Essential Mathematics for Global Leaders I

Statistics Spring 2019

Lecture 12: 2019 July 29

PART II: Statistical inference

6: Simple Linear Regression

（単）線形回帰

6.1 Introduction p.2

6.2 Estimation for simple linear regression p.13

6.3 Confidence Intervals for a and b p. 19

6.4 Prediction interval p. 23

6.5 Relationship with correlation

6.5 Relationship with correlation

- **Regression:** Does one family of data vary with respect to another family according to a “random” model function ?
  帰無回帰：データの第一属は、モデル関数にそって
  データの第二属によって変動するのか。
  - Can be used for prediction

- **Correlation:** Measure how much two random variables on two populations are “linearly” independent.
  相関関係：二つの母集団上の確率変数間の線形関係の
  度合いを示す指標である。

- Correlation and linear regression are strongly related.

A word on (discrete) joint distribution
同時（離散）確率分布の一言で

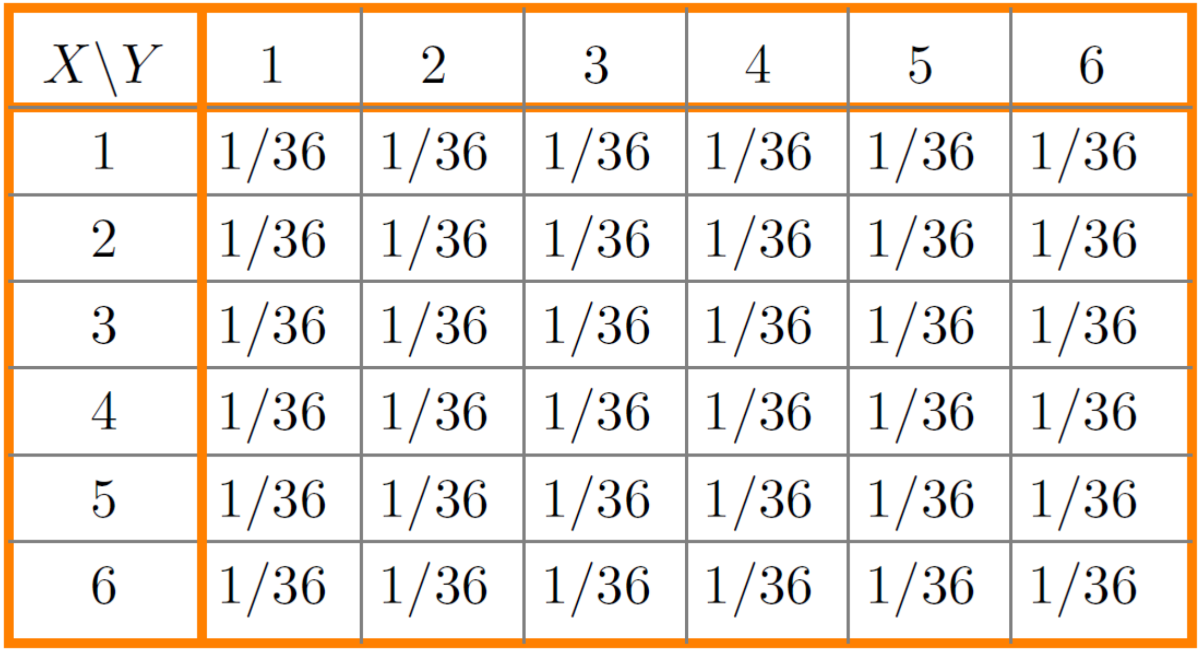
- Random variable X on event space $\Omega = \{\omega_1, \dots, \omega_n\}$.
- Random variable Y on event space $\Gamma = \{\gamma_1, \dots, \gamma_r\}$.
- Product even space: $\Omega \times \Gamma$ is represented by a table.
  The events are all couples $(\omega_i, \gamma_j) = (\ \omega_i$ and $\gamma_j\ )$

$\Gamma \downarrow$	$\Omega \rightarrow$	ω_1	ω_2	$\dots$	$\dots$	ω_n
γ_1		(ω_1, γ_1)	(ω_2, γ_1)			(ω_n, γ_1)
$\vdots$		$\vdots$				$\vdots$
γ_r		(ω_1, γ_r)	$\dots$	$\dots$	$\dots$	(ω_n, γ_r)

- Joint distribution of X and Y ↔ assign probabilities to all events (ω_i, γ_j)

Joint probability mass function: example

- Roll two dice: X is the number on the first die.
- Y is the number on the second die.
- The joint probability table of X and Y is:



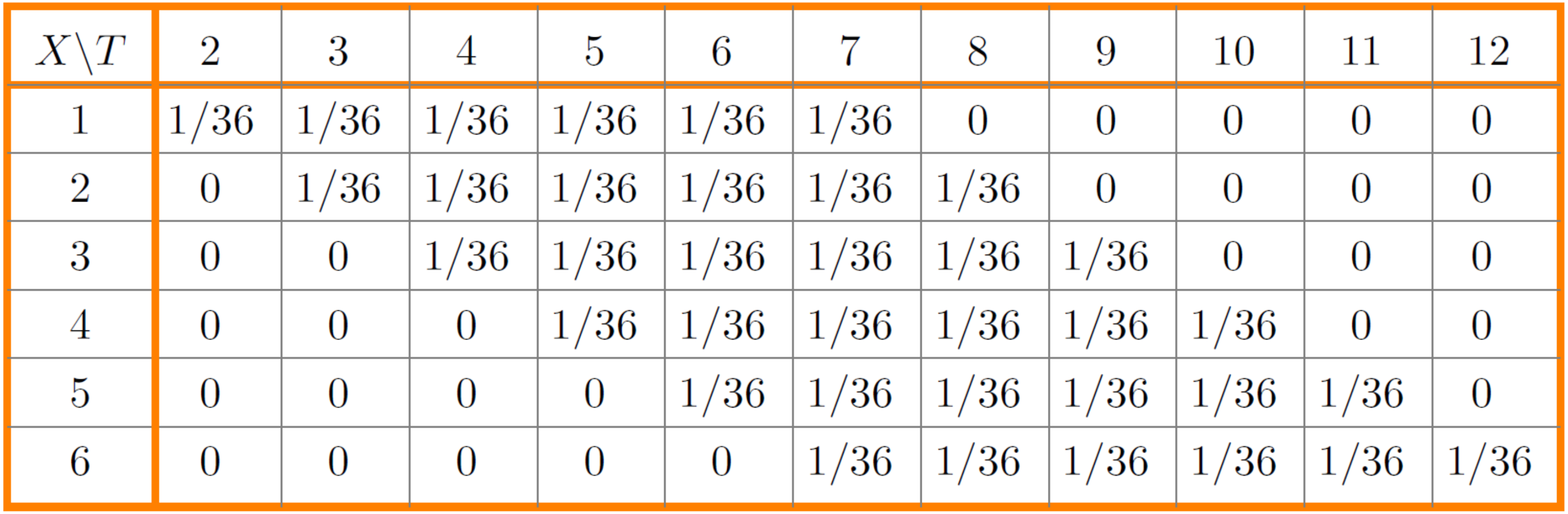
$X\backslash Y$	1	2	3	4	5	6
1	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$
2	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$
3	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$
4	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$
5	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$
6	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$

- $p(i,j) = 1/36$ for all i and j.

$\sum_i \sum_j p(i,j) = 1$

Joint probability mass function: example II

- X is the number on the first die
- T is the total sum of the two dices



$X\backslash T$	2	3	4	5	6	7	8	9	10	11	12
1	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	0	0
2	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	0
3	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0
4	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0
5	0	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0
6	0	0	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36

- Again $\sum_i \sum_j p(i,j) = 1$

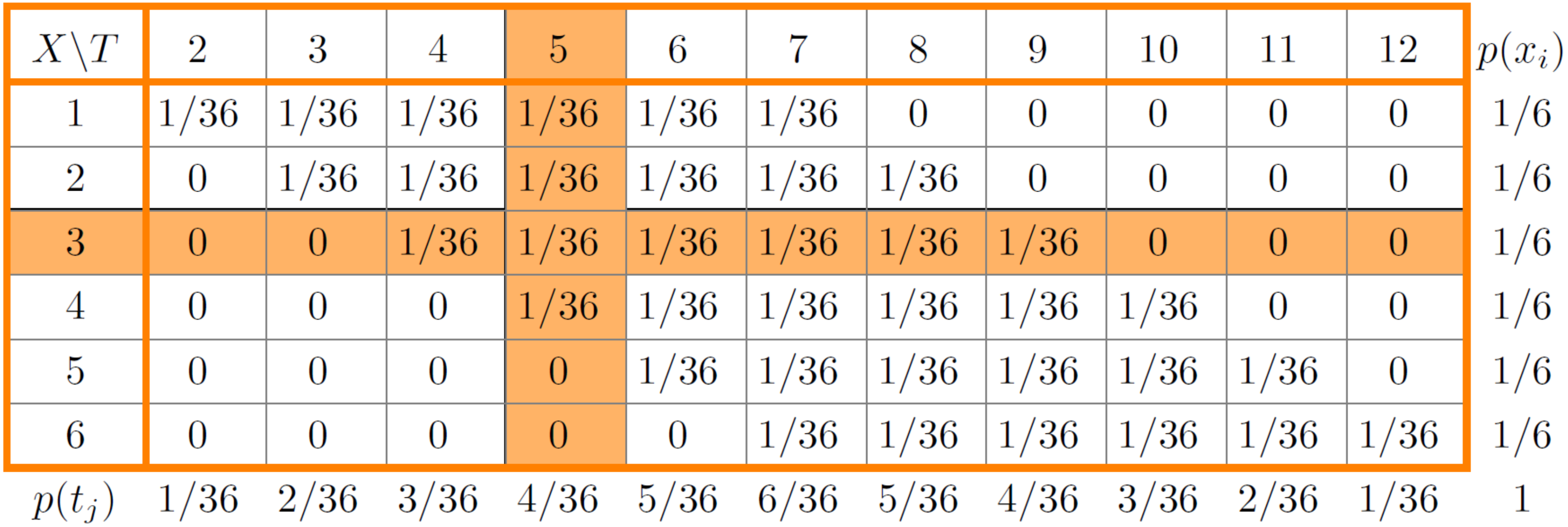
Marginal prob. mass function 周辺分布

- X is the number on the first die
- T is the total sum of the two dices
- Marginal pmf of X: 行の和 sum the rows. $p_X(x)$
- Marginal pmf of T: 列の和 sum the columns $p_T(t)$

$X \backslash T$	2	3	4	5	6	7	8	9	10	11	12	$p(x_i)$
1	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	0	0	0	0	0	$1/6$
2	0	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	0	0	0	0	$1/6$
3	0	0	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	0	0	0	$1/6$
4	0	0	0	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	0	0	$1/6$
5	0	0	0	0	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	0	$1/6$
6	0	0	0	0	0	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/6$
$p(t_j)$	$1/36$	$2/36$	$3/36$	$4/36$	$5/36$	$6/36$	$5/36$	$4/36$	$3/36$	$2/36$	$1/36$	1

Independence

- X and Y are independent if $p(x_i, y_j) = p_X(x_i) \cdot p_Y(y_j)$
- Question: Are X ant T independent?



$X \backslash T$	2	3	4	5	6	7	8	9	10	11	12	$p(x_i)$
1	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	0	0	0	0	0	$1/6$
2	0	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	0	0	0	0	$1/6$
3	0	0	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	0	0	0	$1/6$
4	0	0	0	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	0	0	$1/6$
5	0	0	0	0	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	0	$1/6$
6	0	0	0	0	0	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/6$
$p(t_j)$	$1/36$	$2/36$	$3/36$	$4/36$	$5/36$	$6/36$	$5/36$	$4/36$	$3/36$	$2/36$	$1/36$	1

Covariance　共分散

- Measures to which extent two random variables vary together (example: weight and height).
２つの確率変数は同様に変動する度合いを示す指標である（例：体重と身長）
- X, Y random variables on population Ω, Γ respectively.

$$cov(X,Y) := E\big((X-\mu_X)\cdot(Y-\mu_Y)\big)$$

where μ_X is the (population Ω) mean of X and μ_Y is the (population Γ) mean of Y.
ただし、μ_XはXの（母集団の）平均、μ_YはYの（母集団の）平均である。

$$cov(X,Y) = \sum_{\omega}\sum_{\gamma}\big(X(\omega)-E(X)\big)\big(Y(\gamma)-E(Y)\big)P(\omega \text{ and } \gamma)$$

Some handy properties of covariance
共分散の使いやすい性質

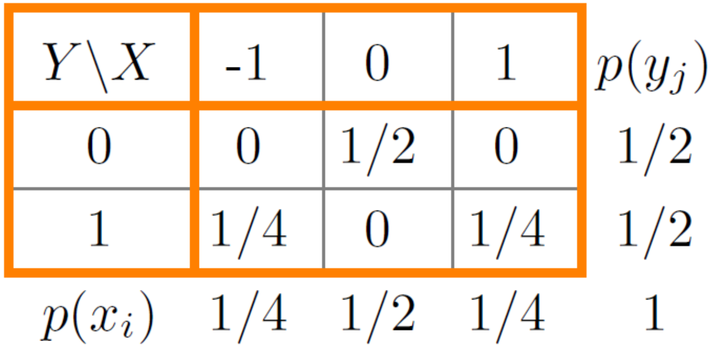
The following properties can be verified easily using the properties of expected value.

以下の性質を、期待値$E(.)$の性質を使って簡単に確認できる。

1. $cov(aX + b, cY + d) = a \cdot\ c \cdot\ cov(X, Y)$
2. $cov(X_1 + X_2, Y) = cov(X_1, Y) + cov(X_2, Y)$
3. $cov(X, X) = Var(X)$
4. $cov(X, Y) = E(XY) - \mu_X \cdot \mu_Y$
5. If X and Y are independent then $cov(X, Y) = 0$
6. !! If $cov(X, Y) = 0$ then X and Y may be dependent !!

Question

- Here is a joint probability table.



$Y\backslash X$	-1	0	1	$p(y_j)$
0	0	$1/2$	0	$1/2$
1	$1/4$	0	$1/4$	$1/2$
$p(x_i)$	$1/4$	$1/2$	$1/4$	1

- Compute

1. $E(X), E(Y)$
2. $E(XY)$
3. Deduce $cov(X,Y) = E(XY) - E(X) \cdot E(Y)$

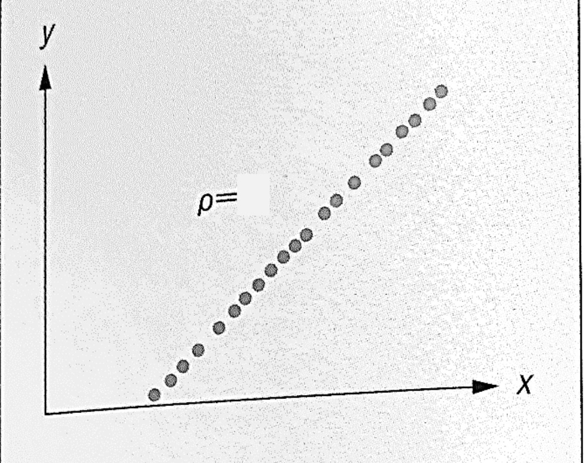
4. Can we say that X and Y are independent?

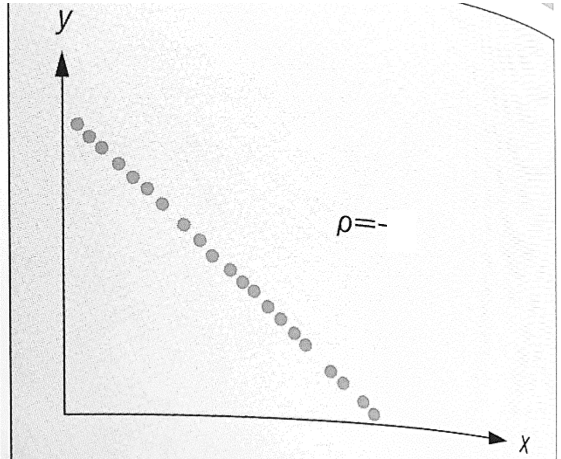
Covariance in linear regression

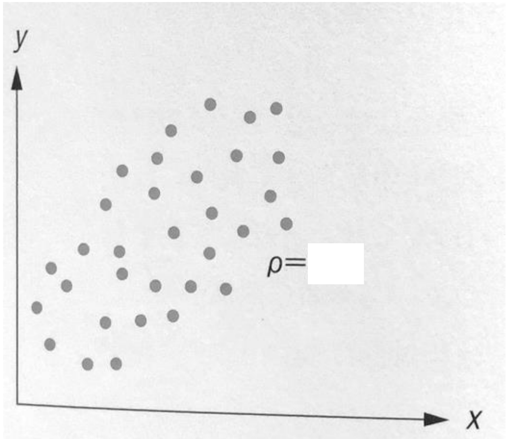
- In linear regression, data are pairs of explanatory/response values:
$$\{(x_1, y_1), \dots, (x_n, y_n)\}$$
- These data define two random variables X and Y and the sample space $\Omega = \{\,1,2,\dots,n\}$ with the uniform distribution: $P(i) = \frac{1}{n}$, by $X(i) = x_i$ and $Y(i) = y_i$.
- Note that if $i \neq j$, $P\big(X = x_i, Y = y_j\big) = 0$ (because no point $\left(x_i, y_j\right)$
- Therefore $P(X = i, Y = i) = \frac{1}{n}$. Finally
- $cov(X,Y) = \frac{1}{n}\sum_i^n \big(X(i) - E(X)\big)\big(Y - E(Y)\big)$ and
- $cov(X,Y) = \frac{1}{n}\sum_i^n (x_i - \bar{x})(y_i - \bar{y})$ (in regression)

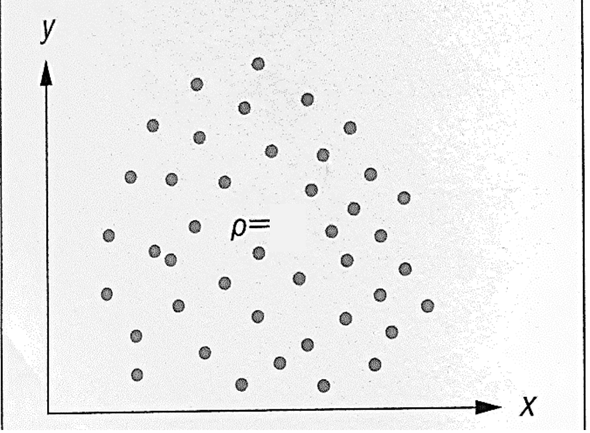
Correlation coefficient （母集団）相関係数

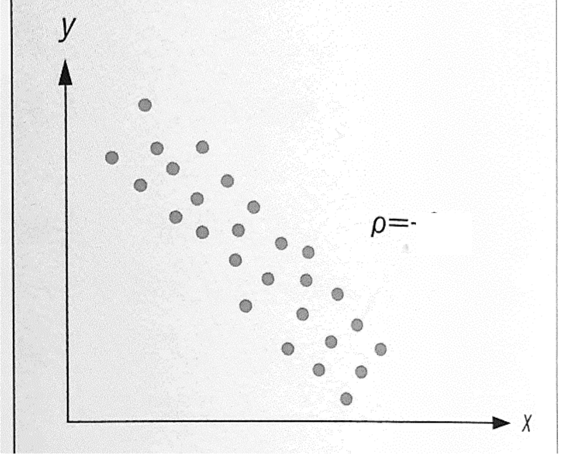
- Covariance does not measure the extent to which 2 random variables are linearly related.
  共分散は２確率変数間の線形関係の度合いを示す指標ではない。
  - ☛ For that, we divide by the standard deviations of X and Y .
  - ☛そのため、XとYの標準偏差で割り切る。
- $\rho = \rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{E\left((X-\mu_x)(Y-\mu_Y)\right)}{\sigma_X \cdot \sigma_Y}$
  - $-1 \leq \rho \leq 1$
  - $\rho = 1$ if and only if $Y = aX + b$ with $a > 0$ and
  - $\rho = -1$ if and only if $Y = aX + b$ with $a < 0$.

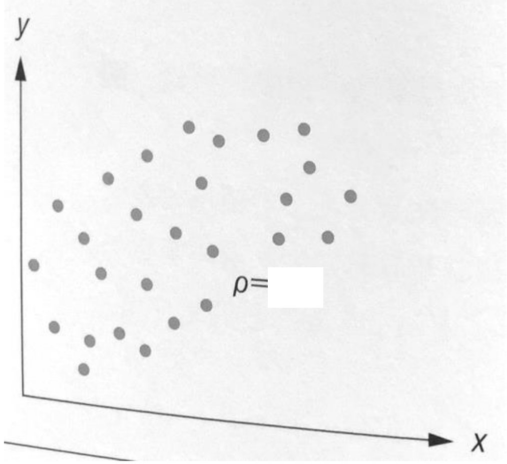


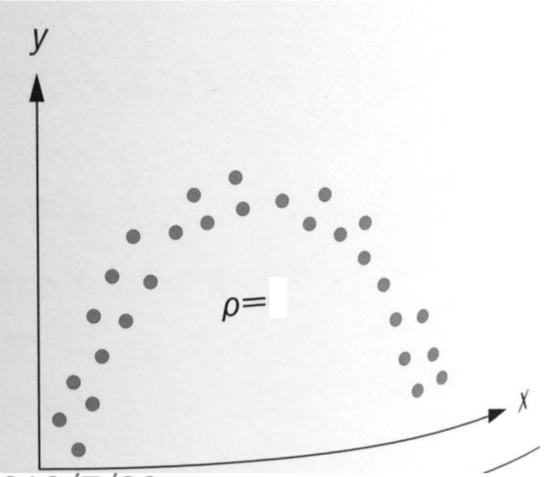




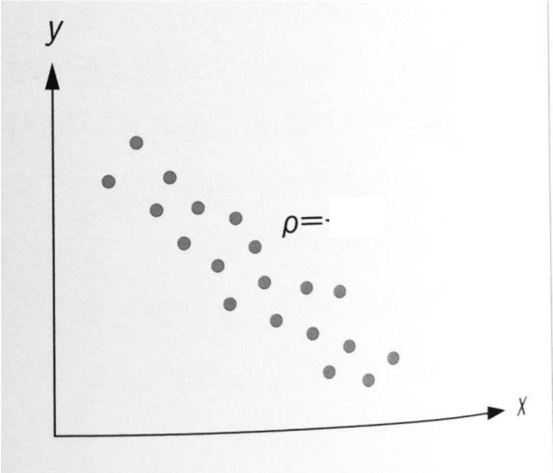








Attribute the correlation coefficient to the corresponding images:
0 , 1 , -0.8 , -0.9 , 0.4 , 0.6 , -1



Point estimator for the population correlation 母相関係数の点推定量：標本相関係数

- Remember the quantities Lecture 11 p.16.
- $S_{xx} = \sum_k (x_k - \bar{x})^2$ $\qquad$ $S_{xY} = \sum_k (x_k - \bar{x})(Y_k - \bar{Y})$
- $cov(X,Y)$の推定量は $S_{xy}/(n-1)$ ← estimator
- σ_Xの推定量は$\sqrt{S_{xx}/(n-1)}$ ← estimator
- σ_Yの推定量は$\sqrt{S_{yy}/(n-1)}$ ← estimator
- 従って、$\rho = cov(X,Y)/\sigma_X \cdot \sigma_Y$の推定量：

$$\boxed{r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}}$$

is called Sample Correlation coefficient 標本相関係数

And is an estimator for the population’s covariance.

Test if X and Y are not linearly correlated XとY間に線形関係があるかを調べる

- Assumption 想定 :
  - Both population are normally distributed.
  - 量母集団は正規分布に従う
- $H_0 : \rho = 0$ 帰無仮説 Null Hypothesis
- $H_\mathrm{A} : \rho \neq 0$ 対立検定 Alternative Hypothesis

Theorem: $T = \frac{r}{\sqrt{(1-r^2)/(n-2)}} \sim t_{n-2}$ Student distribution

- **Reject $\boldsymbol{H_0}$** at significance level α if $|T| \geq c = t_{n-2}\left(1 - \frac{\alpha}{2}\right)$

- **Rmk:** Cannot be used to test $H_0 : \rho = \rho_0$ for some $\rho_0 \neq 0$ (Cannot be used neither for a confidence interval).

Practice problem

販売交渉に費やされた時間と利益の間に線形関係があるかどうかを調べる調査が行われた。無作為に２７市場取引データを標本抽出し、各々の取引の売約までに費やされた時間と利益が記録された。
標本相関係数は$r = 0.424$と計算された。
☛交渉の長さと利益の間に線形関係があるといえるか。

A survey to determine if the relation between the time spent for negotiations and the benefit of the transaction is linear, has been conducted randomly over 27 market transaction's data. For each transaction is recorded the time spent until the sale and the benefit.

The sample correlation coefficient computed is $r = 0.424$.

☛Can we say that the relation between the length of the negotiations and the benefit is linear?

Correlation does not imply causation
相関は因果性を含意しない

- Over time, amount of ice cream consumption is correlated with number of pool drownings.
  やがて、アイスクリームの摂取量と水泳プールで溺れて死んだ人数に相関関係にある。
- In 90% of bar fights ending in a death the person who started the fight died.
  死を及ぼした酒場でのけんかのなか、けんかを始めた人が死んだ場合の率は９割である。
- In a 1685 study (and today !) being a student is the most dangerous profession.
  1685年（今でも！）の調査によると、学生は最も危険な職業である。

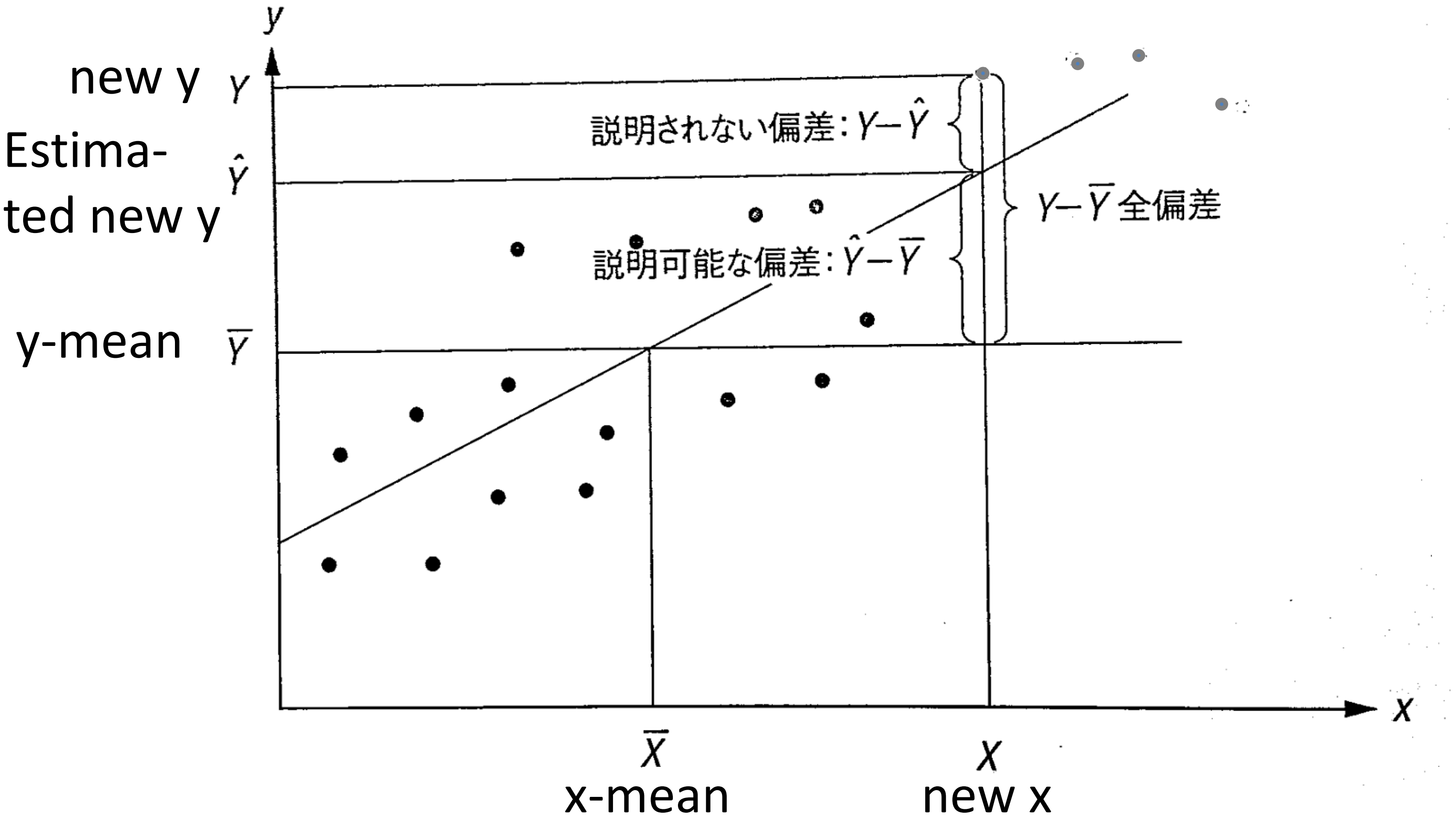
“Covariation is a necessary but not sufficient condition for causality “

Goodness of fit of the linear regression
回帰の当てはまりのよさ

- Aim: Is the regression line $y = \hat{a}x + \hat{b}$ found good ?
- 求めた回帰直線の適度性はどう？
- Need a parameter value that measures relatively the degree of variation around the line
  回帰直線のまわりのデータの変動の度合いを相対的に測ることができる尺度が必要。

Definition (coefficient of determination 決定係数)
r^2 square of the sample correlation coefficient
標本相関関数の２乗

回帰関係の強さ、つまりどのくらい回帰直線がデータに適合しているかを示す指標。



new y

Estima-
ted new y

y-mean

x-mean

new x

Residuals and computation of the r^2
残差と決定係数の計算

- Data: $(x_1, y_1), \dots, (x_n, y_n)$
- $y_k - \hat{y} = y_k - \hat{a}x_k - \hat{b}$ for $k = 1, \dots, n$ = true value – estimated value by regression are called residuals 残差。
- $y - \bar{y} \quad = \quad y - \hat{y} \quad + \quad \hat{y} - \bar{y}$

  Total deviation= residuals + inferred by regression

  線偏差 ＝ 残差 ＋ 回帰直線による推測

$$\sum_{k=1}^{n}(y_k - \bar{y})^2 = \sum_{k=1}^{n}(y_k - \widehat{y_k})^2 + \sum_{k=1}^{n}(\widehat{y_k} - \bar{y})^2$$

$$SST = SSE + SSR$$

全平方和 ＝ 残差平方和 ＋ 回帰平方和

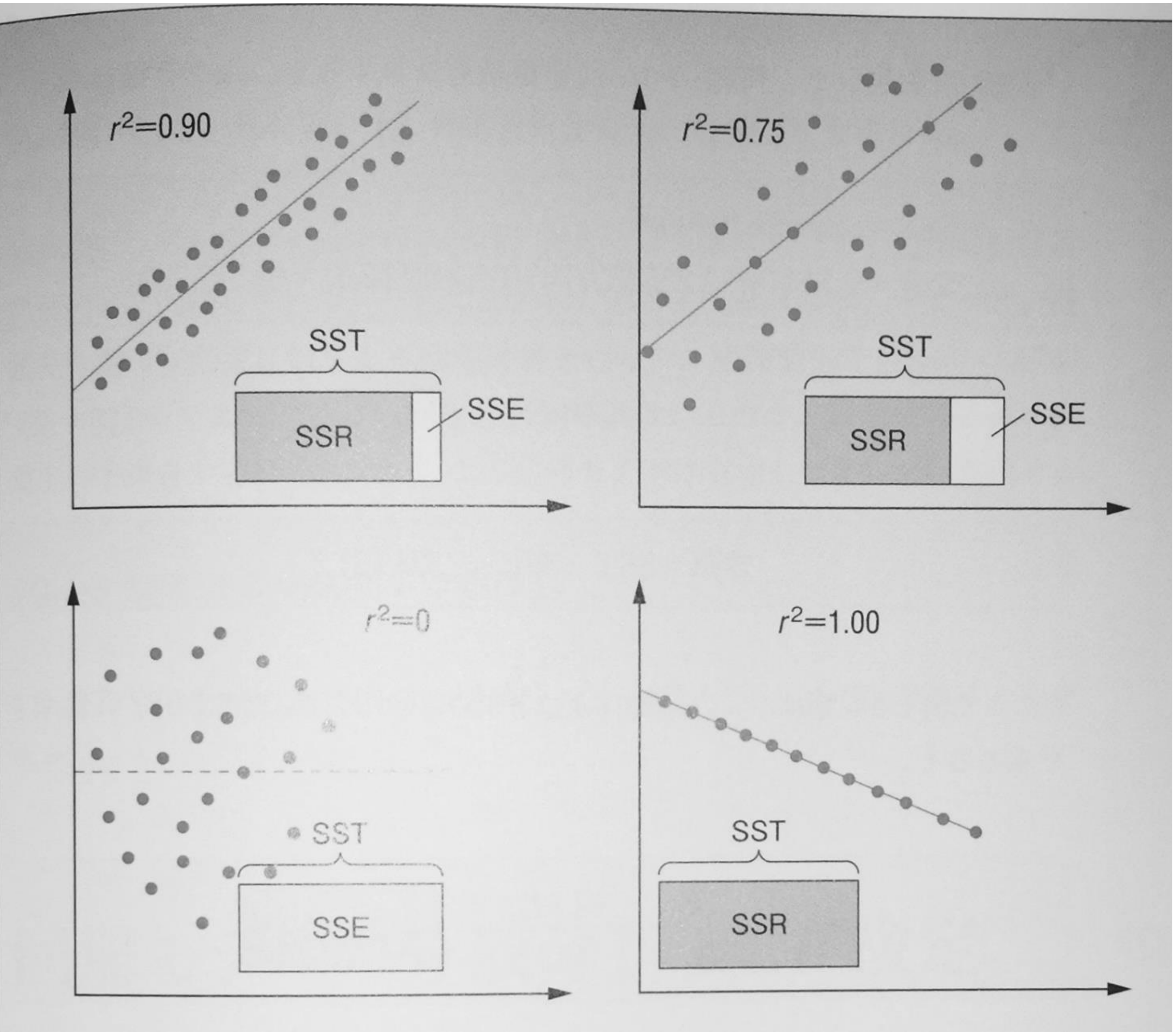
Total sum of ■ = error sum of ■ + regression sum of ■

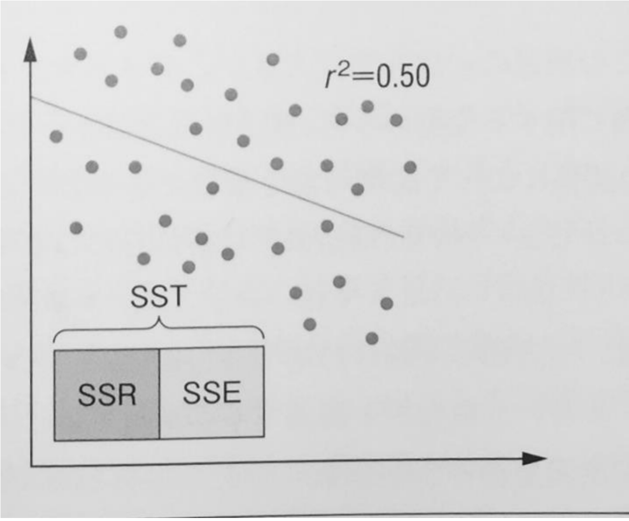
Formula for determination coefficient
決定係数を計算するために公式

- $SSE = \sum_{k=1}^{n} (y_k - \widehat{y_k})^2 = \sum_{k=1}^{n} \left(y_k - \hat{a} - \hat{b} x_k\right)^2$
$$= S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$
- $SST = S_{yy}$
- $SSR = SST - SSE = \frac{S_{xy}^2}{S_{xx}}$

$$r^2 = 1 - \frac{SSE}{SST} = \frac{S_{xy}^2}{S_{xx} S_{yy}}$$

More Examples





Example:

- Hubble’s data (homework for Linear Regression):
- We had $S_{xy} = 4348,\ \ S_{xx} = 9583,\ \ S_{yy} = 3168124$
- Whence, $r^2 = \frac{4348^2}{9583 \times 3168124} = 0.62$
- Most of the variation (around 60%) can be explained by the estimated linear relationship.
  ということは、変動の度合い（ほぼ60%)が回帰による推測された線形関係に、説明されるといえる。