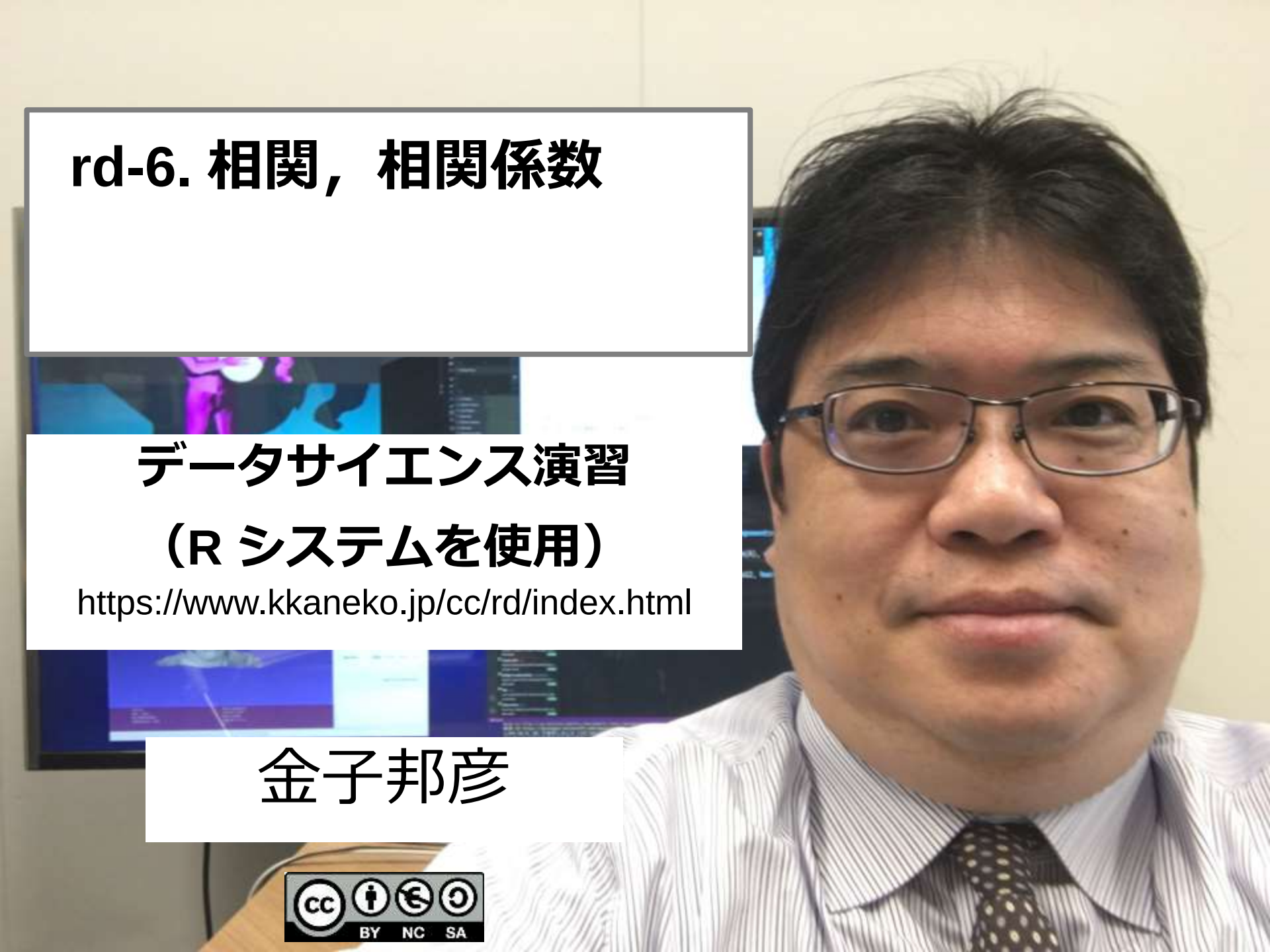
rd-6. 相関，相関係数



データサイエンス演習

（R システムを使用）

https://www.kkaneko.jp/cc/rd/index.html

金子邦彦

アウトライン

6-1. 相関

6-2. 相関係数

6-1 相関

相関

- **相関**は，2つの変数の間に関連性があるか
  （**一方が変化すれば、もう一方も変化する関係**）

- **相関あり**

  X が増えると，Yが増えている

  X が増えると，Y が減っている

- **相関なし**

  X と Y に関係がない

6-2 相関係数

相関係数

- **相関係数**は，**相関**を算出した数値

１や－１に近い値：　相関**あり**

１に近い値：　正の相関関係

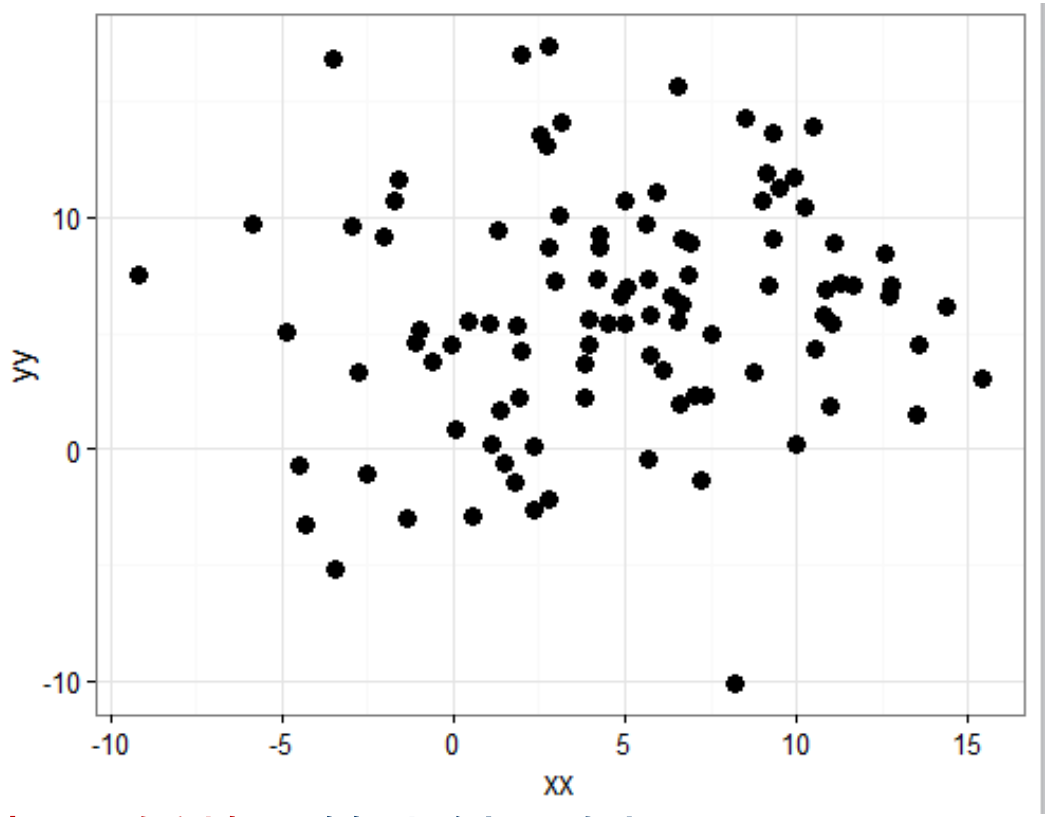
－１に近い値：　負の相関関係

０に近い値：　相関**なし**

R システムでは cor を用いて，相関係数を算出

２つの変数の例

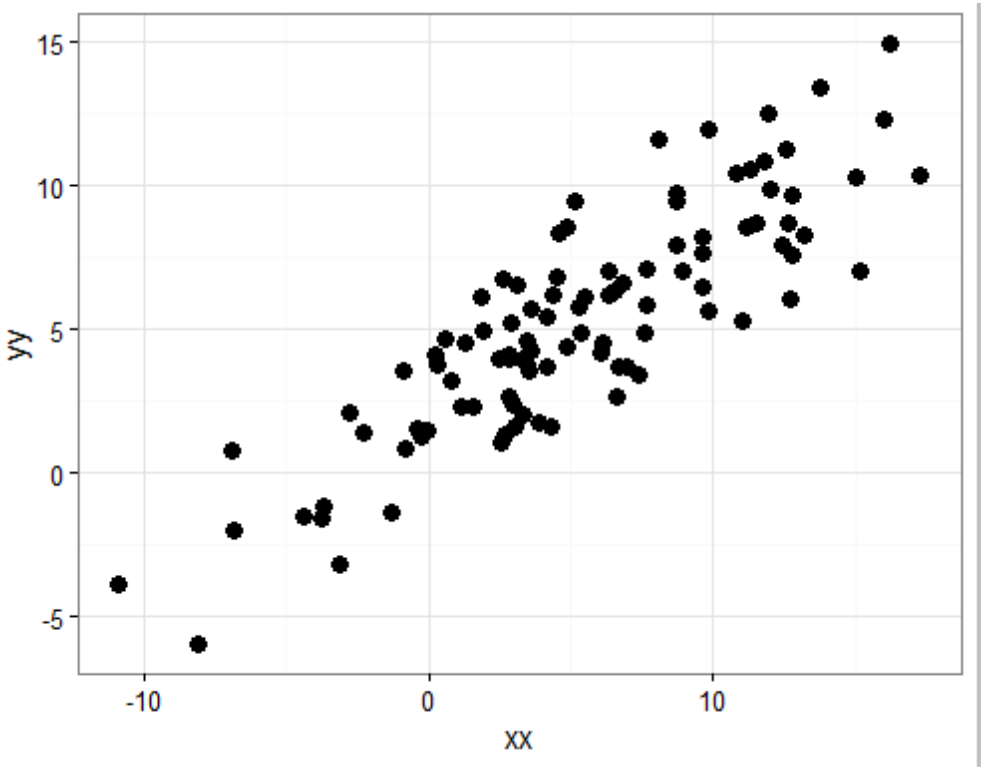
- ２つの変数 xx, yy の散布図



相関係数の算出結果例

```
> cor( d6$xx, d6$yy )
[1] 0.1252164
```

２つの変数の例



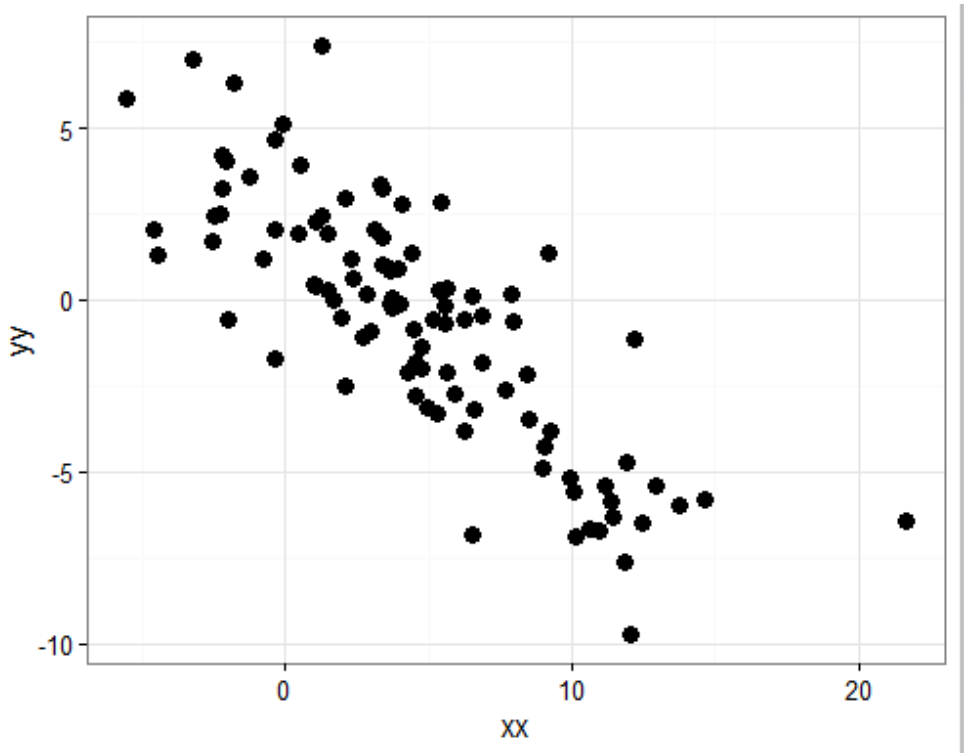
- ２つの変数 xx, yy が互いに相関関係を持つ場合.

xx の値が**増える**と
yy の値が**増える**という
正の相関関係

相関係数の算出結果例

```
> cor(d7$xx, d7$yy)
[1] 0.8620027
>
```

２つの変数の例



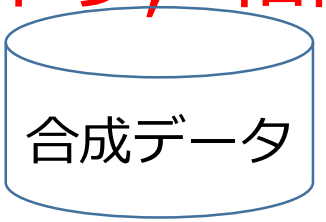
- ２つの変数 xx, yy が互いに相関関係を持つ場合.

xx の値が**増える**と
yy の値が**減る**という
負の相関関係

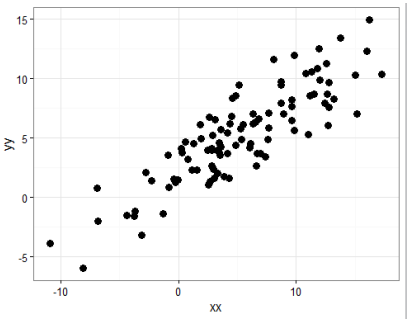
相関係数の算出結果例

```
> cor(d8$xx, d8$yy)
[1] -0.8502535
>
```

合成データからランダムに100個選び標本を作り，相関係数を求める



サイズ１００
の標本を２セット



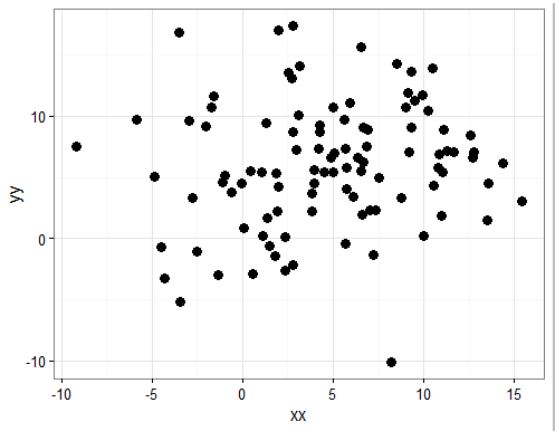
タイプ：数値（整数化しない）
サイズ：100,000

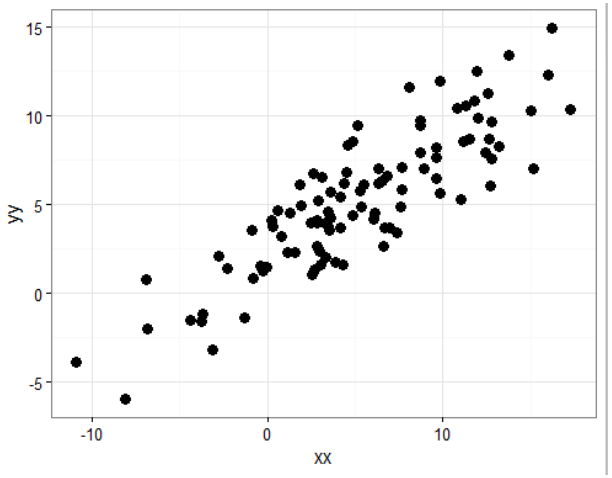
```
x <- rnorm(100000, mean=5, sd=5)
y <- rnorm(100000, mean=5, sd=5)
d7 <- data.frame( xx=x[floor( runif(100, 1, 100000+1) )],
  yy=y[floor( runif(100, 1, 100000+1) )] )
d7$yy <- d7$yy + (d7$xx - d7$yy) * 0.6
library(ggplot2)
ggplot(d7, aes(x=xx)) +
  geom_point( aes(y=yy), size=3 ) + theme_bw()
cor(d7$xx, d7$yy)
```

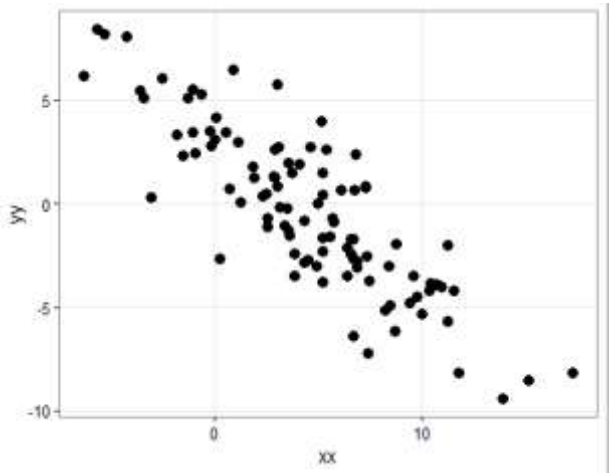
合成データに正の相関関係をもたせる

相関係数

- **１や－１に近い値：　相関あり**
  - １に近い値：　正の相関関係
  - －１に近い値：　負の相関関係
- **０に近い値：　相関なし**








```
> cor( d6$xx, d6$yy )
[1] 0.1252164
```

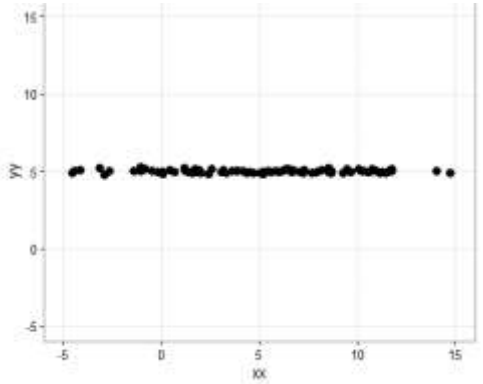
```
> cor(d7$xx, d7$yy)
[1] 0.8620027
>
```

```
> cor(d8$xx, d8$yy)
[1] -0.8502535
```

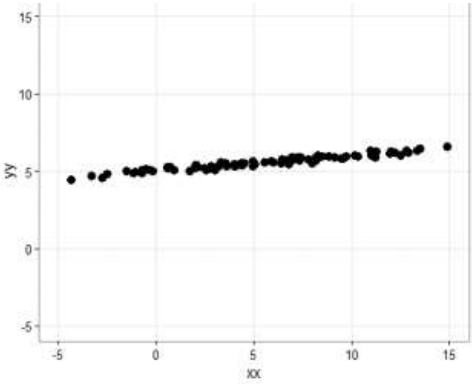
正の相関関係　　**負の相関関係**

相関係数の性質

「**相関の強弱**」の尺度である．「傾き」では**ない**

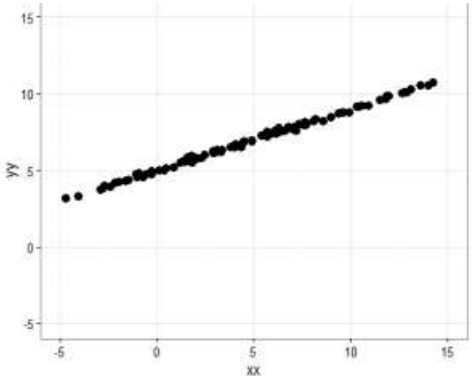



```
> cor(d9$xx, d9$yy)
[1] -0.06027409
```




```
> cor(d10$xx, d10$yy)
[1] 0.9743955
```

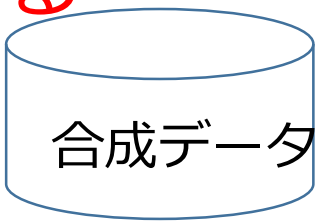
1 に近い値




```
> cor(d11$xx, d11$yy)
[1] 0.998944
>
```

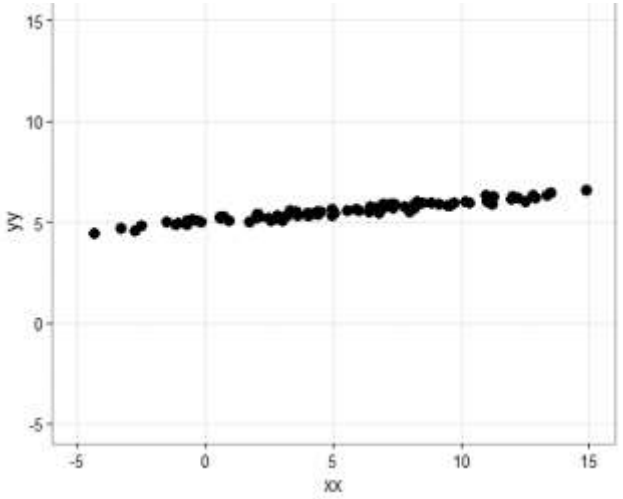
1 に近い値

合成データからランダムに100個選び標本を作る



サイズ１００
の標本を２セット

タイプ：数値（整数化しない）
サイズ：100,000

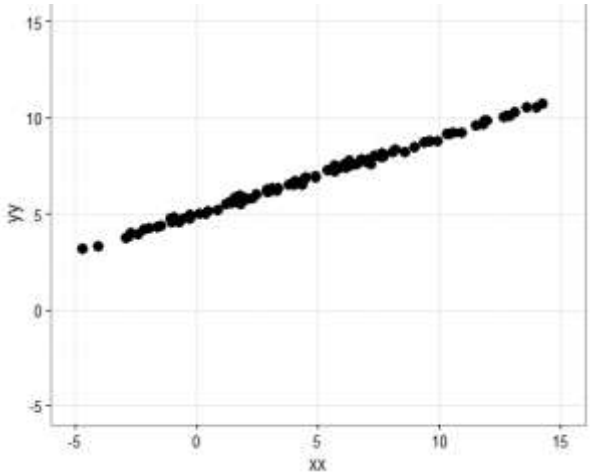



```
x2 <- rnorm(100000, mean=5, sd=5)
y2 <- rnorm(100000, mean=5, sd=0.1)
d10 <- data.frame( xx=x2[floor( runif(100, 1, 100000+1) )],
  yy=y2[floor( runif(100, 1, 100000+1) )] )
d10$yy <- 0.1 * d10$xx + d10$yy
library(ggplot2)
ggplot(d10, aes(x=xx)) +
  geom_point( aes(y=yy), size=3 ) + xlim(-5, 15) + ylim(-5, 15) +
  theme_bw()
cor(d10$xx, d10$yy)
```

合成データに，
正の相関関係をもたせる

合成データからランダムに100個選び標本を作る

サイズ１００
の標本を２セット

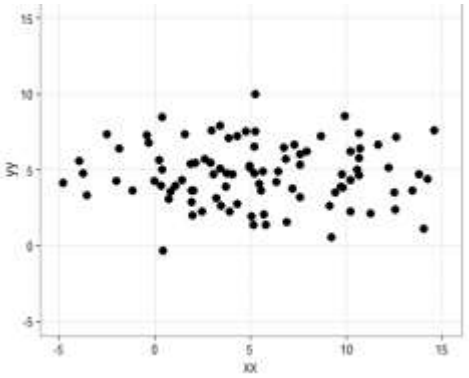


タイプ：数値（整数化しない）
サイズ：100,000

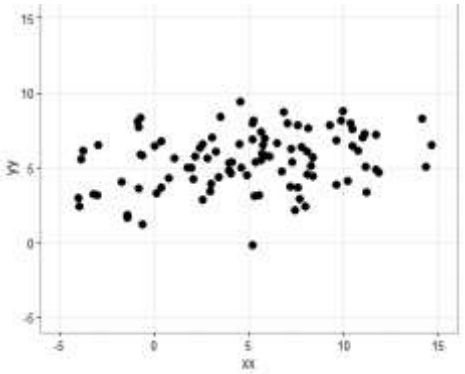
```
x2 <- rnorm(100000, mean=5, sd=5)
y2 <- rnorm(100000, mean=5, sd=0.1)
d11 <- data.frame( xx=x2[floor( runif(100, 1, 100000+1) )],
  yy=y2[floor( runif(100, 1, 100000+1) )] )
d11$yy <- 0.4 * d11$xx + d11$yy
library(ggplot2)
ggplot(d11, aes(x=xx)) +
  geom_point( aes(y=yy), size=3 ) + xlim(-5, 15) + ylim(-5, 15) +
  theme_bw()
cor(d11$xx, d11$yy)
```

合成データに，
正の相関関係をもたせる

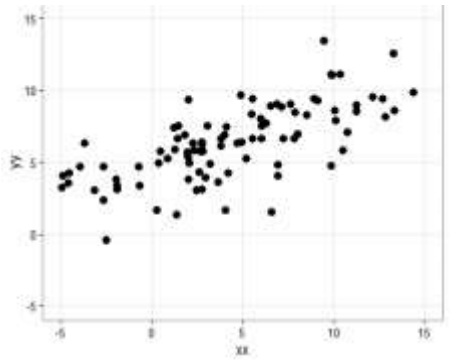
相関係数の例




```
> cor(d12$xx, d12$yy)
[1] -0.02723688
>
```




```
> cor(d13$xx, d13$yy)
[1] 0.2808435
>
```




```
> cor(d14$xx, d14$yy)
[1] 0.7268933
>
```

おわりに

- **相関**は，2つの変数の間に関連性があるか
  （**一方が変化すれば、もう一方も変化する関係**）
- **相関係数**は，**相関**を算出した数値

- ３つ以上の変数があるとき、相関係数は多数求まる

  変数　A, B, C　に対して

  A と B の相関係数，

  B と C の相関係数，

  C と A の相関係数