BeginneR Session
-- Regression analysis --

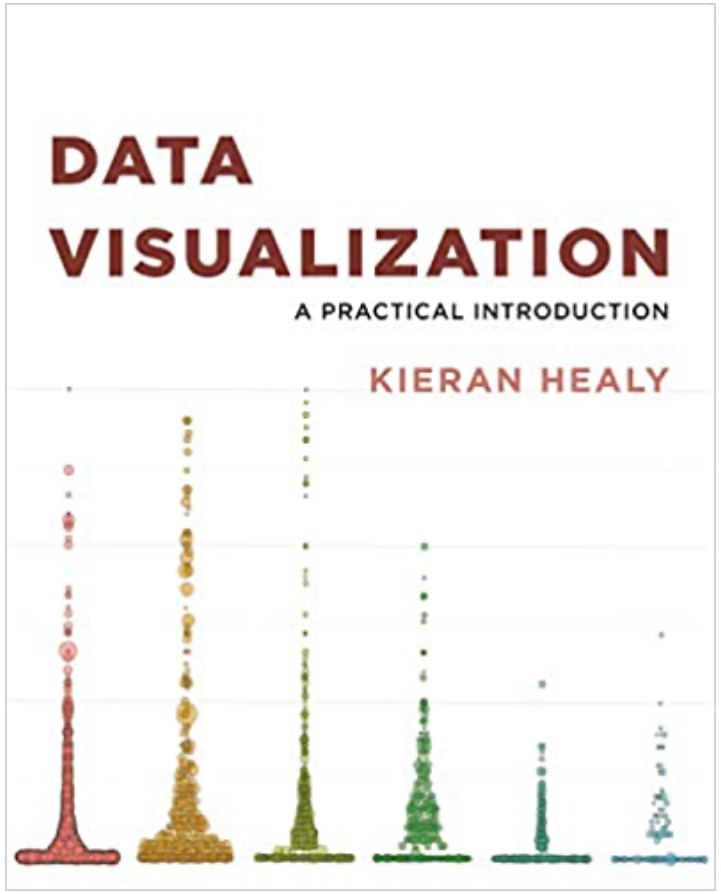
@kilometer00

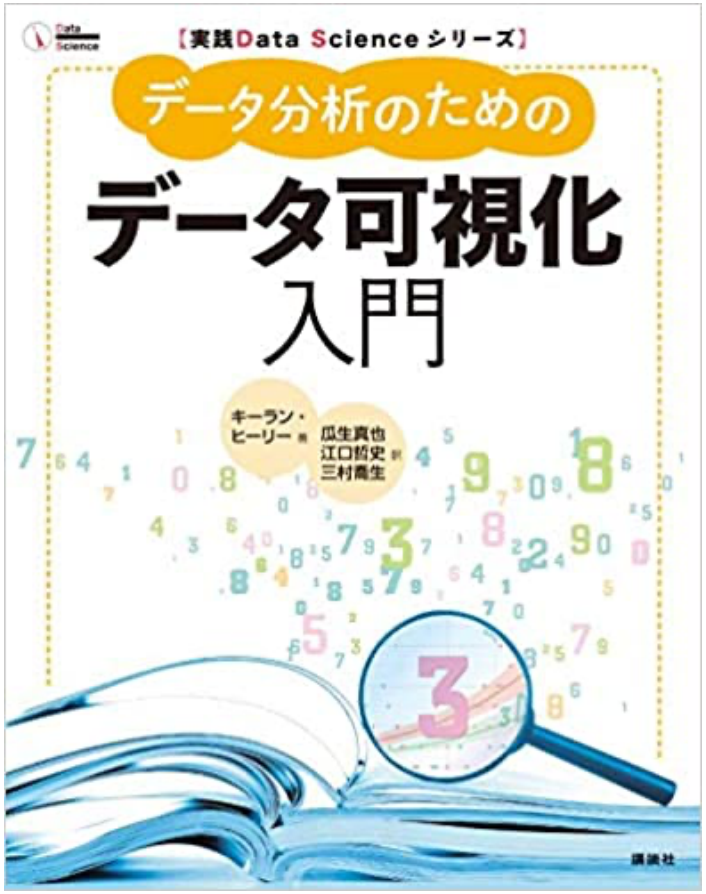
Who！？

Who！？

- @kilometer
- Postdoc Researcher (Ph.D. Eng.)
- Neuroscience
- Computational Behavior
- Functional brain imaging
- R： ~ 10 years

宣伝!!（書籍の翻訳に参加しました。）

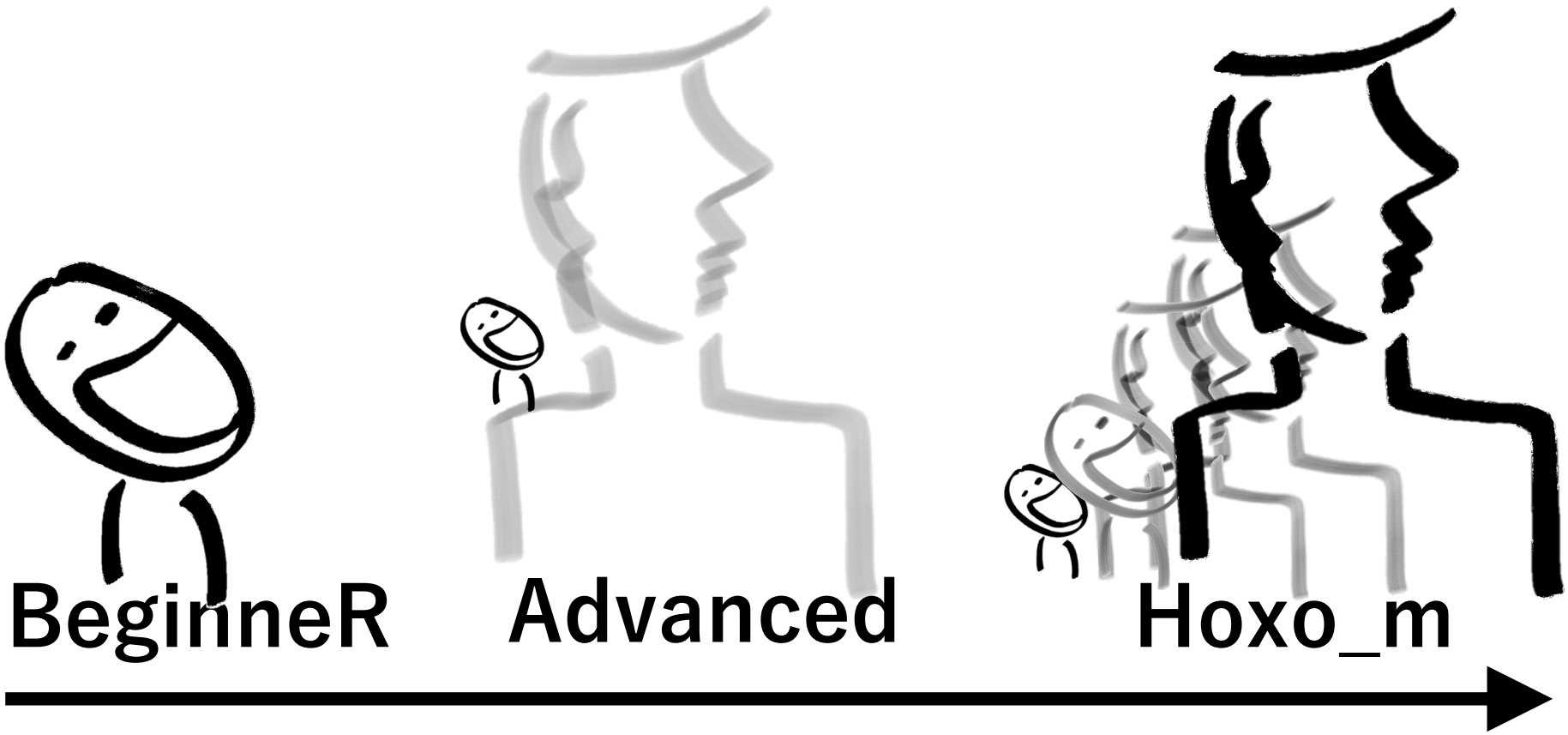




絶賛販売中！

BeginneR Session

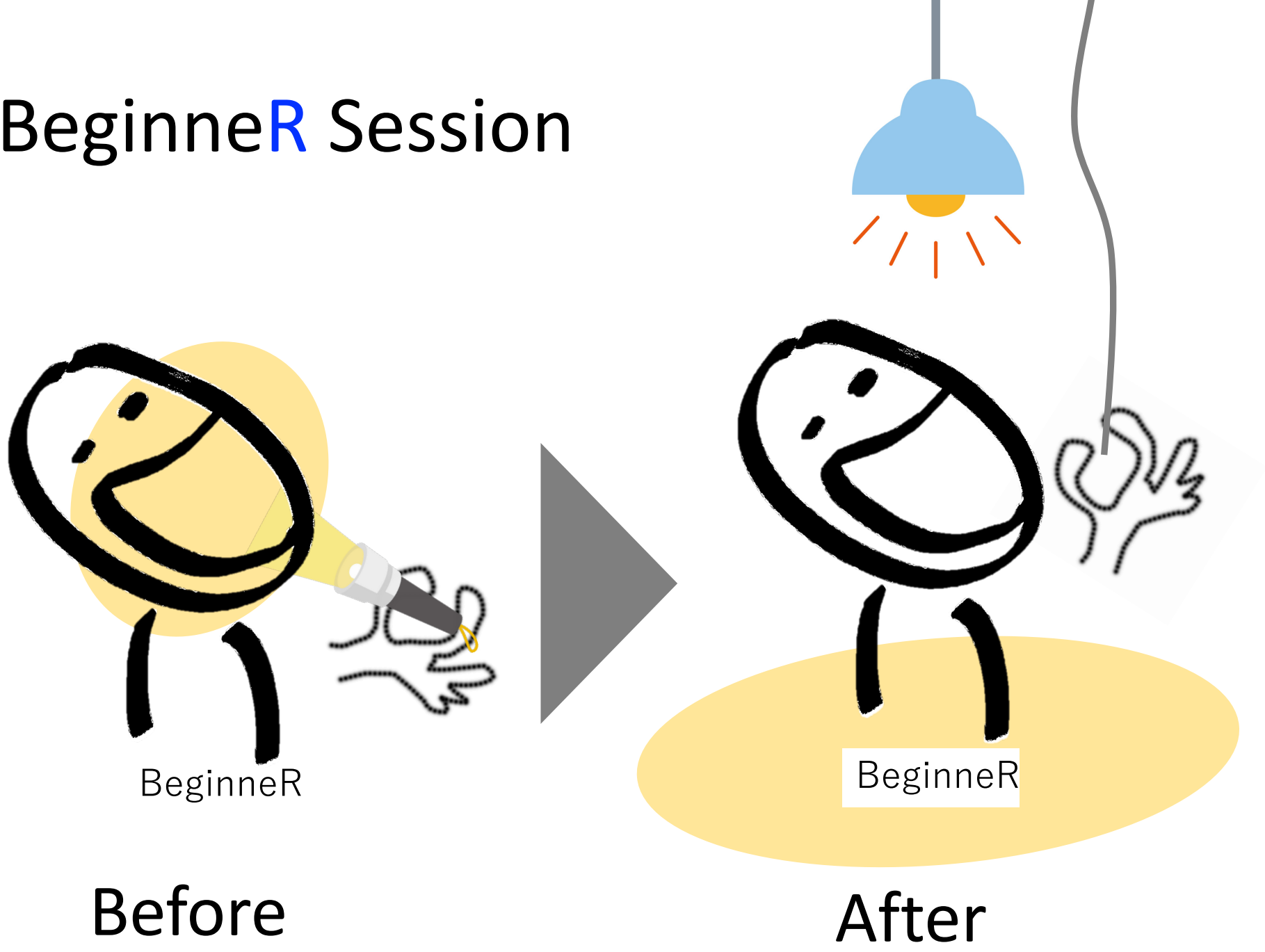
BeginneR



If I have seen further it is by standing on the shoulders of Giants.

-- Sir Isaac Newton, 1676

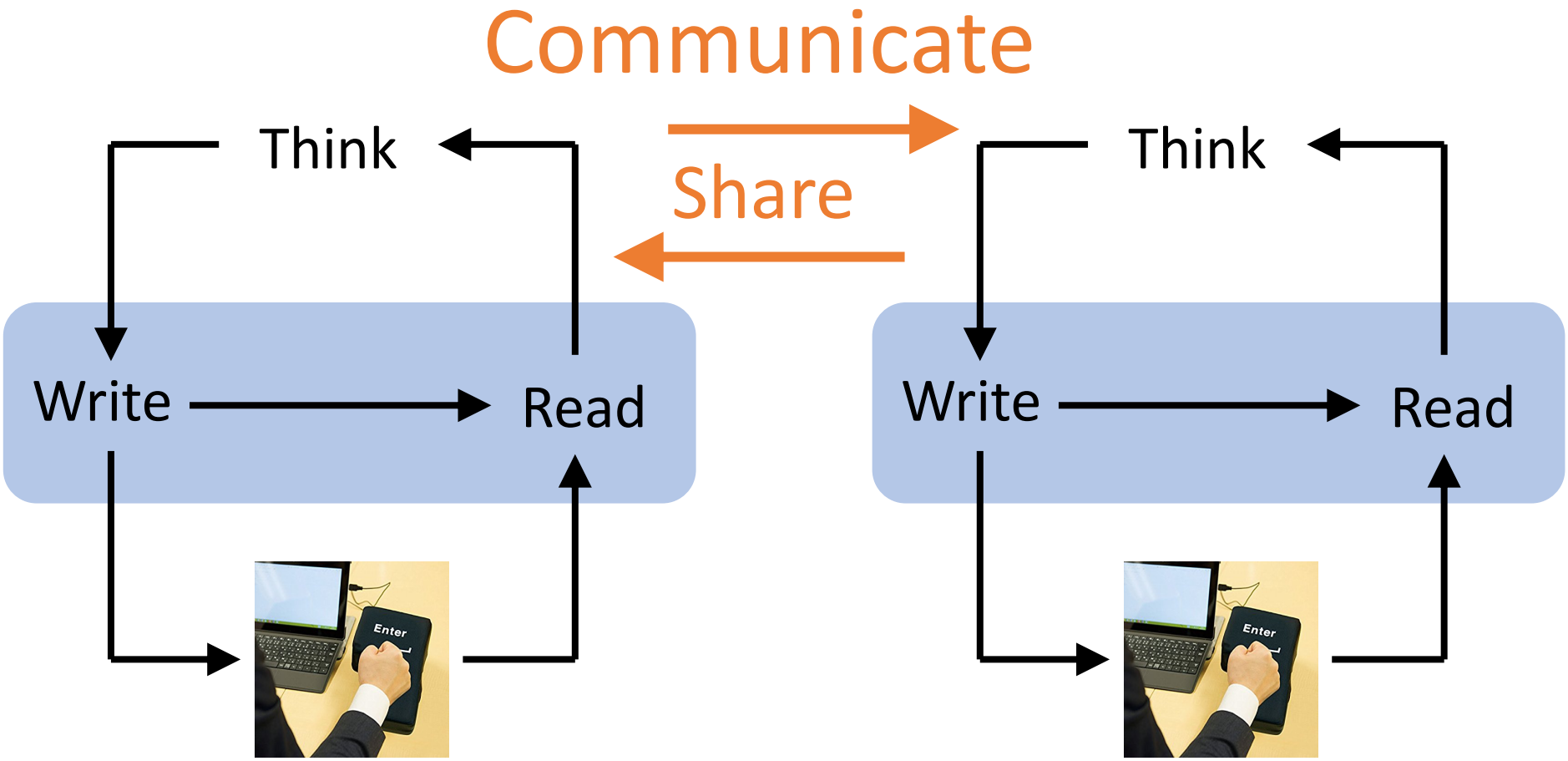
BeginneR Session



Before

After

Programing


Communicate
Think
Share
Think
Write
Read
Write
Read
Enter
Enter

BeginneR Session
-- Regression analysis --

@kilometer00


packages
you
R

Install

```
install.paclages("tidyverse")
```

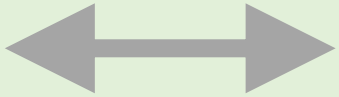
Attach

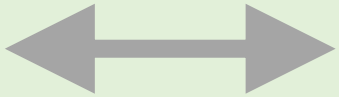
```
library("tidyverse")
```

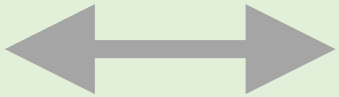
Pipe algebra %>%

{magrittr}

X %>% f  f(X)

X %>% f(y)  f(X, y)

X %>% f %>% g  g(f(X))

X %>% f(y, .)  f(y, X)

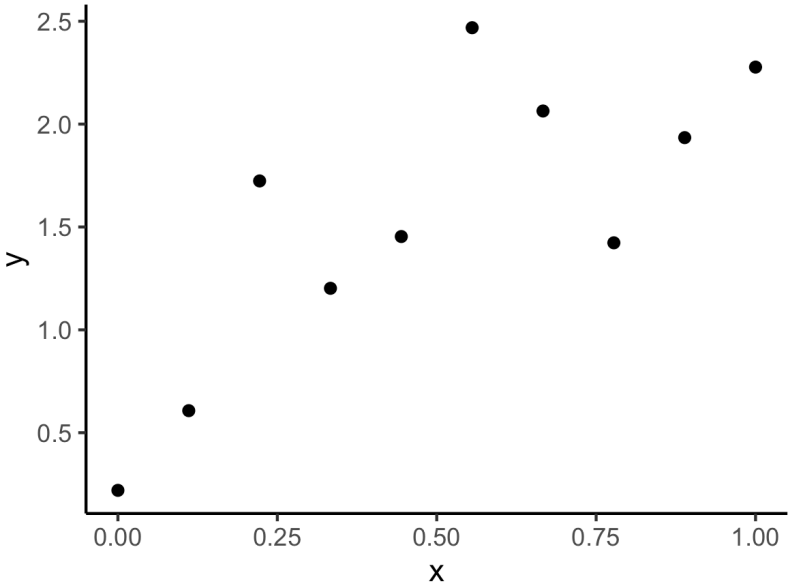
「dplyr再入門（基本編）」yutanihilation

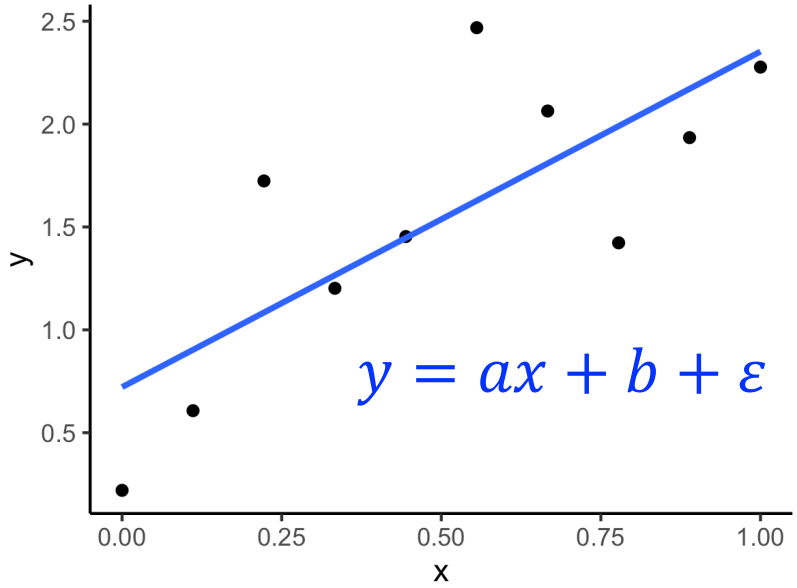
https://speakerdeck.com/yutannihilation/dplyrzai-ru-men-ji-ben-bian

線形回帰分析

- 回帰直線(最小二乗法)
- 誤差の確率モデル
- 決定係数と相関係数
- 回帰モデルの仮説検定

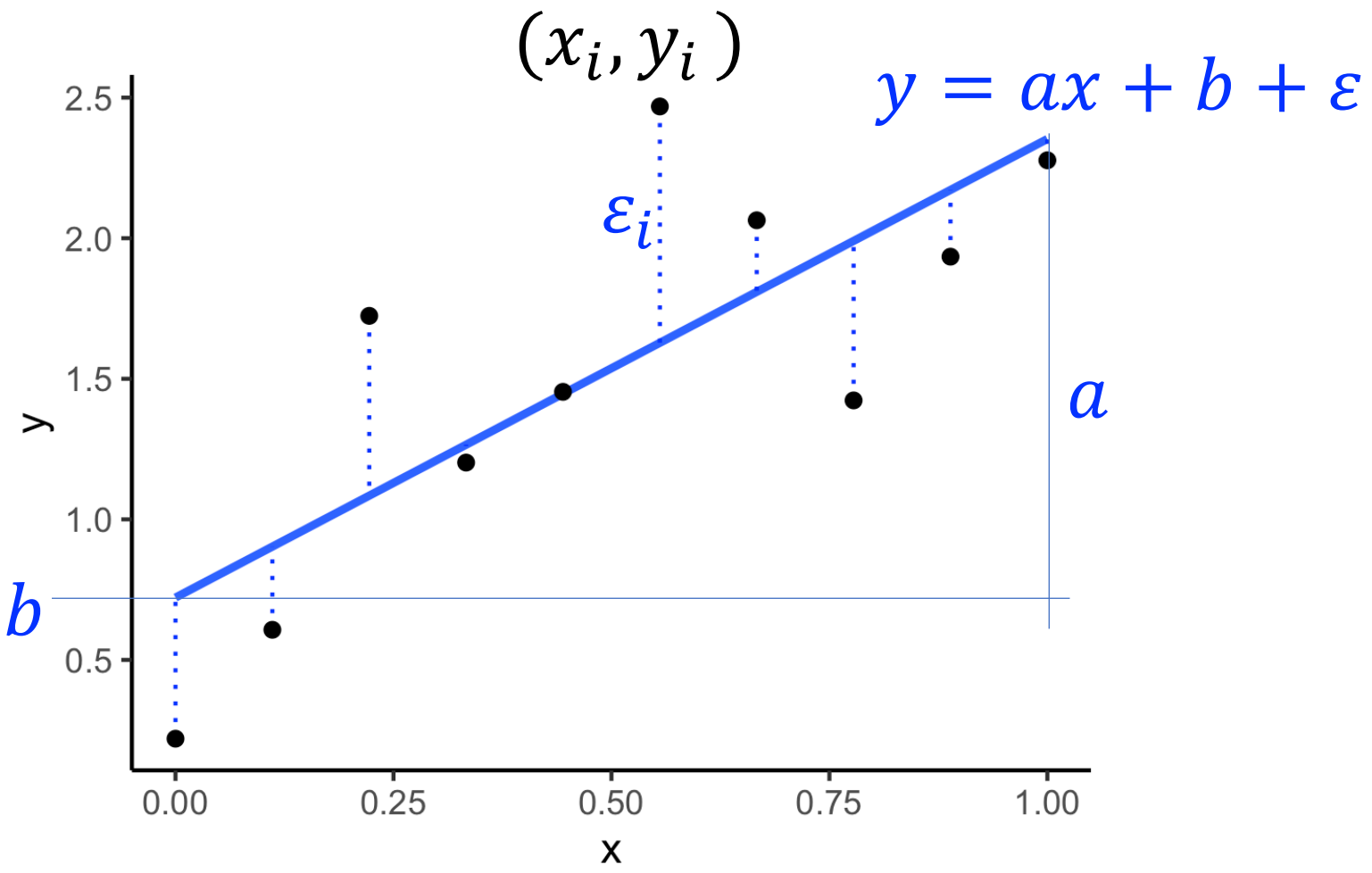
線形回帰 (Linear Regression)





線形回帰 (Linear Regression)

$$\mathrm{argmin}_{(a,b)} \sum_{i=1}^{n} \varepsilon^2 \quad \Rightarrow \quad \hat{a} = \frac{S_{xy}}{S_{xx}}, \quad \hat{b} = \bar{y} - \frac{S_{xy}}{S_{xx}} \bar{x}$$



線形回帰 (Linear Regression)

$$\mathrm{argmin}_{(a,b)} \sum_{i=1}^{n} \varepsilon^2 \quad \Rightarrow \quad \hat{a} = \frac{S_{xy}}{S_{xx}}, \quad \hat{b} = \bar{y} - \frac{S_{xy}}{S_{xx}} \bar{x}$$

$$S_{xx} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

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

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

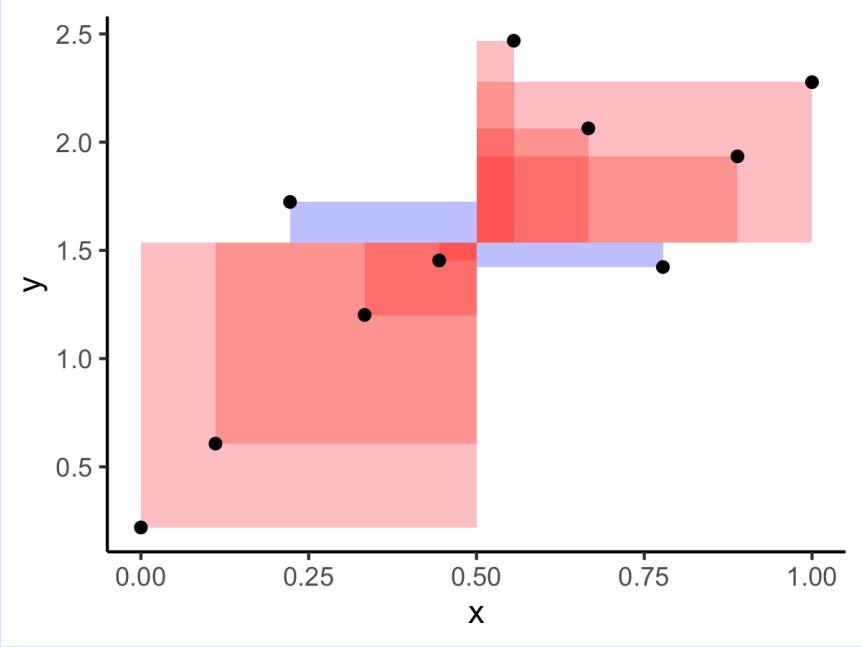
線形回帰 (Linear Regression)

$$\mathrm{argmin}_{(a,b)} \sum_{i=1}^{n} \varepsilon^2 \quad \Rightarrow \quad \hat{a} = \frac{S_{xy}}{S_{xx}}, \quad \hat{b} = \bar{y} - \frac{S_{xy}}{S_{xx}} \bar{x}$$

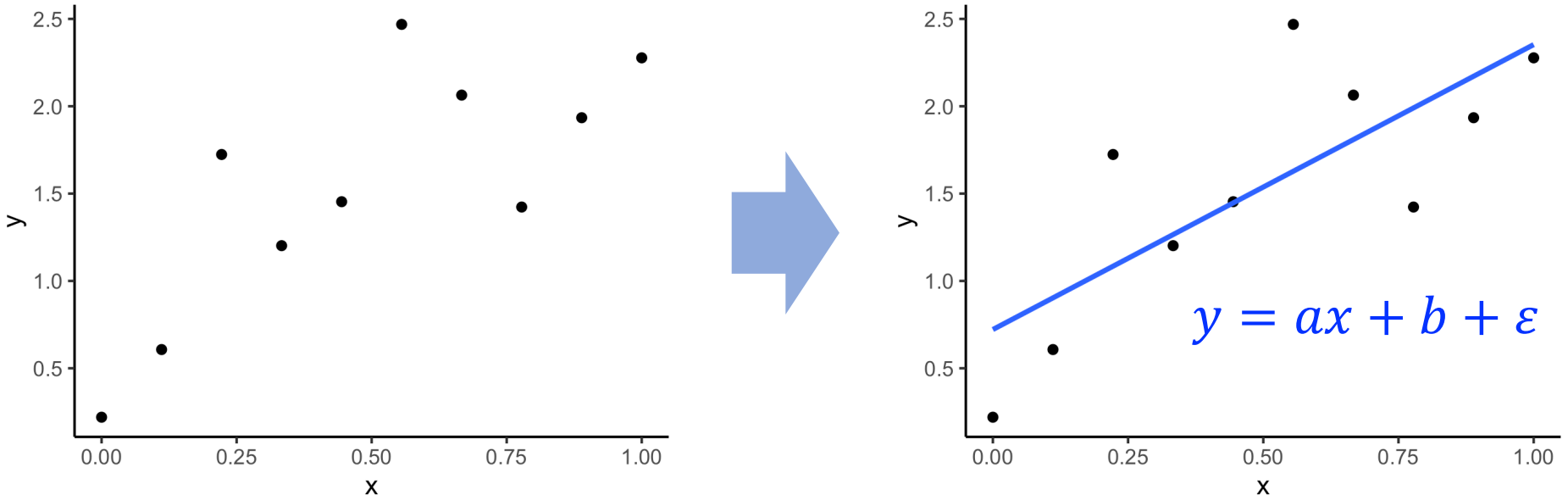
$$S_{xx} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

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

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$



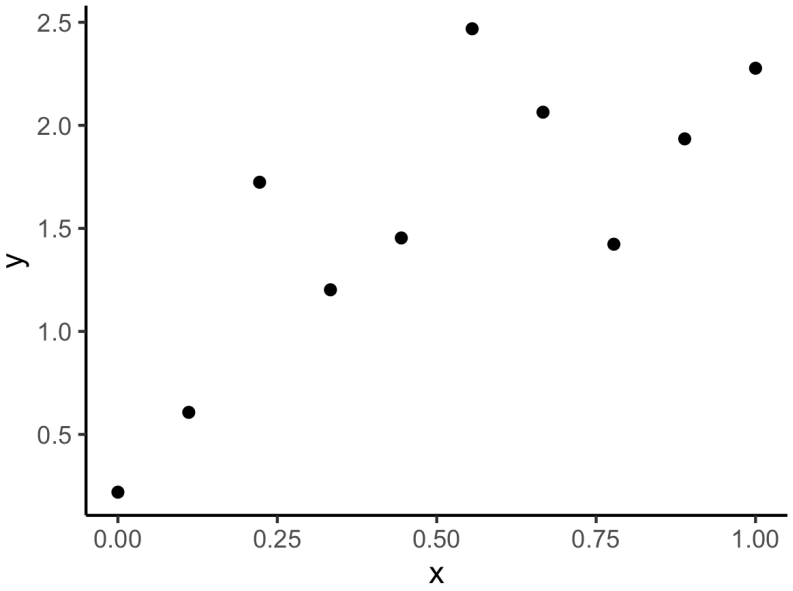
線形回帰 (Linear Regression)

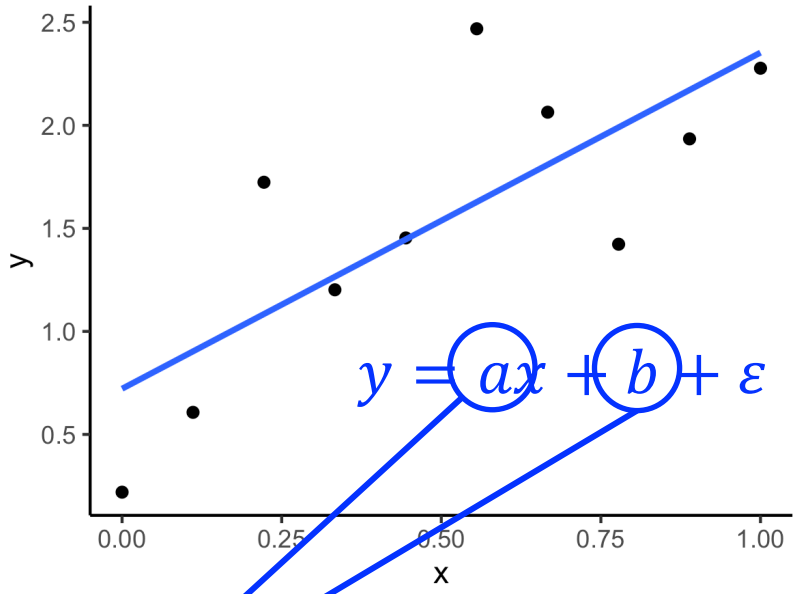



```
dat_lm <- dat %>% lm(y ~ x, data = .)
```

```
## lm(formula = y ~ x, data = .)
##
## Coefficients:
## (Intercept)            x
##      0.7217       1.6311
```

線形回帰 (Linear Regression)






```
dat_lm <- dat %>% lm(y ~ x, data = .)
```

```
## lm(formula = y ~ x, data = .)
##
## Coefficients:
## (Intercept)            x
##      0.7217       1.6311
```

線形回帰 (Linear Regression)

```
dat %>% lm(y ~ x, data = .) %>% summary()
```

```
##
## Call:
## lm(formula = y ~ x, data = .)
##
## Coefficients:
##             Estimate  Std.Error  t value   Pr(>|t|)
## (Intercept)  0.7217    0.2871    2.514    0.03613 *
## x            1.6311    0.4839    3.371    0.00978 **
## ---
## Residual standard error: 0.4884 on 8 degrees of freedom
## Multiple R-squared: 0.5868, Adjusted R-squared: 0.5351
## F-statistic: 11.36 on 1 and 8 DF, p-value: 0.009778
```

線形回帰モデル (Linear Regression Model)

回帰直線

$$\hat{y}_i = ax_i + b,$$
$$\varepsilon_i = y_i - \hat{y}_i$$

$$\xrightarrow{\operatorname{argmin}_{(a,b)} \sum_{i=1}^{n} \varepsilon^2}$$

$$\hat{a} = \frac{S_{xy}}{S_{xx}},$$
$$\hat{b} = \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x}$$

線形回帰モデル (Linear Regression Model)

回帰直線

$$\hat{y}_i = ax_i + b, \qquad \varepsilon_i = y_i - \hat{y}_i$$

$$\xrightarrow{\operatorname{argmin}_{(a,b)} \sum_{i=1}^{n} \varepsilon^2}$$

$$\hat{a} = \frac{S_{xy}}{S_{xx}},$$

$$\hat{b} = \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x}$$

線形回帰モデル

$$Y_i = \alpha X_i + \beta + u_i,$$

$$u_i \sim N(0, \sigma^2)$$

線形回帰モデル (Linear Regression Model)

回帰直線

$$\hat{y}_i = ax_i + b,$$
$$\varepsilon_i = y_i - \hat{y}_i$$

$$\xrightarrow{\operatorname{argmin}_{(a,b)} \sum_{i=1}^{n} \varepsilon^2}$$

$$\hat{a} = \frac{S_{xy}}{S_{xx}},$$
$$\hat{b} = \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x}$$

線形回帰モデル

$$Y_i = \alpha X_i + \beta + u_i,$$
$$u_i \sim N(0, \sigma^2)$$

$$E[\hat{a}] = \alpha,$$
$$E[\hat{b}] = \beta$$

線形回帰 (Linear Regression Model)

```
dat %>% lm(y ~ x, data = .) %>% summary()
```

```
##
## Call:
## lm(formula = y ~ x, data = .)
##
## Coefficients:
##             Estimate  Std.Error  t value   Pr(>|t|)
## (Intercept)  0.7217    0.2871    2.514    0.03613 *
## x            1.6311    0.4839    3.371    0.00978 **
## ---
## Residual standard error: 0.4884 on 8 degrees of freedom
## Multiple R-squared: 0.5868, Adjusted R-squared: 0.5351
## F-statistic: 11.36 on 1 and 8 DF, p-value: 0.009778
```

線形回帰モデル (Linear Regression Model)

決定係数R2と相関係数r

$$\mathrm{R2} := \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$
$$= \cdots$$
$$= \left(\frac{\mathrm{Cov}[x,y]}{\sigma_x \sigma_y} \right)^2$$
$$= r^2$$

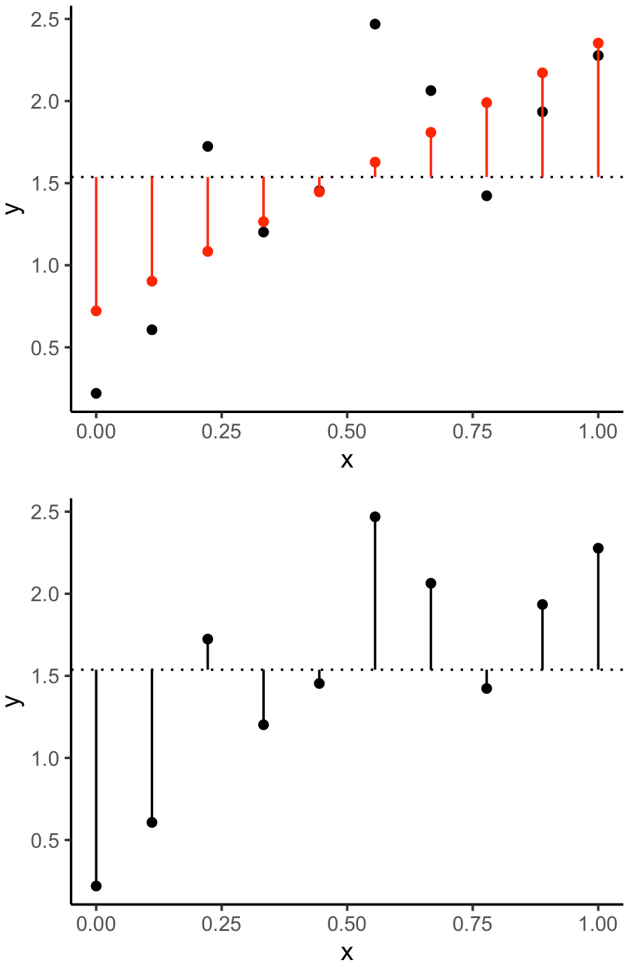
$$\mathrm{Cov}[x, y] := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

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

線形回帰モデル (Linear Regression Model)

決定係数R2と相関係数r

$$\mathrm{R2} := \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$
$$= \cdots$$
$$= \left(\frac{\mathrm{Cov}[x,y]}{\sigma_x \sigma_y}\right)^2$$
$$= r^2$$



$$\mathrm{Cov}[x,y] := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

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

線形回帰 (Linear Regression Model)

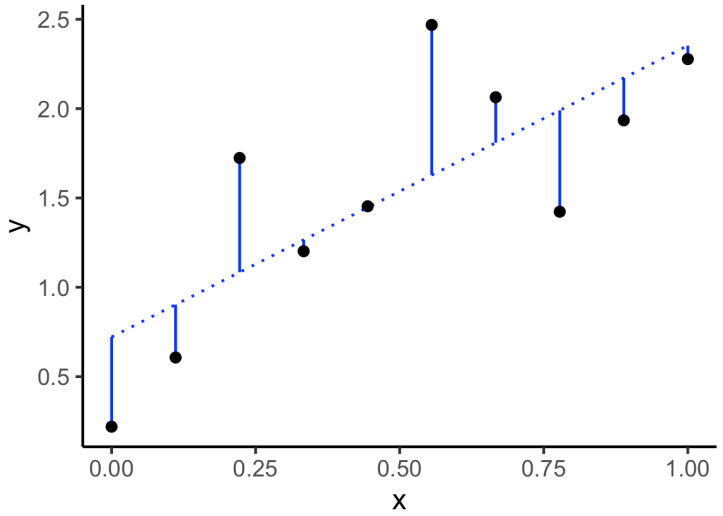
```
dat %>% lm(y ~ x, data = .) %>% summary()
```

```
##
## Call:
## lm(formula = y ~ x, data = .)
##
## Coefficients:
##             Estimate  Std.Error  t value   Pr(>|t|)
## (Intercept)  0.7217    0.2871    2.514    0.03613 *
## x            1.6311    0.4839    3.371    0.00978 **
## ---
## Residual standard error: 0.4884 on 8 degrees of freedom
## Multiple R-squared: 0.5868, Adjusted R-squared: 0.5351
## F-statistic: 11.36 on 1 and 8 DF, p-value: 0.009778
```

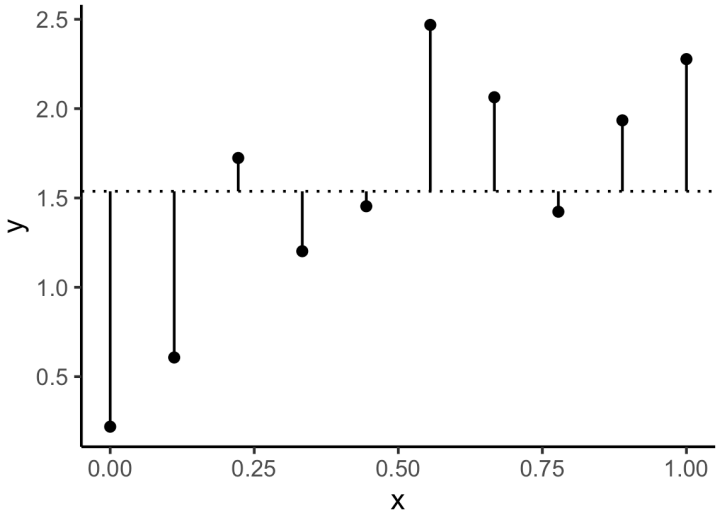
線形回帰モデル (Linear Regression Model)

F検定

$H_1: a = \hat{a}$



$H_0: a = 0$



残差平方和 $SSR_1 = \sum_i (y_i - \hat{y})^2$ $SSR_0 = \sum_i (y_i - \bar{y})^2$

$$\frac{SSR_0 - SSR_1}{SSR_1}$$

が十分にゼロから離れているかを検討する。

線形回帰モデル (Linear Regression Model)

F検定

$$x \sim N(0,1) \Rightarrow \sum_{i=1}^{n} x^2 \sim \chi^2(n)$$

(標準)正規分布 → カイ二乗分布 → F分布

$$N(0,1) \qquad \chi^2(n) \qquad F(n_1, n_2)$$

$$\left.\begin{array}{l} x_1 \sim \chi^2(n_1) \\ x_2 \sim \chi^2(n_2) \end{array}\right. \Rightarrow \frac{x_1/n_1}{x_2/n_2} \sim F(n_1, n_2)$$

線形回帰モデル (Linear Regression Model)

F検定

$$x \sim N(0,1) \Rightarrow \sum_{i=1}^{n} x^2 \sim \chi^2(n)$$

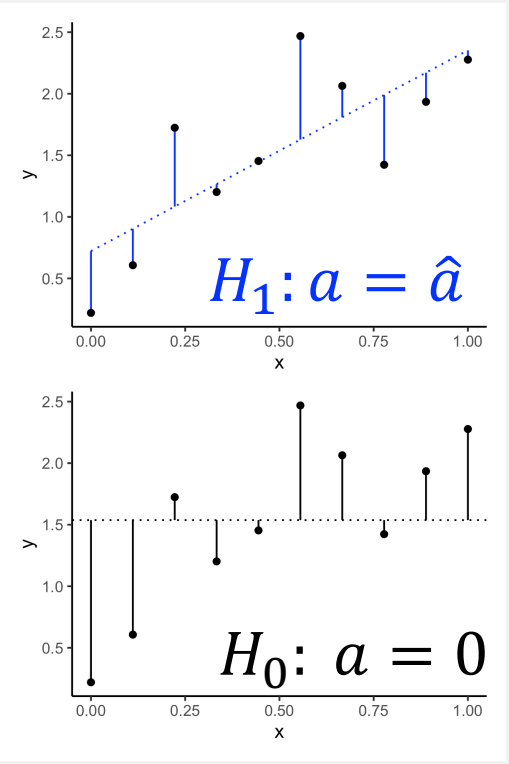
$$SSR_1 = \sum_i (y_i - \hat{y})^2$$

$$\sim \chi^2(n-k-1)$$

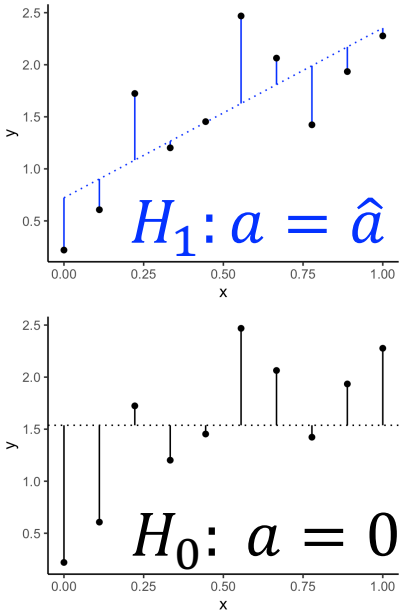
$$SSR_0 - SSR_1 = \sum_i (y_i - \bar{y})^2 - \sum_i (y_i - \hat{y})^2$$

$$\sim \chi^2(k)$$

k , number of estimated parameters in the model



線形回帰モデル (Linear Regression Model)



F検定

$$SSR_0 - SSR_1 \sim \chi^2(k)$$

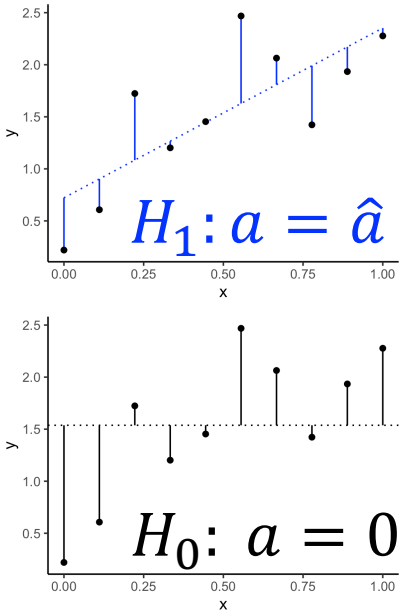
$$SSR_1 \sim \chi^2(n-k-1)$$

(標準)正規分布 → カイ二乗分布 → F分布

$$\chi^2(n) \qquad F(n_1, n_2)$$

$$\begin{matrix} x_1 \sim \chi^2(n_1) \\ x_2 \sim \chi^2(n_2) \end{matrix} \Rightarrow \frac{x_1/n_1}{x_2/n_2} \sim F(n_1, n_2)$$

線形回帰モデル (Linear Regression Model)



F検定

$$SSR_0 - SSR_1 \sim \chi^2(k)$$

$$SSR_1 \sim \chi^2(n-k-1)$$

(標準)正規分布 → カイ二乗分布 → F分布

$$\chi^2(n) \qquad F(n_1, n_2)$$

$$\frac{(SSR_0 - SSR_1)/k}{SSR_1/(n-k-1)} \sim F(k, n-k-1)$$

f-statistic

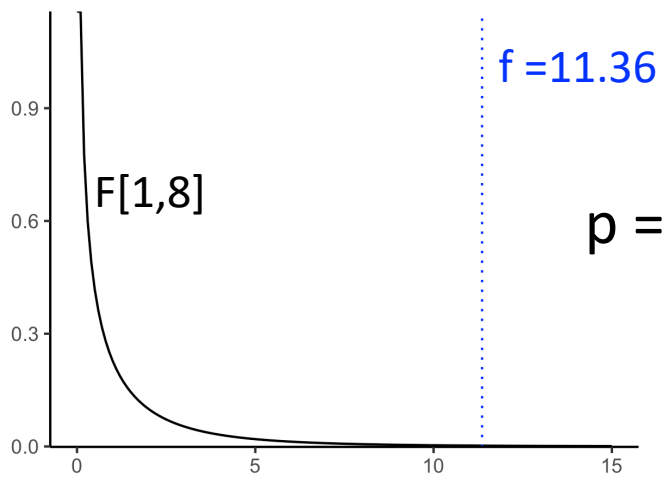
線形回帰 (Linear Regression Model)

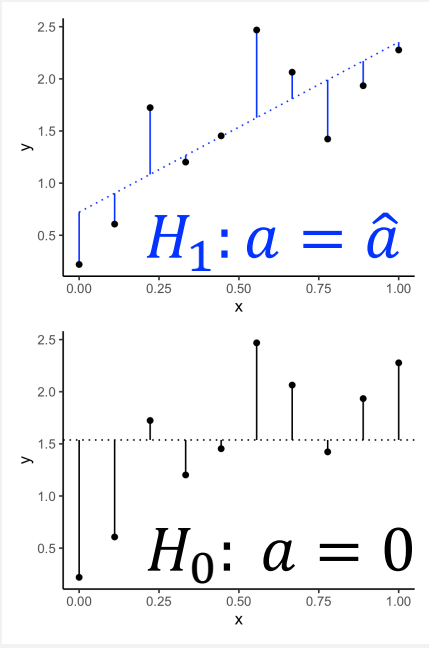
```
dat %>% lm(y ~ x, data = .) %>% summary()
```

F-statistic: 11.36 on 1 and 8 DF, p-value: 0.009778

f-statistic

$$\frac{(SSR_0 - SSR_1)/k}{SSR_1/(n-k-1)} \sim F(k, n-k-1)$$



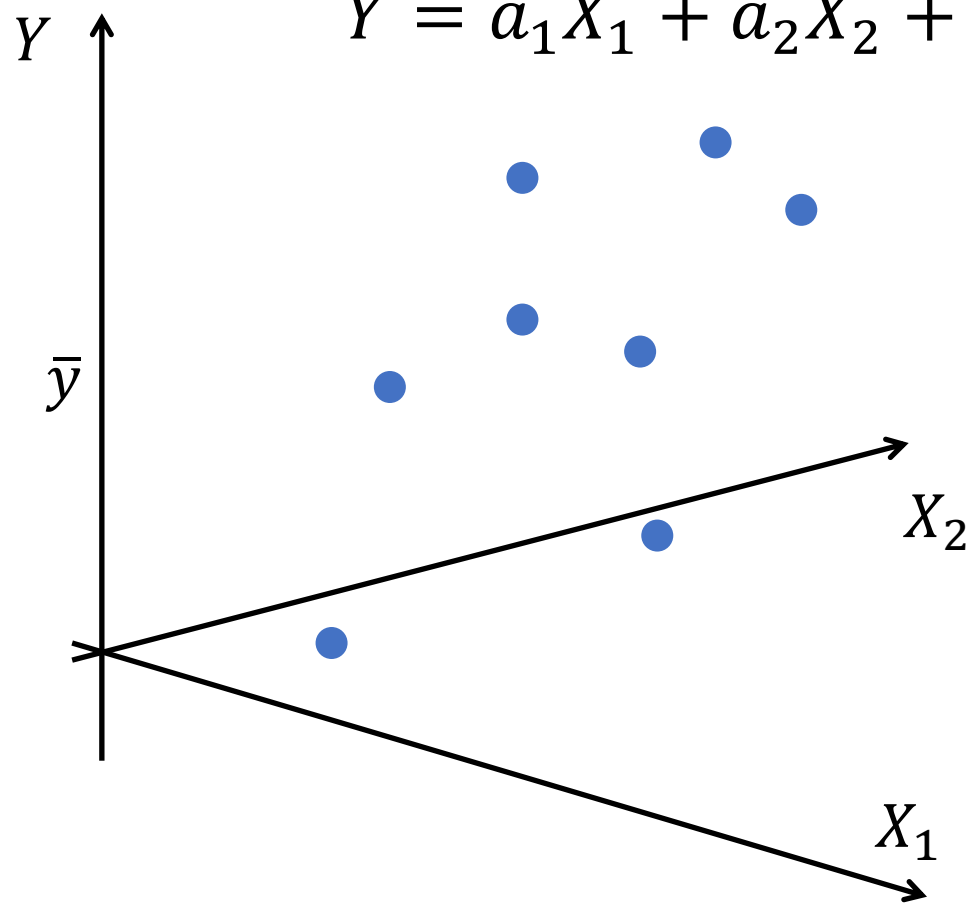


線形回帰 (Linear Regression Model)

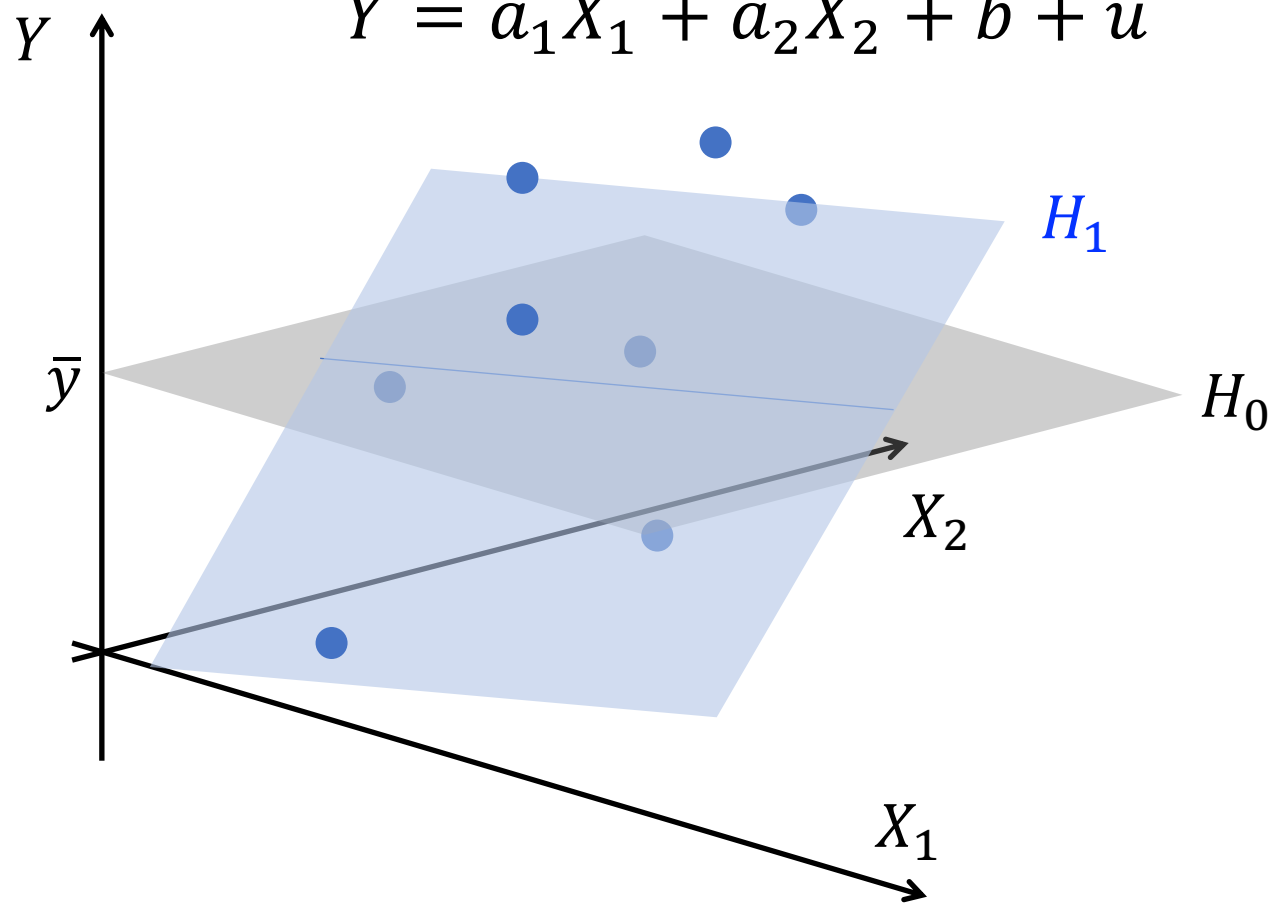
```
dat %>% lm(y ~ x, data = .) %>% summary()
```

```
##
## Call:
## lm(formula = y ~ x, data = .)
##
## Coefficients:
##             Estimate  Std.Error  t value   Pr(>|t|)
## (Intercept)  0.7217    0.2871    2.514    0.03613 *
## x            1.6311    0.4839    3.371    0.00978 **
## ---
## Residual standard error: 0.4884 on 8 degrees of freedom
## Multiple R-squared: 0.5868, Adjusted R-squared: 0.5351
## F-statistic: 11.36 on 1 and 8 DF, p-value: 0.009778
```

$$Y = a_1 X_1 + a_2 X_2 + b + u$$



$$Y = a_1X_1 + a_2X_2 + b + u$$



線形回帰 (Linear Regression Model)

```
dat_lm <-
  dat %>% lm(y ~ x, data = .)
```

```
extract_rsq <- function(lm_model){
  lm_model %>% .$r.squaread
}

dat_lm %>% extract_rsq()
```

```
## [1] 0.5867894
```

線形回帰 (Linear Regression Model)

```
dat_lm <-
  dat %>% lm(y ~ x, data = .)
```

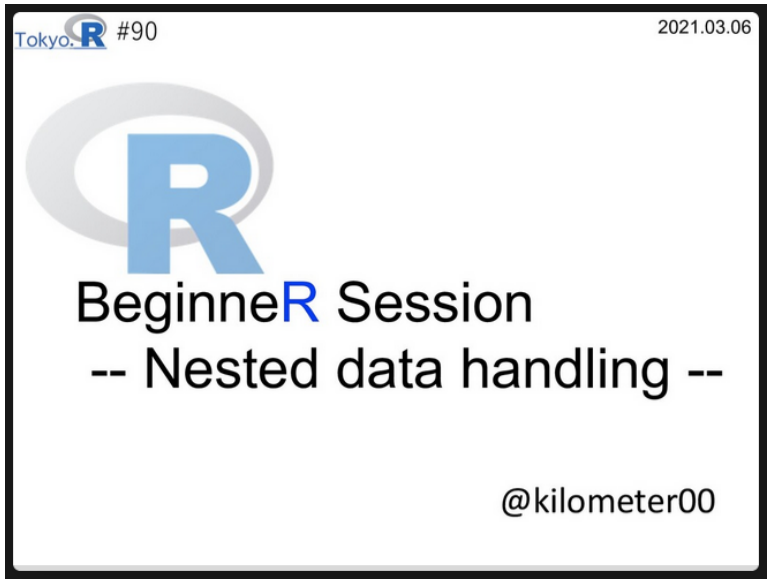
```
extract_p <- function(lm_model){
  f <-
    lm_model %>%
    summary() %>%
    .$fstatistic

  pf(f[1], f[2], f[3], lower.tail = F)
}
dat_lm %>% extract_p()
```

```
##         value
## 0.009777651
```

線形回帰 (Linear Regression Model)

```
dat_lm_nest <-
  dat %>%
  group_nest() %>%
  mutate(lm = map(data,
                  ~ lm(y ~ x, data = .)))
```

```
# A tibble: 1 x 2
                data          lm
                <list>        <list>
1 <df[,2] [10 × 2]>      <lm>
```



線形回帰 (Linear Regression Model)

```
dat_lm_nest <-
  dat %>%
  group_nest() %>%
  mutate(lm = map(data, ~ lm(y ~ x, data = .)))

dat_lm_nest %>%
  mutate(a = map_dbl(lm, ~ .$coefficients[2]),
         b = map_dbl(lm, ~ .$coefficients[1]),
         rsq = map_dbl(lm, extract_rsq),
         pval = map_dbl(lm, extract_p))
```

```
# A tibble: 1 x 6
          data      lm       a       b     rsq      pval
  <list>          <lis>   <dbl>   <dbl>   <dbl>    <dbl>
1 <df[,2] [10 …    <lm>     1.63  0.722  0.587  0.00978
```

線形回帰分析

- 回帰直線(最小二乗法)
- 誤差の確率モデル
- 決定係数と相関係数
- 回帰モデルの仮説検定

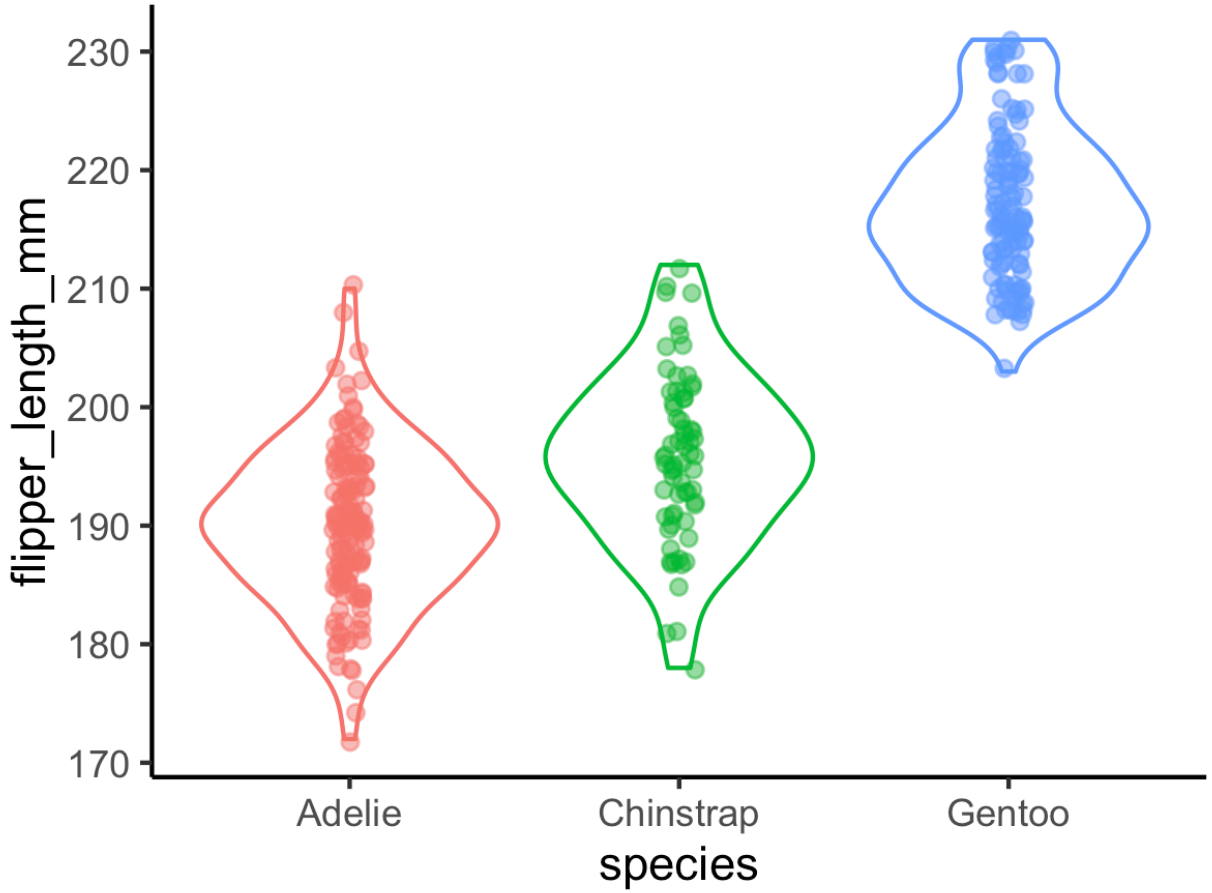
分散分析

- 回帰モデルとの接続性
- One-way ANOVA
- Two-way ANOVA
- Tukey HSD Test

```
library(palmerpenguins)
```

```
penguins
```

```
# A tibble: 344 x 8
   species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g sex     year
   <fct>   <fct>              <dbl>         <dbl>             <int>       <int> <fct>  <int>
 1 Adelie  Torgersen           39.1          18.7               181        3750 male    2007
 2 Adelie  Torgersen           39.5          17.4               186        3800 female  2007
 3 Adelie  Torgersen           40.3          18                 195        3250 female  2007
 4 Adelie  Torgersen           NA            NA                  NA          NA NA      2007
 5 Adelie  Torgersen           36.7          19.3               193        3450 female  2007
 6 Adelie  Torgersen           39.3          20.6               190        3650 male    2007
 7 Adelie  Torgersen           38.9          17.8               181        3625 female  2007
 8 Adelie  Torgersen           39.2          19.6               195        4675 male    2007
 9 Adelie  Torgersen           34.1          18.1               193        3475 NA      2007
10 Adelie  Torgersen           42            20.2               190        4250 NA      2007
# … with 334 more rows
```




```
penguins %>%
  ggplot() +
  aes(species, flipper_length_mm, color = species)+
  geom_violin()+
  geom_jitter(alpha = 0.5, width = 0.05)+
  theme(legend.position = "none")
```

分散分析（ANOVA）

```
penguins_aov <-
  penguins %>%
  aov(flipper_length_mm ~ species, data = .)
```

```
penguins_aov <-
  penguins %>%
  lm(flipper_length_mm ~ species, data = .) %>%
  aov()
```

分散分析（ANOVA）

```
penguins_aov <-
  penguins %>%
  lm(flipper_length_mm ~ species, data = .) %>%
  aov()

penguins_aov %>%
  summary()
```

```
             Df  Sum Sq  Mean Sq  F value  Pr(>F)
species       2   52473    26237    594.8  <2e-16
Residuals   339   14953       44

species    ***
Residuals
```

分散分析（ANOVA）

```
penguins %>%
  lm(flipper_length_mm ~ species, data = .) %>%
  summary()
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	189.9536	0.5405	351.454	< 2e-16	***
speciesChinstrap	5.8699	0.9699	6.052	3.79e-09	***
speciesGentoo	27.2333	0.8067	33.760	< 2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.642 on 339 degrees of freedom
(2 observations deleted due to missingness)
Multiple R-squared: 0.7782, Adjusted R-squared: 0.7769
F-statistic: 594.8 on 2 and 339 DF, p-value: < 2.2e-16

分散分析（ANOVA）

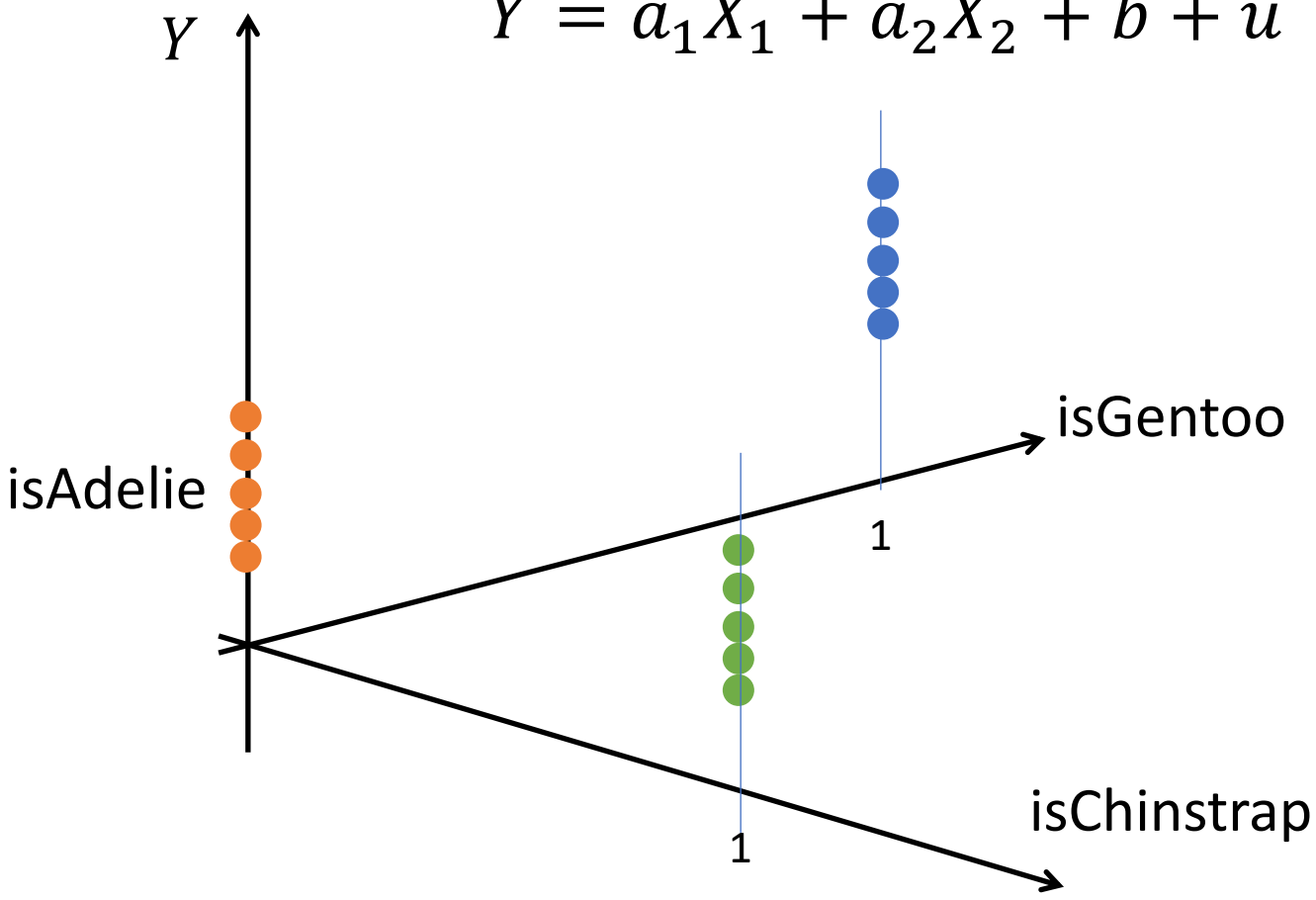
```
penguins_species <-
  penguins %>%
  select(flipper_length_mm, species) %>%
  mutate(isAdelie = if_else(species == "Adelie", 1, 0),
         isChinstrap = if_else(species == "Chinstrap", 1, 0),
         isGentoo = if_else(species == "Gentoo", 1, 0))

penguins_species %>%
  lm(flipper_length_mm ~ isChinstrap + isGentoo, data = .)
```

Call:
lm(formula = flipper_length_mm ~ isGentoo + isChinstrap, data = .)

Coefficients:

(Intercept)	isChinstrap	isGentoo
189.95	5.87	27.23

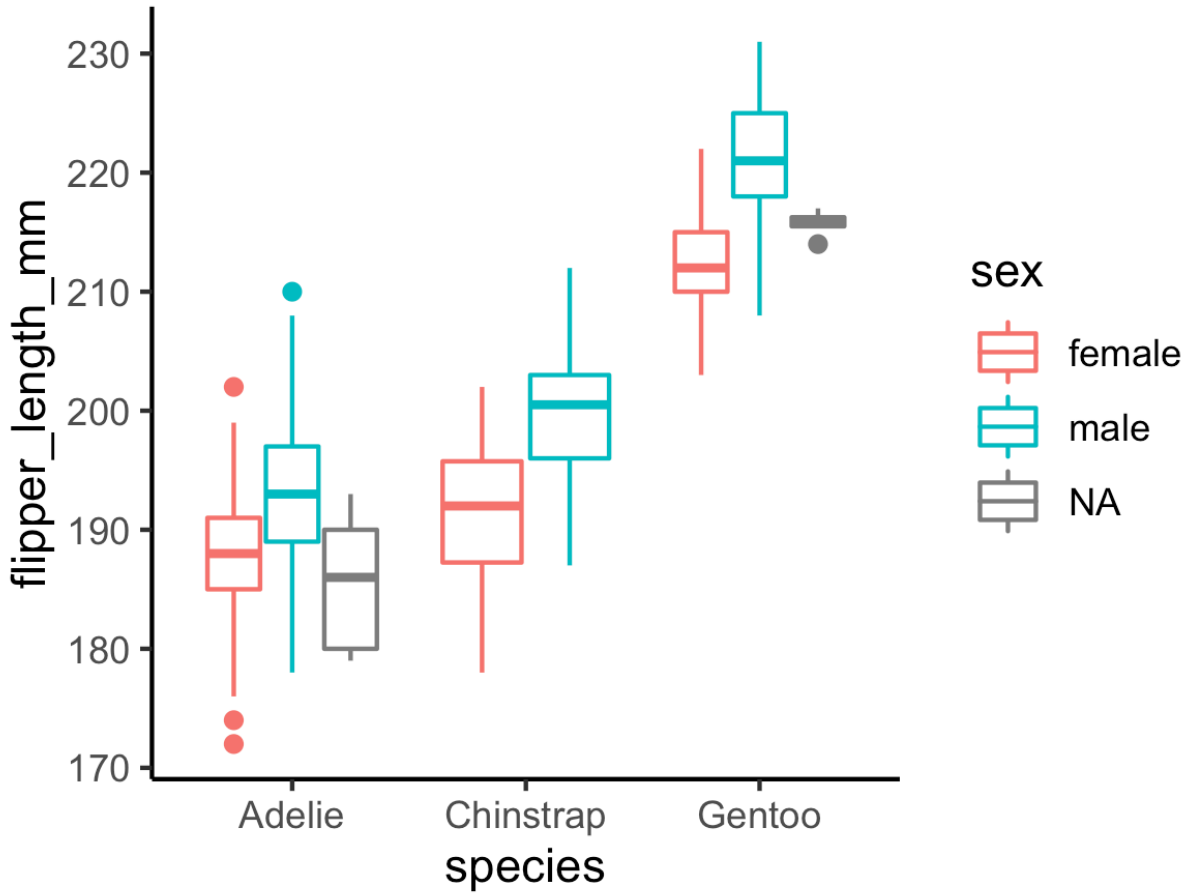
$$Y = a_1 X_1 + a_2 X_2 + b + u$$



分散分析（ANOVA）

```
penguins_aov <-
  penguins %>%
  lm(flipper_length_mm ~ species, data = .) %>%
  aov()

penguins_aov %>%
  summary()
```

```
             Df  Sum Sq  Mean Sq  F value  Pr(>F)
species       2   52473    26237    594.8  <2e-16
Residuals   339   14953       44

species    ***
Residuals
```




```
penguins %>%
  ggplot() +
  aes(species, flipper_length_mm, color = sex)+
  geom_boxplot()
```

分散分析（Two-way ANOVA）

```
penguins %>%
  lm(flipper_length_mm ~
     species + sex + species * sex,
     data = .) %>%
  aov() %>%
  summary()
```

```
              Df   Sum Sq Mean Sq F value  Pr(>F)
species        2    50526   25263 789.912 < 2e-16 ***
sex            1     3906    3906 122.119 < 2e-16 ***
species:sex    2      329     165   5.144 0.00631 **
Residuals    327    10458      32
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
11 observations deleted due to missingness
```

Tukey HSD test

```
penguins %>%
  lm(flipper_length_mm ~
     species + sex + species * sex,
     data = .) %>%
  aov() %>%
  TukeyHSD()
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = .)

$species
                      diff       lwr       upr p adj
Chinstrap-Adelie  5.72079  3.76593  7.675649     0
Gentoo-Adelie    27.13255 25.48814 28.776974     0
Gentoo-Chinstrap 21.41176 19.38766 23.435867     0
...
```

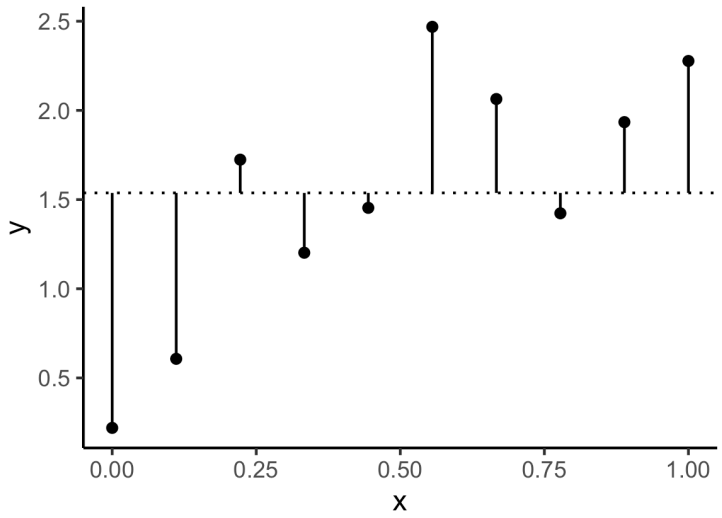
線形回帰分析

- 回帰直線(最小二乗法)
- 誤差の確率モデル
- 決定係数と相関係数
- 回帰モデルの仮説検定
- **ブートストラップ法**

分散分析

- 回帰モデルとの接続性
- One-way ANOVA
- Two-way ANOVA
- Tukey HSD Test

H_0: $a = 0$ のもとで



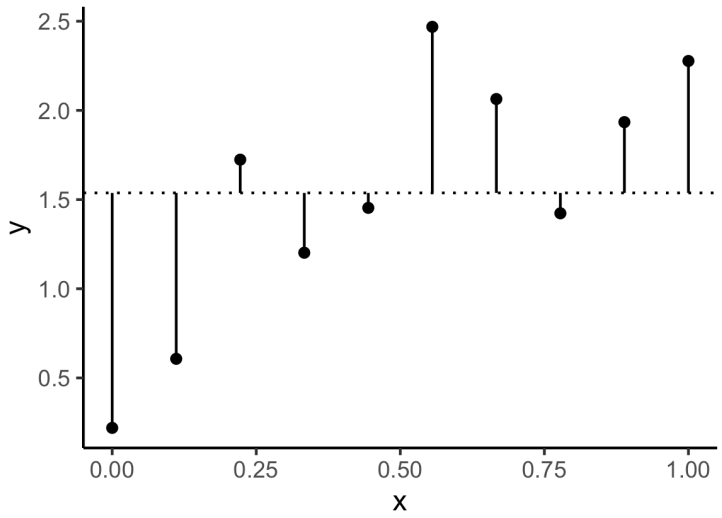
データ (x, y)

推定

$y \sim N(\bar{y}, \sigma^2)$

母集団 (X, Y)

H_0: $a = 0$ のもとで



データ (x, y)

推定

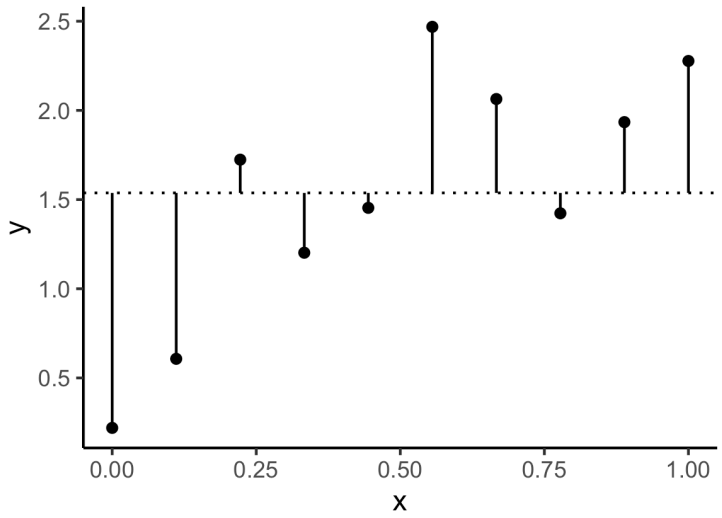
$y \sim N(\bar{y}, \sigma^2)$

母集団 (X, Y)

標本 (x', y')

回帰パラメータ a', b'

H_0: $a = 0$ のもとで



データ (x, y)

推定

$y \sim N(\bar{y}, \sigma^2)$

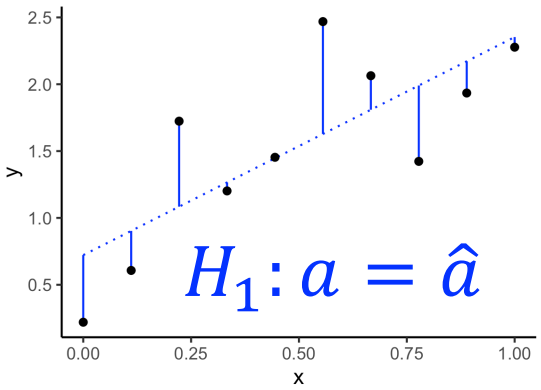
母集団 (X, Y)

標本 (x', y')

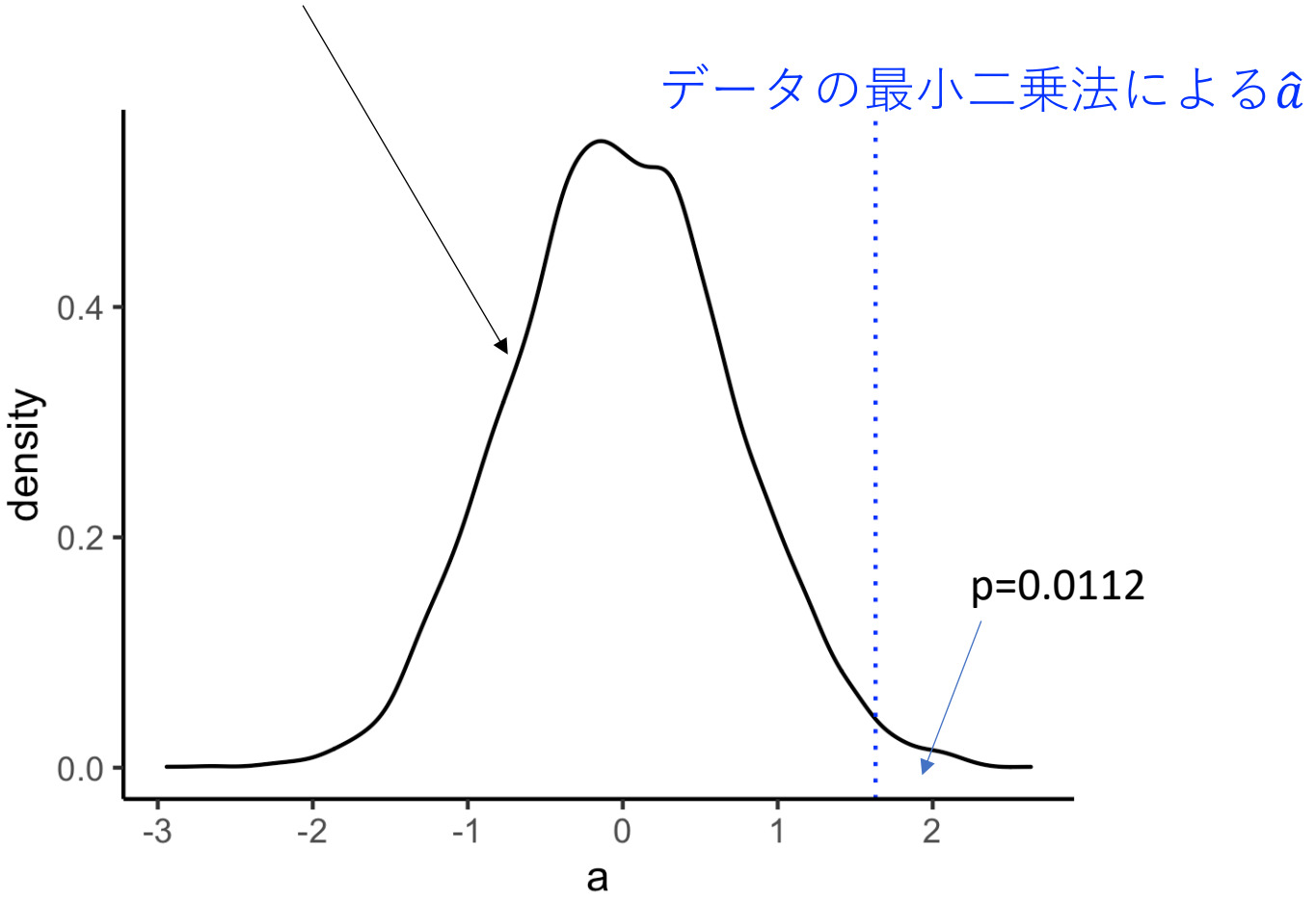
回帰パラメータ a', b'

リサンプリング

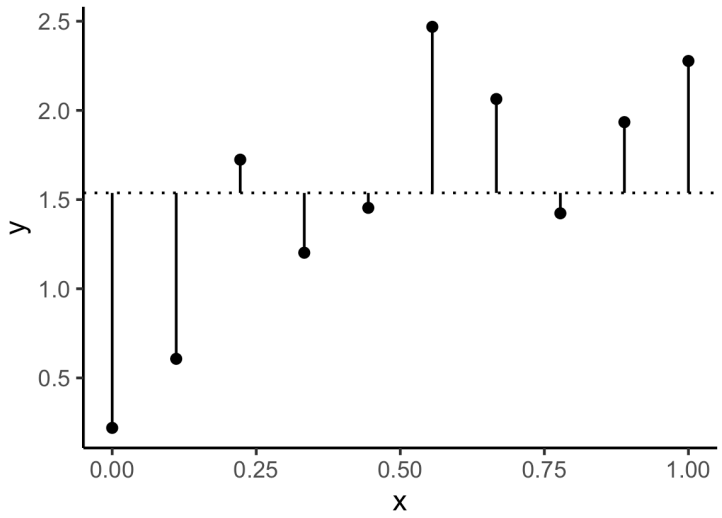
検証



H_0: $a = 0$ のもとで5000回リサンプルされたa'の確率密度関数



H_0: $a = 0$ のもとで



データ (x, y)

推定

$y \sim N(\bar{y}, \sigma^2)$

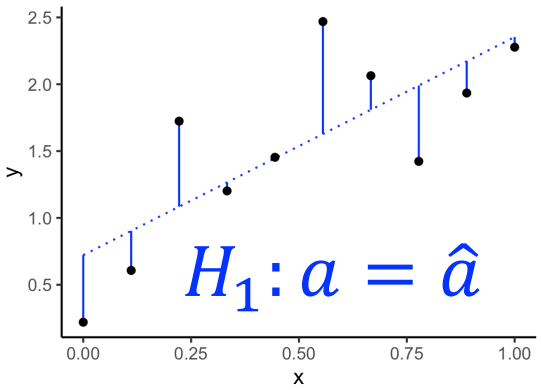
母集団 (X, Y)

標本 (x', y')

回帰パラメータ a', b'

リサンプリング

検証



線形回帰分析

- 回帰直線(最小二乗法)
- 誤差の確率モデル
- 決定係数と相関係数
- 回帰モデルの仮説検定
- ブートストラップ法

分散分析

- 回帰モデルとの接続性
- One-way ANOVA
- Two-way ANOVA
- Tukey HSD Test


Enjoy!!