Rで統計解析入門

(7) 調整済み平均値

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/fkhud708/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

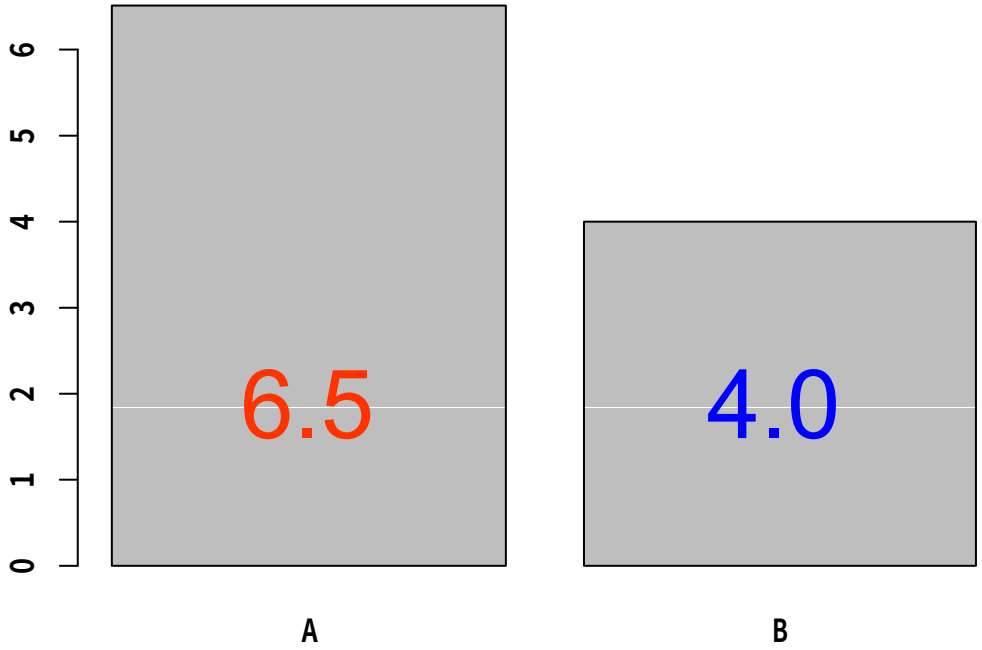
本日のメニュー

1. 調整済み平均値
   - イントロ
   - 薬剤と前治療の有無（カテゴリ変数）の場合
   - 薬剤と罹病期間（連続変数）の場合
2. 傾向スコア

【おさらい】QOL の平均値の比較

- 「薬剤ごとの QOL の平均」に関するグラフを描く

```
> MEAN <- by(AB$QOL, AB$GROUP, mean)  # 各薬剤の平均値を算出
> barplot(MEAN)                       # 平均値の棒グラフ
```



【おさらい】QOL の平均値の比較

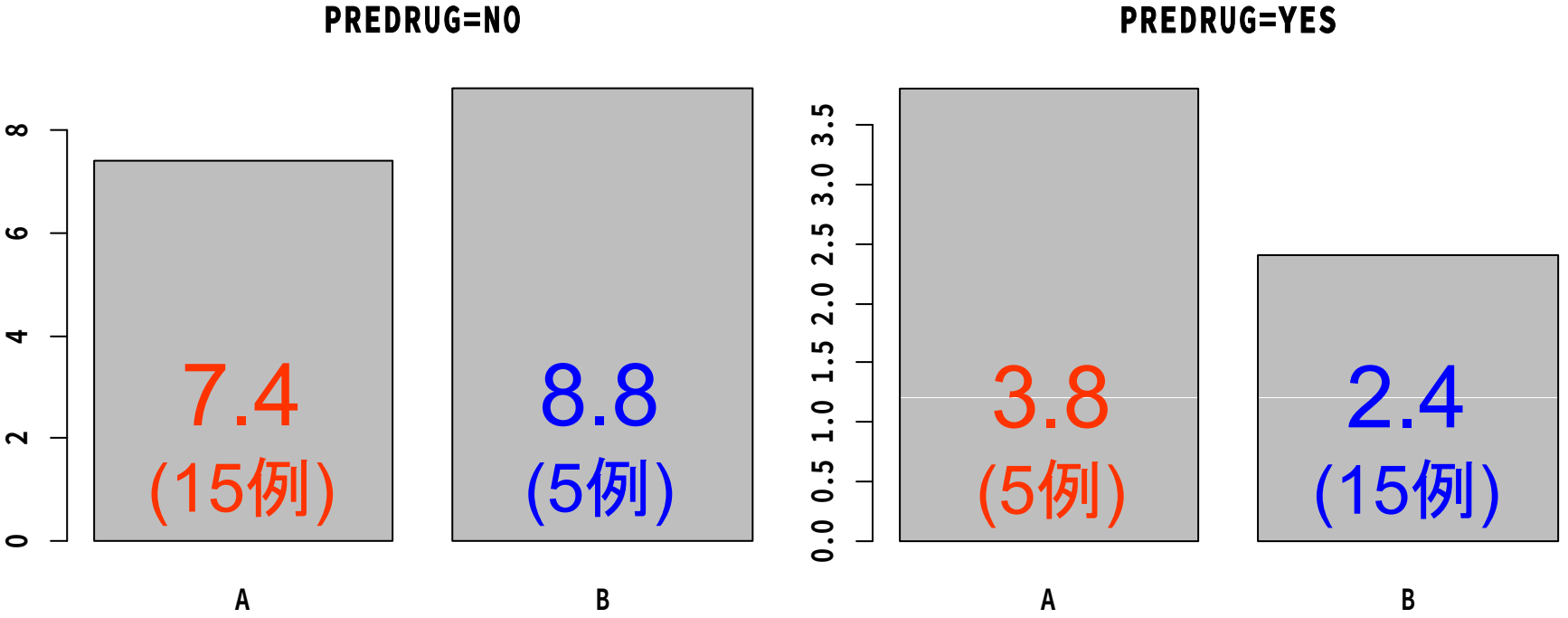
- 「薬剤 A の QOL スコアの平均」と「薬剤 B の QOL スコアの平均」が等しいかどうかを検定する
  - 検定結果は p = 4.7%（有意水準 5% ） ⇒「QOL の平均は等しくない」
  - 平均は 薬剤 A の方が高い

```
> t.test(QOL ~ GROUP, data=AB, var=T)

        Two Sample t-test

data:  QOL by GROUP
t = 2.0503, df = 38, p-value = 0.04728 ← 検定結果（ p 値 = 約 4.7 %）
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.031532 4.968468
sample estimates:
mean in group A mean in group B
            6.5             4.0
```

【おさらい】薬剤と前治療の有無の関係

- 前治療の有無別に「薬剤ごとの QOL の平均」の棒グラフを描く

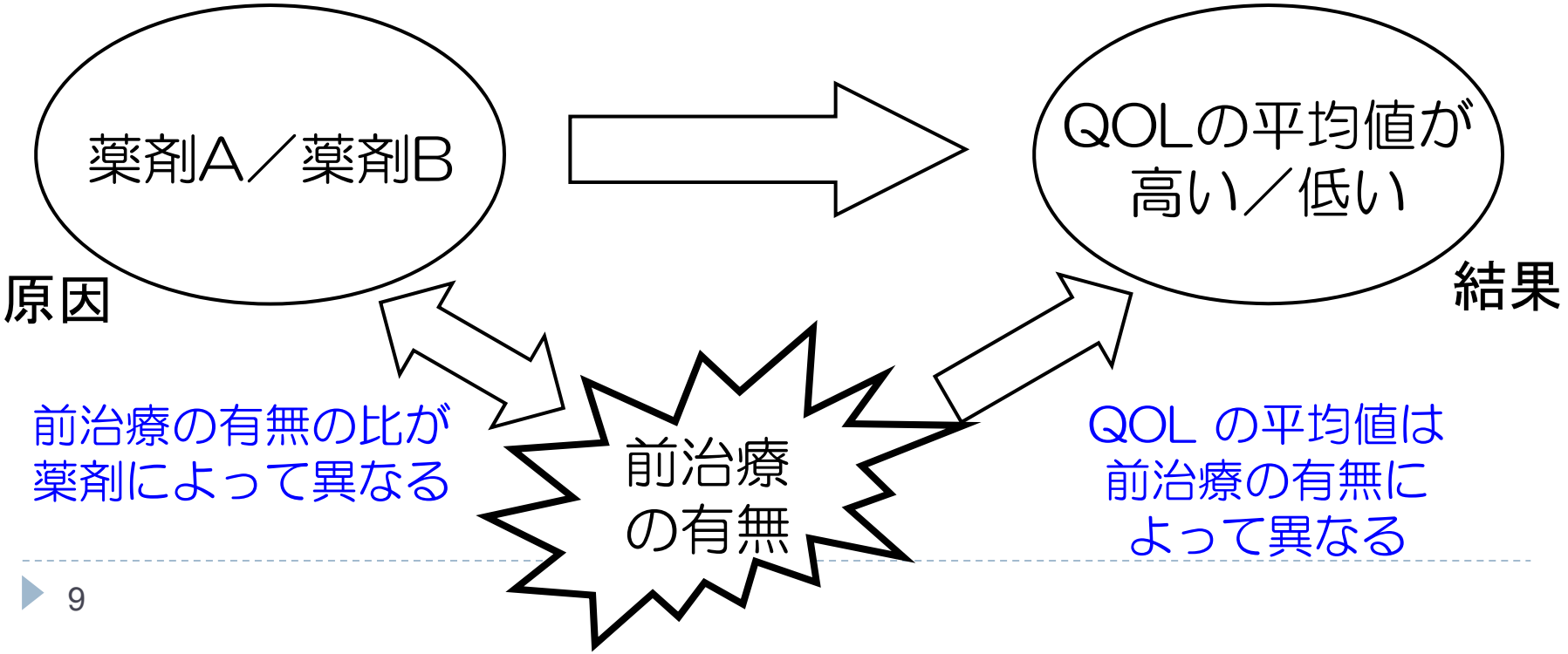
```
> MEAN2 <- tapply(AB$QOL, AB[,c("GROUP","PREDRUG")], mean)
> barplot(MEAN2[,1], main="PREDRUG=NO")     # 前治療なし
> barplot(MEAN2[,2], main="PREDRUG=YES")    # 前治療あり
```



【おさらい】薬剤と前治療の有無の関係

- 前治療の有無の例数の比が薬剤によって異なる：
  薬剤 A の前治療なし：あり = 3：1，薬剤 B の比 = 1：3
- QOL の平均値が前治療の有無によって異なる

⇒ QOL の平均値に影響している「前治療の有無」という要因を無視して（前治療の有無をまとめて全体だけで）解釈をするとおかしな結論に



【おさらい】薬剤と前治療の有無の関係

	QOL の平均値	例数
A	6.5	20
B	4.0	20

前治療なし

	平均値	例数
A	7.4	15
B	8.8	5

前治療あり

	平均値	例数
A	3.8	5
B	2.4	15

- 「薬剤 A のなし：あり　3：1」　≠「薬剤 B のなし：あり　1：3」
- 「前治療なしの QOL の平均値の差」≠「前治療ありの QOL の平均値の差」
  ⇒ 交絡が起きているっぽい ⇒ 一応，回帰分析でも確かめる

【おさらい】薬剤と前治療の有無の関係

```
> 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 *

> result <- lm(QOL ~ GROUP+PREDRUG, data=AB)            # 薬剤＋前治療の有無のモデル
> summary(result)                                       # 結果の要約を表示
Coefficients:
              Estimate  Std. Error t value Pr(>|t|)
(Intercept)  7.750e+00   1.129e+00   6.863 4.32e-08 ***
GROUPA       8.654e-16※  1.166e+00   0.000 1.000000
PREDRUGYES  -5.000e+00   1.166e+00  -4.287 0.000124 ***
```

- 薬剤のみのモデル　　：群間差 = 2.500
- 薬剤+前治療の有無のモデル：群間差 = ほぼ 0 ⇒傾きが変わっている
- 交絡が起きている（前治療の有無は交絡因子）

※ 8.654e-16 = 8.654×10^{-16}　　0

交絡が起きた原因

- 「全体」「前治療薬なし」「前治療薬あり」の結果が全て同じであれば気持ちが良い結果だが，実際はそうなっていない（交絡が起きている）
  - 全体：薬剤 B に比べて薬剤 A の方が QOL が高い
  - 前治療薬がない患者さん：薬剤 A に比べて薬剤 B の方が QOL が高い
  - 前治療薬がある患者さん：薬剤 B に比べて薬剤 A の方が QOL が高い
- 原因は「薬剤間の前治療薬の有無の割合の不均衡」
  - 前治療薬がない患者さん：薬剤 A にとって不利，薬剤 B にとって有利
  - 薬剤 B にとっては，前治療薬がない患者さんが少なくなると不利になる
- 「前治療薬の有無の割合が薬剤間で等しい」場合は「全体」「前治療薬なし」「前治療薬あり」の結果が全て同じとなるが，「前治療薬の有無の割合が薬剤間で異なる」場合は，割合によって「効果がない薬剤」なのに効果がある様に見えてしまう場合がある…

【参考】不均衡がなければ交絡は起きない

	QOL の平均値	例数
A	5.6	10
B	5.6	10

前治療なし

	平均値	例数
A	7.4	5
B	8.8	5

前治療あり

	平均値	例数
A	3.8	5
B	2.4	5

- 「薬剤 A のなし：あり」も「薬剤 B のなし：あり」も 1：1 にしてみる
- 全体の QOL の平均値が同じになった！⇒ 交絡が起きてない

交絡の影響をかわしながらデータを解釈する

- 交絡の影響をかわしながらデータを解釈する方法は
  とりあえず以下の 2 つ
  - 前治療薬の有無ごとに結果を出す
  - 調整済み平均値（Least Square Means：LS Means）で解釈する
- 前治療薬の有無ごとに結果を出すのは済んでいるので，
  以下では調整済み平均値（LS Means）を計算する方法を
  紹介

本日のメニュー

1. 調整済み平均値
   - イントロ
   - 薬剤と前治療の有無（カテゴリ変数）の場合
   - 薬剤と罹病期間（連続変数）の場合
2. 傾向スコア

調整済み平均値（調整因子：カテゴリ変数）

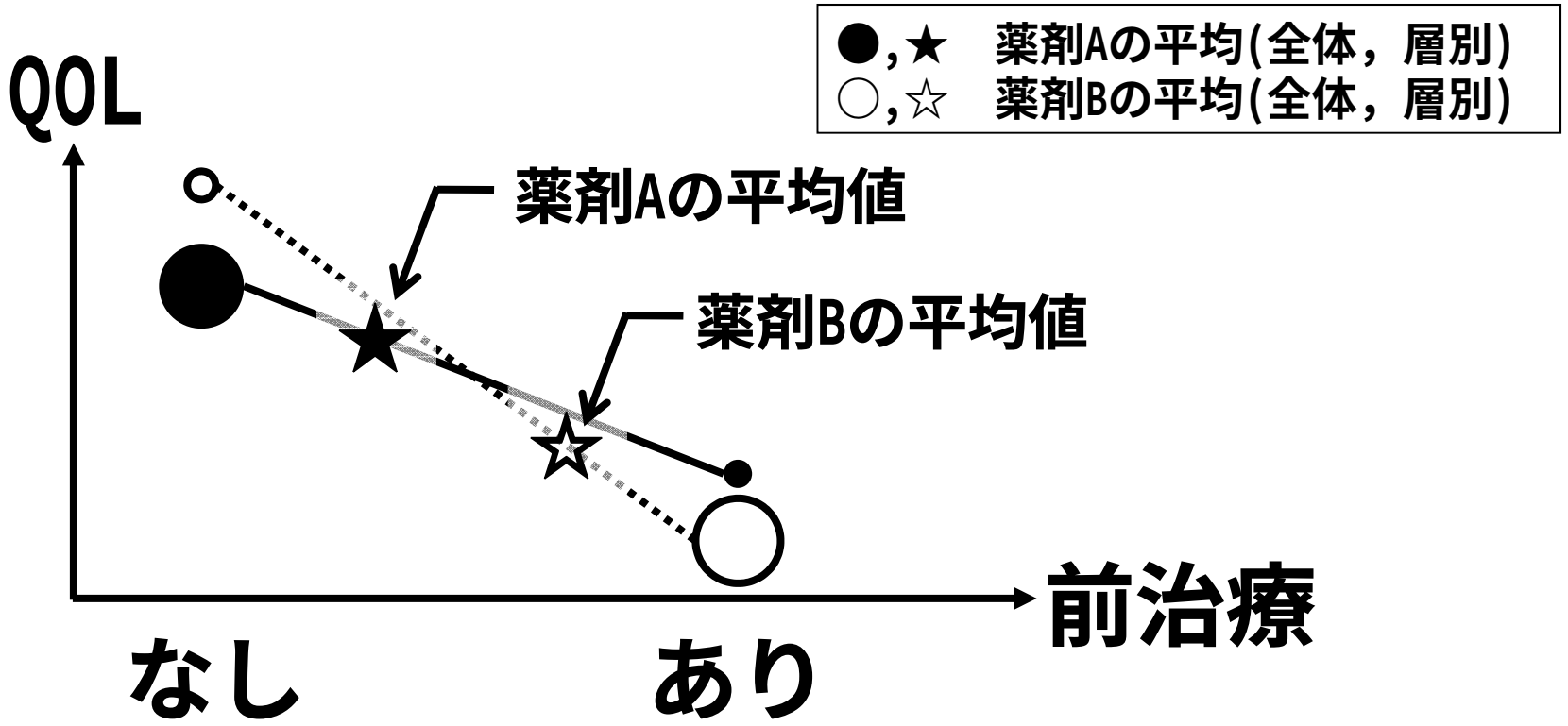
- 問題となっているのが「薬剤間の前治療薬の有無の割合の不均衡」
- 「前治療薬の有無の割合の不均衡」がある状態で単純に平均値を求める（重みづけ平均にする）と・・・
  - 薬剤 A の QOL の平均値 ＝ (7.4 × 15 ＋ 3.8 × 5) ÷ 20 ＝ 6.5
  - 薬剤 B の QOL の平均値 ＝ (8.8 × 5 ＋ 2.4 × 15) ÷ 20 ＝ 4.0

⇒「前治療薬の有無の割合の不均衡」をモロに受けてしまう･･･（交絡の原因）

QOL	薬剤 A	薬剤 B
前治療薬なしの平均	7.4（n = 15）	8.8（n = 5）
前治療薬ありの平均	3.8（n = 5）	2.4（n = 15）
全体の平均	6.5（n = 20）	4.0（n = 20）

調整済み平均値（調整因子：カテゴリ変数）

- 「前治療薬の有無の割合の不均衡」があるのに単純に平均値を求めると
  - 薬剤 A の平均値：割合が大きい「前治療なし」の平均値に引っ張られる
  - 薬剤 B の平均値：割合が大きい「前治療あり」の平均値に引っ張られる



調整済み平均値（調整因子：カテゴリ変数）

- 「前治療薬の有無の割合の不均衡」の影響をかわすため，重みなし平均
  ：（前治療薬なしの平均値 ＋ 前治療薬ありの平均値）÷ 2 を求める
  - 薬剤 A の QOL の平均値 ＝ (7.4 × 1 ＋ 3.8 × 1) ÷ 2 ＝ 5.6
  - 薬剤 B の QOL の平均値 ＝ (8.8 × 1 ＋ 2.4 × 1) ÷ 2 ＝ 5.6

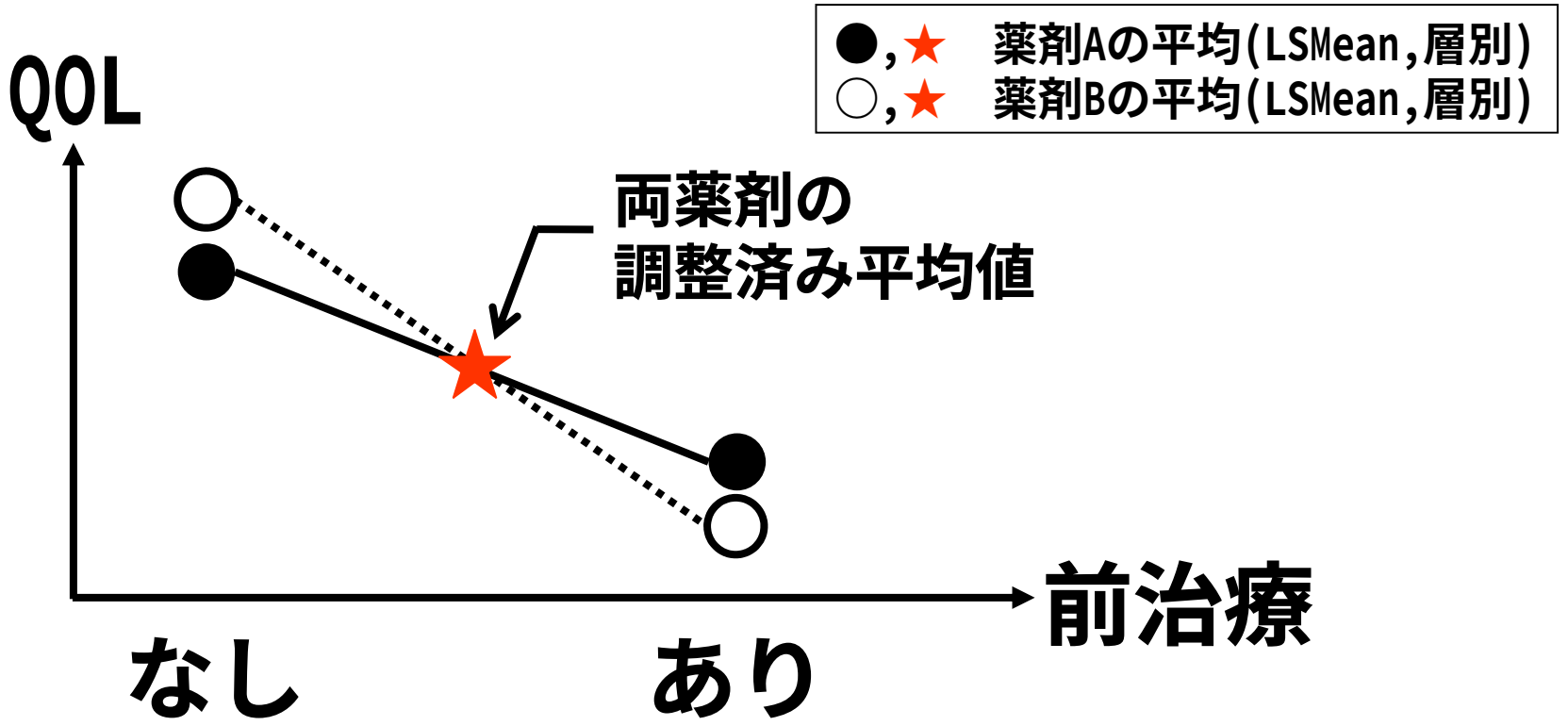
  ⇒「前治療薬の有無の割合の不均衡」の影響を「ある程度」かわすことが出来る
- ↑が調整済み平均値(LS Means)，前治療の有無を「調整因子」と呼ぶ

QOL	薬剤 A	薬剤 B
前治療薬なしの平均	7.4（n = 1）	8.8（n = 1）
前治療薬ありの平均	3.8（n = 1）	2.4（n = 1）
調整済み平均値	5.6（n = 2）	5.6（n = 2）

※ 両薬剤の調整済み平均値が一致したのは偶然です

調整済み平均値（調整因子：カテゴリ変数）

- 調整済み平均値（重みなし平均）を用いると
  - 薬剤 A の平均値：割合が大きい「前治療なし」の平均値に引っ張られることはない
  - 薬剤 B の平均値：割合が大きい「前治療あり」の平均値に引っ張られることはない



※ 両薬剤の調整済み平均値が一致したのは偶然です

調整済み平均値の算出プログラム

```
> result  <- lm(QOL ~ GROUP*PREDRUG, data=AB)
> result2 <- dummy.coef(result)
> result2                                          # 推定値（Estimate）が格納されている

Full coefficients are
(Intercept):       8.8
GROUP:               B    A
                   0.0 -1.4
PREDRUG:            NO  YES
                   0.0 -6.4
GROUP:PREDRUG:    B:NO A:NO B:YES A:YES
                   0.0  0.0   0.0   2.8

> result2$"(Intercept)"  +                       # 切片に関する推定値
+ result2$GROUP          +                       # 薬剤に関する推定値
+ mean(result2$PREDRUG)  +                       # 前治療の有無の推定値の重みなし平均値
+ c(mean(result2$"GROUP:PREDRUG"[c(1,3)]),       # 薬剤×前治療の有無の重みなし平均値（薬剤A）
+   mean(result2$"GROUP:PREDRUG"[c(2,4)]))       # 薬剤×前治療の有無の重みなし平均値（薬剤B）

  B   A
5.6 5.6
```

回帰モデルを用いた調整済み平均値の算出

```
> result  <- lm(QOL ~ GROUP*PREDRUG, data=AB)
> result

Coefficients:
     (Intercept)            GROUPA          PREDRUGYES  GROUPA:PREDRUGYES
             8.8              -1.4              -6.4                 2.8
```

- 以下のモデル「薬剤＋前治療の有無＋薬剤×前治療の有無」に対して分析を行う（薬剤：*A* は 1，B は 0，前治療：なしは 0，ありは 1）：

  $QOL = \beta_0 + \beta_1 \times$薬剤 $+ \beta_2 \times$前治療の有無$+ \beta_3 \times$薬剤×前治療の有無

	[切片]	[薬剤]	[前治療]	[薬剤×前治療]	
薬剤*A*＆前治療薬なしのQOL =	8.8	−1.4	+0.0	+0.0	= 7.4
薬剤*A*＆前治療薬ありのQOL =	8.8	−1.4	−6.4	+2.8	= 3.8
薬剤B＆前治療薬なしのQOL =	8.8	+0.0	+0.0	+0.0	= 8.8
薬剤B＆前治療薬ありのQOL =	8.8	+0.0	−6.4	+0.0	= 2.4

※ 薬剤: *A* は -1.4×1=-1.4，B は -1.4×0=0，前治療: なしは -6.4×0=0，ありは -6.4×1=-6.4
薬剤×前治療: 薬剤 *A* かつ前治療ありは 2.8×1=2.8，それ以外は 2.8×0=0

回帰モデルを用いた調整済み平均値の算出

	[切片]	[薬剤]	[前治療]	[薬剤×前治療]	
薬剤A＆前治療薬なしのQOL =	8.8	−1.4	+0.0	+0.0	= 7.4
薬剤A＆前治療薬ありのQOL =	8.8	−1.4	−6.4	+2.8	= 3.8
薬剤B＆前治療薬なしのQOL =	8.8	+0.0	+0.0	+0.0	= 8.8
薬剤B＆前治療薬ありのQOL =	8.8	+0.0	−6.4	+0.0	= 2.4

- 上記の回帰式（値）は薬剤別・前治療薬の有無別の「QOL の平均値」と一致する！

QOL	薬剤 A	薬剤 B
前治療薬なしの平均	7.4（n = 15）	8.8（n = 5）
前治療薬ありの平均	3.8（n = 5）	2.4（n = 15）
全体の平均	6.5（n = 20）	4.0（n = 20）

※ 薬剤: A は -1.4×1=-1.4，B は -1.4×0=0，前治療: なしは -6.4×0=0，ありは -6.4×1=-6.4
薬剤×前治療: 薬剤 A かつ前治療ありは 2.8×1=2.8，それ以外は 2.8×0=0

回帰モデルを用いた調整済み平均値の算出

	[切片]	[薬剤]	[前治療]	[薬剤×前治療]	
薬剤A＆前治療薬なしのQOL =	8.8	−1.4	+0.0	+0.0	= 7.4
薬剤A＆前治療薬ありのQOL =	8.8	−1.4	−6.4	+2.8	= 3.8
薬剤B＆前治療薬なしのQOL =	8.8	+0.0	+0.0	+0.0	= 8.8
薬剤B＆前治療薬ありのQOL =	8.8	+0.0	−6.4	+0.0	= 2.4

- 薬剤 A の調整済み平均値：薬剤 A に関する上記 2 つの回帰式の平均
- 薬剤 B の調整済み平均値：薬剤 B に関する上記 2 つの回帰式の平均

QOL	薬剤 A	薬剤 B
前治療薬なしの平均	7.4（n = 1）	8.8（n = 1）
前治療薬ありの平均	3.8（n = 1）	2.4（n = 1）
調整済み平均値	5.6（n = 2）	5.6（n = 2）

回帰モデルを用いた調整済み平均値の算出

```
> result2$"(Intercept)"      # 切片の推定値
(Intercept)
        8.8
> result2$GROUP              # 薬剤の推定値：ベクトル result2$GROUP の第2成分
   B    A
 0.0 -1.4
> result2$PREDRUG            # 前治療の有無の推定値
  NO  YES
 0.0 -6.4
> result2$"GROUP:PREDRUG"    # 薬剤×前治療の有無の推定値
 B:NO  A:NO B:YES A:YES      # ：ベクトルresult2$"GROUP:PREDRUG" の第2成分と第4成分
  0.0   0.0   0.0   2.8
```

	[切片]	[薬剤]	[前治療]	[薬剤×前治療]	
薬剤A＆前治療薬なしのQOL =	8.8	−1.4	＋0.0	＋0.0	= 7.4
薬剤A＆前治療薬ありのQOL =	8.8	−1.4	−6.4	＋2.8	= 3.8

- 薬剤 A の調整済み平均値は上の 2 式の平均なので・・・
  - 切片と薬剤の推定値：そのまま足す，前治療の有無：推定値÷2 を足す
  - 薬剤×前治療の有無の推定値：(0.0 + 2.8) ÷ 2 を足す

回帰モデルを用いた調整済み平均値の算出

```
> result2$"(Intercept)"      # 切片の推定値
(Intercept)
        8.8
> result2$GROUP              # 薬剤の推定値：ベクトル result2$GROUP の第1成分
   B    A
 0.0 -1.4
> result2$PREDRUG            # 前治療の有無の推定値
  NO  YES
 0.0 -6.4
> result2$"GROUP:PREDRUG"    # 薬剤×前治療の有無の推定値
 B:NO  A:NO B:YES A:YES      # ：ベクトルresult2$"GROUP:PREDRUG" の第1成分と第3成分
  0.0   0.0   0.0   2.8
```

	[切片]	[薬剤]	[前治療]	[薬剤×前治療]	
薬剤B＆前治療薬なしのQOL =	8.8	＋0.0	＋0.0	＋0.0	= 8.8
薬剤B＆前治療薬ありのQOL =	8.8	＋0.0	－6.4	＋0.0	= 2.4

- 薬剤 *A* の調整済み平均値は上の 2 式の平均なので・・・
  - 切片と薬剤の推定値：そのまま足す，前治療の有無：推定値÷2 を足す
  - 薬剤×前治療の有無の推定値：(0.0 + 0.0) ÷ 2 なので 0（足す必要なし）

調整済み平均値の算出プログラム（再掲）

```
> result  <- lm(QOL ~ GROUP*PREDRUG, data=AB)
> result2 <- dummy.coef(result)
> result2                                          # 推定値（Estimate）が格納されている

Full coefficients are
(Intercept):       8.8
GROUP:               B    A
                   0.0 -1.4
PREDRUG:            NO  YES
                   0.0 -6.4
GROUP:PREDRUG:    B:NO A:NO B:YES A:YES
                   0.0  0.0   0.0   2.8

> result2$"(Intercept)"  +                       # 切片に関する推定値
+ result2$GROUP          +                       # 薬剤に関する推定値
+ mean(result2$PREDRUG)  +                       # 前治療の有無の推定値の重みなし平均値
+ c(mean(result2$"GROUP:PREDRUG"[c(2,4)]),       # 薬剤×前治療の有無の重みなし平均値（薬剤A）
+   mean(result2$"GROUP:PREDRUG"[c(1,3)]))       # 薬剤×前治療の有無の重みなし平均値（薬剤B）

  B   A
5.6 5.6
```

各薬剤の
調整済み平均値

続・回帰モデルを用いた調整済み平均値の算出

```
> result  <- lm(QOL ~ GROUP + PREDRUG, data=AB)
> round(result$coefficients, 2)

(Intercept)      GROUPA  PREDRUGYES
       7.75        0.00       -5.00
```

- 交互作用項を抜いたモデル「薬剤＋前治療の有無」により調整済み平均値を算出する（薬剤：A は 1，B は 0，前治療：なしは 0，ありは 1）：

  $QOL = \beta_0 + \beta_1 \times$薬剤 $+ \beta_2 \times$前治療の有無

	[切片]	[薬剤]	[前治療]	
薬剤A＆前治療薬なしのQOL =	7.75	＋0.00	＋0.00	= 7.75
薬剤A＆前治療薬ありのQOL =	7.75	＋0.00	－5.00	= 2.75
薬剤B＆前治療薬なしのQOL =	7.75	＋0.00	＋0.00	= 7.75
薬剤B＆前治療薬ありのQOL =	7.75	＋0.00	－5.00	= 2.75

※ 薬剤: A も B も 0，前治療: なしは -5.00×0=0，ありは -5.00×1=-5.00

続・調整済み平均値の算出プログラム

```
> result  <- lm(QOL ~ GROUP + PREDRUG, data=AB)
> round(result$coefficients, 2)
(Intercept)      GROUPA  PREDRUGYES
       7.75        0.00       -5.00
> result2 <- dummy.coef(result)
> result2                                          # 推定値（Estimate）が格納されている
Full coefficients are

(Intercept):            7.75
GROUP:                     B            A
                0.000000e+00 8.653656e-16
PREDRUG:                  NO          YES
                           0           -5
> result2$"(Intercept)"  +                         # 切片に関する推定値
+ result2$GROUP          +                         # 薬剤に関する推定値
+ mean(result2$PREDRUG)                           # 前治療の有無の推定値の重みなし平均値

   B    A
5.25 5.25
```

各薬剤の
調整済み平均値

本日のメニュー

1. 調整済み平均値
   - イントロ
   - 薬剤と前治療の有無（カテゴリ変数）の場合
   - 薬剤と罹病期間（連続変数）の場合
2. 傾向スコア

調整済み平均値（調整因子：連続変数）

```
> result  <- lm(QOL ~ GROUP*DURATION, data=AB)
> result

Coefficients:
    (Intercept)           GROUPA         DURATION  GROUPA:DURATION
         5.4872           6.2316          -0.2051          -0.8386
```

- とりあえず「薬剤＋罹病期間＋薬剤×罹病期間」のモデルに対して分析を行う（薬剤：*A* は 1，B は 0，罹病期間を *x* で表す）

  ⇒ 各薬剤の回帰式を求める際，罹病期間が連続変数なので「傾き×変数（傾き×罹病期間）」となることに注意しつつ・・・

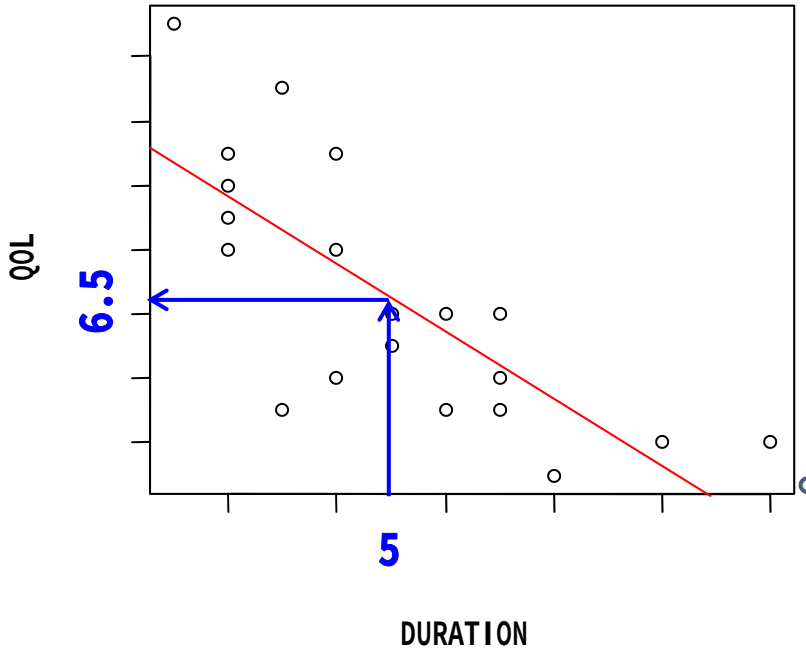
$$QOL = \beta_0 + \beta_1 \times 薬剤 + \beta_2 \times 罹病期間 + \beta_3 \times 薬剤 \times 罹病期間$$

	[切片]	[薬剤]	[罹病期間]	[薬剤×罹病期間]	
薬剤 *A* の QOL =	5.48	＋6.23	−0.20 *x*	−0.83 *x*	= 11.71−1.03 *x*
薬剤 B の QOL =	5.48	＋0.00	−0.20 *x*	−0.00 *x*	= 5.48−0.20 *x*

※ 薬剤: *A* は 6.23×1=6.23，B は 6.23×0=0，罹病期間: *A* も B も -0.20 *x*（罹病期間：*x*）
薬剤×罹病期間: 薬剤 *A* は -0.83 *x*，薬剤 B は 0

【おさらい】回帰分析：回帰式の性質

- 回帰式：QOL = 11.71 - 1.03×罹病期間（DURATION）
- 回帰式の罹病期間に「罹病期間の平均」を入れれば「QOL の平均値」が得られる
  - 罹病期間が 5 年（平均）：QOL = 11.71 - 1.03×5 = 6.5 （平均）



QOL の平均と一致

調整済み平均値（調整因子：連続変数）

- 「薬剤 *A* の回帰式」に「薬剤 *A* の罹病期間の平均値」を代入すると，回帰式から推定された QOL は薬剤 *A* の QOL の平均値と等しくなる
- そこで，以下を算出
  - 各薬剤の罹病期間の要約統計量を算出
  - ついでに全体の罹病期間の要約統計量を算出

罹病期間	薬剤 *A*	薬剤 B
各薬剤の平均	5.00	7.25
全体の平均	6.125	

前頁の要約統計量を算出するプログラム

```
> by(AB$DURATION, AB$GROUP, mean)    # 各薬剤の罹病期間の平均値を算出
AB$GROUP: B
[1] 7.25
------------------------------------------------------------
AB$GROUP: A
[1] 5
> mean(AB$DURATION)                  # 罹病期間の平均値を算出
[1] 6.125
```

罹病期間	薬剤 A	薬剤 B
各薬剤の平均	5.00	7.25
全体の平均	6.125	

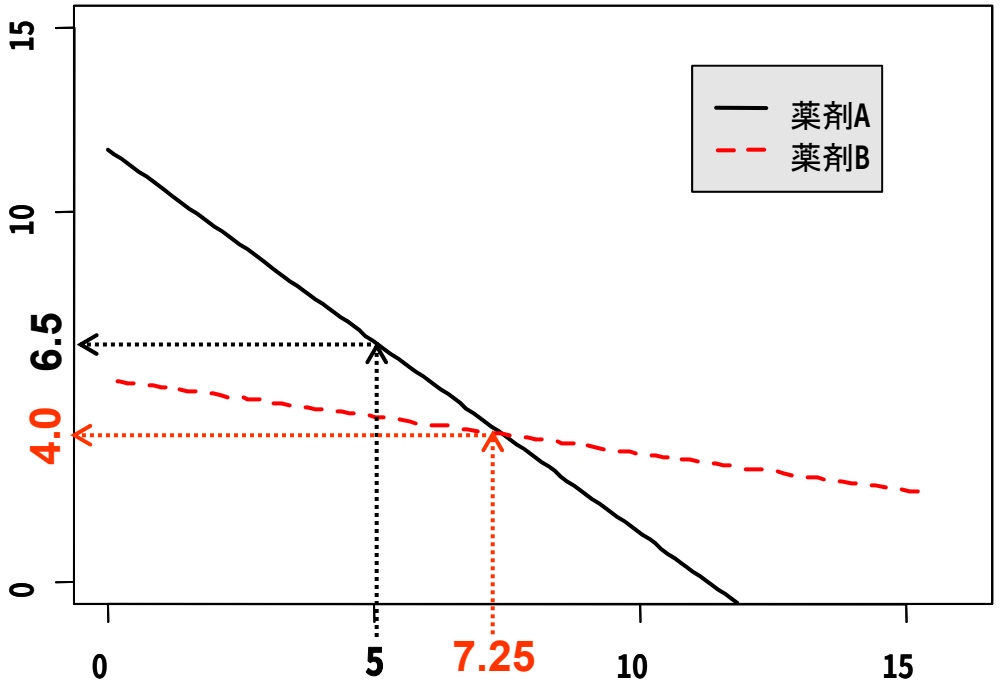
調整済み平均値（調整因子：連続変数）

薬剤 A の QOL = 11.71 − 1.03×5.00 = 6.5

薬剤 B の QOL = 5.48 − 0.20×7.25 = 4.0

- 横軸である罹病期間の平均値によって QOL の推定値が異なる

  ⇒ いずれの薬剤も，罹病期間が短い方が QOL が高くなる傾向あり

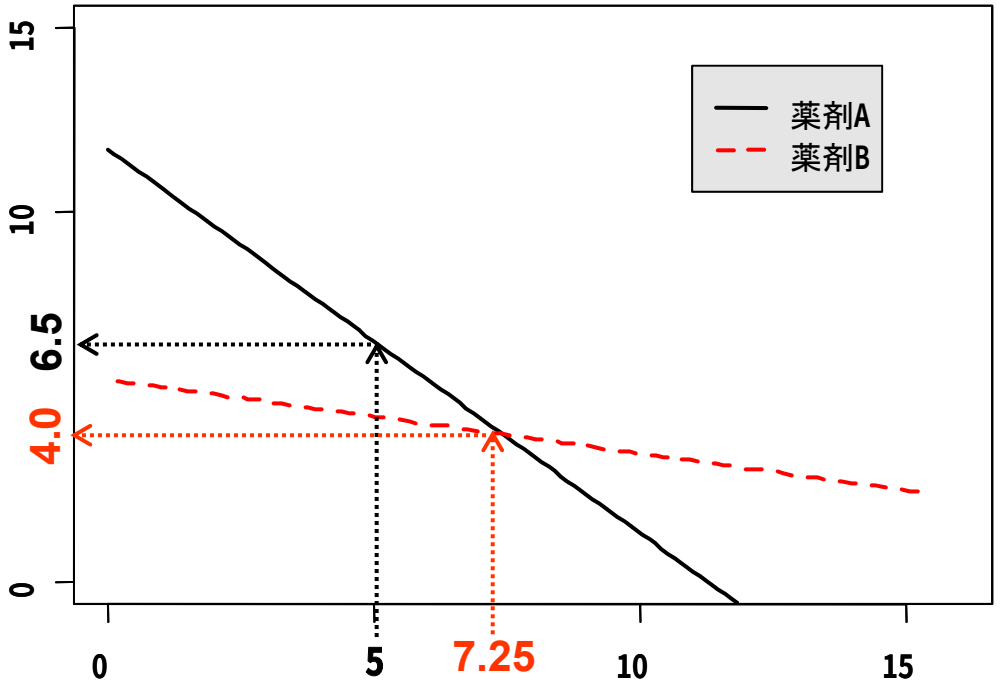


前頁のグラフを描くプログラム

```
> # 回帰直線
> A <- function(x) 11.71-1.04*x
> B <- function(x)  5.48-0.20*x
> 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")
```

調整済み平均値（調整因子：連続変数）

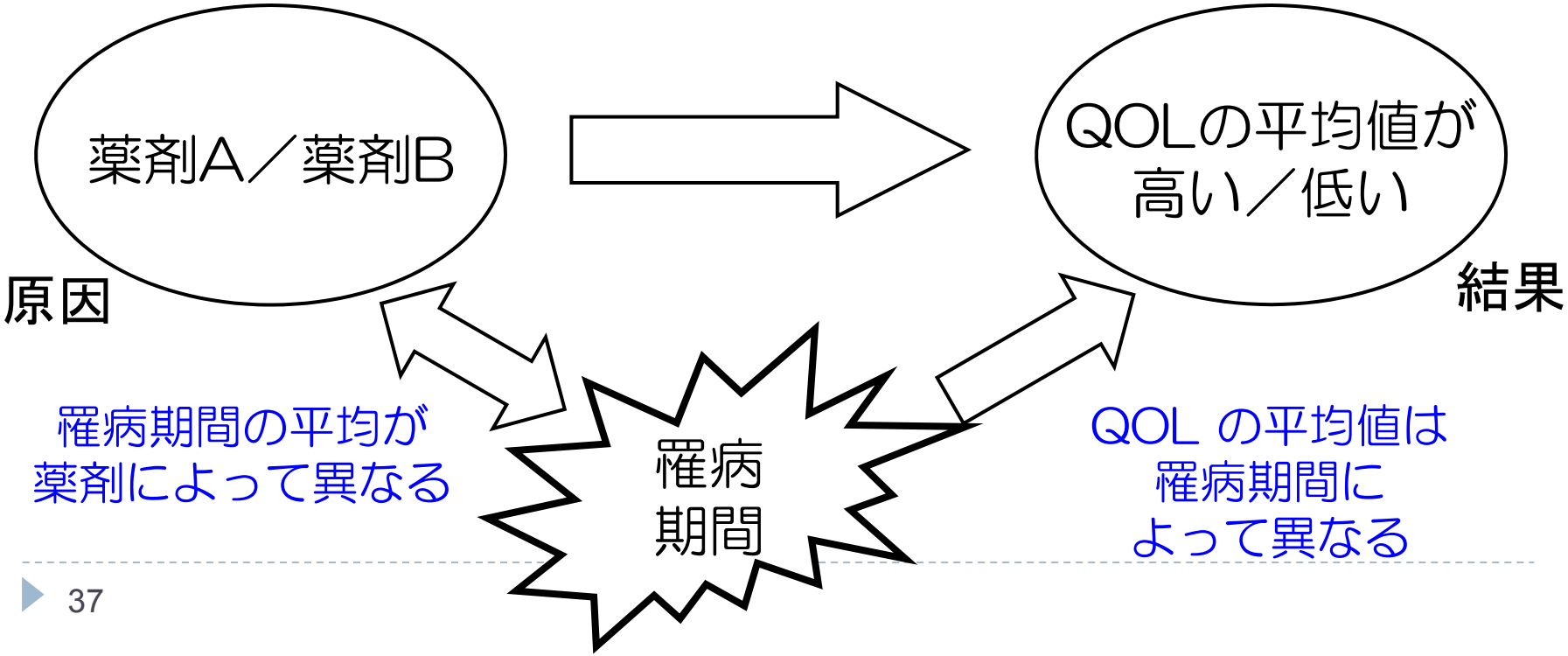
- 横軸である罹病期間の平均値によって QOL の推定値が異なる
  - 罹病期間が短い方が QOL が高くなる傾向あり
  - 薬剤 A は罹病期間の平均値が短い ⇒ QOL が高くなる傾向（有利）
  - 薬剤 B は罹病期間の平均値が長い ⇒ QOL が低くなる傾向（不利）



【おさらい】薬剤と前治療の有無の関係

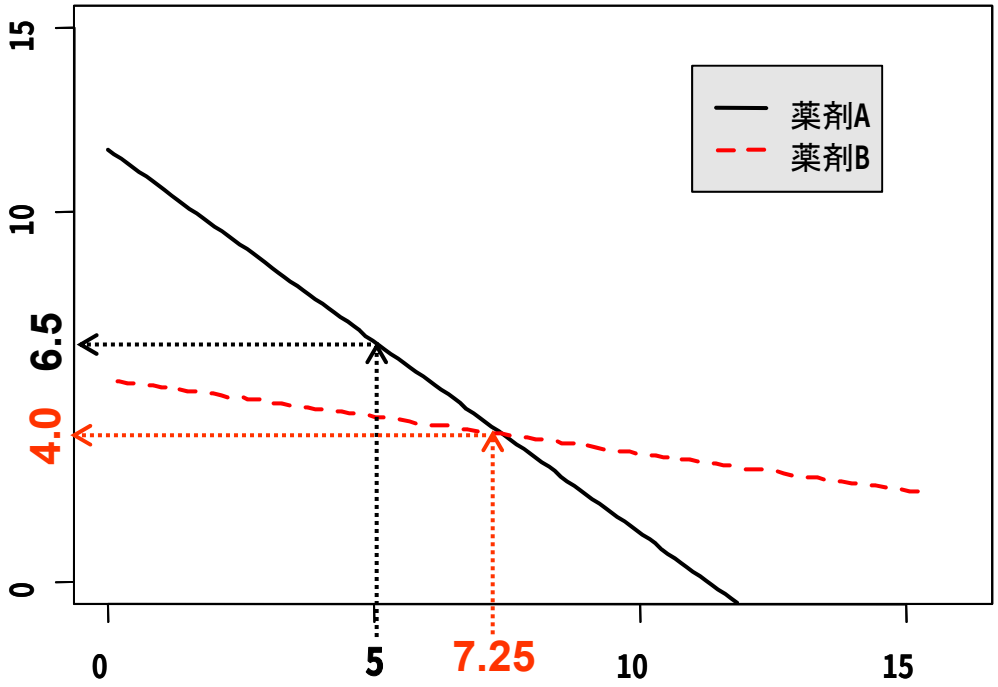
- 罹病期間の平均が薬剤によって異なる：
  薬剤 A の罹病期間の平均 = 5，薬剤 B の罹病期間の平均 = 7.25
- QOL の平均値が罹病期間によって異なる

⇒ QOL の平均値に影響している「罹病期間」という要因を無視して
（罹病期間の平均の違いを考慮せずに）解釈をするとおかしな結論に



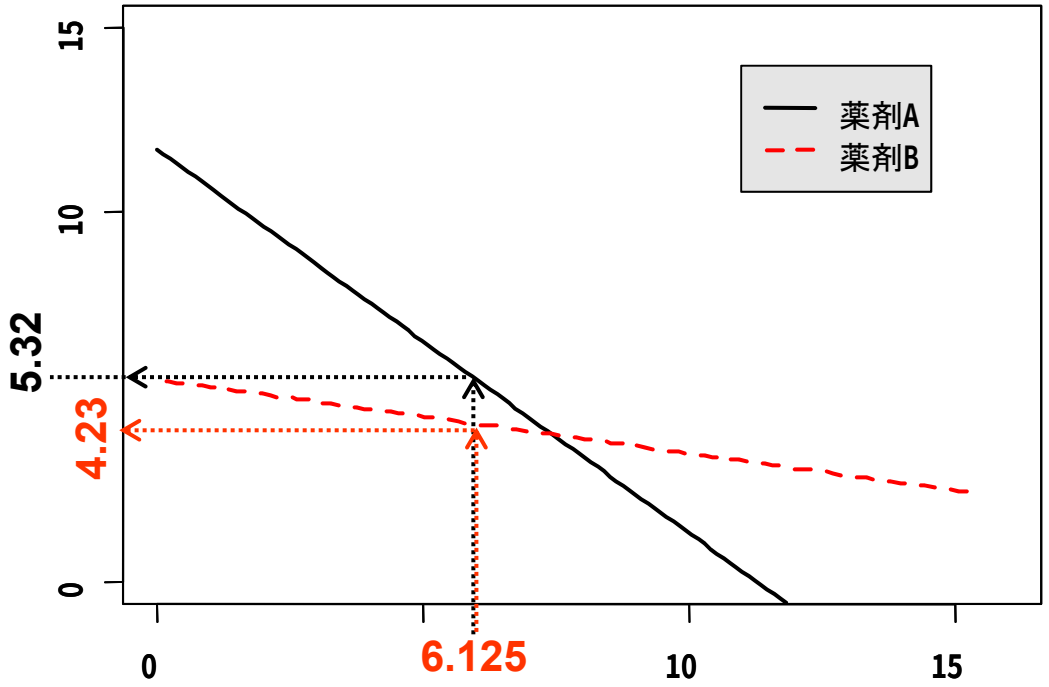
調整済み平均値（調整因子：連続変数）

- 「罹病期間が短いほど QOL は高い」という傾向がある場合，薬剤間で罹病期間の平均値がズレてしまうと，ある薬剤に有利な方に偏ってしまう場合がある
- もし「罹病期間が短くても長くても QOL の値は変わらない」場合，すなわち「回帰直線が横軸とほぼ並行」であれば「薬剤間で罹病期間の平均値がズレるとある薬剤に有利な方に偏る」ような妙な現象は起きない



調整済み平均値（調整因子：連続変数）

- 「罹病期間が短いほど QOL は高い」という傾向があるのはどうにもならない
- せめて「薬剤間での罹病期間の平均値のズレ」を何とかしたい
- 両薬剤の回帰式の罹病期間に「全体の罹病期間の平均値」を代入して計算する
- このようにして算出した QOL を調整済み平均値（LS Means）とする

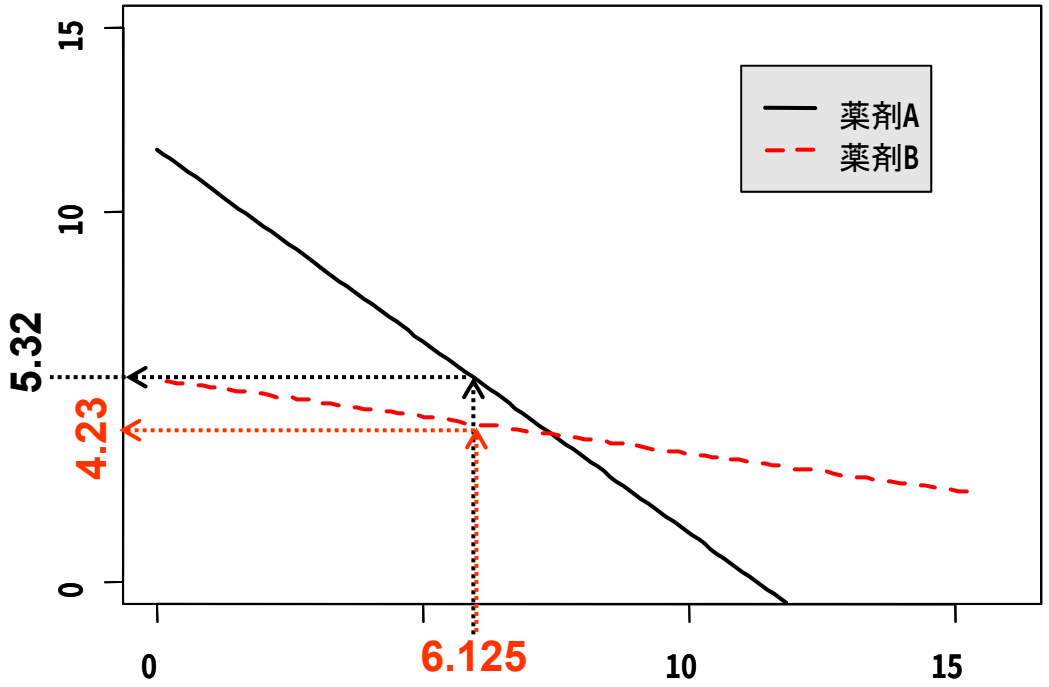


調整済み平均値（調整因子：連続変数）

薬剤 A の QOL = 11.71 − 1.03×6.125 = 5.32

薬剤 B の QOL = 5.48 − 0.20×6.125 = 4.23

- 両薬剤の回帰式の罹病期間に 6.125（全体の罹病期間の平均値）を代入し，得られた QOL の推定値を調整済み平均値とする
- ↑が調整済み平均値(LS Means)，罹病期間を「調整因子」と呼ぶ



調整済み平均値の算出プログラム

```
> result  <- lm(QOL ~GROUP*DURATION, data=AB)
> result2 <- dummy.coef(result)
> result2                                          # 推定値（Estimate）が格納されている
Full coefficients are

(Intercept):         5.487179
GROUP:                     B          A
                    0.000000   6.231571
DURATION:           -0.2051282
GROUP:DURATION:            B          A
                    0.0000000 -0.8386218
> result2$"(Intercept)"  +                        # 切片に関する推定値
+ result2$GROUP          +                        # 薬剤に関する推定値
+ result2$DURATION*mean(AB$DURATION) +            # 罹病期間の推定値の重みなし平均値
+ result2$"GROUP:DURATION"*mean(AB$DURATION)
       B        A
4.230769 5.325781
```

各薬剤の
調整済み平均値

続・調整済み平均値（調整因子：連続変数）

```
> result  <- lm(QOL ~ GROUP + DURATION, data=AB)
> result

Coefficients:
(Intercept)       GROUPA     DURATION
     7.8966       1.2907      -0.5375
```

- 交互作用項を抜いたモデル「薬剤＋罹病期間」のモデルに対して分析を行う（薬剤：A は 1，B は 0，罹病期間を x で表す）

  ⇒ 各薬剤の回帰式を求める際，罹病期間が連続変数なので「傾き×変数（傾き×罹病期間）」となることに注意しつつ・・・

  $QOL = \beta_0 + \beta_1 \times$薬剤 $+ \beta_2 \times$罹病期間

	[切片]	[薬剤]	[罹病期間]	
薬剤 A の QOL =	7.89	＋1.29	−0.53 x	= 9.18−0.53 x
薬剤 B の QOL =	7.89	＋0.00	−0.53 x	= 7.89−0.53 x

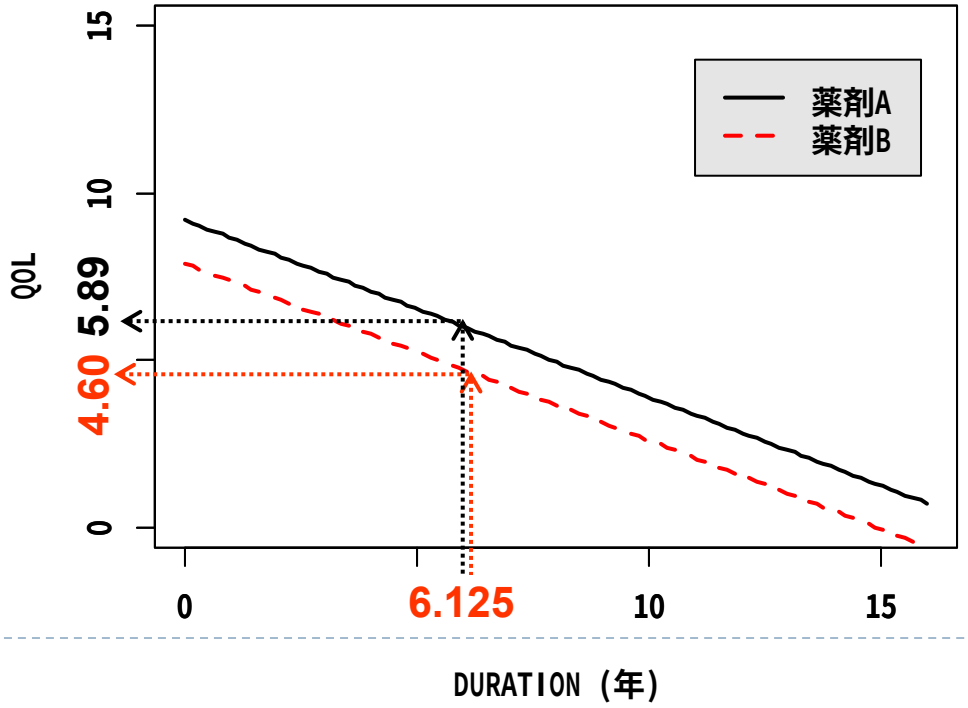
※ 薬剤: A は 1.29×1=1.29，B は 1.29×0=0，罹病期間: A も B も -0.53 x（罹病期間：x）

続・調整済み平均値（調整因子：連続変数）

薬剤 A の QOL = 9.18－0.53 x ⇒ LSMean = 9.18 － 0.53×6.125 = 5.89

薬剤 B の QOL = 7.89－0.53 x ⇒ LSMean = 7.89 － 0.53×6.125 = 4.60

- 各薬剤の直線が平行であると仮定して求めた調整済み平均値（LS Means）が得られる



前頁のグラフを描くプログラム

```
> # 回帰直線
> A <- function(x) 9.18-0.53*x
> B <- function(x) 7.89-0.53*x
> 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")
```

続・調整済み平均値の算出プログラム

```
> result  <- lm(QOL ~ GROUP+DURATION, data=AB)
> result2 <- dummy.coef(result)
> result2                                          # 推定値（Estimate）が格納されている
Full coefficients are


(Intercept):      7.896594
GROUP:                 B        A
                0.000000 1.290712
DURATION:       -0.5374613
> result2$"(Intercept)"  +                         # 切片に関する推定値
+ result2$GROUP          +                         # 薬剤に関する推定値
+ result2$DURATION*mean(AB$DURATION)               # 罹病期間の推定値の重みなし平均値
       B        A
4.604644 5.895356
```

各薬剤の
調整済み平均値

【参考】調整済み平均値と傾きの推定値

```
> result  <- lm(QOL ~ GROUP + DURATION, data=AB)
> result

Coefficients:
(Intercept)        GROUPA      DURATION
     7.8966        1.2907       -0.5375
```

- 交互作用項を抜いたモデル「薬剤＋罹病期間」のモデル：

  $QOL = \beta_0 + \beta_1 \times$薬剤 $+ \beta_2 \times$罹病期間

  から得られた調整済み平均値の薬剤間差（薬剤 A － 薬剤 B）は，

  共分散分析で得られた薬剤の傾きの値と一致する

薬剤 A の QOL = 9.18－0.53 x ⇒ LSMean = 9.18 － 0.53×6.125 = 5.89

薬剤 B の QOL = 7.89－0.53 x ⇒ LSMean = 7.89 － 0.53×6.125 = 4.60

LS Mean の差 = 1.29

【参考】交互作用項を入れる？入れない？

- 「薬剤×前治療薬の有無」の項を
  - 除いた場合：『「薬剤×前治療薬の有無」の交互作用はない』と仮定して求めた調整済み平均値（LS Means）が得られる
  - 含めた場合：『「薬剤×前治療薬の有無」の交互作用がある』と仮定して求めた調整済み平均値（LS Means）が得られる
- 「薬剤×罹病期間」の項を
  - 除いた場合：「薬剤×罹病期間の交互作用はない」と仮定（各薬剤の直線が平行であると仮定）して求めた調整済み平均値（LS Means）が得られる
  - 含めた場合：「薬剤×罹病期間」の交互作用がある」と仮定（各薬剤の直線が平行でないと仮定）して求めた調整済み平均値（LS Means）が得られる

- 調整済み平均値を求める際は交互作用項を含めないモデルを用い，その後，交互作用の有無を確認するために交互作用項を含めたモデルを用いるのが個人的な好み ⇒ 検定で言えば，主効果の平方和交互作用の分の平方和

【参考】交互作用項を入れる？入れない？

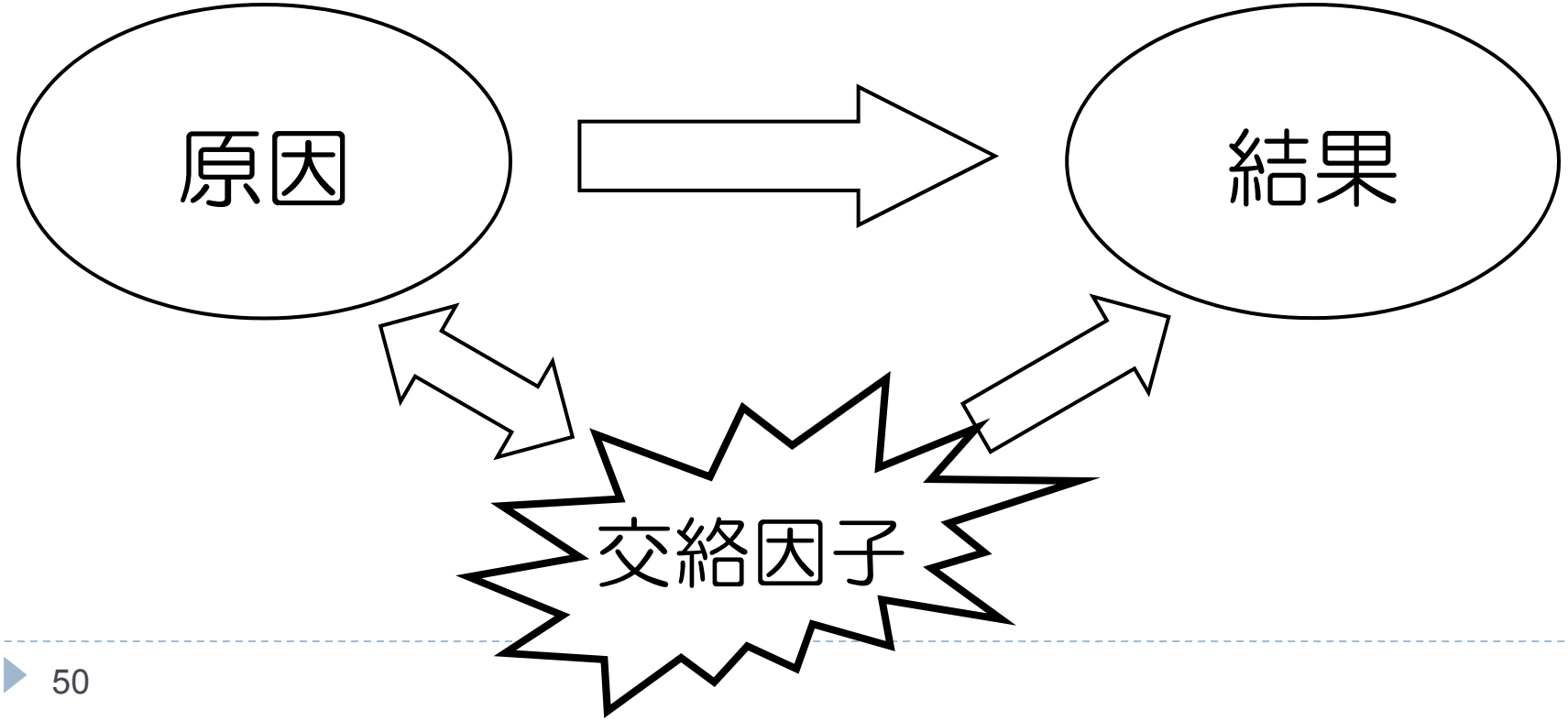
- POINTS TO CONSIDER ON ADJUSTMENT FOR BASELINE COVARIATES より
  - 主要な解析において共変量（調整因子）を入れる場合は，事前に定義しておくこと
  - 主要な解析に交互作用項は入れないこと
    （本質的な交互作用があることが事前に分かっている場合は，その因子のカテゴリ別にデータをとり，層別解析が出来るようなデザインにすること）
  - ただし，主要な解析とは別に，交互作用の有無を確認すること
    （ICH ガイドライン E9 でも推奨されている）

本日のメニュー

1. 調整済み平均値
   - イントロ
   - 薬剤と前治療の有無（カテゴリ変数）の場合
   - 薬剤と罹病期間（連続変数）の場合
2. 傾向スコア

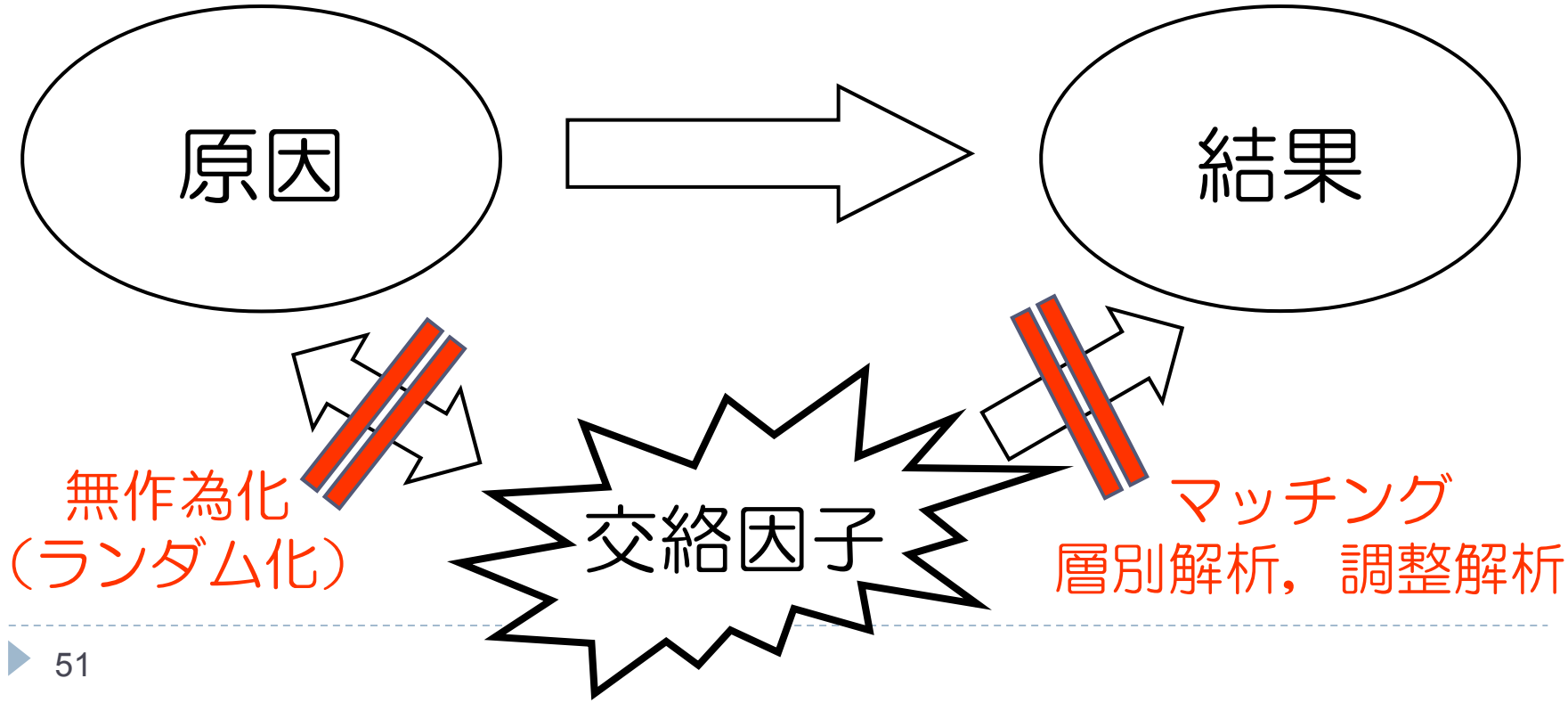
【おさらい】交絡と交絡因子

- 交絡：原因の結果への影響を調べる際，この 2 つの両方に影響を及ぼす因子があるため原因と結果の関係が正しく解釈できない状態
- 交絡因子：原因と結果の両方に影響を及ぼす因子
  ⇒ 原因と結果の関係（因果関係）が正しく解釈できない要因になりえる



交絡の影響をかわす方法

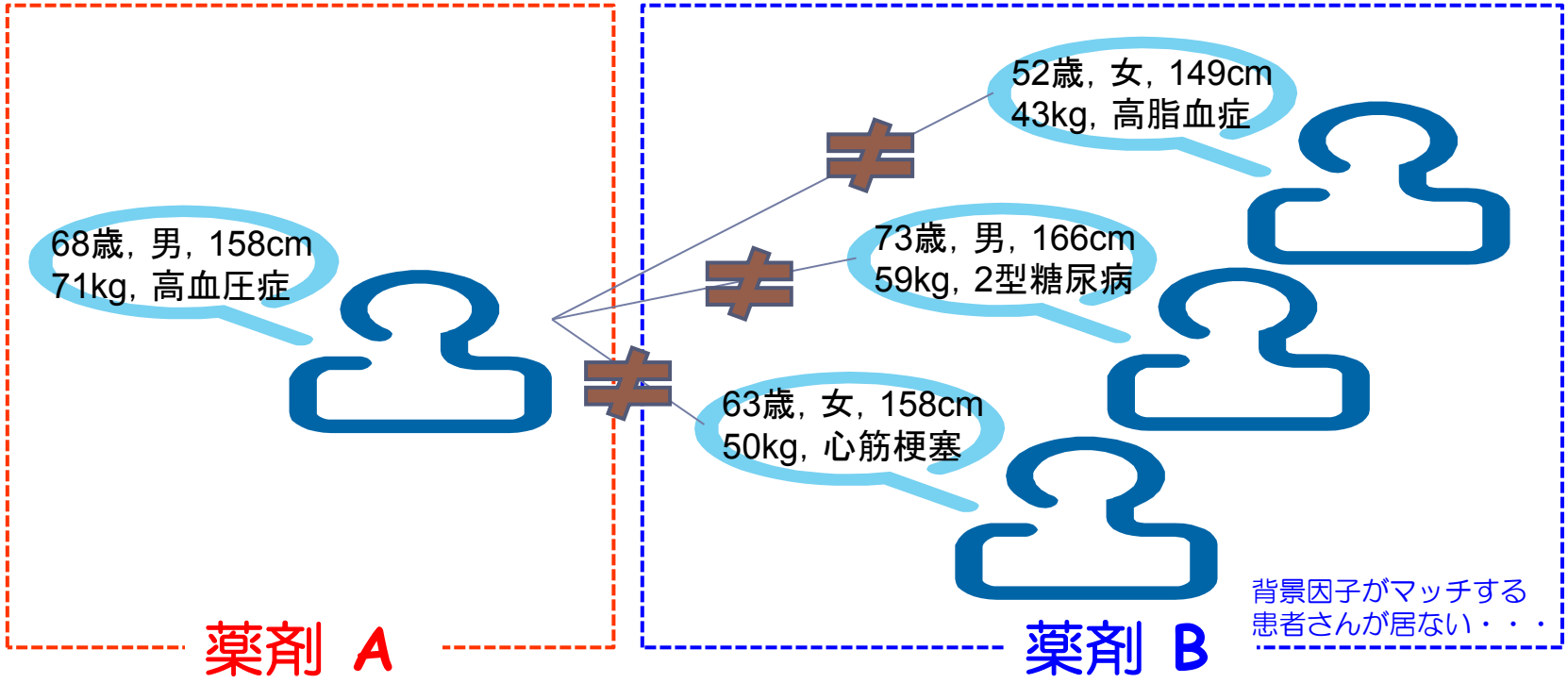
- 研究を行う前の対処方法（デザイン）：
  無作為化（ランダム化）
- 研究を行った後の対処方法（解析）：
  マッチング，層別解析，背景因子をモデルに入れた調整解析



交絡の影響をかわす方法

薬剤 A（処理群）と薬剤 B の 2 群比較を行う場合

1. 無作為化（ランダム化）試験を行う
   ⇒ 背景因子の分布が群間で均一になる方向になるが実施が大変
2. 背景因子のマッチング，層別解析，調整解析を行う（観察研究など）
   ⇒ 背景因子やマッチングする因子が多くなってしまうと破綻する



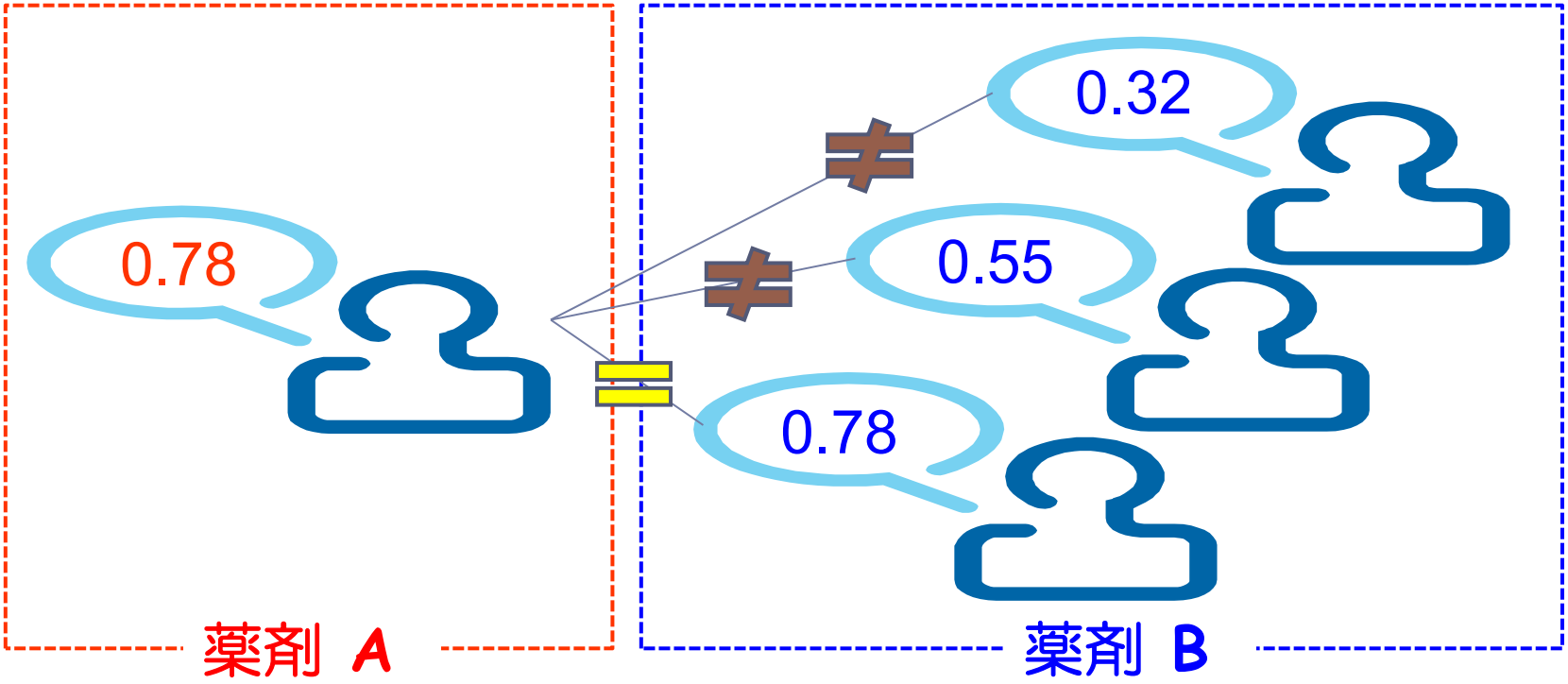
交絡の影響をかわす方法

3. 傾向スコア（propensity score）を用いる

⇒「調整したい背景因子を与えた時に薬剤 A の処理を受ける確率」とし
この確率（傾向スコア）を全ての患者について算出する

⇒「傾向スコアが同じ患者は背景因子が似る」という性質がある

⇒よって「傾向スコアが同じ集団」を「背景因子が同じ集団」とみなし
マッチングや層別解析，傾向スコアを調整因子とした調整解析を行う



傾向スコア（propensity score）

- 各患者さんの薬剤 A に属する確率（患者さんごとに 0% ～ 100%）

＝ 複数の背景因子を 1 つの因子に縮約出来るので，バランスが取りやすい

- propensity score による解析手順は以下の通り
  1. 以下のロジスティックモデルにより「薬剤 A に属する確率」を算出
     （この確率が propensity score ）
     投与群 = 背景因子1 + 背景因子2 + 背景因子3 + ・・・
  2. 以下のいずれかにより解析を行う（他にも多数の方法がある※）
     - マッチング：2 つの群で傾向スコアが等しいとみなせる患者さんをペアとし，そのペアの差の平均を推定値とする
     - 層別解析：傾向スコアの大小によっていくつかの層に分け（ 5 つ程度），各層で 2 群の平均を算出した後，それらを併合した効果の推定値を算出
     - 調整解析：傾向スコアを共変量とした調整解析を行う

※ 傾向スコアによる重み付け推定，重み付き一般化推定方程式，重み付けM推定量，等

傾向スコアを用いたマッチングの例

```
> AB$GROUP <- ifelse(AB$GROUP=="A", 1, 0)   # 群の変数を 0,1 に変換
> result <- glm(GROUP ~ PREDRUG + DURATION, family=binomial, data=AB)
> summary(result)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.37675    0.77863   1.768   0.0770 .
PREDRUG[T.YES] -1.97272    0.84833  -2.325   0.0201 *
DURATION      -0.06415    0.13160  -0.487   0.6259
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> ( X <- result$fitted )   # 各患者さんの傾向スコア
        1         2         3         4         5         6         7         8
0.7879483 0.7657202 0.7770317 0.7540164 0.7770317 0.7770317 0.7540164 0.7770317
 ・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・

> Y <- AB$QOL
> G <- AB$GROUP
> lm(QOL ~ GROUP, data=AB) # 調整前の結果

Coefficients:
(Intercept)        GROUP
        4.0          2.5
```

傾向スコアを用いたマッチングの例

```
> install.packages("Matching", dep=T)    # マッチング用のパッケージのインストール
> library(Matching)                      # マッチング用のパッケージの呼び出し
> matching <- Match(Y=Y, Tr=G, X=X, M=1, estimand="ATE")
> summary(matching)                      # 傾向スコアを用いたマッチング

Estimate...  -0.1375                    # 調整後の平均値の群間差
AI SE......  1.1646
T-stat.....  -0.11807
p.val......  0.90601

Original number of observations..............  40
Original number of treated obs...............  20
Matched number of observations...............  40
Matched number of observations  (unweighted).  55
```

- 以下のモデルにより「薬剤 *A* に属する確率（傾向スコア）」を算出
  投与群 = 背景因子1 + 背景因子2 + 背景因子3 + ・・・
- 応答変数 Y に QOL，変数 *G* に投与群，変数 X に傾向スコアを代入する（ベクトル化する）
- 関数 Match によりマッチングを行う（結果を変数 matching に代入する）
  → マッチした同士で差（群間差）を求め，その平均を取った値が効果の差
- 変数 matching を使うと，背景の偏りの調整度合いも確認できる（次頁）

背景因子の偏りの修正の度合いを確認

```
> MatchBalance(GROUP ~ PREDRUG + DURATION, data=AB, match.out=matching)
***** (V1) PREDRUG[T.YES] *****
                       Before Matching         After Matching
mean treatment........      0.25                    0.25
mean control.........       0.75                    0.25
std mean diff.........    -112.55                     0    # 調整前後の背景因子の差
mean raw eQQ diff.....       0.5                      0
med  raw eQQ diff.....       0.5                      0
max  raw eQQ diff.....         1                      0

 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
```

関数 Match() の引数について

- M=1：1:1 マッチングを行う
- estimand："ATE":群間差，"ATT":処理群（薬剤 A），"ATC":対照群（薬剤 B）の推定値
- caliper：傾向スコアでマッチングを行う際，「傾向スコアが最も近い患者さん」同士をマッチングするが，あまりにも差が大きい場合はマッチングするとよろしくない
  何も指定しない場合は，どれだけ傾向スコアが離れているマッチングも全て解析に含めるが，例えば caliper=0.25 とすると，傾向スコア X の標準偏差 0.25 以上離れている同士のマッチングを除外する

傾向スコアを用いた層別解析

1. 両薬剤群合わせて，傾向スコアの小さい順に 5 つの層に分ける
   ⇒ [0%, 20%) [20%, 40%) [40%, 60%) [60%, 80%) [80%, 100%]
   と，パーセント点を使って 5 つの層に分ける
2. 1 の各層の重みを算出する
3. 以下のモデルにより共分散分析を行う
   $QOL = \beta_0 + \beta_1 \times$薬剤 $+ \beta_2 \times$層$+ \beta_3 \times$薬剤$\times$層
4. 傾向スコアを用いて調整した効果の推定値を計算する
   （k：層の番号，N_k：第 k 層の例数，N：全体の例数，
   $\bar{y}_{ik}$：第 k 層の第 i 群（i=1：薬剤 A，i=0：薬剤 B ）の平均値）

$$\text{効果の推定値} = \sum_{k=1}^{5} \frac{N_k}{N} \left(\bar{y}_{1k} - \bar{y}_{0k} \right)$$

※推定値の分散 $= \sum_{k=1}^{5} \left(\frac{N_k}{N} \right)^2 Var[\bar{y}_{1k} - \bar{y}_{0k}]$

傾向スコアを用いた層別解析

```
> strata <- quantile(X, probs=c(0.0, 0.2, 0.4, 0.6, 0.8, 1.0))
> S      <- cut(X, breaks=strata, label=c("1st","2nd","3rd","4th","5th"), include.lowest=T)
> W <- table(S)/length(S)                  # 各層の重み
> result <- lm(Y ~ G * S)
> ( result2 <- dummy.coef(result) )
Full coefficients are
(Intercept):      2.666667
G:               -0.6666667
S:                     1st        2nd        3rd        4th        5th
                  0.000000  -0.500000   2.133333   9.333333   4.833333
G:S:                   1st        2nd        3rd        4th        5th
                 0.0000000  3.5000000 -0.3833333 -4.4761905  3.7666667
> adjust <- result2$"(Intercept)"  +       # 切片に関する推定値
+ c(result2$G,                             # 薬剤に関する推定値（薬剤A）
+   0)                             +       # 薬剤に関する推定値（薬剤B）
+ weighted.mean(result2$S, W)      +       # 傾向スコアの推定値の重み付け平均値
+ c(weighted.mean(result2$"G:S",W),        # 薬剤×傾向スコアの推定値の重み付け平均値（薬剤A）
+   0)                                     # 薬剤×傾向スコアの推定値の重み付け平均値（薬剤B）
> names(adjust) <- c("A","B")             # ラベルをつける
> adjust; diff(-adjust)                   # 各群の効果
       A        B
5.470179 5.759167
```


※ 調整済み平均値を求めた際の計算方法の変法

【参考】傾向スコアによる重み付け推定（IPW）

- N：全体の例数，y_i：第 i 群の反応，z_i：薬剤群※，e_i：傾向スコア

群1の効果の推定値 $= \sum_{i=1}^{N} \frac{z_i y_i}{e_i} \Big/ \sum_{i=1}^{N} \frac{z_i}{e_i}$

群0の効果の推定値 $= \sum_{i=1}^{N} \frac{(1-z_i) y_i}{1-e_i} \Big/ \sum_{i=1}^{N} \frac{(1-z_i)}{1-e_i}$

```
> Y <- AB$QOL
> Z <- AB$GROUP
> E <- glm(GROUP ~ PREDRUG + DURATION, family=binomial, data=AB)$fitted
> sum(Z*Y/E) / sum(Z/E)                    # 群1（薬剤A）の効果
[1] 5.435364
> sum((1-Z)*Y/(1-E)) / sum((1-Z)/(1-E))    # 群0（薬剤B）の効果
[1] 5.492924
```

※ 薬剤群 → 1：薬剤 A，0：薬剤 B

傾向スコアについて

- 傾向スコアによる調整は，通常のモデルによる解析に比べて利点が多い
  （背景因子の縮約，仮定が弱い，モデルの誤設定に対して頑健，等）
- 傾向スコアは，データにある背景因子についてはバランスを取ることが出来るが
  データにない背景因子についてはバランスを取ることは出来ない
  （共分散分析も同じ）
- 傾向スコアを算出する際，説明変数に入れた因子がひとつでも欠測になると，
  傾向スコアが算出されない（　解析から除かれる）
- 2 群比較の場合は比較的簡単に実行出来るが，3 群以上の場合は解析を行うこと
  が難しい（手法はいくつか提案されている）
- 患者さんの数が薬剤群間で異なる場合，単純な 1：1 マッチングでは，
  「患者さんが少ない方の薬剤群」の患者さんは全て解析に使われるが，
  「患者さんが多い方の薬剤群」の患者さんは余ってしまう可能性がある
- 傾向スコアを用いた調整方法のうち，「傾向スコアを共変量とした調整解析」は
  「傾向スコアと目的変数との間に線形な関係がある」ことを仮定しているが，
  この仮定が成り立つことは保証されていない

本日のメニュー

1. 調整済み平均値
   - イントロ
   - 薬剤と前治療の有無（カテゴリ変数）の場合
   - 薬剤と罹病期間（連続変数）の場合
2. 傾向スコア

参考文献

- 統計学（白旗 慎吾 著，ミネルヴァ書房）
- 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）
- SAS/STAT 9.2 User's Guide Second Edition
- POINTS TO CONSIDER ON ADJUSTMENT FOR BASELINE COVARIATES（EMEA，現 EMA）
- 「臨床試験のための統計的原則」について（ ICH ガイドライン E9 ）
- 調査観察データの統計科学（星野 崇宏 著，岩波書店）

Rで統計解析入門

終