6. 回帰分析による統計的検定と推論

honocat

2025-12-29

回帰分析における仮説検定

データの準備

```
HR1996 <- read_csv('data/hr-data.csv') |>
  mutate(experience = as.numeric(status == '現職' | status == '元職'),
         expm = exp / 10 ^ 6) |>
  filter(year == 1996)
```

単回帰の例(1)

1996 年の選挙データを使って、「議員経験(experience; X)が得票率(voteshare; V)に影響する」という仮説を検証する。この仮説を、統計モデルとして以下のように表現する。

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

このモデルは分析上の仮定(assumption)であり、正しいとは限らないことに注意。

ここで検証する帰無仮説と対立仮説は以下の通り。

- 帰無仮説：$\beta = 0$
- 対立仮説：$\beta \neq 0$

これを回帰分析で検証する。

まず、`lm()` 関数を使って回帰式を推定する。

```
fit1 <- lm(voteshare ~ experience,
           data = HR1996)
```

結果を確認する。

```
broom::tidy(fit1)
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic   p.value
  <chr>          <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)     16.0     0.461      34.7 5.66e-186
2 experience      22.8     0.789      28.9 1.68e-141
```

estimate（推定値）の列の、(Intercept) の行にある数値が α の推定値 a、experience の行にある数値が β の推定値 b である。よって、この結果から、

$$\hat{V}_i = 16.0 + 22.8 X_i$$

という予測が得られる。

しかし、推定値は標本を取り直すごとに変わる。したがって、ここで得られた値をそのまま信じるわけには行かない。これらの値は、偶然得られただけで、真の値とはかけ離れているかもしれない。そこで、統計的検定を行う。検定には t 分布を使う。

私たちが立てた帰無仮説は、$\beta = 0$ である。推定量の標準誤差(standard error)は分析結果の std.error の列に表示されている。よって t 値は、

$$T = \frac{b - \tilde{\beta}}{\mathrm{SE}(b)} = \frac{b}{\mathrm{SE}(b)} \approx \frac{22.8274}{0.7891} \approx 28.93$$

である。R で b の値を取り出すには、

```
coef(fit1)[2] # or summary(fit1)$coefficients[2, 1]
```

```
experience
  22.82744
```

とする。また、$\mathrm{SE}(b)$ は、

```
summary(fit1)$coefficients[2, 2]
```

```
[1] 0.7891199
```

なので、t 値は、

```
with(summary(fit1), coefficients[2, 1] / coefficients[2, 2])
```

```
[1] 28.92772
```

である。この値が上の表の statistic の列に表示されている。この値を t 分布の臨界値と比較する。

有意水準を 5 %にして検定を実施する。利用する t 分布の自由度は、$N - K - 1$ である。N は、

```
(N1 <- nrow(HR1996))
```

```
[1] 1261
```

である。単回帰なので $K = 1$。よって、求める臨界値は、

```
(c1 <- qt(p = 0.05 / 2, df = N1 - 1 - 1, lower.tail = FALSE))
```

```
[1] 1.96185
```

である。

$$|T| = 28.93 > 1.96 = |c|$$

となるので、有意水準 5% で帰無仮説を棄却する。よって、$\beta \neq 0$ である。

ここから、過去の議員経験は得票率に影響すると考える。$b \approx 22.8$ なので、議員経験がない場合に比べ、議員経験がある場合には**平均すると** 22.8 ポイント得票が増えることが期待される。22.8 ポイントの差は実質的に大きな差であり、議員経験が実質的にも重要な意味を持っていると言えそうだ(ただし、この結論は仮定した統計モデルに依存しているという点に注意)。

続いて、β の 95% 信頼区間を求める。95% 信頼区間の下限値は、

```
coef(fit1)[2] - c1 * summary(fit1)$coefficients[2, 2]
```

```
experience
   21.2793
```

上限値は、

```
coef(fit1)[2] + c1 * summary(fit1)$coefficients[2, 2]
```

```
experience
  24.37557
```

である。よって、求める 95% 信頼区間は、[21.28, 24.38] である。

この区間は、`confint()` によって求めることもできる。

```
confint(fit1, level = 0.95)
```

```
               2.5 %   97.5 %
(Intercept) 15.10294 16.91102
experience  21.27930 24.37557
```

`broom::tidy()` にもある。

```
broom::tidy(fit1, conf.int = TRUE, conf.level = 0.95)
```

```
# 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
```

単回帰の例(2)

次に、「選挙費用 [単位：百万円]（`expm`; M)が得票率(`voteshare`; V)に影響する」という仮説を検証する。この仮説を、統計モデルとして以下のように表現する。

$$V_i \sim \text{Normal}(\alpha + \beta M_i, \sigma)$$

このモデルは分析上の仮定(assumption)であり、正しいとは限らない。

ここで検証する帰無仮説と対立仮説は以下の通り。

- 帰無仮説：$\beta = 0$
- 対立仮説：$\beta \neq 0$

これを回帰分析で検証する。

まず、`lm()` 関数を使って回帰式を推定する。

```
fit2 <- lm(voteshare ~ expm,
           data = HR1996)
```

結果を確認する。

```
broom::tidy(fit2)
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic   p.value
  <chr>          <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)     7.44    0.665       11.2 9.82e- 28
2 expm            1.88    0.0609      30.8 3.80e-154
```

この結果から、

$$\hat{V}_i = 7.4 + 1.9 M_i$$

という予測が得られる。

繰り返しになるが、推定値は標本を取り直すごとに変わる。したがって、ここで得られた値をそのまま信じるわけには行かない。これらの値は、偶然得られただけで、真の値とはかけ離れているかもしれない。そこで、統計的検定を行う。検定には、t 分布を利用する。

私たちが立てた帰無仮説は $\beta = 0$ である。よって、t 値は、

$$T = \frac{b - \tilde{\beta}}{\mathrm{SE}(b)} = \frac{b}{\mathrm{SE}(b)} \approx \frac{1.87687}{0.06086} \approx 30.84$$

である。この値が、上の表の `statistic` の列に表示されている。この値を t 分布の臨界値と比較する。

有意水準 7% にして検定を実施する。利用する t 分布の自由度は、$N - K - 1$ である。N は、

```
(N2 <- length(fit2$fitted.values))
```

```
[1] 1198
```

である（`expm` に欠損があるので、数が変わる）。単回帰なので $K = 1$ である。よって、求める臨界値は、

```
(c2 <- qt(p = 0.07 / 2, df = N2 - 1 - 1, lower.tail = FALSE))
```

```
[1] 1.813534
```

である。

$$|T| = 30.84 > 1.81 = |c|$$

だから、有意水準 7% で帰無仮説を棄却する。よって、$\beta \neq 0$ である。ここから、選挙費用は得票率に影響すると考える。$b \approx 1.9$ なので、選挙費用を 100 万円増やすごとに**平均すると** 1.9 ポイント得票が増えることが期待

される。得票を 10 ポイント上昇させるには、$\frac{10}{1.9} \approx 5.3$ 百万円選挙費用を増やせば良いことに成る。得票率は、選挙の支出によってある程度変化すると言えるかもしれない。

93% 信頼区間を求める。下限値は、

```
coef(fit2)[2] - c2 * summary(fit2)$coefficients[2, 2]
```

```
    expm
1.766503
```

上限値は、

```
coef(fit2)[2] + c2 * summary(fit2)$coefficients[2, 2]
```

```
    expm
1.987246
```

である。

この区間は、`confint()` によって求めることもできる。

```
confint(fit2, level = 0.93)
```

```
                3.5 %   96.5 %
(Intercept) 6.238630 8.650744
expm        1.766503 1.987246
```

重回帰の例

「過去の議員経験(`experience`; X)と選挙費用 [単位：百万円]（`expm`; M)が得票率(`voteshare`; V)に影響する」という仮説を検証する。この仮説を統計モデルとして以下のように表現する。

$$V_i \sim \text{Normal}(\beta_0 + \beta_1 X_i + \beta_2 M_i, \sigma)$$

■包括的仮説検定（結合仮説の検定）

以下の帰無仮説と対立仮説を利用する。

- 帰無仮説：$\beta_1 = \beta_2 = 0$
- 対立仮説：$\beta_1 \neq 0$ または $\beta_2 \neq 0$

■個別的仮説検定

以下の帰無仮説と対立仮説を利用する。

- 議員経験に関する仮説
  - 帰無仮説 1：$\beta_1 = 0$
  - 対立仮説 1：$\beta_1 \neq 0$
- 選挙費用に関する仮説
  - 帰無仮説 2：$\beta_2 = 0$
  - 対立仮説 2：$\beta_2 \neq 0$

回帰式を推定する。回帰式自体は、包括的仮説検定でも個別的仮説検定でも同じである。

```
fit3 <- lm(voteshare ~ experience + expm,
           data = HR1996)
```

まず、包括的仮説(結合仮説)検定を考える。summary() で結果を表示する。

```
summary(fit3)
```

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

Residuals:
    Min      1Q  Median      3Q     Max
-31.919  -7.419  -0.936   6.088  53.340

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.77407    0.59742   13.01   <2e-16 ***
experience  13.61318    0.80134   16.99   <2e-16 ***
expm         1.31223    0.06396   20.52   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.34 on 1195 degrees of freedom
  (63 observations deleted due to missingness)
Multiple R-squared:  0.5513,    Adjusted R-squared:  0.5506
F-statistic: 734.2 on 2 and 1195 DF,  p-value: < 2.2e-16
```

包括的仮説検定では、この結果のうち `Residual standard error` ブロックに表示される値を利用する。帰無仮説が正しい場合、F 値

$$F_0 = \frac{R^2}{1-R^2}\frac{N-K-1}{K}$$

が、第 1 自由度 K、第 2 自由度 $N-K-1$ の F 分布に従って分布する。この回帰分析における F 値は、上の結果の一番下に行に $F-statistic$ として表示されている。すなわち、$F_0 = 734.2$ である。この値を、F 分布の臨界値と比較する。

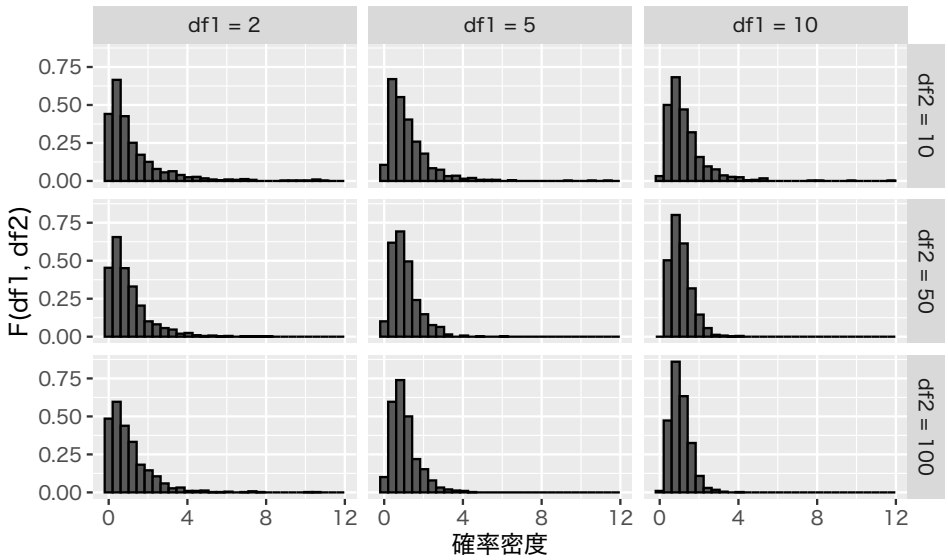
実際に検定を行う前に、F 分布がどんな分布に成るか確認しておこう。F 分布の母数(パラメタ)は 2 つの自由度である。

$$x \sim F(\mathrm{df1}, \mathrm{df2})$$

として、F 分布からランダムに抽出した x の分布の例を示す。

```
seq1 <- c(2, 5, 10)
seq2 <- c(10, 50, 100)
Fdist <- tibble()
for (i in seq_along(seq1)) {
  for (j in seq_along(seq2)) {
    Fdist <- tibble(
      df1 = seq1[i],
      df2 = seq2[j],
      x = rf(1000, df1 = seq1[i], df2 = seq2[j])
    ) |>
      bind_rows(Fdist)
  }
}
Fdist <- Fdist |>
  mutate(df1 = paste('df1 =', df1),
         df2 = paste('df2 =', df2),
         df1 = factor(df1,
                      levels = c('df1 = 2',
                                 'df1 = 5',
                                 'df1 = 10')),
         df2 = factor(df2,
                      levels = c('df2 = 10',
                                 'df2 = 50',
                                 'df2 = 100')))
```

```
plt_F <- ggplot(Fdist, aes(x = x, y = after_stat(density))) +
  geom_histogram(color = 'black') +
  facet_grid(df2 ~ df1) +
  labs(x = '確率密度', y = 'F(df1, df2)')
plot(plt_F)
```



このように、F 分布は正の値しか取らない。したがって、F 分布を使った検定では、F 値は臨界値の絶対値を考える必要はない。

仮説検定で利用する F 分布の自由度も結果の最終行に表示されており、第 1 自由度は 2（説明変数の数 $K = 2$）、第 2 自由度は 1195 である。有意水準 5 %として、検定の臨界値を求める。F 分布の臨界値は `qf()` で求める。

```
qf(0.05, df1 = 2, df2 = 1195, lower.tail = FALSE)
```

```
[1] 3.003255
```

この結果を使うと、

$$F_0 = 734.2 > 3.00 = c$$

なので、有意水準 5% で帰無仮説を棄却する。よって、$\beta_1 \neq 0$ または $\beta_2 \neq 0$ である。

続いて、個別的仮説検定を行う。そのために、`broom::tidy()` で結果を表示する。

```
broom::tidy(fit3, conf.int = TRUE)
```

```
# A tibble: 3 x 7
  term        estimate std.error statistic  p.value conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
1 (Intercept)     7.77    0.597       13.0 2.67e-36     6.60      8.95
2 experience     13.6     0.801       17.0 3.83e-58    12.0      15.2
3 expm            1.31    0.0640      20.5 2.22e-80     1.19      1.44
```

議員経験 experience について私たちが立てた帰無仮説は $\beta_1 = 0$ である。よって、t 値は、

$$T_1 = \frac{b_1 - \tilde{\beta}_1}{\mathrm{SE}(b_1)} = \frac{b_1}{\mathrm{SE}(b_1)} \approx \frac{13.613177}{0.08013375} \approx 16.99$$

である。この値が、上の表の statistic の列に表示されている。この値を t 分布の臨界値と比較する。

有意水準を 4% にして、検定を実施しよう。利用する t 分布の自由度は $N - K - 1$ である。N は、

```
(N3 <- length(fit3$fitted.values))
```

```
[1] 1198
```

である。説明変数が 2 つあるので $K = 2$ である。よって求める臨界値は、

```
(c3 <- qt(p = 0.04 / 2, df = N3 - 2 - 1, lower.tail = FALSE))
```

```
[1] 2.055993
```

である。

$$|T_1| = 16.99 > 2.06 = |c|$$

だから、有意水準 4% で帰無仮説を棄却する。よって、$\beta_1 \neq 0$ である。

同様に、選挙費用 expm について、私たちが立てた帰無仮説は、$\beta_2 = 0$ である。よって、t 値は、

$$T_2 = \frac{b_2 - \tilde{\beta}_2}{\mathrm{SE}(b_2)} = \frac{b_2}{\mathrm{SE}(b_2)} \approx \frac{1.312231}{0.0639582} \approx 20.52$$

である。この値が上の表の statistic の列に表示されている。この値を t 分布の臨界値と比較する。有意水準 4% の臨界値は上で求めた c3 である。

$$|T_2| = 20.52 > 2.06 = |c|$$

だから、有意水準 4% で帰無仮説を棄却する。よって、$\beta \neq 0$ である。

ここから、議員経験と選挙費用はどちらも得票率に影響すると考える。$b1 \approx 13.6$ なので、**選挙費用が同じ候補者同士を比べる**と、議員経験がある者のほうが経験がない者よりも**平均して** 13.6 ポイント得票率が高いと予測できる。同様に、$b_2 \approx 1.3$ なので、**議員経験の有無が同じ場合**には、選挙費用を 100 万円増やすごとに**平均すると** 1.3 ポイント得票が増えることが予測される。

シミュレーションで回帰分析を理解する

回帰分析のしくみ(特に信頼区間の意味)を理解するために、簡単なモンテカルロシミュレーションを行う。シミュレーションでは、自分で母数(パラメタ)を設定し、データを生成する。そのうえで、特徴を知りたい分析手法(今回の場合は線形回帰を用いた回帰分析)を生成したデータに当てはめ、母数をうまく推測できるかどうか確認する。

今回は、単回帰を例にシミュレーションを行う。シミュレーションを行う主な目的は以下の 3 つである。

1. 線形回帰が想定するデータ生成過程(data generating process)を理解する
2. 線形回帰の推定量の基本的な性質を理解する
3. 信頼区間の意味を理解する

単回帰モデルは、以下の通りである。

$$Y_i \sim \text{Normal}(\beta_1 + \beta_2 X_i, \sigma)$$

したがって、設定する母数は 3 つ(β_1, β_2, σ)である。

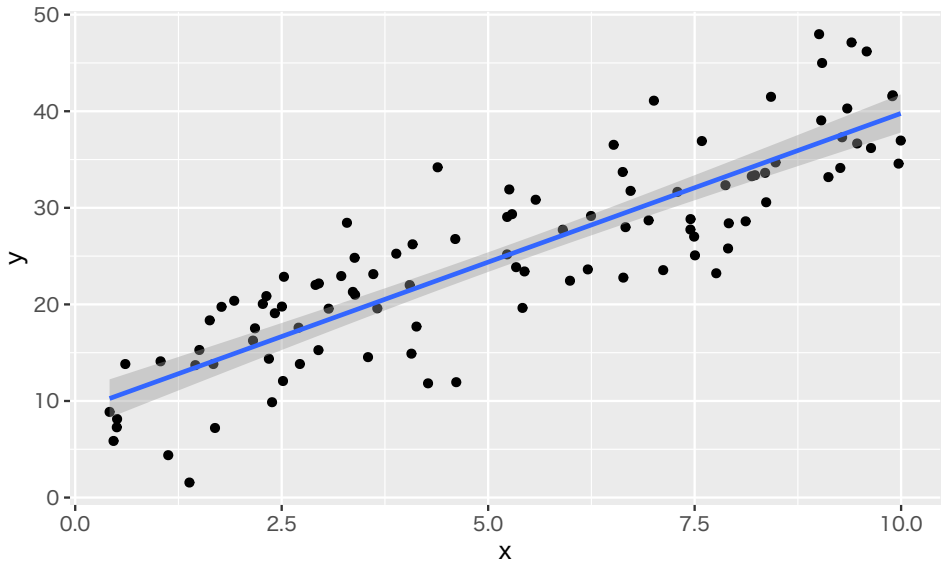
```
beta1 <- 10
beta2 <- 3
sigma <- 6
```

次に、単回帰モデルが想定する**データ生成過程**に従って、データを生成する。

```
N <- 100
x <- runif(N, min = 0, max = 10)
y <- rnorm(N, beta1 + beta2 * x, sd = sigma)
```

ここで、手に入れたデータを散布図にして直線を当てはめてみる。

```r
df <- data.frame(y, x)
scat <- ggplot(df, aes(x, y)) +
  geom_point() +
  geom_smooth(method = 'lm')
plot(scat)
```



回帰式を推定すると、

```r
eg_1 <- lm(y ~ x, data = df)
broom::tidy(eg_1)
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)     8.97     1.06       8.44 2.89e-13
2 x               3.08     0.179     17.2  1.90e-31
```

3 つの母数のうち、関心がある β_1, β_2 に対する推定値は、それぞれ 8.97, 3.08 であることがわかる。

このとき、係数の 95% 信頼区間は、

```r
confint(eg_1)
```

```
               2.5 %    97.5 %
(Intercept) 6.863256 11.082868
x           2.726403  3.435404
```

で求められる。95% 信頼区間は、切片が [6.86, 11.08]、傾きが [2.73, 3.44] であり、どちらの信頼区間も母数を含んでいる。つまり、ここで得られた信頼区間が母数を含む確率は 1（100%）である！

以上の過程を、母数とサンプルサイズは変えずに何度も繰り返す。

```
sim_ols1 <- function(beta, sigma, n = 100, trials = 10000,
                     x_rng = c(0, 10)) {
  ## 単回帰をシミュレートする関数
  ## 引数
  ##   beta:   係数パラメタのベクトル
  ##   sigma:  誤差の標準偏差
  ##   n:      標本サイズ
  ##   trials: シミュレーションの繰り返し回数
  ##   x_rng:  説明変数 x の値の範囲
  ## 返り値
  ##   df: 以下の列を含むデータフレーム
  ##     (1) パラメタの推定値
  ##     (2) 各パラメタの推定値の標準誤差
  ##     (3) 各パラメタの 95% 信頼区間

  ## 結果を保存するためのデータフレームを作る
  col_names <- c('b1', 'b1_se', 'b1_lower', 'b1_upper',
                 'b2', 'b2_se', 'b2_lower', 'b2_upper',
                 'sigma_hat')
  df <- as.data.frame(matrix(rep(NA, trials * length(col_names)),
                             ncol = length(col_names)))
  names(df) <- col_names

  for (i in 1 : trials) {
    x <- runif(n, x_rng[1], x_rng[2])
    y <- rnorm(n, mean = beta[1] + beta[2] * x, sd = sigma)

    fit <- lm(y ~ x)

    sigma_hat <- summary(fit)$sigma

    b1 <- coef(fit)[1]
    b2 <- coef(fit)[2]

    b1_se <- sqrt(summary(fit)$cov.unscaled[1, 1]) * sigma_hat
```

```
    b2_se <- sqrt(summary(fit)$cov.unscaled[2, 2]) * sigma_hat

    b1_ci95 <- confint(fit)[1,]
    b2_ci95 <- confint(fit)[2,]

    df[i,] <- c(b1, b1_se, b1_ci95,
                b2, b2_se, b2_ci95,
                sigma_hat)
  }
  return(df)
}
```

この関数を利用して、実際にシミュレーションを行ってみる。自分で母数とサンプルサイズを指定し、データの生成と回帰式の推定を 1,000 回繰り返すことにする。

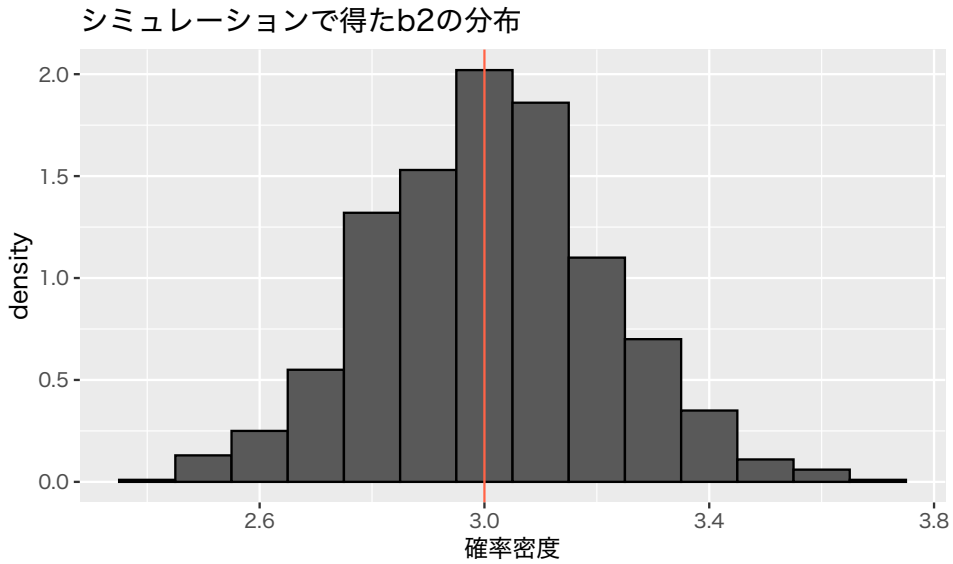
```
beta1 <- 10
beta2 <- 3
sigma <- 6
sim_1 <- sim_ols1(beta = c(beta1, beta2), sigma = sigma, trials = 1000)
head(sim_1)
```

```
         b1    b1_se b1_lower b1_upper       b2     b2_se b2_lower b2_upper
1 10.889735 1.219974 8.468737 13.31073 2.822854 0.2126011 2.400954 3.244754
2 10.062474 1.199309 7.682484 12.44246 2.846796 0.2179424 2.414296 3.279295
3 10.799442 1.141257 8.534655 13.06423 2.820502 0.1840057 2.455348 3.185655
4  8.147758 1.256388 5.654496 10.64102 3.213706 0.1960152 2.824720 3.602692
5 10.571428 1.128475 8.332005 12.81085 2.845771 0.1920806 2.464594 3.226949
6  9.403279 1.208117 7.005809 11.80075 3.174001 0.2117837 2.753723 3.594279
  sigma_hat
1  5.698033
2  6.069712
3  5.129270
4  5.936476
5  5.356491
6  6.090690
```

係数の推定値を理解する

説明変数 x の係数の推定値 b2 の分布を確認する。

```
hist_b2 <- ggplot(sim_1, aes(x = b2, y = after_stat(density))) +
  geom_histogram(binwidth = 0.1, color = 'black')
hist_b2 <- hist_b2 +
  labs(x = '確率密度', title = 'シミュレーションで得たb2の分布') +
  geom_vline(xintercept = 3, color = 'tomato')
plot(hist_b2)
```

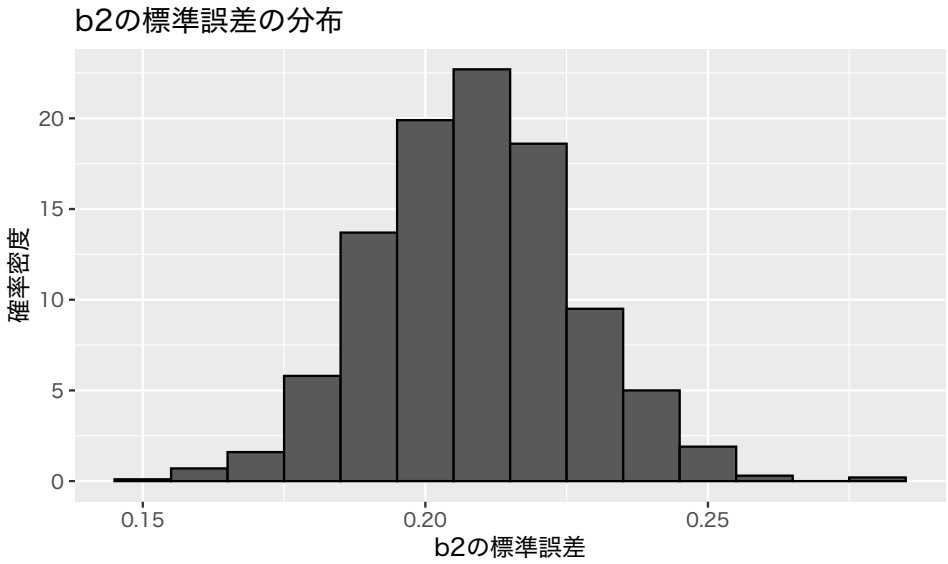


このヒストグラムが示すように、線形回帰は**平均すると**、母数をうまく推定してくれる（**不偏性, unbiasedness**）。しかし、常に正しい推定をするわけではなく、係数の値を過小推定することもあれば、過大推定することもある。実際の分析では、データセットが 1 つしかないのが普通であり、自分のデータ分析が係数を「正しく」推定しているとは限らない。そのために、推定の不確実性を明示することが求められるのである。

標準誤差を理解する

次に、b2 の標準誤差（standard error）をヒストグラムにしてみる。

```
hist_se <- ggplot(sim_1, aes(x = b2_se, y = after_stat(density))) +
  geom_histogram(binwidth = .01, color = 'black') +
  labs(x = 'b2の標準誤差', y = '確率密度',
       title = 'b2の標準誤差の分布')
plot(hist_se)
```



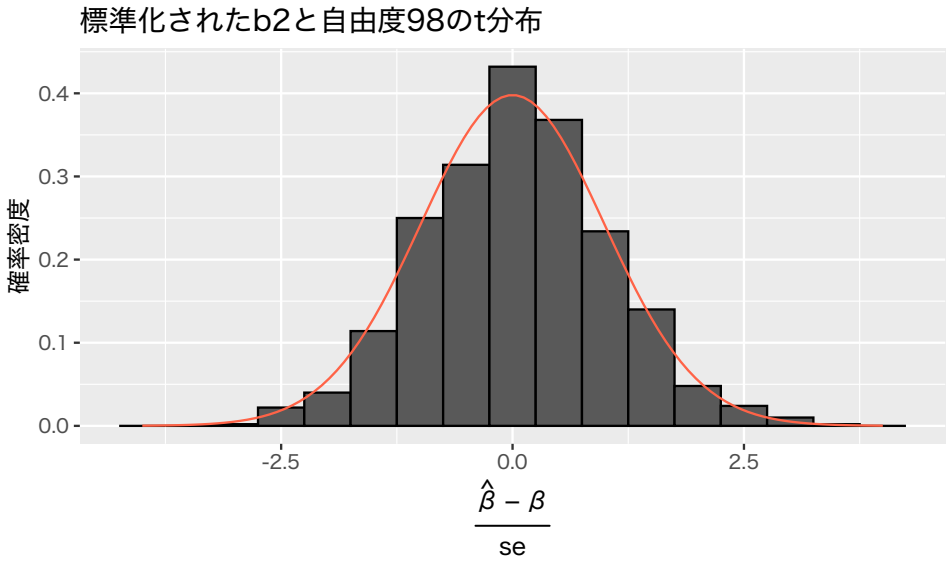
このように、標準誤差自体が推定量なので、値はサンプルごとに異なる（運ぶする）。

標準誤差を se とすると、

$$\frac{\hat{\beta} - \beta}{\text{se}}$$

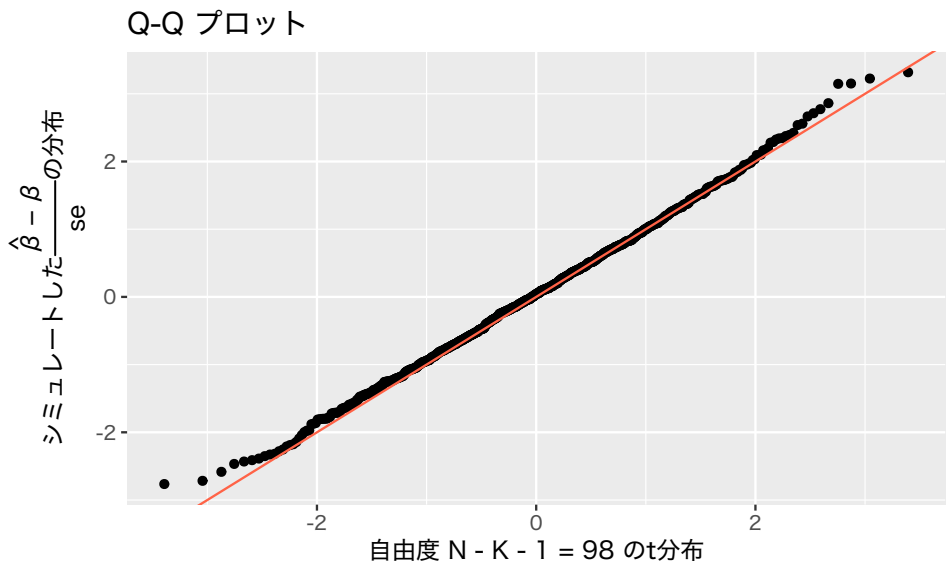
は自由度 $N - K - 1$（ここでは、$100 - 1 - 1 = 98$）の t 分布に従うはずである。まず、この値をヒストグラムにして、ジユ度 98 の t 分布の確率密度曲線を重ねてみる。

```
sim_1 <- sim_1 |>
  mutate(b2_t = (b2 - beta2) / b2_se)
true_t <- data.frame(x = seq(-4, 4, length = 100)) |>
  mutate(density = dt(x, df = 98))
hist_t <- ggplot(sim_1, aes(x = b2_t, y = after_stat(density))) +
  geom_histogram(binwidth = 0.5, color = 'black') +
  geom_line(data = true_t, aes(x = x, y = density), color = 'tomato')
hist_t <- hist_t +
  labs(x = expression(frac(hat(beta) - beta, 'se')),
       y = '確率密度',
       title = '標準化された b2 と自由度 98 の t 分布')
plot(hist_t)
```



この図から、t 分布に近い分布であることがわかる。

```
qqplot_t <- ggplot(sim_1, aes(sample = b2_t)) +
  stat_qq(distribution = qt, dparams = list(df = 98)) +
  geom_abline(intercept = 0, slope = 1, color = 'tomato') +
  labs(title = 'Q-Q プロット',
       x = '自由度 N - K - 1 = 98 の t 分布',
       y = expression(paste('シミュレートした',
                            frac(hat(beta) - beta, 'se'),
                            'の分布')))
plot(qqplot_t)
```



Q-Q プロットの点がほぼ一直線に並んでおり、$(\hat{\beta} - \beta)/\text{se}$ が t 分布に従っていることがわかる。ただし、分布の裾では、理論値との乖離が大きいことに注意。

今回のシミュレーションで得られた標準誤差の平均値は、

```
mean(sim_1$b2_se)
```

```
[1] 0.2087778
```

である。標準誤差は推定値の標準偏差のはずだが、本当にそうなっているか。

```
sd(sim_1$b2)
```

```
[1] 0.2053017
```

これで、実際に標準誤差は推定値の標準偏差に(ほぼ)一致することがわかった。

信頼区間を理解する

係数の推定値 b2 の 95% 信頼区間を例として考える。シミュレーションで得られた 1 つ目の信頼区間は、

```
sim_1[1, c('b2_lower', 'b2_upper')]
```

```
  b2_lower b2_upper
1 2.400954 3.244754
```

すなわち、[2.4, 3.24] が b2[1] お 95% 信頼区間である。この区間は母数である $\beta_2 = 3$ を含んでいる。したがって、**この信頼区間が母数を含む確率は 1（100%）である**。

しかし、信頼区間が母数を区間内に含んでいない場合がある。つまり、**信頼区間が母数を含む確率は 0 である**。

シミュレーションで得た 1,000 個の信頼区間のうち、どの信頼区間が母数を含んでるか調べてみる。母数を信頼区間内に含むのは、信頼区間の下限値が母数以下かつ上限値が母数以上のものである。

```
check_ci <- sim_1$b2_lower <= beta2 & beta2 <= sim_1$b2_upper
```

この結果、**TRUE** となっているものが母数を区間内に捉えているもの、**FALSE** がそうでないものである。これを表にすると、

```
table(check_ci)
```

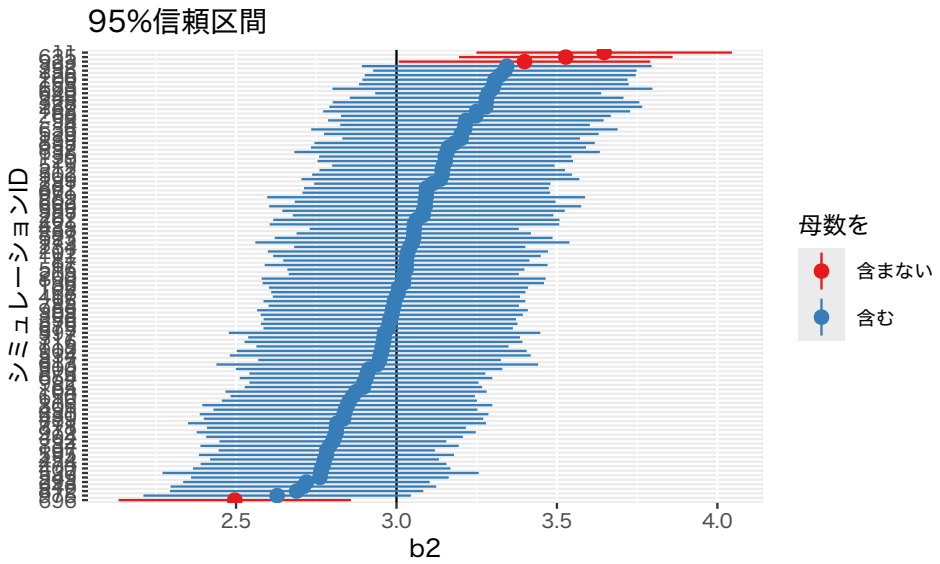
```
check_ci
FALSE  TRUE
```

```
    45    955
```

となる。つまり、1,000 個の 95% 信頼区間のうち、955 個(95.5%)は母数を区間内に捉え、残りの 45 個が捉えていないということである。このように、一定のデータ生成過程から得られた異なるデータセットに対し、信頼区間を求める作業を繰り返したとき、「得られた信頼区間のうち何 % が母数を区間内に含むか」というのが、信頼区間の信頼度である。

これを図にしてみる。無作為に 100 個だけ選ぶ。

```
sim_1 <- sim_1 |>
  mutate(id = 1 : n())
sim_1$check_ci <- check_ci
sim_1_sub <- sim_1 |>
  slice_sample(n = 100)
ciplt <- ggplot(sim_1_sub,
                aes(x = reorder(id, b2), y = b2,
                    ymin = b2_lower, ymax = b2_upper,
                    color = check_ci)) +
  geom_hline(yintercept = beta2, linethpe = 'dashed') +
  geom_pointrange() +
  labs(x = 'シミュレーション ID', y = 'b2', title = '95% 信頼区間') +
  scale_color_brewer(palette = 'Set1',
                     name    = '母数を',
                     labels  = c('含まない', '含む')) +
  coord_flip()
plot(ciplt)
```



このように、約 95% の信頼区間が、母数である 3 をまたいでいる。