回帰分析ではlm()ではなくestimatr::lm_robust()を使おう

2022/07/23
第100回R勉強会@東京（#TokyoR）
森下光之助（@dropout009）

自己紹介

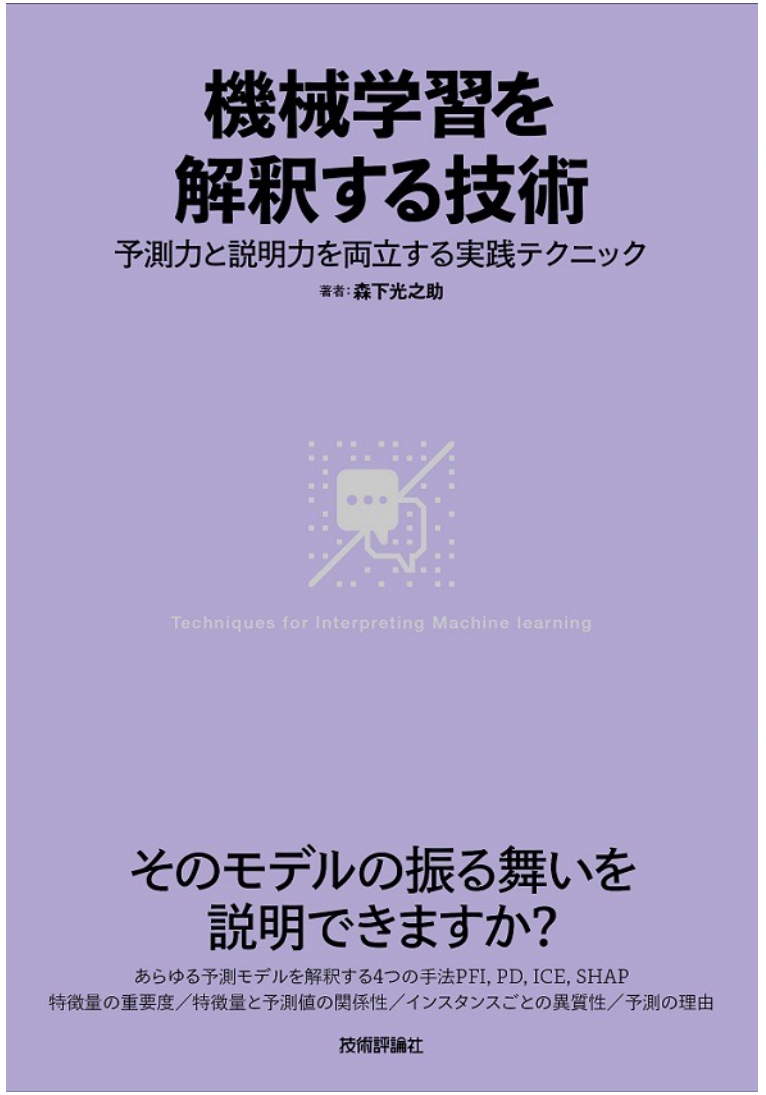
森下光之助

TVISION INSIGHTS株式会社
データサイエンティスト
執行役員（データ・テクノロジー担当）

テレビの視聴行動を分析しています

データの利活用、マネジメント、組織づくり、
因果推論、機械学習の解釈手法などに興味があ
ります

Twitter: @dropout009
Speaker Deck: dropout009
Blog: https://dropout009.hatenablog.com/



Rで回帰分析をするときは

標準の`lm()`だと

不均一分散に対応できないので

`estimatr::lm_robust()`を使おう

という話をします

lm()と回帰係数の標準誤差

シミュレーションデータをもとに、lm()の挙動を確認していく

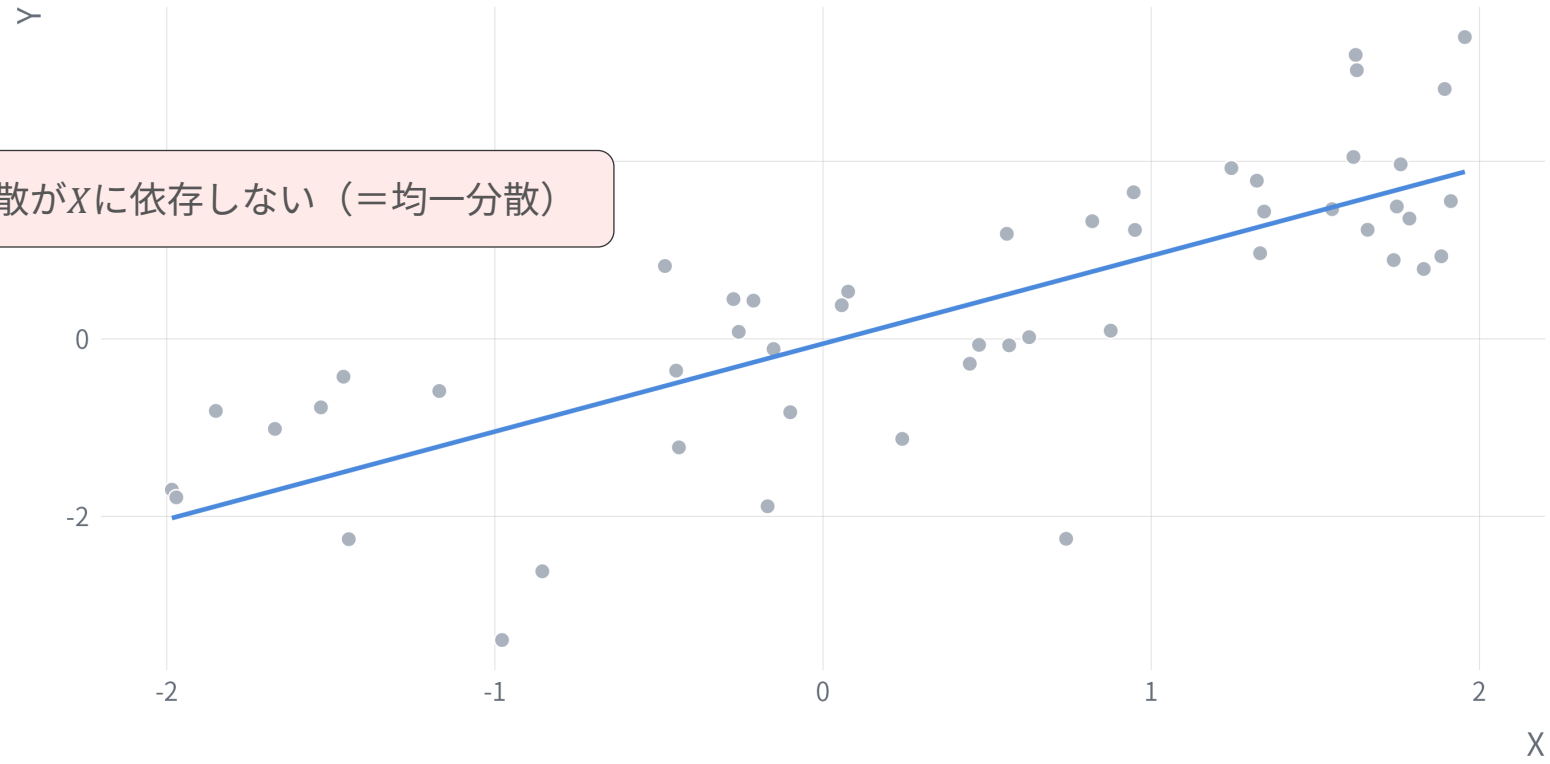
以下のようなシミュレーションデータに対して、線形回帰モデルをあてはめる

$$Y = X + U$$
$$X \sim \mathrm{Uniform}(-2, 2)$$
$$U \sim \mathcal{N}(0, 1)$$

誤差項Uの分散がXに依存しない（＝均一分散）

```
N <- 50
df <- tibble(
  x = runif(N, -2, 2),
  u = rnorm(N, 0, 1),
  y = x + e
)
```

シミュレーションデータと回帰直線



シミュレーションデータをもとに、lm()の挙動を確認していく

シミュレーションデータに単回帰モデル

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

をあてはめる。lm()の出力結果から回帰係数の大きさや標準誤差がわかる

信頼区間やp値はこの標準誤差を使って計算される

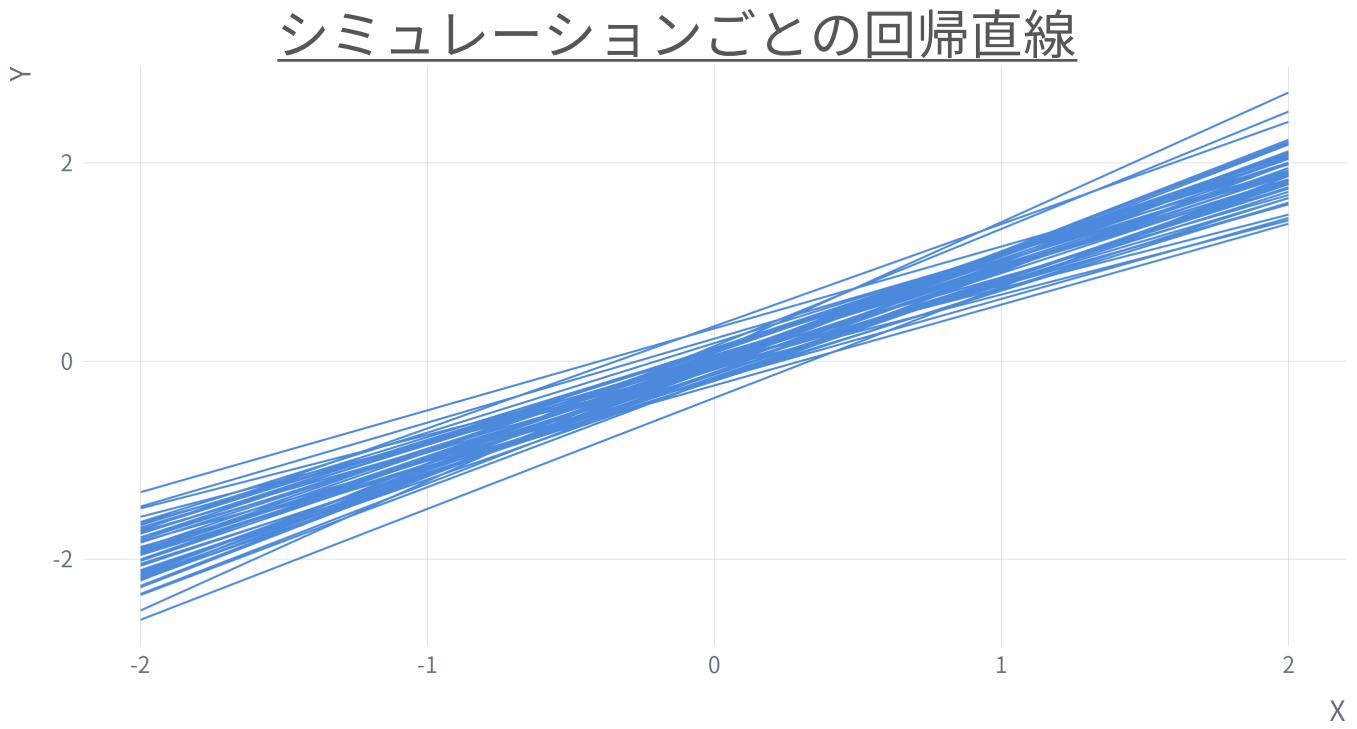
```
> df %>%
+   lm(y ~ x, data = .) %>%
+   tidy()

# A tibble: 2 × 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)  -0.0542     0.146    -0.371 7.12e- 1
2 x             0.990      0.115     8.58  2.93e-11
```

回帰係数$\hat{\beta}$の標準誤差

回帰係数の標準誤差ってなに？

- 観測されたデータはあくまで母集団から抽出されたサンプルで、回帰分析ではそこから母集団のパラメータを推測している
- なので、同じシミュレーション設定で何度もデータを生成し回帰直線を引くと、生成されるデータのばらつきに応じて少しずつ違う回帰直線が引かれる
- この回帰直線のばらつき（標準偏差）を推定しているのが回帰係数の標準誤差



実際に何度もデータを生成して回帰分析の結果を確認していく

```
N <- 50
M <- 5000

df <- tibble(
  m = rep(seq_len(M), each = N),
  x = runif(N * M, -2, 2),
  u = rnorm(N * M, 0, 1),
  y = x + u
)

df_result <- df %>%
  nest(data = c(x, u, y)) %>%
  mutate(result = map(data, ~ tidy(lm(y ~ x, data = .)))) %>%
  select(m, result) %>%
  unnest(c(m, result)) %>%
  filter(term == "x")
```

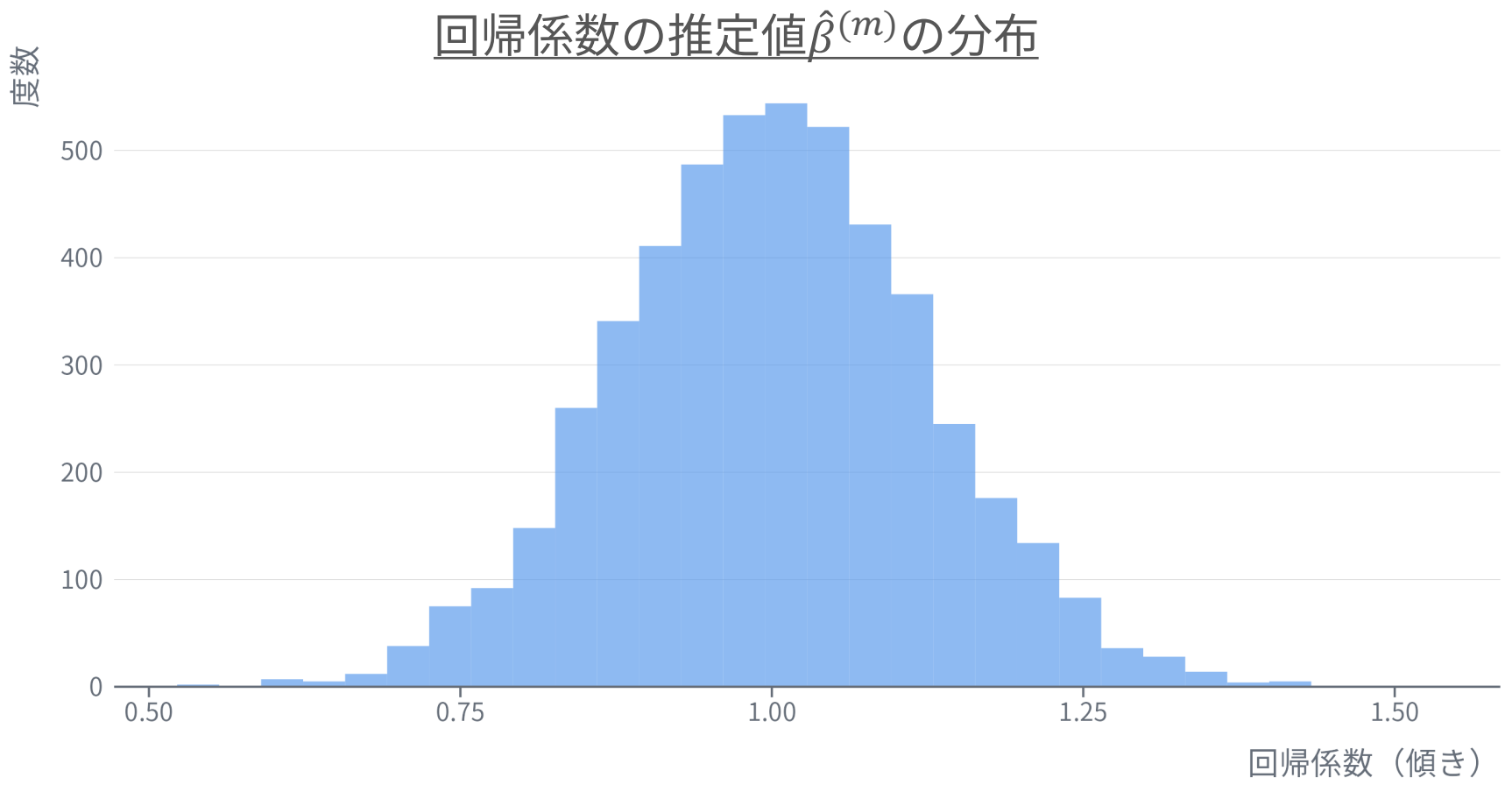
$Y = X + U$
$X \sim \mathrm{Uniform}(-2, 2)$
$U \sim \mathcal{N}(0, 1)$

$\left\{\left(Y_i^{(m)}, X_i^{(m)}\right)\right\}_{i=1}^{N}$ から単回帰で
$\left(\hat{\alpha}^{(m)}, \hat{\beta}^{(m)}\right)$を推定

傾き$\hat{\beta}^{(m)}$に注目

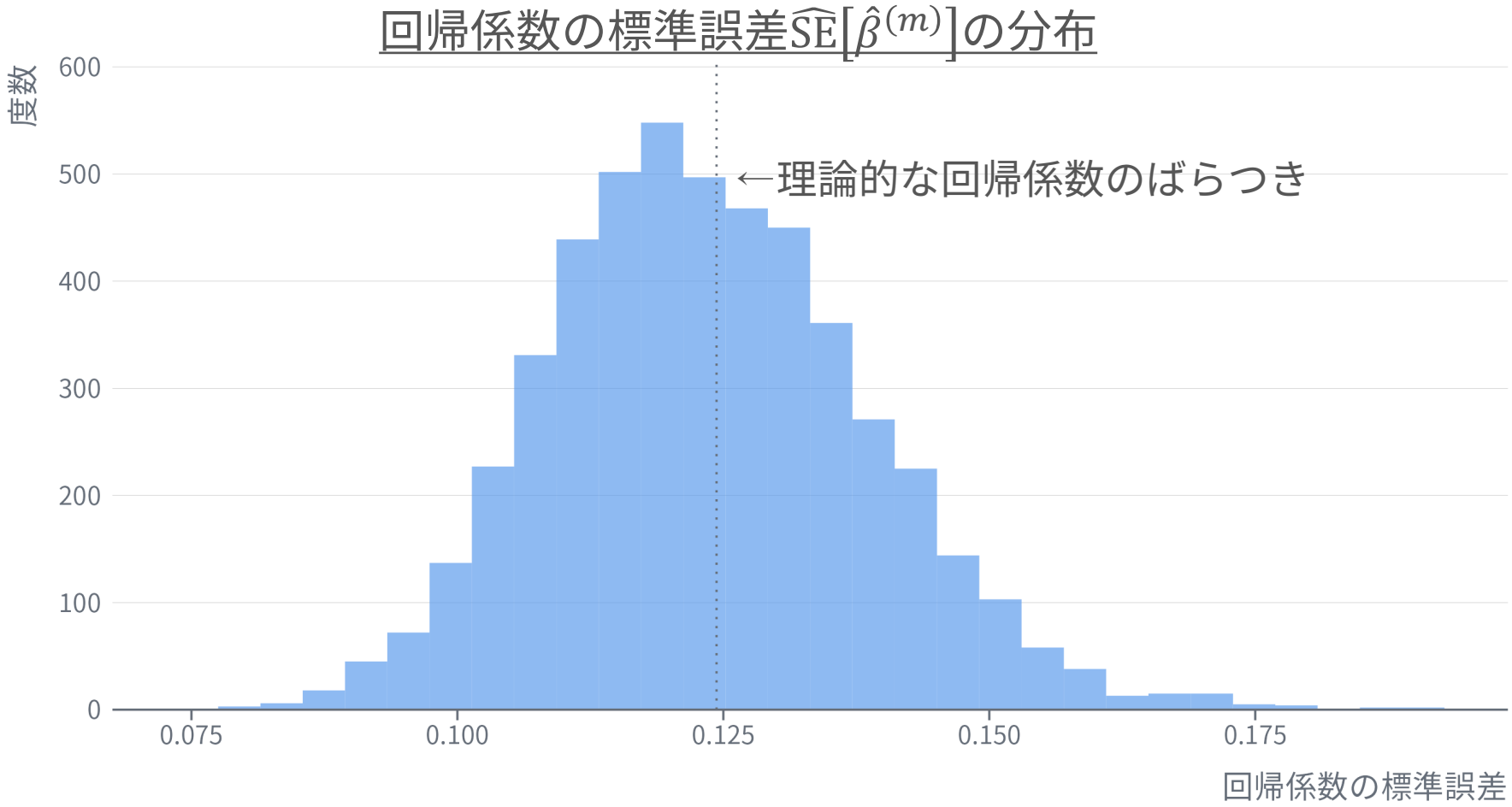
シミュレーションごとに回帰係数の推定値$\hat{\beta}^{(m)}$の分布を確認

- 回帰係数はシミュレーション毎に少し異なる値になるので、結果の分布を確認
- この分布の標準偏差$\mathrm{SD}\left[\hat{\beta}^{(1)}, \ldots, \hat{\beta}^{(M)}\right]$が回帰係数の理論的なばらつき（のシミュレーションによる近似）
- 回帰係数の標準誤差$\widehat{\mathrm{SE}}\left[\hat{\beta}^{(m)}\right]$は回帰係数は理論的なばらつきを推定している



lm()の標準誤差は回帰係数のばらつきをうまく推定できている

- 回帰係数の標準誤差はシミュレーション毎に少し異なる値になるので、結果の分布を確認
- 標準誤差は理論的な回帰係数のばらつきを中心に分布しており、ばらつきを概ねうまく推定できていることがわかる（ほんのちょっと小さめに推定してはいる）



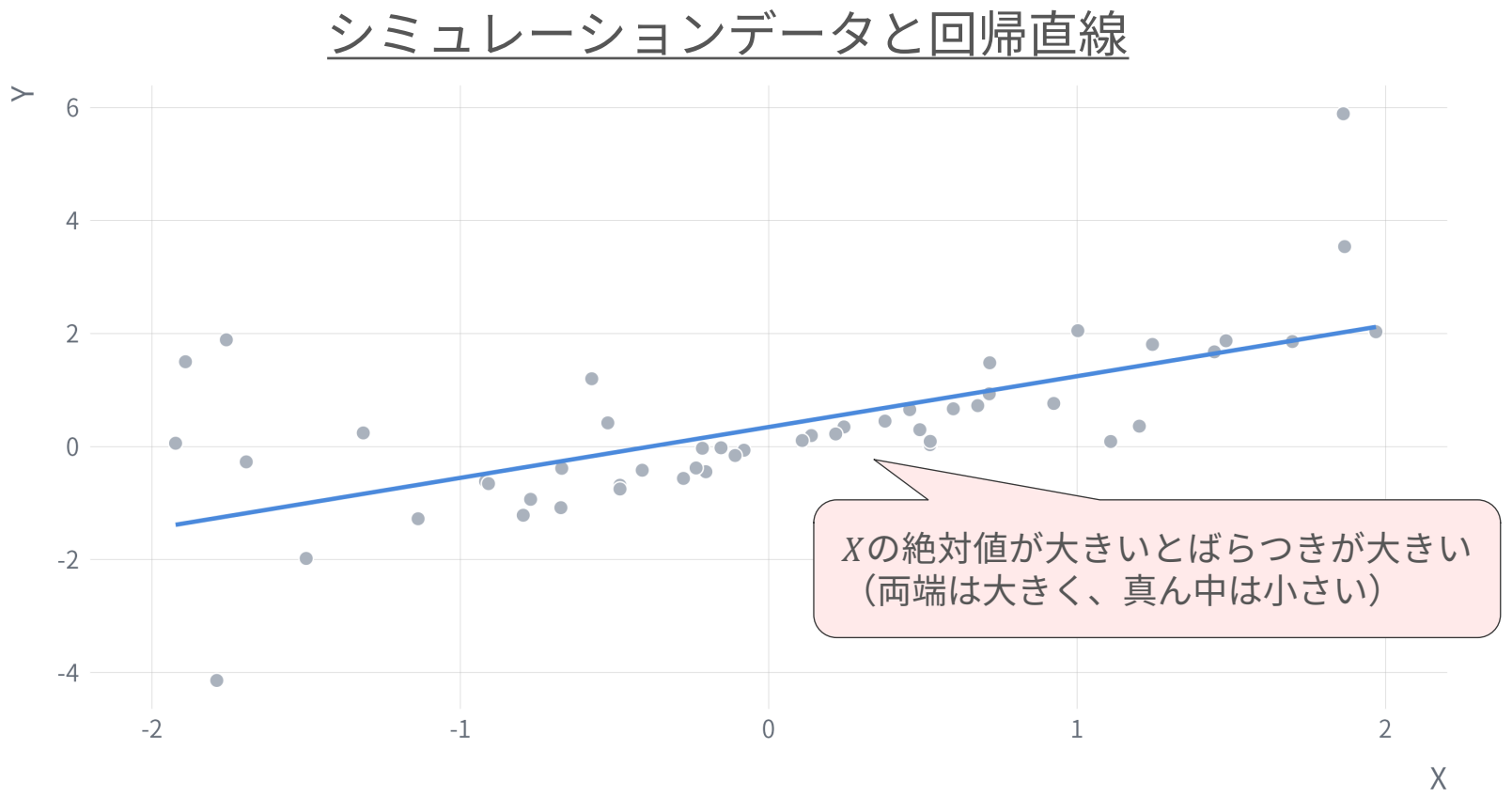
不均一分散と lm()の標準誤差

誤差項Uの分散がXに依存している（＝分散が一定でない）場合を考える

誤差項Uの分散がXに依存（＝不均一分散）

$$Y = X + U$$
$$X \sim \text{Uniform}(-2, 2)$$
$$U \sim \mathcal{N}(0, X^2)$$

```
N <- 50
df_heterogeneous <- tibble(
  x = runif(N, -2, 2),
  u = rnorm(N, 0, abs(x)),
  y = x + u
)
```



実際に何度もデータを生成して回帰分析の結果を確認

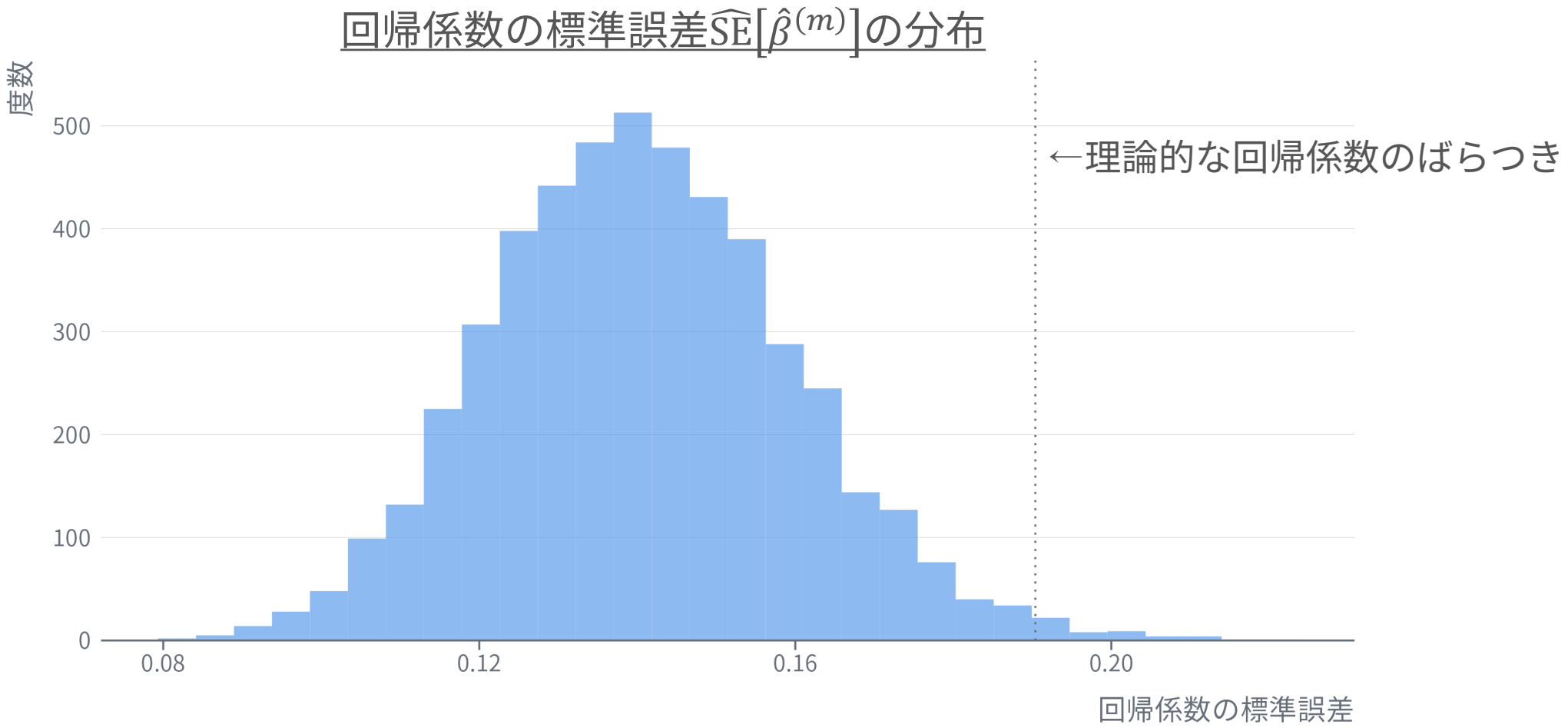
```
N <- 50
M <- 5000

df_simulation_heterogeneous <- tibble(
  m = rep(seq_len(M), each = N),
  x = runif(N * M, -2, 2),
  u = rnorm(N * M, 0, abs(x)),
  y = x + u
)

df_result_heterogeneous <- df_simulation_heterogeneous %>%
  nest(data = c(x, u, y)) %>%
  mutate(result = map(data, ~ tidy(lm(y ~ x, data = .)))) %>%
  select(m, result) %>%
  unnest(c(m, result)) %>%
  filter(term == "x")
```

lm()の標準誤差は回帰係数のばらつきを過小評価してしまっている

- 回帰係数の標準誤差はシミュレーション毎に少し異なる値になるので、結果の分布を確認
- 標準誤差は理論的な回帰係数のばらつきを過小評価してしまっている



lm()で計算される標準誤差は

均一分散を仮定しているから

lm()の標準誤差は
どうやって計算されているのか

そもそもlm()の標準誤差はどうやって計算されているのか？

回帰係数の標準誤差を考えるため、線形回帰モデル

$$Y = \boldsymbol{X}'\boldsymbol{\beta} + U$$

を考える

ここで、 $\boldsymbol{X}$と$\boldsymbol{\beta}$は共に$K \times 1$の行列

$$\boldsymbol{X} = (X_1, \dots, X_k, \dots, X_K)'$$
$$\boldsymbol{\beta} = (\beta_1, \dots, \beta_k, \dots, \beta_K)'$$

とする

典型的な仮定として$\mathbb{E}[U \mid \boldsymbol{X}] = 0$を置く。このとき、誤差項の条件付き分散$\sigma^2(\boldsymbol{X})$は

$$\sigma^2(\boldsymbol{X}) = \mathrm{Var}[U \mid \boldsymbol{X}] = \mathbb{E}[U^2 \mid \boldsymbol{X}]$$

となる

加えて以下も仮定しておく

$$\mathbb{E}[Y^2] < \infty, \mathbb{E}[\|\boldsymbol{X}\|] < \infty, \mathbb{E}[\boldsymbol{X}\boldsymbol{X}'] > 0$$

最小二乗法による回帰係数の推定

ランダムサンプリングされたデータ$\{(Y_1, \boldsymbol{X}_1), \ldots, (Y_i, \boldsymbol{X}_i), \ldots, (Y_N, \boldsymbol{X}_N)\}$から最小二乗法で回帰係数を推定する

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{b}}{\operatorname{argmin}} \sum_{i=1}^{N} (Y_i - \boldsymbol{X}_i' \boldsymbol{b})^2 = \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i' \right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{X}_i Y_i \right)$$

この最小二乗推定量$\widehat{\boldsymbol{\beta}}$の信頼性（標準誤差）を知りたい
後々便利なので、 $\widehat{\boldsymbol{\beta}}$を以下のように変形できることを確認しておく

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i' \right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{X}_i Y_i \right) \\
&= \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i' \right)^{-1} \left(\sum_{i=1}^{N} X_i (\boldsymbol{X}_i' \boldsymbol{\beta} + U_i) \right) \\
&= \boldsymbol{\beta} + \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i' \right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{X}_i U_i \right)
\end{aligned}$$

（参考）回帰係数の条件付き期待値

最小二乗法推定量の信頼性（標準誤差）を見る前に、まずは条件付き期待値を確認しておく

条件付き期待値は

$$\mathbb{E}\left[\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X_1}, \ldots, \boldsymbol{X_N}\right] = \mathbb{E}\left[\beta + \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{X}_i U_i\right) \mid \boldsymbol{X_1}, \ldots, \boldsymbol{X_N}\right]$$

$$= \boldsymbol{\beta} + \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{X}_i \mathbb{E}[U_i \mid \boldsymbol{X}_i]\right)$$

$$= \boldsymbol{\beta}$$

となり、最小二乗法推定量は（条件付き）不偏性を満たすことがわかる

回帰係数の条件付き分散

回帰係数の標準誤差は、回帰係数の条件付き分散の平方根をとったものなので、とりあえず条件付き分散を考える

$$\mathrm{Var}\left[\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\right] = \mathrm{Var}\left[\boldsymbol{\beta} + \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{X}_i U_i\right) \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\right]$$

$$= \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i' \mathrm{Var}[U_i \mid \boldsymbol{X}_i]\right) \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1}$$

$$= \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i' \mathbb{E}\left[U_i^2 \mid \boldsymbol{X}_i\right]\right) \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1}$$

$$= \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i' \sigma_i^2\right) \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1}$$

$i \neq j$の場合は

$$\mathrm{Cov}\left[U_i, U_j \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\right] = \mathbb{E}\left[U_i U_j \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\right] = \mathbb{E}[U_i \mid \boldsymbol{X}_i]\mathbb{E}\left[U_j \mid \boldsymbol{X}_j\right] = 0$$

ここで、 $\sigma_i^2 = \sigma^2(\boldsymbol{X}_i) = \mathbb{E}\left[U_i^2 \mid \boldsymbol{X}_i\right]$としている

回帰係数の分散の推定：均一分散を仮定する場合

N個のσ_i^2を推定するのは大変そうなので均一分散$\sigma_i^2 = \sigma^2$を仮定すると話が単純になる

$$\mathrm{Var}\left[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\right] = \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i' \sigma_i^2\right) \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1}$$

均一分散の仮定による近似（成り立たない場合は推定がうまくいかない）

$$\approx \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i' \sigma^2\right) \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1}$$

$$= \sigma^2 \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1}$$

分散σ^2を

$$s^2 = \frac{1}{N-K} \sum_{i=1}^{N} \hat{U}_i^2 = \frac{1}{N-K} \sum_{i=1}^{N} \left(Y_i - \boldsymbol{X}_i' \hat{\boldsymbol{\beta}}\right)^2$$

で推定して置き換えると、（均一分散の仮定のもとでの）回帰係数の分散が推定できる

$$\widehat{\mathrm{Var}}\left[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\right] = s^2 \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1}$$

あとは平方根を取れば標準誤差になる。 `lm()`ではこうして標準誤差を計算している

不均一分散だとなぜ標準誤差をうまく推定できないのか

話をわかりやすくするために、単回帰モデル

$$Y = \alpha + \beta X + U$$

を考える。単回帰モデルの最小二乗推定量は

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})\,(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

分散を$(X_i - \bar{X})^2$で重みをつけて平均していると解釈できる（平均から乖離しているほど重くする）

なので、傾き$\hat{\beta}$の分散は以下になる。

$$\mathrm{Var}\left[\hat{\beta} \mid X_1, \ldots, X_N\right] = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^2\sigma_i^2}{\left(\sum_{i=1}^{N}(X_i - \bar{X})^2\right)^2} = \left(\frac{1}{\sum_{i=1}^{N}(X_i - \bar{X})^2}\right)\frac{\sum_{i=1}^{N}(X_i - \bar{X})^2\sigma_i^2}{\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

一方、均一分散での近似は

$$\mathrm{Var}\left[\hat{\beta} \mid X_1, \ldots, X_N\right] \approx \left(\frac{1}{\sum_{i=1}^{N}(X_i - \bar{X})^2}\right)\sigma^2$$

なので、平均からの乖離でσ_i^2を加重平均するべきところをワンナンバーσ^2で近似している。
実際に均一分散が成り立つなら問題ないが、不均一分散の場合は乖離が生まれてしまう。

不均一分散が標準誤差に与える影響をシミュレーションで確認

均一分散、平均周りがばらついている不均一分散①、平均から離れた部分がばらついている不均一分散②の3種類を考える。全体としてのばらつきは同じ。

$$Y = X + U$$
$$X \sim \mathrm{Uniform}(\{1, 2, 3\})$$
$$U \sim \mathcal{N}\left(0, \sigma^2(X)\right)$$

誤差項Uの分散がXに依存

$$\sigma^2(X) = \begin{cases} 16^2 \text{ if } X = 1 \\ 16^2 \text{ if } X = 2 \\ 16^2 \text{ if } X = 3 \end{cases}$$

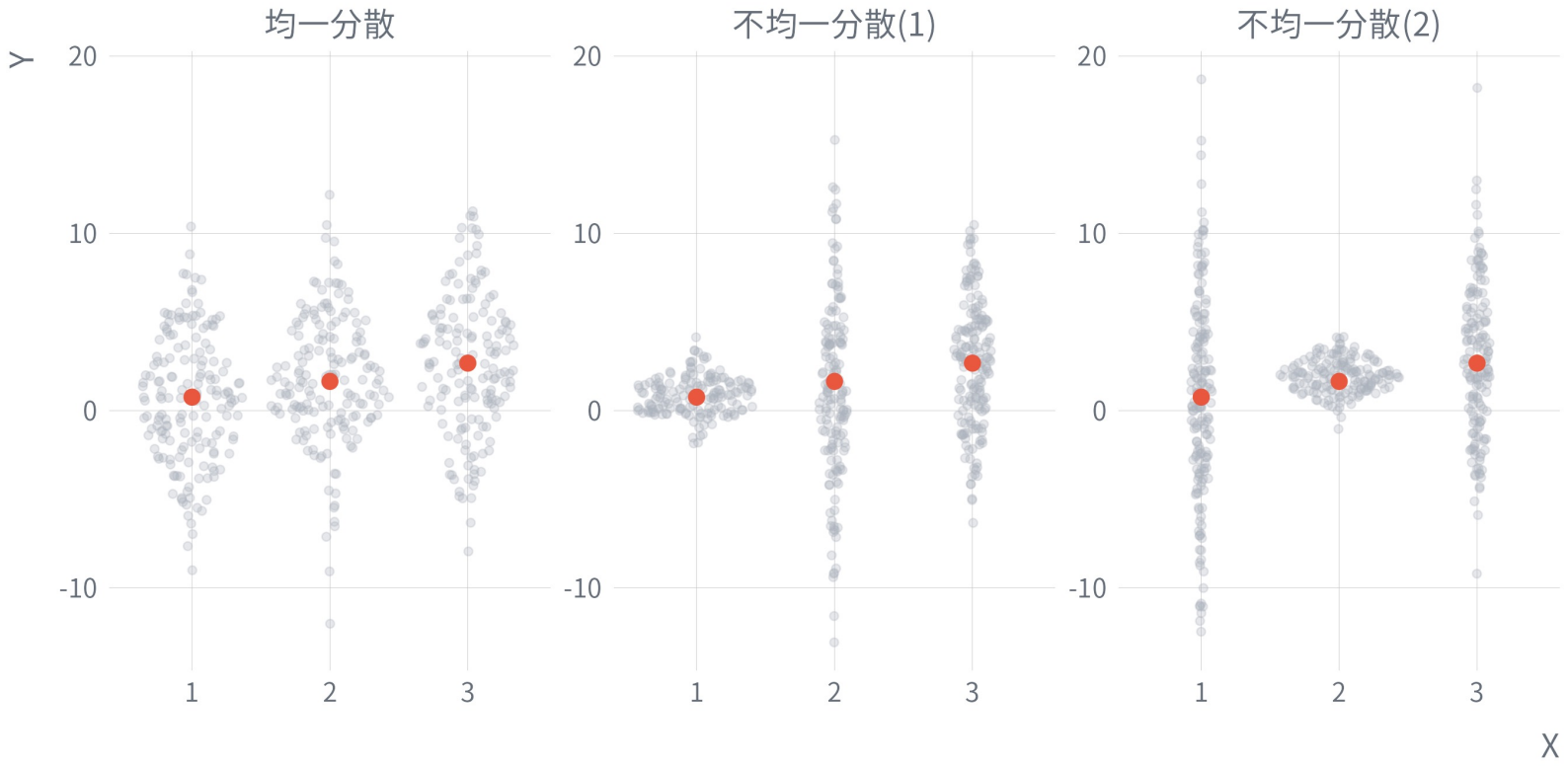
均一分散

$$\sigma^2(X) = \begin{cases} 1^2 \text{ if } X = 1 \\ 31^2 \text{ if } X = 2 \\ 16^2 \text{ if } X = 3 \end{cases}$$

不均一分散(1)

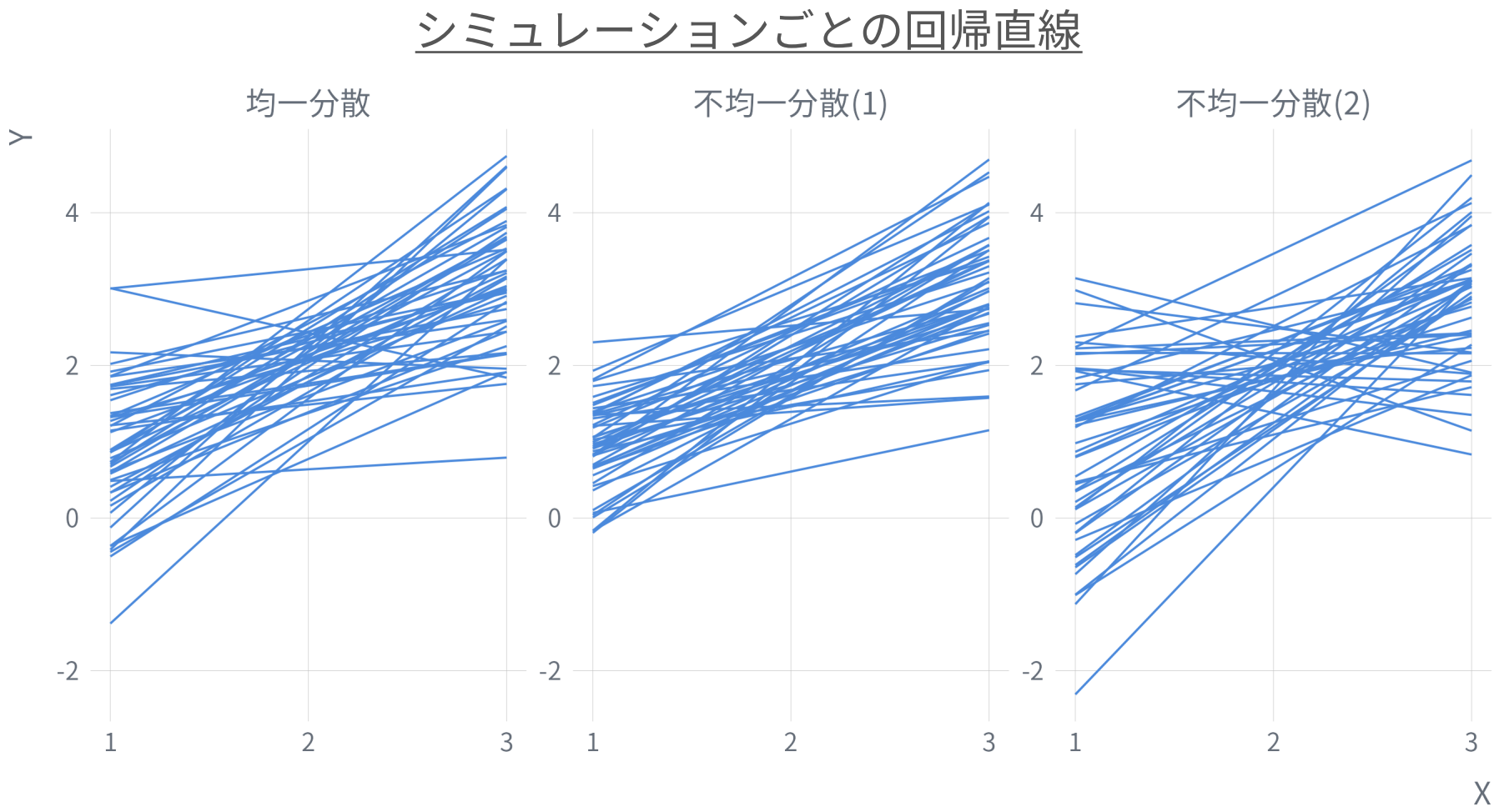
$$\sigma^2(X) = \begin{cases} 31^2 \text{ if } X = 1 \\ 1^2 \text{ if } X = 2 \\ 16^2 \text{ if } X = 3 \end{cases}$$

不均一分散(2)



均一分散と不均一分散で回帰直線のばらつきを比較

均一分散と比較して、不均一分散①（平均周りの分散大、左端の分散小）の場合は回帰直線のばらつきが小さく、不均一分散②（左端の分散大、平均周りの分散小）の場合はばらつきが大きい



均一分散を仮定した標準誤差は理論的なばらつきと乖離する

全体としては同程度の分散でも、不均一分散がある場合は回帰係数の理論的なばらつきは変わってくる

lm()の推定では均一分散を仮定しているので、全体の分散が同程度なら同じ標準誤差を返してしまう

分散タイプごとのlm()の出力

分散タイプ	回帰係数の 理論的なばらつき	lm()で推定された 標準誤差の平均
均一分散	0.70	0.70
不均一分散①	0.52	0.70
不均一分散②	0.84	0.69

Robust Standard Errorsで不均一分散に対処する

不均一分散を考慮した標準誤差の推定を考える

回帰係数の分散

$$\mathrm{Var}\left[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\right] = \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i' \mathbb{E}\left[U_i^2 \mid \boldsymbol{X}_i\right]\right) \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1}$$

を推定するにあたって、均一分散を仮定する場合は誤差項の平均的な分散を推定したs^2で置き換えていた

$$s^2 = \frac{1}{N-K} \sum_{i=1}^{N} \hat{U}_i^2$$

ここで、$\mathbb{E}\left[U_i^2 \mid \boldsymbol{X}_i\right]$をワンナンバー s^2で近似するのではなく、直接残差の二乗$\hat{U}_i^2$で置き換えるやり方も考えられる

$$\widehat{\mathrm{Var}}\left[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\right] = \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i' \hat{U}_i^2\right) \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1}$$

個別の$\hat{U}_i^2$はサンプルサイズ1なので誤差があるが、 Nが十分に大きくなれば$\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i' \hat{U}_i^2$はうまく推定できることを利用

上記の平方根をとったものをHeteroskedasticity-Consistent Standard ErrorsやRobust Standard Errorsと言う。Robust SEはいくつか手法が提案されていて、これは「HC0」と呼ばれているもの

RでRobust SEを計算するには？

`estimatr::lm_robust()`を

使えば計算できる

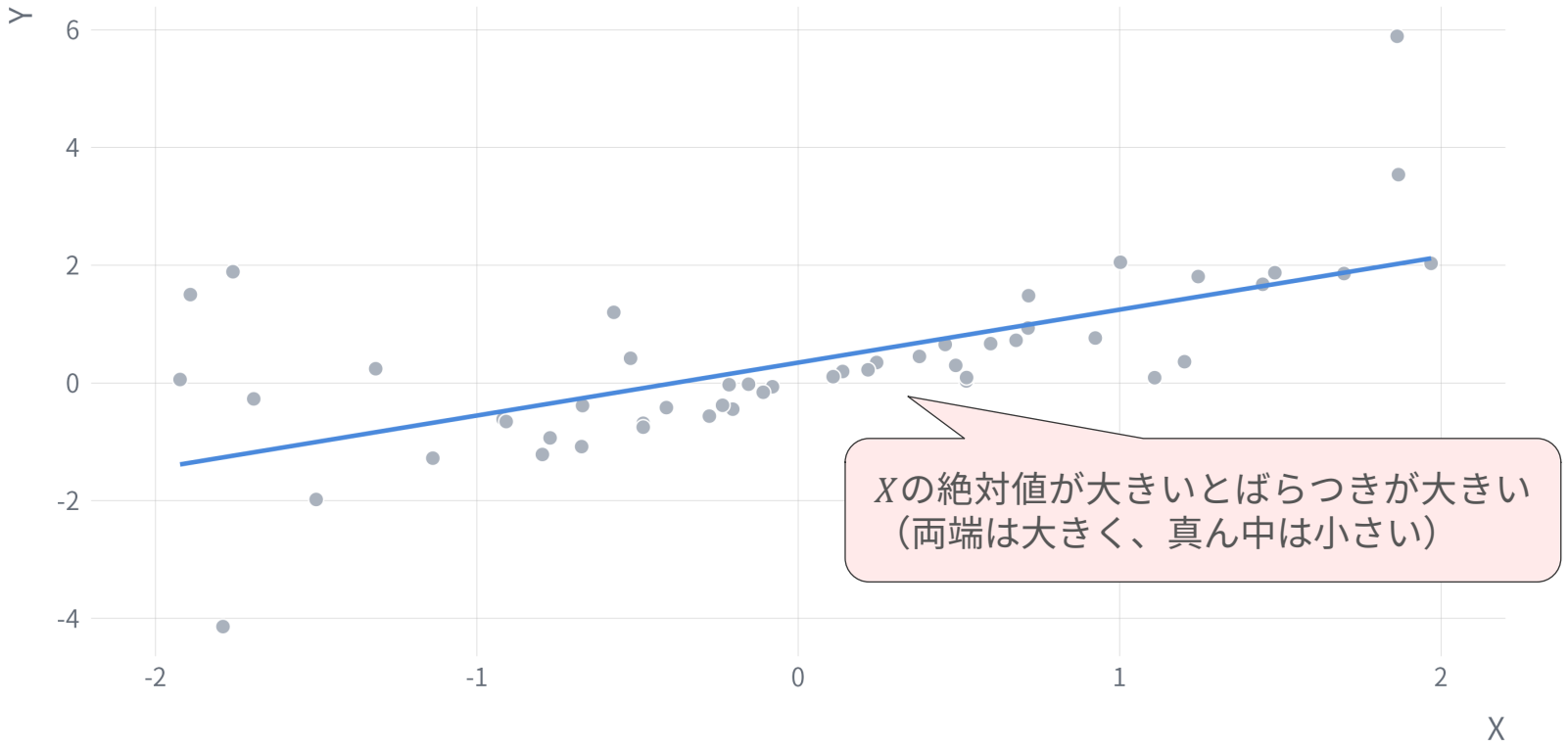
（再掲）不均一分散のシミュレーションデータ

誤差項Uの分散がXに依存（＝不均一分散）

$$Y = X + U$$
$$X \sim \text{Uniform}(-2, 2)$$
$$U \sim \mathcal{N}(0, X^2)$$

```
N <- 50
df_heterogeneous <- tibble(
  x = runif(N, -2, 2),
  u = rnorm(N, 0, abs(x)),
  y = x + u
)
```

シミュレーションデータと回帰直線



Xの絶対値が大きいとばらつきが大きい
（両端は大きく、真ん中は小さい）

estimatr::lm_robust()の挙動を確認

estimatr::lm_robust()でse_typeを指定することでRobust SEを計算できる
標準誤差を確認すると、Robust SEは誤差を大きめに見積もっている

```
> df_heterogeneous %>%
+   estimatr::lm_robust(y ~ x, data = ., se_type = "classical") %>%
+   tidy()
         term  estimate std.error statistic      p.value   conf.low conf.high df outcome
1 (Intercept) 0.3456885 0.1594783  2.167621 3.517795e-02 0.02503593  0.666341 48       y
2           x 0.8999665 0.1531487  5.876423 3.869962e-07 0.59204038  1.207893 48       y
```

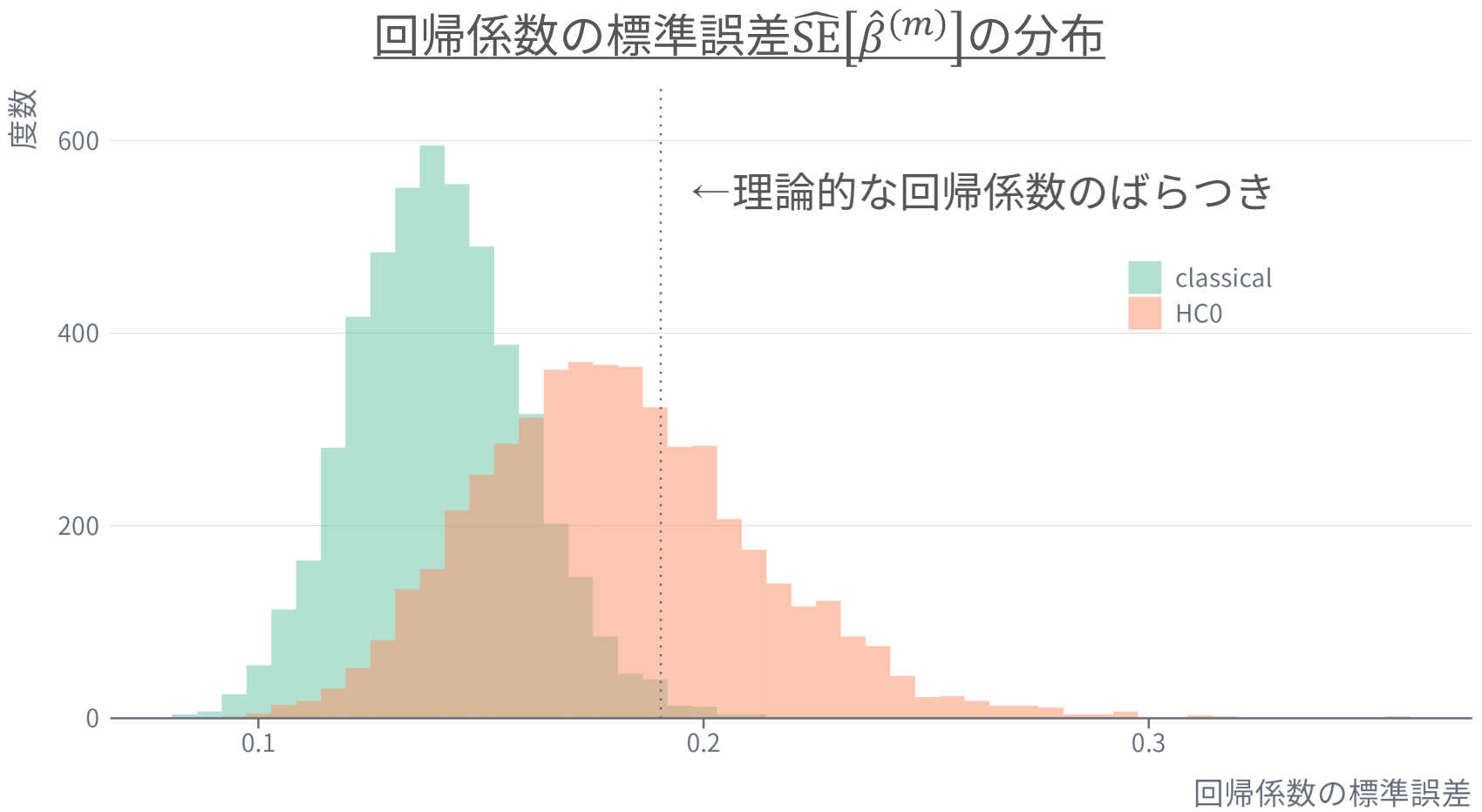
均一分散を仮定した標準誤差
（lm()の出力と同じ）

```
> df_heterogeneous %>%
+   estimatr::lm_robust(y ~ x, data = ., se_type = "HC0") %>%
+   tidy()
         term  estimate std.error statistic     p.value   conf.low conf.high df outcome
1 (Intercept) 0.3456885 0.1563591  2.210862 0.031843538 0.03130734 0.6600696 48       y
2           x 0.8999665 0.2389852  3.765784 0.000453233 0.41945455 1.3804784 48       y
```

不均一分散対してRobustな標準誤差

Robust SEを使うことで標準誤差の過小評価を軽減できる

- 均一分散を仮定した標準誤差（classical）と比べて、Robust SE（HC0）は標準誤差を大きく見積もっている
- Robust SEは理論的な回帰係数のばらつきよりも少し小さい傾向。これはサンプルサイズが50と小さいことによる



回帰係数の標準誤差をさらにうまく推定する

- HC0を改良した推定量がいくつか提案されている
- サンプルサイズが十分に大きければHC0 - HC3はどれも似たような値をとるが、小さい場合はHC2やHC3がベター。HC0 < HC2 < HC3なので、HC3がもっとも保守的
- `estimatr::lm_robust()`のデフォルトはHC2（ちなみにStataのデフォルトはHC1）

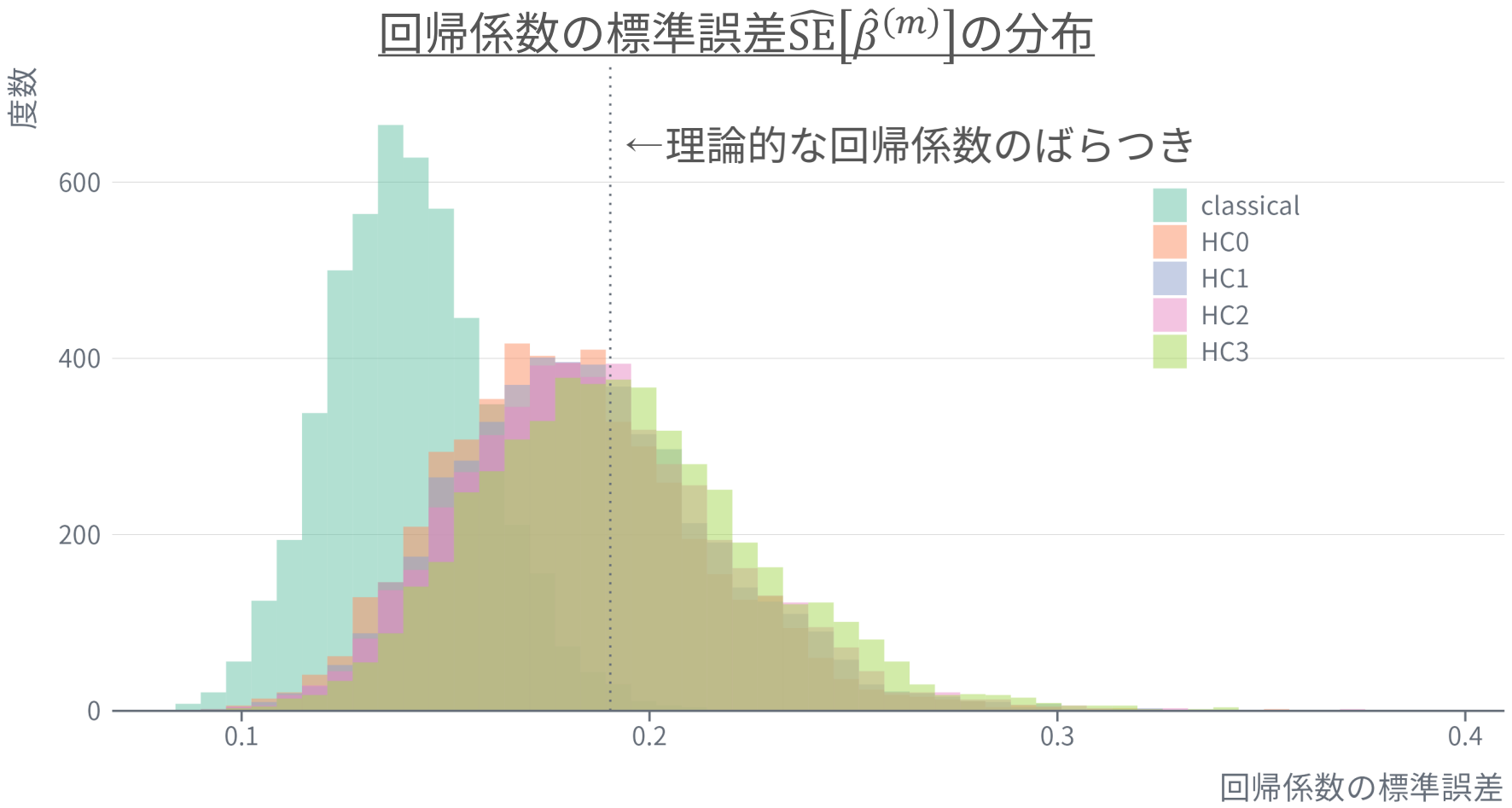
名称	計算式
HC0	$\left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i' \widehat{U}_i^2\right) \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1}$
HC1	$\frac{N}{N-K} \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i' \widehat{U}_i^2\right) \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1}$
HC2	$\left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i' \frac{\widehat{U}_i^2}{1-h_{ii}}\right) \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1}$
HC3	$\left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i' \frac{\widehat{U}_i^2}{(1-h_{ii})^2}\right) \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1}$

$$h_{ii} = \boldsymbol{X}_i' \left(\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1} \boldsymbol{X}_i$$

はleverageと呼ばれていて、観測値X_iの他の観測値と比較してどのくらい外れているかを表している。 $0 \leq h_{ii} \leq 1$なので、HC0<HC2<HC3。詳細はHansen(2022)を参照

Robust SEの比較

- HC0-HC3は似たような傾向を示す
- HC1-HC3はHC0よりもやや標準誤差を大きめに見積もる。このデータ例では、より保守的な標準誤差を出すHC3がやや優勢か



まとめ

まとめ

- Rは線形回帰分析ができる関数lm()をデフォルトで備えているが、lm()で計算される回帰係数の標準誤差は均一分散を仮定している。よって、不均一分散の状況下では標準誤差の推定にバイアスが生じてしまう
- 不均一分散を考慮した標準誤差としてRobust SEがある。実際のところ均一分散の仮定が成り立つことはあまりないと考えられるので、とりあえずRobust SEを使っておくことが推奨されている
- Rでは、estimatrパッケージのlm_robust()を使えばRobust SEを計算できる。sandwichパッケージのvcovHC()でも計算できるが、lm_robust()ならlm()を置き換えるだけで済むので便利
- Robust SEにはHC0, HC1, HC2, HC3のようにいくつかタイプがある。サンプルサイズが十分に大きければほとんど同じ値になるが、サンプルサイズが小さい場合はHC2やHC3がベターと言われている

参考文献

参考文献

- Hansen, Bruce E. "Econometrics." (2022). https://www.ssc.wisc.edu/~bhansen/econometrics/.
- Blair G, Cooper J, Coppock A, Humphreys M, Sonnet L. (2022). estimatr: Fast Estimators for Design-Based Inference. https://declaredesign.org/r/estimatr/. https://github.com/DeclareDesign/estimatr.

Rを使えるデータサイエンティスト、エンジニアを大募集