高知工科大学 経済・マネジメント学群

計量経済学

6. 回帰分析による統計的推測 II
仮説を検証する

やない ゆうき
矢内 勇生

https://yukiyanai.github.io
yanai.yuki@kochi-tech.ac.jp

このトピックの目標

- 回帰分析で「統計的検定」と「統計的推定」を行うための準備を整える
  - 母集団における回帰直線と標本の回帰直線を区別する
  - 回帰分析の帰無仮説と対立仮説を理解する
- 回帰分析で仮説を検証する方法を理解する
  - 回帰係数の統計的検定手続きを理解する
  - 「統計的に有意」の意味を理解する

母集団の回帰直線と
標本の回帰直線

回帰分析による推定

- データから作った散布図への直線（平面）の当てはめは、標本データの要約
- 興味があるのは母集団の特徴

★ どのような方法で、標本から母集団を推定する？

統計モデルをつくる

- 自分が観察しているデータが生み出される過程をモデル化する
  - データ生成過程（data generating process; DGP）
  - モデル：目的に応じた現象の単純化
    - 本質的に「正しくない」
    - 「正しいかどうか」ではなく、「役に立つかどうか」で評価する

"All models are wrong, but some are useful."

–George E. P. Box

Cf. Box, George. 1976. "Science and Statistics." *Journal of the American Statistical Association,* 71(356): 791-799.

単回帰

- 母集団における単回帰

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

  - α, β：パラメタ, 母数（推定の対象）
  - ε：誤差 (error)
    - 説明変数以外で応答変数に影響を与えるもの
    - 平均すると0

誤差をモデル化する

- 誤差 ε の分布を以下のように**仮定**する
  - $\varepsilon_i \sim \text{Normal}(0, \sigma)$
    - 誤差の平均は 0
    - 誤差は、1つの正規分布から生み出される
      - 誤差の標準偏差 σ は、i によらず一定

単回帰モデル

- 単回帰モデル：単回帰が想定するDGP
  - まず、X_i $(i = 1,2,\ldots)$の値が決まる
  - 次に、Y_i $(i = 1,2,\ldots)$の値が以下のように決まる

$$Y_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta X_i$$

単回帰モデルの書き換え

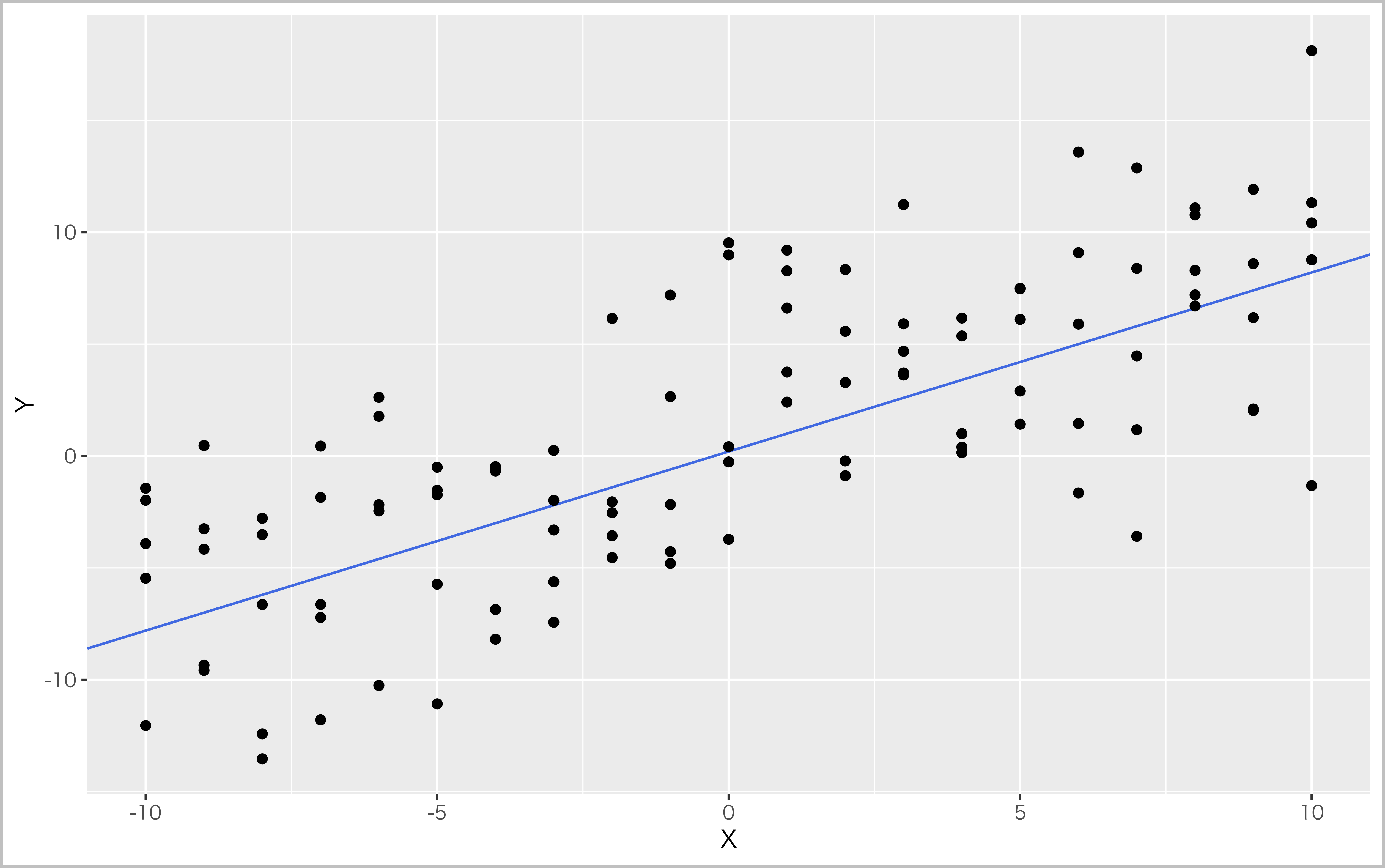
- 以下のような表記が使われることも多い（意味はどれも同じ）
    - 別表記（1）

$$Y_i \sim \text{Normal}(\alpha + \beta X_i,\ \sigma)$$

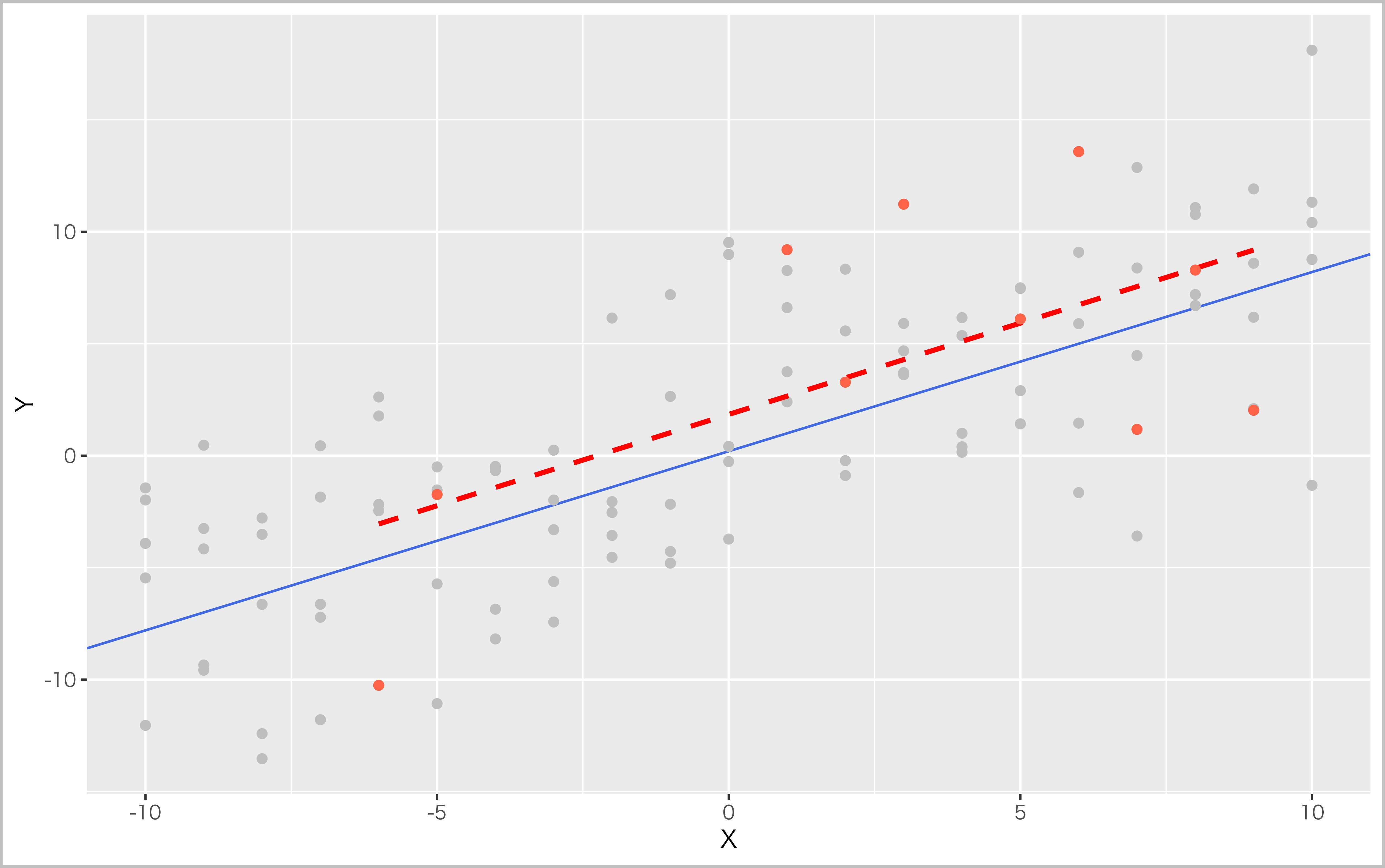
    - 別表記 (2)

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$
$$\varepsilon_i \sim \text{Normal}(0, \sigma)$$

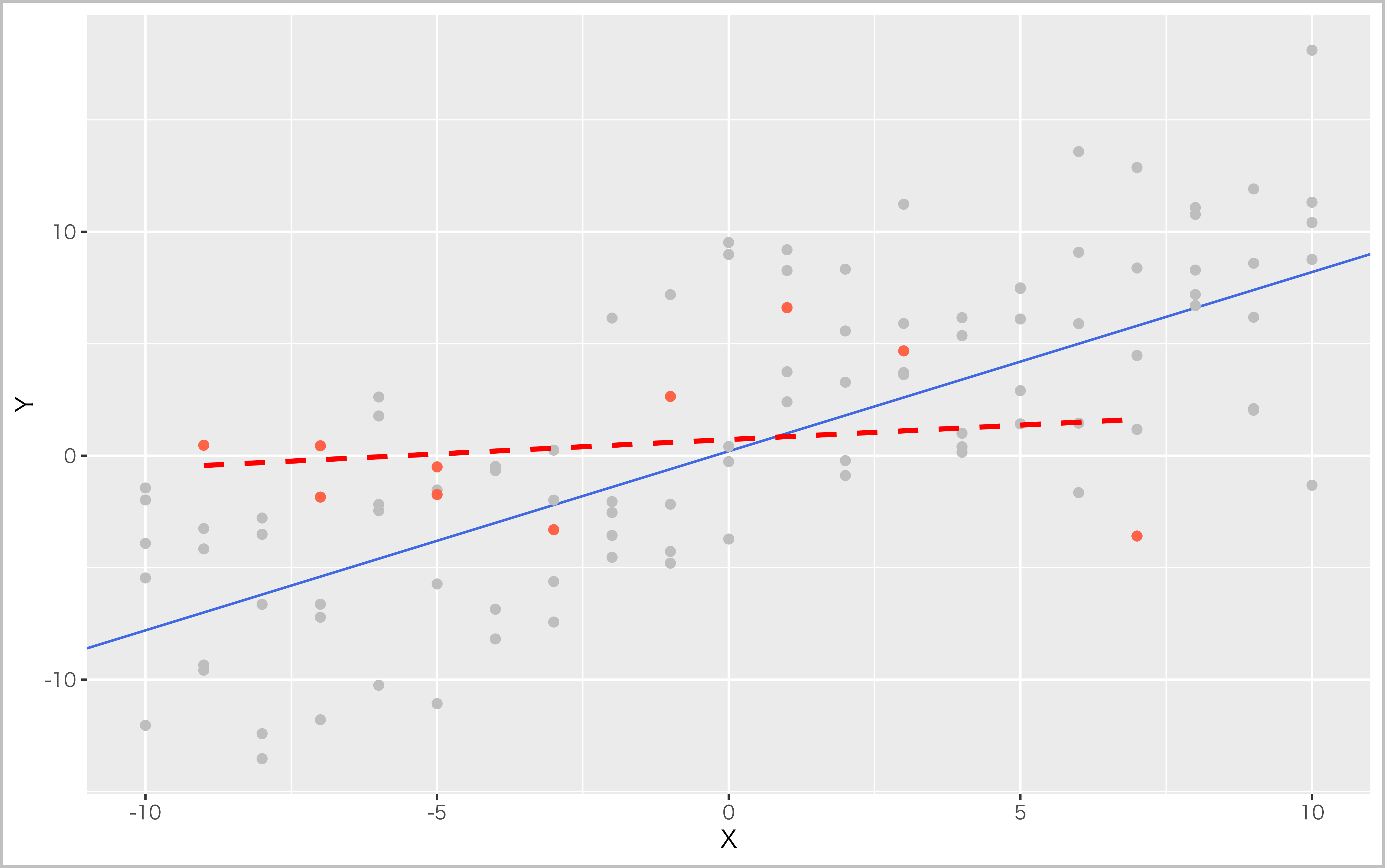
母集団の回帰直線



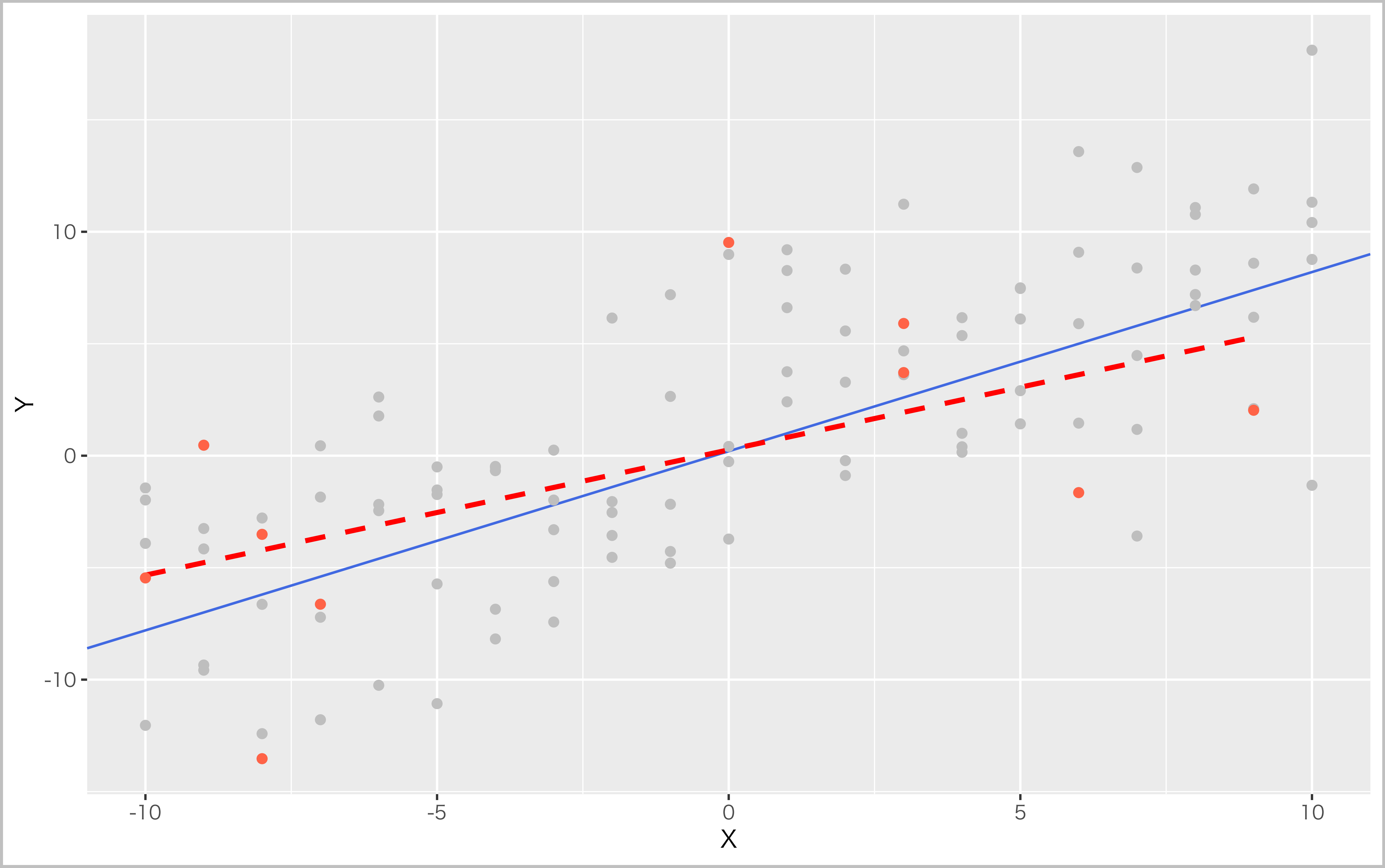
標本の回帰直線 (1)



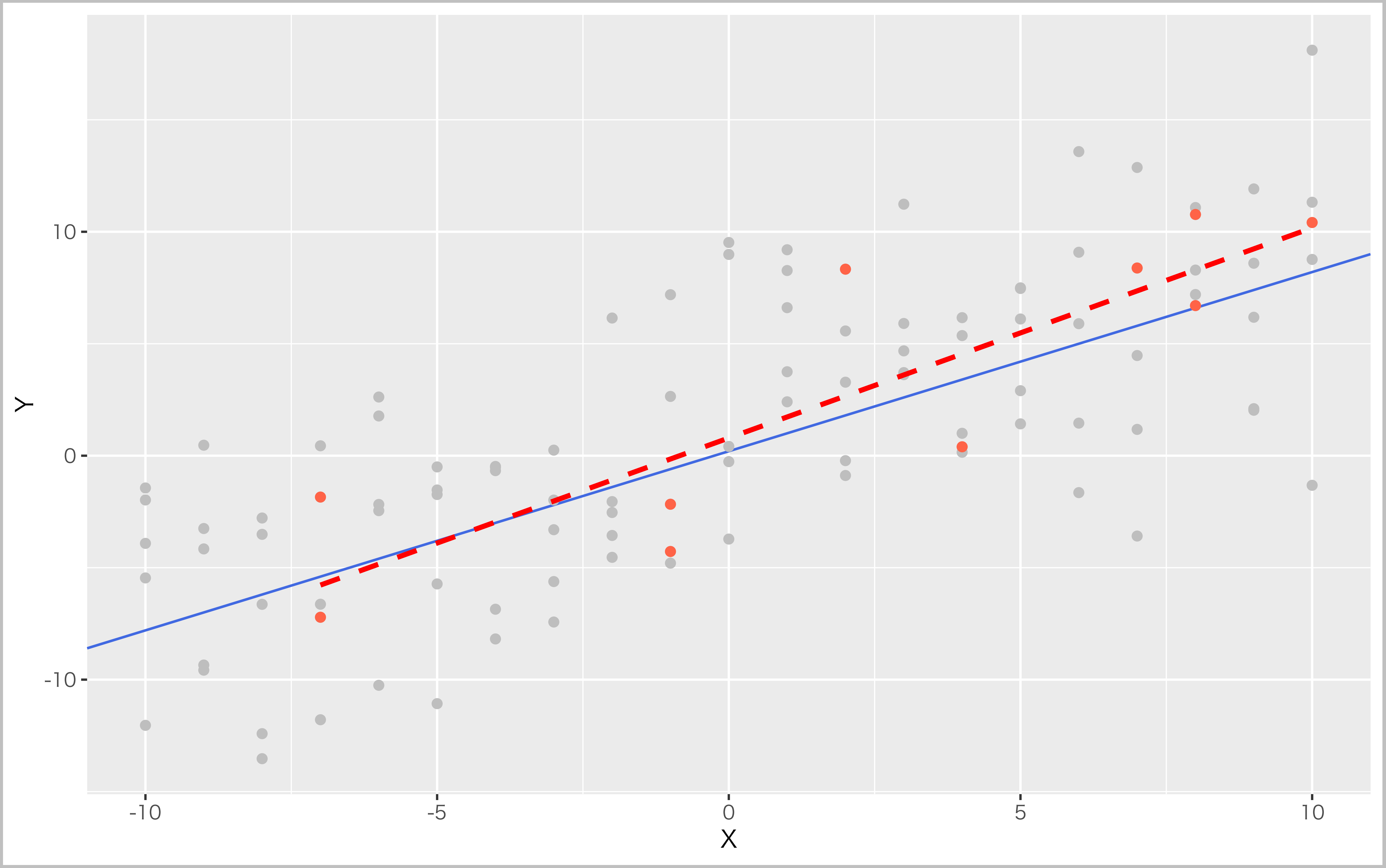
標本の回帰直線 (2)



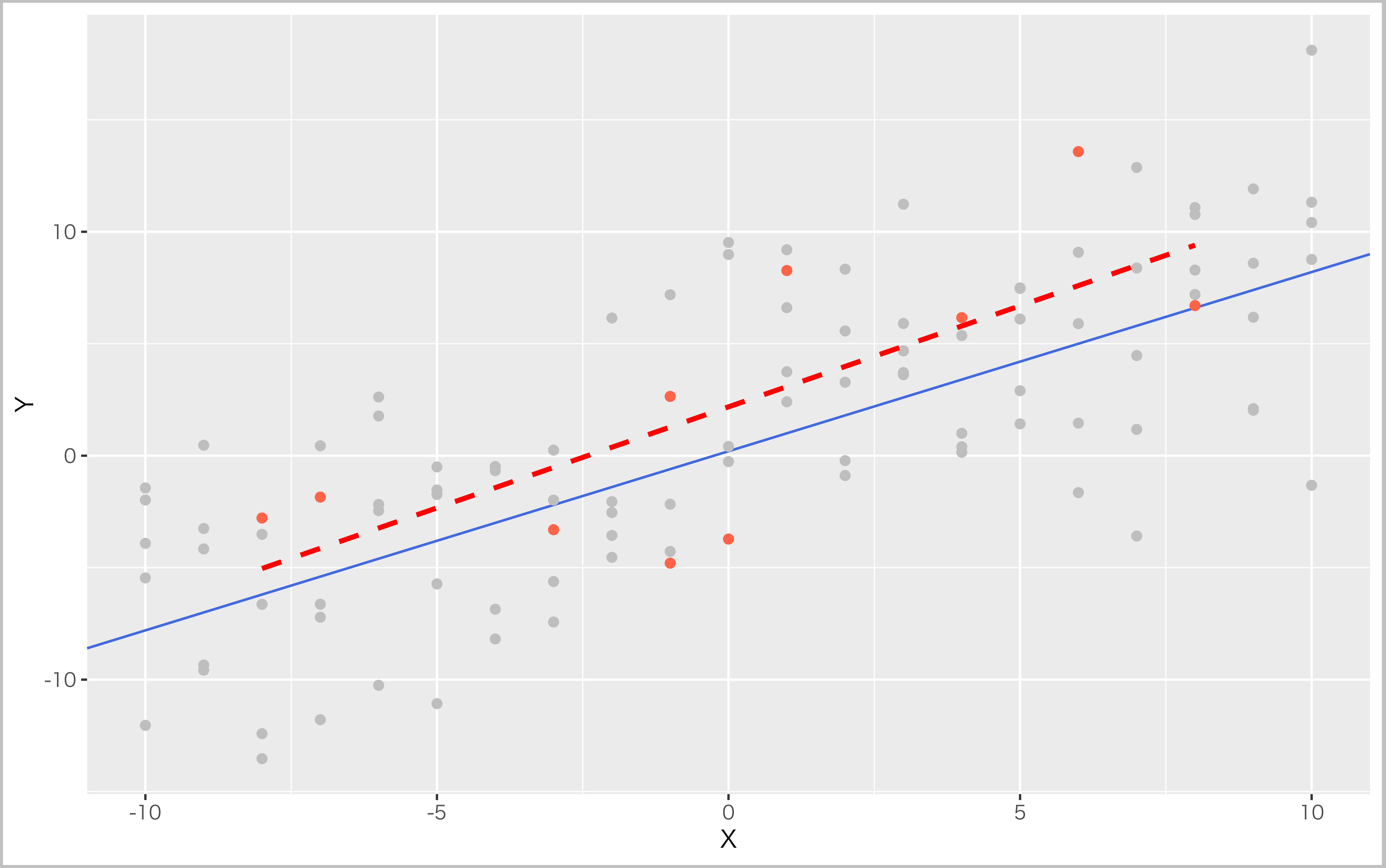
標本の回帰直線 (3)



標本の回帰直線 (4)



標本の回帰直線 (5)



最小二乗法による母数の推定：単回帰の場合

- 標本データを使い、最小二乗法よって求めた回帰係数 a, b は、単回帰モデルに登場する α, β の点推定値

- 最小二乗推定量は以下の望ましい性質をもつ
  - 不偏性 (unbiasedness)：$\mathbb{E}[a] = \alpha, \quad \mathbb{E}[b] = \beta$
  - 一致性 (consistency)：標本サイズを無限大にすると、推定値は母数に一致する

重回帰

- 母集団における重回帰

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_K X_{iK} + \varepsilon_i$$

- β_k ：パラメタ, 母数（推定の対象）、$k = 0, 1, 2, \ldots, K$
- ε ：誤差
  - $\varepsilon_i \sim \text{Normal}(0, \sigma)$

重回帰モデル

- 重回帰モデル：重回帰が想定するDGP
  - まず、X_{ik} $(i = 1,2,\ldots; k = 0,2,\ldots,K)$の値が決まる
  - 次に、Y_i $(i = 1,2,\ldots)$の値が以下のように決まる

$$Y_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_K X_{iK}$$

最小二乗法による母数の推定：重回帰の場合

- 標本データを使い、最小二乗法によって求めた回帰係数 $b_0, b_1, \ldots, b_K$ は、$\beta_0, \beta_1, \ldots, \beta_K$ の推定値である
  - 不偏性：$\mathbb{E}[b_k] = \beta_k$（ $k = 0,1,2,\ldots,K$ ）
  - 一致性

回帰分析の
帰無仮説と対立仮説

何のために回帰分析を行うのか

- 目的：理論（理論仮説）を検証したい
  - そのために作業仮説を用意する
  - 回帰分析で検証可能な作業仮説を用意する
    - 1つの応答変数
    - 1つ以上の説明変数
    - 説明変数が応答変数に与える影響についての仮説
      - 例：「X が Y を増加させる」

帰無仮説と対立仮説

- 帰無仮説：「説明変数は応答変数に影響を与えない」
- 対立仮説：「説明変数が応答変数に影響する」
  - 自分が「正しい」ことを示したい理論の作業仮説を対立仮説にする
- 統計的検定（方法は後で説明する）で帰無仮説が棄却されたとき、「作業仮説が統計的に正しい」と判断する
  - 作業仮説が正しいと考えられるので、操作化がうまくできていれば、理論仮説の蓋然性が高まる
    - 操作化（作業仮説と理論仮説の類似度）が重要

単回帰の場合

- モデル：$Y_i \sim \text{Normal}(\alpha + \beta X_i,\ \sigma)$
- 検証する仮説
  - 帰無仮説：$\beta = 0$
  - 対立仮説：$\beta \neq 0$

重回帰の場合 (1) 包括的検定

- モデル：$Y_i \sim \text{Normal}(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_K X_{iK}, \sigma)$
- 検証する仮説のパタン1
  - 帰無仮説：$\beta_1 = \beta_2 = \cdots = \beta_K = 0$
  - 対立仮説：「$\beta_1, \beta_2, \ldots, \beta_K$ のうち、少なくとも1つについて $\beta_k \neq 0$」

重回帰の場合 (2) 個別的検定

- モデル：$Y_i \sim \text{Normal}(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_K X_{iK}, \sigma)$
- 検証する仮説のパタン2

	β_1の仮説	β_2の仮説		β_Kの仮説
帰無仮説：	$\beta_1 = 0$	$\beta_2 = 0$	…	$\beta_K = 0$
対立仮説：	$\beta_1 \neq 0$	$\beta_2 \neq 0$		$\beta_K \neq 0$

- 実際は、すべての k について仮説を立てて検証するわけではなく、理論における「原因」とみなされるものについてのみ個別に仮説を検証する

「影響がない」を検証する？？？

- 通常、「影響がない」は帰無仮説
  - 「影響がない」を対立仮説にすると、帰無仮説「影響がある」は棄却できない（検証する対象が無限にある）
  - 「影響がない」という帰無仮説を棄却できなくても、それは「影響がない」ことを意味しない
    - 「影響がある」という証拠が見つからないだけ
    - 「証拠の不在」は「不在の証拠」ではない！

★「影響がない」ことを主張する理論は、（これまで勉強してきた）統計的分析では検証不可能

回帰分析による
統計的検定と推測

回帰分析における仮説検定

- 回帰分析では、説明変数が応答変数に影響を与えているかどうかに関心がある
  - 帰無仮説：説明変数の影響はない（影響が0である）
  - 対立仮説：説明変数の影響がある（影響が0ではない）

単回帰の例

- 単回帰モデル：$Y_i \sim \mathrm{Normal}\left(\alpha + \beta X_i, \sigma\right)$
  - 帰無仮説：$\beta = \tilde{\beta}$
  - 対立仮説：$\beta \neq \tilde{\beta}$
- 標本 (y, x) から求めた回帰直線：$\hat{y}_i = a + bx_i$

推定値のばらつき

- b: β の点推定量
  - b の値は標本によってばらつく
  - 標本ごとに異なる b の標準偏差：標準誤差 (SE)

$$\mathrm{SE}(b) = \sqrt{\frac{\hat{V}_1}{N}}$$

$$\hat{V}_1 = \frac{\frac{1}{N}\sum_{i=1}^{N}\left[(x_i - \bar{x})^2 e_i^2\right]}{\left[\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^2}$$

ただし、e_i は残差：$e_i = y_i - \hat{y}_i = y_i - (a + bx_i)$

- 詳しくは、西山ほか (2019)『計量経済学』（有斐閣）：第4章を参照

推定量 b の分布

$$\frac{b - \tilde{\beta}}{\mathrm{SE}(b)} \sim t\,(N - K - 1)$$

- $\tilde{\beta}$：帰無仮説が想定する β
  - 帰無仮説が正しいなら、$\mathbb{E}[b] = \tilde{\beta}$
- $t\,(N - K - 1)$：自由度 $N - K - 1$ の t 分布
  - N：標本サイズ
  - K：説明変数の数（切片は含まない）

t 統計量を用いた仮説検定

t 統計量：$T = \dfrac{b - \tilde{\beta}}{\mathrm{SE}(b)}$

- 特定の有意水準のもとで、自由度 $N - K - 1$ のt 分布の臨界値 c を求め、

$$|T| > |c|$$

となるとき、帰無仮説を棄却する

t 統計量を用いた仮説検定 (続)

- 帰無仮説が $\beta = 0$（つまり、$\tilde{\beta} = 0$）のとき、

$$T = \frac{b - \tilde{\beta}}{\mathrm{SE}(b)} = \frac{b}{\mathrm{SE}(b)}$$

- この T の値は、Rで回帰分析結果に t value または statistic として表示される
- 有意水準が5パーセントのとき、検定の臨界値は約2
  - よって、係数を標準誤差で割った値の絶対値が2より大きければ、有意水準5%で帰無仮説を棄却する

Rで回帰分析

- `lm()` 関数を使う
  - 例、myd という名前のデータセット（データフレーム, tibble）に含まれる変数を使い、y を x1 とx2 に回帰する

```
fit <- lm(y ~ x1 + x2,
          data = myd)
```

summary() による結果の表示

- `lm()` で推定した後、`summary()` で結果を確認する
- 例：`summary(fit)`
  - Estimate: パラメタの点推定値
  - Std. Error：標準誤差（推定の不確実性）
  - t value: t 検定で使う検定統計量
  - Pr(>|t|)：p 値

summary() による結果の表示（続）

```
> summary(fit1)

Call:
lm(formula = voteshare ~ experience, data = HR1996)

Residuals:
    Min      1Q  Median      3Q     Max 
-38.334 -10.007  -2.207   8.593  67.393 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  16.0070     0.4608   34.74   <2e-16
experience   22.8274     0.7891   28.93   <2e-16

Residual standard error: 13.28 on 1259 degrees of freedom
Multiple R-squared:  0.3993,	Adjusted R-squared:  0.3988 
F-statistic: 836.8 on 1 and 1259 DF,  p-value: < 2.2e-16
```

broom::tidy() で結果を確認する

- broom パッケージの `tidy()` 関数でも結果を確認できる
- 以下のようにすると、95パーセント信頼区間も表示できる（95パーセント以外にするには、conf.level を変える）

```
tidy(fit,
     conf.int = TRUE,
     conf.level = 0.95)
```

broom::tidy() で結果を確認する（続）

```
> tidy(fit1, conf.int = TRUE)
# A tibble: 2 x 7
  term        estimate std.error statistic   p.value conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>
1 (Intercept)     16.0     0.461      34.7 5.66e-186     15.1      16.9
2 experience      22.8     0.789      28.9 1.68e-141     21.3      24.4
```

Rで信頼区間を求める

- `lm()` を実行した後、`confint()` 関数を使うと、係数の信頼区間を求めることができる。
  - 例
    - 95%信頼区間：`confint(fit)`
    - 50%信頼区間：`confint(fit, level = 0.5)`
    - 68%信頼区間：`confint(fit, level = 0.68)`
  - 上のコマンドを実行すると、信頼区間の下限値と上限値が表示される

信頼区間の図示

- ggplot2 を使えば、以下のものが図示できる
  - 回帰直線 + 95%信頼区間

```
geom_smooth(method = "lm")
```

  - 回帰直線 + 89%信頼区間

```
geom_smooth(method = "lm", level = 0.89)
```

  - 回帰直線のみ

```
geom_smooth(method = "lm", se = FALSE)
```

信頼区間

- 回帰分析による点推定値は、1つの標本（データ）から得られたもの

➡ 母数に一致するとは限らない（実際の標本サイズは有限なので）

- 統計量はばらつく（シミュレーションで確認する！）
  - 標準誤差：統計量のばらつき

➡ 信頼区間を求める!

信頼区間の意味（1）

- 95%信頼区間とは何か？
  - よくある**誤解**：「得られた信頼区間に、真の値が入っている確率が95%」
  - 「真の値」があるなら、「得られた信頼区間に、真の値が入っている確率」は、

    - 100%（実際に入っている）

      または

    - 0%（入っていない）

    しかあり得ない

信頼区間の意味（2）

- では、95%信頼区間とは何なのか？

1. データを生成する（新たに観測する）
2. データを分析する
3. 95%信頼区間を求める

- 95%信頼区間：上の1〜3までを何度も何度も繰り返し行うと、そのうち95%は「真の値を含む信頼区間」が得られるだろう

信頼区間の信頼度（1）

- 信頼区間の長さ
  - 信頼度が高いほど区間が長くなる
  - 信頼度が低いほど区間が短くなる
- なぜ？
  - 区間を長くすれば、取りこぼしの確率が小さくなる
  - 区間を短くすれば、取りこぼしの確率は大きくなる

信頼区間の信頼度（2）

- では、信頼区間は長い方がいいのか？
  - **No!**
  - 同じ信頼度で、信頼区間が短いほうが推定の不確実性が小さい
  - 信頼区間の長さ：標準誤差に依存
    - 標準誤差が大きい：信頼区間が長い
    - 標準誤差が小さい：信頼区間が短い

統計的に有意とは？

統計的に有意とは？(1)

- 「統計的に有意」な結果を見せられたとき、私たちはどのように反応すべきか？
  - 「**だから何？**」「統計的に有意だと**何が嬉しいの**？」
- 統計的に有意：効果が0ではない
  - 「ゼロでない効果」には色々ある
    - 計量経済学に関する自習時間を1日10時間増やすと、期末試験の点数が5点上がる
    - 計量経済学に関する自習時間を1日に10分増やすと、期末試験の点数が25点上がる

統計的に有意とは？(2)

- 効果が「ゼロではない」と信じるに足る証拠がある
  - それだけ！
- 「ゼロではない」≠ 重要
- 研究においては、「重要である」ことを示すことが求めらる
  - 実質的重要性 (substantive significance) を示すことが必要（**浅野・矢内 2018: pp. 165-168** を参照）
- **係数の値そのもの（効果量, effect size）を議論することが絶対に必要！！！**

やってはいけない (1)

- **「統計的に有意であること」を論文（あるいは統計分析の）の結論のように書いてはいけない！**
  - 統計的に有意であることは、分析結果の一部に過ぎない
  - そこから「論文で扱っている特定の研究対象について」何が言えるのか掘り下げ、リサーチクエスチョンに答える必要がある
- **結論は、リサーチクエスチョン (RQ) に対する答え**

ダメな例

- RQ:「計量経済学」の成績を上げるにはどうしたらいいか？
- 理論：「Rを使いこなすと、成績が上がる」
- 作業仮説：「Rを1時間以上利用する日数が増えると、成績（100点満点）が上昇する」
- 回帰分析で検証：統計的に有意
- 結論：「Rの使用日数が成績に与える効果は、統計的に有意だ」

★　読者：？？？？？？？？？？？？？？？？？？？？？？

ダメな例を改善する：パタン1

- RQ:「計量経済学」の成績を上げるにはどうしたらいいか？
- 理論：「Rを使いこなすと、成績が上がる」
- 作業仮説：「Rを1時間以上利用する日数が増えると、成績（100点満点）が上昇する」
- 回帰分析で検証：統計的に有意
  - 使用日数が1日増えるごとに、点数が1点上がる
  - 1Qは60日ある：最大で60点成績アップが可能
  - 分析の結論：「Rの使用日数は成績を上げる」
- 結論：「計量経済学」の成績を上げるためには、1時間以上Rを使う日をできるだけ増やせばよい

★　読者：！！！

ダメな例を改善する：パタン2

- RQ:「計量経済学」の成績を上げるにはどうしたらいいか？
- 理論：「Rを使いこなすと、成績が上がる」
- 作業仮説：「Rを1時間以上利用する日数が増えると、成績（100点満点）が上昇する」
- 回帰分析で検証：**統計的に有意**
  - 使用日数が1日増えるごとに、点数が0.05点上がる
  - 1Qは60日ある：最大で3点成績アップが可能
  - 分析の結論：「Rの使用日数を増やしても成績は**あまり変わらない**」
- 結論：Rを1時間以上使う日数を増やしただけでは「計量経済学」の成績をよくするのは難しいので、他の方法を考える必要がある

矛盾しない！

★ 読者：…

効果がないことを証明できる？

- 効果がないことを証明したいとき、$\beta = 0$ という帰無仮説が保留（受容）されることは証拠として使える？

➡ 使えない！

- （通常の）統計的仮説検定の方法では、効果がない証拠を見つけることは不可能（以下のいずれかの方法が必要）
  - 同等性の検定 (test of equivalence) を実施する
  - ROPE [region of practical equivalence] というものを設定し、ベイズ統計分析を実行する

やってはいけない (2)

- **「影響がない」ことを（これまで習った）統計分析の結論として述べてはいけない**
  - 通常の統計的検定の枠組みでは、「影響がない」ことは示せない
    - 「神がいる」という証拠がないことは、「神がいない」ことの証明にはならない
- **結論は、以下の3つのうちのどれか：**
  - **「意味のある影響がある（統計的に有意で実質的にも有意）」**
  - **「影響はある（統計的に有意）が実質的には無意味」**
  - **「影響があるという証拠がない（統計的に有意ではない）」**

次のトピック

回帰分析の応用