今日のはなし

DAY-4 7/7　(7)(8) 多変量正規分布: 多次元の正規分布と線形代数(ゼロから理解する正規分布)

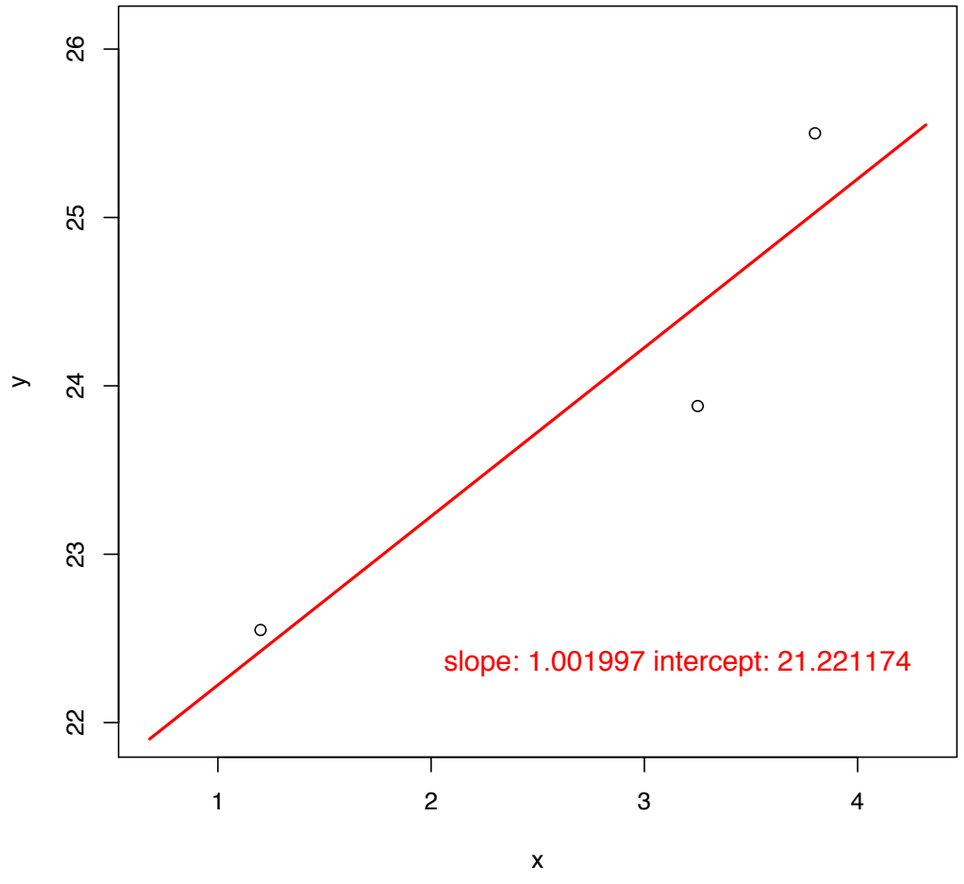
[午前]
- 回帰分析の流れを具体的にみる
- 決定係数
- 相関係数
- 残差プロット

[午後]
- 単変量の正規分布の復習
- 多変量の正規分布

例題：単回帰を計算してみよう

この6個だけ最もシンプルなデータで何が言えるのか？

→ 回帰分析の手順と結果を具体的に見てみよう！



x	y
3.8	25.5
1.2	22.55
3.25	23.88

Rでの回帰分析の実行例

```
> x <- c(3.8, 1.2, 3.25)
> y <- c(25.5, 22.55, 23.88)
> res <- lm(y~x)
> summary(res)

:
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.2212     1.1821  17.953   0.0354 *
x             1.0020     0.3981   2.517   0.2408
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7715 on 1 degrees of freedom
Multiple R-squared:  0.8636, Adjusted R-squared:  0.7273
F-statistic: 6.333 on 1 and 1 DF,  p-value: 0.2408
```

主結果

x	y
3.8	25.5
1.2	22.55
3.25	23.88

Excelでの回帰分析の実行例

	A	B
1	概要	
2		
3	回帰統計	
4	重相関 R	0.929322
5	重決定 R2	0.863639
6	補正 R2	0.727278
7	標準誤差	0.771526
8	観測数	3

10	分散分析表								
11		自由度	変動	分散	測された分散	有意 F			
12	回帰	1	3.770015	3.770015	6.333481	0.240785			
13	残差	1	0.595252	0.595252					
14	合計	2	4.365267						
15									
16		係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
17	切片	21.22117	1.18205	17.95285	0.035424	6.2018	36.24055	6.2018	36.24055
18	X 値 1	1.001997	0.398149	2.516641	0.240785	-4.05696	6.060957	-4.05696	6.060957

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.2212     1.1821  17.953   0.0354 *
x             1.0020     0.3981   2.517   0.2408
```

x	y
3.8	25.5
1.2	22.55
3.25	23.88

$y = a\,x + b$

説明変数	係数値	標準偏差	t比	p値
定数項b	21.2212	1.1821	17.953	0.0354
係数a	1.0020	0.3981	2.517	0.2408

説明変数	係数値	標準偏差	t比	p値
定数項b	21.2212	1.1821	17.953	0.0354
係数a	1.0020	0.3981	2.517	0.2408

$$\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} 1.0 \\ 21.2 \end{bmatrix}$$

x	y
3.8	25.5
1.2	22.55
3.25	23.88

$$\boldsymbol{X} = \begin{bmatrix} 3.8 & 1 \\ 1.2 & 1 \\ 3.25 & 1 \end{bmatrix}, \boldsymbol{y} = \begin{bmatrix} 25.5 \\ 22.55 \\ 23.88 \end{bmatrix}$$

$$\boldsymbol{X}^\top \boldsymbol{X} = \begin{bmatrix} 26.4 & 8.25 \\ 8.25 & 3.0 \end{bmatrix}, \boldsymbol{X}^\top \boldsymbol{y} = \begin{bmatrix} 202.0 \\ 71.9 \end{bmatrix}$$

$$\begin{bmatrix} 26.4 & 8.25 \\ 8.25 & 3.0 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 202.0 \\ 71.9 \end{bmatrix}$$ より

説明変数	係数値	標準偏差	t比	p値
定数項b	21.2212	1.1821	17.953	0.0354
係数a	1.0020	0.3981	2.517	0.2408

$$\sqrt{\mathrm{var}(\hat{b})} = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) = 1.18205$$

$$\sqrt{\mathrm{var}(\hat{a})} = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = 0.3981487$$

$$\hat{\sigma}^2 = 0.5952517$$

$$S_{xx} = 3.755$$

❶〜❸を計算すること

残差分散の不偏推定量

❶ $$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - (\hat{a}x_i + \hat{b}))^2$$

$$\mathbb{E}\{\hat{\sigma}^2\} = \sigma^2$$

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

σ^2 **は未知量なので**

$\hat{\sigma}^2$ **で代用**

傾きの推定量

(分布が正規分布→t分布に)

❸ $$\mathrm{var}(\hat{a}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

切片の推定量

❹ $$\mathrm{var}(\hat{b}) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

単回帰の推定量のまとめ (1)

データの背後に置く仮定	$y = ax + b + \varepsilon, \varepsilon \sim N(0, \sigma^2)$
回帰式	$y = ax + b$

① 係数の推定量

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

② 切片の推定量

$$\hat{b} = \bar{y} - \hat{a}\bar{x} = \sum_{i=1}^{n}\left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}}\right)y_i$$

③ 残差分散の不偏推定量

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - (\hat{a}x_i + \hat{b}))^2$$

単回帰の推定量のまとめ (2)

注意：教科書では最初から平均値を引いてあるので $\bar{x} = 0$

① 係数の推定量

$$\mathbb{E}\{\hat{a}\} = a$$

$$\mathrm{var}(\hat{a}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

→ 教科書(4.62)

② 切片の推定量

$$\mathbb{E}\{\hat{b}\} = b$$

$$\mathrm{var}(\hat{b}) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right)$$

→ 教科書(4.61)

$$\mathrm{cov}(\hat{a}, \hat{b}) = \frac{-\sigma^2 \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

→ 教科書(4.63)

③ 残差分散の不偏推定量

$$\mathbb{E}\{\hat{\sigma}^2\} = \sigma^2$$

説明変数	係数値	標準偏差	t比	p値
定数項b	21.2212	1.1821	17.953	0.0354
係数a	1.0020	0.3981	2.517	0.2408

$$T_b = \frac{\hat{b}}{\sqrt{\mathrm{var}(\hat{b})}} = 17.95285$$

$$T_a = \frac{\hat{a}}{\sqrt{\mathrm{var}(\hat{a})}} = 2.516641$$

不偏推定量なので

$$\mathbb{E}\{\hat{a}\} = a,\ \mathbb{E}\{\hat{b}\} = b$$

上記T比はa=0やb=0の検定統計量

分散の定義

$$\mathrm{var}\{X\} := \mathbb{E}\{(X - \mathbb{E}\{X\})^2\}$$

$$X \sim N(\mathbb{E}\{X\}, \mathrm{var}\{X\})$$

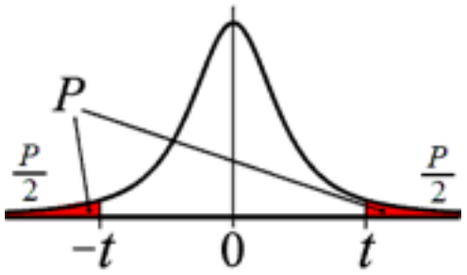
標準化 ⟶ $$\frac{X - \mathbb{E}\{X\}}{\sqrt{\mathrm{var}\{X\}}} \sim N(0,1)$$

varを標本推定量で置換すると

⟶ $$\frac{X - \mathbb{E}\{X\}}{\sqrt{\widehat{\mathrm{var}}\{X\}}} \sim t(n-2)$$

説明変数	係数値	標準偏差	t比	p値
定数項b	21.2212	1.1821	17.953	0.0354
係数a	1.0020	0.3981	2.517	0.2408

$$T \sim t(1)$$

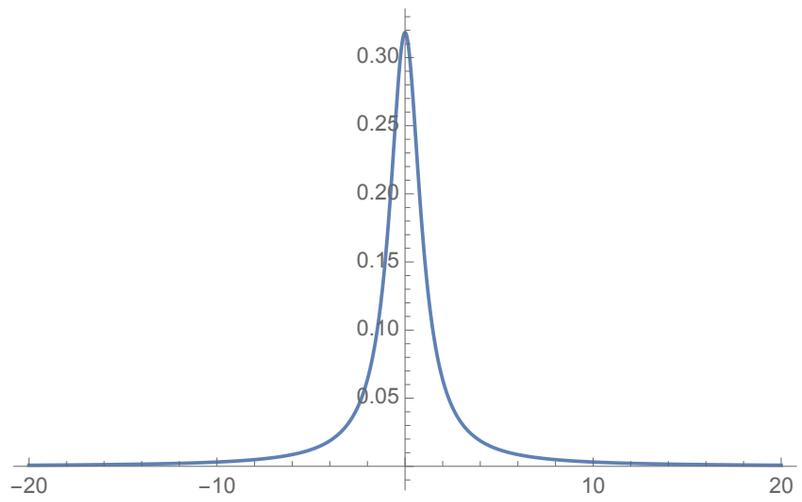


$$n = 3$$

$$t(n-2) = t(1)$$

$$P(|T| > 17.953) = P(T > 17.953) + P(T < -17.953)$$
$$= 2 \times P(T < -17.953) = 0.0354$$

$$P(|T| > 2.517) = P(T > 2.517) + P(T < -2.517)$$
$$= 2 \times P(T < -2.517) = 0.2408$$



degrees of freedom f	two tails probability P				
	0.1	0.05	0.02	0.01	0.001
1	6.313752	12.70620	31.82052	63.65674	636.6192

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   21.2212     1.1821  17.953   0.0354 *
x              1.0020     0.3981   2.517   0.2408
```

degrees of freedom f	two tails probability P				
	0.1	0.05	0.02	0.01	0.001
1	6.313752	12.70620	31.82052	63.65674	636.6192

x	y
3.8	25.5
1.2	22.55
3.25	23.88

$y = a\,x + b$

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

説明変数	係数値	標準偏差	t比	p値
定数項b	21.2212	1.1821	17.953	0.0354
係数a	1.0020	0.3981	2.517	0.2408

```
> x <- c(3.8, 1.2, 3.25)
> y <- c(25.5, 22.55, 23.88)
> res <- lm(y~x)
> summary(res)

:

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.2212     1.1821  17.953   0.0354 *
x             1.0020     0.3981   2.517   0.2408
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7715 on 1 degrees of freedom
Multiple R-squared:  0.8636, Adjusted R-squared:  0.7273
F-statistic: 6.333 on 1 and 1 DF,  p-value: 0.2408
```

標準誤差

$$\hat{\sigma}^2 = 0.5952517$$

$$\sqrt{\hat{\sigma}^2} = 0.7715256$$

x	y	$\bar{y}$	$\hat{y}$
3.8	25.5	23.98	25.03
1.2	22.55	23.98	22.42
3.25	23.88	23.98	24.48

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - (\hat{a}x_i + \hat{b}))^2$$

$$= \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$n - 2 = 3 - 2 = 1$$

```
> x <- c(3.8, 1.2, 3.25)
> y <- c(25.5, 22.55, 23.88)
> res <- lm(y~x)
> summary(res)

:

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.2212     1.1821  17.953   0.0354 *
x             1.0020     0.3981   2.517   0.2408
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7715 on 1 degrees of freedom
Multiple R-squared:  0.8636, Adjusted R-squared:  0.7273
F-statistic: 6.333 on 1 and 1 DF,  p-value: 0.2408
```

自由度調整済み決定係数

$$R^{*2} := 1 - \frac{0.5952517/1}{4.365267/2} = 0.7272782$$

決定係数

$$R^2 := \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{3.770015}{4.365267} = 0.8636391$$

x	y	$\bar{y}$	$\hat{y}$
3.8	25.5	23.98	25.03
1.2	22.55	23.98	22.42
3.25	23.88	23.98	24.48

$$= 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{0.5952517}{4.365267} = 0.8636391$$

決定係数(または寄与率)

平方和の分解 (教科書p.48)

yの全変動　　回帰変動　　残差

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 + \sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

↳ 回帰で説明できた部分

$$R^2 := \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

y と $\hat{y}$ の相関係数 r の二乗 r^2 に等しい

(自由度調整済み版) $n = 3, p = 1$　　**各々平方和の自由度で割る**

$$R^{*2} := 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-p-1)}{\sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)}$$

← $\hat{a}, \hat{b}$ により2つ減る

← $\bar{y}$ により1つ減る

自由度 = 自由に値をとることができる変数の数

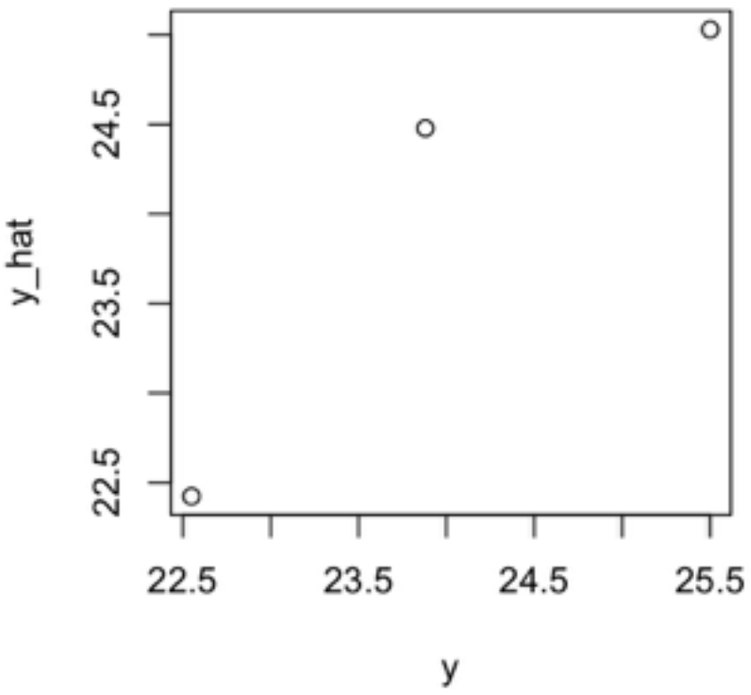
例)　$X_1, X_2, X_3, \bar{X} = (X_1 + X_2 + X_3)/3$ の3変数に対し

平方和　$(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + (X_3 - \bar{X})^2$

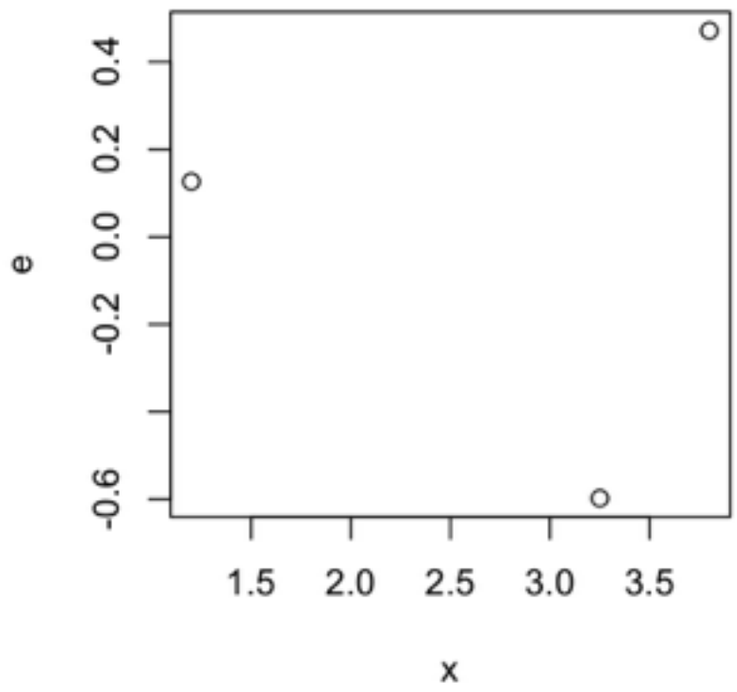
を考える。$\bar{X}$ が決まっているとすると実質動ける変数は２つしかない。このことを指して、この平方和の自由度は２であると言う。 $\bar{X} = 10, X_1 = 9, X_2 = 10 \Rightarrow X_3 = 11$

x	y	$\bar{y}$	$\hat{y}$	$\varepsilon = y - \hat{y}$
3.8	25.5	23.98	25.03	0.471
1.2	22.55	23.98	22.42	0.126
3.25	23.88	23.98	24.48	-0.598

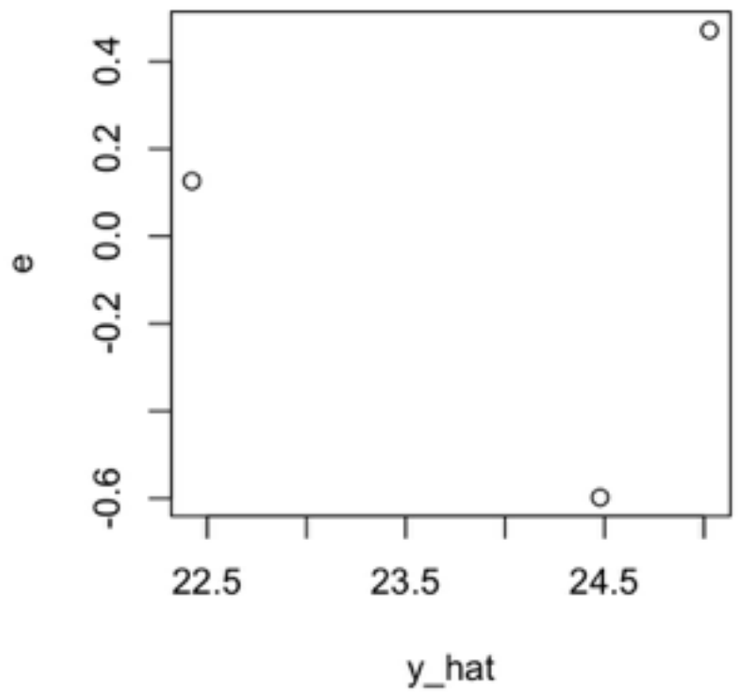
残差や予測のプロット



重回帰でもOK



重回帰ではNG



重回帰でもOK

```
> x <- c(3.8, 1.2, 3.25)
> y <- c(25.5, 22.55, 23.88)
> res <- lm(y~x)
> summary(res)

:

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.2212     1.1821  17.953   0.0354 *
x             1.0020     0.3981   2.517   0.2408
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7715 on 1 degrees of freedom
Multiple R-squared:  0.8636, Adjusted R-squared:  0.7273
F-statistic: 6.333 on 1 and 1 DF,  p-value: 0.2408
```

仮説「回帰係数がすべて0」＝「回帰に意味がない」の検定

→ 単回帰では係数1個しかないので「傾き=0」の検定と同じ

注意：ただしt検定ではなくF検定を用いる。

t検定とF検定の関係(自由度kのt分布に従う確率変数の2乗は
自由度(1,k)のF分布に従う)より同じ検定(p値は同じ)となる。

F統計量　$n = 3, p = 1$

各々平方和の自由度で割る

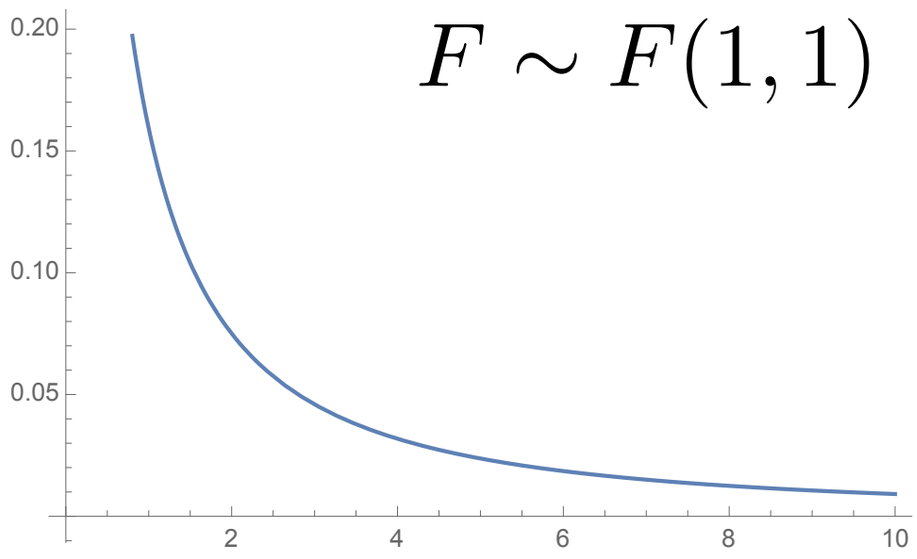
$$F = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})/p}{\sum_{i=1}^{n}(y_i - \hat{y}_i)/(n-p-1)} = \frac{3.770015/1}{0.5952517/(3-1-1)} = 6.333481$$

この量は　$\beta_1 = \beta_2 = \cdots = \beta_p = 0$　の仮説のもとで

自由度 $(p, n-p-1)$ のF分布に従う

$$F \sim F(1,1)$$

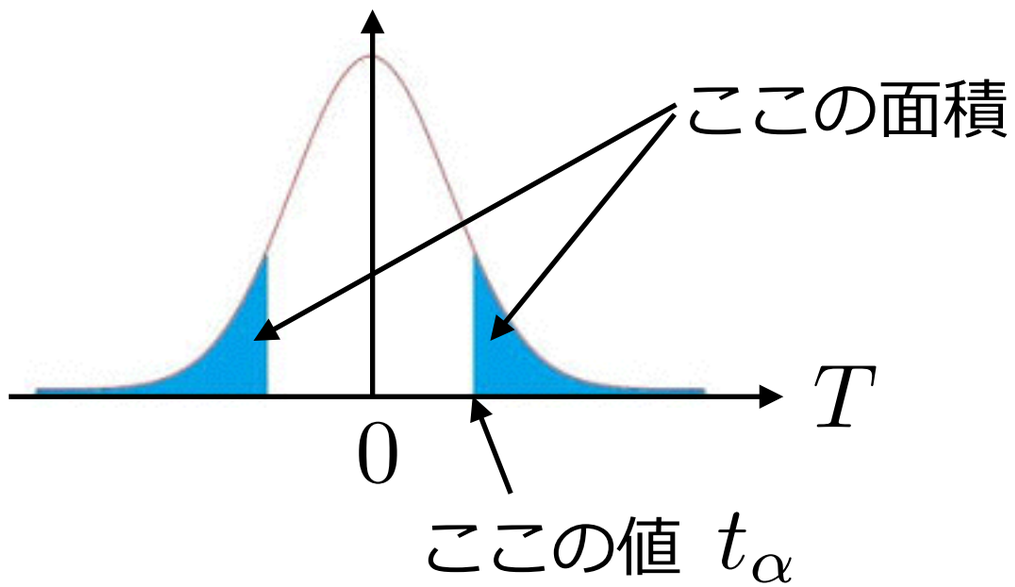
$$P(F > 6.333481) = 0.24078$$



仮説検定と区間推定は同じことの表と裏の関係にある

5%

裾確率　$P(|T| \geqslant t_\alpha) = \alpha$



95%

確率 $1-\alpha$ で $|T| < t_\alpha$

$$\Leftrightarrow \left| \frac{X - \mathbb{E}\{X\}}{\sqrt{\widehat{\mathrm{var}}\{X\}}} \right| < t_\alpha$$

$$\Leftrightarrow X - t_\alpha \sqrt{\widehat{\mathrm{var}}\{X\}} < \mathbb{E}\{X\} < X + t_\alpha \sqrt{\widehat{\mathrm{var}}\{X\}}$$

$$X \sim N(\mathbb{E}\{X\}, \mathrm{var}\{X\})$$

標準化 ⟶ $\dfrac{X - \mathbb{E}\{X\}}{\sqrt{\mathrm{var}\{X\}}} \sim N(0,1)$

varを標本推定量で置換すると

⟶ $\dfrac{X - \mathbb{E}\{X\}}{\sqrt{\widehat{\mathrm{var}}\{X\}}} \sim t(n-2)$

```
> x <- c(3.8, 1.2, 3.25)
> y <- c(25.5, 22.55, 23.88)
> res <- lm(y~x)
> summary(res)

Call:
lm(formula = y ~ x)

Residuals:
      1       2       3
 0.4712  0.1264 -0.5977

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.2212     1.1821  17.953   0.0354 *
x             1.0020     0.3981   2.517   0.2408
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7715 on 1 degrees of freedom
Multiple R-squared:  0.8636, Adjusted R-squared:  0.7273
F-statistic: 6.333 on 1 and 1 DF,  p-value: 0.2408
```

10	分散分析表								
11		自由度	変動	分散	測された分散	有意 F			
12	回帰	1	3.770015	3.770015	6.333481	0.240785			
13	残差	1	0.595252	0.595252					
14	合計	2	4.365267						
15									
16		係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
17	切片	21.22117	1.18205	17.95285	0.035424	6.2018	36.24055	6.2018	36.24055
18	X 値 1	1.001997	0.398149	2.516641	0.240785	-4.05696	6.060957	-4.05696	6.060957

$$\bar{x} = 2.75$$

$$t(1, 0.05) = 12.7062$$

```
> coef(res)
(Intercept)           x
  21.221174    1.001997
> confint(res)
                2.5 %    97.5 %
(Intercept)  6.201800 36.240548
x           -4.056962  6.060957
> predict(res,data.frame(x=2.0),interval="confidence")
       fit      lwr      upr
1 23.22517 16.41121 30.03913
> predict(res,data.frame(x=2.0),interval="prediction")
       fit      lwr      upr
1 23.22517 11.28649 35.16385
```

$$\hat{b} \pm t(1, 0.05)\sqrt{\mathrm{var}(\hat{b})}$$

$$\hat{a} \pm t(1, 0.05)\sqrt{\mathrm{var}(\hat{a})}$$

$$2.0\hat{a} + \hat{b} \pm t(1, 0.05)\sqrt{\hat{\sigma}^2\left(\frac{1}{3} + \frac{(2.0 - \bar{x})^2}{S_{xx}}\right)}$$

$$2.0\hat{a} + \hat{b} \pm t(1, 0.05)\sqrt{\hat{\sigma}^2\left(1 + \frac{1}{3} + \frac{(2.0 - \bar{x})^2}{S_{xx}}\right)}$$

Excelでの実行の補足

Excel2003以前: 「ツール」→「分析ツール」をクリック
Excel2007: 「データ」タブの中にある「データ分析」をクリック

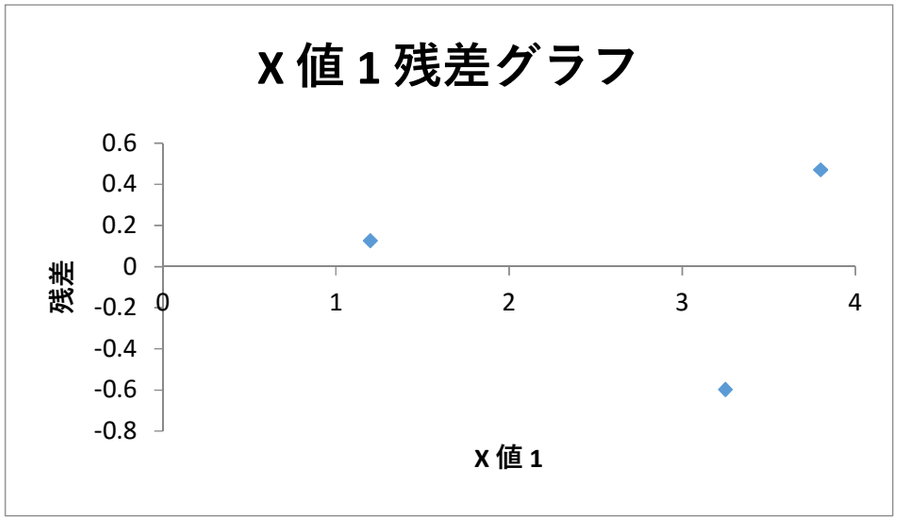
Mac版(Excel 2011)にはこの機能ないので注意

概要								
回帰統計								
重相関 R	0.951759							
重決定 R2	0.905845							
補正 R2	0.901838							
標準誤差	8.612262							
観測数	50							
分散分析表								
	自由度	変動	分散	観測された分散比	有意 F			
回帰	2	33538.42	16769.21	226.0882953	7.679E-25			
残差	47	3486.04	74.17106					
合計	49	37024.46						
	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	21.96217	9.353082	2.348121	0.023127974	3.146194	40.77814	3.146194	40.77814
x_α	3.202912	0.230515	13.89456	2.9214 3E-18	2.7391744	3.666649	2.739174	3.666649
x_β	4.609189	0.505876	9.111303	5.9088E-12	3.5914981	5.62688	3.591498	5.62688

概要

回帰統計	
重相関 R	0.929322
重決定 R2	0.863639
補正 R2	0.727278
標準誤差	0.771526
観測数	3

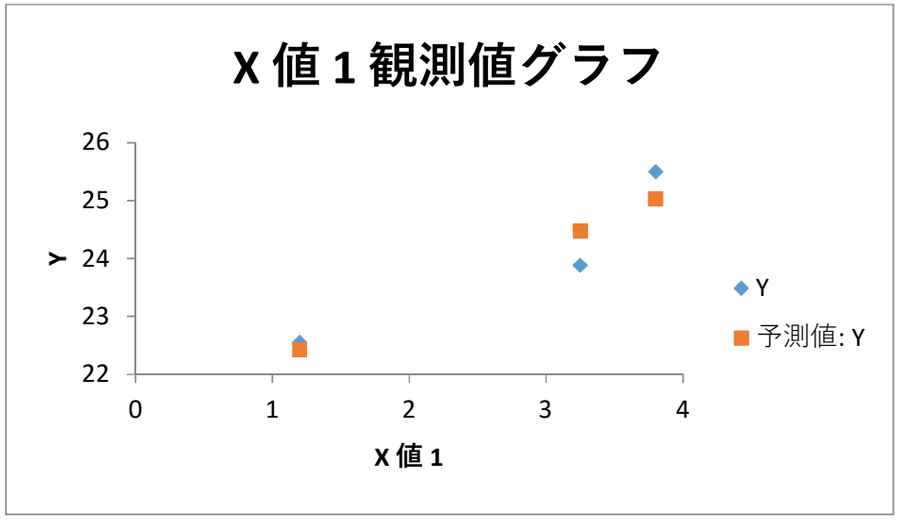
x	y
3.8	25.5
1.2	22.55
3.25	23.88



分散分析表

	自由度	変動	分散	測された分散	有意 F
回帰	1	3.770015	3.770015	6.333481	0.240785
残差	1	0.595252	0.595252		
合計	2	4.365267			

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 90.0%	上限 90.0%
切片	21.22117	1.18205	17.95285	0.035424	6.2018	36.24055	13.758	28.68435
X 値 1	1.001997	0.398149	2.516641	0.240785	-4.05696	6.060957	-1.51181	3.51581

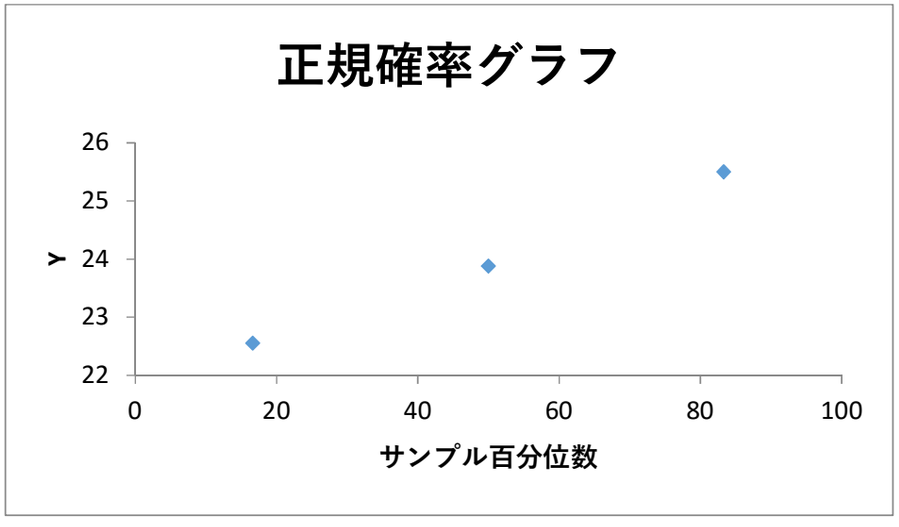


残差出力

観測値	予測値: Y	残差	標準残差
1	25.02876	0.471236	0.86378
2	22.42357	0.126429	0.231746
3	24.47767	-0.59767	-1.09553

確率

百分位数	Y
16.66667	22.55
50	23.88
83.33333	25.5



Excelの分析ツールの出力(フル)

参考

回帰分析 (統計ライブラリー), 佐和隆光 著, 朝倉書店, 1979.

https://www.amazon.co.jp/dp/4254125135/

回帰と相関，知っているようで知らない，その本質: Excel の回帰分析を例として

http://www.ab.auone-net.jp/~biology/regline/regline.htm

「データ解析」(下平英寿) 検定と信頼区間

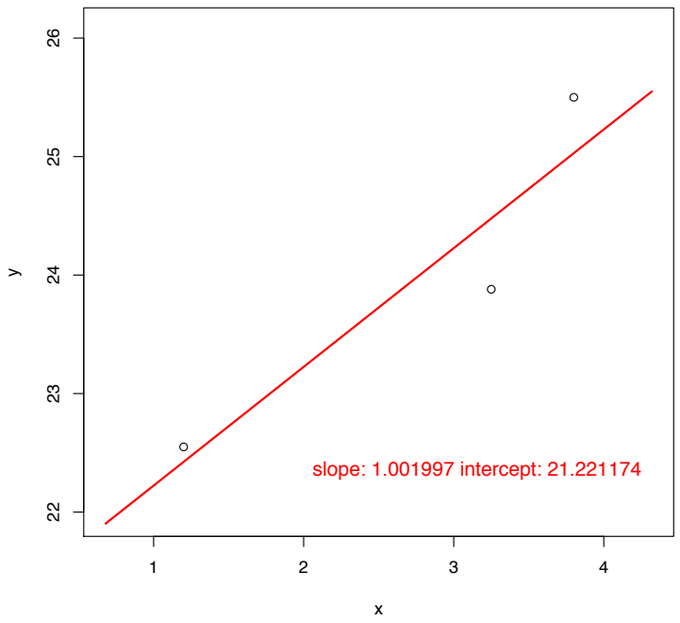
http://www.is.titech.ac.jp/~shimo/class/dk2005/san07.pdf

重回帰分析へ

単回帰

説明変数1個

回帰＝直線(1次元平面)



x	y
3.8	25.5
1.2	22.55
3.25	23.88

重回帰

説明変数p個

回帰＝p次元(超)平面

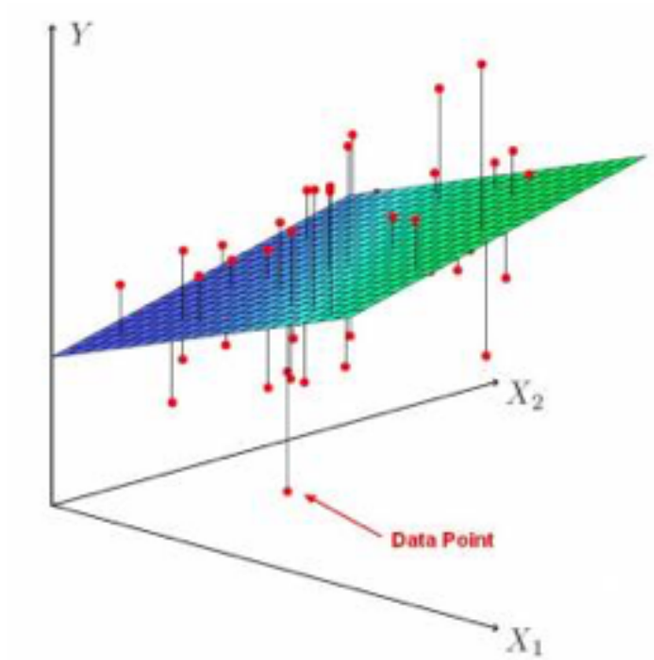


表 1.3　中古マンションのデータ

サンプル No.	広さ x_1 (m^2)	築年数 x_2 (年数)	価格 y (千万円)
1	51	16	3.0
2	38	4	3.2
3	57	16	3.3
4	51	11	3.9
5	53	4	4.4
6	77	22	4.5
7	63	5	4.5
8	69	5	5.4
9	72	2	5.4
10	73	1	6.0

- 参考) 正規分布する変数を線形変換しても正規分布する

証明「数理統計学ー基礎から学ぶデータ解析ー(鈴木武・山田作太郎著)」p.121 定理4.2

定理 4.2 $\boldsymbol{X}$ が k 次元正規分布 $N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ に従うとき

$$\boldsymbol{Y}=\boldsymbol{A}\boldsymbol{X}+\boldsymbol{a},$$

$\boldsymbol{A}$ は $p\times k$ の定数行列，$\boldsymbol{a}$ は p 次元定数ベクトル

とおくと，$\boldsymbol{Y}$ は $N_p(\boldsymbol{A}\boldsymbol{\mu}+\boldsymbol{a}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}')$ に従う．

証明 $\boldsymbol{t}=(t_1, \cdots, t_p)'$ に対して

$$M_Y(\boldsymbol{t})=E(e^{\boldsymbol{t}'(\boldsymbol{A}\boldsymbol{X}+\boldsymbol{a})})=e^{\boldsymbol{t}'\boldsymbol{a}}E(e^{\boldsymbol{t}'\boldsymbol{A}\boldsymbol{X}})$$

$$=e^{\boldsymbol{t}'\boldsymbol{a}}\exp\left\{\boldsymbol{t}'\boldsymbol{A}\boldsymbol{\mu}+\frac{1}{2}\boldsymbol{t}'\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}'\boldsymbol{t}\right\}$$

$$=\exp\left\{\boldsymbol{t}'(\boldsymbol{A}\boldsymbol{\mu}+\boldsymbol{a})+\frac{1}{2}\boldsymbol{t}'\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}'\boldsymbol{t}\right\}$$

を得る．積率母関数は確率分布と1対1に対応するので，(4.51)より $\boldsymbol{Y}$ は p 次元正規分布 $N_p(\boldsymbol{A}\boldsymbol{\mu}+\boldsymbol{a}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}')$ に従う． □

定理 4.2 では $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}'$ は対称行列ではあるが，必ずしも正値定符号ではない．$p\leqq k$ のとき $\boldsymbol{Y}$ の分布が(4.45)の意味での p 次元正規分布であるためには，$\boldsymbol{A}$ のランク(rank)が p であればよい．

相関係数: 標準化変数の共分散

共分散　$\sigma_{x_i,x_j} := \mathbb{E}\{(x_i - \mu_i)(x_j - \mu_j)\}$

相関係数　$r_{i,j} := \mathbb{E}\left\{\left(\frac{x_i - \mu_i}{\sigma_i}\right) \cdot \left(\frac{x_j - \mu_j}{\sigma_j}\right)\right\}$

$= \mathbb{E}\{z_i \cdot z_j\}$

標準化 (平均0,分散1の確率変数に変換)

$$z_i = \frac{x_i - \mu_i}{\sigma_i} \quad \Rightarrow \quad \mathbb{E}\{z_i\} = 0$$

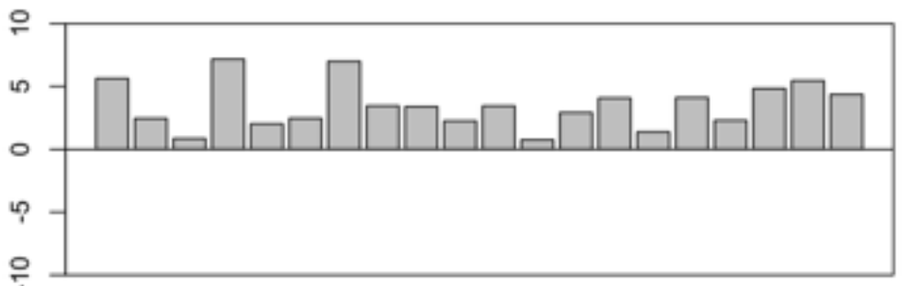
$$\mathrm{var}\{z_i\} = \mathbb{E}\{(z_i - 0)^2\} = 1$$

相関係数の直感的イメージ

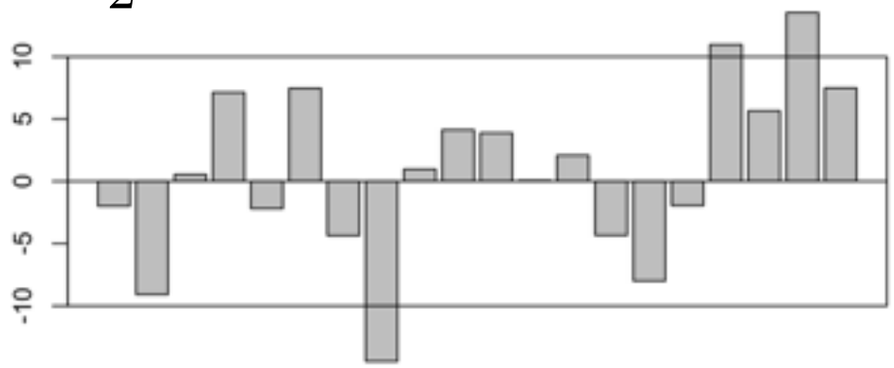
相関係数　$r_{i,j} := \mathbb{E}\left\{\left(\frac{x_i - \mu_i}{\sigma_i}\right) \cdot \left(\frac{x_j - \mu_j}{\sigma_j}\right)\right\}$

$= \mathbb{E}\{z_i \cdot z_j\}$

x_1

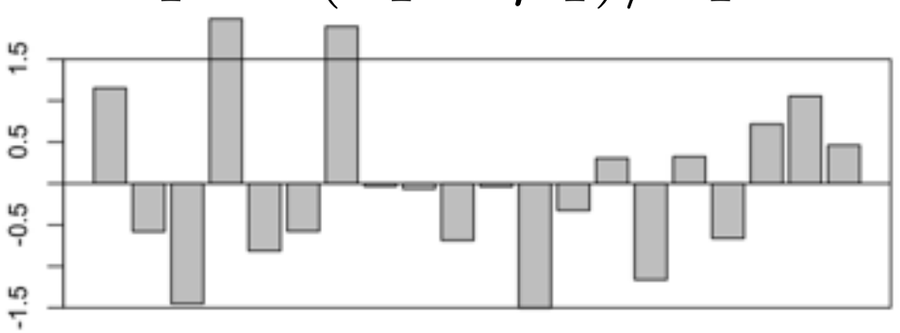


x_2

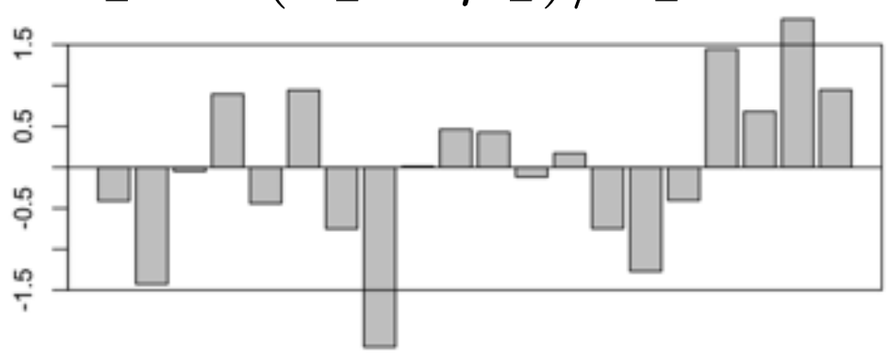


標準化　分布の真ん中(平均)とブレ幅(分散)を整えた後の変動の類似

$z_1 := (x_1 - \mu_1)/\sigma_1$



$z_2 := (x_2 - \mu_2)/\sigma_2$



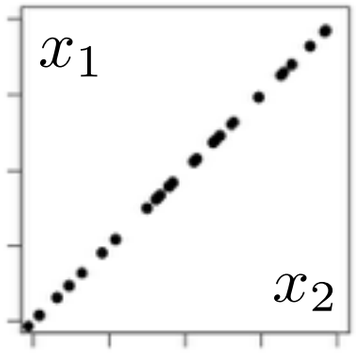
$r_{i,j} = 1$

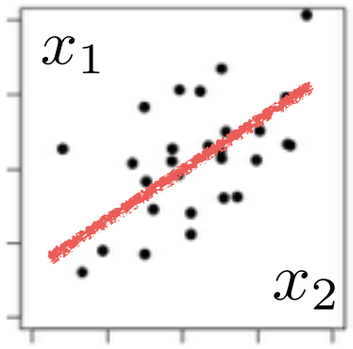
$r_{i,j} > 0$

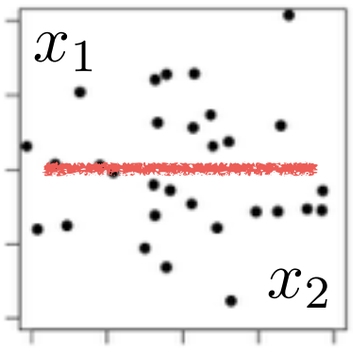
$r_{i,j} = 0$

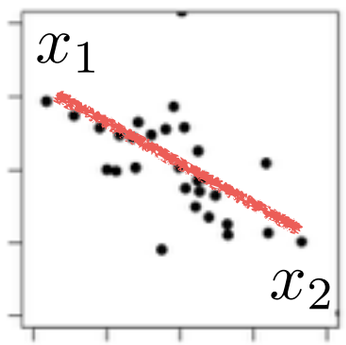
$r_{i,j} < 0$

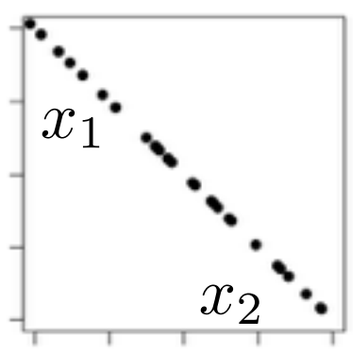
$r_{i,j} = -1$

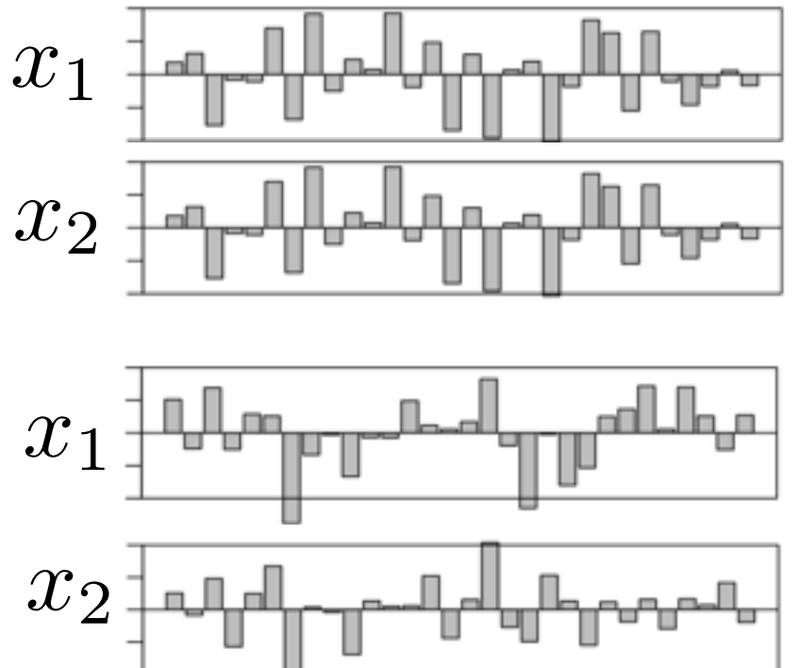








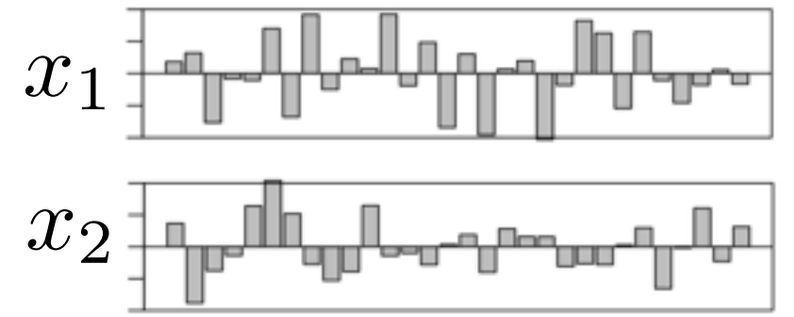




正の相関

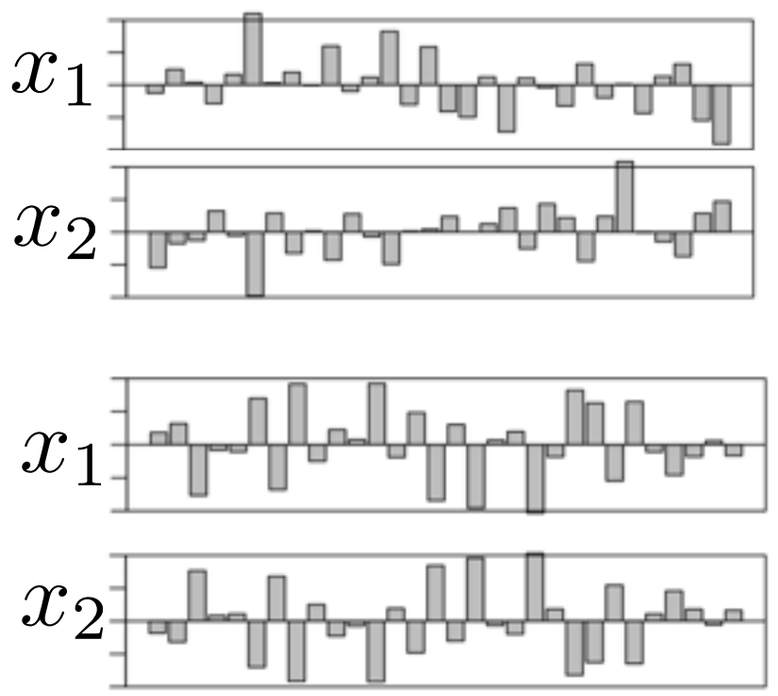
$x_1 \downarrow, x_2 \downarrow$

$x_1 \uparrow, x_2 \uparrow$



無相関

(パターンなし)



$x_1 \uparrow, x_2 \downarrow$

$x_1 \downarrow, x_2 \uparrow$

負の相関

重回帰の推定量のまとめ　　p 変量とする (説明変数が p 個)

データの背後に置く仮定	$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I)$
回帰式	$\boldsymbol{y} = X\boldsymbol{\beta}$

① 偏回帰係数の推定量

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \|\boldsymbol{\epsilon}\|^2 = (X'X)^{-1}X'\boldsymbol{y}$$

$\boldsymbol{y}, \boldsymbol{\epsilon} \in \mathbb{R}^n$
$X \in \mathbb{R}^{n \times p}$
$\boldsymbol{\beta} \in \mathbb{R}^p$

$$\mathbb{E}\{\hat{\boldsymbol{\beta}}\} = \boldsymbol{\beta}$$

教科書p.58の(3)より

$$\mathbb{E}\{(\hat{\boldsymbol{\beta}} - \mathbb{E}\{\boldsymbol{\beta}\})(\hat{\boldsymbol{\beta}} - \mathbb{E}\{\boldsymbol{\beta}\})'\} = \mathbb{E}\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\}$$
$$= \sigma^2(X'X)^{-1}$$

よって、各々の偏回帰係数の分布は
$\hat{\beta}_i \sim N(\beta_i, \sigma^2 / S_{x_i x_i})$

② 残差分散の不偏推定量 (後述)

$$\hat{\sigma}^2 := \frac{1}{n-p-1}\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|^2$$

（3）統計量の分布

※教科書p.58-59

単回帰モデル (4.49) より

$$\begin{aligned} E(\boldsymbol{y}) &= X\boldsymbol{\beta} \\ V(\boldsymbol{y}) &= V(\boldsymbol{\varepsilon}) = \sigma^2 I_n \end{aligned} \tag{4.57}$$

に注意する．(4.57) 式および (3.50) 式と (3.51) 式を用いて $\hat{\boldsymbol{\beta}}$ の期待値と分散共分散行列は次のようになる．

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E((X'X)^{-1}X'\boldsymbol{y}) \\ &= (X'X)^{-1}X'\,E(\boldsymbol{y}) \\ &= (X'X)^{-1}X'\,X\boldsymbol{\beta} = \boldsymbol{\beta} \end{aligned} \tag{4.58}$$

$$\begin{aligned} V(\hat{\boldsymbol{\beta}}) &= V((X'X)^{-1}X'\boldsymbol{y}) \\ &= (X'X)^{-1}X'\,V(\boldsymbol{y})\,X(X'X)^{-1} \\ &= (X'X)^{-1}X'\,\sigma^2 I_n\,X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \end{aligned} \tag{4.59}$$

これより，

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(X'X)^{-1}) \tag{4.60}$$

$$\boldsymbol{X} = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}, \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \\ 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & n \end{bmatrix}$$

$$\boldsymbol{X}'\boldsymbol{y} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \\ 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n y_i \end{bmatrix}$$

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} & \frac{-\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \frac{-\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} & \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{1}{n} \end{bmatrix}$$

$$\mathrm{var}\{\boldsymbol{\beta}\} = \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}$$

$$\mathrm{var}(\hat{a}) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\mathrm{var}(\hat{b}) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)$$

$$\min \|\boldsymbol{\epsilon}\|^2 = (\boldsymbol{y} - X\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})$$

$$= (\boldsymbol{y} - X(X'X)^{-1}X'\boldsymbol{y})'(\boldsymbol{y} - X(X'X)^{-1}X'\boldsymbol{y})$$

$$= \boldsymbol{y}'(I - X(X'X)^{-1}X')'(I - X(X'X)^{-1}X')\boldsymbol{y}$$

$$= \boldsymbol{y}'(I - X(X'X)^{-1}X')\boldsymbol{y}$$

$P := (I - X(X'X)^{-1}X')$

$PP = P$ (ベキ等)

$P = P'$ (対称)

$$= (X\boldsymbol{\beta} + \boldsymbol{\epsilon})'(I - X(X'X)^{-1}X')(X\boldsymbol{\beta} + \boldsymbol{\epsilon})$$

$X'(I - X(X'X)^{-1}X') = O$

$(I - X(X'X)^{-1}X')X = O$

$$= \boldsymbol{\epsilon}'(I - X(X'X)^{-1}X')\boldsymbol{\epsilon}$$

$\boldsymbol{v}'A\boldsymbol{v} = \mathrm{tr}\{A\boldsymbol{v}\boldsymbol{v}'\}$

※教科書p.37 (3.30)

$$= \mathrm{tr}\{(I - X(X'X)^{-1}X')\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\}$$

$$
\begin{aligned}
\mathbb{E}\{\min \|\boldsymbol{\epsilon}\|^2\} &= \mathbb{E}\{\mathrm{tr}\{(I - X(X'X)^{-1}X')\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\}\} \\
&= \mathrm{tr}\{(I - X(X'X)^{-1}X')\mathbb{E}\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\}\} \\
&= \mathrm{tr}\{(I - X(X'X)^{-1}X')(\sigma^2 I)\} \\
&= \sigma^2 \mathrm{tr}\{(I - X(X'X)^{-1}X')\} \\
&= \sigma^2(\underbrace{\mathrm{tr}\{I\}}_{=\mathrm{rank}(I)\ =n} - \underbrace{\mathrm{tr}\{X(X'X)^{-1}X'\}}_{=\mathrm{rank}(X(X'X)^{-1}X')\ =p+1}\}
\end{aligned}
$$

$A\boldsymbol{v} = \lambda\boldsymbol{v}$ という
固有値の定義から
対称ベキ等行列の
固有値は1か0なので

tr A = rank A

$$
= (n - p - 1)\sigma^2
$$

よって、p変量の時の誤差分散 σ^2 の**不偏**推定量は

$$
\hat{\sigma}^2 := \frac{1}{n - p - 1}\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|^2
$$

単回帰のときn-2で
割ったのに相当(導出
の詳細は前回資料参照)

$$\hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}} = \underbrace{X(X'X)^{-1}X'}_{= H}\boldsymbol{y}$$

ハット行列　　$\mathrm{Im}(X)$への直交射影行列

重相関係数　　$\boldsymbol{y}$と$\hat{\boldsymbol{y}}$の相関係数

$$R = \frac{(\boldsymbol{y}-\bar{\boldsymbol{y}})'(\hat{\boldsymbol{y}}-\bar{\boldsymbol{y}})}{\|\boldsymbol{y}-\bar{\boldsymbol{y}}\| \cdot \|\hat{\boldsymbol{y}}-\bar{\boldsymbol{y}}\|}$$

$$= \frac{(\boldsymbol{y}-\bar{\boldsymbol{y}})'H(\boldsymbol{y}-\bar{\boldsymbol{y}})}{\|\boldsymbol{y}-\bar{\boldsymbol{y}}\| \cdot \|\hat{\boldsymbol{y}}-\bar{\boldsymbol{y}}\|} = \frac{(\boldsymbol{y}-\bar{\boldsymbol{y}})'H^2(\boldsymbol{y}-\bar{\boldsymbol{y}})}{\|\boldsymbol{y}-\bar{\boldsymbol{y}}\| \cdot \|\hat{\boldsymbol{y}}-\bar{\boldsymbol{y}}\|} = \frac{\|H(\boldsymbol{y}-\bar{\boldsymbol{y}})\|^2}{\|\boldsymbol{y}-\bar{\boldsymbol{y}}\| \cdot \|\hat{\boldsymbol{y}}-\bar{\boldsymbol{y}}\|}$$

$$= \frac{\|\hat{\boldsymbol{y}}-\bar{\boldsymbol{y}}\|^2}{\|\boldsymbol{y}-\bar{\boldsymbol{y}}\| \cdot \|\hat{\boldsymbol{y}}-\bar{\boldsymbol{y}}\|} = \frac{\|\hat{\boldsymbol{y}}-\bar{\boldsymbol{y}}\|}{\|\boldsymbol{y}-\bar{\boldsymbol{y}}\|}$$

偏差平方和（全変動）　回帰による平方和　残差平方和

$$\|\boldsymbol{y}-\bar{\boldsymbol{y}}\|^2 = \|\hat{\boldsymbol{y}}-\bar{\boldsymbol{y}}\|^2 + \underbrace{\|\hat{\boldsymbol{y}}-\boldsymbol{y}\|^2}_{= \|\boldsymbol{\epsilon}\|^2}$$

※教科書p.48,p.69

決定係数(寄与率)

$$R^2 = \frac{\|\hat{\boldsymbol{y}}-\bar{\boldsymbol{y}}\|^2}{\|\boldsymbol{y}-\bar{\boldsymbol{y}}\|^2} = \frac{\|\boldsymbol{y}-\bar{\boldsymbol{y}}\|^2 - \|\boldsymbol{\epsilon}\|^2}{\|\boldsymbol{y}-\bar{\boldsymbol{y}}\|^2} = 1 - \frac{\|\boldsymbol{\epsilon}\|^2}{\|\boldsymbol{y}-\bar{\boldsymbol{y}}\|^2}$$

分散分析とF値：回帰式が役に立つかどうかのF検定

回帰分散　$V_{\hat{\boldsymbol{y}}} := \dfrac{1}{p}\|\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}}\|^2$

誤差分散　$V_{\boldsymbol{\epsilon}} := \dfrac{1}{n-p-1}\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2 = \dfrac{1}{n-p-1}\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|^2$

分散比 $F := \dfrac{V_{\hat{\boldsymbol{y}}}}{V_{\boldsymbol{\epsilon}}}$ は、「仮説 $\boldsymbol{\beta} = \mathbf{0}$ が正しい」とき、

自由度 $p, n-p-1$ のF分布に従う。

※ $\boldsymbol{\beta} = \mathbf{0}$ なら投入した説明変数は目的変数の説明に何の役にも立っていないということなので、回帰式は無意味になる

F値と決定係数の関係

回帰分散　$V_{\hat{\boldsymbol{y}}} := \dfrac{1}{p}\|\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}}\|^2$

誤差分散　$V_{\boldsymbol{\epsilon}} := \dfrac{1}{n-p-1}\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2 = \dfrac{1}{n-p-1}\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|^2$

決定係数　$R^2 = \dfrac{\|\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}}\|^2}{\|\boldsymbol{y} - \bar{\boldsymbol{y}}\|^2} = \dfrac{\|\boldsymbol{y} - \bar{\boldsymbol{y}}\|^2 - \|\boldsymbol{\epsilon}\|^2}{\|\boldsymbol{y} - \bar{\boldsymbol{y}}\|^2} = 1 - \dfrac{\|\boldsymbol{\epsilon}\|^2}{\|\boldsymbol{y} - \bar{\boldsymbol{y}}\|^2}$

$$1 - R^2 = \frac{\|\boldsymbol{\epsilon}\|^2}{\|\boldsymbol{y} - \bar{\boldsymbol{y}}\|^2} = \frac{\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2}{\|\boldsymbol{y} - \bar{\boldsymbol{y}}\|^2}$$

$$F := \frac{V_{\hat{\boldsymbol{y}}}}{V_{\boldsymbol{\epsilon}}} = \frac{n-p-1}{p}\frac{\|\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}}\|^2}{\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2} = \frac{n-p-1}{p}\frac{R^2}{1-R^2}$$

※標準化偏回帰係数

偏回帰係数の大きさは各々の変量のダイナミックレンジに依存する！たとえば、x1は1～100までの値をとり、x2は-1～1までの場合、β1のほうがβ2より小さくなりやすいがこれは必ずしもβ1のほうがβ2より目的変数を説明しないことを意味しない。

なので、目的変数、および、各説明変数を標準化してから計算した偏回帰係数を**標準化偏回帰係数**と言う。

標準化偏回帰係数 $\tilde{\beta}_i = \hat{\beta}_i \dfrac{\mathrm{sd}(x_i)}{\mathrm{sd}(y)}$ として非標準化係数から計算できる

※自由度調整済み決定係数

決定係数は説明変数の数が増えると大きくなるため
自由度で調整したものも用いられる。(教科書p.80)

$$R^{*2} := 1 - \frac{S_e/(n-p-1)}{S_y y/(n-1)}$$

※標準化残差

残差を $\frac{\epsilon_i}{\sqrt{V_\epsilon}}$ で標準化すると近似的にN(0,1)に従う。

ヒストグラム　①はずれ値→データミスをチェック
②左右非対称→重要な説明変数の見逃し
③多峰・高原型→何らかの層別が必要

アンスコムの数値例と残差プロット

表 6.2　アンスコムの数値例

i	(a)～(c) X	(a) Y	(b) Y	(c) Y	(d) X	(d) Y
1	10.0	8.04	9.14	7.46	8.0	6.58
2	8.0	6.95	8.14	6.77	8.0	5.76
3	13.0	7.58	8.74	12.74	8.0	7.71
4	9.0	8.81	8.77	7.11	8.0	8.84
5	11.0	8.33	9.26	7.81	8.0	8.47
6	14.0	9.96	8.10	8.84	8.0	7.04
7	6.0	7.24	6.13	6.08	8.0	5.25
8	4.0	4.26	3.10	5.39	19.0	12.50
9	12.0	10.84	9.13	8.15	8.0	5.56
10	7.0	4.82	7.26	6.42	8.0	7.91
11	5.0	5.68	4.74	5.73	8.0	6.89

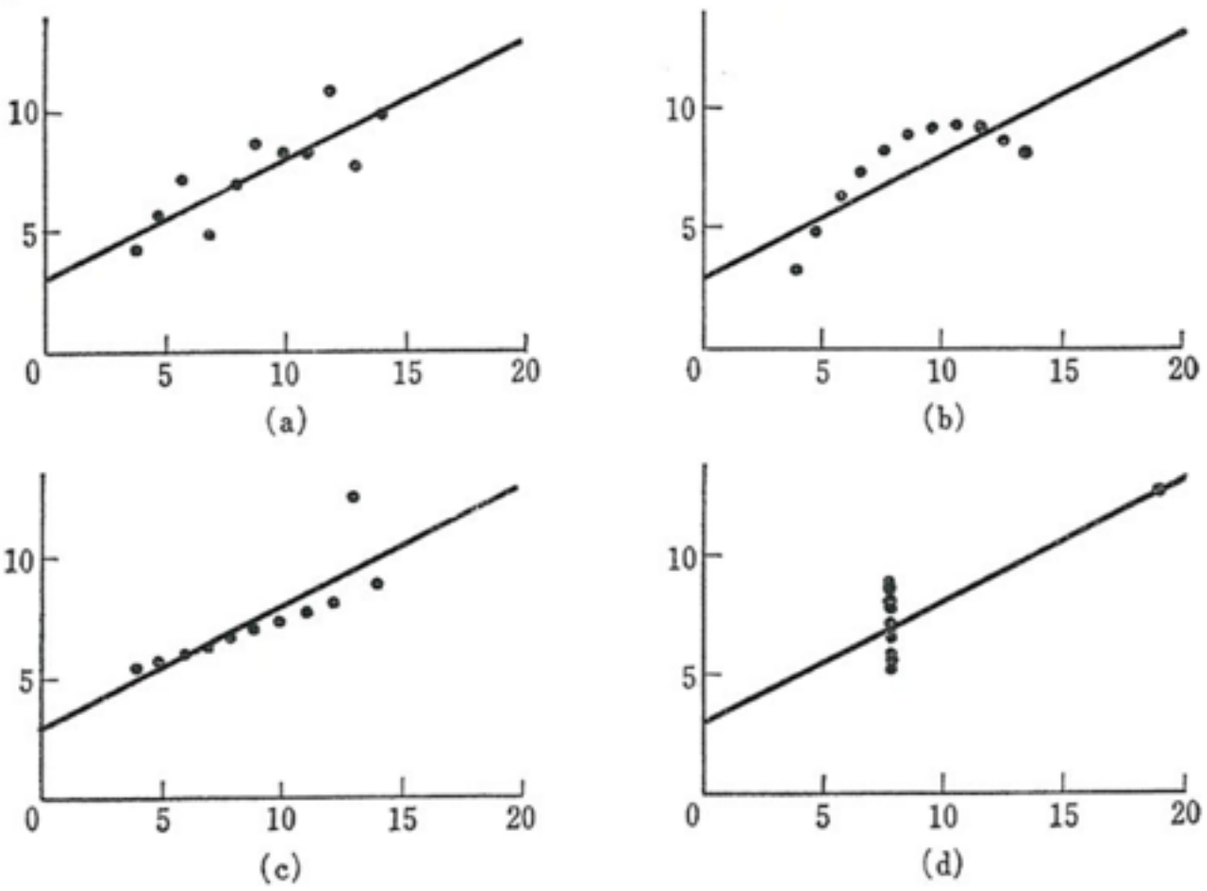


図 6.4　アンスコムの数値例

(a)～(d)のデータの推定結果はすべて同じになる!

$$\hat{y} = 0.5x + 3.0$$

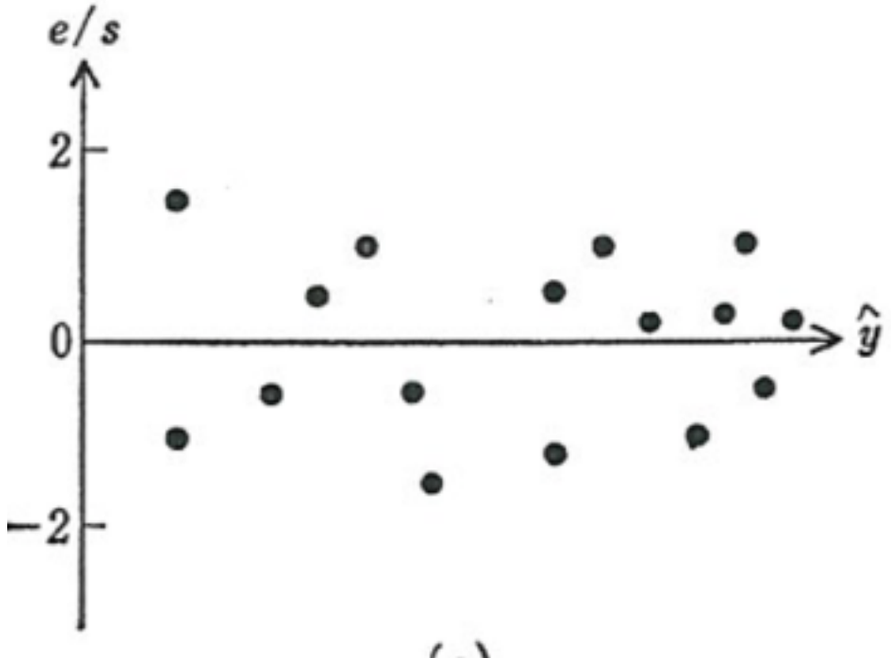
回帰係数の標準誤差 0.118

決定係数 0.677

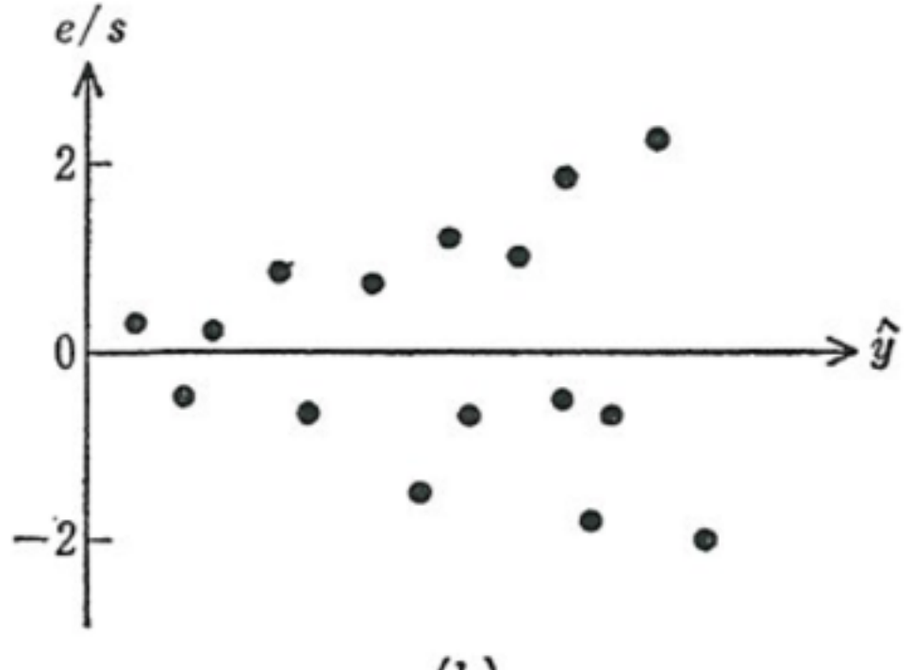
$$\bar{x} = 9.0, \bar{y} = 7.5$$

$$\sum_i (x_i - \bar{x})^2 = 110.0, \sum_i (y_i - \bar{y})^2 = 41.25$$

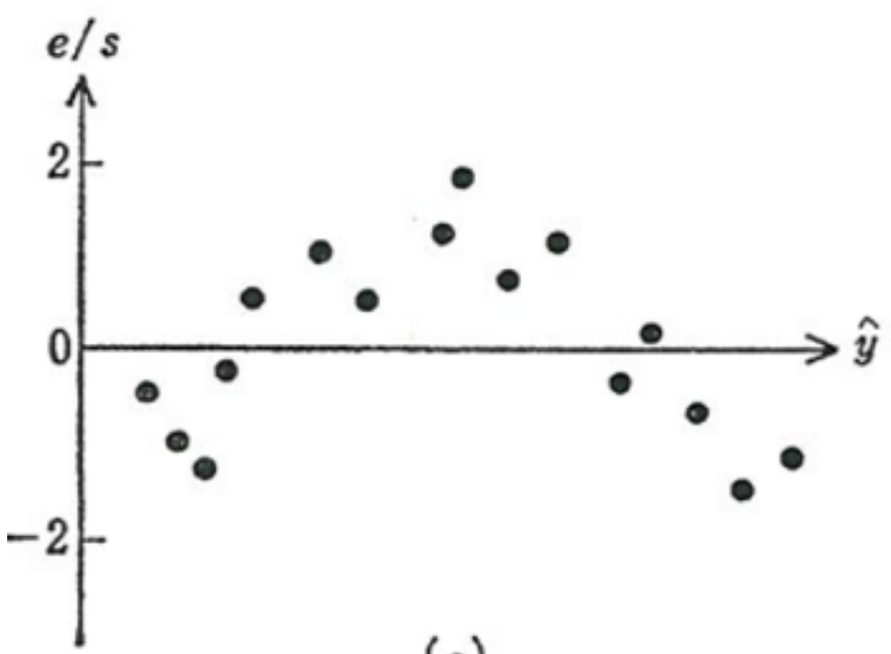
F. J. Anscombe, *The American Statistician*, 27(1), 1973, pp.17-21.



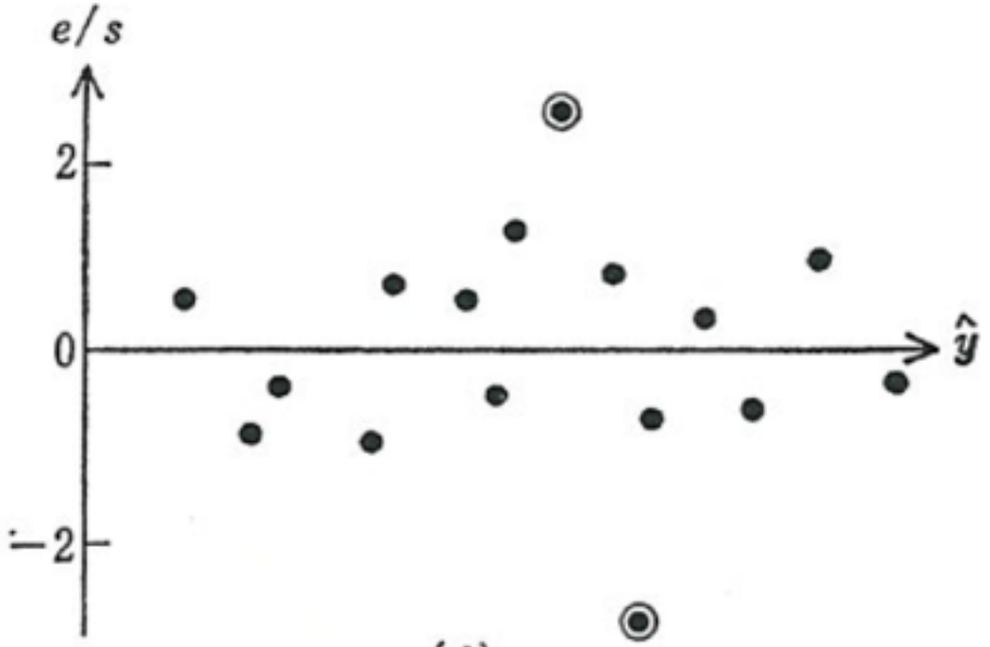
(a)



(b)



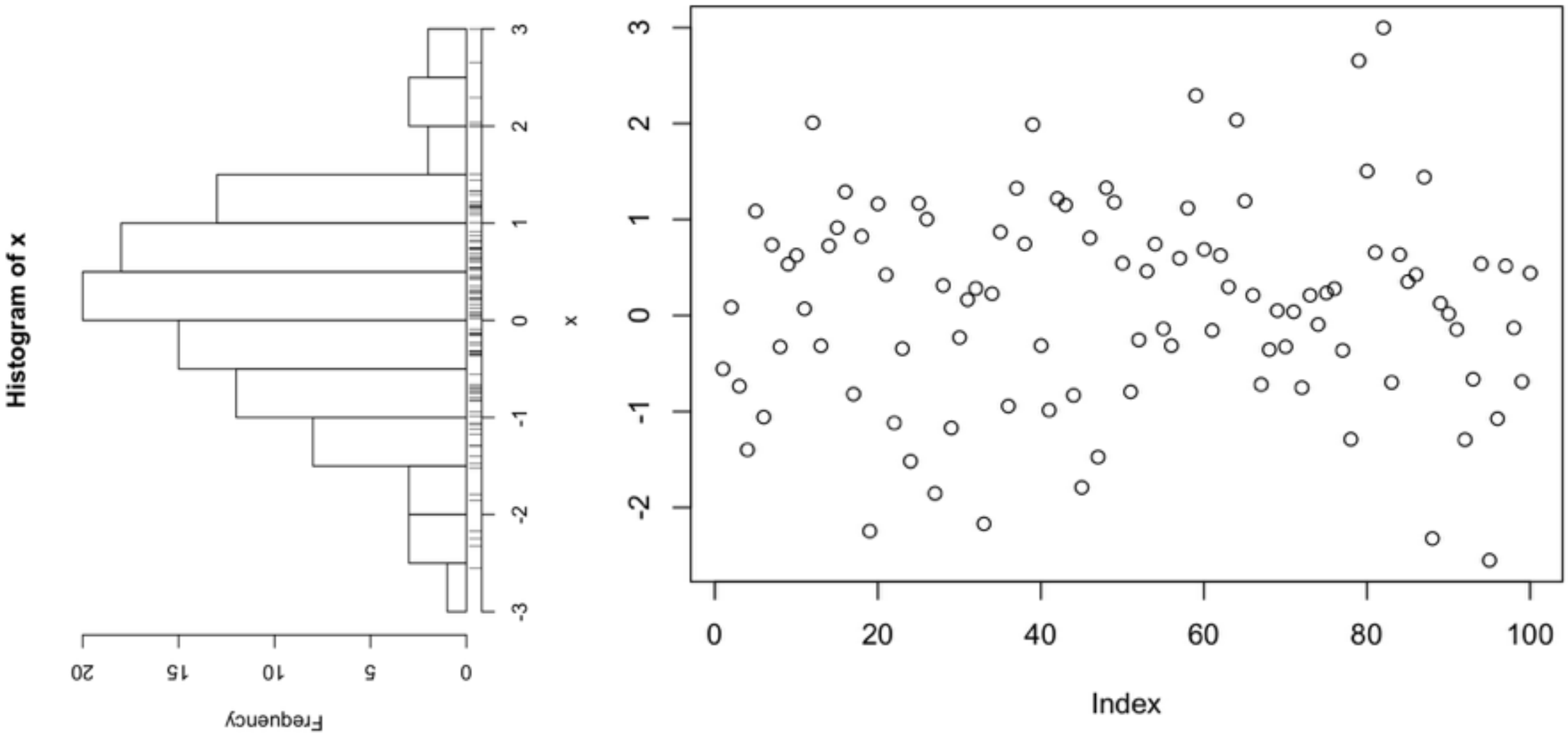
(c)

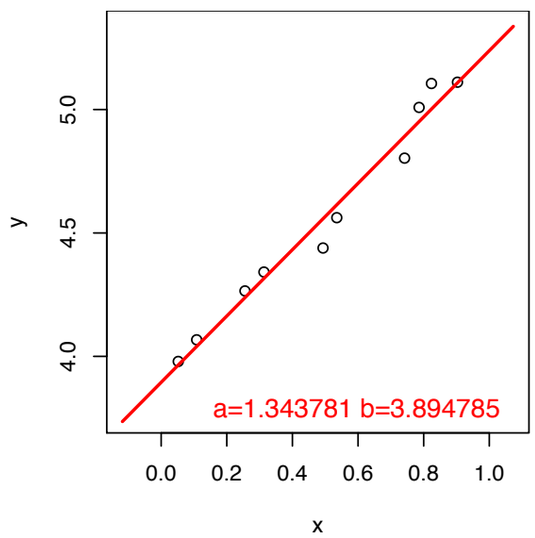


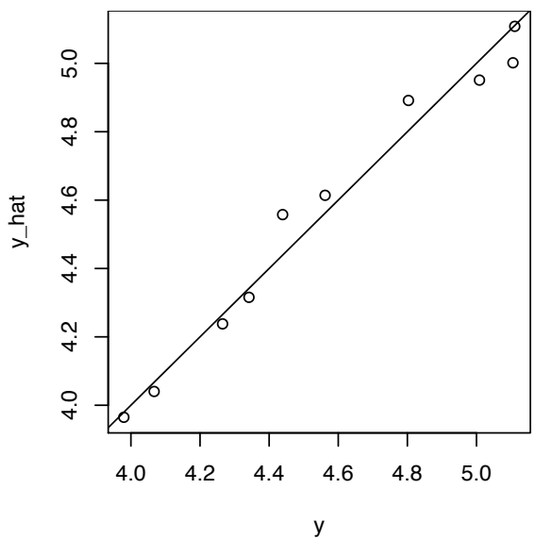
(d)

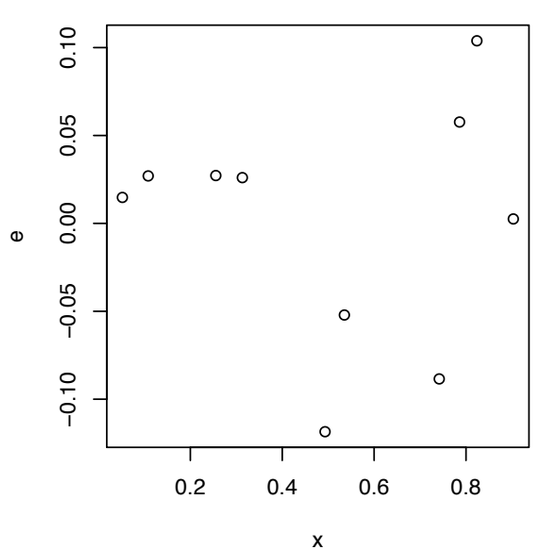
図 6.3 標準化された残差のプロット
（a） 正常な場合 （b） 誤差分散の増大傾向が認められる場合
（c） 非線形の可能性が認められる場合 （d） 異常値のある場合

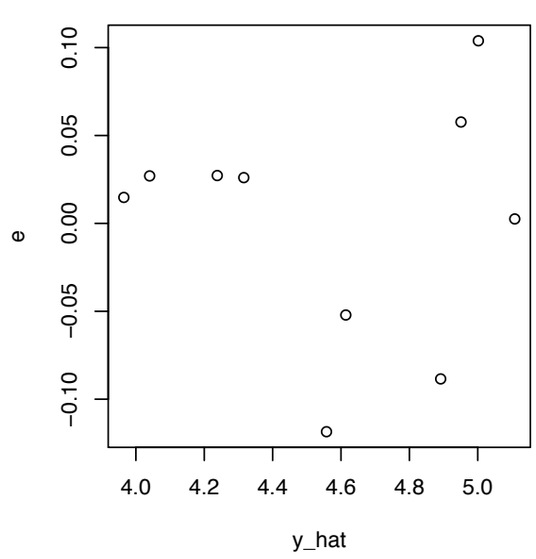
iid正規乱数の残差プロット

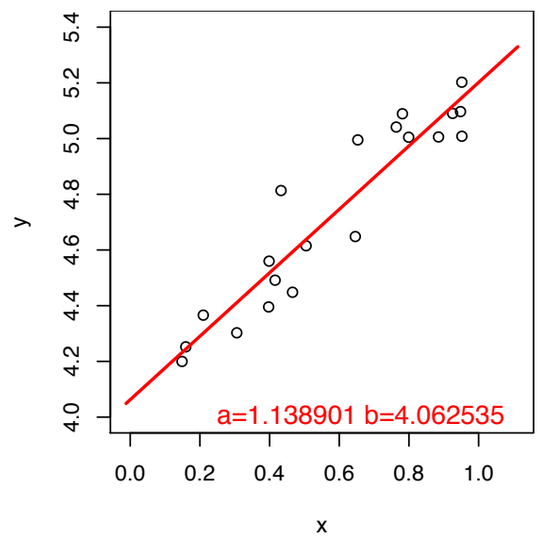


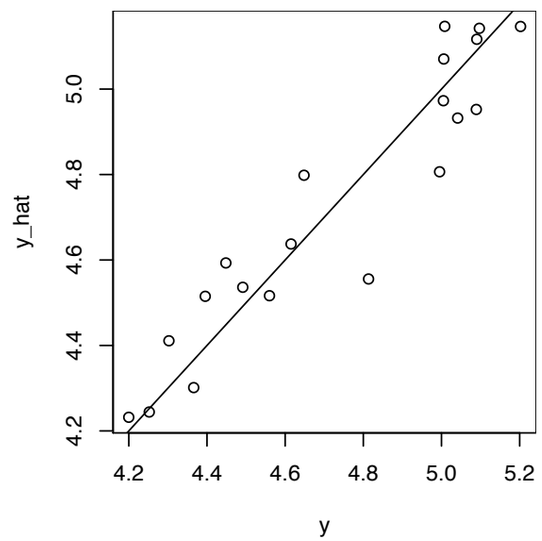

a=1.343781 b=3.894785
y
x

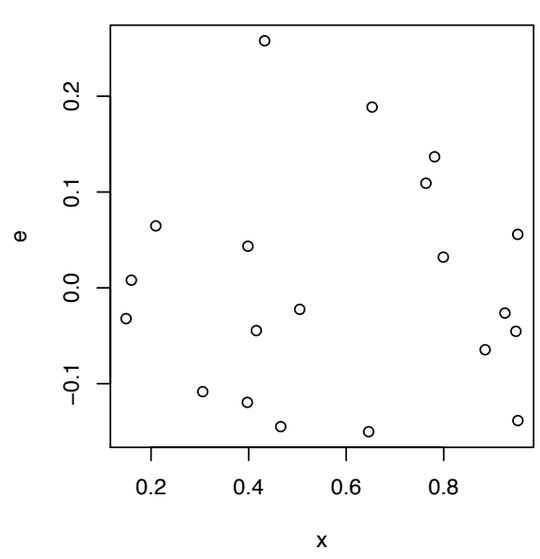

y_hat
y

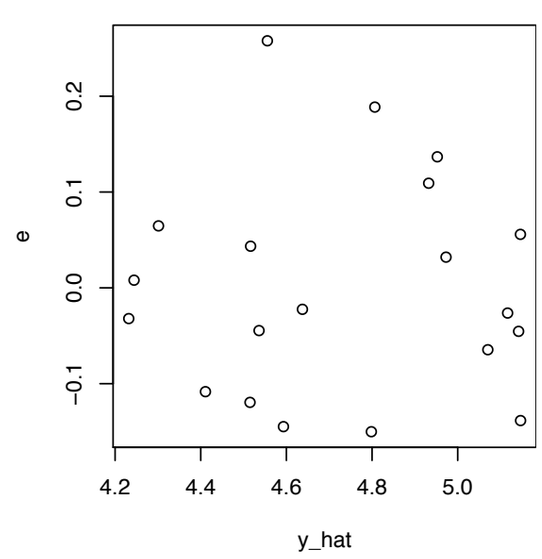

e
x

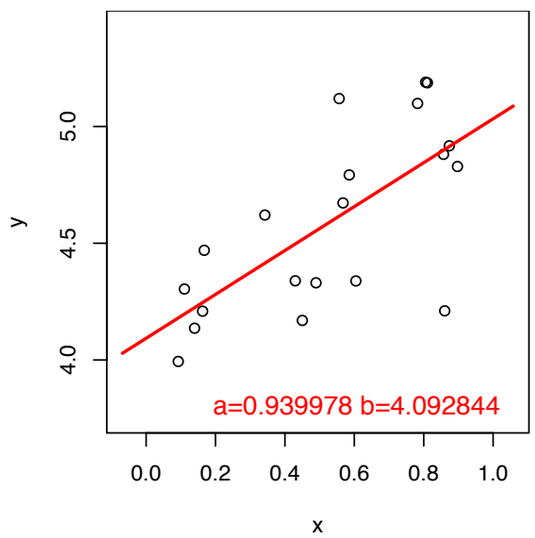

e
y_hat

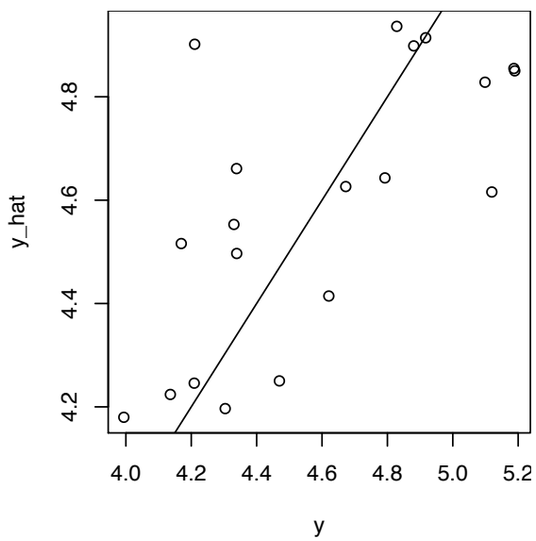

a=1.138901 b=4.062535
y
x

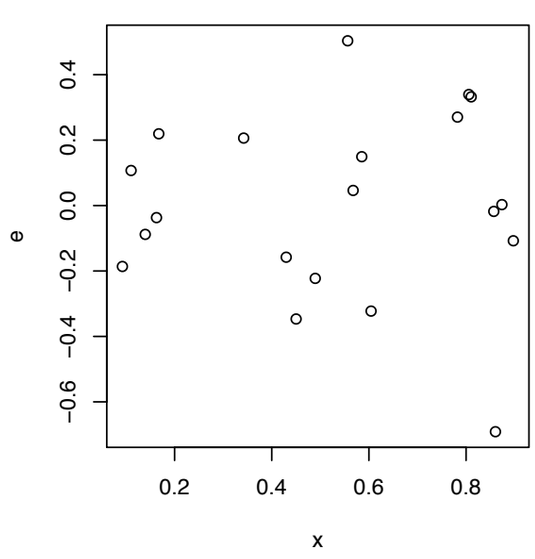

y_hat
y

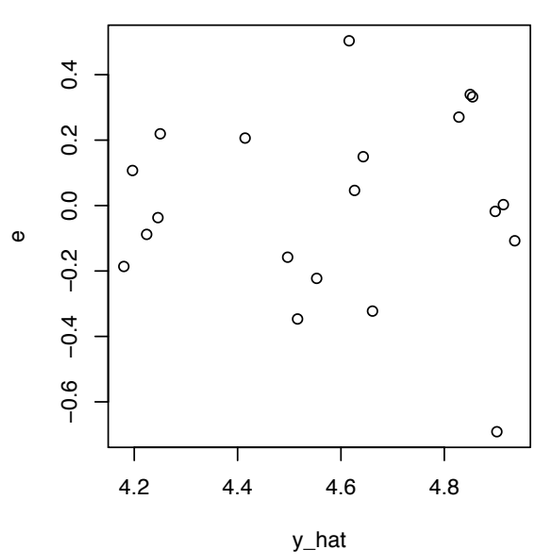

e
x

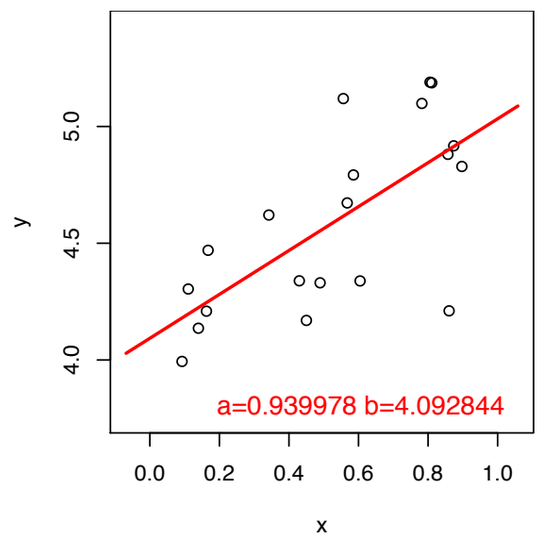

e
y_hat

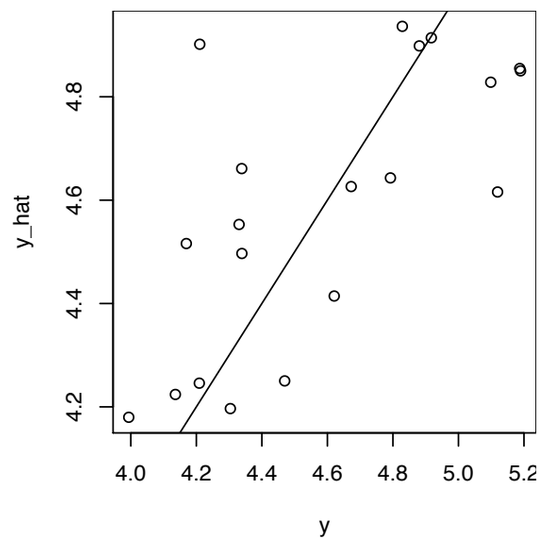

a=0.939978 b=4.092844
y
x

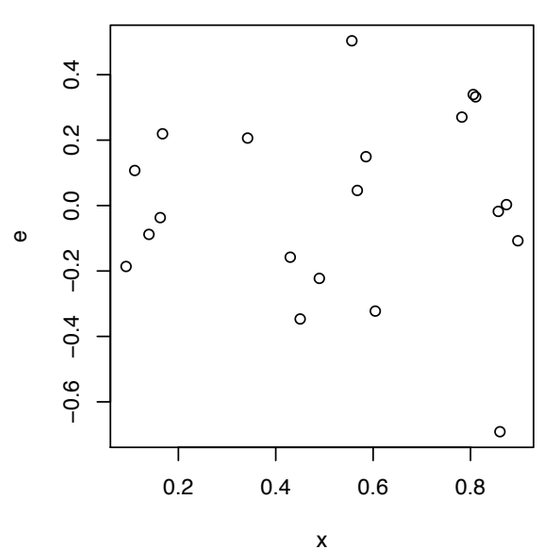

y_hat
y

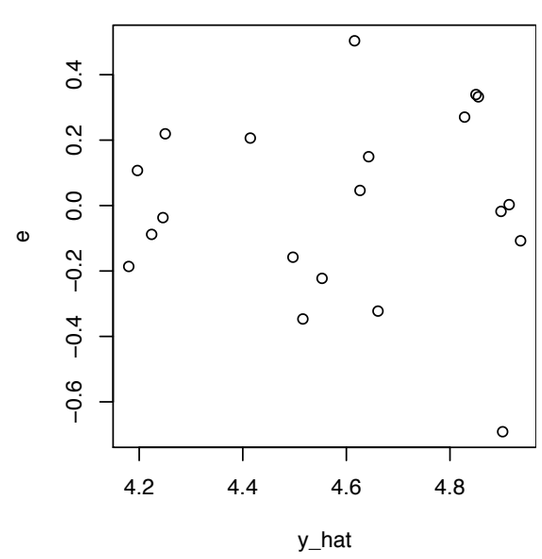

e
x

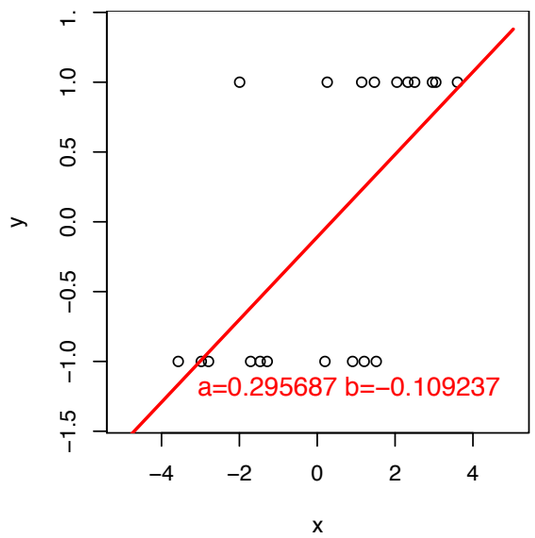

e
y_hat

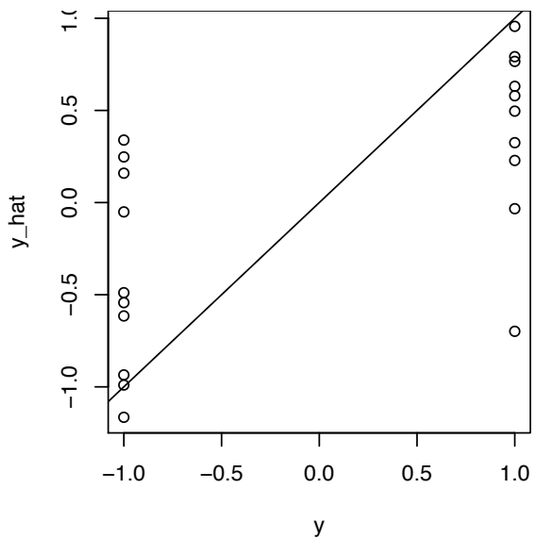

a=0.939978 b=4.092844
y
x

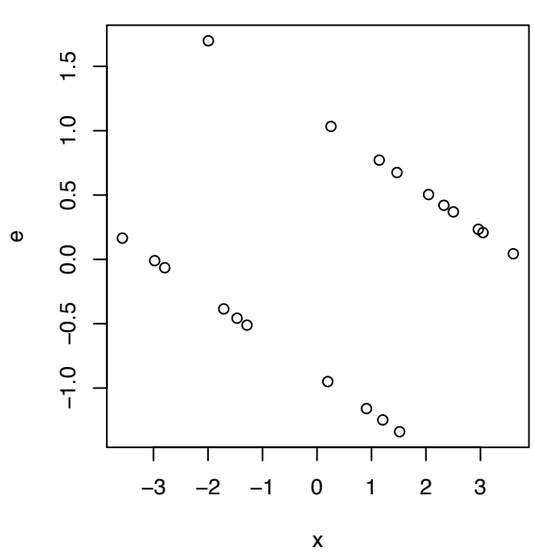

y_hat
y

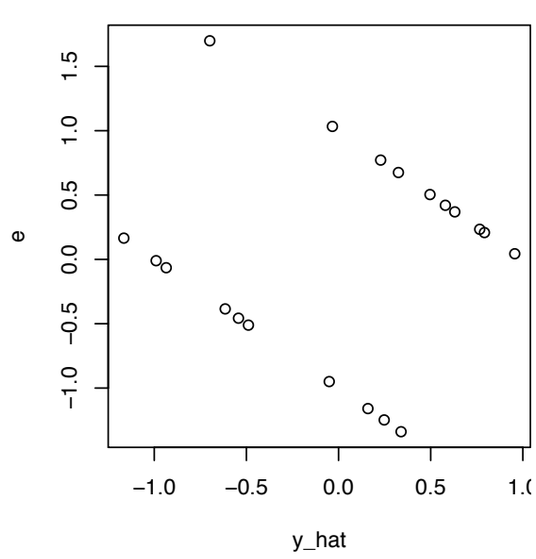

e
x

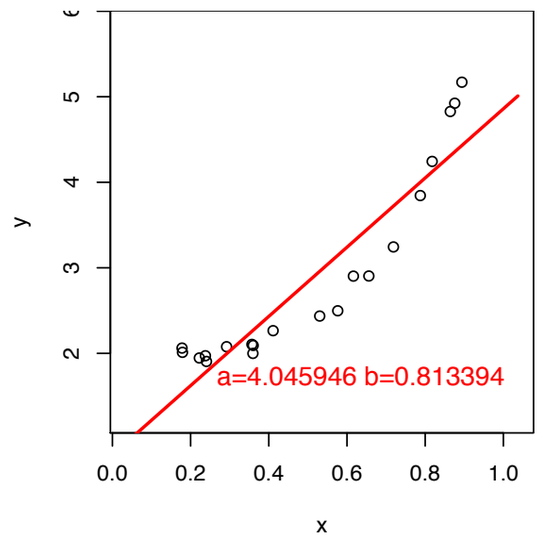

e
y_hat

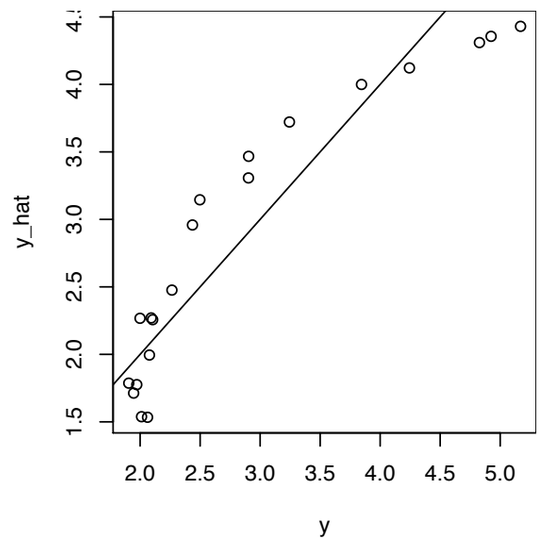

a=0.295687 b=−0.109237
x
y

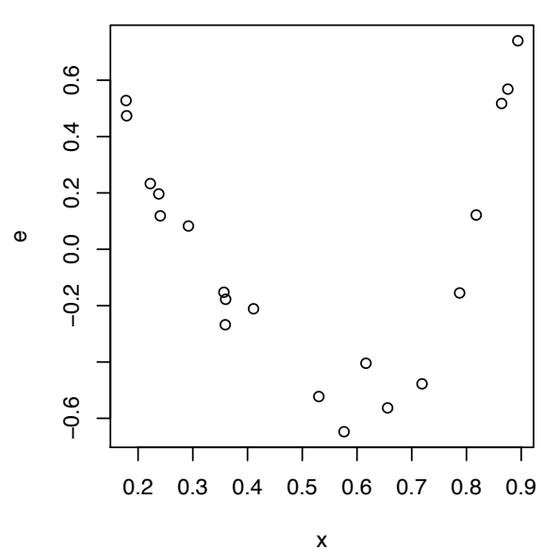

y
y_hat

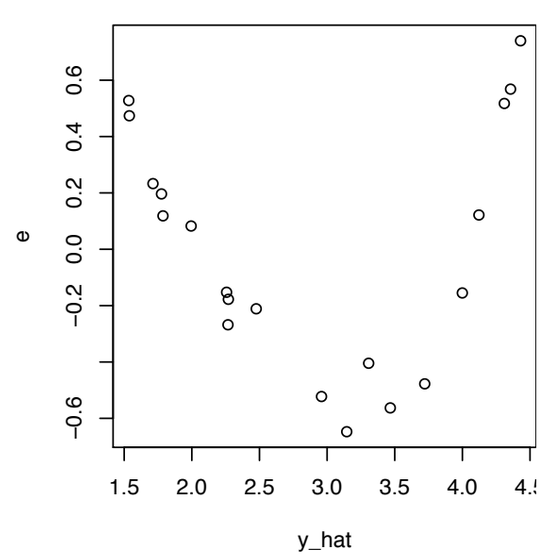

x
e

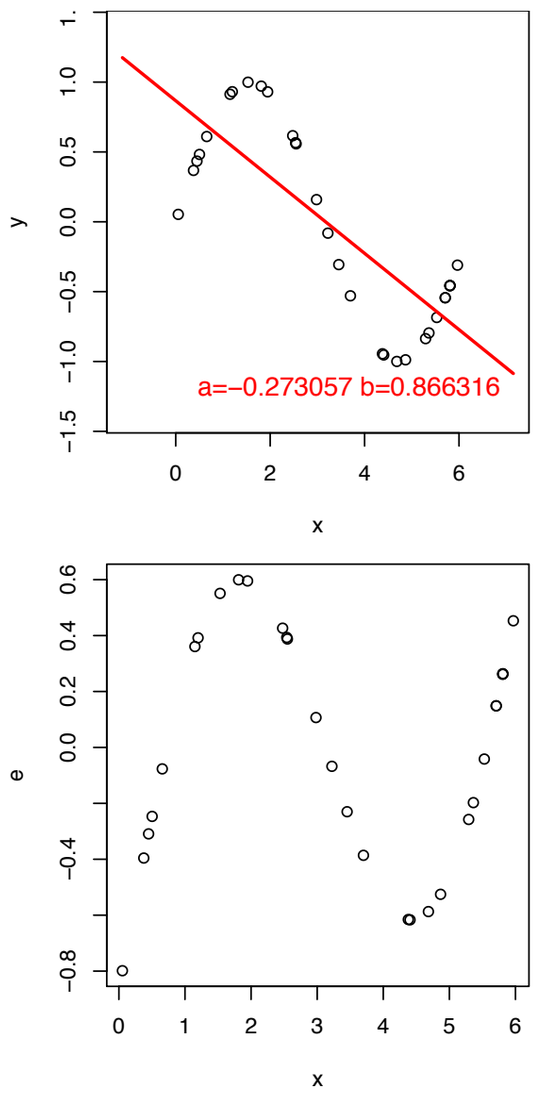

y_hat
e

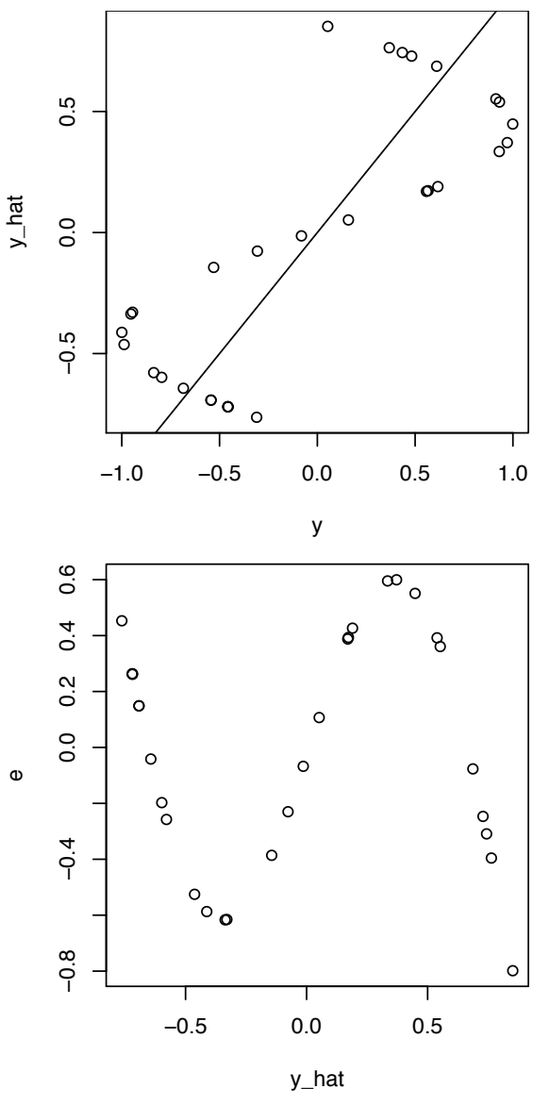

a=4.045946 b=0.813394
x
y

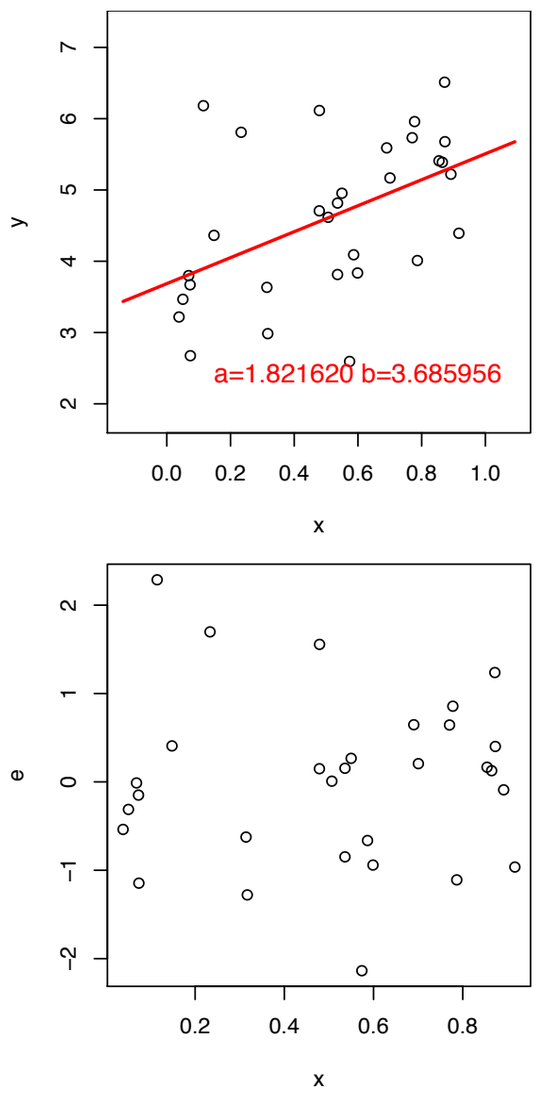

y
y_hat

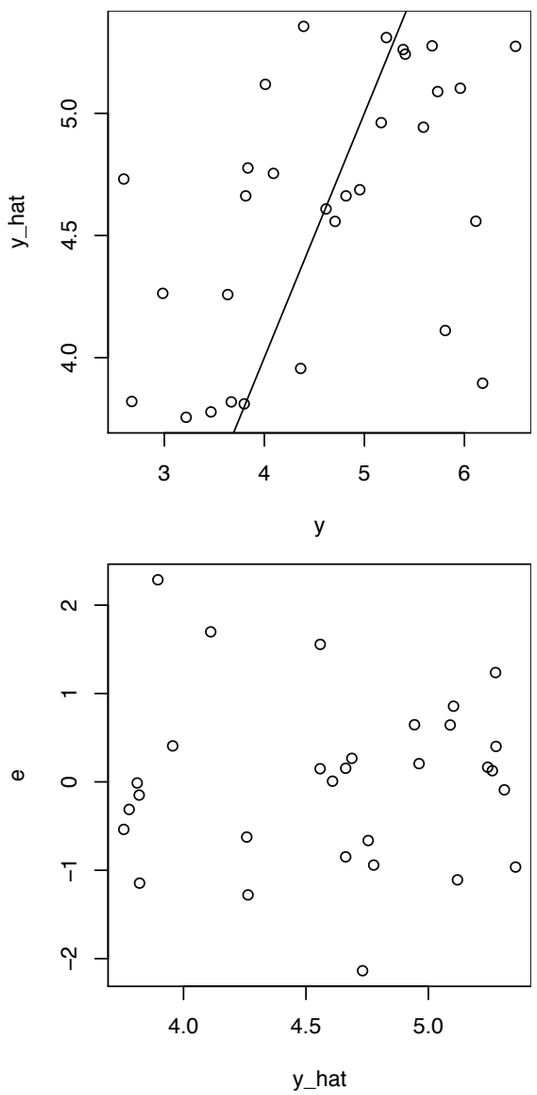

x
e


y_hat
e


a=−0.273057 b=0.866316
y
x


y_hat
y


a=1.821620 b=3.685956
y
x

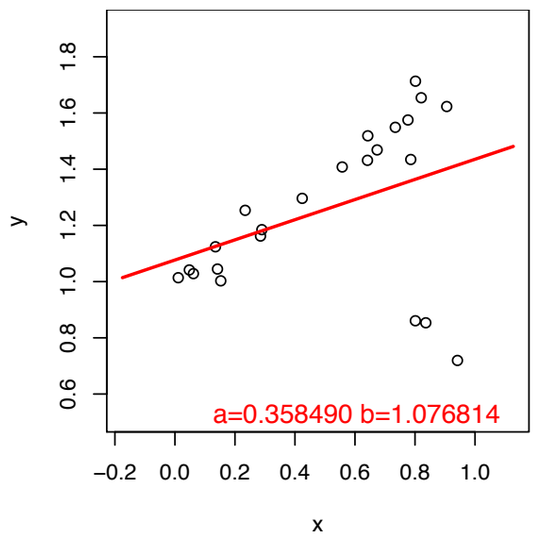

y_hat
y

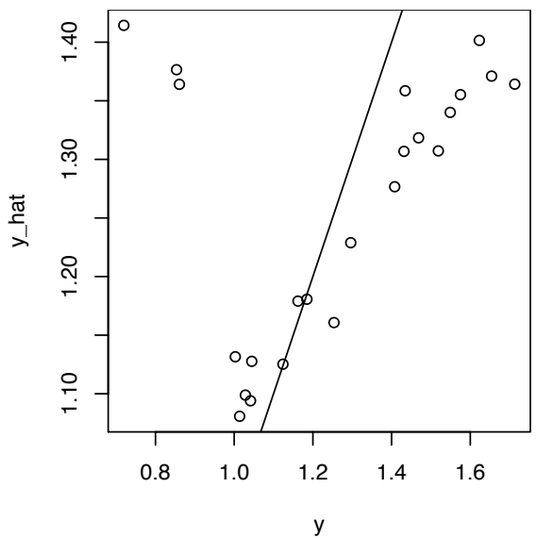
e
x

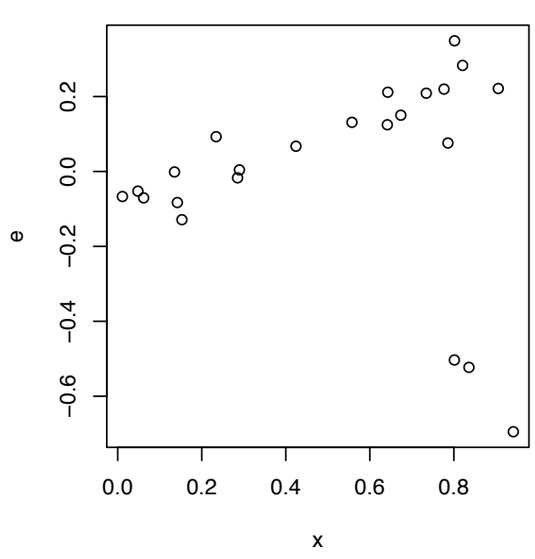
e
y_hat

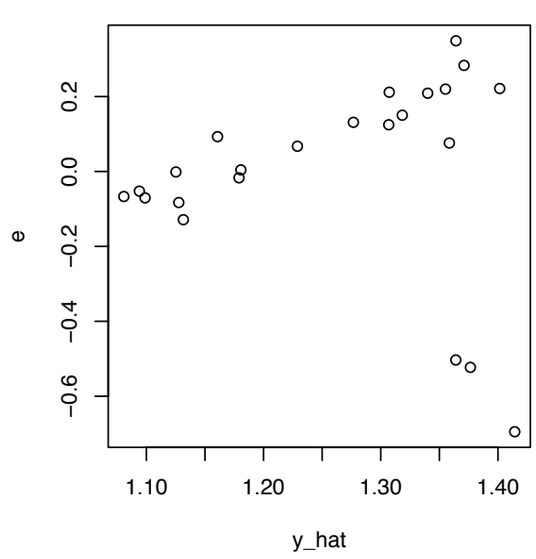
e
x

e
y_hat


y
x
a=0.358490 b=1.076814
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
0.6
0.8
1.0
1.2
1.4
1.6
1.8


y_hat
y
0.8
1.0
1.2
1.4
1.6
1.10
1.20
1.30
1.40


e
x
0.0
0.2
0.4
0.6
0.8
−0.6
−0.4
−0.2
0.0
0.2


e
y_hat
1.10
1.20
1.30
1.40
−0.6
−0.4
−0.2
0.0
0.2