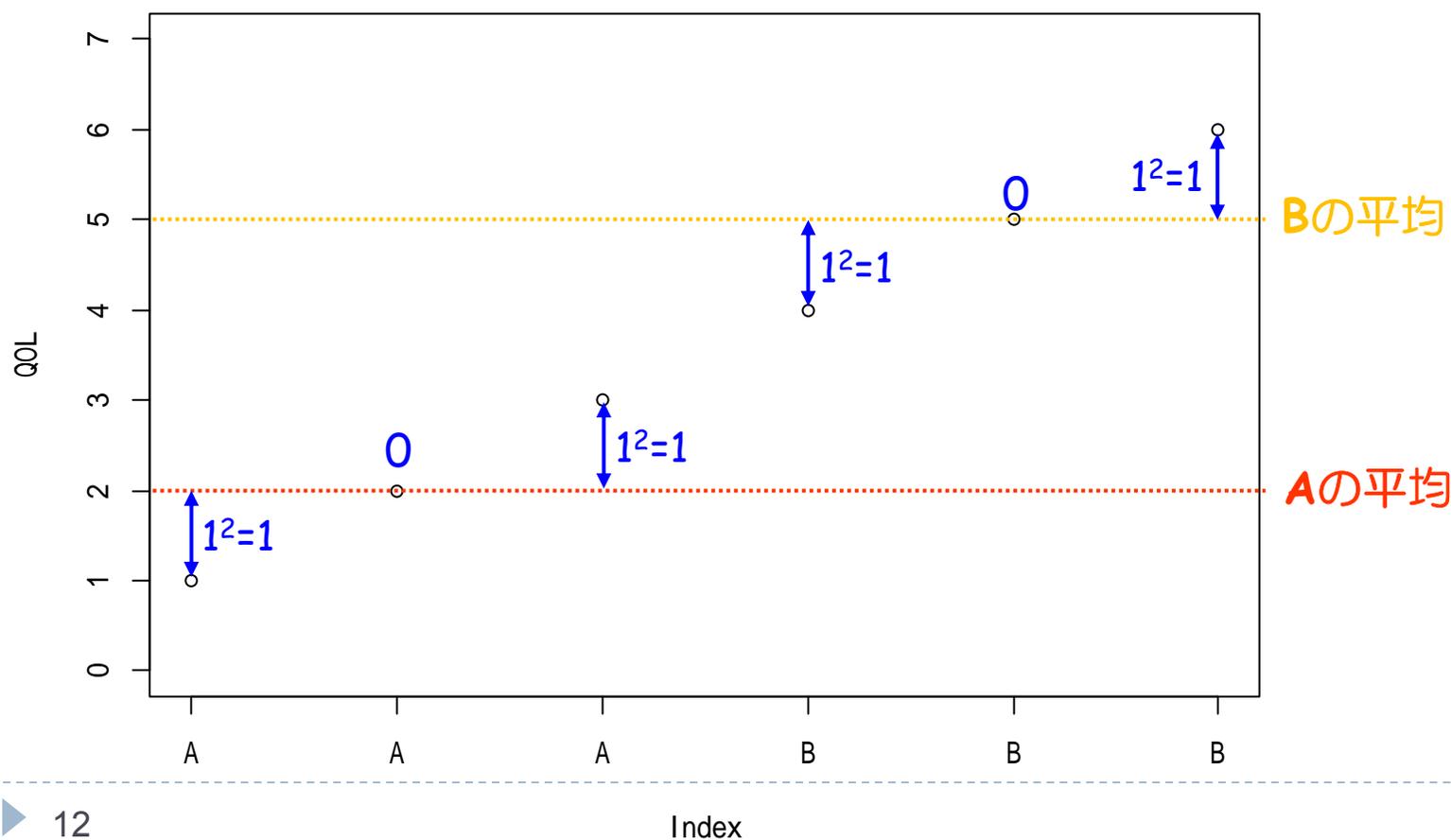
- ▶ **薬剤の平方和 (SSG)** : (各薬剤の平均 - 全平均)² の総和
「データが属する薬剤の平均値」と「全平均」の距離の2乗を足し算
値が小さい: 薬剤の効果なし, 値が大きい: 薬剤の効果あり





図による平方和の説明

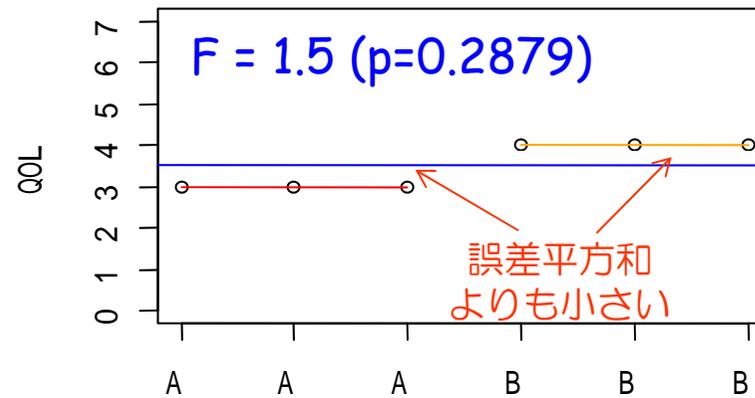
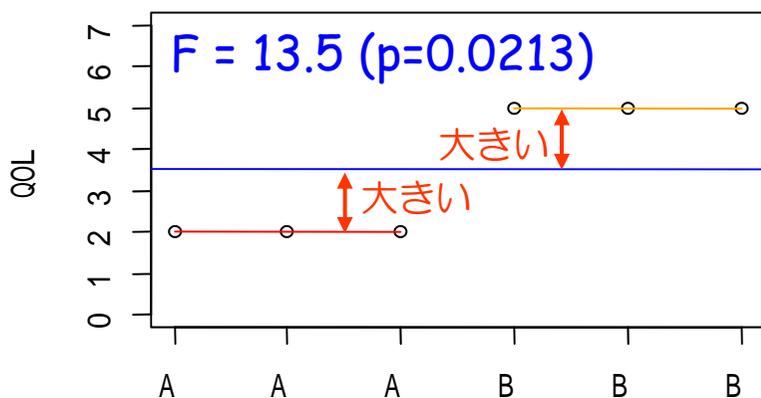
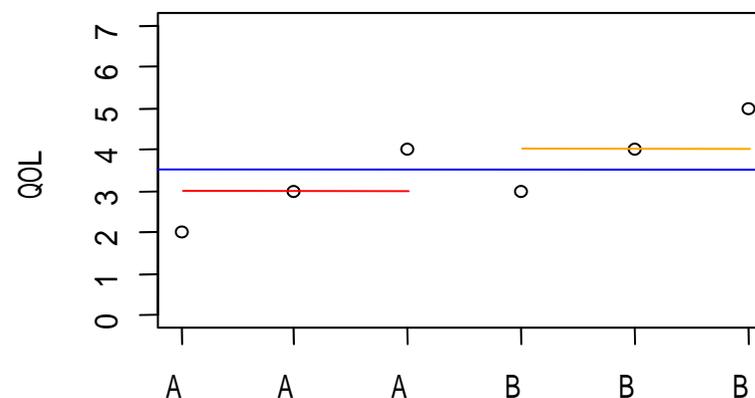
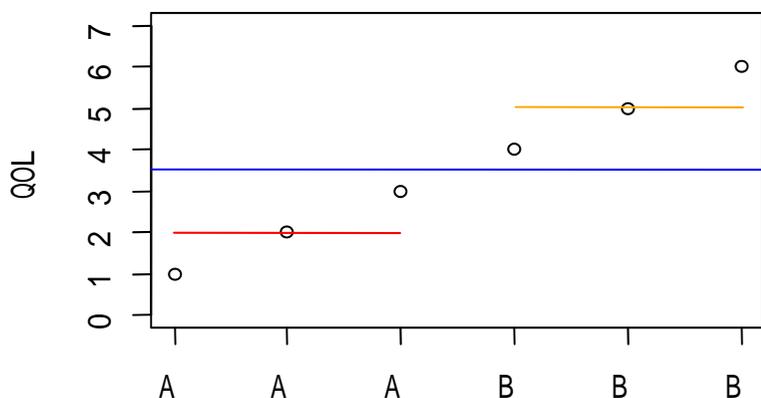
- ▶ 誤差の平方和 (SSE) : (各データー各薬剤の平均)² の総和
「データ」と「データが属する薬剤の平均値」の距離の2乗を足し算
「総平方和から薬剤の平方和を引いた残り (ゴミ)」の方が理解しやすい?





図による平方和の説明

- ▶ 薬剤の平方和が誤差の平方和 **よりも大きい** = 薬剤の **効果あり**
「QOLの変動（平方和）」を「薬剤の平均からの変動（平方和）」を代わりに使うことで、ある程度説明することが出来るという状態





【参考】 前の頁のグラフを作成するプログラム

```
> par(mfrow=c(2,1))
> Q0L  <- c(1,2,3,4,5,6)
> plot(Q0L, ylim=c(0,7), xaxt="n")
> axis(1, 1:6, GROUP)
> abline(h=3.5, col="blue")
> Q0L  <- c(2,2,2,5,5,5)
> plot(Q0L, ylim=c(0,7), xaxt="n")
> axis(1, 1:6, GROUP)
> segments(1,2, 3,2, col="red")
> segments(4,5, 6,5, col="orange")
> abline(h=3.5, col="blue")
```

```
> Q0L  <- c(2,3,4,3,4,5)
> plot(Q0L, ylim=c(0,7), xaxt="n")
> axis(1, 1:6, GROUP)
> abline(h=3.5, col="blue")
> Q0L  <- c(3,3,3,4,4,4)
> plot(Q0L, ylim=c(0,7), xaxt="n")
> axis(1, 1:6, GROUP)
> segments(1,3, 3,3, col="red")
> segments(4,4, 6,4, col="orange")
> abline(h=3.5, col="blue")
```



本日のメニュー

1. 分散分析

- ▶ イントロ

- ▶ データ「DEP」による例示

2. 共分散分析



準備：データ「DEP」の読み込み

1. データ「DEP」を以下からダウンロードする
<http://www.cwk.zaq.ne.jp/fkhd708/files/dep.csv>
2. ダウンロードした場所を把握する　ここでは「c:/temp」とする
3. R を起動し，2. の場所に移動し，データを読み込む
4. データ「DEP」から薬剤 A と B のデータを抽出

```
> setwd("c:/temp") # dep.csv がある場所に移動
> getwd() # 移動できたかどうか確認
> DEP <- read.csv("dep.csv") # dep.csv を読み込む
> AB <- subset(DEP, GROUP != "C") # 薬剤 A と B のデータを抽出
> AB$GROUP <- factor(AB$GROUP) # 薬剤の水準を 2 カテゴリに
> AB$GROUP <- relevel(AB$GROUP, ref="B") # ベースを「B」に変更
```



準備：架空のデータ「DEP」の変数

- ▶ **GROUP**：薬剤の種類（A, B, C）
- ▶ **QOL**：QOL の点数（数値） 点数が大きい方が良い
- ▶ **EVENT**：改善の有無（1：改善あり，2：改善なし）
QOLの点数が5点以上である場合を「改善あり」とする
- ▶ **DAY**：観察期間（数値，単位は日）
- ▶ **PREDRUG**：前治療薬の有無（YES：他の治療薬を投与したことあり，
NO：投与したことなし）
- ▶ **DURATION**：罹病期間（数値，単位は年）



準備：架空のデータ「DEP」（一部）

GROUP	QOL	EVENT	DAY	PREDRUG	DURATION
A	15	1	50	NO	1
A	13	1	200	NO	3
A	11	1	250	NO	2
A	11	1	300	NO	4
A	10	1	350	NO	2
A	9	1	400	NO	2
A	8	1	450	NO	4
A	8	1	550	NO	2
A	6	1	600	NO	5
A	6	1	100	NO	7
A	4	2	250	NO	4
A	3	2	500	NO	6
A	3	2	750	NO	3
A	3	2	650	NO	7
A	1	2	1000	NO	8
A	6	1	150	YES	6
A	5	1	700	YES	5
A	4	2	800	YES	7
A	2	2	900	YES	12
A	2	2	950	YES	10
B	13	1	380	NO	9
B	12	1	880	NO	5
B	11	1	940	NO	2
B	4	2	20	NO	7
B	4	2	560	NO	2
B	5	1	320	YES	11
B	5	1	940	YES	3
B	4	2	80	YES	6
B	3	2	140	YES	7
B	3	2	160	YES	13



【復習】 交絡の有無の判定方法

興味のある因子が薬剤，「前治療の有無」が交絡因子かどうかを判定する

- ▶ 以下のモデルで回帰分析し，薬剤の効果（薬剤に関する傾き β_1 ）が変わる場合，「前治療の有無」は交絡因子
 - ▶ 「薬剤のみ」のモデル： $QOL = \beta_0 + \beta_1 \times \text{薬剤}$
 - ▶ 「薬剤+前治療の有無」のモデル： $QOL = \beta_0 + \beta_1 \times \text{薬剤} + \beta_2 \times \text{前治療の有無}$
- ▶ 薬剤の効果が変わる = 「分散分析の結果，薬剤の効果が変わる」と読み替えて，分散分析表から交絡の有無を判定してみる



交絡が起きているかの判定

```
> result <- lm(QOL ~ GROUP, data=AB) # 薬剤のみのモデル
> Anova(result, Type="II") # 分散分析表 (Type II 平方和)
Response: QOL
      Sum Sq Df F value Pr(>F)
GROUP    62.5  1  4.2035 0.04728 *
Residuals 565.0 38

> result <- lm(QOL ~ GROUP+PREDRUG, data=AB) # 薬剤 + 前治療の有無のモデル
> Anova(result, Type="II") # 分散分析表 (Type II 平方和)
Response: QOL
      Sum Sq Df F value Pr(>F)
GROUP     0.0  1  0.000 1.0000000
PREDRUG  187.5  1 18.378 0.0001243 ***
Residuals 377.5 37
```

- ▶ 薬剤のみのモデル : 薬剤の p 値 = 0.04728 (薬剤の効果あり)
 - ▶ 薬剤+前治療の有無のモデル : 薬剤の p 値 = 1.00... (薬剤の効果なし)
- 「効果あり」から「効果なし」に変わっているので交絡が起きている



【復習】 交互作用があるかどうかの判定方法

「薬剤×前治療の有無」の交互作用があるかどうかを判定する場合

- ▶ 以下のモデルで回帰分析し，交互作用項の効果（傾き β_3 ）が 0 でない変わる場合，交互作用あり
- ▶ 「薬剤＋前治療の有無＋薬剤×前治療の有無」のモデル：
$$QOL = \beta_0 + \beta_1 \times \text{薬剤} + \beta_2 \times \text{前治療の有無} + \beta_3 \times \text{薬剤} \times \text{前治療の有無}$$
- ▶ 交互作用項の効果が 0 でない
＝「分散分析の結果，交互作用の効果がある」と読み替えて，
分散分析表から交互作用の有無を判定してみる



交互作用があるかどうかの判定

```
> result <- lm(QOL ~ GROUP*PREDRUG, data=AB) # 交互作用モデル (薬剤 × 前治療)
> Anova(result, Type="II") # 分散分析表 (Type II 平方和)
Anova Table (Type II tests)

Response: QOL

          Sum Sq Df F value    Pr(>F)
GROUP          0.0  1  0.0000 1.0000000
PREDRUG       187.5  1 18.6053 0.0001196 ***
GROUP:PREDRUG  14.7  1  1.4587 0.2350191
Residuals     362.8 36
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- ▶ 検定の帰無仮説 $H_0 : F = 1$ (交互作用の効果はない)
- ▶ 交互作用項の p 値 = 0.2350 (交互作用の効果なし)
交互作用はなさそう



本日のメニュー

1. 分散分析

- ▶ イントロ
- ▶ データ「DEP」による例示

2. 共分散分析



分散分析と共分散分析

前頁までで紹介した手法で QOL に影響を与える因子を調べる際、

- ▶ モデルの中の因子が全てカテゴリ変数（薬剤，性別，前治療の有無等）

分散分析

- ▶ カテゴリ変数が 1 つだけ（薬剤）入っている：一元配置分散分析
- ▶ カテゴリ変数が 2 つ（薬剤+前治療の有無）入っている：二元配置分散分析
- ▶ カテゴリ変数が 3 つ入っている場合：三元配置分散分析
- ▶ カテゴリ変数が 4 つ・・・：四元配置分散分析
- ▶ モデルの中にカテゴリ変数と連続変数（年齢や罹病期間等）が混在

共分散分析

- ▶ ここではモデルに「薬剤」と「罹病期間」を入れてみる



QOL に関する共分散分析

- 以下のモデルについて共分散分析を行い回帰式と分散分析表を求める

$$QOL = \beta_0 + \beta_1 \times \text{薬剤} + \beta_2 \times \text{罹病期間} \quad (\text{薬剤} \quad 1 : A, 0 : B)$$

$$QOL = 7.89 + 1.29 \times \text{薬剤} - 0.53 \times \text{罹病期間} \quad \text{となった}$$

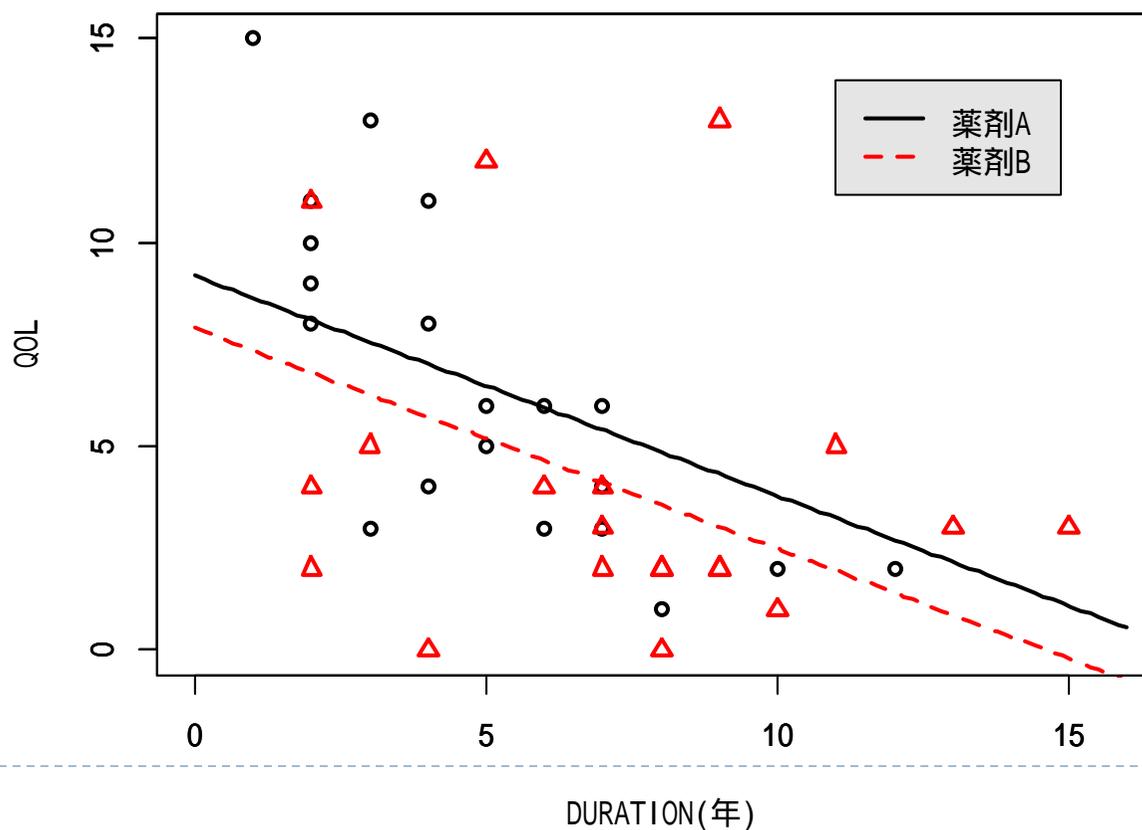
```
> result <- lm(QOL ~ GROUP+DURATION, data=AB) # 薬剤 + 罹病期間のモデル
> summary(result) # 結果の要約を表示

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.8966     1.4777   5.344 4.85e-06 ***
GROUPA        1.2907     1.1678   1.105 0.27619
DURATION     -0.5375     0.1732  -3.102 0.00367 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



QOL に関する共分散分析

- ▶ 回帰式 $QOL = 7.89 + 1.29 \times \text{薬剤} - 0.53 \times \text{罹病期間}$ (薬剤 1 : A, 0 : B)
- ▶ 薬剤 A の回帰式 : $QOL = 9.18 - 0.53 \times \text{罹病期間}$
- ▶ 薬剤 B の回帰式 : $QOL = 7.89 - 0.53 \times \text{罹病期間}$





前頁のグラフを描くプログラム

```
> # 散布図と回帰直線
> A <- function(x) 7.89+1.29*x-0.54*x
> B <- function(x) 7.89+1.29*0-0.54*x
> plot(QOL ~ DURATION, data=AB, pch=ifelse(GROUP=="A",1,2),
+      col=ifelse(GROUP=="A",1,2),
+      xlim=c(0,16), ylim=c(0,15), lwd=2, lty=1, ann=F)
> par(new=T)
> curve(A, xlim=c(0,16), ylim=c(0,15), lwd=2, col=1, lty=1, ann=F)
> par(new=T)
> curve(B, xlim=c(0,16), ylim=c(0,15), lwd=2, col=2, lty=2,
+      xlab="DURATION(年)", ylab="QOL")
> legend(11, 14, c("薬剤A ", "薬剤B "), lwd=2, col=1:2, lty=1:2,
+      ncol=1, cex=1.0, bg="gray90")
```



QOL に関する共分散分析

- ▶ 薬剤間のQOLの平均値の差：
薬剤 A と薬剤 B の回帰式の引き算から，QOL の平均値の差が求まる
回帰式の「薬剤の傾きの推定値 (GROUPA : 1.29)」を見ればよい
- ▶ 薬剤間の QOL の平均値の差に対する「Pr(>|t|)」の意味：
- ▶ 「薬剤 B の QOL の平均値を 0 としたときの，
薬剤 A のQOL の平均値が 0 かどうかの検定」の結果
= 「薬剤 A と薬剤 B の QOL の平均値の差が 0 かどうかの検定」
結果は「Pr(>|t|) : 0.276」となっており，5% よりも大きいので
「帰無仮説は間違っていない」 QOL の平均値に差があるとはいえない

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	7.8966	1.4777	5.344	4.85e-06	***
GROUPA	1.2907	1.1678	1.105	0.27619	
DURATION	-0.5375	0.1732	-3.102	0.00367	**

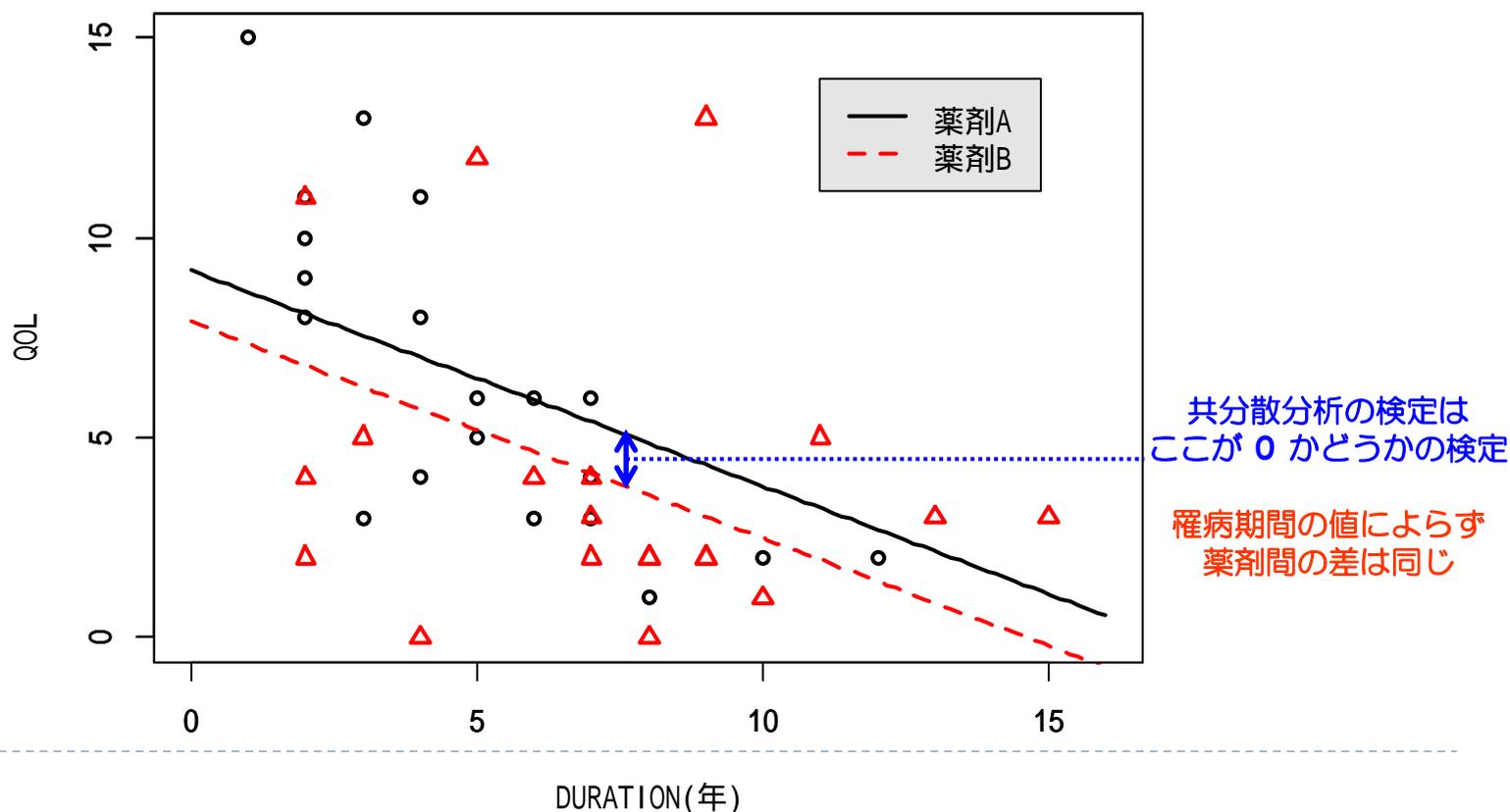


QOL に関する共分散分析

- ▶ 薬剤間の QOL の平均値の差に対する「Pr(>|t|)」の意味：

「薬剤 A と薬剤 B の QOL の平均値の差が 0 かどうかの検定」

結果は「Pr(>|t|) : 0.276」となっており、5% よりも大きいので
「帰無仮説は間違っていない」 QOL の平均値に差があるとはいえない





【参考】 回帰直線と平均値の関係

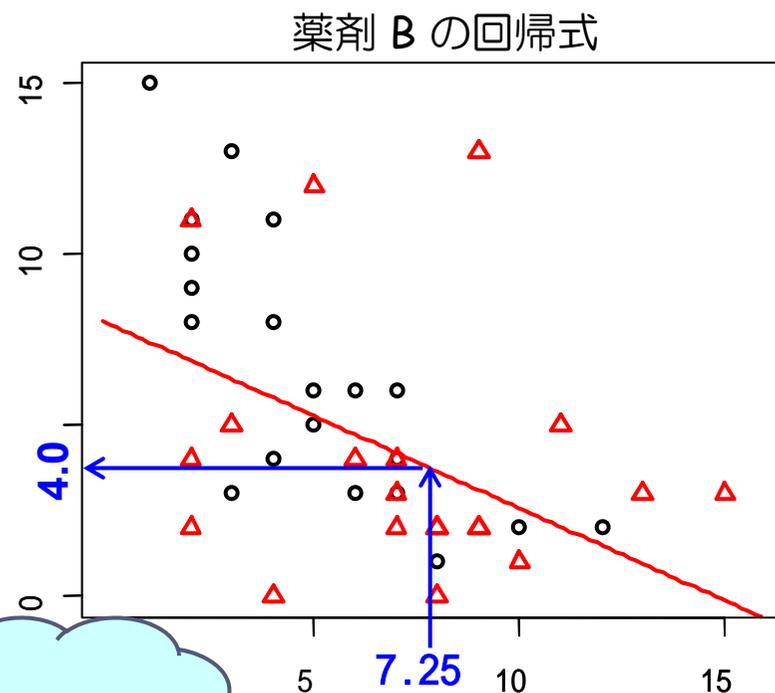
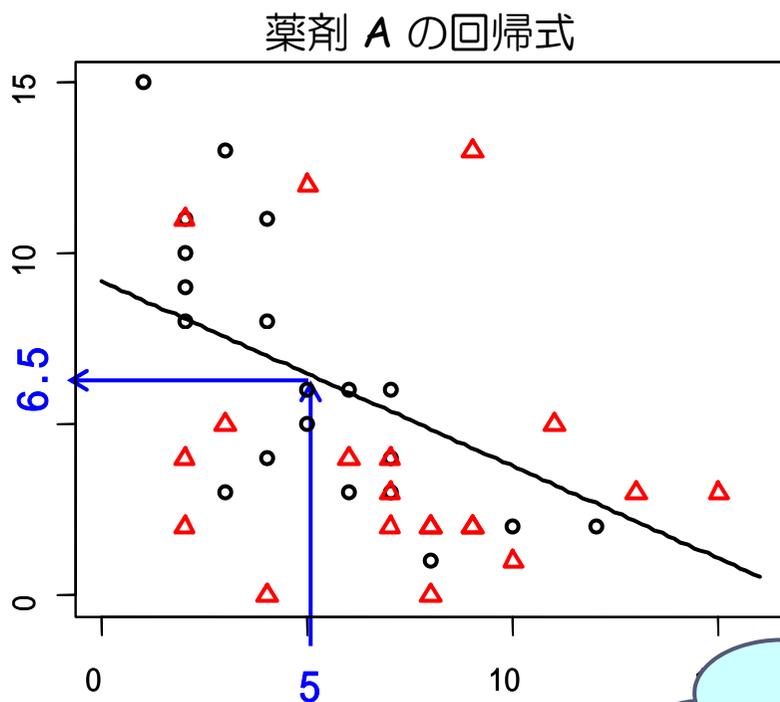
- ▶ 薬剤 A の回帰式： $QOL = 9.18 - 0.53 \times \text{罹病期間}$
- ▶ 薬剤 B の回帰式： $QOL = 7.89 - 0.53 \times \text{罹病期間}$
- ▶ ここで、各薬剤の罹病期間の平均を算出 **A : 5 年, B : 7.25 年**
- ▶ 各薬剤の罹病期間の平均を各薬剤の回帰式に代入すると・・・
 - ▶ 薬剤 A の回帰式： $QOL = 9.18 - 0.53 \times 5 = 6.5$ **QOL の平均と一致**
 - ▶ 薬剤 B の回帰式： $QOL = 7.89 - 0.53 \times 7.25 = 4.0$ **QOL の平均と一致**

```
> by(AB$DURATION, AB$GROUP, mean) # 各薬剤の罹病期間の平均
AB$GROUP: B
[1] 7.25
-----
AB$GROUP: A
[1] 5
```



【参考】 回帰直線と平均値の関係

- ▶ 各薬剤の罹病期間の平均を各薬剤の回帰式に代入すると・・・
 - ▶ 薬剤 A の回帰式： $QOL = 9.18 - 0.53 \times 5 = 6.5$ QOL の平均と一致
 - ▶ 薬剤 B の回帰式： $QOL = 7.89 - 0.53 \times 7.25 = 4.0$ QOL の平均と一致



QOL の平均と一致



ある因子が交絡因子かどうかの判定方法

興味のある因子が薬剤、「罹病期間」が交絡因子かどうかを判定する

- ▶ 以下のモデルで共分散分析し、薬剤の効果が変わる場合、「罹病期間」は交絡因子
 - ▶ 「薬剤のみ」のモデル：
$$QOL = \beta_0 + \beta_1 \times \text{薬剤}$$
 - ▶ 「薬剤＋罹病期間」のモデル：
$$QOL = \beta_0 + \beta_1 \times \text{薬剤} + \beta_2 \times \text{罹病期間}$$

【参考】前回の「薬剤」と「前治療の有無」の場合は、全体と層別の結果（前治療なしの結果と、前治療ありの結果）を比較することでも確認することが出来たが、連続変数の場合は一旦カテゴリ化（例えば、罹病期間が5年以上、5年以下）した上で層別の結果を出せばよいが、カテゴリ化する際の閾値（5年？6年？7年？）の設定の仕方によって結果がコロコロ変わる場合があるので注意が必要



【再掲】 薬剤のみのモデル

```
> result <- lm(QOL ~ GROUP, data=AB) # 薬剤のみのモデル
> summary(result) # 結果の要約を表示

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.0000     0.8622   4.639 4.07e-05 ***
GROUPA       2.5000     1.2194   2.050  0.0473 *

> Anova(result, Type="II") # 分散分析表 (Type II 平方和)

Response: QOL
            Sum Sq Df F value  Pr(>F)
GROUP       62.5  1  4.2035 0.04728 *
Residuals  565.0 38
```

- ▶ 薬剤の傾き : 2.50 (傾きが 0 かどうかの検定の p 値 = 0.04728)
- ▶ 薬剤の平方和 : p 値 = 0.04728 (薬剤の効果あり)



ある因子が交絡因子かどうかの判定方法

```
> result <- lm(QOL ~ GROUP+DURATION, data=AB) # 薬剤 + 罹病期間のモデル
> summary(result) # 結果の要約を表示
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.8966     1.4777    5.344 4.85e-06 ***
GROUPA       1.2907     1.1678    1.105 0.27619
DURATION     -0.5375     0.1732   -3.102 0.00367 **

> Anova(result, Type="II") # 分散分析表 (Type II 平方和)
Response: QOL
              Sum Sq Df F value  Pr(>F)
GROUP         14.80  1  1.2216 0.276187
DURATION     116.63  1  9.6243 0.003667 **
Residuals   448.37 37
```

- ▶ 薬剤のみのモデル : 群間差 = 2.50
- ▶ 薬剤 + 罹病期間のモデル : 群間差 = 1.29

傾きが変わっているので交絡が起きている



ある因子が交絡因子かどうかの判定方法

```
> result <- lm(QOL ~ GROUP+DURATION, data=AB) # 薬剤 + 罹病期間のモデル
> summary(result) # 結果の要約を表示
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.8966     1.4777   5.344 4.85e-06 ***
GROUP         1.2907     1.1678   1.105 0.27619
DURATION      -0.5375     0.1732  -3.102 0.00367 **

> Anova(result, Type="II") # 分散分析表 (Type II 平方和)
Response: QOL
              Sum Sq Df F value  Pr(>F)
GROUP         14.80  1  1.2216 0.276187
DURATION     116.63  1  9.6243 0.003667 **
Residuals   448.37 37
```

- ▶ 薬剤のみのモデル : (平均) 平方和の検定の p 値 = 0.0213
- ▶ 薬剤 + 罹病期間のモデル : (平均) 平方和の検定の p 値 = 0.2761
薬剤の効果が「あり なし」に変わっているので交絡が起きている



交互作用とは

- ▶ **交互作用**：複数の変数の組み合わせにより生じる作用のこと
- ▶ **交互作用がある**：2つの要因（例えば「薬剤×罹病期間」）が互いに影響を及ぼし合っている状態のこと
 - 「薬剤×罹病期間」を、「薬剤」と「罹病期間」との交互作用を表すこととし**交互作用項**と呼ぶことにする
 - 「薬剤×罹病期間」の交互作用がある場合、この要因である「罹病期間」を**効果修飾因子**と呼ぶ
- ▶ 「薬剤×連続変数」の交互作用が「なし」の状態と「あり」の状態を解説した後、「薬剤×罹病期間」の交互作用があるかどうか調べる

【参考】前回の「薬剤」と「前治療の有無」の場合は、全体と層別の結果

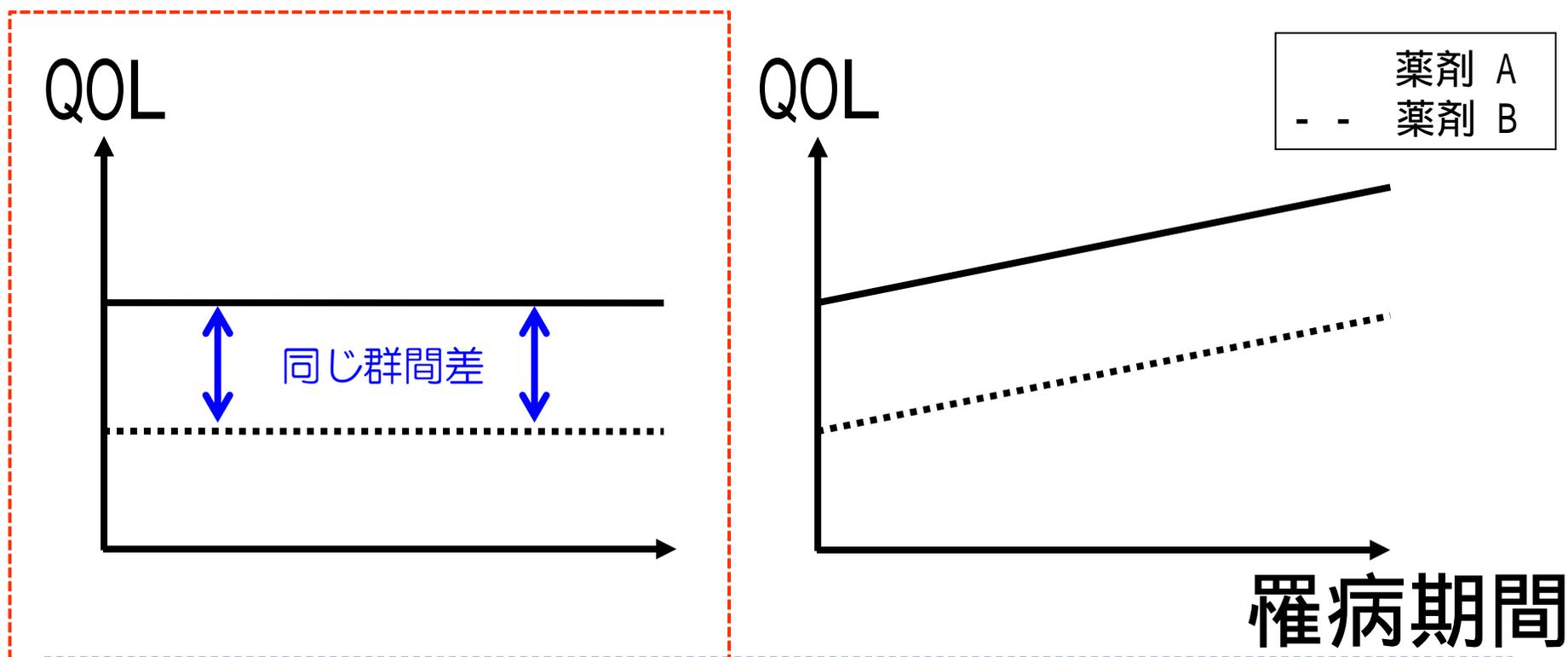
（前治療なしの結果と、前治療ありの結果）を比較することでも確認することが出来たが、連続変数の場合は一旦カテゴリ化（例えば、罹病期間が5年以上、5年以下）した上で層別の結果を出せばよい

が、カテゴリ化する際の閾値（5年？6年？7年？）の設定の仕方で結果がコロコロ変わる場合があるので注意が必要



交互作用がない状態 (●, ○ : QOL の平均値)

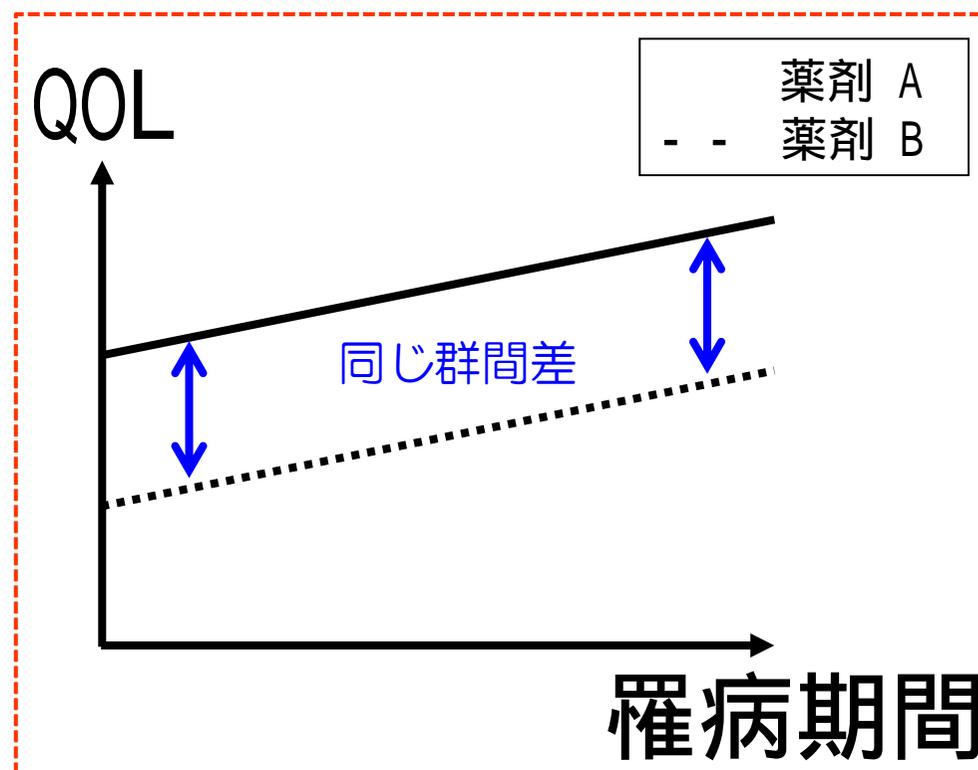
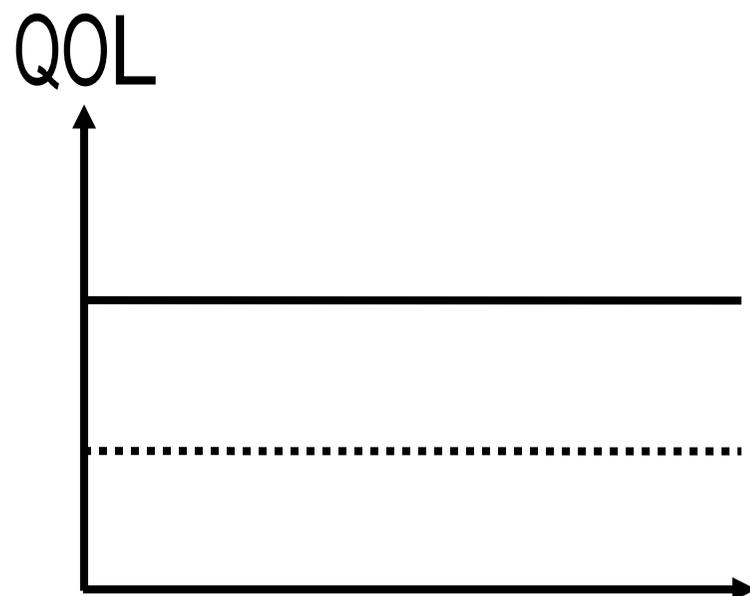
- ▶ 左下の図は以下の特徴がある
 - ▶ 「薬剤×罹病期間」の交互作用がない
 - ▶ 罹病期間が QOL に影響を及ぼしていない
罹病期間の値が大きくても小さくても、薬剤間の平均値の差は同じ





交互作用がない状態 (●, ○ : QOL の平均値)

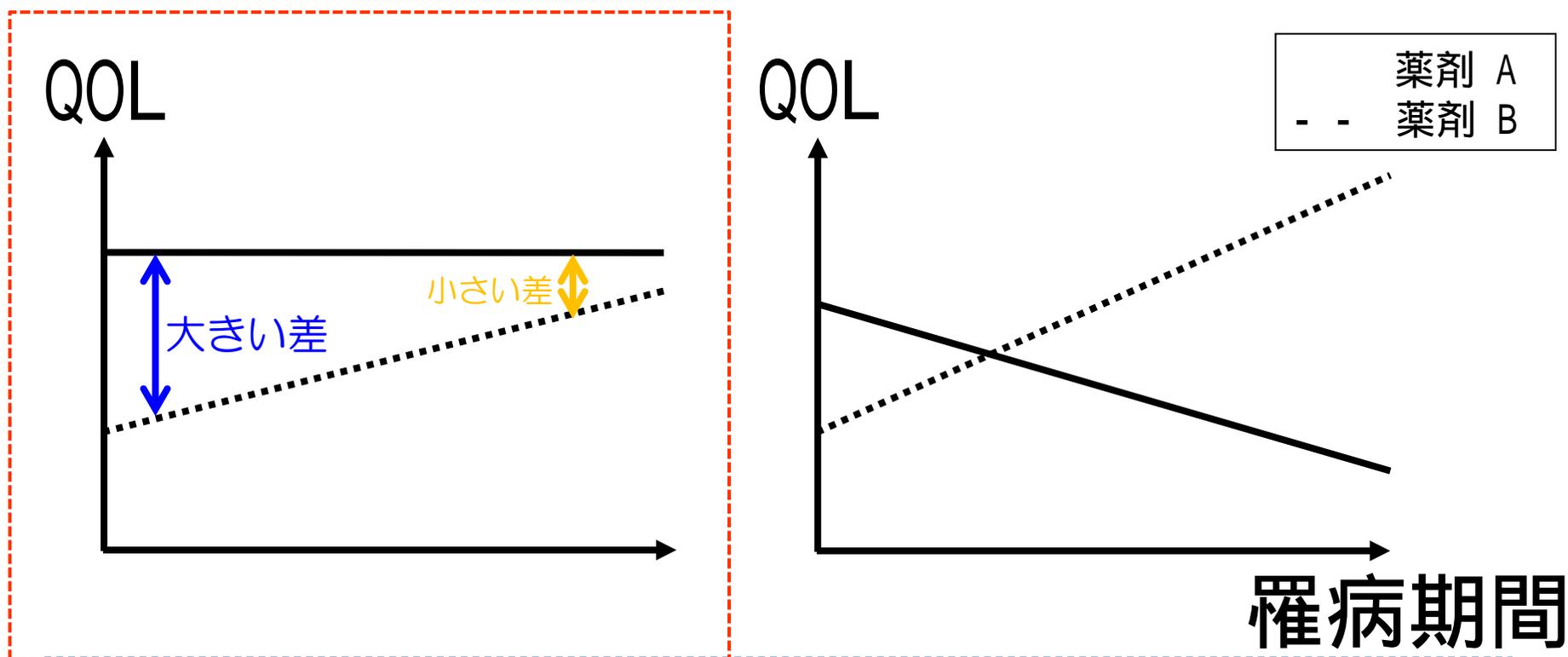
- ▶ 右下の図は以下の特徴がある
 - ▶ 「薬剤×罹病期間」の交互作用がない
 - ▶ 罹病期間が QOL に影響を及ぼしている
罹病期間の値が大きくても小さくても、薬剤間の平均値の差は同じ





交互作用がある状態（●，○：QOL の平均値）

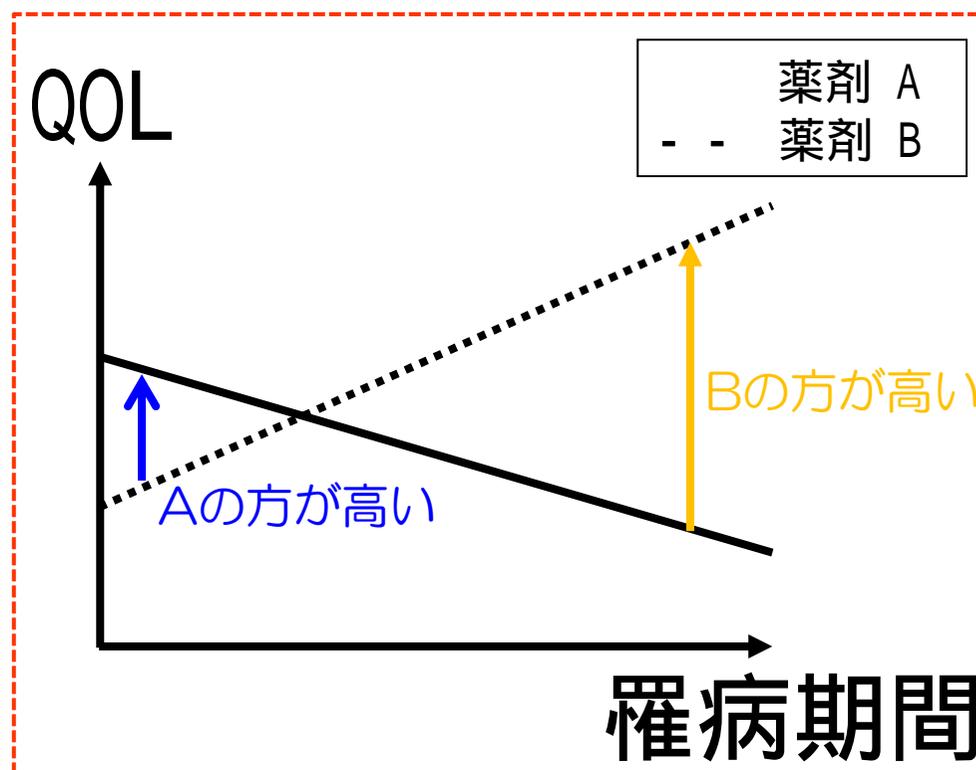
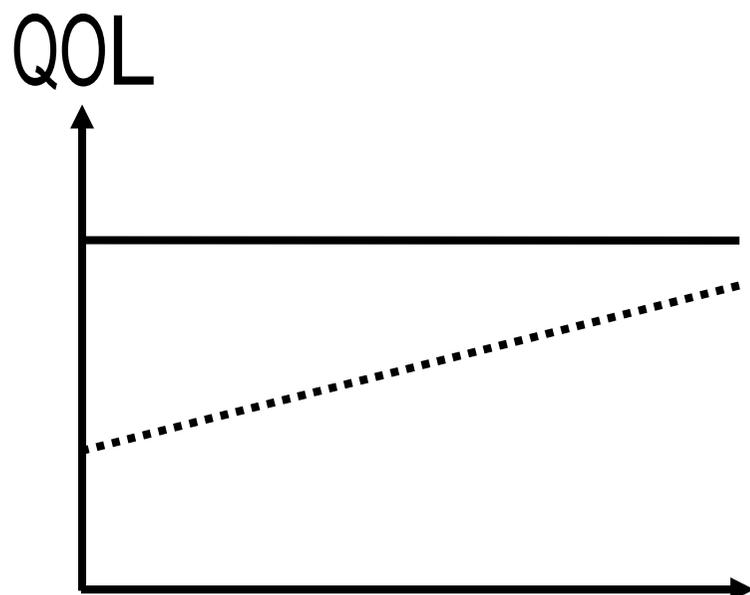
- ▶ 左下の図は以下の特徴がある 量的な交互作用と呼ぶ
 - ▶ 「薬剤×罹病期間」の交互作用がある
 - ▶ 罹病期間の値が小さいほど，薬剤 A の平均値の方が高い
 - ▶ 罹病期間の値によって薬剤間の平均値の差が異なる





交互作用がある状態（●，○：QOLの平均値）

- ▶ 右下の図は以下の特徴がある 質的な交互作用と呼ぶ
 - ▶ 「薬剤×罹病期間」の交互作用がある
 - ▶ 罹病期間の値が小さい：薬剤 A の方が高い，大きい：薬剤 B の方が高い
- 罹病期間によって薬剤間の平均値の差が異なる上，逆転現象が起こっている





交互作用がある状態

- ▶ 前頁の図はいずれも「薬剤×罹病期間」の交互作用がある状態
罹病期間の値によって薬剤間の平均値の差が異なる

「薬剤」と「罹病期間」が互いに影響を及ぼし合っているため

左図：QOL の差の大小はあれど、罹病期間の値が大きい場合も小さい場合も薬剤 A の平均値の方が高い（大小関係の逆転は起こっていない）

この状態を「量的な交互作用あり」の状態と呼ぶ

右図：QOL の差の違いがあり、かつ罹病期間の値によって大小関係の逆転が起こっている

この状態を「質的な交互作用あり」の状態と呼ぶ

- ▶ 2つの要因の間に交互作用がある場合は「薬剤」と「罹病期間」の両方を考慮して結果の解釈をする必要がある



ある因子が効果修飾因子かどうかの判定方法

興味のある因子が薬剤、「罹病期間」が効果修飾因子かどうかを判定する

「薬剤」と「罹病期間」の交互作用があるかどうかを判定する場合

2. 以下のモデルで共分散分析し、交互作用項の効果が0でない場合、

「罹病期間」は効果修飾因子

▶ 「薬剤＋罹病期間＋薬剤×罹病期間」のモデル：

$$QOL = \beta_0 + \beta_1 \times \text{薬剤} + \beta_2 \times \text{罹病期間} + \beta_3 \times \text{薬剤} \times \text{罹病期間}$$



ある因子が効果修飾因子かどうかの判定方法

```
> result <- lm(QOL ~ GROUP*DURATION, data=AB) # 交互作用モデル (薬剤 × 罹病期間)
> Anova(result, Type="II") # 分散分析表 (Type II 平方和)
```

Response: QOL

	Sum Sq	Df	F value	Pr(>F)	
GROUP	14.80	1	1.4008	0.244346	
DURATION	116.63	1	11.0364	0.002058	**
<u>GROUP:DURATION</u>	<u>67.93</u>	<u>1</u>	<u>6.4284</u>	<u>0.015716</u>	<u>*</u>
Residuals	380.44				

- ▶ 交互作用項の（平均）平方和の検定の p 値 = 0.0157

検定結果が 5% よりも小さいので「交互作用あり」



本日のメニュー

1. 分散分析

- ▶ イントロ
- ▶ データ「DEP」による例示

2. 共分散分析

3. おまけ：平方和の Type について



【おさらい】 交絡の有無の判定方法

興味のある因子が薬剤，「前治療の有無」が交絡因子かどうかを判定した際

- ▶ 以下のモデルで回帰分析し，分散分析表から交絡の有無を判定した
 - ▶ 「薬剤のみ」のモデル： $QOL = \beta_0 + \beta_1 \times \text{薬剤}$
 - ▶ 「薬剤＋性別」のモデル： $QOL = \beta_0 + \beta_1 \times \text{薬剤} + \beta_2 \times \text{前治療の有無}$

```
> result <- lm(QOL ~ GROUP, data=AB)
> Anova(result, Type="II")
Response: QOL
      Sum Sq Df F value  Pr(>F)
GROUP    62.5  1  4.2035 0.04728 *
Residuals 565.0 38
```

薬剤のみのモデル
分散分析表 (Type II 平方和)

Type II って何?

```
> result <- lm(QOL ~ GROUP+PREDRUG, data=AB)
> Anova(result, Type="II")
Response: QOL
      Sum Sq Df F value  Pr(>F)
GROUP    0.0  1  0.000 1.0000000
PREDRUG  187.5  1 18.378 0.0001243 ***
Residuals 377.5 37
```

薬剤＋前治療の有無のモデル
分散分析表 (Type II 平方和)

Type II って何?



Type I 平方和と Type II 平方和の違い

- ▶ 本資料では一貫して「Type II 平方和」を用いて解釈を行ったが、他に「Type I 平方和」「Type III 平方和」「Type IV 平方和」等がある
- ▶ 例として前頁のモデルを用いる
 - ▶ 「薬剤＋前治療の有無」のモデル： $QOL = \beta_0 + \beta_1 \times \text{薬剤} + \beta_2 \times \text{前治療の有無}$
 - ▶ 「前治療の有無＋薬剤」のモデル： $QOL = \beta_0 + \beta_1 \times \text{前治療の有無} + \beta_2 \times \text{薬剤}$
- ▶ Type I 平方和：
 - ▶ 普通の統計の教科書で出てくる計算方法
 - ▶ 最初に計算した平方和が大きくなる（p 値が小さくなる）傾向があり、「モデルに指定した因子の順番」で結果（平方和や p 値）が変わってしまう
- ▶ Type II 平方和：
 - ▶ 少し難しめの統計の本で出てくる計算方法
 - ▶ 平方和は、計算の順番（モデルに指定した順番）によらない



Type I 平方和

```
> result <- lm(QOL ~ GROUP+PREDRUG, data=AB) # 薬剤 + 前治療の有無のモデル
> anova(result) # 分散分析表 (Type I 平方和)
Response: QOL
      Df Sum Sq Mean Sq F value Pr(>F)
GROUP  1  62.5    62.500   6.1258 0.0180277 *
PREDRUG 1 187.5   187.500  18.3775 0.0001243 ***
Residuals 37 377.5   10.203

> result <- lm(QOL ~ PREDRUG+GROUP, data=AB) # 前治療の有無 + 薬剤のモデル
> anova(result) # 分散分析表 (Type I 平方和)
Response: QOL
      Df Sum Sq Mean Sq F value Pr(>F)
PREDRUG 1 250.0   250.000  24.503 1.646e-05 ***
GROUP  1   0.0    0.000   0.000 1
Residuals 37 377.5   10.203
```

- ▶ 「薬剤」と「前治療の有無」の順番を入れ替えると結果が変わる・・・
- ▶ モデルに指定する順番で結果が変わるのはよろしくない・・・



Type II 平方和

```
> result <- lm(QOL ~ GROUP+PREDRUG, data=AB) # 薬剤 + 前治療の有無のモデル
> Anova(result, Type="II") # 分散分析表 (Type II 平方和)
Response: QOL
      Sum Sq Df F value    Pr(>F)
GROUP      0.0  1  0.000 1.0000000
PREDRUG  187.5  1 18.378 0.0001243 ***
Residuals 377.5 37
> result <- lm(QOL ~ PREDRUG+GROUP, data=AB) # 前治療の有無 + 薬剤のモデル
> Anova(result, Type="II") # 分散分析表 (Type II 平方和)
Response: QOL
      Sum Sq Df F value    Pr(>F)
PREDRUG  187.5  1 18.378 0.0001243 ***
GROUP      0.0  1  0.000 1.0000000
Residuals 377.5 37
```

- ▶ 「薬剤」と「前治療の有無」の順番を入れ替えても結果は変わらない
- ▶ この理由（モデルに指定する順番で結果が変わらない）により、本資料では **Type II** 平方和を用いています



本日のメニュー

1. 分散分析
 - ▶ イントロ
 - ▶ データ「DEP」による例示
2. 共分散分析
3. おまけ：平方和の `Type` について



参考文献

- ▶ 統計学（白旗 慎吾 著，ミネルヴァ書房）
- ▶ R によるやさしい統計学（山田 剛史 他著，オーム社）
- ▶ ロスマンの疫学（Kenneth J. Rothman 著，矢野 栄二 他翻訳，篠原出版新社）
- ▶ The R Tips 第 2 版（オーム社）
- ▶ R 流！イメージで理解する統計処理入門（カットシステム）

Rで統計解析入門

終