rd-6. 相関係数

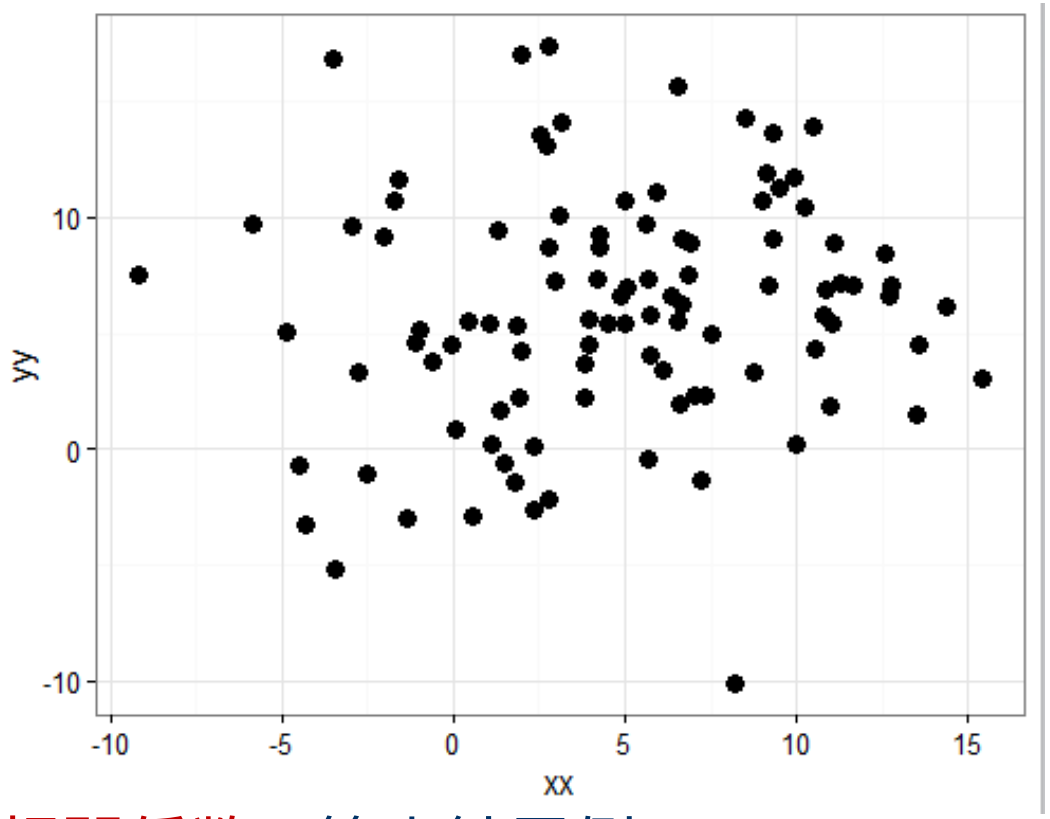
（Rシステムでデータサイエンス演習）

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

金子邦彦

２つの数値データの例

２つの数値データ xx, yy の散布図

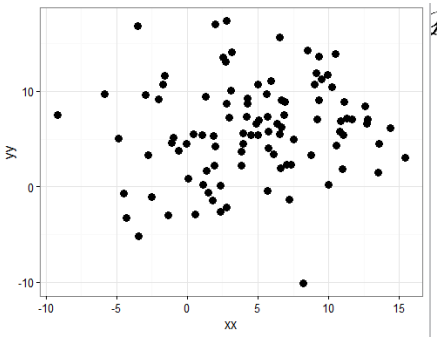


相関係数の算出結果例

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

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

```
x <- rnorm(100000, mean=5, sd=5)
y <- rnorm(100000, mean=5, sd=5)
d6 <- data.frame( xx=x[floor( runif(100, 1, 100000+1) )],
  yy=y[floor( runif(100, 1, 100000+1) )] )
library(ggplot2)
ggplot(d6, aes(x=xx)) +
  geom_point( aes(y=yy), size=3 ) + theme_bw()
cor(d6$xx, d6$yy)
```

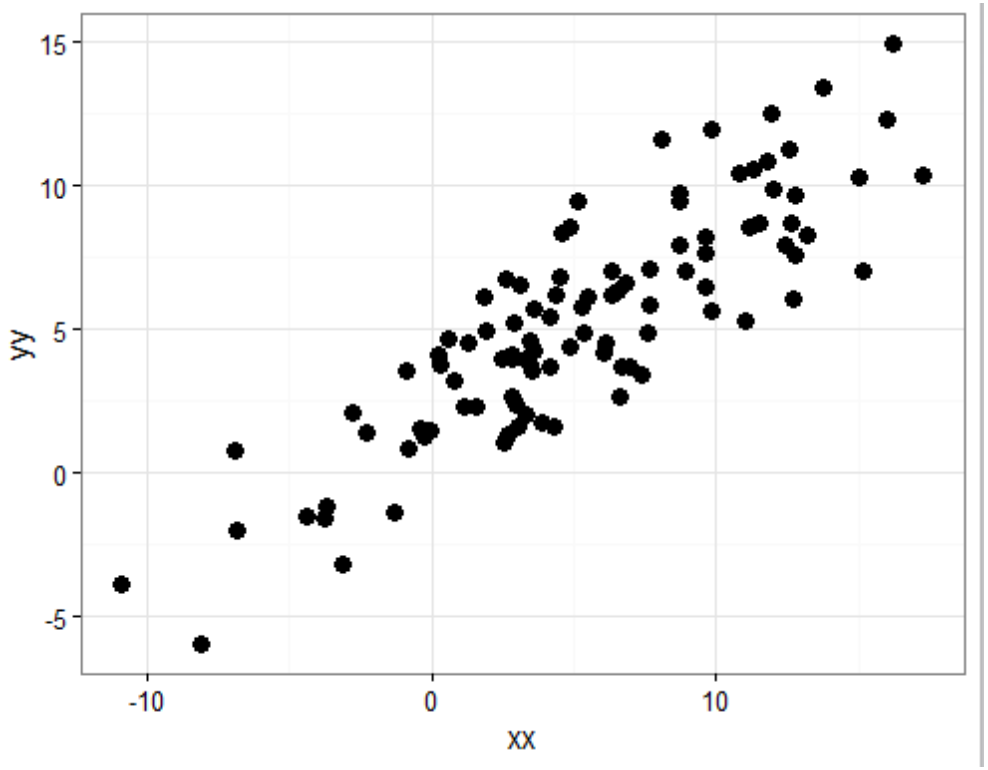
合成データの生成

「x=xx」 グラフの x 軸は，属性 xx

「y=yy」 グラフの y 軸は，属性 yy

２つの数値データの例

２つの数値データ xx, yy が互いに相関関係を持つ場合.

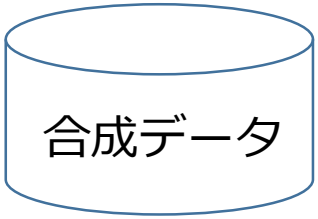


xx の値が**増える**と
yy の値が**増える**という
正の相関関係

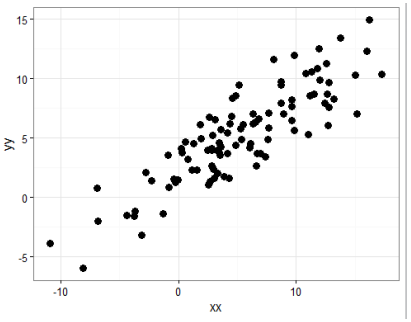
相関係数の算出結果例

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

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



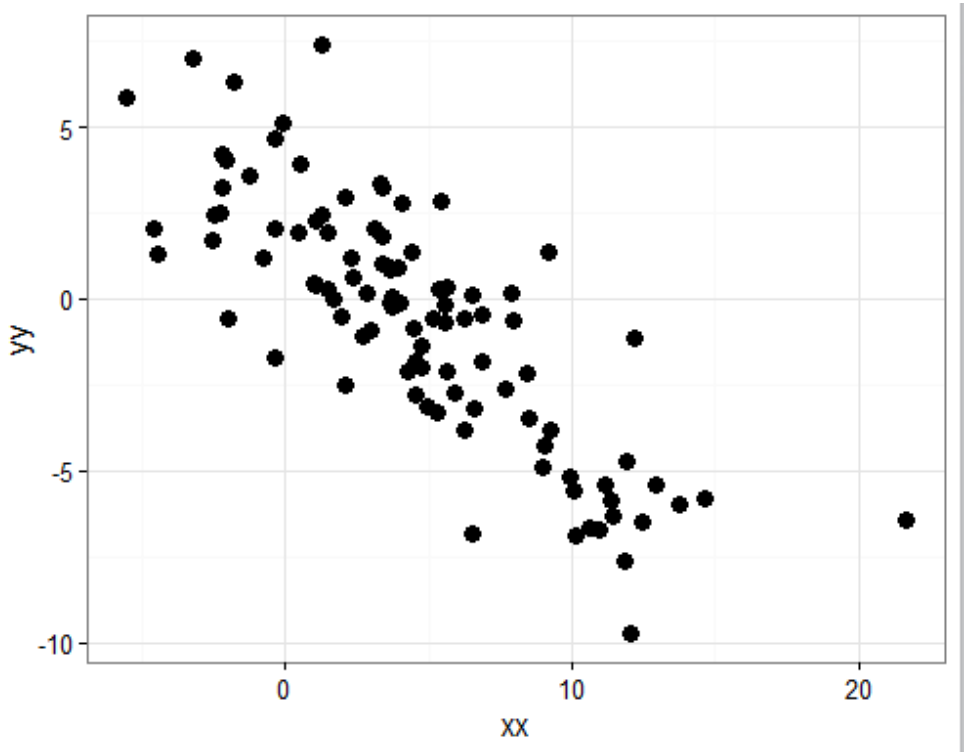
タイプ：数値（整数化しない）
サイズ：100,000

```
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

２つの数値データの例

２つの数値データ xx, yy が互いに相関関係を持つ場合.



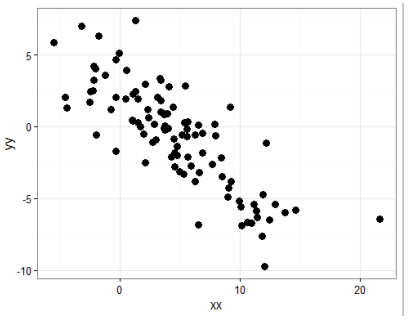
xx の値が**増える**と
yy の値が**減る**という
負の相関関係

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



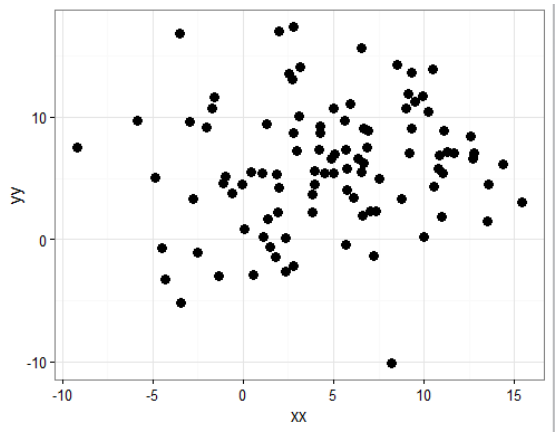
タイプ：数値（整数化しない）
サイズ：100,000

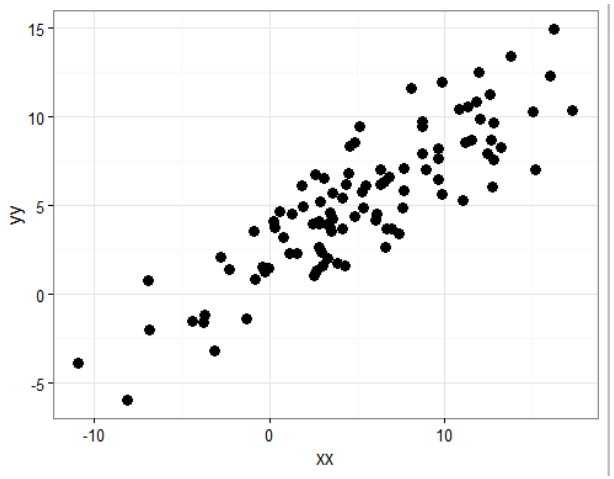
```
x <- rnorm(100000, mean=5, sd=5)
y <- rnorm(100000, mean=5, sd=5)
d8 <- data.frame( xx=x[floor( runif(100, 1, 100000+1) )],
  yy=y[floor( runif(100, 1, 100000+1) )] )
d8$yy <- d8$yy - (d8$xx + d8$yy) * 0.6
library(ggplot2)
ggplot(d8, aes(x=xx)) +
  geom_point( aes(y=yy), size=3 ) + theme_bw()
cor(d8$xx, d8$yy)
```

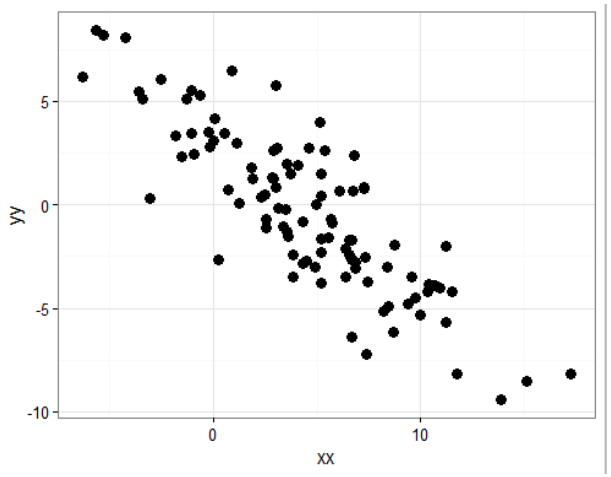
合成データに
負の相関関係をもたせる

相関係数

- 相関係数は，２つのデータの間の**関係の強弱**をはかるための指標








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

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

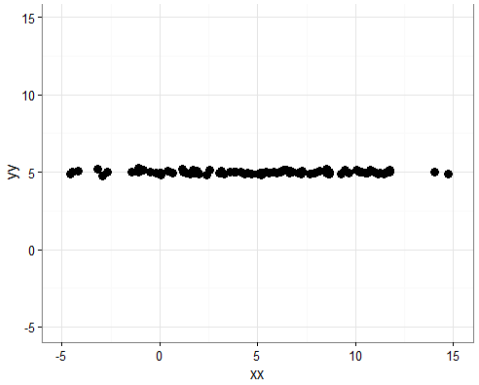
```
> cor(d8$xx, d8$yy)
[1] -0.8502535
```

正

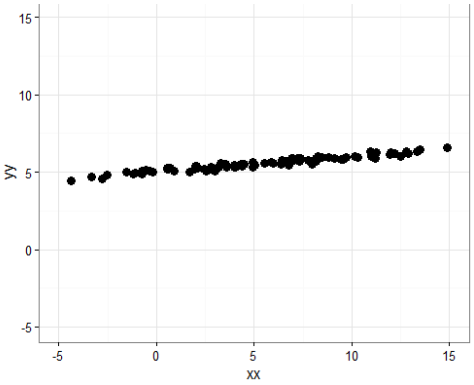
負

相関係数の性質

「**相関の強弱**」の尺度である．「傾き」ではない

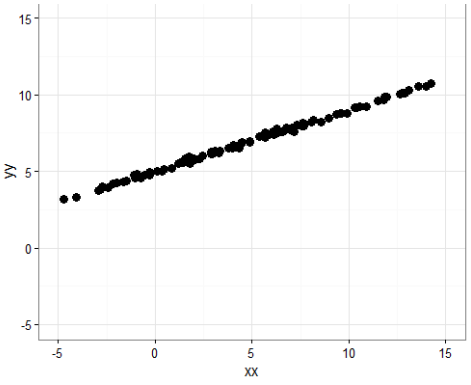



```
> cor(d9$xx, d9$yy)
[1] -0.06027409
```



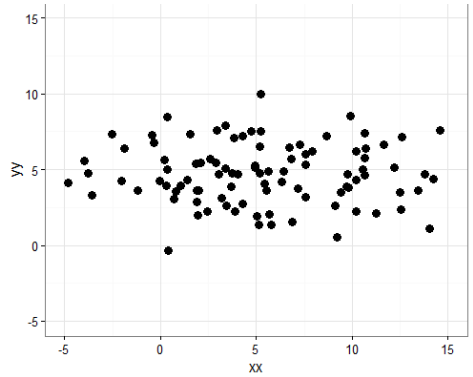

```
> cor(d10$xx, d10$yy)
[1] 0.9743955
```

1 に近い値

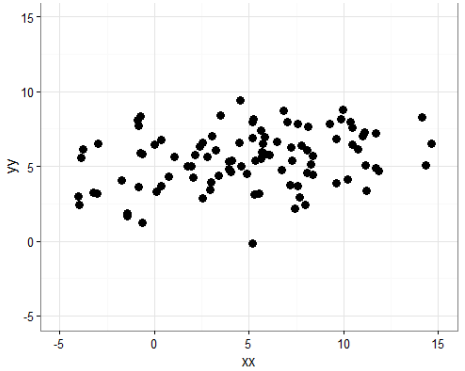



```
> cor(d11$xx, d11$yy)
[1] 0.998944
>
```

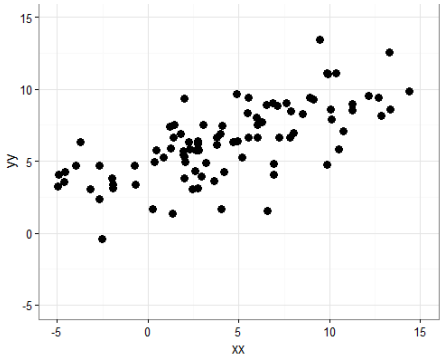
1 に近い値




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

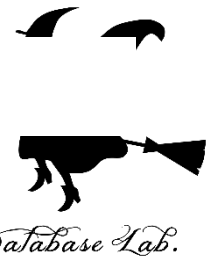


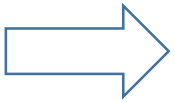

```
> cor(d13$xx, d13$yy)
[1] 0.2808435
>
```



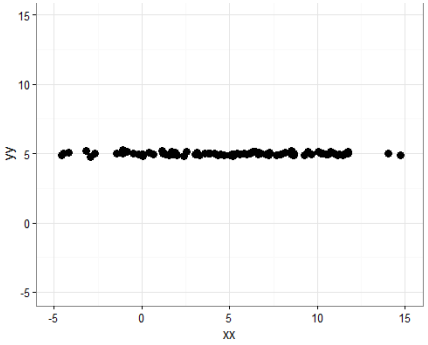

```
> cor(d14$xx, d14$yy)
[1] 0.7268933
>
```

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

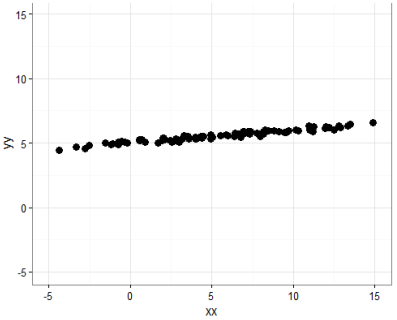


タイプ：数値（整数化しない）
サイズ：100,000

```
x2 <- rnorm(100000, mean=5, sd=5)
y2 <- rnorm(100000, mean=5, sd=0.1)
d9 <- data.frame( xx=x2[floor( runif(100, 1, 100000+1) )],
  yy=y2[floor( runif(100, 1, 100000+1) )] )
library(ggplot2)
ggplot(d9, aes(x=xx)) +
  geom_point( aes(y=yy), size=3 ) + xlim(-5, 15) + ylim(-5, 15) +
  theme_bw()
cor(d9$xx, d9$yy)
```

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

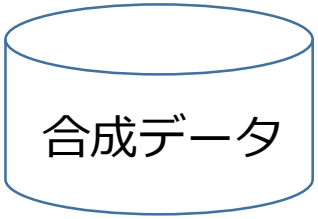


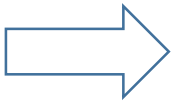
タイプ：数値（整数化しない）
サイズ：100,000

```
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)
```

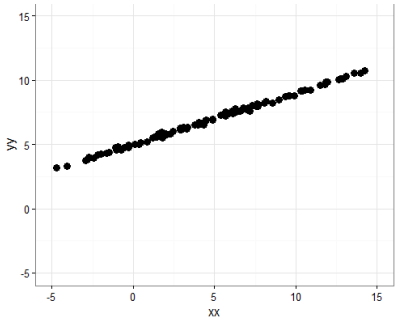
合成データに，
正の相関関係をもたせる

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



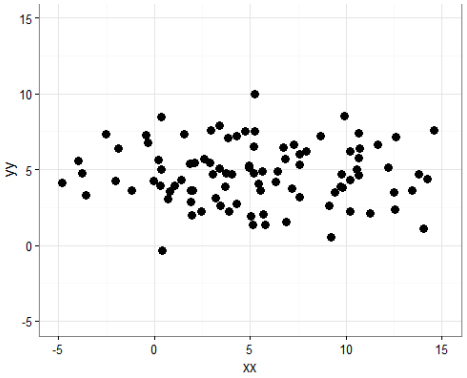
タイプ：数値（整数化しない）
サイズ：100,000

```
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

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



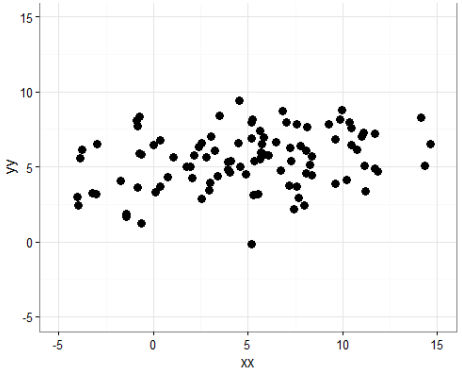
タイプ：数値（整数化しない）
サイズ：100,000

```
x3 <- rnorm(100000, mean=5, sd=5)
y3 <- rnorm(100000, mean=5, sd=2)
d12 <- data.frame( xx=x3[floor( runif(100, 1, 100000+1) )],
  yy=y3[floor( runif(100, 1, 100000+1) )] )
library(ggplot2)
ggplot(d12, aes(x=xx)) +
  geom_point( aes(y=yy), size=3 ) + xlim(-5, 15) + ylim(-5, 15) +
  theme_bw()
cor(d12$xx, d12$yy)
```

合成データからランダムに100個選び標本を作る

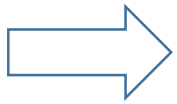
サイズ１００
の標本を２セット

タイプ：数値（整数化しない）
サイズ：100,000

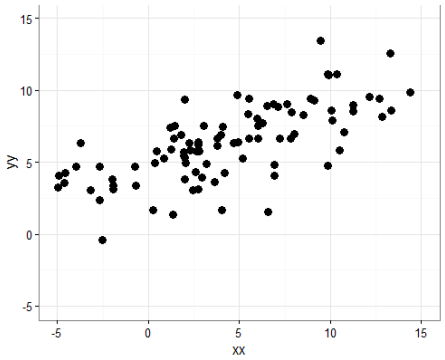



```
x3 <- rnorm(100000, mean=5, sd=5)
y3 <- rnorm(100000, mean=5, sd=2)
d13 <- data.frame( xx=x3[floor( runif(100, 1, 100000+1) )],
  yy=y3[floor( runif(100, 1, 100000+1) )] )
d13$yy <- 0.1 * d13$xx + d13$yy
library(ggplot2)
ggplot(d13, aes(x=xx)) +
  geom_point( aes(y=yy), size=3 ) + xlim(-5, 15) + ylim(-5, 15) +
  theme_bw()
cor(d13$xx, d13$yy)
```

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

```
x3 <- rnorm(100000, mean=5, sd=5)
y3 <- rnorm(100000, mean=5, sd=2)
d14 <- data.frame( xx=x3[floor( runif(100, 1, 100000+1) )],
  yy=y3[floor( runif(100, 1, 100000+1) )] )
d14$yy <- 0.4 * d14$xx + d14$yy
library(ggplot2)
ggplot(d14, aes(x=xx)) +
  geom_point( aes(y=yy), size=3 ) + xlim(-5, 15) + ylim(-5, 15) +
  theme_bw()
cor(d14$xx, d14$yy)
```

おわりに

- 相関係数は、２つの変数の**相関の強弱**をはかるための指標
- ２つの変数をプロットしたときの「傾き」の指標ではない

- ３つ以上の変数があるとき、相関係数はたくさん求まる

  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