Rで統計解析入門

(4) 散布図と回帰直線と相関係数

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 のデータのみ抽出

```
> 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
> A <- subset(DEP, GROUP=="A")      # 薬剤 A のデータを抽出
> head(A)
 GROUP QOL EVENT DAY PREDRUG DURATION
1     A  15     1  50      NO        1
2     A  13     1 200      NO        3
:     :   :     :   :       :        :
```

準備：架空のデータ「DEP」の変数

- GROUP：薬剤の種類（A，B，C）⇒ A のみ
- 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

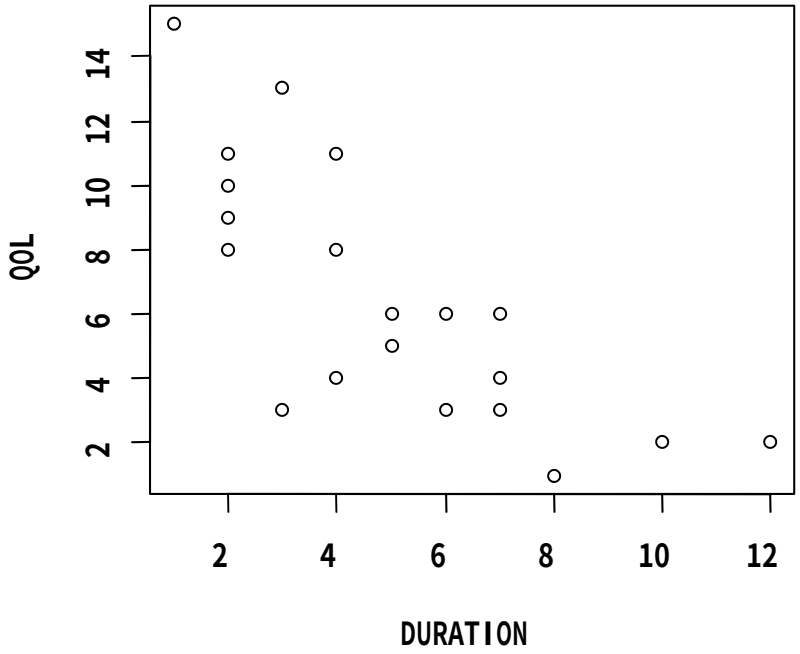
本日のメニュー

1. 散布図と相関係数
2. 回帰直線
3. 相関係数と回帰直線

2 つの連続変数の関係

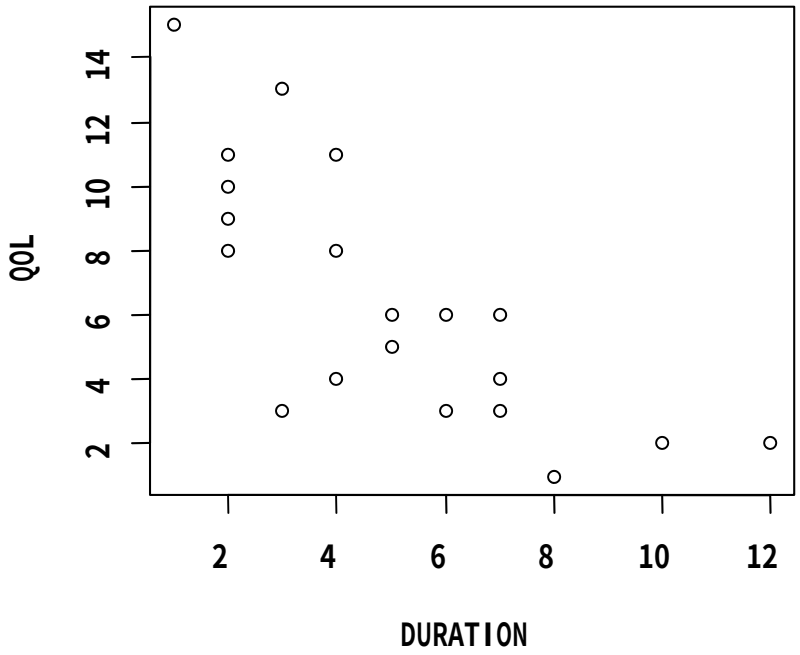
- 罹病期間（DURATION）と QOL がどんな関係かを調べる
- 手っ取り早い方法は散布図を描く
  （ y = f(x) のような感じで QOL ~ DURATION とする）

```
> plot(QOL ~ DURATION, data=A)  # 横軸：罹病期間，縦軸：QOL
```



2 つの連続変数の関係

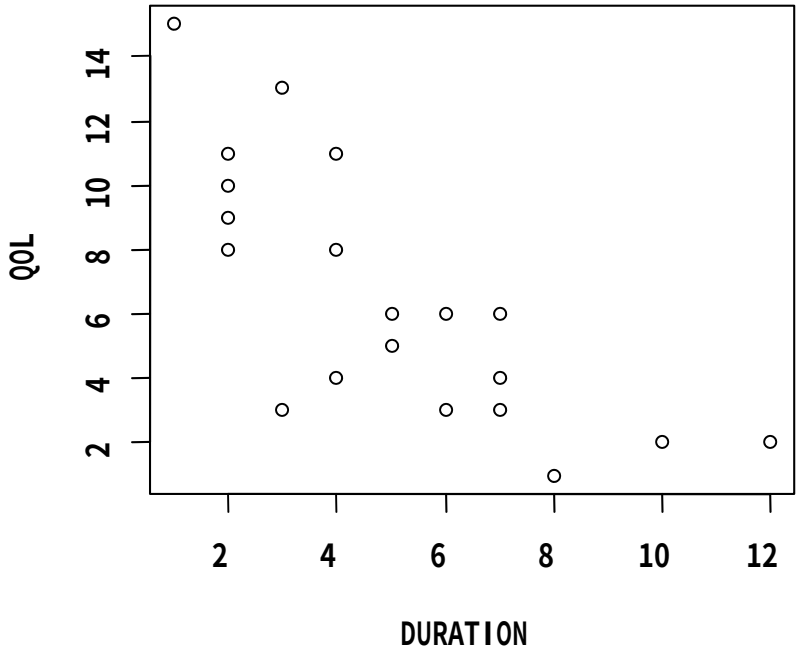
- 散布図より「罹病期間（DURATION）が増えると QOL が下がる」ような感じだが，はっきりしない
- 2 つの連続変数の関係を定量的に表す方法が相関係数



2 つの連続変数の関係

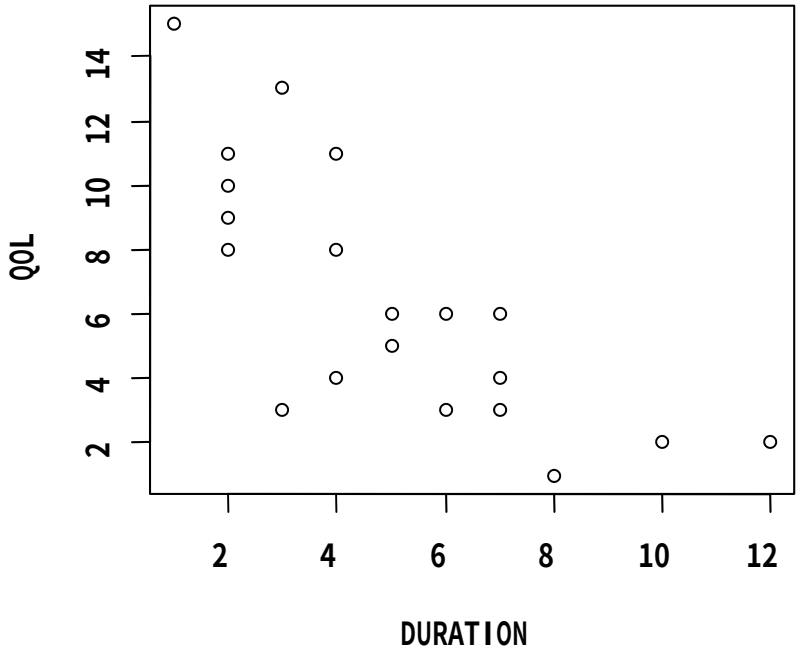
- ピアソンの相関係数：-0.76，スピアマンの相関係数：-0.80

```
> cor(A$DURATION, A$QOL, method="pearson")     # ピアソンの相関係数
[1] -0.76098
> cor(A$DURATION, A$QOL, method="spearman")    # スピアマンの相関係数
[1] -0.8039636
```



2 つの連続変数の関係

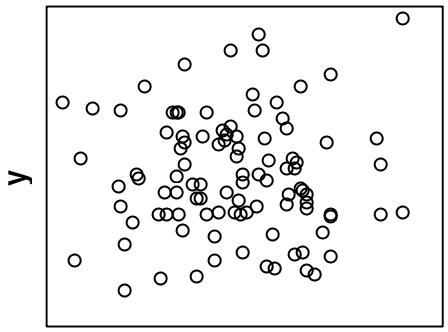
- ピアソンの相関係数　：良く使われるが，外れ値の影響を受けやすい
- スピアマンの相関係数：データを順位データに変換して相関係数を算出
  （外れ値の影響を受けにくい）
- -0.76 とか -0.8 がどうなのかが分からない…⇒ 次頁で判断基準を示す



正の相関（横軸が増えると縦軸も増える傾向）

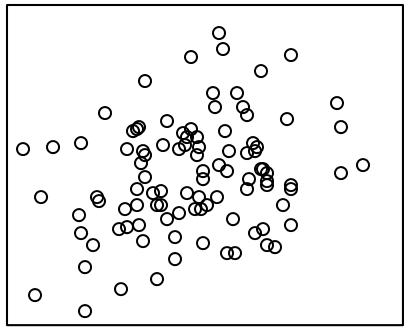
関連なし

r = 0



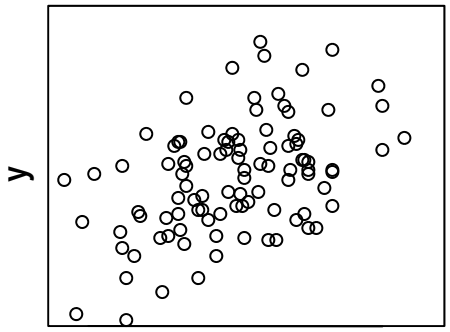
弱い関連

r = 0.3



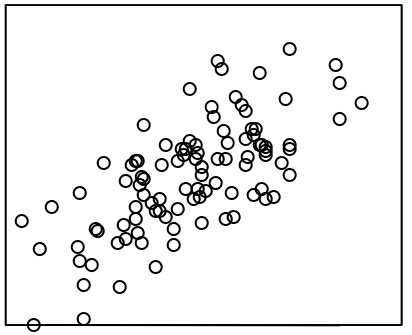
関連あり

r = 0.5



強い関連

r = 0.7

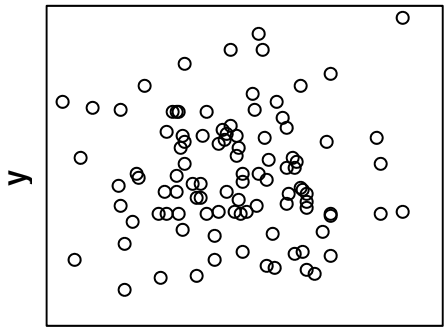



```
> library(TeachingDemos)
> run.old.cor.examp(n = 100)
```

負の相関（横軸が増えると縦軸は減る傾向）

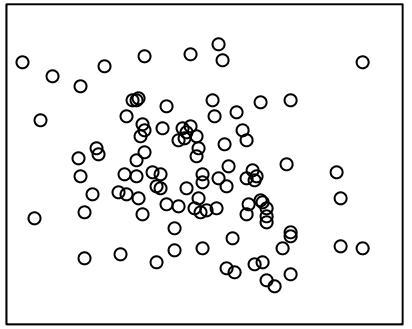
関連なし

r = 0



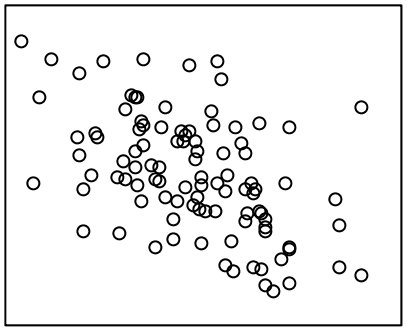
弱い関連

r = - 0.3



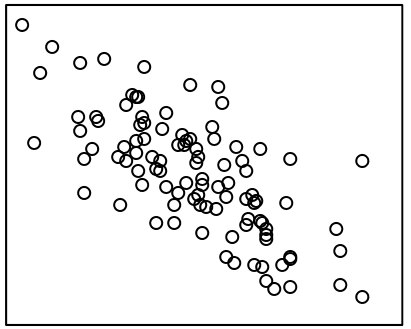
関連あり

r = - 0.5



強い関連

r = - 0.7



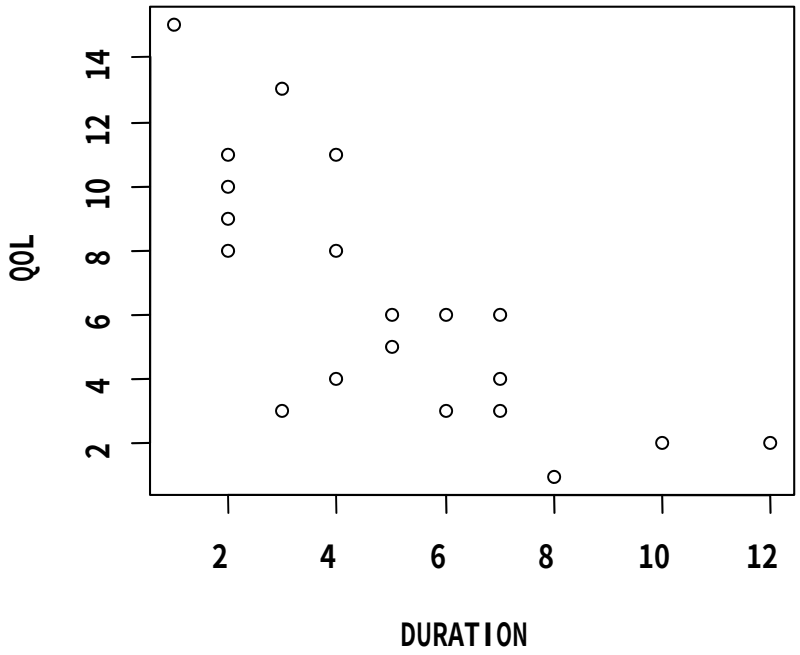

```
> library(TeachingDemos)
> run.old.cor.examp(n = 100)
```

2 つの連続変数の関係

- ピアソンの相関係数：-0.76，スピアマンの相関係数：-0.80

  ⇒ 強い負の相関あり

  ⇒「罹病期間（DURATION）が増えると QOL が下がる」



本日のメニュー

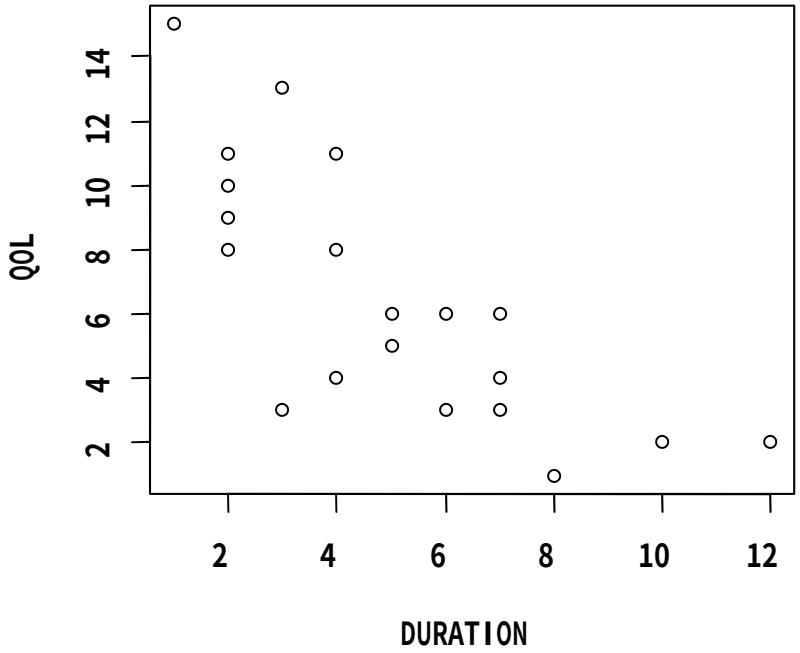
1. 散布図と相関係数
2. 回帰直線
3. 相関係数と回帰直線

2 つの連続変数の関係

- 罹病期間（DURATION）と QOL がどんな関係かを調べる

  ⇒ 散布図にはいろんな点があるせいでどの点を見れば良いか分からない

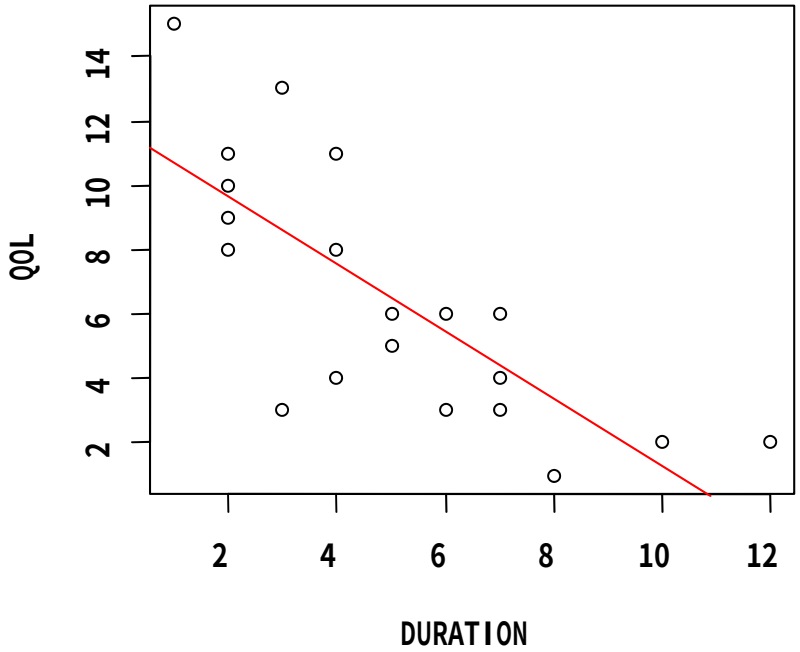
  ⇒ 相関係数から関係の度合いは分かるが，罹病期間（DURATION）がどうなったら QOL がどうなるか，までは分からない



回帰分析：回帰直線・単回帰分析

- 回帰直線を描くことで「2つの連続変数の平均的な推移を直線で表す」ことが出来る ⇒ パッと傾向をつかむことが出来る（単回帰分析）

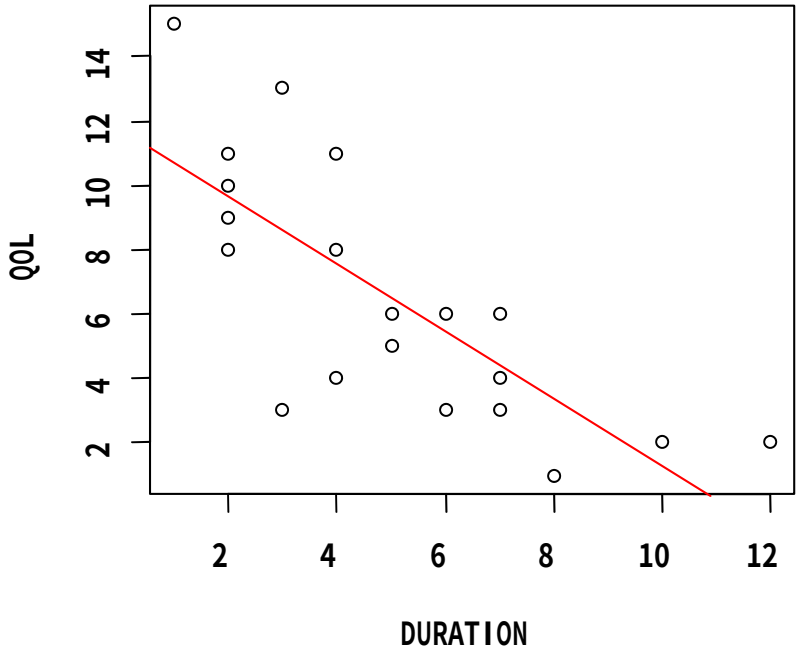
```
> plot(QOL ~ DURATION, data=A)  # 横軸：罹病期間，縦軸：QOL
> abline(lm(QOL ~ DURATION, data=A), col="red") # 回帰直線
```



回帰分析：回帰式

- 回帰式：QOL = 11.7 - 1.04×罹病期間（DURATION）

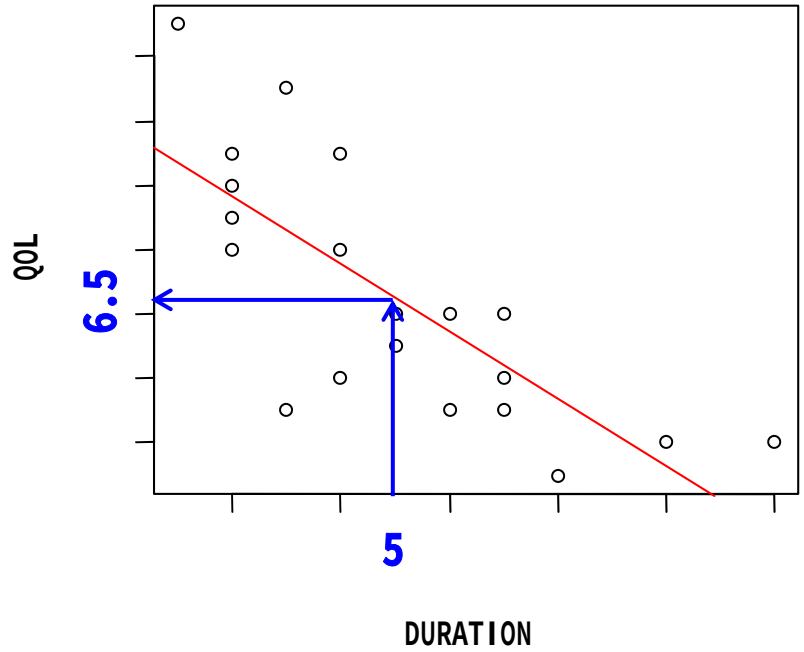
```
> lm(QOL ~ DURATION, data=A)    # 回帰式

Coefficients:
(Intercept)     DURATION
     11.719       -1.044
```



回帰分析：回帰式の性質（1）

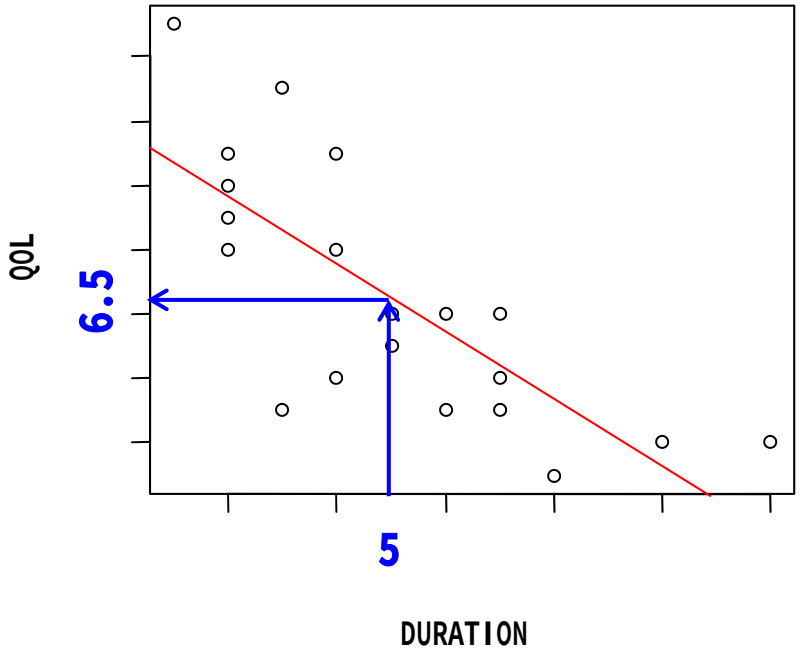
- 回帰式：QOL = 11.7 - 1.04×罹病期間（DURATION）
- 罹病期間が 1 年増えた時に QOL がどう変わるかが予測できる
  - 罹病期間が 0 年：QOL = 11.7 - 1.04×0 = 11.7
  - 罹病期間が 1 年：QOL = 11.7 - 1.04×1 = 10.66

1.04だけ減少



回帰分析：回帰式の性質（2）

- 回帰式：QOL = 11.7 - 1.04×罹病期間（DURATION）
- ある罹病期間の値を入れれば QOL の値が予測できる
  - 罹病期間が 0 年のときの QOL = 11.7 - 1.04×0 = 11.7
  - 罹病期間が 5 年のときの QOL = 11.7 - 1.04×5 = 6.5



【寄り道】データ A の要約統計量

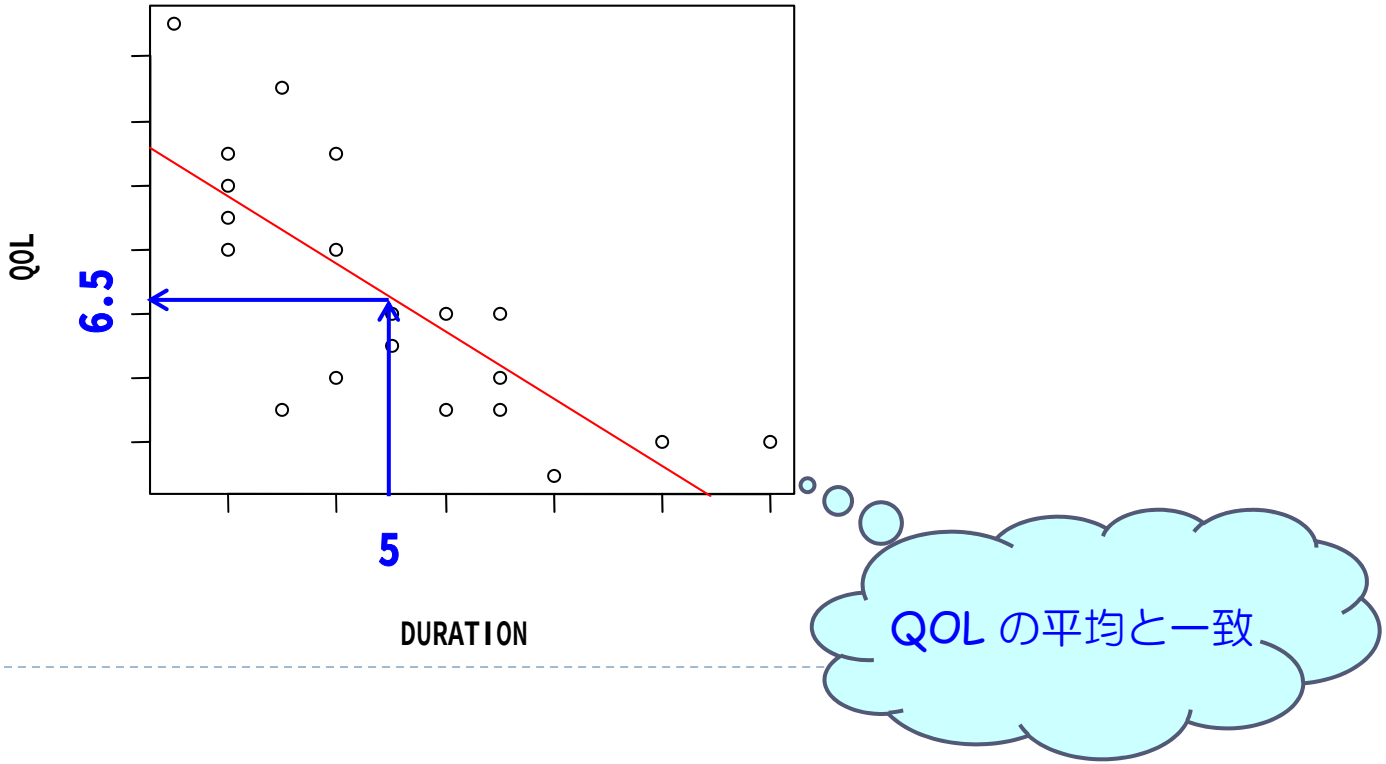
- データ A の要約統計量をパッと出したい場合は関数 summary() を使う

```
> summary(A)

 GROUP       QOL              EVENT           DAY           PREDRUG
 A:20   Min.   : 1.00   Min.   :1.0   Min.   :  50.0   NO :15
 B: 0   1st Qu.: 3.00   1st Qu.:1.0   1st Qu.: 250.0   YES: 5
 C: 0   Median : 6.00   Median :1.0   Median : 475.0
        Mean   : 6.50   Mean   :1.4   Mean   : 495.0
        3rd Qu.: 9.25   3rd Qu.:2.0   3rd Qu.: 712.5
        Max.   :15.00   Max.   :2.0   Max.   :1000.0
    DURATION
 Min.   : 1.00
 1st Qu.: 2.75
 Median : 4.50
 Mean   : 5.00
 3rd Qu.: 7.00
 Max.   :12.00
```

QOL の平均：6.5
罹病期間(DURATION)の平均：5

回帰分析：回帰式の性質（3）

- 回帰式：QOL = 11.7 - 1.04×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 = 11.7 - 1.04×5 = 6.5 （平均）



本日のメニュー

1. 散布図と相関係数
2. 回帰直線
3. 相関係数と回帰直線

相関係数と回帰直線

- 相関係数と回帰直線はどちらも「2つの連続データの関係を見る道具」
  - 相関係数：2 つの連続変数の関連の度合いを -1 ～ 1 の範囲で表したもの
  - 回帰直線：2 つの連続変数の平均的な推移を直線で表したもの

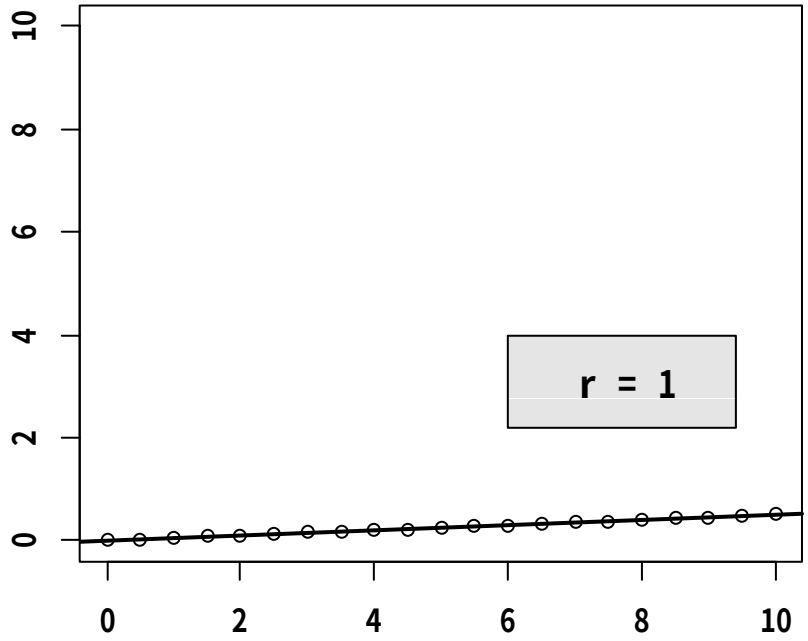
注意：相関係数が 1 や -1 に近い場合は「関連の度合いが強い」ことを表すが，必ずしも回帰直線の傾きが急であることを表さない！

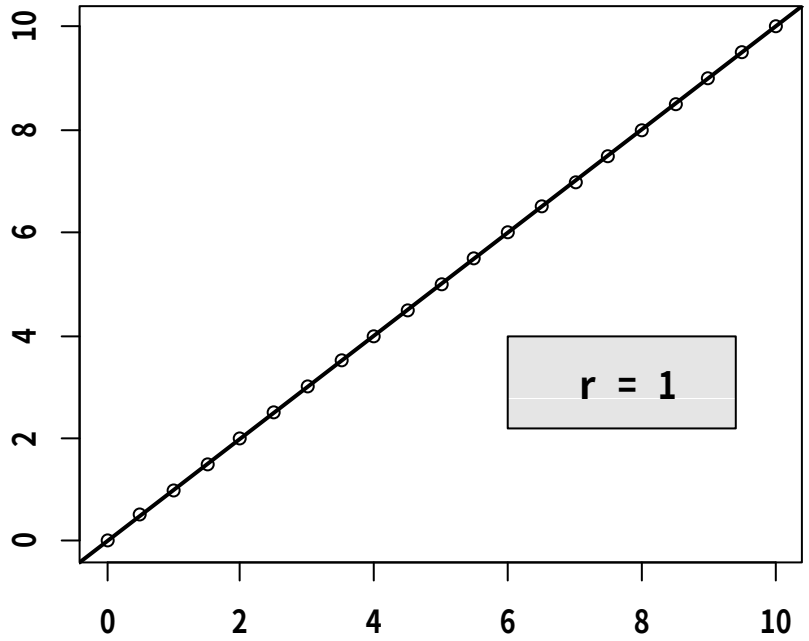
  - 相関係数が 1 や -1 に近い：データが回帰直線からほとんど離れていない
  - 相関係数が 0 に近い：データが回帰直線から離れている

例 1：相関係数の大きさと回帰直線の傾き

- データ（散布図の点）が回帰直線の上にピタッと乗っている
- 「データが回帰直線からほとんど離れていない」ため相関係数が 1
- しかし，回帰直線の傾きは必ずしも急ではない点に注意！

（以下，直線：回帰直線，r：ピアソンの相関係数の値）



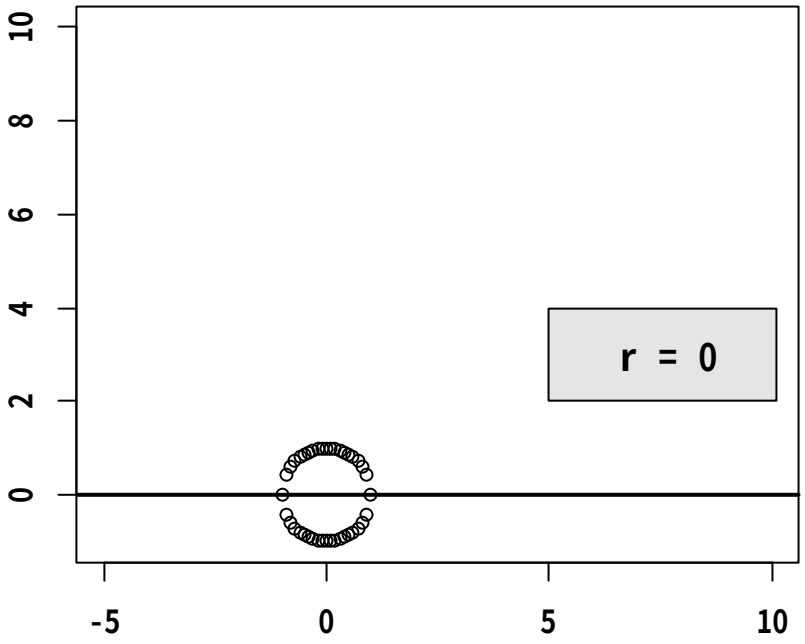


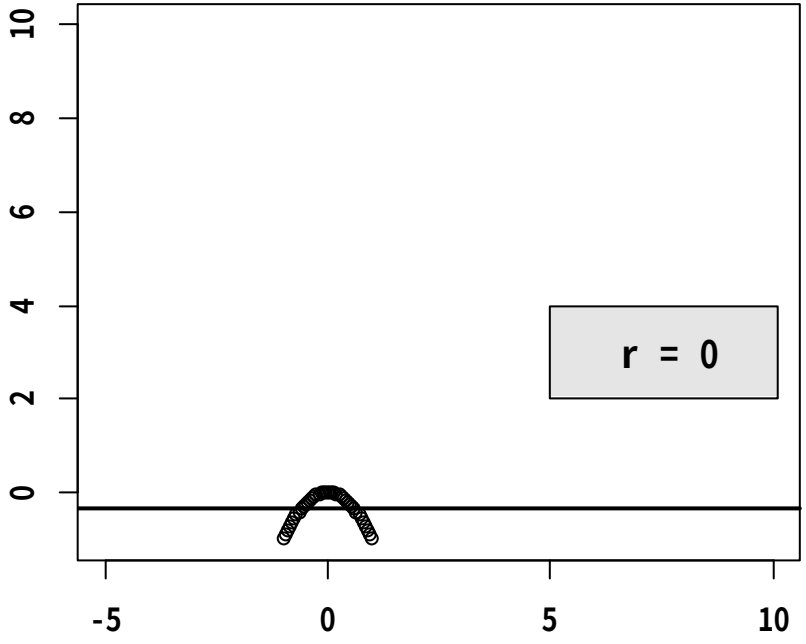
前頁のグラフを描くプログラム

```
> x <- seq(0, 10, length=21)
> y <- seq(0, 0.5, length=21)
> cor(x, y)
[1] 1
> r <- round(cor(x,y), 1)
> plot(x, y, xlim=c(0,10), ylim=c(0,10), xlab="", ylab="")
> abline(lm(y ~ x), lwd=2)
> legend(6, 4, paste("r =",r," "), cex=1.2, bg='gray90')

> x <- seq(0, 10, length=21)
> y <- seq(0, 10, length=21)
> cor(x, y)
[1] 1
> r <- round(cor(x,y), 1)
> plot(x, y, xlim=c(0,10), ylim=c(0,10), xlab="", ylab="")
> abline(lm(y ~ x), lwd=2)
> legend(6, 4, paste("r =",r," "), cex=1.2, bg='gray90')
```

例 2：2 変数の関係を表すが…

- 回帰直線：「関係を直線で表す」ため「曲線的な関係」はつかめない
- 相関係数：「関連の度合いを表す」が「曲線的な関係」はつかめない
- 以下の図では，円形や $y = -x^2$ という関係があるが，相関係数は 0
  （関連なし ⇒ 曲線的な関係はとらえられず）



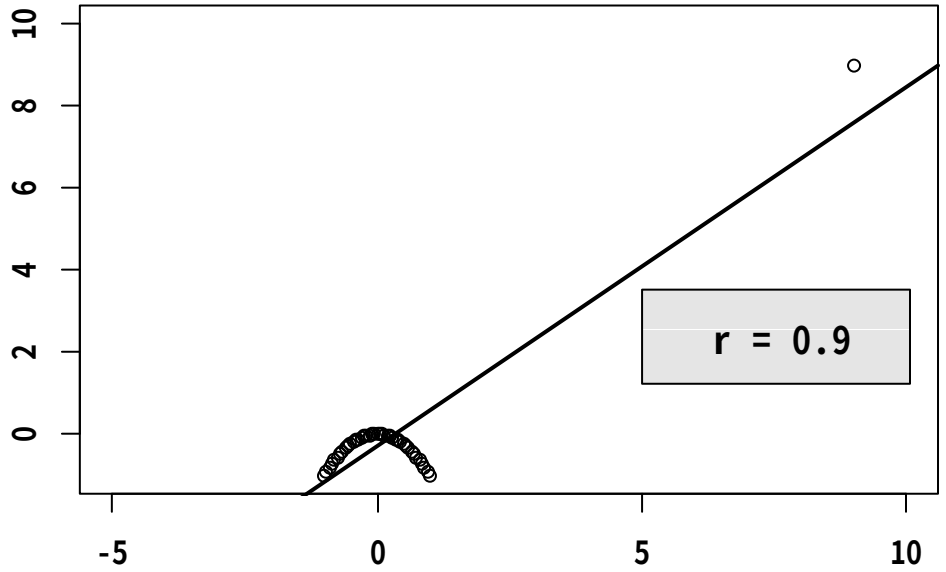


前頁のグラフを描くプログラム

```
> x0 <- c( seq(-1, 1, length=21), seq(-0.9, 0.9, length=19))
> y0 <- c( -sqrt(1-x0[1:21]^2), sqrt(1-x0[22:40]^2) )
> r <- round(cor(x0,y0), 1)
> plot(x0, y0, xlim=c(-5,10), ylim=c(-1,10), xlab="", ylab="")
> abline(lm(y0 ~ x0), lwd=2)
> legend(5, 4, paste("r =",r," "), cex=1.2, bg='gray90')

> x1 <- seq(-1, 1, length=41)
> y1 <- -x1^2
> r <- round(cor(x1,y1), 1)
> plot(x1, y1, xlim=c(-5,10), ylim=c(-1,10), xlab="", ylab="")
> abline(lm(y1 ~ x1), lwd=2)
> legend(5, 4, paste("r =",r," "), cex=1.2, bg='gray90')
```

例 3：点（9, 9）という外れ値の影響

- 例 2 の右の図に点（9, 9）を追加する
- 回帰直線は大きく傾く，ピアソンの相関係数が 0 ⇒ 0.9 になる
- ⇒ 回帰直線やピアソンの相関係数は「外れ値」があると 2 変数間の関係を上手くとらえることが出来なくなる
- 数値の算出の前にグラフ（散布図など）を描くことが重要



前頁のグラフを描くプログラム

```
> x2 <- c(x1, 9)
> y2 <- c(y1, 9)
> r <- round(cor(x2,y2), 1)
> plot(x2, y2, xlim=c(-5,10), ylim=c(-1,10), xlab="", ylab="")
> abline(lm(y2 ~ x2), lwd=2)
> legend(5, 3.5, paste("r =",r," "), cex=1.2, bg='gray90')
```

【参考】例 3 の相関係数（ 2 種類）

- ピアソンの相関係数は外れ値（点 (9,9) ）の影響を大きく受けた
- スピアマンの相関係数は外れ値（点 (9,9) ）の影響をあまり受けない
  ⇒ 0 付近の値となっている

```
> x2 <- c(x1, 9)
> y2 <- c(y1, 9)
> cor(x2 ,y2, method="pearson")    # ピアソンの相関係数
[1] 0.8991118
> cor(x2 ,y2, method="spearman")   # スピアマンの相関係数
[1] 0.048225
```

本日のメニュー

1. 散布図と相関係数
2. 回帰直線
3. 相関係数と回帰直線

参考文献

- 統計学（白旗 慎吾 著，ミネルヴァ書房）
- 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