回帰分析（単回帰モデル）

東京国際大学 データサイエンス教育研究所 竹田 恒

2022-08-06

データ

```
x <- 1:30
n <- length(x)
b0 <- 20
b1 <- 1.2
set.seed(2)
e <- rnorm(n, mean = 0, sd = 5)
y <- b0 + b1 * x + e
ybar <- mean(y)

# Test data
#x <- c(1, 2, 3, 4, 5)
#y <- c(2, 2, 4, 3, 5)

d <- data.frame(x, y)
str(d)
```

```
## 'data.frame':    30 obs. of  2 variables:
##  $ x: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ y: num  16.7 23.3 31.5 19.1 25.6 ...
```

要約統計量

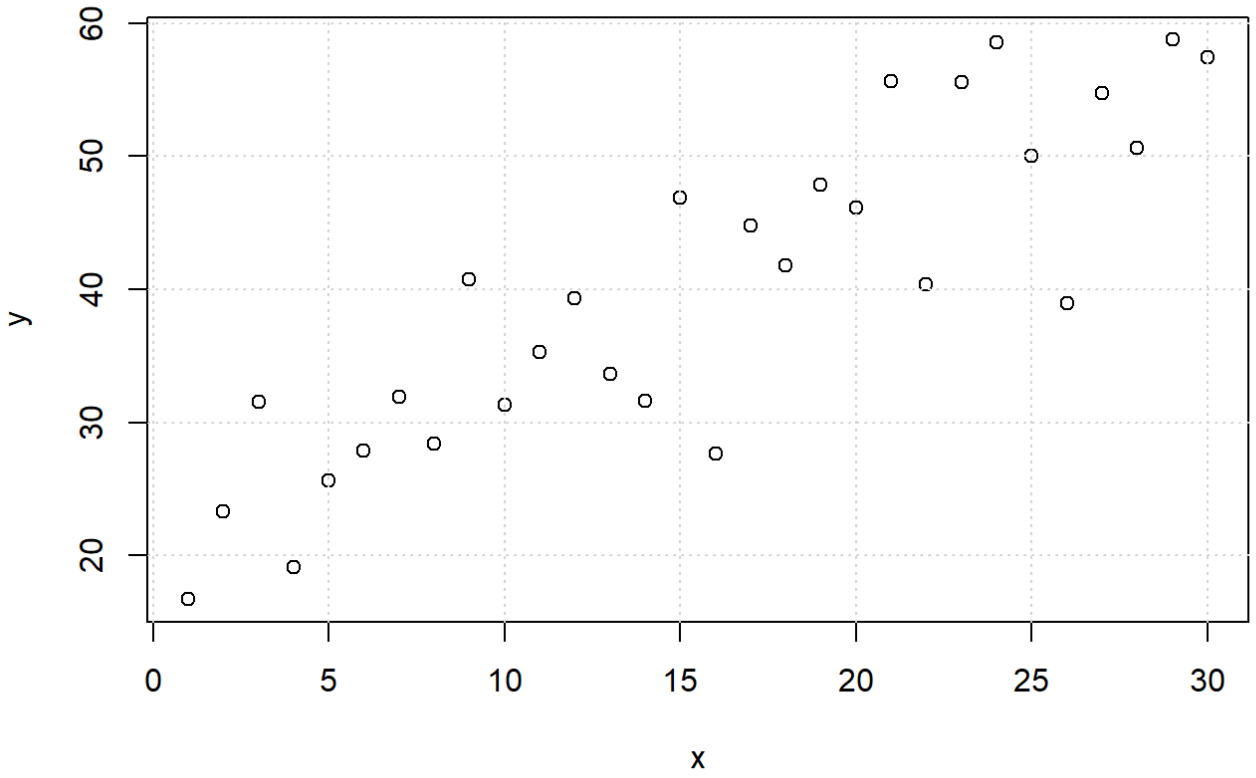
```
summary(d)
```

```
##        x               y        
##  Min.   : 1.00   Min.   :16.72  
##  1st Qu.: 8.25   1st Qu.:31.36  
##  Median :15.50   Median :39.85  
##  Mean   :15.50   Mean   :39.74  
##  3rd Qu.:22.75   3rd Qu.:49.48  
##  Max.   :30.00   Max.   :58.76
```

```
# カラーパレット
# (https://www.rapidtables.com/web/color/RGB_Color.html)
COL <- c(rgb(255,   0,   0,  255, max = 255), # 赤
         rgb(  0,   0, 255,  255, max = 255), # 青
         rgb(  0, 155,   0,  255, max = 255)) # 緑
```

散布図

```
matplot(x, y, pch = 1)
grid()
```



1. 切片なし単回帰モデル（Model1）

$$y = \beta x + e$$
$$e \sim \mathbb{N}(0, \sigma^2)$$

1.1 計算式による回帰係数の推定

```
options(digits = 3)

b <- sum(x * y) / sum(x * x)
```

$$\hat{\beta} = 2.244$$

切片なし単回帰モデルのMultiple R_0^2（第2項の分母で平均値が引かれていない）

$$R_0^2 = \frac{\sum_i \hat{y_i}^2}{\sum_i y_i^2} = 1 - \frac{\sum_i (y_i - \hat{y_i})^2}{\sum_i y_i^2}$$

```
yhat <- b * x
RSS = sum((y - yhat)^2)
TSS = sum(y^2)
R2 = 1 - RSS/TSS
```

$$R_0^2 = 0.921$$

観測値と推定値の相関係数の二乗値とは異なる値となる.

```
cor(y, yhat)^2
```

```
## [1] 0.77
```

1.2 Rによる回帰係数の推定

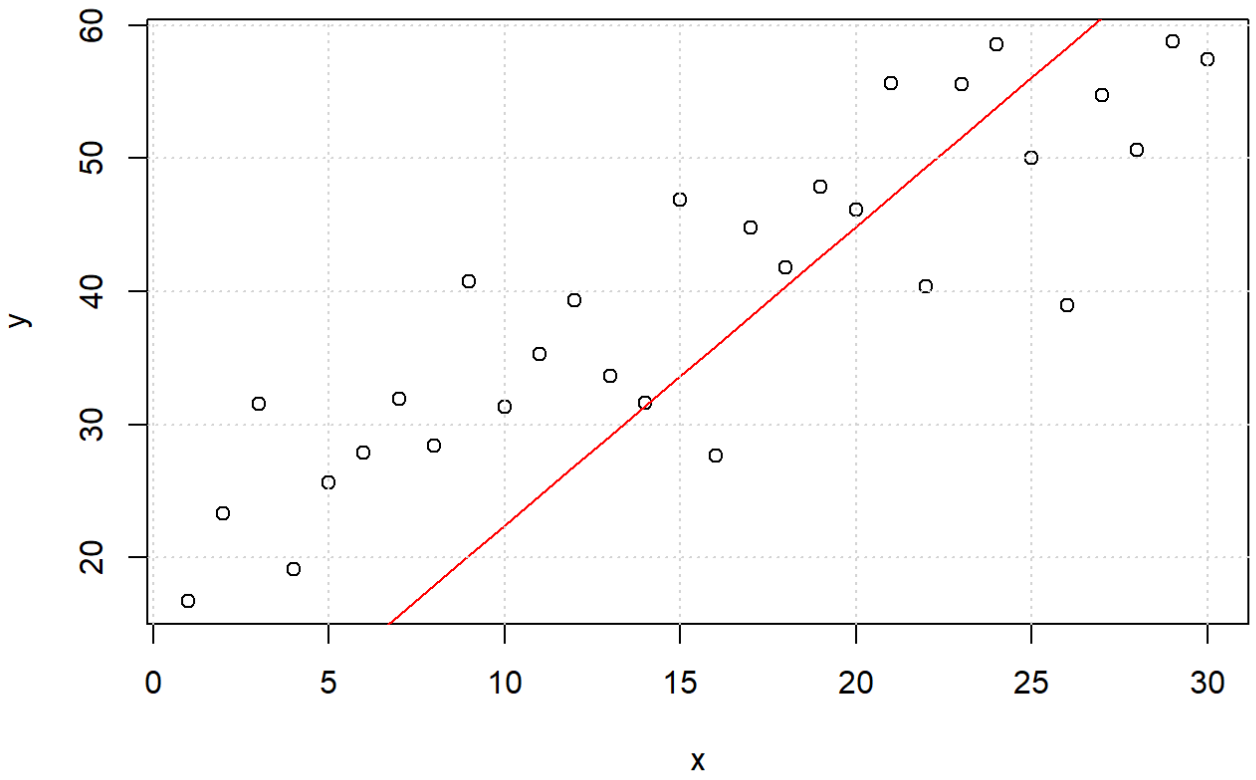
```
fit1 <- lm(y ~ x - 1, data = d)
sum(yhat - fit1$fitted)
```

```
## [1] -1.6e-14
```

```
summary(fit1)
```

```
##
## Call:
## lm(formula = y ~ x - 1, data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -19.41  -4.31   5.93  13.03  24.81
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## x    2.244      0.122    18.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.9 on 29 degrees of freedom
## Multiple R-squared:  0.921,  Adjusted R-squared:  0.918
## F-statistic:  337 on 1 and 29 DF,  p-value: <2e-16
```

```
matplot(x, y, pch = 1)
matlines(x, fit1$fitted, col = COL[1])
grid()
```



2. 切片あり単回帰モデル（Model2）

$$y = \beta_0 + \beta_1 x + e$$
$$e \sim \mathbb{N}(0, \sigma^2)$$

2.1 計算式による回帰係数の推定

```
n <- nrow(d)
xbar <- mean(x)
ybar <- mean(y)
b1 <- (sum(x * y) - n * xbar * ybar) / (sum(x * x) - n * xbar^2)
b0 <- ybar - b1 * xbar
```

$$\hat{\beta}_0 = 20.853$$
$$\hat{\beta}_1 = 1.219$$

切片あり単回帰モデルのMultiple R^2（一般的な定義）

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

```
yhat <- b0 + b1 * x
RSS = sum((y - yhat)^2)
TSS = sum((y - mean(y))^2)
R2 = 1 - RSS/TSS
```

$$R^2 = 0.77$$

観測値と推定値の相関係数の二乗値と同じ値となる.

```
cor(y, yhat)^2
```

```
## [1] 0.77
```

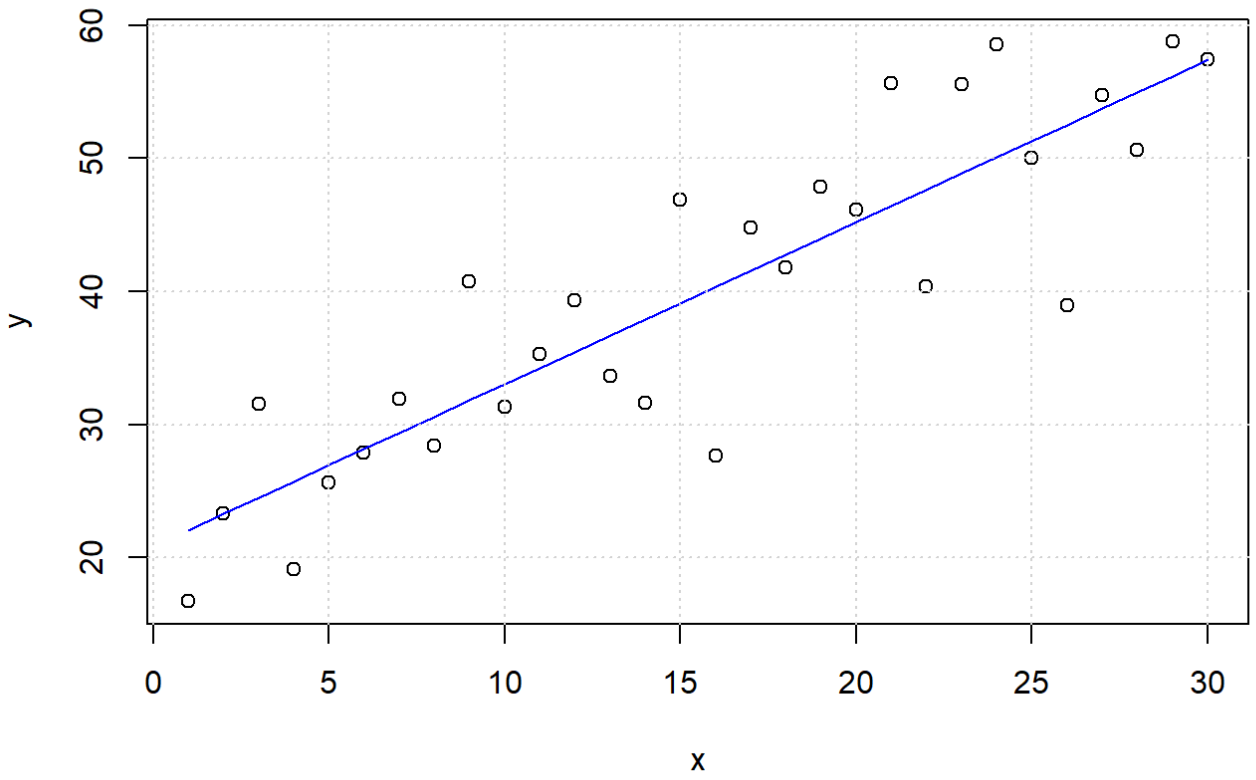
2.2 Rによる回帰係数の推定

```
fit2 <- lm(y ~ x, data = d)

summary(fit2)
```

```
##
## Call:
## lm(formula = y ~ x, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.599  -2.845   0.033   3.679   9.208
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   20.853      2.236    9.33  4.4e-10 ***
## x              1.219      0.126    9.68  2.0e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.97 on 28 degrees of freedom
## Multiple R-squared:  0.77,   Adjusted R-squared:  0.762
## F-statistic: 93.7 on 1 and 28 DF,  p-value: 1.97e-10
```

```
matplot(x, y, pch = 1)
matlines(x, fit2$fitted, col = COL[2])
grid()
```



3 回帰モデルの比較（Model1 vs Model2）

```
library(sjPlot)
library(sjmisc)
library(sjlabelled)
tab_model(fit1, fit2, dv.labels = c("Model1", "Model2"))
```

	Model1			Model2		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
x	2.24	1.99 – 2.49	**<0.001**	1.22	0.96 – 1.48	**<0.001**
(Intercept)				20.85	16.27 – 25.43	**<0.001**
Observations	30			30		
R^2 / R^2 adjusted	0.921 / 0.918			0.770 / 0.762		

【注意】決定係数Multiple R^2をみると，Model1の説明力がModel2より高くなっている．切片の有無により決定係数の計算方法が異なるため，この指標では切片の有無によるモデル性能の違いを評価できない．
Removal of statistically significant intercept term increases R2 in linear model (https://stats.stackexchange.com/questions/26176/removal-of-statistically-significant-intercept-term-increases-r2-in-linear-mo)

単回帰のときは，決定係数のような平均値からの変動を評価する指標ではなくて平均二乗誤差平方RMSE(root mean squared error)や残差標準誤差RSE（residual standard error）など残差から評価する指標を代わりに使おう．

```
(RMSE1 <- sqrt(mean((y - fit1$fitted)^2)))
```

```
## [1] 11.7
```

```
(RMSE2 <- sqrt(mean((y - fit2$fitted)^2)))
```

```
## [1] 5.77
```

```
matplot(x, y, pch = 1)
grid()
matlines(x, fit1$fitted, col = COL[1])
matlines(x, fit2$fitted, col = COL[2])

library(latex2exp)
text(10, 20, adj = 0, TeX(sprintf("$RMSE_1 = %1.2f$", RMSE1)))
text( 3, 35, adj = 0, TeX(sprintf("$RMSE_2 = %1.2f$", RMSE2)))

legend("topleft", lty = 1, col = COL,
       legend = c(TeX("$Model_1 ( \\hat{y} = \\beta x )$"),
                  TeX("$Model_2 ( \\hat{y} = \\beta_0 + \\beta_1 x )$")))
```