

# 情報科学入門

## 第10回: データの関係性

瓜生真也（デザイン型AI教育研究センター・助教）

# 講義内容

1. ガイダンス

2. 情報社会への理解

3. 情報社会を支える仕組みと特徴

4. 情報セキュリティ

5. データサイエンス・AIの歴史

6. AI活用の現状と展望

7. プログラミング基礎

8. データの記述
9. データの可視化

10. データの関係性

11. プログラミング演習

12. レポート作成

13. プログラミング応用

14. プレゼンテーション1

15. プレゼンテーション2

16. まとめ・振り返り

講義に関する資料（スライド、補足資料等）を  GitHubに置いておきます

 <https://github.com/uribo/INFO1010>




ダウンロード可能

Preview

5.86 MB

Raw



# 今日の目標

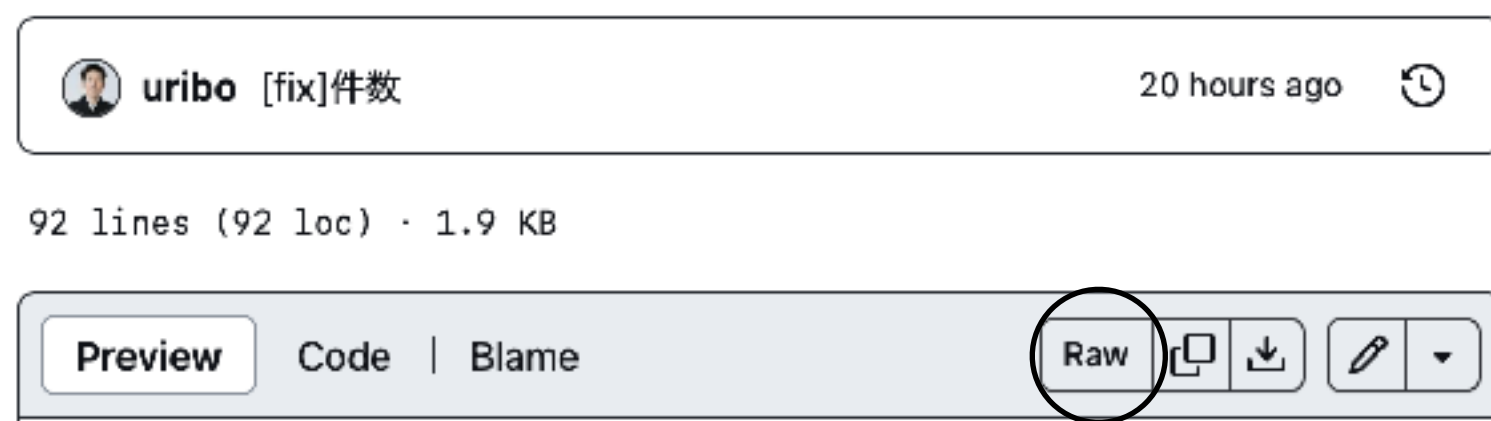
データの関係を数値的に把握する

# 【課題】 関係・比較についてのクイズ

提出期限: 次回 (2024-01-10) の講義開始前まで **manabaのレポートとして提出してください**

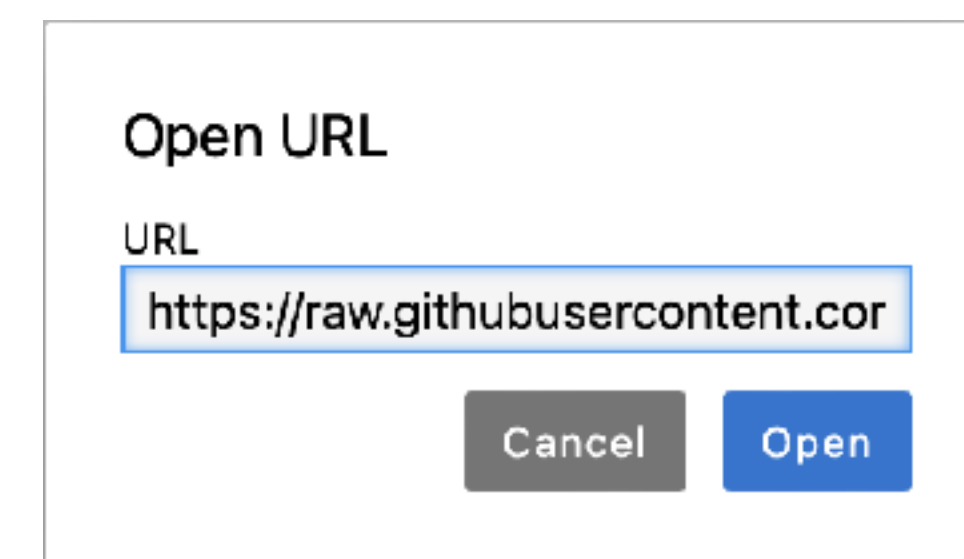
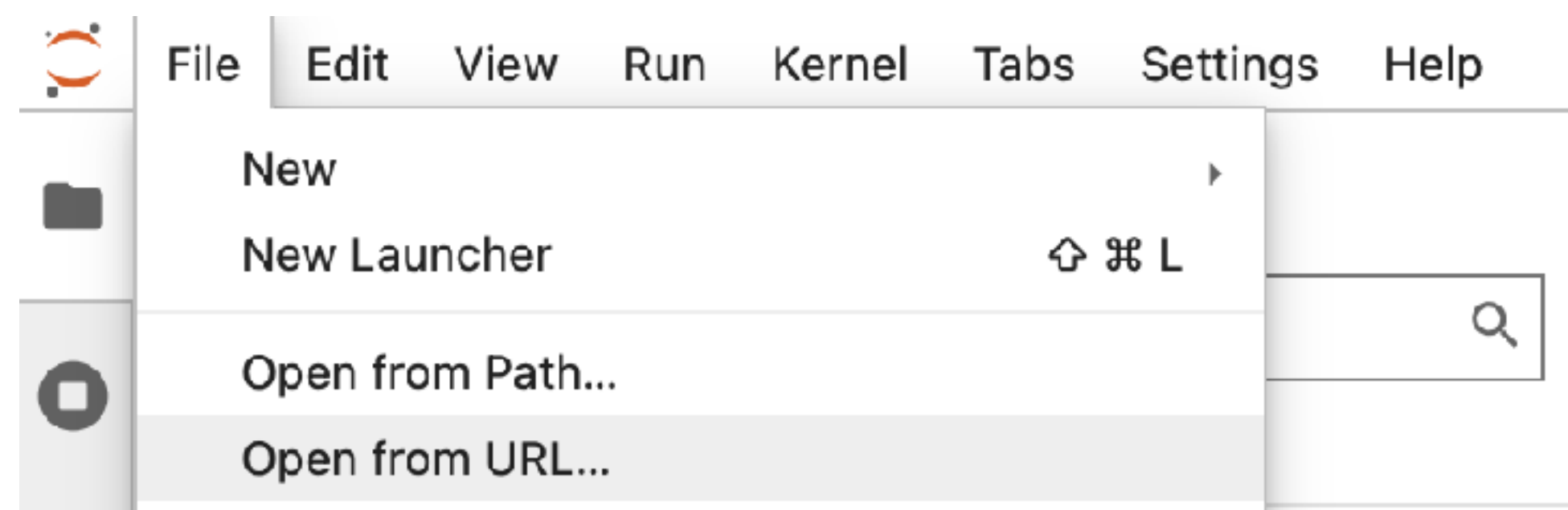
GitHubからweek10\_your\_turn.ipynbをアップロードして記載

exeai / week04 / answer.ipynb



Rawをクリックして表示先のURLをコピー

JupyterHubのサーバを起動、メニューのFileから コピーしたURLを貼り付け  
“Open from URL…” を選択



## 注意: ファイル名は英数字のみにすること

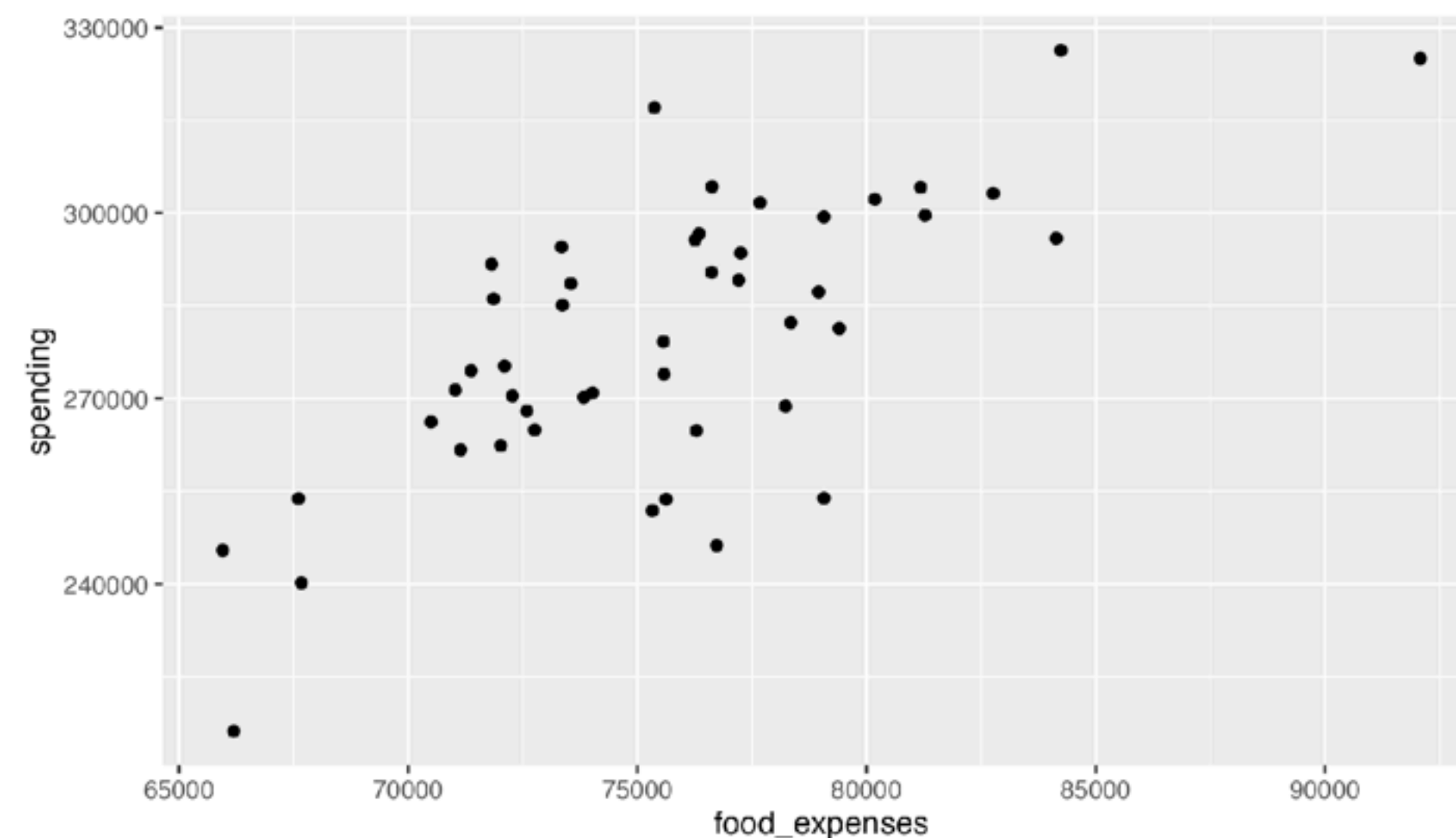
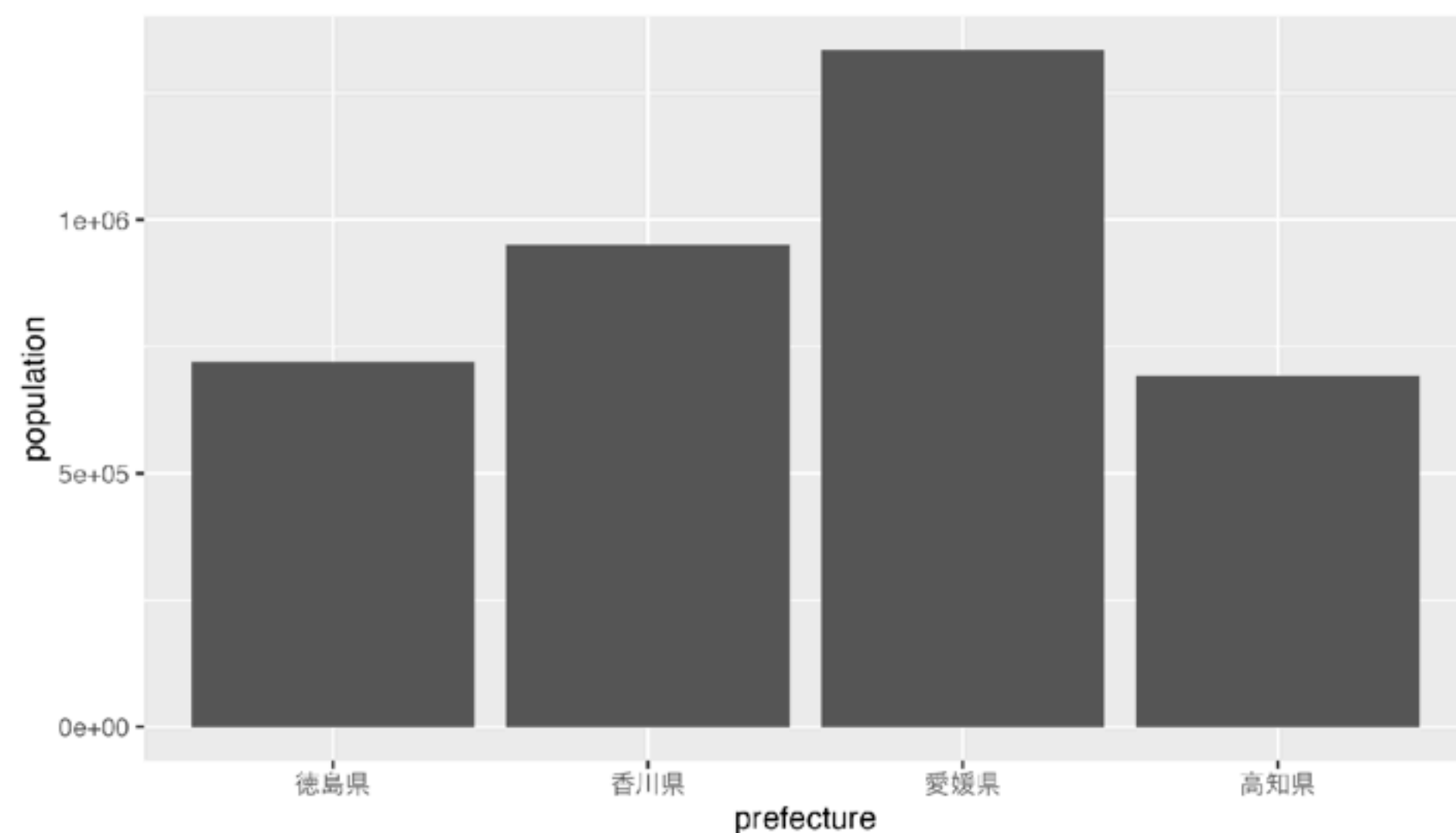
日本語（漢字、片仮名、平仮名）、全角英数字、スペース、記号等は使わない

## ファイルをダウンロードしても開けなくても問題ない（気にしない）

内容の確認、編集はJupyterHub上で行う

# 可視化によって関係・比較を可能にする

表、グラフ（棒グラフ、散布図）    パターンの発見、直感の検証、問題の特定



差がありそう、関連がありそう… 主観的な「ありそう」を客観的に評価するには？

## 相関分析

2つの変数の間の関係を調べるための手法、関連の程度の推定

## 統計的仮説検定

データを用いた統計的推論により仮説が真であるかを評価



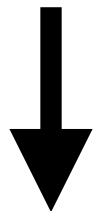
# 分割表

質的変数間の関係を集計、表形式で表現したもの  
2次元分割表… 2つの変数（例えば、学生の学部と進路についてのデータ）

学籍番号	学部	進路	出身
10020230	工学部	就職	愛知県
10030447	文学部	不明	大阪府
...			
10040268	理学部	進学	東京都

学部、進路、出身はいずれも質的変数

学部と進路の各組みあわせをカウント



↓ 学部		進学	就職	不明	計	→ 進路
	理学部	123	80	3	206	
	工学部	152	146	2	300	
	法学部	26	147	7	180	
	文学部	15	154	13	182	
	計	316	524	25	868	

# データ分析における2つの関係

複数の変数がともに変化する状態

データ分析では**相関関係**と**因果関係**の2つの関係を扱う（似て非なるもの）

## 相関関係

ある出来事や物事と別の出来事や物事の間に関係があるもの

ペンギン個体の翼の長さ  $\longleftrightarrow$  ペンギン個体のくちばしの長さ

## 因果関係

ある出来事や物事が**原因**となって、別の出来事や物事（**結果**）が起こるもの

ある水道会社の利用を止める  $\longrightarrow$  水道を利用していた地域のコレラ患者が減る

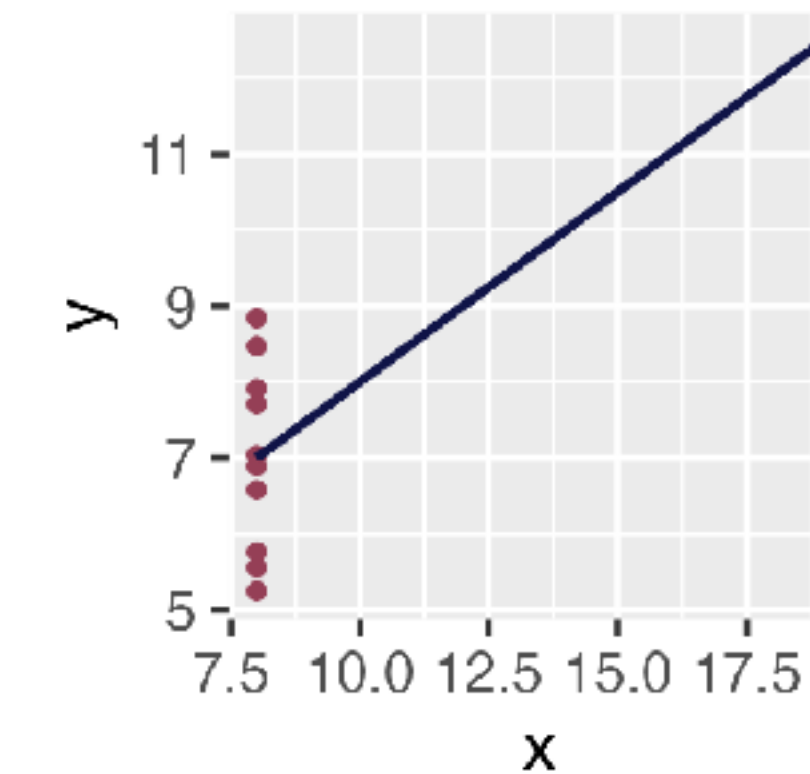
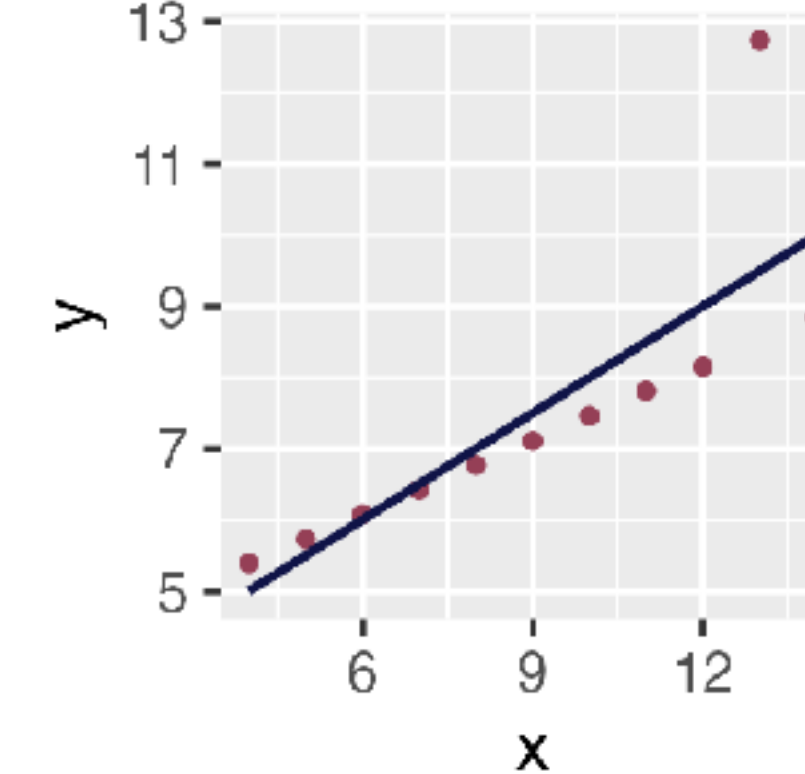
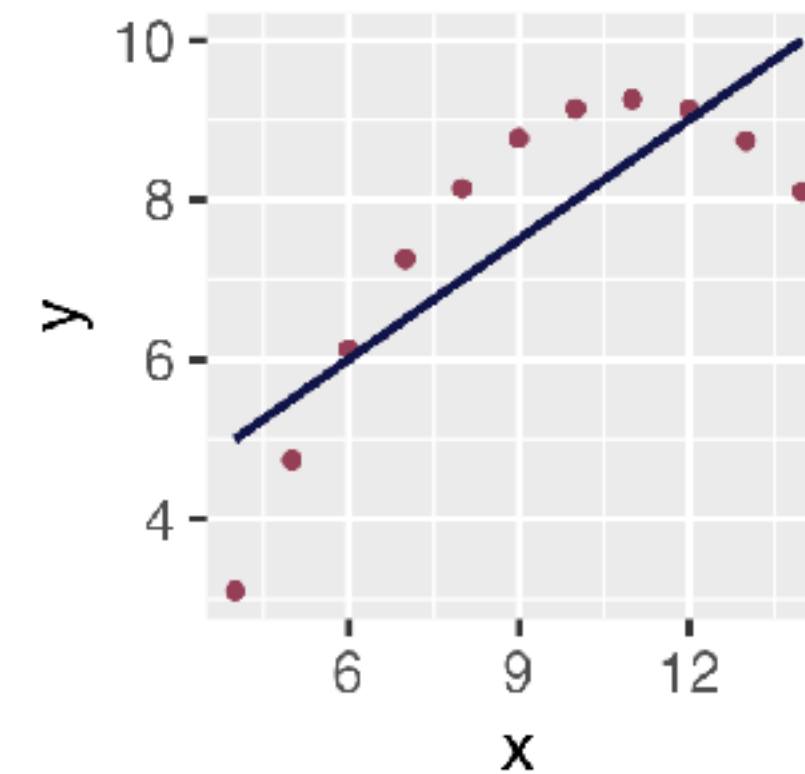
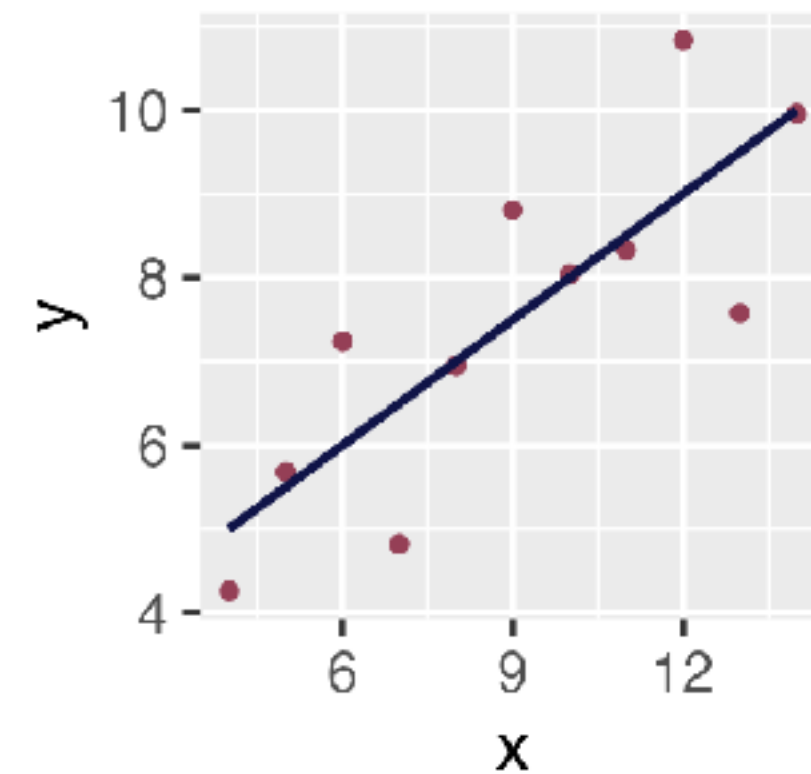
相関関係があるからと言って必ずしも因果関係があるわけではない  
→ 因果関係を調べる「因果推論」

# 【再】アンスコム の例

参考) 第9回: データの可視化

統計量、相関係数がほぼ同じ値になる4種のデータセット  
→ 散布図を描くとデータの傾向が大きく異なる

x1	x2	x3	x4	y1	y2	y3	y4
10	10	10	8	8.04	9.14	7.46	6.58
8	8	8	8	6.95	8.14	6.77	5.76
13	13	13	8	7.58	8.74	12.74	7.71
9	9	9	8	8.81	8.77	7.11	8.84
11	11	11	8	8.33	9.26	7.81	8.47
14	14	14	8	9.96	8.10	8.84	7.04
6	6	6	8	7.24	6.13	6.08	5.25
4	4	4	19	4.26	3.10	5.39	12.50
12	12	12	8	10.84	9.13	8.15	5.56
7	7	7	8	4.82	7.26	6.42	7.91
5	5	5	8	5.68	4.74	5.73	6.89



統計量（相関係数など）と可視化を  
セットで評価することが重要

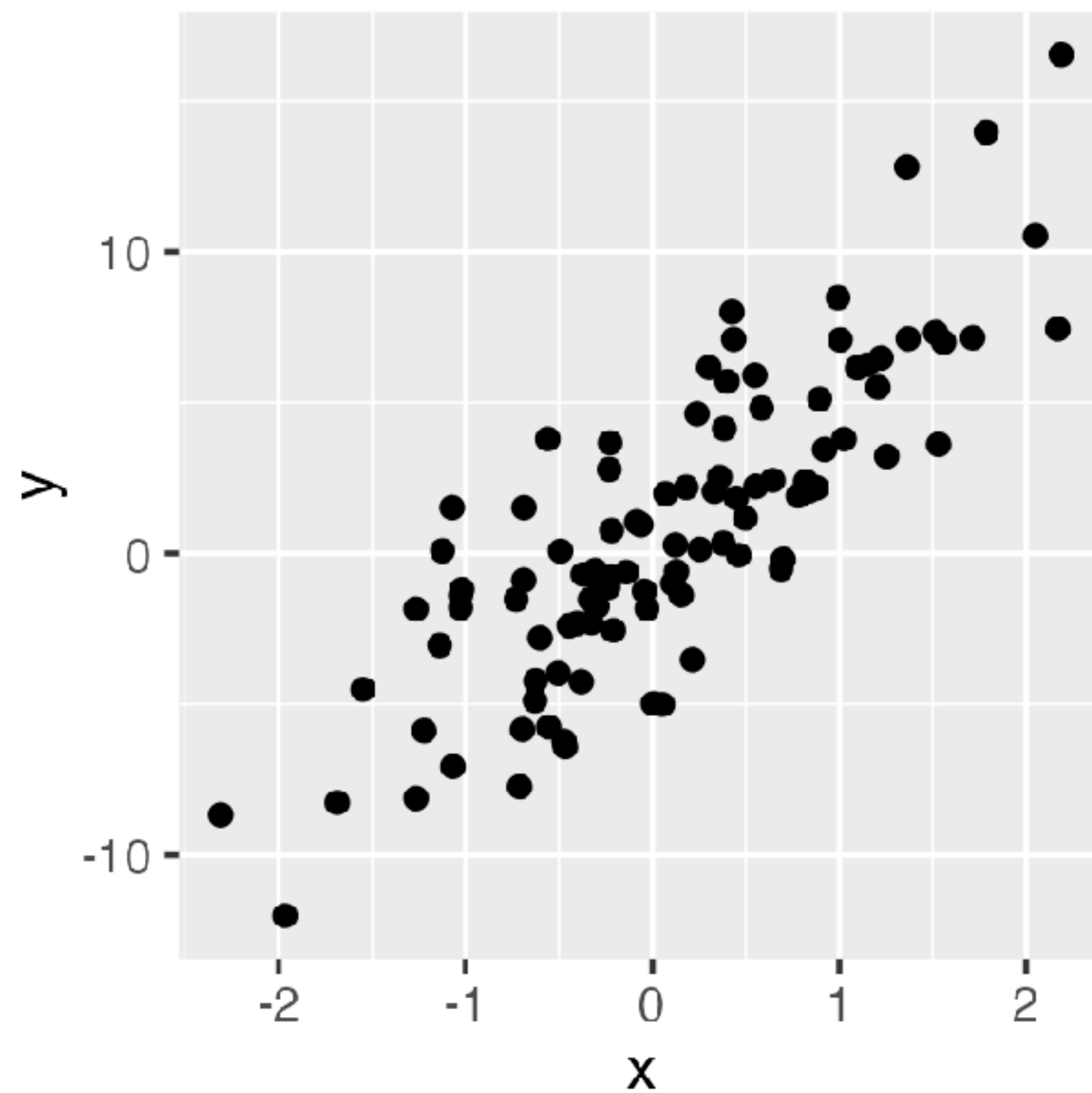


# 相関

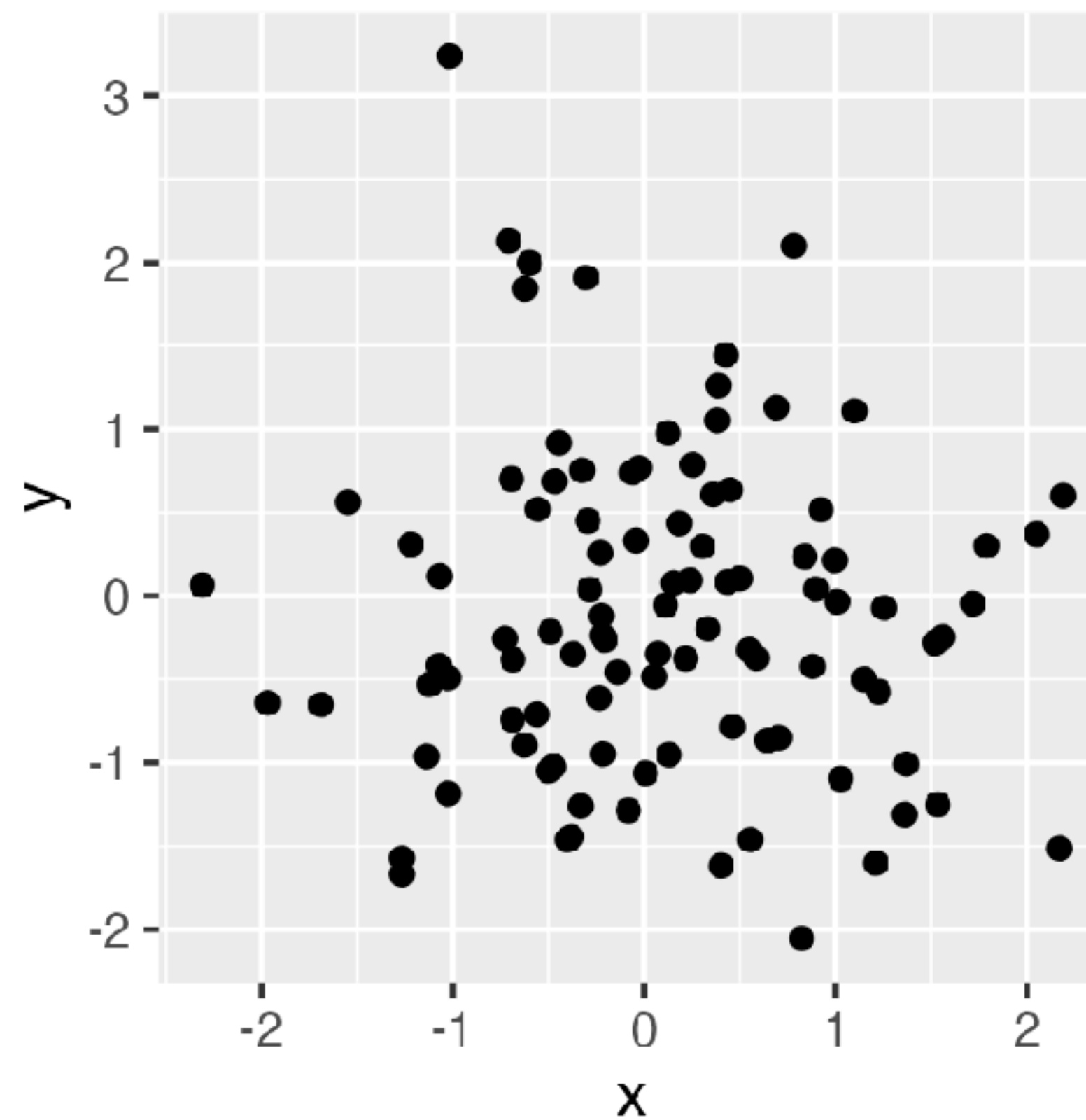
2つの変数間で起こる関係を表す

散布図としてグラフ上に可視化することで傾向を把握しやすくなる

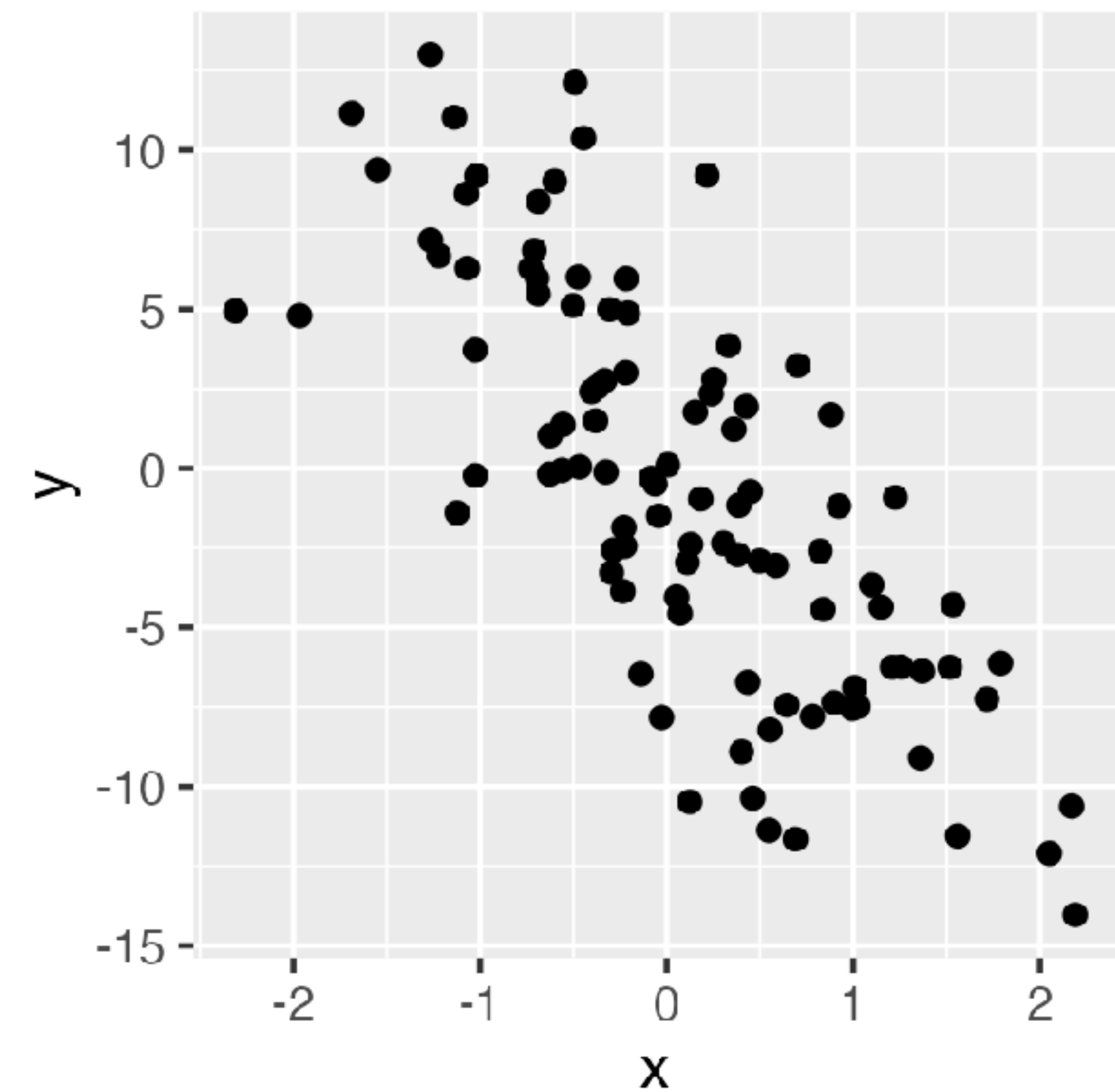
正の相関関係



無相関



負の相関関係



# 関係の数値化1: 共分散 covariance

2つの変数( $x$ と $y$ )についての共分散は次のように求められる

$$Cov_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

## 手順1/2

①、②変数  $x$  ( $y$ ) の値から変数  $x$  ( $y$ ) の平均値を引く → **偏差**

$$Cov_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad \text{①}$$

$$Cov_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad \text{②}$$

$$Cov_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad \text{③ 偏差の積を求める}$$

# 関係の数値化1: 共分散 covariance

## 手順2/2

④  $n$  (すべてのデータ) まで右の処理を行い、それを足し合わせる

$$Cov_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

データの  $i = 1$  番目から

⑤ 変数  $x$  と変数  $y$  の各値に対して偏差を求め、それを掛け合わせたものを足す

$$Cov_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

$n$  (データ数) で割る


```
# Rの標準関数で共分散を求めるとデータの数 - 1で割る不偏共分散になる
cov(penguins$bill_length_mm, penguins$bill_depth_mm, use = "complete.obs")
#> [1] -2.534234
```



# 共分散の特徴

値が大きいほど2変数の関係が強いことを示す  
変数の単位に依存して値が変わる

```
cov(  
  # cm を m に  
  set_units(set_units(df_animal$body_length_cm, cm), m),  
  df_animal$weight_kg,  
  use = "complete.obs")  
#> [1] 66.19572  
cov(  
  # cm を m に  
  set_units(set_units(df_animal$body_length_cm, cm), m),  
  # kg を g に  
  set_units(set_units(df_animal$weight_kg, kg), g),  
  use = "complete.obs")  
#> [1] 66195.72
```



→標準化によって変数間のスケールを揃える

# 関係の数値化2: 相関係数

共分散の単位依存の問題を解消する指標

共分散を各変数の標準偏差の積で割ることで算出される

$$r = \frac{Cov_{xy}}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}}$$

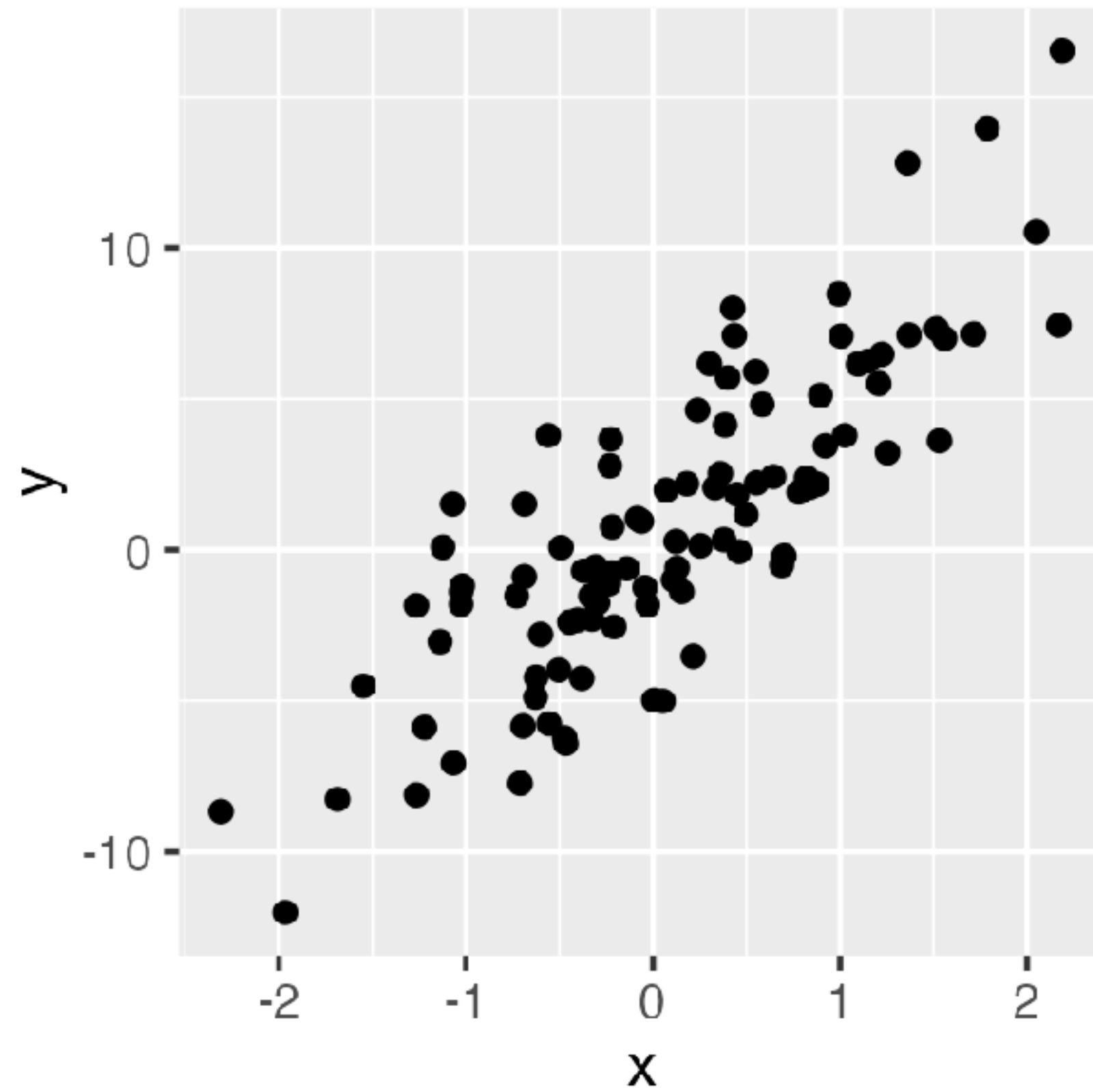
-1 から 1 までの値をとる

変数の関係が強いほど絶対値が1に近づく

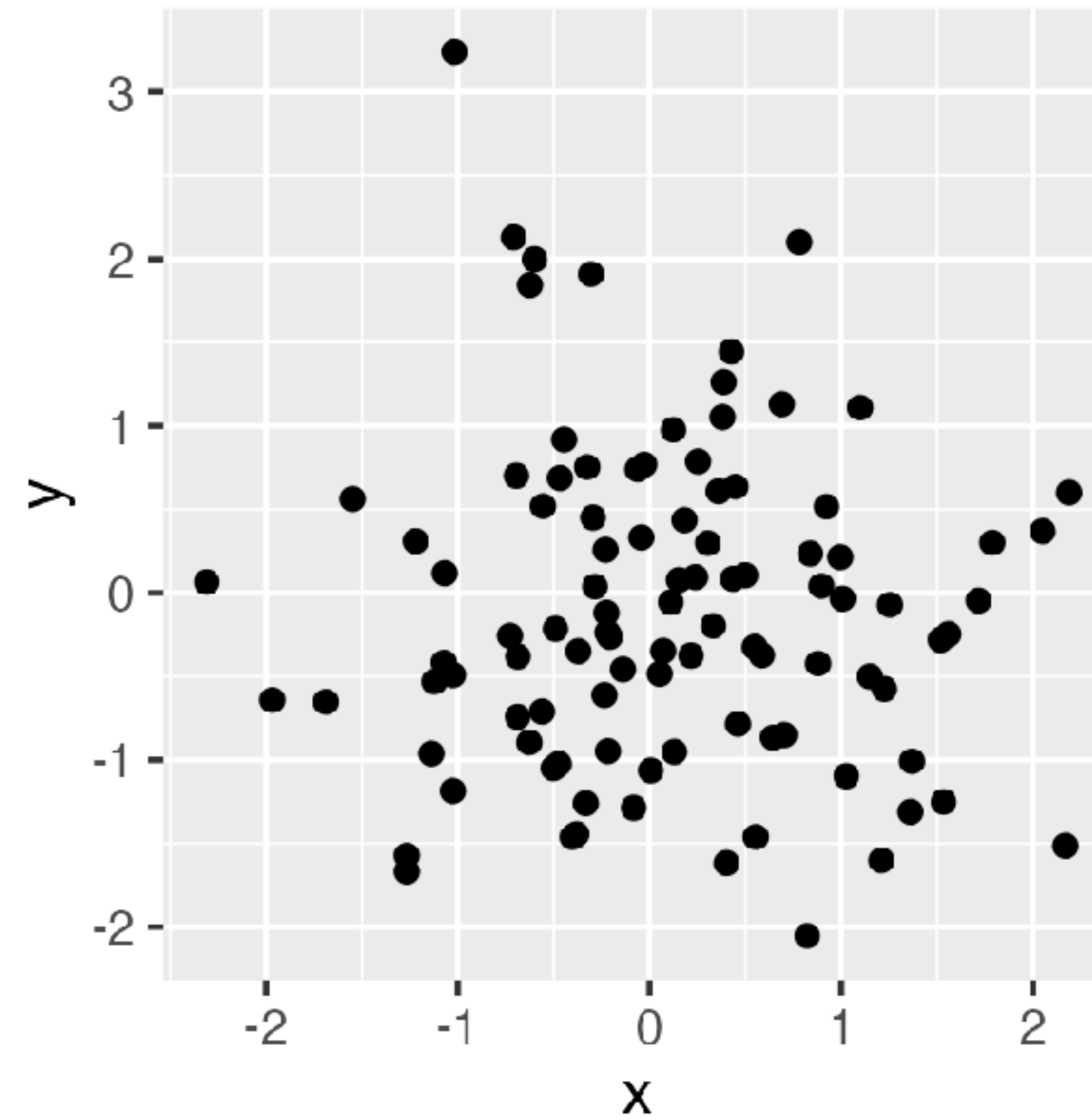
```
cor(penguins$bill_length_mm, penguins$body_mass_g, use = "complete.obs")  
#> [1] 0.5951098
```



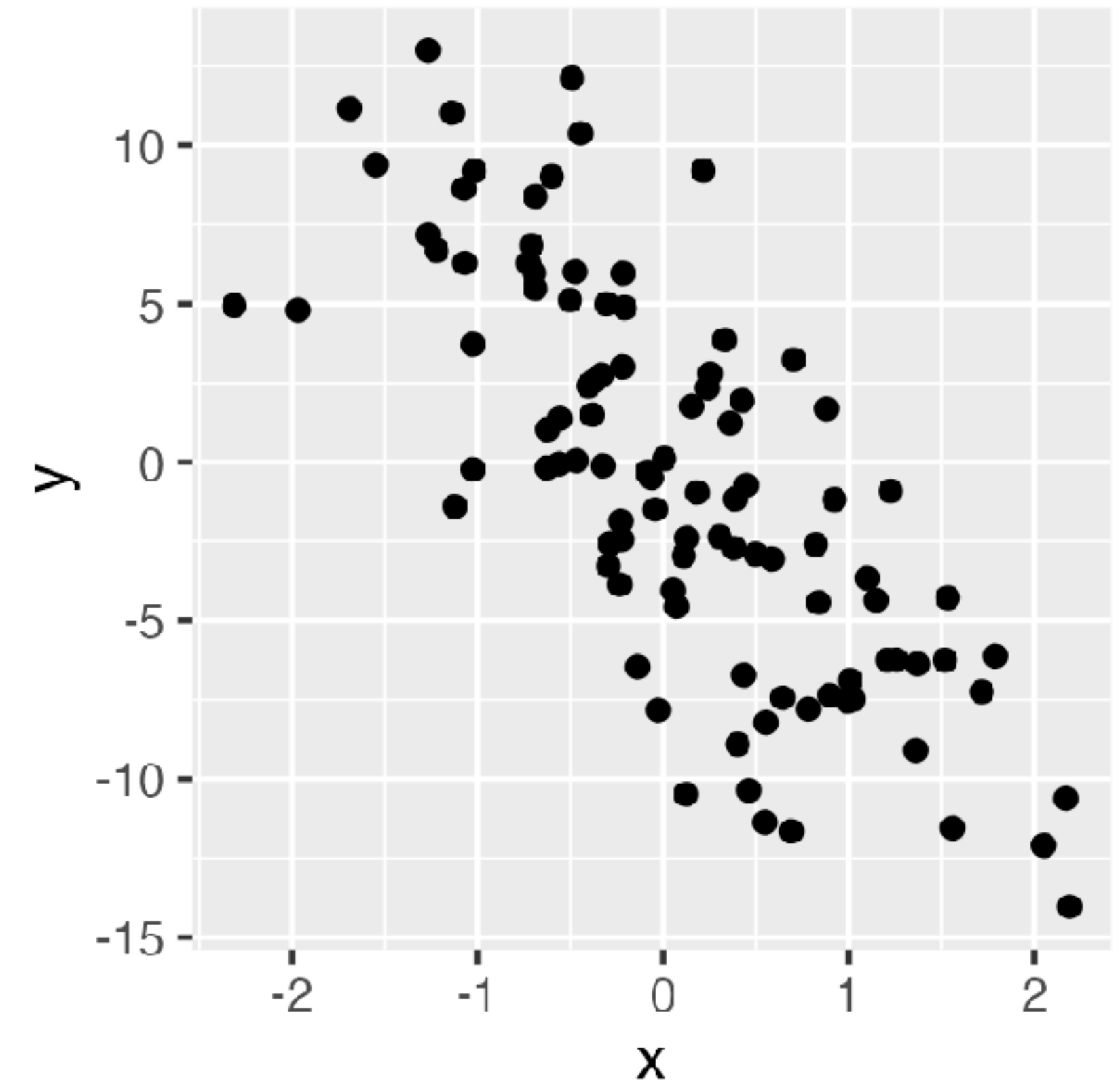
正の相関関係  $r = 0.829446$



無相関  $r = -0.04953215$



負の相関関係  $r = -0.7527922$



## 相関係数の大きさの目安

相関係数	相関の強さ
$\pm 0.7$ 以上	とても強い
$\pm 0.4 \sim 0.7$	やや強い
$\pm 0.2 \sim 0.4$	弱い
$\pm 0.2$ 以下	ほとんどなし

# 相関係数行列

変数のペアごとに計算した相関係数を行列形式で表現する

```
cor(penguins[, c("bill_length_mm", "bill_depth_mm", "flipper_length_mm", "body_mass_g")],  
    use = "complete.obs")
```



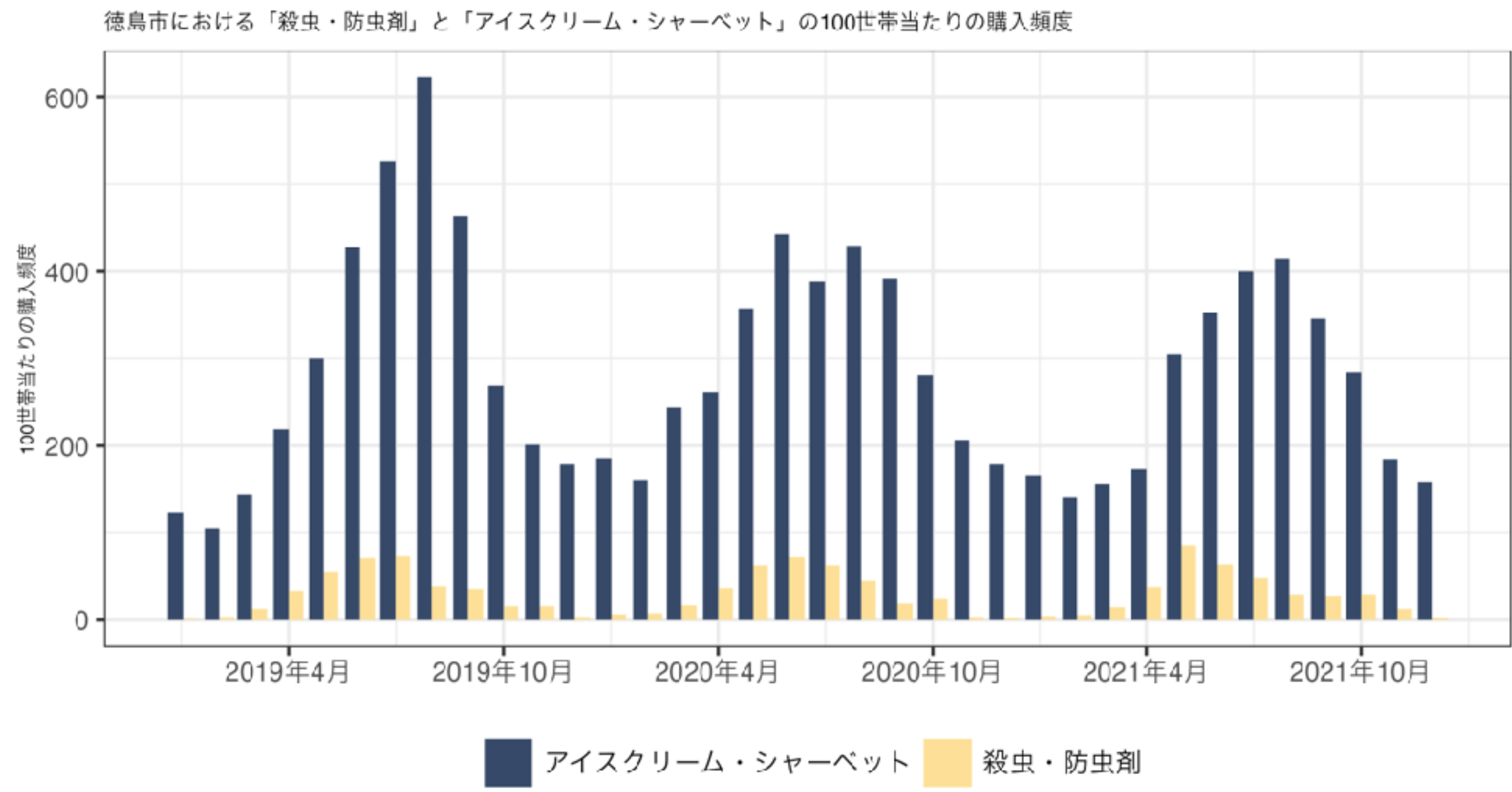
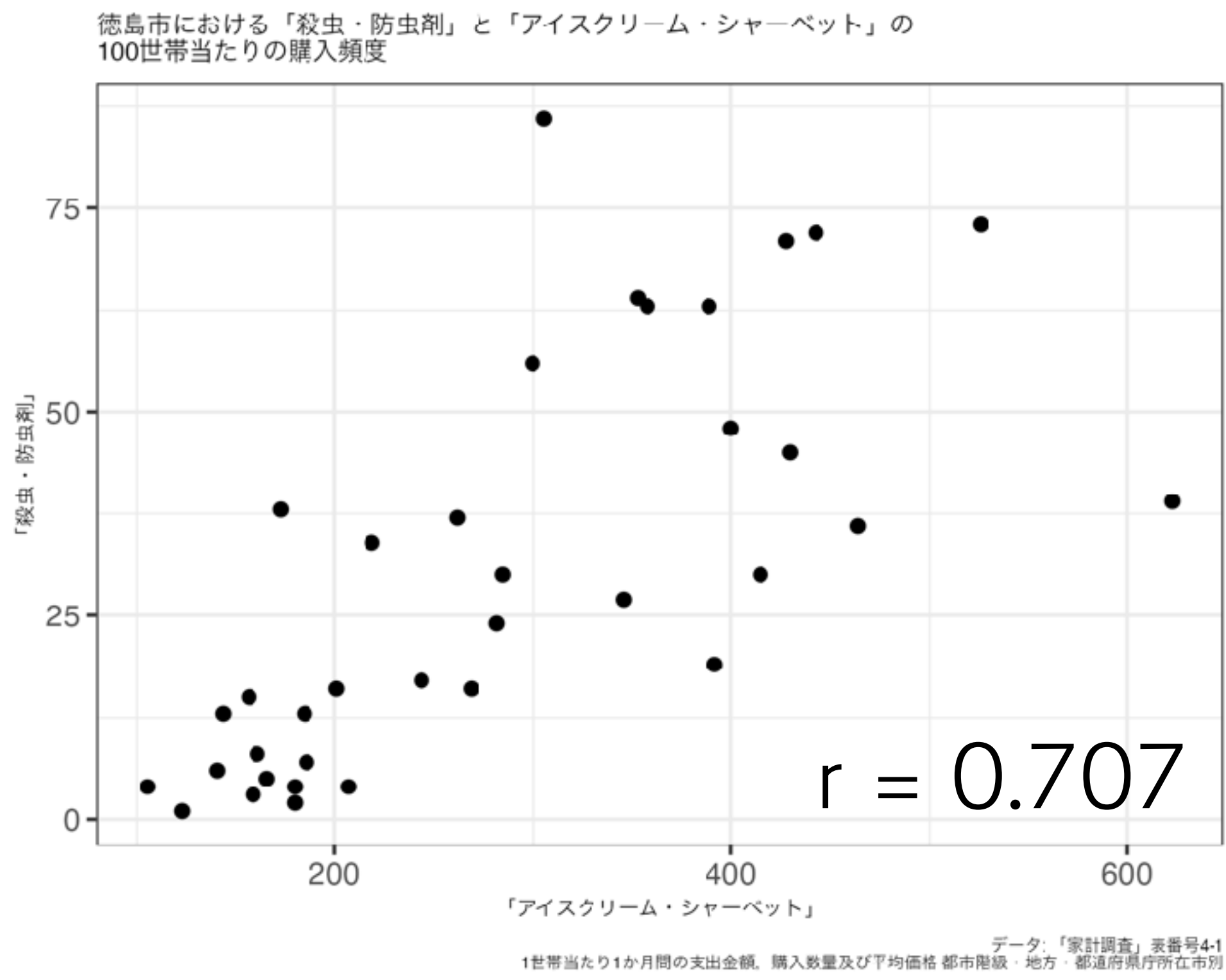
	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
bill_length_mm	1.0000000	-0.2350529	0.6561813	0.5951098
bill_depth_mm	-0.2350529	1.0000000	-0.5838512	-0.4719156
flipper_length_mm	0.6561813	-0.5838512	1.0000000	0.8712018
body_mass_g	0.5951098	-0.4719156	0.8712018	1.0000000

自分自身との相関係数は1

# 疑似相関（見せかけの相関）

因果関係がありそうに見える二変数の関係が、  
観測されていない第三の変数（潜在変数）の効果によってもたらされるもの

例）「殺虫・防虫剤」と「アイスクリーム・シャーベット」の購入頻度… **どちらも気温の影響を受ける**



データ：「家計調査」表番号4-1  
1世帯当たり1か月間の支出金額、購入数量及び平均価格 都市階級・地方・都道府県庁所在市別

→相関から因果関係を導き出すのは難しい。データを得る際の設計・計画が重要



# 参考資料・URL

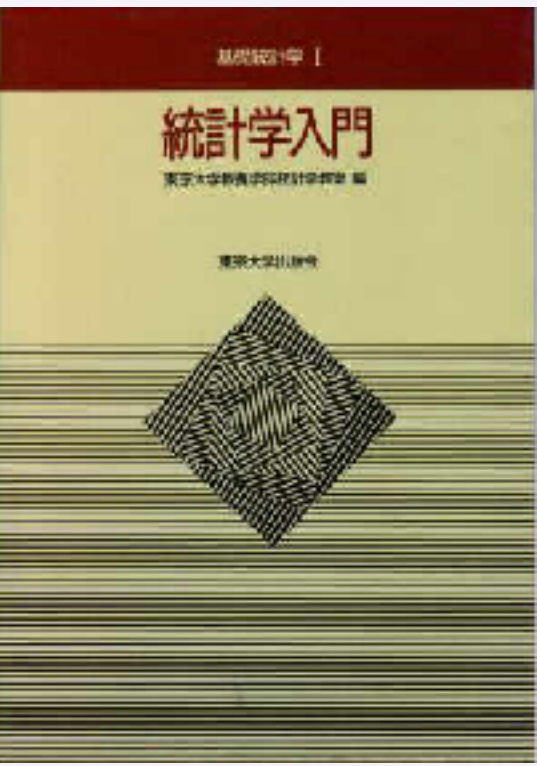
東京大学教養学部統計学教室（編）『基礎統計学I: 統計学入門』（1991）  
東京大学出版会. ISBN: 4-13-042065-8  
瓜生居室: あり、徳大図書館: あり、市立図書館: なし、県立図書館: あり

滋賀大学データサイエンス学部, 長崎大学情報データ科学部（編）『データサイエンスの歩き方』（2022）学術図書出版社. ISBN: 978-4-7806-0936-3  
瓜生居室: あり（電子版）、徳大図書館: あり、市立図書館: なし、県立図書館: なし

中西啓喜（編）耳塚寛明（監修）『教育を読み解くデータサイエンス：データ収集と分析の論理』（2021）ミネルヴァ書房. ISBN: 978-4-623-09172-0  
瓜生居室: あり、徳大図書館: なし、市立図書館: なし、県立図書館: なし

Peter Bruce, Andrew Bruce, Peter Gedeck（著）, 黒川利明（訳）  
『データサイエンスのための統計学入門（第二版）』（2020）オライリー・ジャパン. ISBN: 978-4-87311-926-7  
瓜生居室: あり（電子版）、徳大図書館あり（第一版）、市立図書館: なし、県立図書館: あり

阿部真人『統計学入門：データ分析に必須の知識・考え方』（2021）ソシム.  
ISBN: 978-4-8026-1319-4  
瓜生居室: あり（電子版）、徳大図書館: あり、市立図書館: なし、県立図書館: なし



# 参考資料・URL

📖 日本統計学会（編）『データの分析：日本統計学会公式認定統計検定3級対応』（2020）東京図書.  
ISBN: 978-4-489-02332-3

瓜生居室: あり、徳大図書館: あり（初版）、  
、市立図書館: なし、県立図書館: なし

📖 日本統計学会（編）『統計学基礎：日本統計学会公式認定統計検定2級対応』（2015）東京図書.  
ISBN: 978-4-489-02227-2

瓜生居室: あり、徳大図書館: あり、  
、市立図書館: なし、県立図書館: なし

