Essential Mathematics for Global Leaders I

Statistics Spring 2019

Lecture 11: 2019 Jul. 15-Jul.22

PART II: Statistical inference (推計統計学)

6: Simple Linear Regression （単）線形回帰

6.1 Introduction

- Previously: inference about fixed parameters of a population (mean or variance was assumed fixed)
  先：母集団に関する一定な量を推定した（母平均と母分散などは一定だと想定した）。

- Often scientists need to know how the mean varies with another quantity of the population.
  科学者は母平均が他の母集団の量によってどうやって変わるかをしばしば調べたい。
  - Example: Population = {all rented flats in Tokyo}
    data={rental price}. Mean=average rental price.
  - Question: how the rental price varies with the surface?
    Target data: variation in $$/m^2$ + minimal price

- Goal: find a SIMPLE relation between two data of the population.
  Data 2 depends on Data 1. 量２は量１に依存する。
  目的：母集団の二つ量に対して簡単な関係を推定すること。

Modeling bivariate data as: function + noise (I)
函数+ノイズとする二変数データのモデル

- Bivariate Data: $(x_1, y_1), \ldots, (x_n, y_n)$
  Not a random sample. Just data. （ランダムな標本ではなく、ただのデータ）
- **Model**: x and y are related as:

$$y_i = f(x_i) + E_i$$

$f(x)$ is a function, or model

- E_i: "random" noise (☛Randomness is here! Not in the sample)
  - Typically, models measurement errors or the relation with negligible parameters not considered etc.
    典型的にこのノイズは測定誤差また考慮しない無視できる量との関係などをモデルできるもの。

Example of linear models for regression

- Lines: $y = ax + b + E$
  線
  Simple Linear regression
- Polynomials: $y = ax^2 + bx + c + E$
  多項式
- Other: $y = a/x + b + E$
- Other: $y = a\sin(x) + b + E$
- Goal: find parameters a, b, c ... of the model
  目標：モデルのパラメータ$a, b, c, \ldots$を求めること。

Modeling bivariate data as: function + noise(II)
函数+ノイズとする二変数データのモデル

- Find f that minimizes the Total Squared Error (最小二乗):

$$Min\left(\sum_{i=1}^{n} E_i^2\right) = \mathrm{Min}\left(\sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2\right)$$

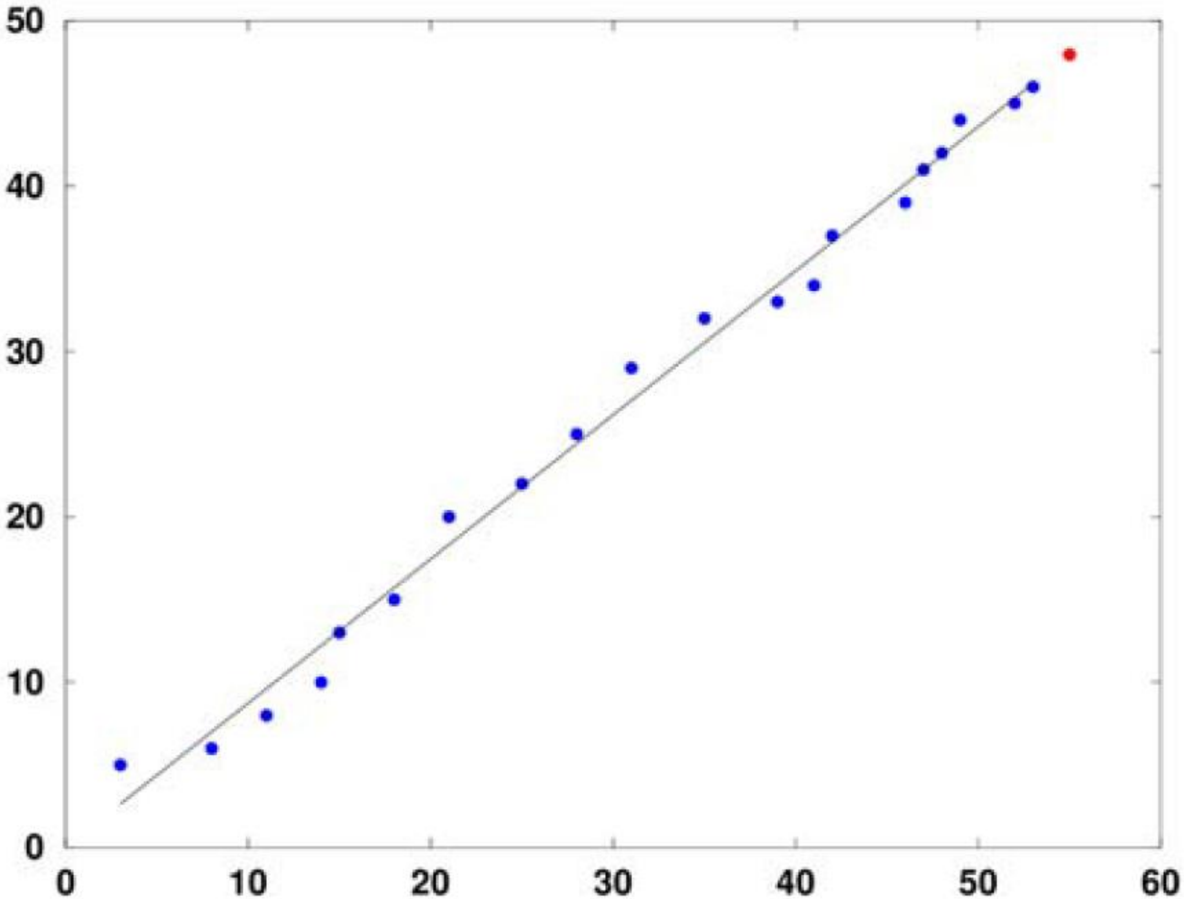
- Use for prediction: 予測のために利用する
  Infer the best function f (within the model) and then predict y given x by $y = f(x)$
  モデルの中での最も適切な函数fを推定してから、データ1のxに対するデータ2のyを予知する

- x: independent or predictor or explanatory variable
  (独立 or 予測 or 説明変数)

- y: dependent or outcome or response variable
  (従属 or 結果 or 反応変数)

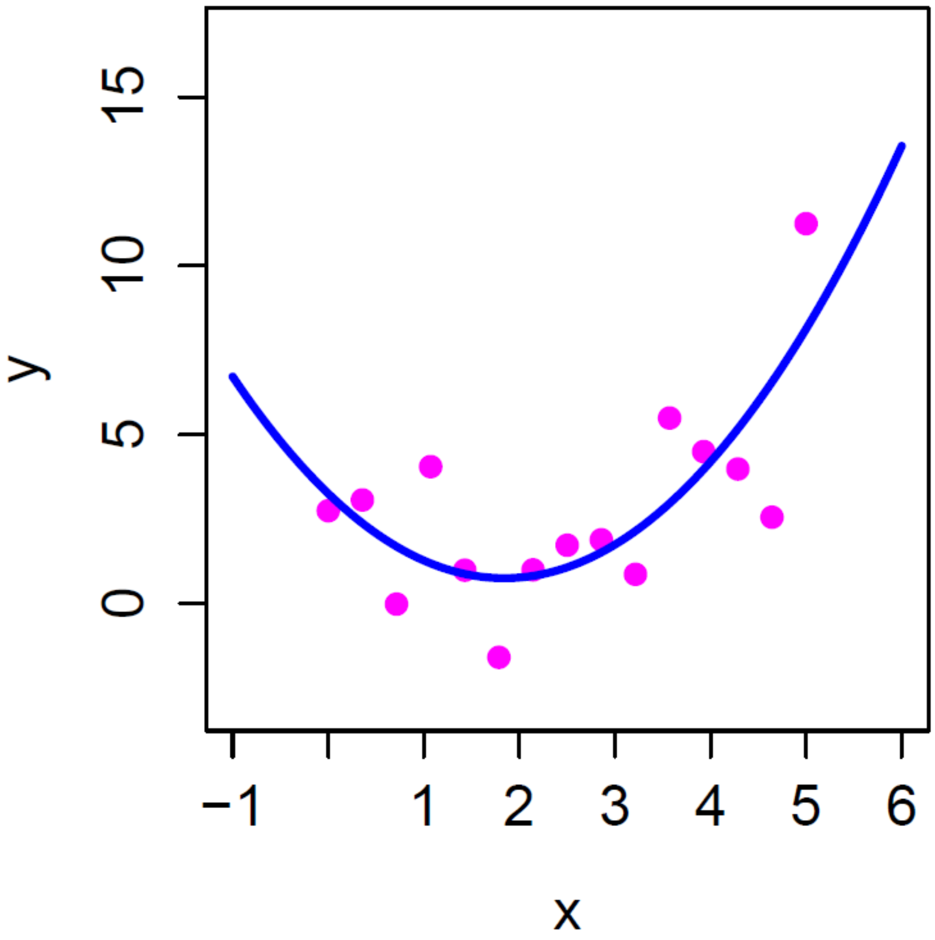
Example: Simple Linear Regression

- Example: Stamp price in US (in Cents) vs. time (years since 1960).



- Red point is predicted cost in 2015.
- 赤点は2015年に予測される値段

Example: Linear Regression by a parabola
放物線を使う線形回帰の例



- This linear regression is not simple.
- それは単回帰ではない

What is linear about linear regression?
線形回帰においての「線形」って、いったいどのことか？

- Linear in the parameters a, b, c, ... of the model
  - $y = ax^2 + bx + c$
  - $y = ax + b$
  - $y = a/x + b$
  - $y = a \sin(x) + b$

- Non-linear (but parametric) model: $y = be^{ax^2} + E$

- It is not because the curve being fit has to be a straight line (simple lin. reg. 単回帰)
  Although it is the most common case

Simple Linear Regression: finding the best fitting line 単回帰: 最良適合線

- Simple Linear Regression: fit a line to the data $y_i = a\,x_i + b + E_i$, where $E_i \sim N(0, \sigma^2)$
  - σ is a fixed value, the **same** for all data points.
  - The r.v. E_i are **independent**
- Total square error: $\sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (y_i - a\,x_i - b)^2$
- Goal: find the values of a and b that give the 'best fitting line'

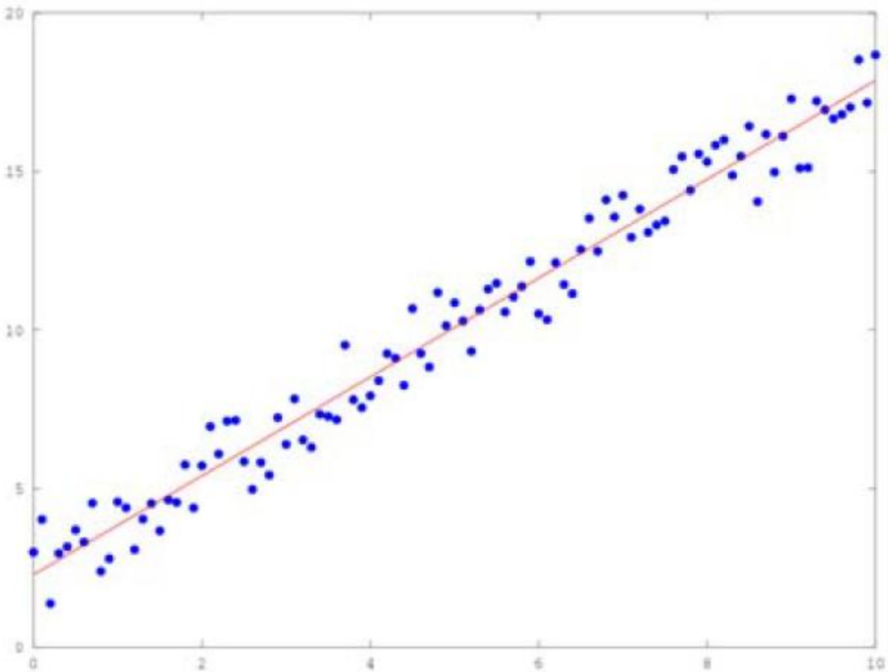
➢ Best fit (least squares)
The value of a and b that minimize the total squared error.

Linear Regression: finding the best fitting degree 2 polynomial　2次多項式の場合

- Linear Regression: fit a parabola to the data
  $y_i = a\,x_i^2 + bx_i + c + E_i$, where $E_i \sim N(0, \sigma^2)$
  - σ is a fixed value, the **same** for all data points.
  - The r.v. E_i are **independent**
- Total square error: $\sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} \left(y_i - a\,x_i^2 - bx_i - c\right)^2$
- Goal: find the values of a, b, c that give the 'best fitting line'
- Best fit (least squares)
  The value of a, b, c that minimize the total squared error.

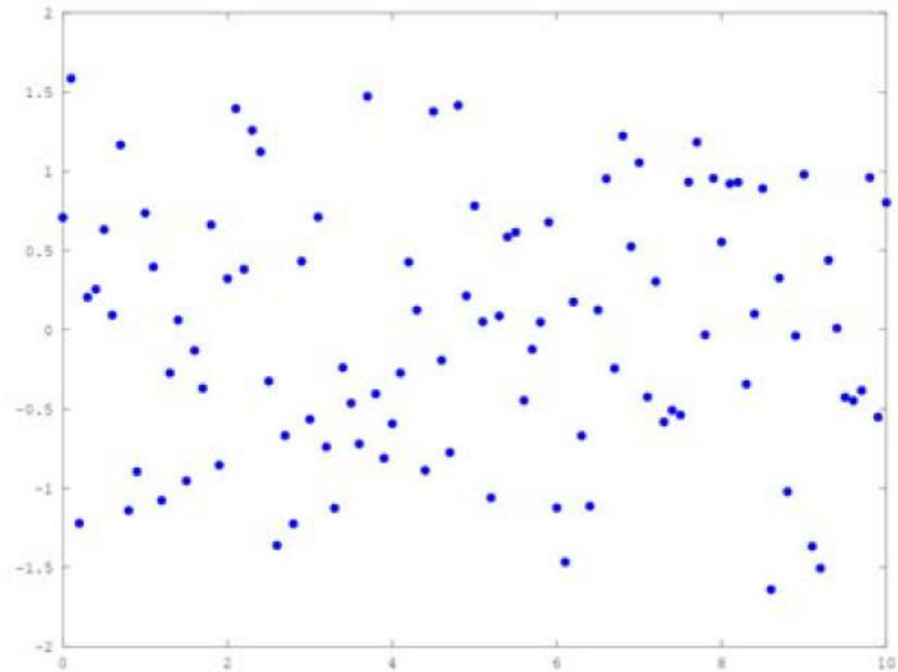
Errors have same variance:

- BIG ASSUMPTION:
  the errors E_i are independent with same variance σ^2.



Data and regression line (left).

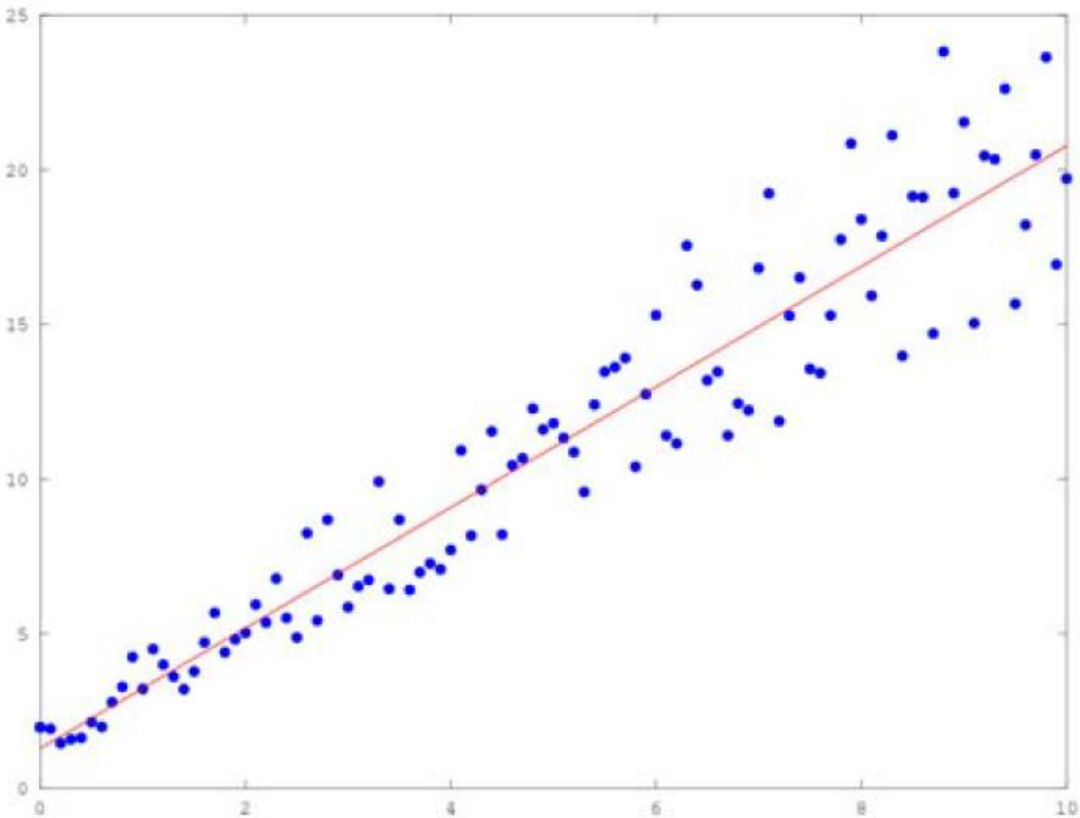
Homoscedastic (!) data



Errors: distance from line to point
誤差：点から直線まだ距離

Homoscedastic vs heteroscedastic
等分散性 vs 不等分散性

Here the variance of the random noise (=error) increases.
ここで、誤差（ノイズ）の分散は増加する.



Heteroscedastic Data

- Remark: Homoscedasticity と heteroscedasticityの発音はものすごく難しい。

6.2 Estimation for simple linear regression

1. Find point estimators $\hat{a}$ and $\hat{b}$ for the slope a and the y-intercept b
   傾きaとy-切片bの点推定$\hat{a}$と$\hat{b}$を求める。
   ☛Maximum Likelihood Estimation (here same as Least Square Estimation)
   最尤推定法を用いて（ここで、最小二乗推定法と一致する）
2. Find an (unbiased) point estimator for the variance σ^2
   分散σ^2に対する不変な点推定量を求める
3. Deduce the distribution of $\hat{a} - a$ and $\hat{b} - b$ under the normality assumption of random noise: $E \sim N(0, \sigma^2)$
   ランダムノイズ$E \sim N(0, \sigma^2)$の仮定下で、$\hat{a} - a$と$\hat{b} - b$の分布を演繹する。
4. And form a confidence interval for a and b
   aとbに対する信頼区間を設定する。

MLE for slope a and y-intercept b 傾きaとy-切片b:最尤推定法

- Not enough time in this course to study MLE in details この授業では、MLEの原理を勉強するために時間がない。
- ……a glimpse into the MLE method: MLEを垣間見したら
  - Distribution of population depends on some parameters (vector θ) "parametric model"
    $f(\cdot|\theta)$ $\theta \rightarrow$ unknown 未知
  - Each sample/data x_i is a random variable following $f(\cdot|\theta)$
    - $f(x_i|\theta)$ conditional probability density function.
  - Joint distribution (multivariate probability function):
    - $f_\theta(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$, (independent assumption)
  - MLE consists to find parameters θ that maximize the likelihood function L (given of n data points $x_1, \ldots, x_n$):
    $$\mathrm{L}(\theta) \coloneqq f_\theta(x_1, \ldots, x_n|\theta)$$

MLE for slope a and y-intercept b
傾きaとy-切片b:最尤推定法

- Model: $Y_k = ax_k + b + E_k,$ $\qquad E_k \sim N(0, \sigma^2)$
- ☛ $\quad Y_k \sim N(ax_k + b, \sigma^2)$
- Thus, $f(Y_k|a,b) = \dfrac{e^{-\left(Y_k-(ax_k+b)\right)^2/2\sigma^2}}{\sigma\sqrt{2\pi}}$
- MLE: Find a and b that maximize:

$$L(a,b) := \prod_{k=1}^{n} f(Y_k|a,b) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\sum_{k=1}^{n}\left(Y_k-(ax_k+b)\right)^2/2\sigma^2}$$

- Same as finding a and b that minimize:

$$T(a,b) := \sum_{k=1}^{n}\left(Y_k - (ax_k + b)\right)^2$$

- Least square estimation !! 最小二乗推定法!!

Formula for $\hat{a}$ and $\hat{b}$

- Solve $\partial T/\partial a=0$ and $\partial T/\partial b=0$. After solving, we find :
- We find: $\hat{a} = \frac{S_{xY}}{S_{xx}}$ and $\hat{b} = \bar{Y} - \hat{a}\bar{x}$

(both $\hat{a}$ and $\hat{b}$ are here random variables)

Where

- $S_{xx} = \sum_k (x_k - \bar{x})^2 = \sum_k x_k^2 - n\bar{x}^2$ where $\bar{x} = \frac{1}{n}(x_1 + \cdots + x_n)$
- $S_{xY} = \sum_k (x_k - \bar{x})(Y_k - \bar{Y}) = \sum_k x_k Y_k - n\bar{x}\bar{Y}$ where $\bar{Y} = \frac{1}{n}(Y_1 + \cdots + Y_n)$

- Remark: S_{xx} is a number S_{xY} is a Random Variable
  - When Y is measured to get data $y_1, \dots, y_n$ we write y instead of Y.
  - $\bar{y} = \frac{1}{n}(y_1 + \cdots + y_n)$ and $S_{xy} = \sum(x_k - \bar{x})(y_k - \bar{y})$

Example: using S_{xx} and S_{xy}

- Example: Consider the data $(1,3), (2,1)\ , (4,4)$
  Find the line $y = ax + b$ that best fits these three points.

- Answer:
  $\bar{x} = \frac{7}{3},\ \bar{y} = \frac{8}{3},\ S_{xx} = \frac{14}{3}\ , S_{xy} = \frac{7}{3}, \hat{a} = \frac{1}{2}, \hat{b} = \frac{3}{2}$

Example: using MLE (or equivalently LSE)

- *(on the blackboard)*

6.3 Confidence Interval for a and b

Properties of $\hat{a}$ and $\hat{b}$ as r.v.

- Estimators $\hat{a}$ and $\hat{b}$ are **unbiased.**
- $E(\hat{a}) = a$ $\qquad$ $E(\hat{b}) = b$
- **Variance**: $Var(\hat{a}) = \dfrac{\sigma^2}{S_{xx}}$ $\qquad$ $Var(\hat{b}) = \dfrac{\sigma^2 \sum_k x_k^2}{nS_{xx}}$
- **Theorem:**

$$\hat{a} \sim N\left(a, \frac{\sigma^2}{S_{xx}}\right) \qquad \hat{b} \sim N\left(b, \frac{\sigma^2 \sum_k x_k^2}{nS_{xx}}\right)$$

Estimation for σ (not known in general)

- MLE estimator for σ^2: $\widehat{\sigma^2} = \frac{1}{n}\sum_k \left(Y_k - \hat{a} - \hat{b}x_k\right)^2$
- But it is biased ただ、不変でない推定値

$$E(\hat{\sigma}^2) = \frac{n-2}{n}\sigma^2$$

- **Unbiased estimator** for σ^2:

$$s^2 = \frac{1}{n-2}\sum_k \left(Y_k - \hat{a} - \hat{b}x_k\right)^2 = \frac{n}{n-2}\hat{\sigma}^2 \quad ☛ \quad E(s^2) = \sigma^2$$

- **Theorem (Helmert/Cochran)** (☛Lecture 8, ch 4.2 p. 5)

$$(n-2)\frac{s^2}{\sigma^2} \sim \chi^2_{n-2}$$

Confidence Interval for a and b (I)

- We can find the law of $\hat{a} - a$ and $\hat{b} - b$ to build Confidence Intervals for a and b.

- By the Theorem on page 19, we have:

$$\frac{\hat{a}-a}{\sigma\sqrt{1/S_{xx}}} \sim N(0,1) \text{ and } \frac{\hat{b}-b}{\sigma\sqrt{Var(\hat{b})}} \sim N(0,1)$$

Theorem (Fisher-Student) (☛ Lecture 8, Ch.4.3, page 11)

- $T_a := \frac{\hat{a}-a}{s}\sqrt{S_{xx}} \sim t_{n-2}$ (Student's Law with n-2 df)
- $T_b := \frac{\hat{b}-b}{s}\sqrt{\frac{nS_{xx}}{\sum_{k=1}^{n} x_k^2}} \sim t_{n-2}$ (n-2自由度Studentのt分布)

Confidence Interval for a and b (II)

- $1 - \alpha$ Confidence Interval for the slope a:

$$\hat{a} \pm t_{\frac{\alpha}{2}} \cdot s \frac{1}{\sqrt{S_{xx}}}$$

- $1 - \alpha$ Confidence Interval for the y-intercept b:

$$\hat{b} \pm t_{\frac{\alpha}{2}} \cdot s \sqrt{\frac{\sum_k x_k^2}{n S_{xx}}}$$

- S_{xx} ☛ page 16 s ☛ page 24

6.4 Prediction interval 予測区間

- Suppose we have found $\hat{a}$ and $\hat{b}$ as well as s from data $(x_1, y_1), \dots, (x_n, y_n)$.
- Given a new data x_{new} we can predict the response Y_{new} by the estimator (or "predictor")

$$\hat{Y}_{new} = \hat{a}\, x_{new} + \hat{b}$$

- Next slide ☞ CI for $\hat{Y}_{new}$ 点推定の信頼区間

"Prediction Interval 予測区間"

- Remark: $Y_{new} = a x_{new} + b + E_{new}$. Since $E(E_{new}) = 0$,

$$E(Y_{new}) = \mu_{Y_{new}} = a\, x_{new} + b$$

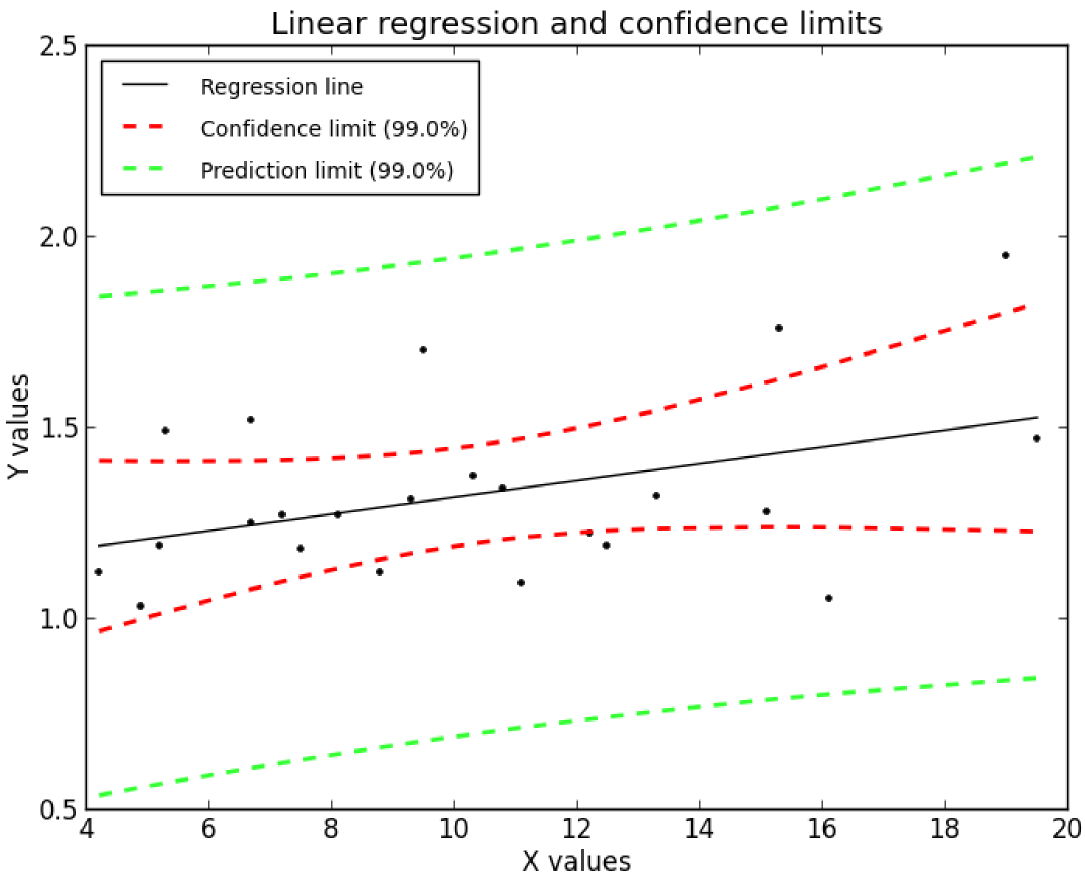
- Don't know a nor b : estimator $\widehat{\mu_{Y_{new}}} = \hat{a}\, x_{new} + \hat{b}$.
- (Since $\hat{a}$ and $\hat{b}$ are random variables $\widehat{\mu_{Y_{new}}}$ also.

- We can show that a $1-\alpha$ confidence interval for Y_{new}

$$\hat{a}x_{new} + \hat{b} \pm t_{\frac{\alpha}{2}} \cdot s\sqrt{1 + Sx_{new}}$$

- We can show that a $1-\alpha$ confidence interval for $\widehat{\mu_{Y_{new}}}$

$$\hat{a}x_{new} + \hat{b} \pm t_{\frac{\alpha}{2}} \cdot s\sqrt{Sx_{new}}$$



- $Sx_{new} = \frac{1}{n} + \frac{(x_{new}-\bar{x})^2}{S_{xx}}$

- where $t_{\frac{\alpha}{2}}$ is the $p = \frac{\alpha}{2}$-value of the Student t distribution with $n-2$ degrees of freedom