

情報科学入門

第13回: データのモデリング

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. 再現可能性

- 9. データの記述
- 10. データの可視化
- 11. データサイエンス応用
- 12. データの関係性
- 13. データのモデリング**
- 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



【課題】 データモデリングについてのクイズ

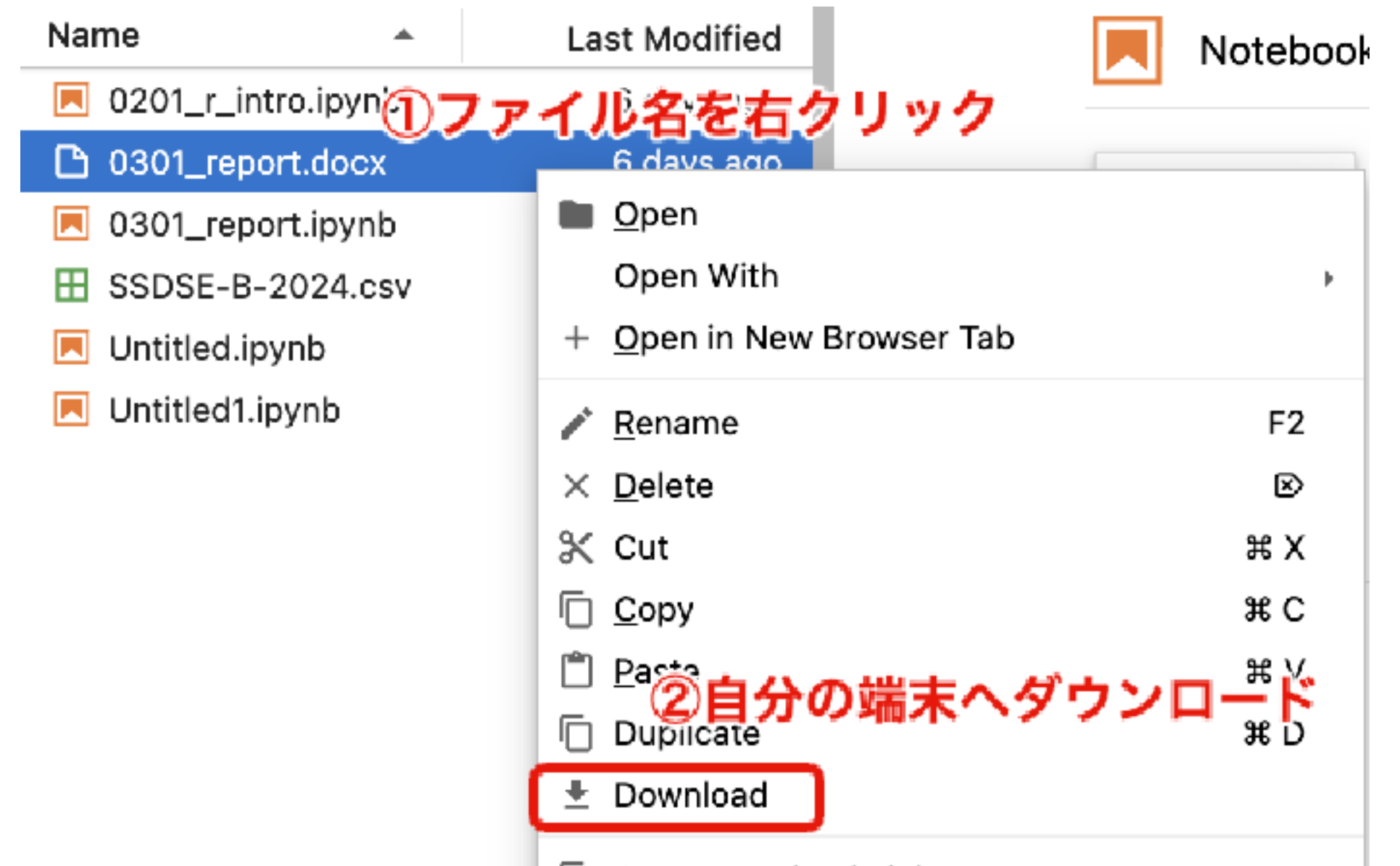
提出期限: 来週の講義開始前まで

manabaのレポートとして提出してください

手順

1. 添付ファイルをダウンロード
2. JupyterHubへアップロード
3. コードやコメントを記述、実行
4. 保存
5. ダウンロードしたファイルをmanabaへアップロード

メニュー上の「ファイル」から「ダウンロード」



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

データモデリングの概要

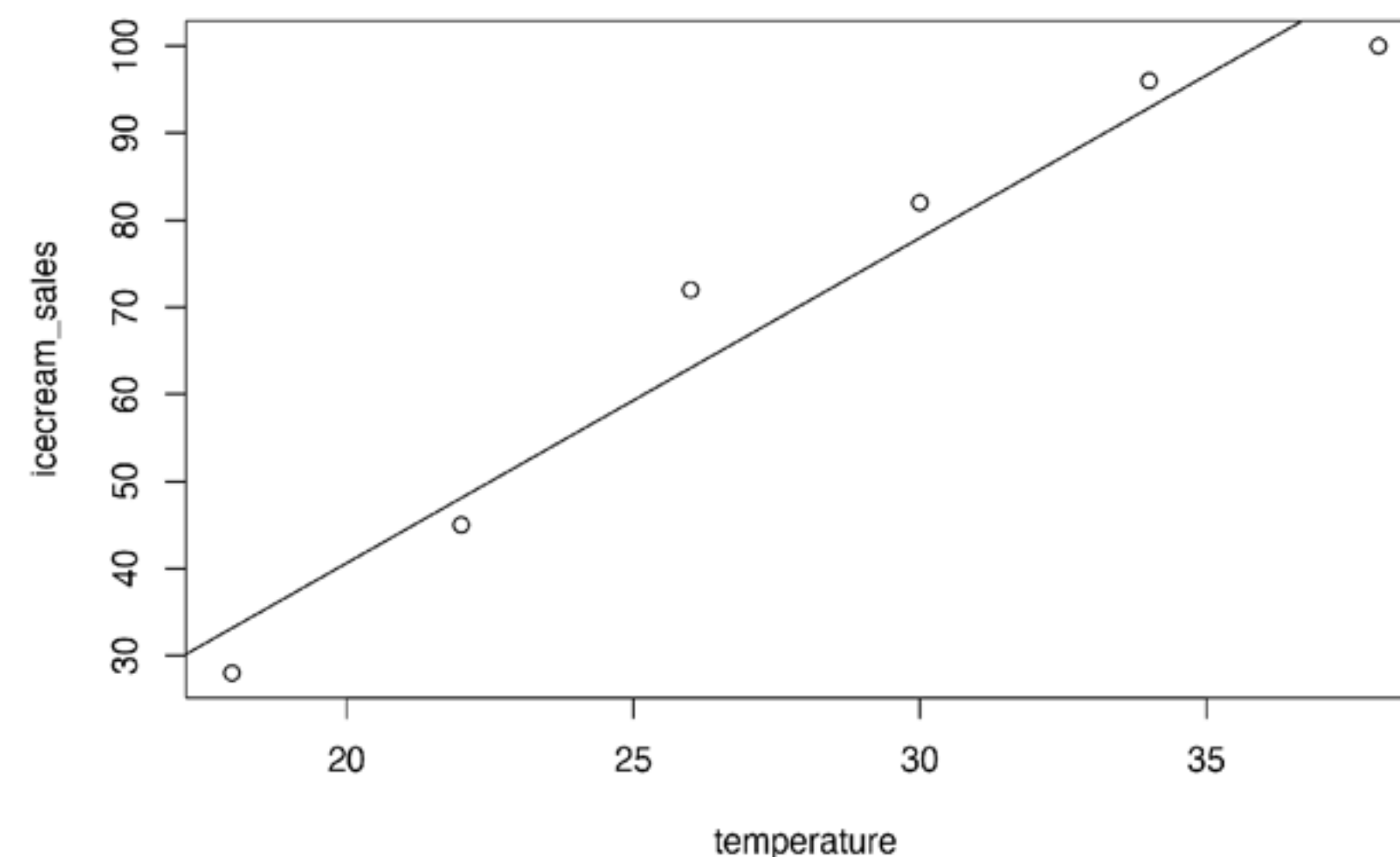
モデリング

モデル… 世の中の複雑な現象を表現する

データの背景にあるメカニズムを推論

をもとに予測

から意思決定



気温が1℃上がるとアイスクリームの売り上げはどう変化する？
気温32℃のときのアイスクリームの売り上げはどのくらいになる？
気温が上がるほどアイスクリームの売り上げは増加する？

「全てのモデルは間違っている、それでもいくつかのモデルは役に立つ」

正しいモデルは存在しない、どこまで突き詰めても現象の近似にほかならない

George E. P. Box

統計モデル

観測されたデータを用いて、現象との関係性を表現する数式を仮定する

→帰納的な手法 ある具体的な観察事例から一般原理や法則を導き出す論理的な推論

なにをする？... （確率分布の形を決める） パラメータを推定

二項分布における成功率

正規分布における平均値、分散

回帰分析でのモデルの係数

どうやって？

最小二乗法… 線形モデル

最尤推定法… 一般化線形モデル

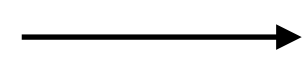
マルコフ連鎖モンテカルロ法… ベイズモデル

回帰分析の概要

事象の関係性を線形モデルで表現

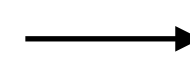
変数を用いて予測を行う

気温 ↗



アイスクリームの売り上げ ↗

学習時間 ↗



テストの成績 ↗

線形モデル

変数の間の関係を等式で表現

回帰分析や分散分析

単純な2変数の関係を表現→単回帰モデル



目的変数

切片

傾き 説明変数

誤差項

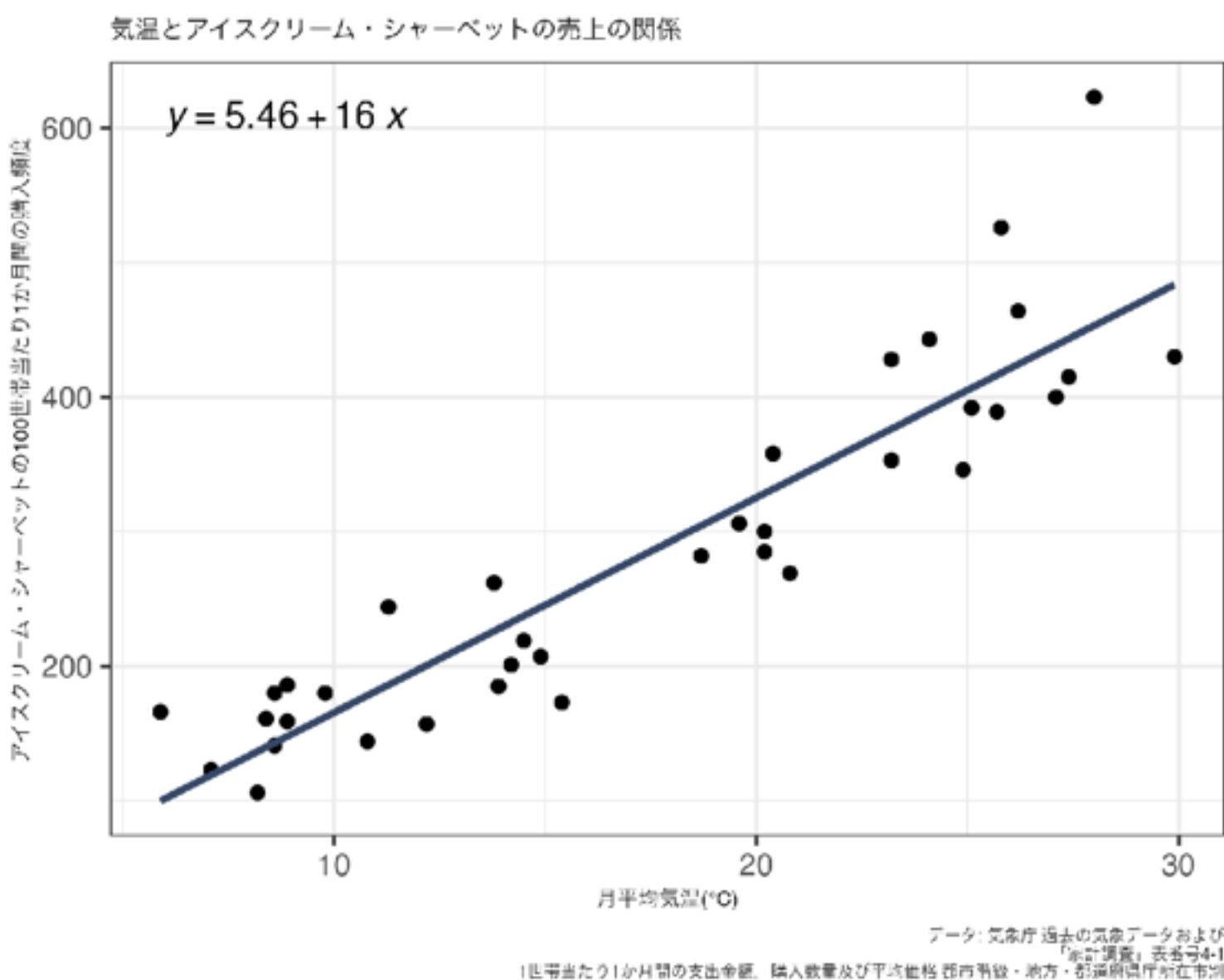
$$y = a + bx + \epsilon$$

x が0のときの y の値

a と b がパラメータ
→ 回帰係数

y 目的変数（被説明変数、従属変数、出力変数）... モデルによって表現される値

x 説明変数（独立変数、入力変数）



回帰直線

「切片aと傾きbからなる回帰直線上のxの値によってyの値が決まる」と考える

平均気温が高くなればアイスの売り上げが上がる？

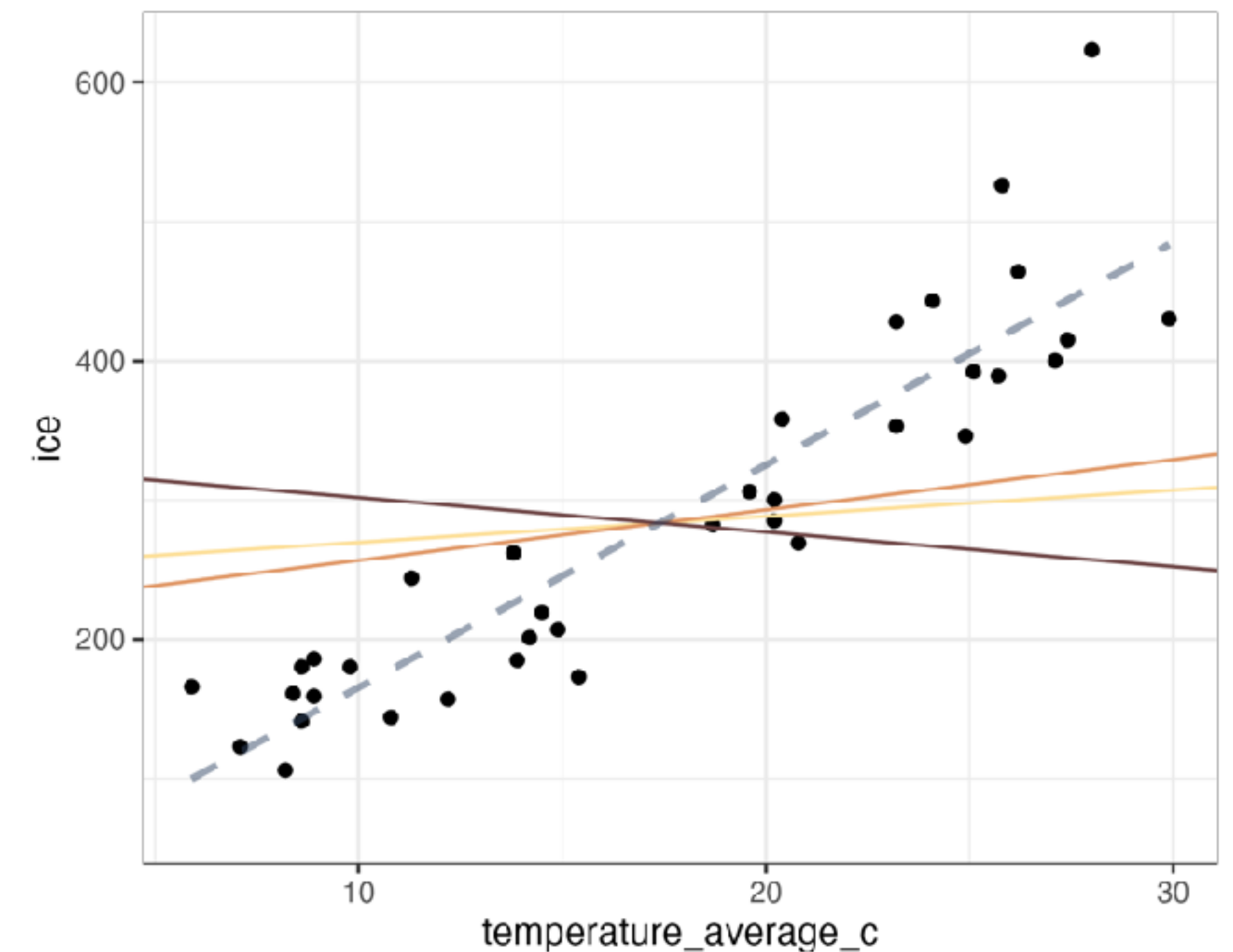
気温が1℃上がればアイスの購入頻度が16増える

回帰直線の係数と気温からアイスの売り上げを予測する

アイスの売り上げ = $5.46 + 16.0 \times 30$ (気温)

切片と傾きが異なれば回帰直線の形も異なる

2つの変数の関係を最も説明できる回帰直線とは？



さまざまな回帰係数からなる回帰直線

最小二乗法

回帰直線から残差 (residual, 観測値 y から予測値 \hat{y} のズレ… 誤差) を求める

予測値 \hat{y} は回帰直線の係数から推定する

$$\hat{y}_i = 5.46 + 16x_i$$

$$x_1 = 7.1$$

$$\hat{y}_1 = 5.46 + 16 \times 7.1$$

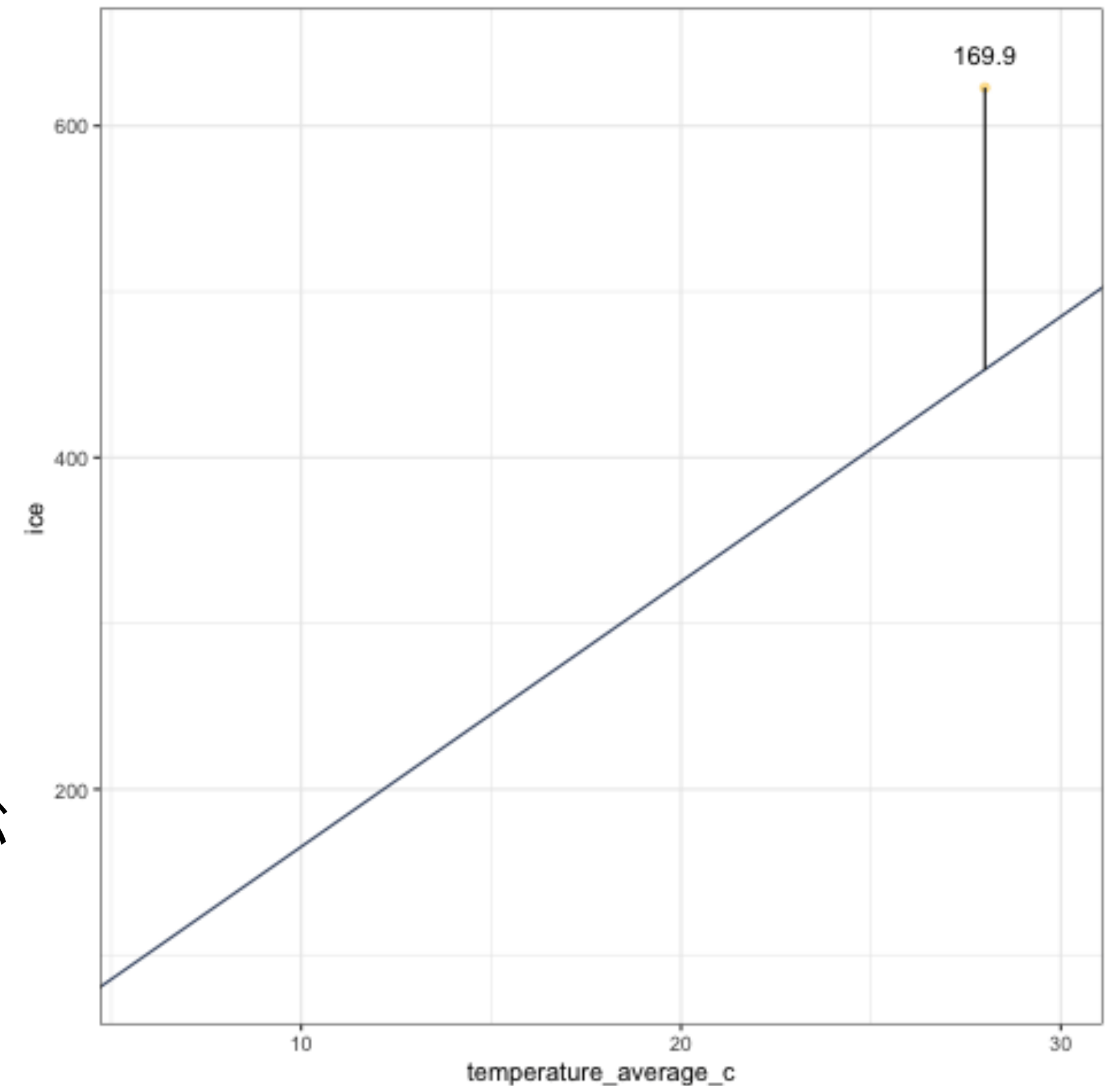
$$\hat{y}_1 = 118.9$$

$$residual_1 = y_1 - \hat{y}_i$$

$$y_1 = 123$$

$$123 - 118.9$$

残差平方和 (各残差を二乗した結果を合計する) が
最小となる定数項 (傾きと切片) を求める

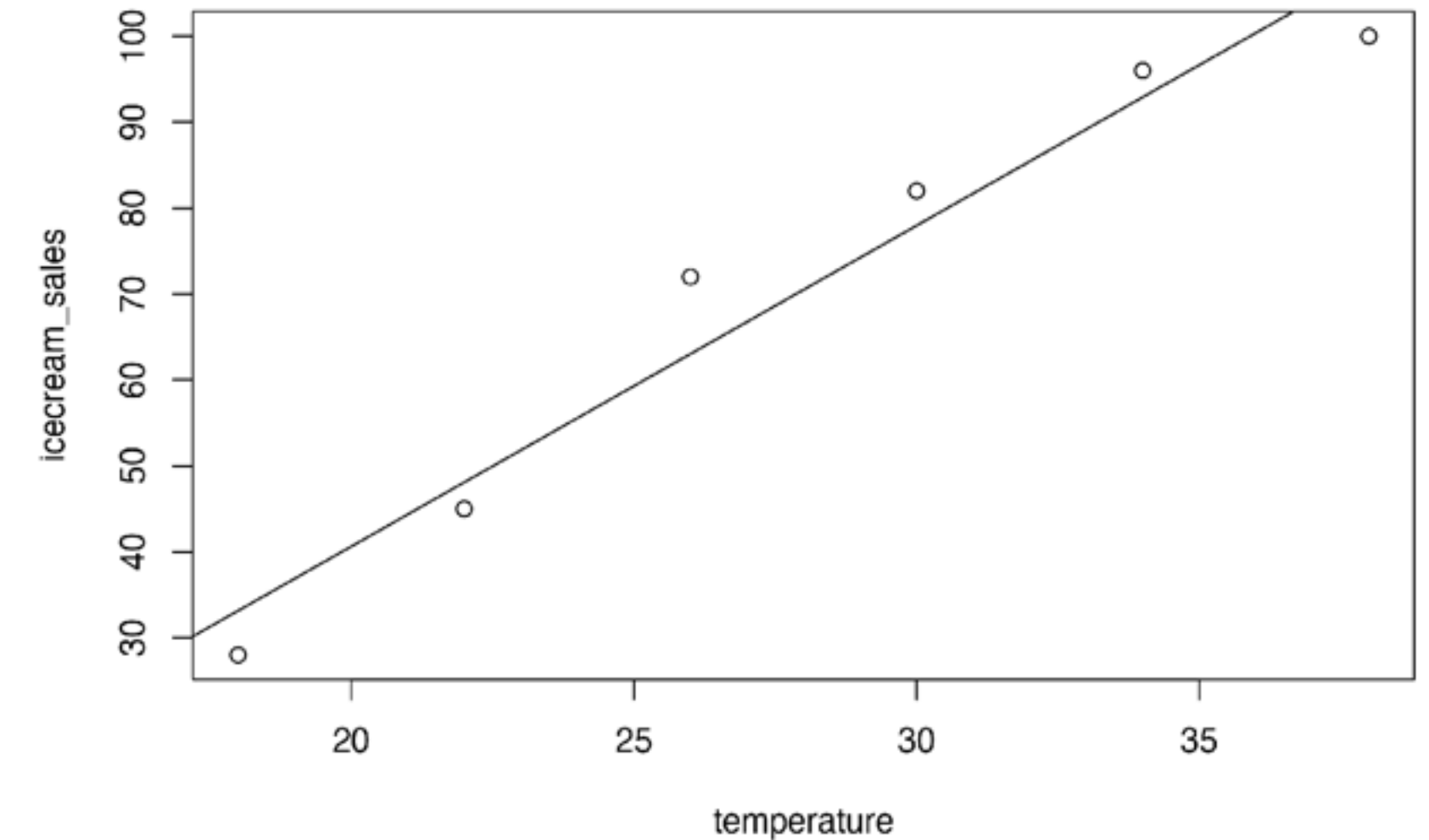


残差平方和を最小にすることにより当てはめた回帰直線と気温。縦の棒が残差を示す。

アイスクリームの売り上げと気温の関係

散布図による関係の可視化

```
# 気温のデータを作成
temperature <- seq(18, 40, 4)
# アイスクリームの売り上げのデータを作成
icecream_sales <- c(28, 45, 72, 82, 96, 100)
# 散布図の作成
plot(temperature, icecream_sales)
```



```
# 単回帰モデルの構築
# lm( )関数の引数に 目的変数 ~ 説明変数 の形式で指定する
model <- lm(icecream_sales ~ temperature)
```



アイスクリームの売り上げと気温の関係

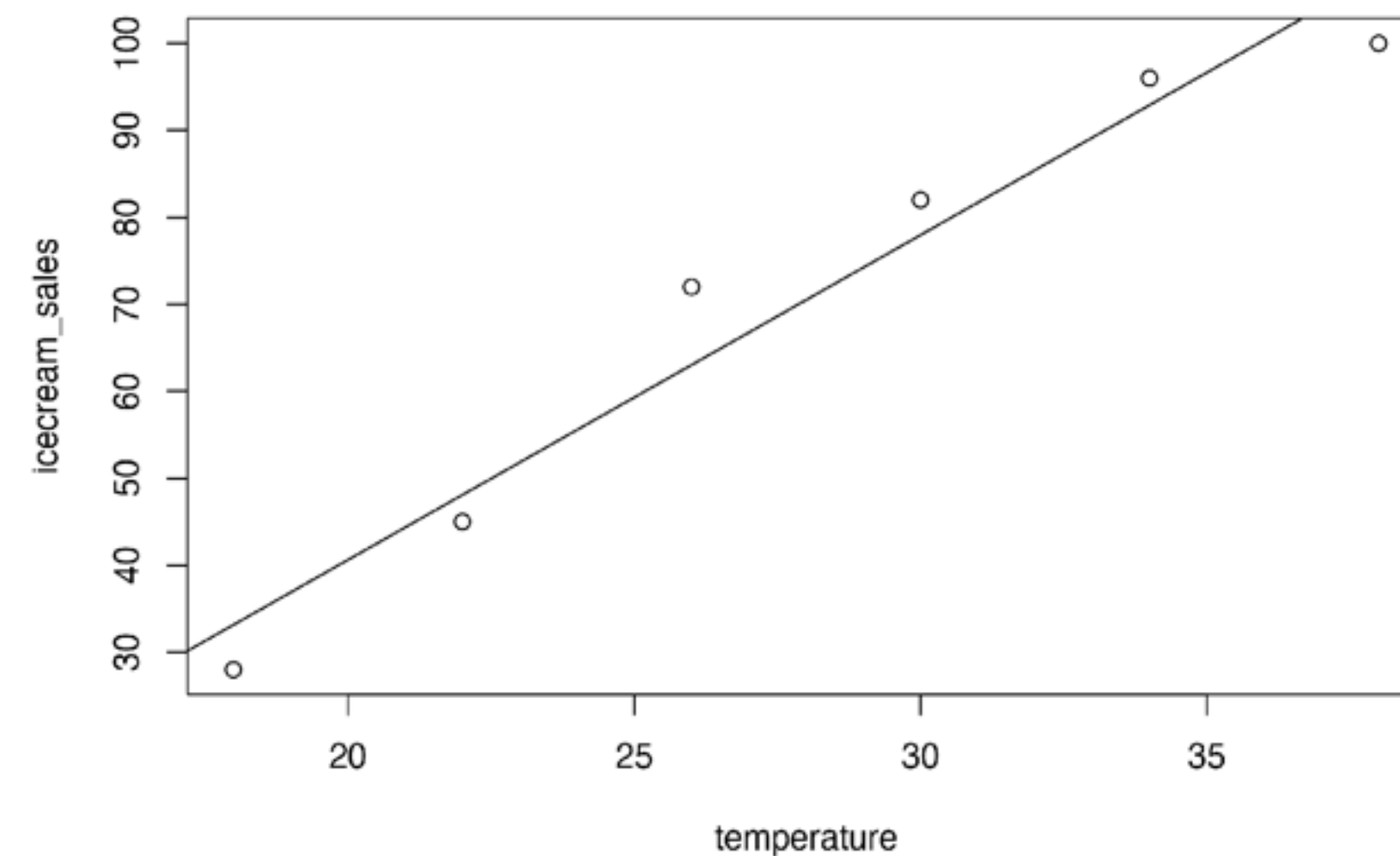
傾きと切片の値を確認

```
# intercept... 切片
# coefficient... 傾き ( 係数 )
coefficients(model)
#> (Intercept) temperature
#> -34.100000      3.735714
```



回帰係数… 説明変数の値が1単位増えたとき、目的変数の値がどれだけ増えるか

```
# 回帰直線の追加
abline(model)
```



アイスクリームの売り上げと気温の関係

新たなデータに対する予測

```
predict(model, data.frame(temperature = 30))  
#> 1  
#> 77.97143
```


回帰モデルにおけるp値の扱い

係数の有意性およびモデルの適合度の評価に用いられる

p値を含む結果の出力

summary(model)

#> Coefficients:

#> Estimate Std. Error t value Pr(>|t|)

#> (Intercept) -34.1000 12.2987 -2.773 0.050194 .

#> temperature 3.7357 0.4267 8.754 0.000938 *** 係数の有意性を示す

#> ---

#> Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#>

#> Residual standard error: 7.14 on 4 degrees of freedom 決定係数

#> Multiple R-squared: 0.9504, Adjusted R-squared: 0.938

#> F-statistic: 76.64 on 1 and 4 DF, p-value: 0.0009384 モデルの適合度を示す



重回帰分析

複数の説明変数を用いて目的変数の予測を行う

$$y = a + b_1 x_1 + \dots + b_k x_k + \epsilon$$

目的変数 切片 係数 (偏回帰係数) 説明変数 係数 (偏回帰係数) 説明変数 誤差項

重回帰モデルの構築



```
humidity <- c(0.65, 0.8, 0.75, 0.85, 0.9, 0.8)
```

```
wind_speed <- c(2, 3, 4, 6, 3, 1)
```

```
model <- lm(icecream_sales ~ temperature + humidity + wind_speed)
```

+演算子を使って説明変数をつなげる

```
coefficients(model)
```

```
#> (Intercept) temperature humidity wind_speed
```

```
#> -43.030028 3.829201 -3.741047 2.928650
```

回帰分析のまとめ

与えられたデータに基づき、説明変数と目的変数の関係を表現する
数学的なモデルを構築する

予測値と実際の値の差が最小となるようなモデルを得るために
最小二乗法による係数を推定する

説明変数の数に応じて、単回帰モデルと重回帰モデルを使い分ける

一般化線形モデル

確率分布の選び方

モデリングを開始する前に、データの性質を理解することが重要

- 説明したいデータの**種類**（離散値か連続値か）
- 説明したいデータの**範囲**（ゼロから無限大までか、あるいは有限の範囲か）
- 説明したいデータの**分散**（標本分散） **と平均**（標本平均） の関係

→データの性質に合った確率分布を選ぶ

最小二乗法は等分散正規分布に従う誤差項を持つ線形回帰モデルに適用される

非等分散性や非正規性を持つデータに対しては、最小二乗法の代わりに、
ロバスト回帰や一般化線形モデルなどの手法を検討する

→非線形な関係の表現を可能にする

正規分布では適切に表現できないデータ

例) 植物個体ごとの種子数 (離散値、非負)

ウェブサイト上での1時間あたりの訪問者数 (離散値、非負)

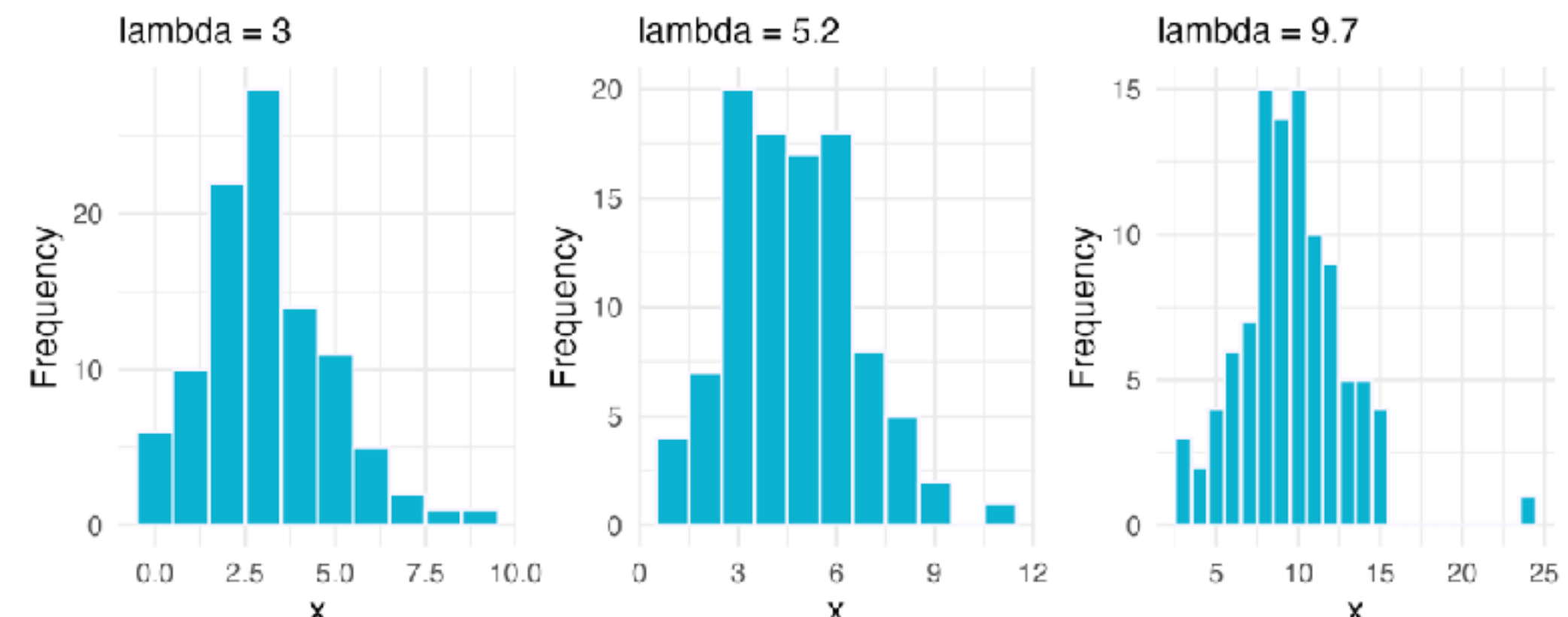
生存率や死亡率といった比率 (連続値、0 から 1 の範囲)

本来、ポアソン分布や二項分布に従うようなデータの確率分布に正規分布 (無限の範囲、負の値を取り得る) を仮定することは不適切

ポアソン分布

$$P(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

パラメータ… 平均値 λ = 分散



ゼロから無限大までの離散値を扱う

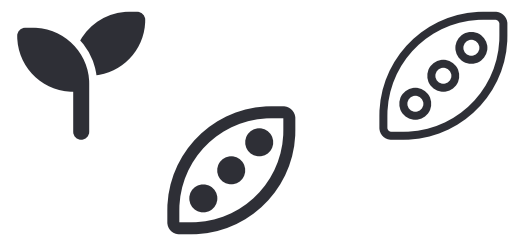
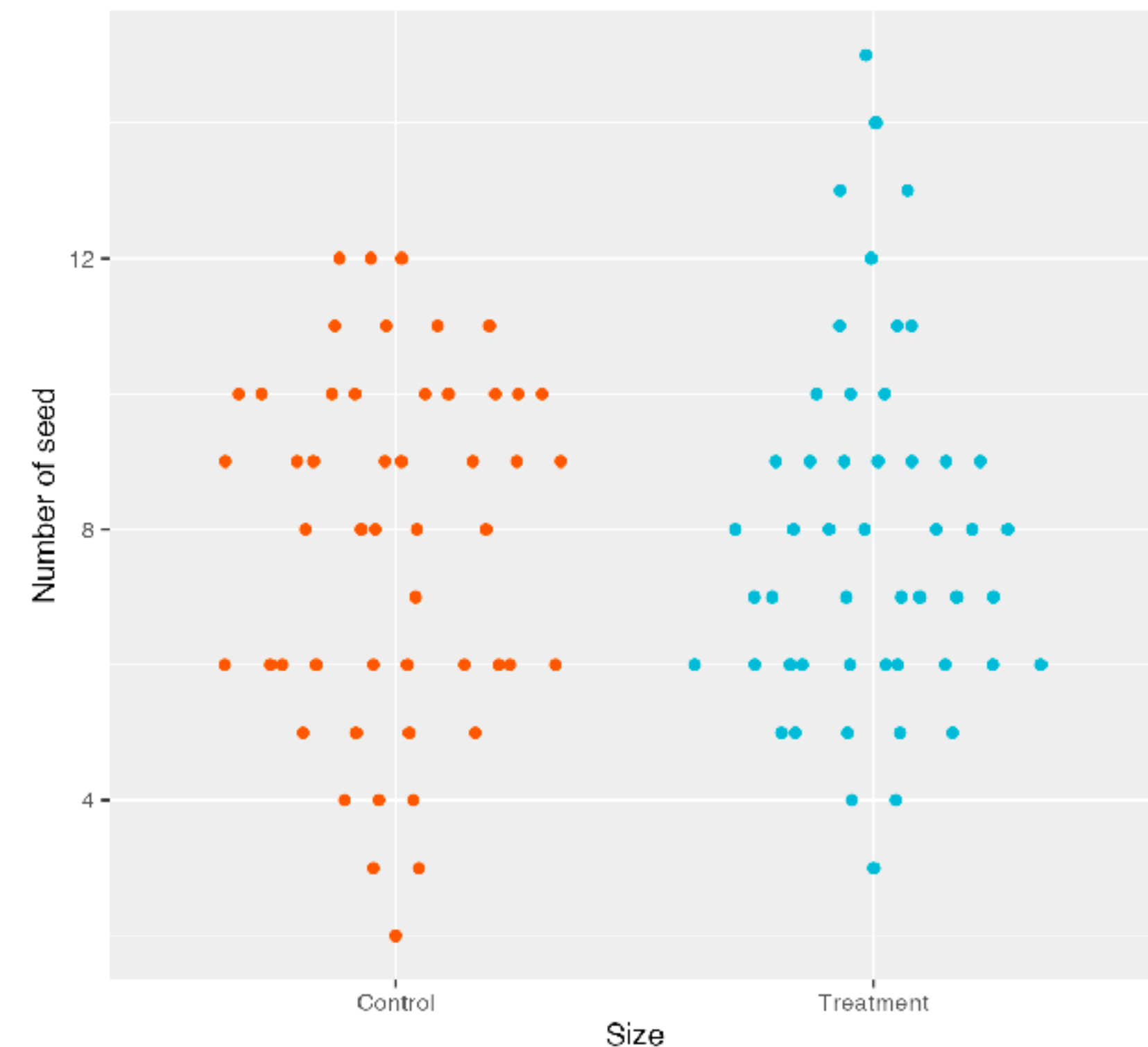
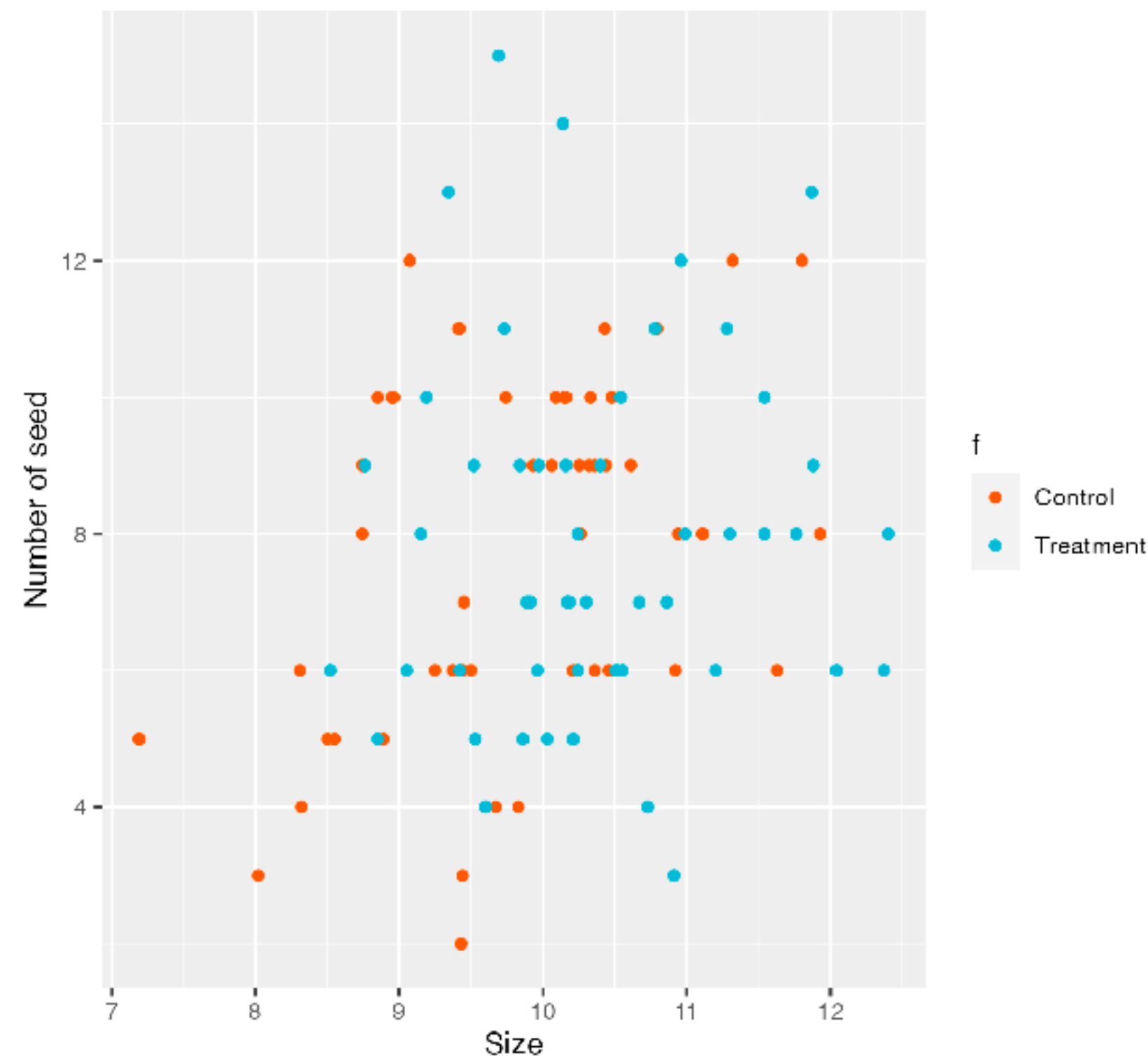
平均値 λ が大きくなると、正規分布に近づく

植物の種子数（カウントデータ）の統計モデリング

```
# 『データ解析のための統計モデリング入門』（久保2012）からデータを利用  
# 植物の種子数と体サイズについての架空のデータ
```



```
d <- read.csv("https://kuboweb.github.io/~kubo/stat/iwanamibook/fig/  
poisson/data3a.csv")
```



一般化線形モデル

線形モデルの拡張（一般化）

最小二乗法… 線形モデルのパラメータ推定。正規分布に従うことを仮定する
→ 正規分布しないデータについてもモデルを考えたい

最尤推定法

ゆうど

尤度が最も高くなるような値をパラメータの推定値とする

尤度… 「当てはまりの良さ」をあらわす統計量

データが特定の確率分布に従うと仮定し、その確率分布の確率関数を使用して尤度関数を構築

尤度関数… パラメータを入力として、観測されたデータが得られる確率を表現する関数

尤度関数を最大化するようなパラメータの値を求める

多様な確率分布を仮定した統計モデルにも適用可能

一般化線形モデルの構成要素

確率分布… 対象のデータを生み出す確率論的な過程を示す、
確率関数と確率変数を取り得る値をまとめたもの

線形予測子… 説明変数の関係を線形結合（Rでは+で表現）によって示したもの

リンク関数… 目的変数と線形予測子を対応付ける関数

一般化線形モデルでは、リンク関数を通じて、線形予測子が目的変数の確率分布のパラメータにマッピングされる

一般化線形モデルの構成要素

確率分布

線形予測子

リンク関数

正規分布

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_1 x_{i3}$$

恒等関数

入力をそのまま出力とする関数

ポアソン分布

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_1 x_{i3}$$

対数関数

逆関数である指数関数の働きにより、
λ が負の値とならないようになる

二項分布

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_1 x_{i3}$$

ロジット関数

逆関数であるロジスティック関数の働きにより、
0から1の値を取る確率として捉えることができる

$$P(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

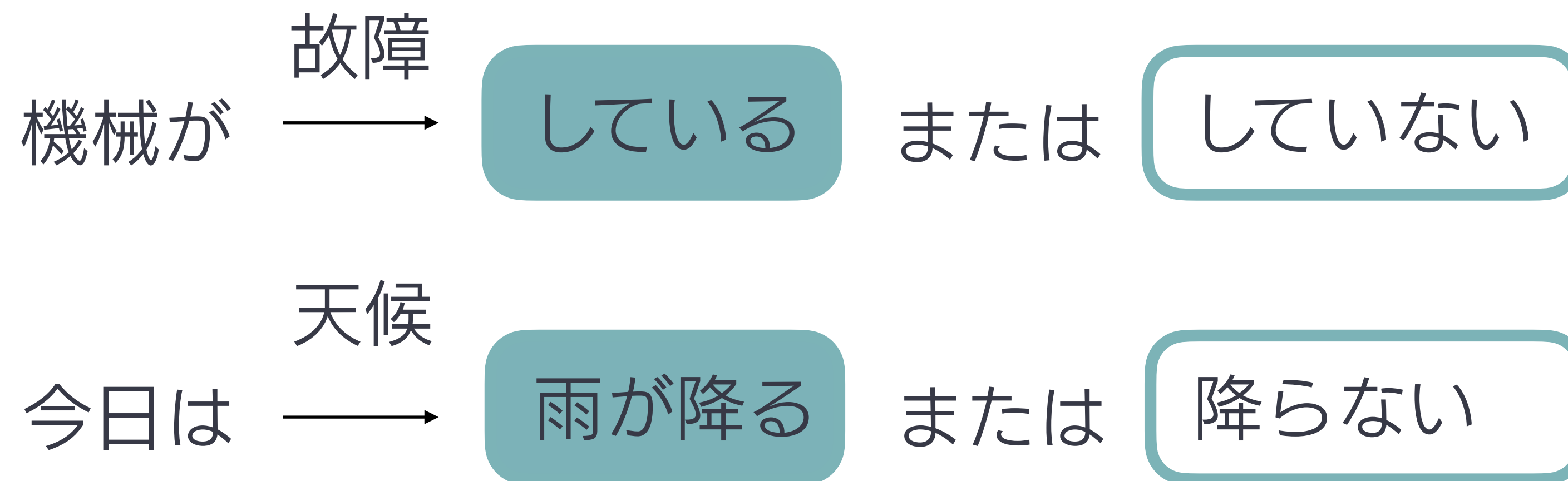
一般化線形モデルの適用と評価

```
model <- glm(y ~ x + f, data = d, family = poisson)
summary(model)
# ...
#> Coefficients:
#>      Estimate Std. Error z value Pr(>|z|)
#> (Intercept)  1.26311     0.36963   3.417 0.000633 ***
#>      x        0.08007     0.03704   2.162 0.030620 *
#>      fT       -0.03200     0.07438  -0.430 0.667035
#> ---
#>      Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> (Dispersion parameter for poisson family taken to be 1)
#>
#> Null deviance: 89.507  on 99  degrees of freedom
#> Residual deviance: 84.808  on 97  degrees of freedom
#> AIC: 476.59
```



分類モデル

二値… 二値のうち、どちらになるかを調べる



ロジスティック回帰

どちらに属するかを確率をもとに判断する

```
dplyr::glimpse(df_weather)
#> Rows: 124
#> Columns: 4
#> $ pressure      <dbl> 1010.3, 1012.0, 1018.5, 1021.3, 1021.5, ...
#> $ humidity      <dbl> 67, 59, 47, 63, 70, 67, 68, 51, 64, 66, 84, ...
#> $ temperature   <dbl> 16.1, 15.2, 15.6, 17.0, 18.3, 20.0, 20.4, ...
#> $ weather       <fct> 雨, 雨, 雨以外, 雨以外, 雨以外, 雨以外, ...
```



気象庁2022年5月の気象データ

気象に関する変数（気圧、湿度、気温）から「雨」か「雨以外」かを予測

ロジスティック回帰

ロジスティック回帰モデルの作成



```
model <-
```

```
  glm(weather ~ temperature + humidity + pressure, data = df_weather,  
       family = binomial)
```

「雨」を1、「雨以外」を0として扱う

```
contrasts(df_weather$weather)
```

```
#>      雨
```

```
#> 雨以外 0
```

```
#> 雨      1
```

ロジスティック回帰

```
summary(model)
#>
#> Call:
#> glm(formula = weather ~ temperature + humidity + pressure, family = binomial,
#>      data = df_weather)
#>
#> Deviance Residuals:
#>      Min       1Q   Median       3Q      Max
#> -1.9405  -0.7035  -0.3729   0.7241   2.5742
#>
#> Coefficients:
#>              Estimate Std. Error z value Pr(>|z|)
#> (Intercept) 140.58410    50.68850   2.773  0.00555 **
#> temperature  -0.52327     0.12171  -4.299 1.71e-05 ***
#> humidity       0.08834     0.02233   3.956 7.62e-05 ***
#> pressure      -0.13524     0.04877  -2.773  0.00555 **
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> (Dispersion parameter for binomial family taken to be 1)
#>
#>      Null deviance: 167.98  on 123  degrees of freedom
#> Residual deviance: 114.58  on 120  degrees of freedom
#> AIC: 122.58
#>
#> Number of Fisher Scoring iterations: 5
```



ロジスティック回帰による分類

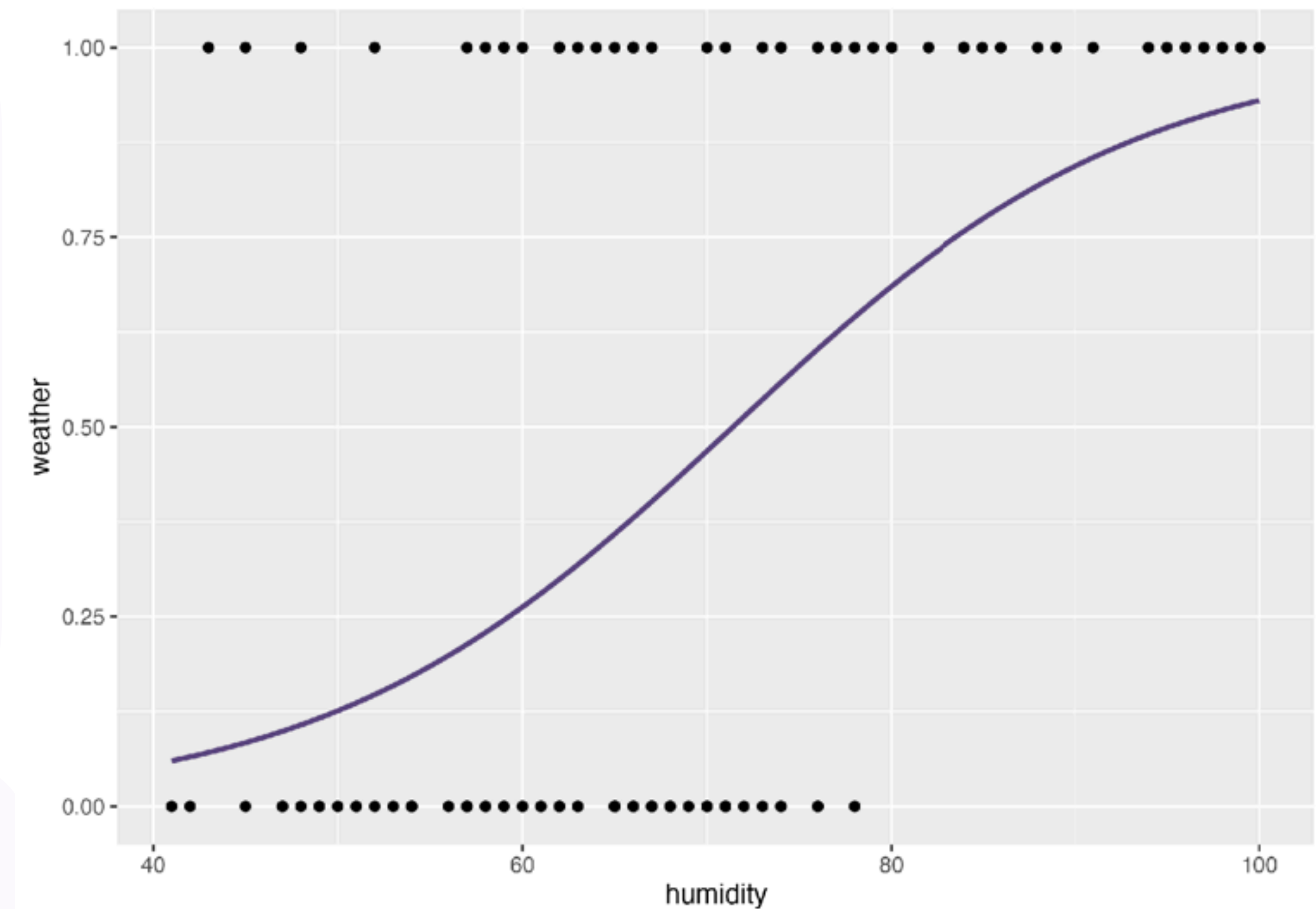
散布図を描画し、データの傾向を確認

```
library(ggplot2)
p <- ggplot(df_weather) +
  aes(humidity, as.numeric(weather)-1) +
  geom_point() +
  ylab("weather")
```

p



```
p + stat_smooth(
  method = "glm",
  method.args = list(family = "binomial"),
  se = FALSE,
  color = "#57467b")
```



ロジスティック回帰による分類

新しいデータをもとに「雨」の確率を求める

```
new_weather <- data.frame(  
  temperature = 14.1,  
  humidity = 88,  
  pressure = 1001)
```



```
predicted_prob <- predict(  
  model,  
  newdata = new_weather,  
  type = "response")  
predicted_prob  
#> 1  
#> 0.9963273
```



「雨」の確率は99%

```
new_weather <- data.frame(  
  temperature = 22.6,  
  humidity = 28,  
  pressure = 1023)
```



```
# 1- にすることで「雨以外」の確率を求める  
1 - predict(  
  model,  
  newdata = new_weather,  
  type = "response")  
#> 1  
#> 0.9991922
```



「雨以外」の確率は99%

参考資料・URL

目 久保拓弥『データ解析のための統計モデリング入門：一般化線形モデル・階層ベイズモデル・MCMC』（2012）岩波書店. ISBN: 978-4-00-006973-1
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

目 松井秀俊, 小泉和之（著）, 竹村彰通（編）『統計モデルと推測』（2019）講談社. ISBN: 978-4-06-517802-7
瓜生居室: あり（電子版）、徳大図書館: なし、市立図書館: なし、県立図書館: なし

目 西内啓『統計学が最強の学問である：データ社会を生き抜くための武器と教養』（2013）ダイヤモンド社. ISBN: 978-4-478-02221-4
瓜生居室: あり、徳大図書館: あり、市立図書館: あり、県立図書館: あり

目 有賀友紀, 大橋俊介（著）『RとPythonで学ぶ実践的データサイエンス&機械学習』（2021）技術評論社. ISBN: 978-4-297-12022-1
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

目 江崎貴裕『データ分析のための数理モデル入門: 本質をとらえた分析のために』（2020）ソシム. ISBN: 978-4-8026-1249-4
瓜生居室: あり（電子版）、徳大図書館: あり、市立図書館: あり、県立図書館: なし

目 馬場真哉『RとStanではじめるベイズ統計モデリングによるデータ分析入門』（2019）講談社. ISBN: 978-4-06-516536-2
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

