回归分析

Regression

吴志勇
清华大学深圳研究生院

主要内容

- 相关分析
- 线性回归
  - 一元线性回归
  - 多元线性回归
  - 多项式回归
- 非线性回归
- 回归分析的统计特性

参考教材

- 《数理统计学习教程》
  陈希孺2009
- 《Statistical Model》
  David Freedman 2010
- 《Statistical Inference》
  George Casella 2002
- 《Linear Regression Analysis》
  George Seba 2003

基础知识

- 研究变量之间的两种关系
  - 确定性关系: 变量之间存在确定性的**因果关系**, 可用函数关系表达
  - 相关性关系: 变量之间存在非确定性的**依赖关系**, 又称统计相关关系

统计相关关系

- 变量之间存在非确定性的依赖关系，但具有统计意义上的相关性
- 涉及的变量是随机变量

例1: **身高和体重**，不存在准确的函数可以由身高计算出体重，但从统计意义上来说，它们之间存在相关性，身高者，体也重

例2: **父母的身高与子女的身高**之间也有一定联系，通常父母高，子女也高

回归分析: Regression Analysis

回归分析：一种研究变量间统计相关关系的统计分析方法.

(研究相关性关系的最基本, 应用最广泛的方法)

回归模型

- $X = (x_1, x_2, \ldots, x_k)$: 非随机变量, **自变量** (independent variable)
- y: 受 $x_1, x_2, \ldots, x_k$ 的随机影响, 随机变量, **因变量** (dependent variable)
- y 对 $x_1, x_2, \ldots, x_k$ 相关关系的数学模型: $y=f(x_1,x_2,\ldots,x_k)+\varepsilon$
  称 y 对 $x_1, x_2, \ldots, x_k$ 的**回归模型**, ε 随机**误差**
- $y=f(x_1, x_2, \ldots, x_k)$
  称 y 对 $x_1, x_2, \ldots, x_k$ 的**回归方程**,
  函数 $f(x_1, x_2, \ldots, x_k)$ 称为**回归函数**, y 对 X 的回归函数

回归分析

- 回归分析:
  根据变量 $x_1, x_2, \ldots, x_k$ 和 y 的具体观察值去估计回归函数 f

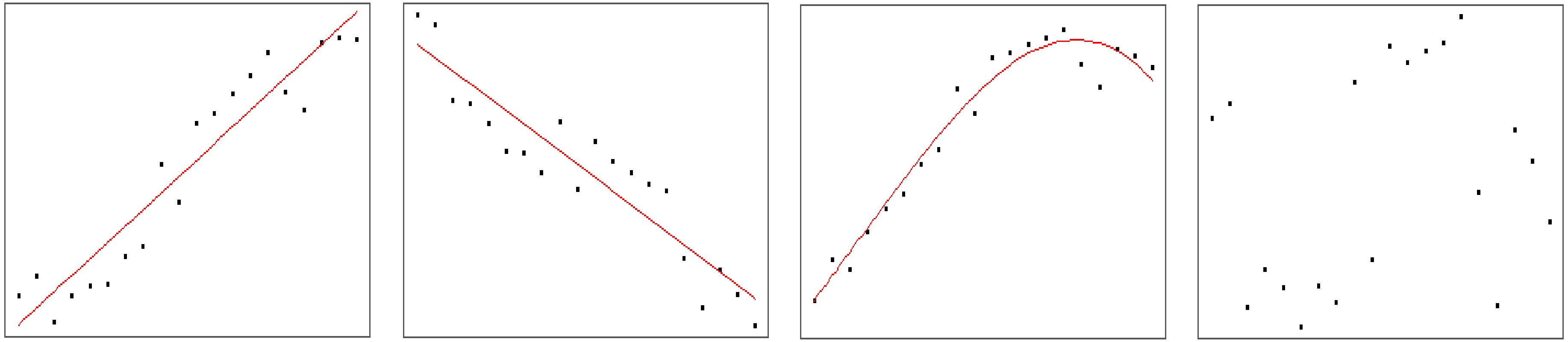
- 一元线性回归分析:
  只有一个自变量的回归分析 (i.e. k=1): $y=ax+b$

- 多元线性回归分析:
  多于一个自变量的回归分析 (i.e. k>1): $y=a_1x_1+a_2x_2+\ldots+a_kx_k+b$

Analysis of Correlation
相关分析

1. 散点图

- 散点图是描述变量之间关系的一种直观方法.
- 我们用坐标的横轴代表自变量 x, 纵轴代表因变量 y, 每组数据 (x_i, y_i) 在坐标系中用一个点表示, 由这些点形成的散点图描述了两个变量之间的大致关系, 从中可以直观地看出变量之间的关系形态及关系强度.



2. 相关系数

- 相关系数是对变量之间关系密切程度的度量
- 总体相关系数
  - 若相关系数是根据总体全部数据计算的, 称为**总体相关系数**, 记为 ρ
- 样本相关系数
  - 若相关系数是根据样本数据计算的, 则称为**样本相关系数**, 记为 r
  - 样本相关系数简称为**相关系数**

2. 相关系数

- 总体相关系数

$$\rho = \frac{COV(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$$

$COV(X,Y)$ 为变量 X 和 Y 的协方差

$D(X)$ 和 $D(Y)$ 分别为 X 和 Y 的方差

- 样本相关系数

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)}}$$

一般情况下，总体相关系数 ρ 是未知的.
通常是将样本相关系数 r 作为 ρ 的近似估计值.

相关系数 r 的性质

- 1) 相关系数的取值范围: $-1 \le r \le 1$.
  若 $0 < r \le 1$, 表明 X 与 Y 之间存在**正线性相关关系**;
  若 $-1 \le r < 0$, 表明 X 与 Y 之间存在**负线性相关关系**.

- 2) 若 $r = 1$, 表明 X 与 Y 之间为**完全**正线性相关关系;
  若 $r = -1$, 表明 X 与 Y 之间为**完全**负线性相关关系;
  若 $r = 0$, 说明二者之间**不存在**线性相关关系.

相关系数 r 的性质

- 3) 当 $-1 < r < 1$ 时, 为说明两个变量之间的相关系的密切程度, 通常将相关程度分为以下几种情况:

  当 $|r| \geq 0.8$ 时, 可视为高度相关;

  当 $0.5 \leq |r| < 0.8$ 时, 可视为中度相关;

  当 $0.3 \leq |r| < 0.5$ 时, 可视为低度相关;

  当 $|r| < 0.3$ 时, 说明两个变量之间的相关程度极弱, 可视为不相关.

建立在对相关系数进行显著性检验的基础之上

3. 相关系数的显著性检验

- 检验总体相关系数 ρ 是否显著为 0
- 通常采用费歇尔 (Fisher) 提出的 t 分布检验, 可用于小样本, 也可用于大样本

1) **提出假设**:
假设样本是从一个不相关的总体中随机抽取的, 即

$$H_0: \rho = 0; H_1: \rho \neq 0$$

2) 由样本观测值计算**检验统计量**:

$$t = |r|\sqrt{\frac{n-2}{1-r^2}} \sim t(n-2)$$

的观测值 t_0 和衡量观测结果极端性的 ***P*值 (P-value)**:

$$\text{P-value} = P\{|t| \geq |t_0|\} = 2P\{t \geq |t_0|\}$$

3) **进行决策**:
比较 P值 和显著性水平 α 作判断:
若 $P < \alpha$, 拒绝原假设 H_0; 若 $P \geq \alpha$, 不能拒绝原假设 H_0.

相关系数的显著性检验

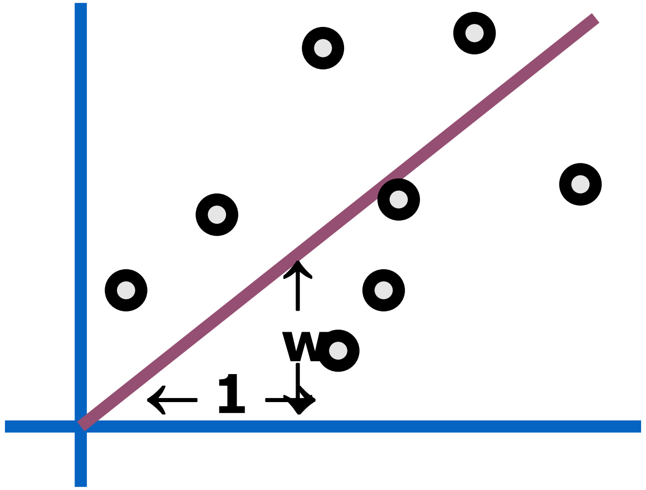
- 若 $P>0.05$, 接受H_0, 相关不显著, 即总体 x 与 y 间不存在相关关系
- 若 $0.01<P<0.05$, 拒绝H_0, 相关显著, 即总体 x 与 y 间存在显著相关关系
- 若 $P<0.01$, 拒绝H_0, 相关极显著, 即总体 x 与 y 间存在极显著的相关关系

Linear Regression
线性回归

Linear Regression (线性回归)

- Linear regression is a linear approach for modelling the relationship between a ***scalar*** dependent variable y (一个因变量) and one or more explanatory variables (or independent variables, 一个或多个自变量) denoted X.
- In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are ***estimated from the data***.
- 线性回归分析是研究一个因变量 y 与一个或多个自变量 X 间关系的统计方法

Linear Regression (线性回归)



inputs	outputs
$x_1 = 1$	$y_1 = 1$
$x_2 = 3$	$y_2 = 2.2$
$x_3 = 2$	$y_3 = 2$
$x_4 = 1.5$	$y_4 = 1.9$
$x_5 = 4$	$y_5 = 3.1$

- Linear regression assumes that the expected value of the output given an input, $E[y|x]$, is **linear**.
- Simplest case: $\text{out}(x) = wx$ for some unknown w.
- Given the data, we can estimate w.

$\hat{y} = E[y|x]$: y对x的回归函数, 简称回归
- x是一维: 一元回归
- x是p维向量: p元回归

Linear Regression

- In linear regression, the case of one independent variable (一个自变量) is called simple linear regression (一元线性回归).

- For more than one independent variable (多个自变量), the process is called multiple linear regression (多元线性回归).

1. Simple Linear Regression (一元线性回归)

Assume that the data is formed by a linear equation:

$$y_i = wx_i + noise_i$$

where…

- the noise signals are independent
- the noise has a normal distribution with mean 0 and unknown variance σ^2

$p(y|w,x)$ has a normal distribution with

- mean wx
- variance σ^2

Bayesian Linear Regression

$$p(y|w,x) = N(wx, \sigma^2)$$

We have a set of data points (x_1,y_1) (x_2,y_2) … (x_n,y_n) which are evidence about w.

We want to infer w from the data.

$$p(w|x_1, x_2, \ldots x_n, y_1, y_2, \ldots y_n)$$

- We can use BAYES rule to work out a posterior distribution for w given the data.
- Or we could do Maximum Likelihood Estimation.

Maximum a Posterior Estimation

- In Bayesian statistics, a **maximum a posteriori probability (MAP) estimation** is an estimate of an unknown quantity, that equals the mode of the posterior distribution.
- The MAP can be used to obtain a point estimate of an unobserved quantity on the basis of empirical data.

Maximum a Posterior Estimation

- MAP estimation

$$p(w|\{x_i, y_i\}) = \frac{p(\{y_i\}|w, \{x_i\})p(w)\cancel{p(\{x_i\})}}{\cancel{p(\{x_i, y_i\})}}$$

$$\arg\max_{w} p(w|\{x_i, y_i\}) = \arg\max_{w} p(\{y_i\}|w, \{x_i\})p(w)$$

需要知道被估计参数w的先验分布概率

Maximum Likelihood Estimation

- In statistics, **maximum likelihood estimation (MLE)** is a method of estimating the parameters of a statistical model, given observations.
- MLE attempts to find the parameter values that **maximize the likelihood function**, given the observations.

Maximum Likelihood Estimation

Asks the question:

"For which value of w is this data most likely to have happened?"

<=>

For what w is $p(y_1, y_2, \ldots, y_n \mid x_1, x_2, \ldots, x_n, w)$ maximized?

$$\arg\max_w p(w|\{x_i, y_i\}) = \arg\max_w p(\{y_i\}|w, \{x_i\})$$

Maximum Likelihood Estimation

Asks the question:

"For which value of w is this data most likely to have happened?"

<=>

For what w is $p(y_1, y_2, \ldots, y_n \mid x_1, x_2, \ldots, x_n, w)$ maximized?

<=>

For what w is $\prod_{i=1}^{n} p(y_i | w, x_i)$ maximized?

Maximum Likelihood Estimation

For what w is $\prod_{i=1}^{n} p(y_i|w, x_i)$ maximized?

$<=>$

For what w is $\prod_{i=1}^{n} \exp\left(-\frac{(y_i - wx_i)^2}{2\sigma^2}\right)$ maximized?

$<=>$

For what w is $\sum_{i=1}^{n} -\frac{(y_i - wx_i)^2}{2\sigma^2}$ maximized?

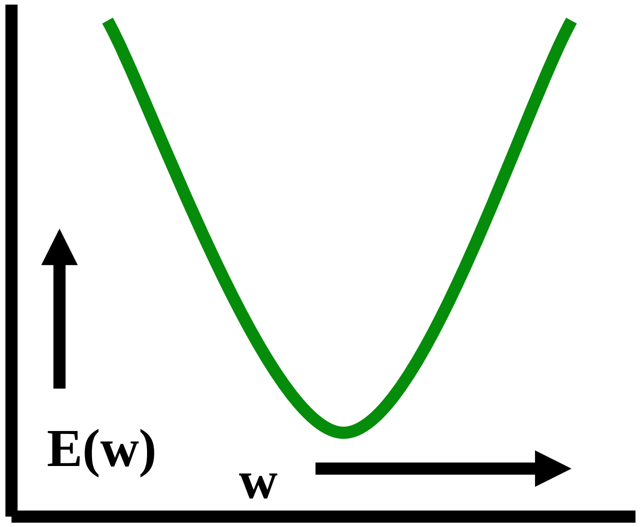
$<=>$

For what w is $\sum_{i=1}^{n} (y_i - wx_i)^2$ minimized?

This is also called
least square regression
(最小二乘估计/回归)

Linear Regression

The maximum likelihood w is the one that minimizes sum-of-squares of *residuals*



$$E = \sum_i (y_i - wx_i)^2 = \sum_i y_i^{\,2} - \left(2\sum x_i y_i\right)w + \left(\sum x_i^{\,2}\right)w^2$$

We want to minimize a quadratic function of w.

Linear Regression

Easy to show that the sum of squares is minimized when

$$w = \frac{\sum x_i y_i}{\sum x_i^{\;2}}$$

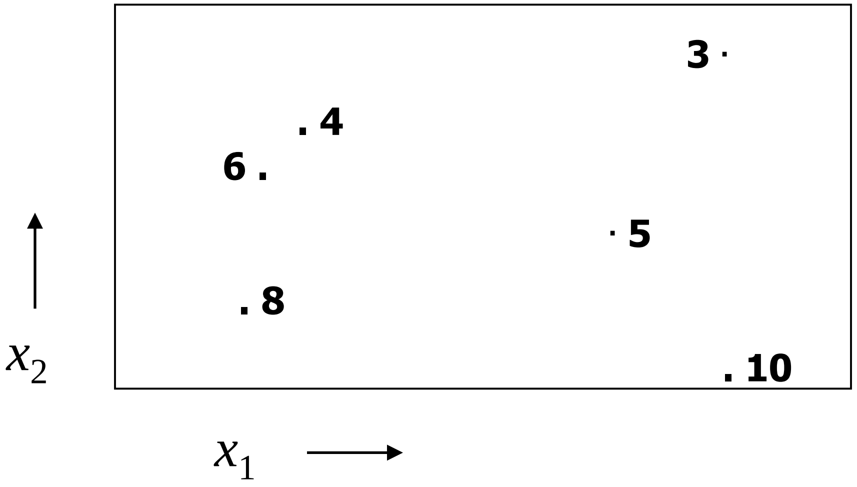
The Maximum Likelihood model is

$$\mathrm{out}(x) = wx$$

We can use it for prediction.

2. Multiple Linear Regression (多元线性回归)

What if the inputs are vectors?



2-d input example

Dataset has the form

$\mathbf{x}_1$	y_1
$\mathbf{x}_2$	y_2
$\mathbf{x}_3$	y_3
…	…
$\mathbf{x}_R$	y_R

Multiple Linear Regression

$$\mathbf{X} = \begin{bmatrix} \cdots\cdots\mathbf{x}_1\cdots\cdots \\ \cdots\cdots\mathbf{x}_2\cdots\cdots \\ \vdots \\ \cdots\cdots\mathbf{x}_R\cdots\cdots \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ & & \vdots & \\ x_{R1} & x_{R2} & \cdots & x_{Rm} \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_R \end{bmatrix}$$

Write the $\mathbf{X}$ matrix and $\mathbf{y}$ vector as above:

- There are R data points, and each input has m components.

- The linear regression model assumes a vector $\mathbf{w}$ such that:

$$out(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} = w_1x_1 + w_2x_2 + \cdots + w_mx_m$$

- The maximum likelihood $\mathbf{w}$ is $\mathbf{w} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{y})$

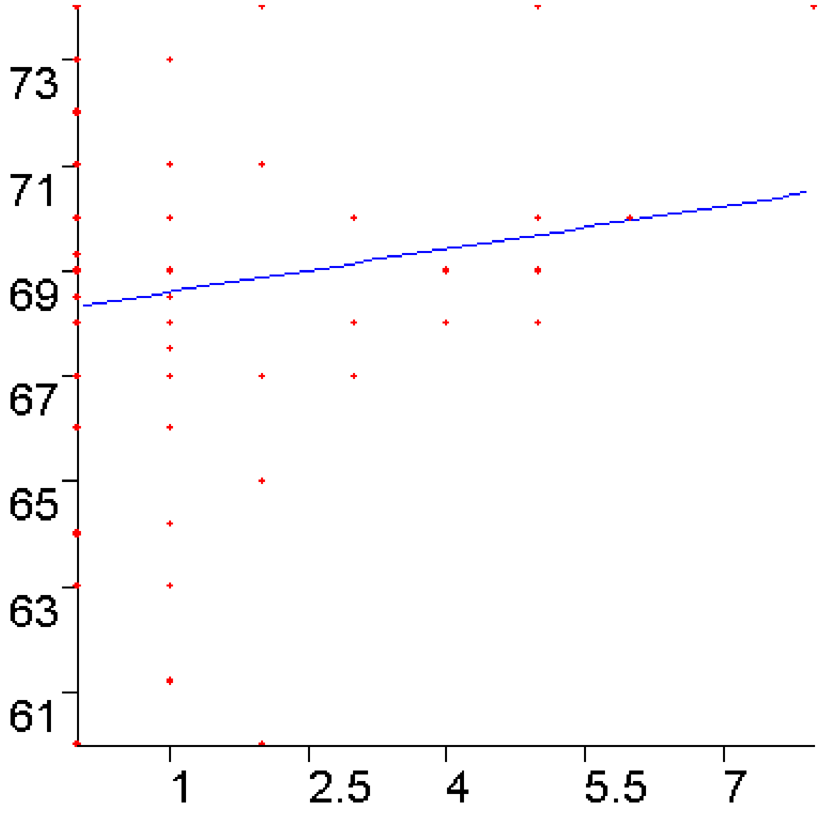
Multiple Linear Regression

- The maximum likelihood $\mathbf{w}$ is $\mathbf{w} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{y})$
- $\mathbf{X}^{\mathrm{T}}\mathbf{X}$: an $m \times m$ matrix with the $(i, j)^{\text{th}}$ term being $\sum_{k=1}^{R} x_{ki}x_{kj}$
- $\mathbf{X}^{\mathrm{T}}\mathbf{y}$: an m-element vector with the i^{th} term being $\sum_{k=1}^{R} x_{ki}y_k$

3. Constant Term in Linear Regression (常数项, *y*截距)

We may expect linear data that does not go through the origin.

Statisticians and Neural Net Folks all agree on a simple obvious hack.



The Constant Term

- The trick is to create a fake input "x_0" that always takes the value 1

x_1	x_2	y
2	4	16
3	4	17
5	5	20

Before:

$y=w_1x_1+ w_2x_2$

A poor model

In this example, you should be able to see the MLE w_0 , w_1 and w_2 by inspection.

x_0	x_1	x_2	y
1	2	4	16
1	3	4	17
1	5	5	20

After:

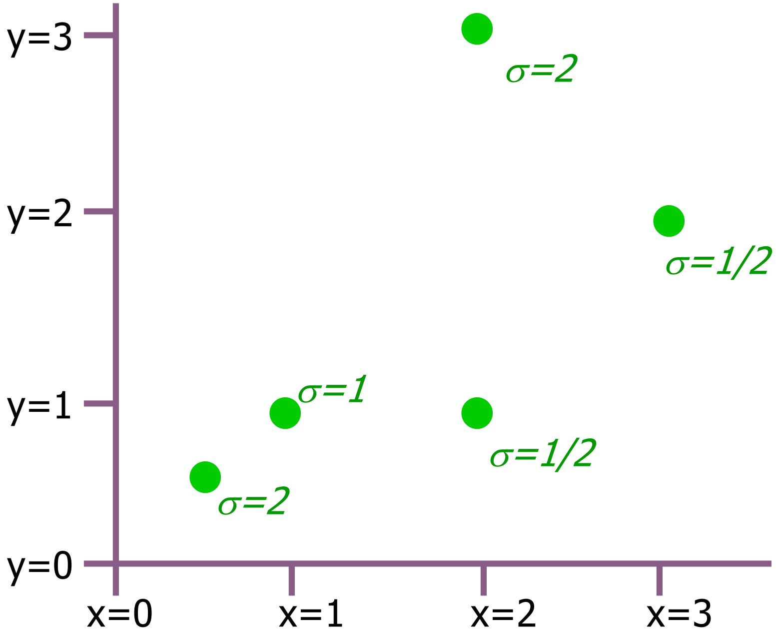
$y= w_0x_0+w_1x_1+ w_2x_2$

$= w_0+w_1x_1+ w_2x_2$

has a fine constant term

4. Regression with Varying Noise

- Suppose you know the variance of the noise that was added to each data point



x_i	y_i	σ_i^2
½	½	4
1	1	1
2	1	1/4
2	3	4
3	2	1/4

Assume $y_i \sim N(wx_i, {\sigma_i}^2)$

What's the MLE estimate of w?

ML Estimation with Varying Noise

Recall: ML estimation: $\arg\max_w p(\{y_i\}|w,\{x_i\})$

$$\arg\max_w \log p(y_1, y_2, \dots, y_R | x_1, x_2, \dots, x_R, \sigma_1^2, \sigma_2^2, \dots, \sigma_R^2, w)$$

$$= \arg\min_w \sum_{i=1}^{R} \left(\frac{y_i - wx_i}{\sigma_i}\right)^2$$

Assuming independence among noise and then plugging in equation for Gaussian and simplifying.

Setting dLL/dw=0

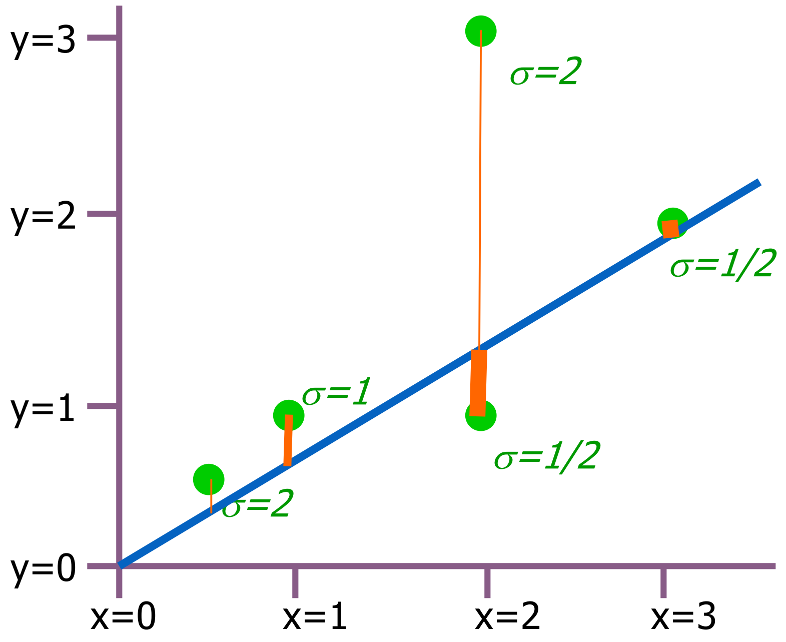
$$\Leftrightarrow \sum_{i=1}^{R} \frac{x_i(y_i - wx_i)}{\sigma_i^2} = 0 \quad \Leftrightarrow \quad w = \frac{\sum_{i=1}^{R} \frac{x_i y_i}{\sigma_i^2}}{\sum_{i=1}^{R} \frac{x_i^2}{\sigma_i^2}}$$

This is Weighted Regression

- We are minimizing the weighted sum of squares of noises

$$\arg\min_{w} \sum_{i=1}^{R} \frac{(y_i - wx_i)^2}{\sigma_i^2}$$

where weight for i^{th} data point is $\frac{1}{\sigma_i^2}$



Tsinghua University

5. Polynomial Regression (多项式回归)

- Polynomial regression is a form of linear regression in which the relationship between the independent variable x and the dependent variable y is modeled as an qth degree polynomial.
- Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y, denoted $E(y|x)$.

$$\hat{\boldsymbol{y}} = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_1 x_2 + w_5 x_2^2$$

Why polynomial regression is linear regression?

5. Polynomial Regression (多项式回归)

Why polynomial regression is linear regression?

- Although polynomial regression *fits a nonlinear model to the data*, as a statistical estimation problem it is linear, in the sense that the regression function $E(y|x)$ is **linear in the unknown parameters** w that are estimated from the data.
- Polynomial regression is considered to be a special case of multiple linear regression.

Polynomial Regression

- So far we've mainly been dealing with linear regression

x_1	x_2	y
3	2	7
1	1	3
:	:	:

$$\mathbf{X}=\begin{bmatrix} 3 & 2 \\ 1 & 1 \\ \vdots & \vdots \end{bmatrix} \quad \mathbf{y}=\begin{bmatrix} 7 \\ 3 \\ \vdots \end{bmatrix}$$

$\mathbf{x}_1=(3,2)..$ $\quad y_1=7..$

$$\mathbf{Z}=\begin{bmatrix} 1 & 3 & 2 \\ 1 & 1 & 1 \\ \vdots & & \vdots \end{bmatrix} \quad \mathbf{y}=\begin{bmatrix} 7 \\ 3 \\ \vdots \end{bmatrix}$$

$\mathbf{z}_1=(1,3,2)..\mathbf{z}_k=(1,x_{k1},x_{k2})$ $\quad y_1=7..$

$$\mathbf{w}=\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{y}\right)$$

$$\hat{\boldsymbol{y}}=w_0+w_1x_1+w_2x_2$$

Quadratic Regression

- It's trivial to do **linear fits** of fixed nonlinear basis functions

x_1	x_2	y
3	2	7
1	1	3
:	:	:

$\mathbf{X}=$

3	2
1	1
:	:

$\mathbf{y}=$

7
3
:

$\mathbf{x}_1=(3,2)$.. $y_1=7$..

$\mathbf{Z}=$

1	3	2	9	6	4
1	1	1	1	1	1
:					:

$\mathbf{y}=$

7
3
:

$\mathbf{z}=(1, x_1, x_2, {x_1}^2, x_1x_2, {x_2}^2)$

$$\mathbf{w} = \left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{y}\right)$$

$$\hat{\mathbf{y}} = w_0 + w_1x_1 + w_2x_2 + w_3x_1^2 + w_4x_1x_2 + w_5x_2^2$$

Quadratic Regression

- Each component of a $\mathbf{z}$ vector is called a term.
- Each column of the $\mathbf{Z}$ matrix is called a term column.
- How many terms in a quadratic regression with m inputs?
  - 1 constant term
  - m linear terms
  - $(m+1)$-choose-2 $= m(m+1)/2$ quadratic terms
  - ➢ $(m+2)$-choose-2 terms in total $= O(m^2)$
- Note that solving $\mathbf{w} = \left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{y}\right)$ is thus $O(m^6)$

Qth-degree Polynomial Regression

x_1	x_2	y
3	2	7
1	1	3
:	:	:

$\mathbf{X}=\begin{bmatrix}3 & 2\\ 1 & 1\\ \vdots & \vdots\end{bmatrix}$ $\mathbf{y}=\begin{bmatrix}7\\ 3\\ \vdots\end{bmatrix}$

$\mathbf{x}_1=(3,2)..$ $y_1=7..$

$\mathbf{Z}=\begin{bmatrix}1 & 3 & 2 & 9 & 6 & 4\\ 1 & 1 & 1 & 1 & 1 & 1\\ \vdots & & & & & \vdots\end{bmatrix}$ $\mathbf{y}=\begin{bmatrix}7\\ 3\\ \vdots\end{bmatrix}$

z=(all products of powers of inputs in which sum of powers is Q or less)

$$\mathbf{w} = (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}(\mathbf{Z}^{\mathrm{T}}\mathbf{y})$$

$$\hat{\boldsymbol{y}} = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^Q + \cdots + w_k x_2^Q$$

- m inputs, degree Q: how many terms?

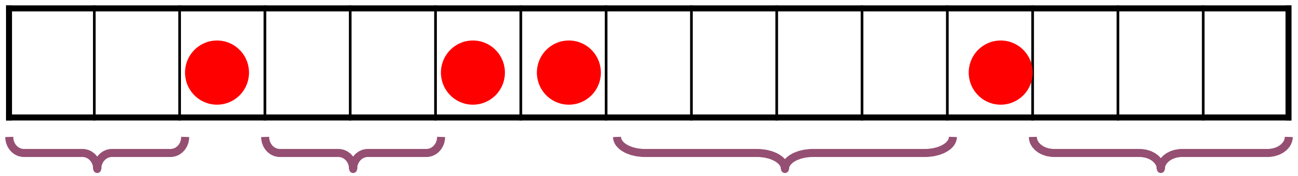
= the number of unique terms of the form $x_1^{q_1} x_2^{q_2} \ldots x_m^{q_m}$ where $\sum_{i=1}^{m} q_i \leq Q$

= the number of unique terms of the form $1^{q_0} x_1^{q_1} x_2^{q_2} \ldots x_m^{q_m}$ where $\sum_{i=0}^{m} q_i = Q$

= the number of lists of non-negative integers $[q_0, q_1, q_2, .. q_m]$ in which $\Sigma q_i = Q$

= the number of ways of placing m disks on a row of squares of length $Q+m$

= $(Q+m)$-choose-Q



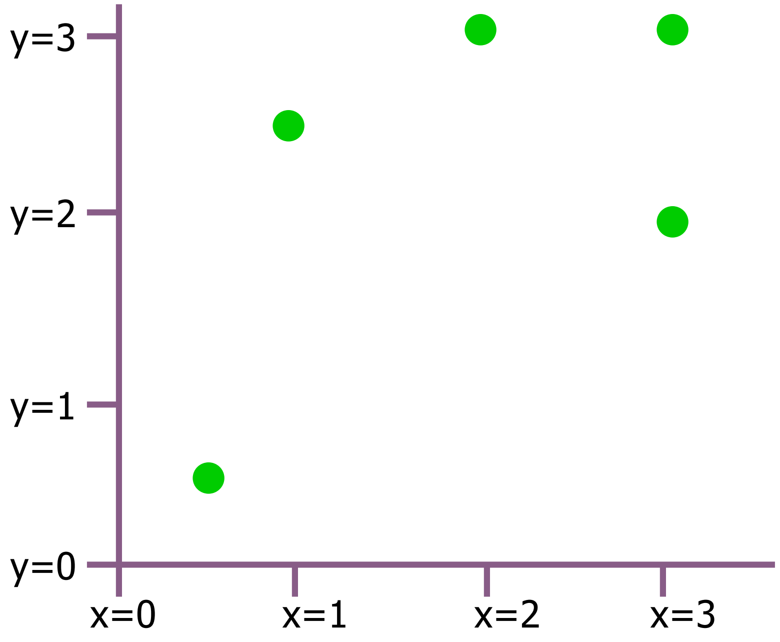
$Q=11,\ m=4$

$q_0=2$ $q_1=2$ $q_2=0$ $q_3=4$ $q_4=3$

Nonlinear Regression
非线性回归

Nonlinear Regression

- Suppose you know that y is related to a function of x in the way that the predicted values have a **non-linear dependence on** w



x_i	y_i
½	½
1	2.5
2	3
3	2
3	3

What's the MLE of w?

Assume $y_i \sim N(\sqrt{w + x_i}, \sigma^2)$

Nonlinear Regression

- Nonlinear regression is a form of regression analysis in which observational data are modeled by a function which is a **nonlinear combination of the model parameters** and depends on one or more independent variables.
- Examples of nonlinear functions include exponential functions, logarithmic functions, etc.

Nonlinear ML estimation

Recall: ML estimation: $\arg\max_w p(\{y_i\}|w,\{x_i\})$

$$\arg\max_w \log p(y_1, y_2, \ldots, y_R | x_1, x_2, \ldots, x_R, \sigma^2, w)$$

$$= \arg\min_w \sum_{i=1}^{R} \left(y_i - \sqrt{w + x_i}\right)^2$$

Assuming i.i.d. among noise and then plugging in equation for Gaussian and simplifying.

$$\Leftrightarrow \text{ select } w \text{ such that} \sum_{i=1}^{R} \frac{y_i - \sqrt{w + x_i}}{\sqrt{w + x_i}} = 0$$

Setting dLL/dw=0

Nonlinear ML estimation

$$\arg\max_{w} \log p(y_1, y_2, \dots, y_R | x_1, x_2, \dots, x_R, \sigma^2, \dots$$

$$= \arg\min_{w} \sum_{i=1}^{R} \left(y_i - \sqrt{w + x_i}\right)^2$$

$$\Leftrightarrow \text{ select } w \text{ such that } \sum_{i=1}^{R} \frac{y_i - \sqrt{w + x_i}}{\sqrt{w + x_i}} = 0$$

Iterative approach:

- Start with an initial guess of w
- Update w in each iteration until convergence

How to select w?

Common (but not only) approach:

Numerical Solutions:

- Line Search
- Simulated Annealing
- Gradient Descent
- Conjugate Gradient
- Levenberg Marquart
- Newton's Method

Also, special statistical-optimization-specific tricks such as EM

Linearization (线性化)

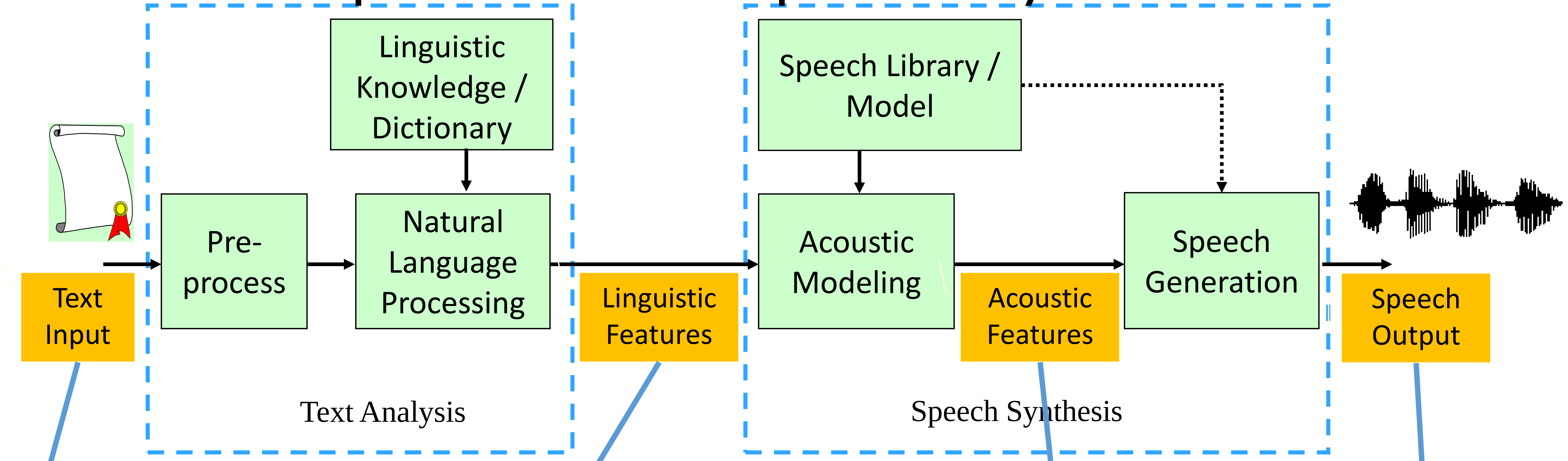
- Sometimes, it may be possible to transform a nonlinear regression function to a linear one.
- Example: $\mu_y(x) = \beta_1 e^{-\beta_2 x}$
  - $\ln\mu_y(x) = \ln\beta_1 - \beta_2 x$
  - Given $\{(y_1, x_1), \cdots, (y_n, x_n)\}$, let $z_i = \ln y_i$
  - Apply linear regression to $\{(z_1, x_1), \cdots, (z_n, x_n)\}$ and estimate $\ln\hat{\beta}_1$ and $\hat{\beta}_2$

Linearization (cont'd)

- Example: $\mu_y(x) = \frac{1}{\beta_1 - \beta_2 x}$
  - $\mu_y^{-1}(x) = \beta_1 - \beta_2 x$
  - Apply linear regression to $\{(y_i^{-1}, x_i)\}$

- Example: $\mu_y(x) = \frac{\beta_1 x}{\beta_2 + x}$
  - $\mu_y^{-1}(x) = \frac{1}{\beta_1} + \frac{\beta_2}{\beta_1} x^{-1}$
  - Apply linear regression to $\{(y_i^{-1}, x_i^{-1})\}$

An Example: Text-to-Speech Synthesis

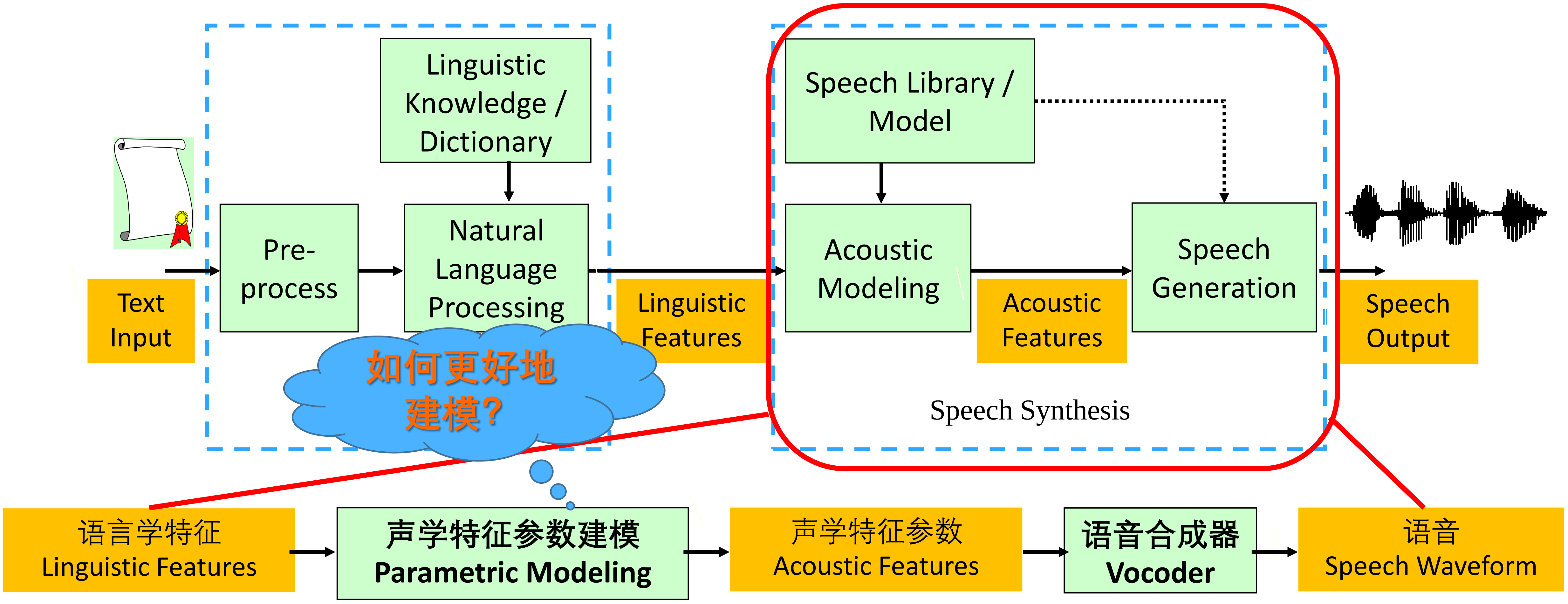


他说他经历了48年来最多的哭哭笑笑

/ 他说 / 他 | 经历了 | 四十八 | 年来 / 最多的 | 哭哭 | 笑笑 /

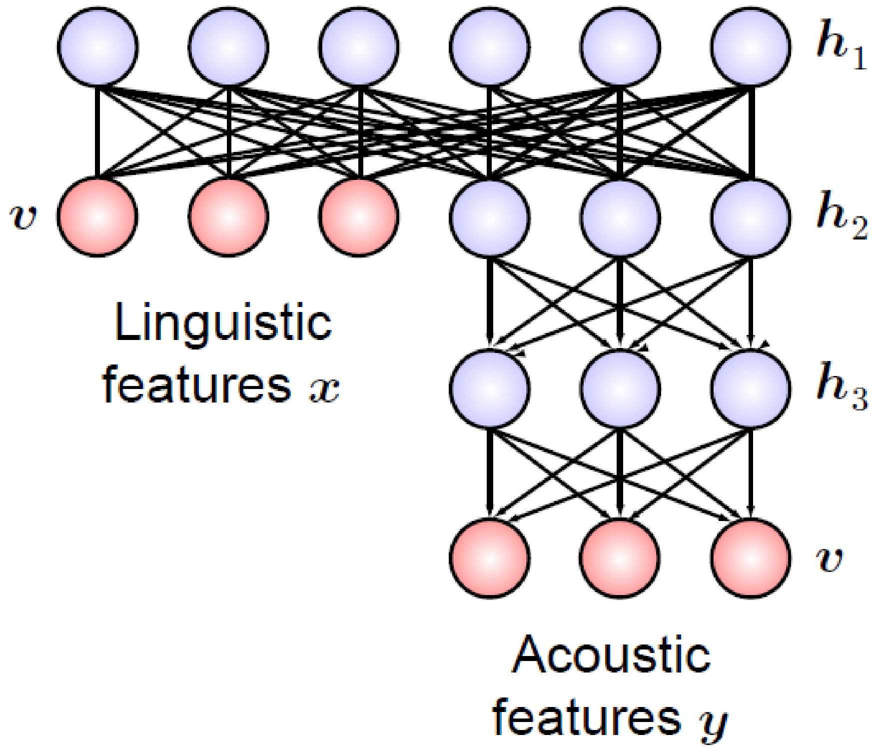
ta1 shuo1 ta1 jing1 li4 le5 si4 shi2 ba1 nian2 lai2 zui4 duo1 de5 ku1 ku1 xiao4 xiao4

Duration Parameters: Duration
Excitation Parameters: Pitch
Spectral Parameters: Spectrum


Text
Input
Pre-
process
Linguistic
Knowledge /
Dictionary
Natural
Language
Processing
Linguistic
Features
Speech Library /
Model
Acoustic
Modeling
Acoustic
Features
Speech
Generation
Speech
Output
Speech Synthesis
如何更好地
建模?
语言学特征
Linguistic Features
声学特征参数建模
Parametric Modeling
声学特征参数
Acoustic Features
语音合成器
Vocoder
语音
Speech Waveform

Speech Synthesis with DBN (Tsinghua-CUHK)

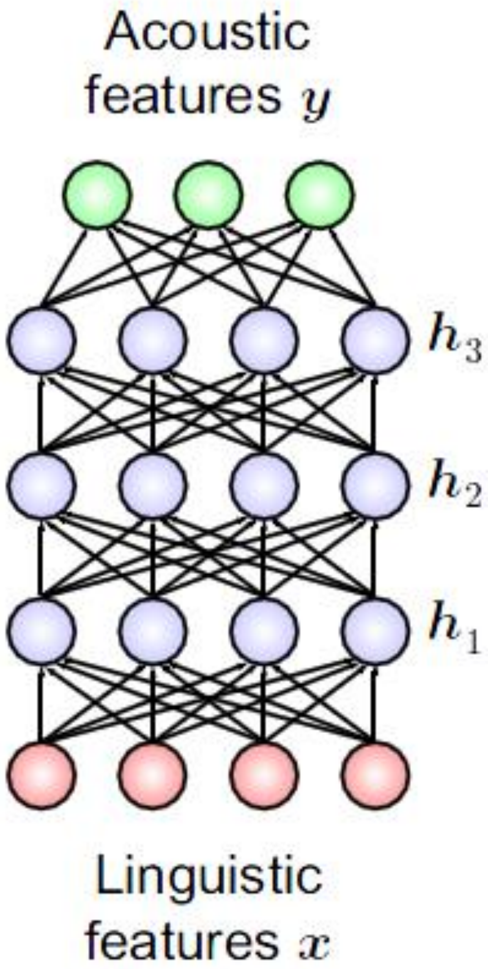
- DBN represents joint distribution of linguistic & acoustic features
  - P(x, y)
  - State-level modeling
- The acoustic features modeled:
  - Cepstral: spectrum
  - F0: pitch



Work done by Tsinghua-CUHK @ 2013

Speech Synthesis with DNN (Google)

- DNN represents conditional distribution of acoustic feature given linguistic feature
  - P(y | x)
  - State-level modeling
- The acoustic features modeled:
  - Cepstral: spectrum
  - F0: pitch



Work done by Google @ 2013

Detailed Analysis and Insights
回归分析的统计特性

Parameter Estimation

- An (unknown) parameter θ to be estimated
  - A point estimator $\hat{\theta}(x)$
  - $\hat{\theta}(x)$: estimate for the observed dataset x
- Unbiased estimator (无偏估计量): $E\left[\hat{\theta}(x)\right] = \theta$
- Consistent estimator (一致估计量):
  - A sequence of estimators $\{\hat{\theta}_n\}$ is a consistent estimator of θ iff $\forall \epsilon > 0, \lim_{n\to\infty} \left[\left|\hat{\theta}_n - \theta\right| < \epsilon\right] = 1$
  - The estimator converges in probability to the true values as the number of data points increases.

Unbiased Estimators

Given independently drawn observations $\{X_i\}$ of random variable X and $\{Y_i\}$ of random variable Y

- Sample mean: $\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$
    - Degree of freedom: n
- Sample variance: $S_X^2 = \frac{1}{n-1}\sum_i (X_i - \bar{X})^2$
    - Bessel correction: use n-1 but not n
    - D.f.: n-1 (constraint on $\bar{X}$)
- Sample covariance: $S_{XY} = \frac{1}{n-1}\sum_i (X_i - \bar{X})(Y_i - \bar{Y})$
    - D.f.: n-1

Linear Regression

- Given $\{(X_i, Y_i)\}$, find a straight line $\hat{Y} = b_0 + b_1 X$ to approximate their relationship
  - Fitted value $\hat{Y}_i$ and **residual** (残差) $e_i = Y_i - \hat{Y}_i$
  - MLE (LSE): $\arg\min_{b_0, b_1} \sum_i e_i^2$

$$\frac{\partial \sum_i \left(Y_i - \hat{Y}_i\right)^2}{\partial b_0} = 0 \Rightarrow \frac{\sum_i e_i}{n} = 0$$

$$\frac{\partial \sum_i \left(Y_i - \hat{Y}_i\right)^2}{\partial b_1} = 0 \Rightarrow \frac{\sum_i e_i X_i}{n} = 0 \Leftrightarrow \begin{cases} Cov(X, e) = 0 \\ Cov(\hat{Y}, e) = 0 \end{cases}$$

The residual has zero sample mean and is uncorrelated to X and therefore $\hat{Y}$.

Least Square Estimation

- Intercept:

$$\bar{e} = 0 \Leftrightarrow \bar{Y} = b_0 + b_1 \bar{X}$$

  - The center of mass $(\bar{X}, \bar{Y})$ is on the regression line.
  - LS finds the center of mass and rotates the line through that point until getting the "right" slope.
- Slope:

$$Cov(X, e) = 0 \Leftrightarrow b_1 = \frac{S_{XY}}{S_X^2} = r_{XY} \times \frac{S_Y}{S_X}$$

  - So, the right slope is the sample correlation coefficient times a scaling factor that ensures the proper units for b_1.

Decomposing the Sum of Squares

- How well does the least square line explain the variation in Y?

$$Y_i = \hat{Y}_i + e_i$$

- Given $\bar{e} = 0$ and $Cov(\hat{Y}, e) = 0$, we have

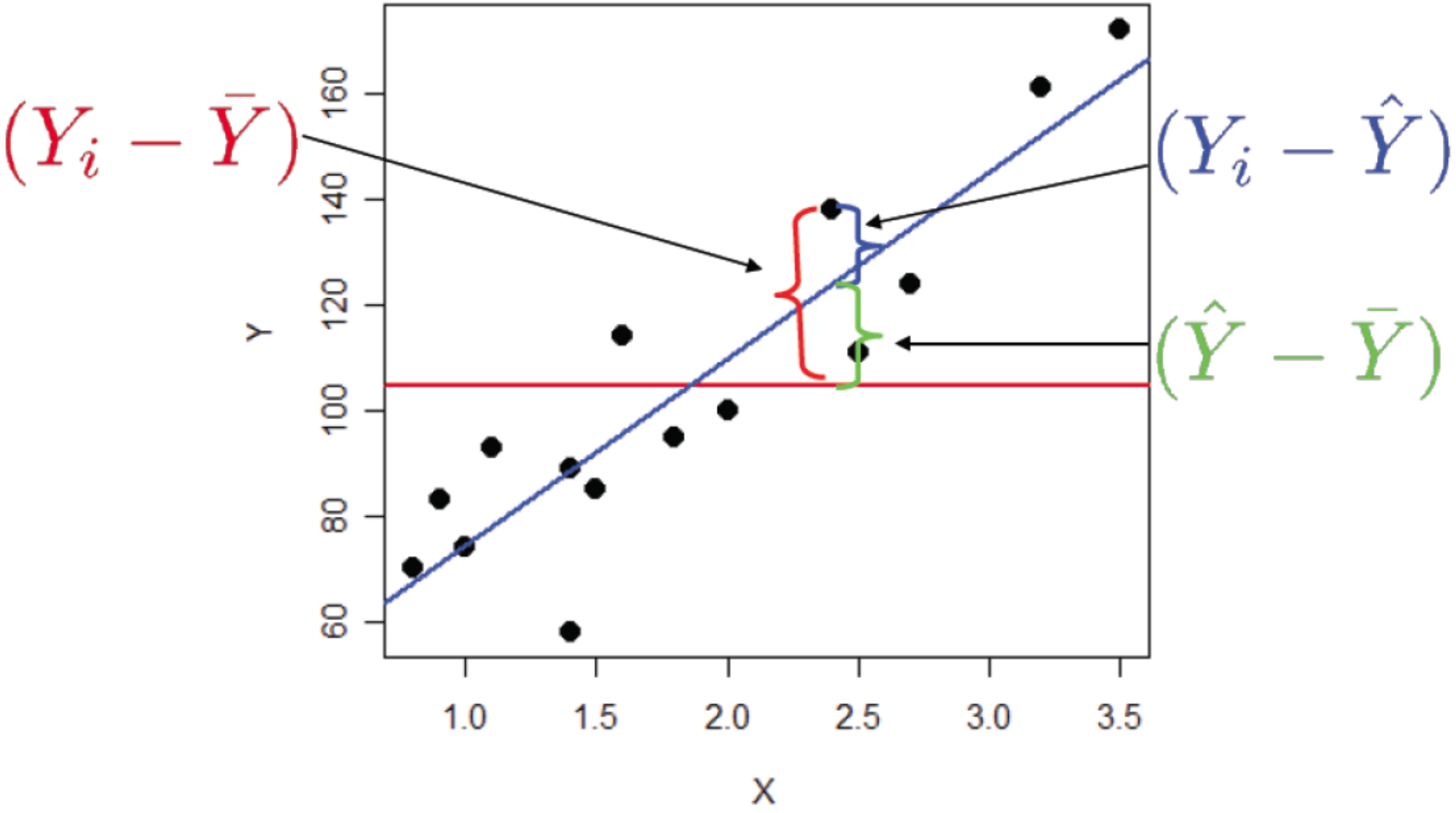
$$\underbrace{\sum_i (Y_i - \bar{Y})^2}_{\substack{\text{Total sum of squares} \\ \text{SST}}} = \underbrace{\sum_i \left(\hat{Y}_i - \bar{Y}\right)^2}_{\substack{\text{Regression SS} \\ \text{SSR}}} + \underbrace{\sum_i e_i^2}_{\substack{\text{Error SS} \\ \text{SSE}}}$$

SSR: variation in Y explained by the regression line (回归平方和)

SSE: variation in Y that is left unexplained (残差平方和)

SST: total variation (总变差平方和) $\quad \text{SSR} = \text{SST} \rightarrow \text{perfect fit}$

On A Scatter Plot



SSE: Error Sum of Squares (残差平方和)

$$SSE = \sum_i e_i^2 = \sum_i \left(Y_i - \hat{Y}_i\right)^2$$

- 反映了除去Y与X间线性关系之后，其他因素引起的数据$Y_1,Y_2,\ldots,Y_n$的波动
  - 如SSE=0, 则每个观测值可由线性关系精确拟合;
  - SSE越大, 观测值与线性拟合的偏差也越大.

SSR: Regression Sum of Squares (回归平方和)

$$SSR = \sum_i (\hat{Y}_i - \bar{Y})^2$$

- 反映了拟合值与其平均值的总偏差, 即由变量$X_1, X_2, \ldots, X_n$的变化引起的$Y_1, Y_2, \ldots, Y_n$的波动
  - 如$SSR=0$, 则每个拟合值均相等, 即Y_i ($i=1,2,\ldots,n$)不随X_i的变化而变化.

SST: Total Sum of Squares (总变差平方和)

$$SST = \sum_i (Y_i - \bar{Y})^2$$

- 反映了数据 $Y_1, Y_2, \ldots, Y_n$ 本身波动性的大小

$$SST = SSR + SSE$$

- SSR越大, 说明由线性回归关系描述的 Y_i 波动的比例就越大, 即 Y 与 X 之间的线性关系就越显著

A Goodness of Fit Measure

- The coefficient of determination (**判定系数**), denoted by R^2, measures goodness of fit:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- $0 \leq R^2 \leq 1$
- The closer R^2 is to 1, the better the fit.
- 可以解释为 y_i 的总变化量 SST 中被线性回归方程所描述的比例
- 反映了回归方程对数据的拟合程度, 是衡量拟合优劣的重要统计量

An Interesting Fact

- $R^2 = r^2_{XY}$ (R^2 is squared correlation coefficient)

$$R^2 = \frac{\sum_i \left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_i (Y_i - \bar{Y})^2}$$

$$= \frac{\sum_i (b_0 + b_1 X_i - b_0 - b_1 \bar{X})^2}{\sum_i (Y_i - \bar{Y})^2}$$

$$= \frac{b_1^2 \sum_i (X_i - \bar{X})^2}{\sum_i (Y_i - \bar{Y})^2} = \frac{b_1^2 S_X^2}{S_Y^2} = r_{XY}^2$$

- No surprise: the higher the sample *correlation coefficient* between X and Y, the better we can do in our regression.

Prediction and the Modeling Goal

- The LS line: a prediction rule $\hat{Y} = b_0 + b_1 X$
  - $\hat{Y}$ is not a perfect prediction
- Forecast accuracy:
  - What value of Y can we expect for a new X?
  - How sure are we about his forecast?
- Our goal is to measure the accuracy of (or equivalently, the uncertainty in) our forecasts.

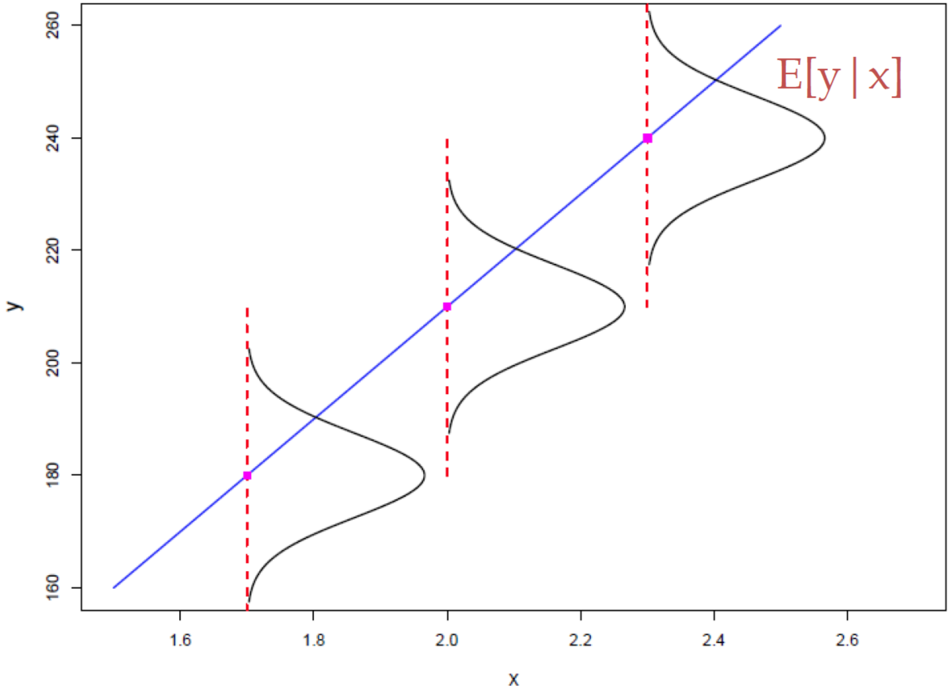
Prediction interval: Probable range for Y given X

Prediction and the Modeling Goal

- **Key Insight**: To construct a prediction interval, we will have to build a probability model, and assess the likely range of error values corresponding to a Y value that has not yet been observed.
- We need to work with the notion of a "true line" and a probability distribution that describes deviation around the line.

Simple Linear Regression Model

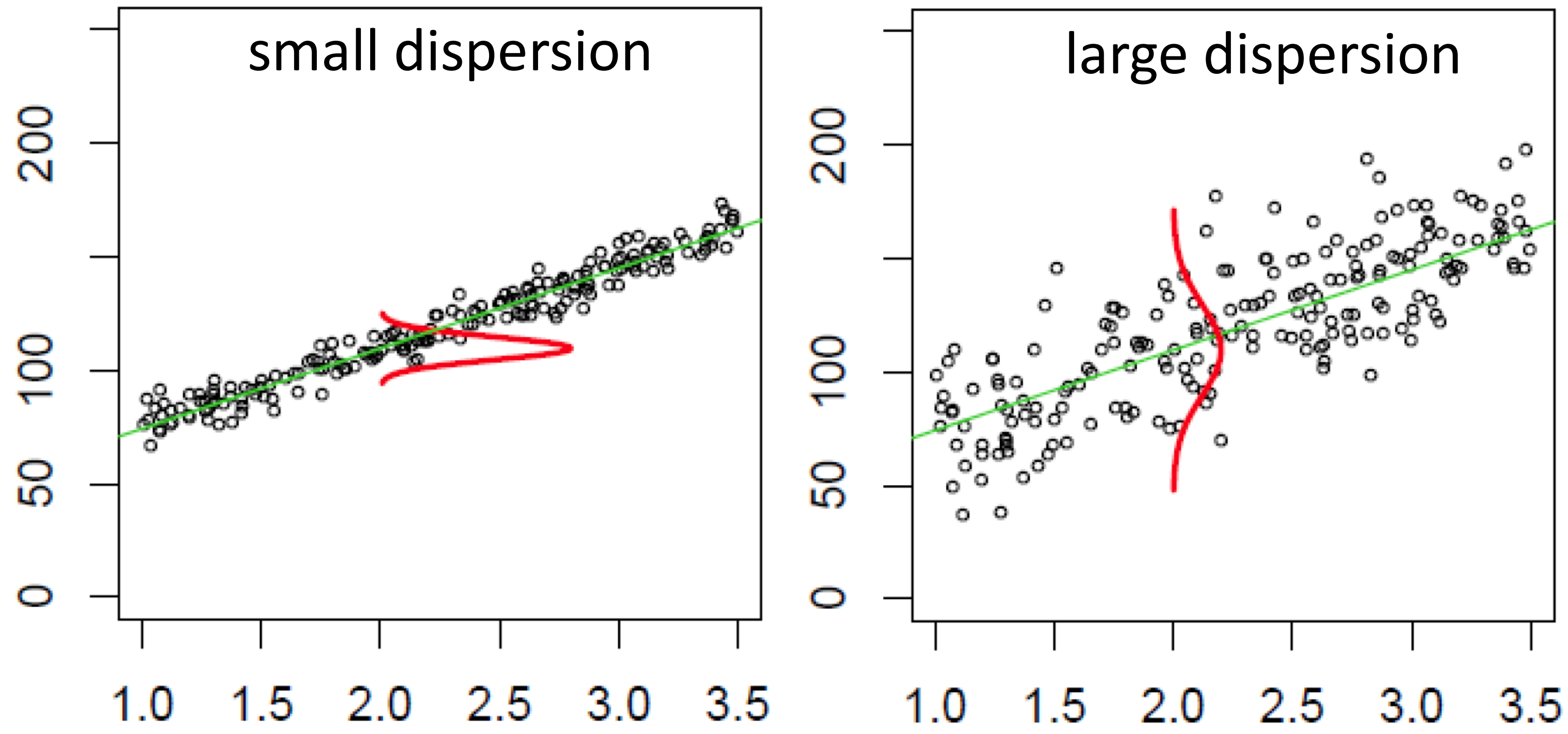
- Assume $Y = \beta_0 + \beta_1 X + \varepsilon$, $\varepsilon\ i.i.d. \sim N(0, \sigma^2)$
  - 3 parameters: β_0 and β_1: (linear pattern),σ (variation around the line)



- $\beta_0 + \beta_1 X$ represents the "true line"; the part of Y that depends on X
- The error term ε is independent "noise"; the part of Y not associated with X

Conditional Distributions

- $(Y|X = x) \sim N(\beta_0 + \beta_1 x, \sigma^2)$
- With prob. 95%, given x,



Estimation of the SLR Model

- We use least squares to estimate β_0 and β_1

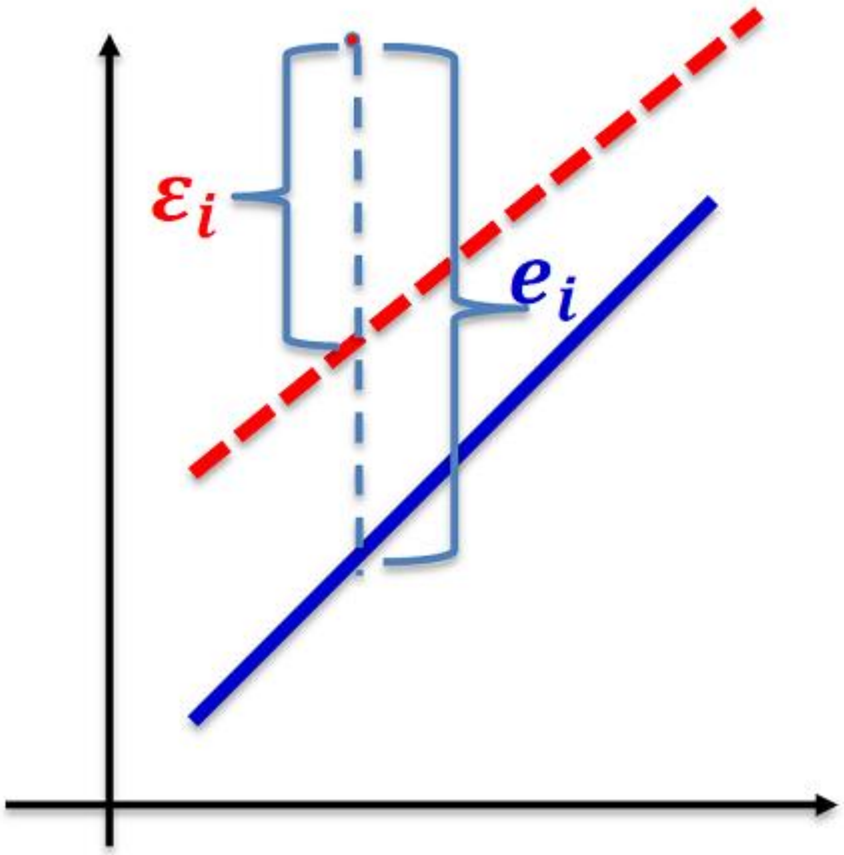
$$\hat{\beta}_1 = b_1 = r_{XY} \times \frac{S_Y}{S_X}$$

$$\hat{\beta}_0 = b_0 = \bar{Y} - b_1\bar{X}$$

Tue line: $E[Y|X] = \beta_0 + \beta_1 X$

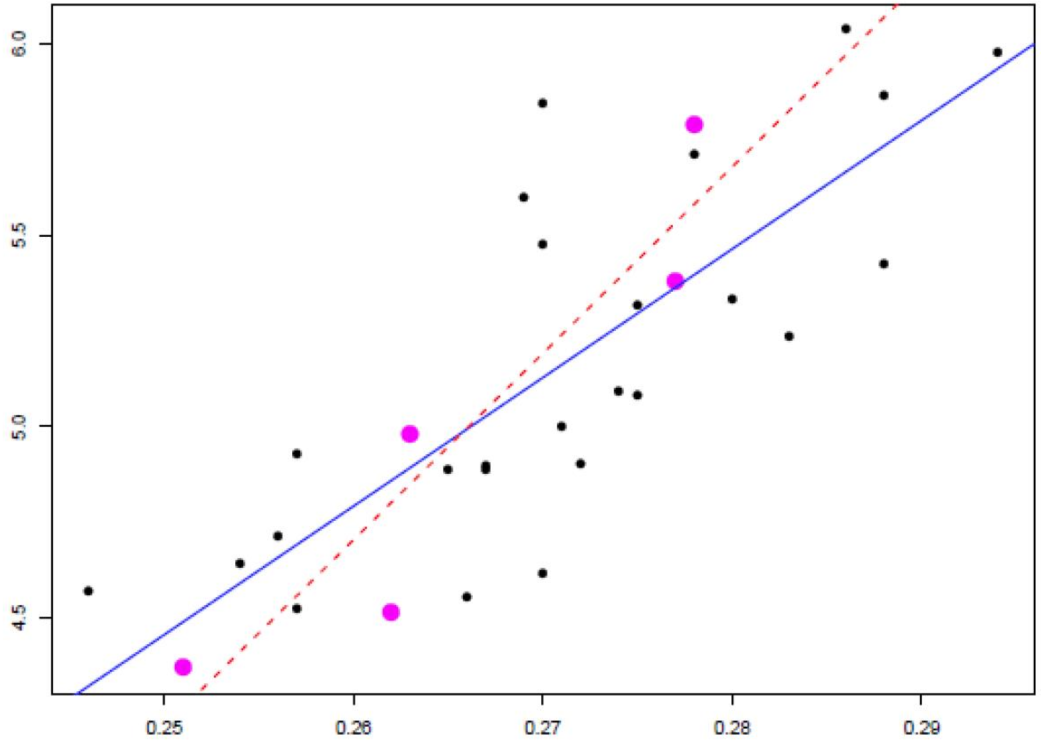
Least square line: $\hat{Y} = b_0 + b_1 X$

NOTE: $b_0 \neq \beta_0, b_1 \neq \beta_1, e_i \neq \varepsilon_i$



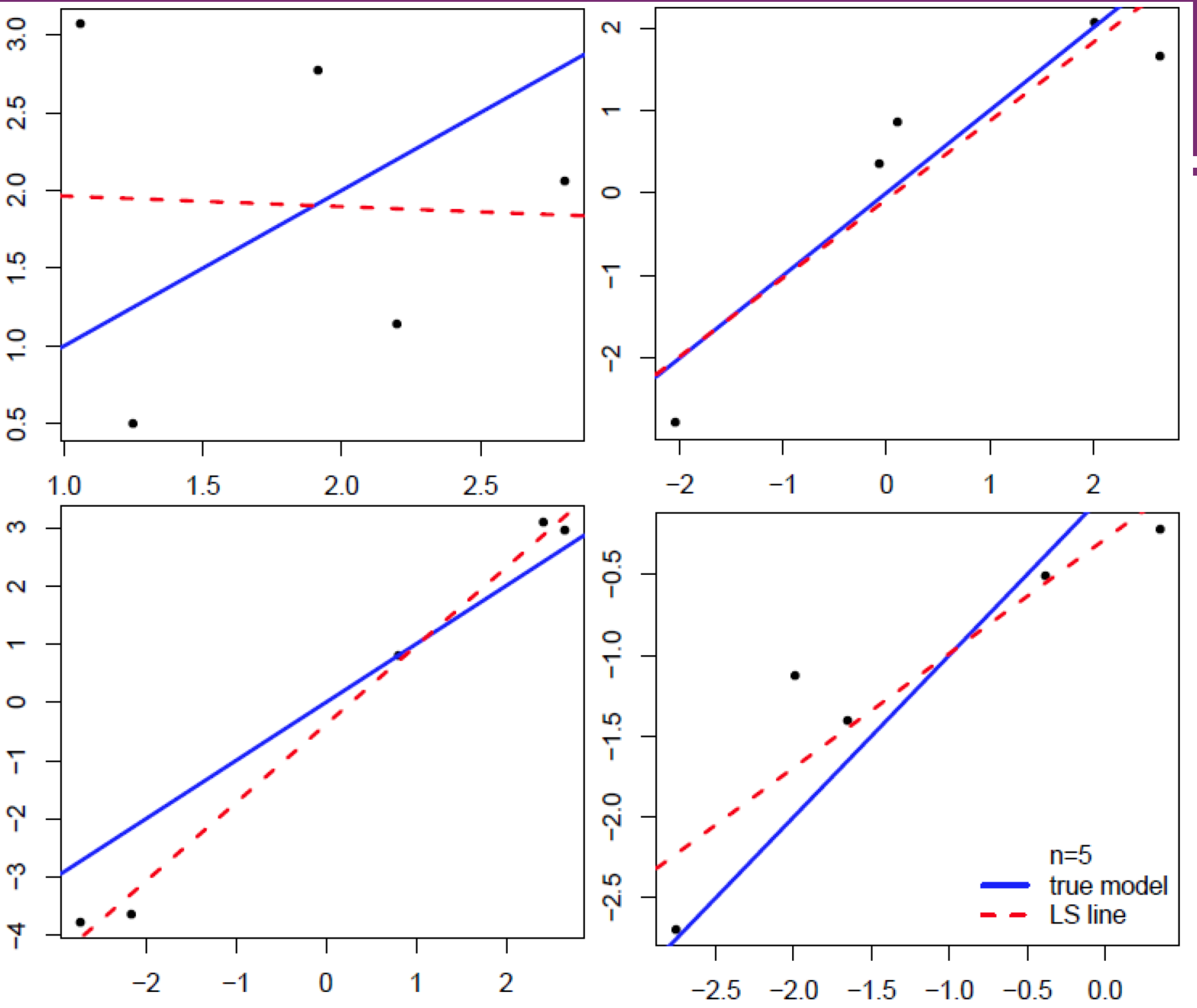
Prediction and the Modeling Goal

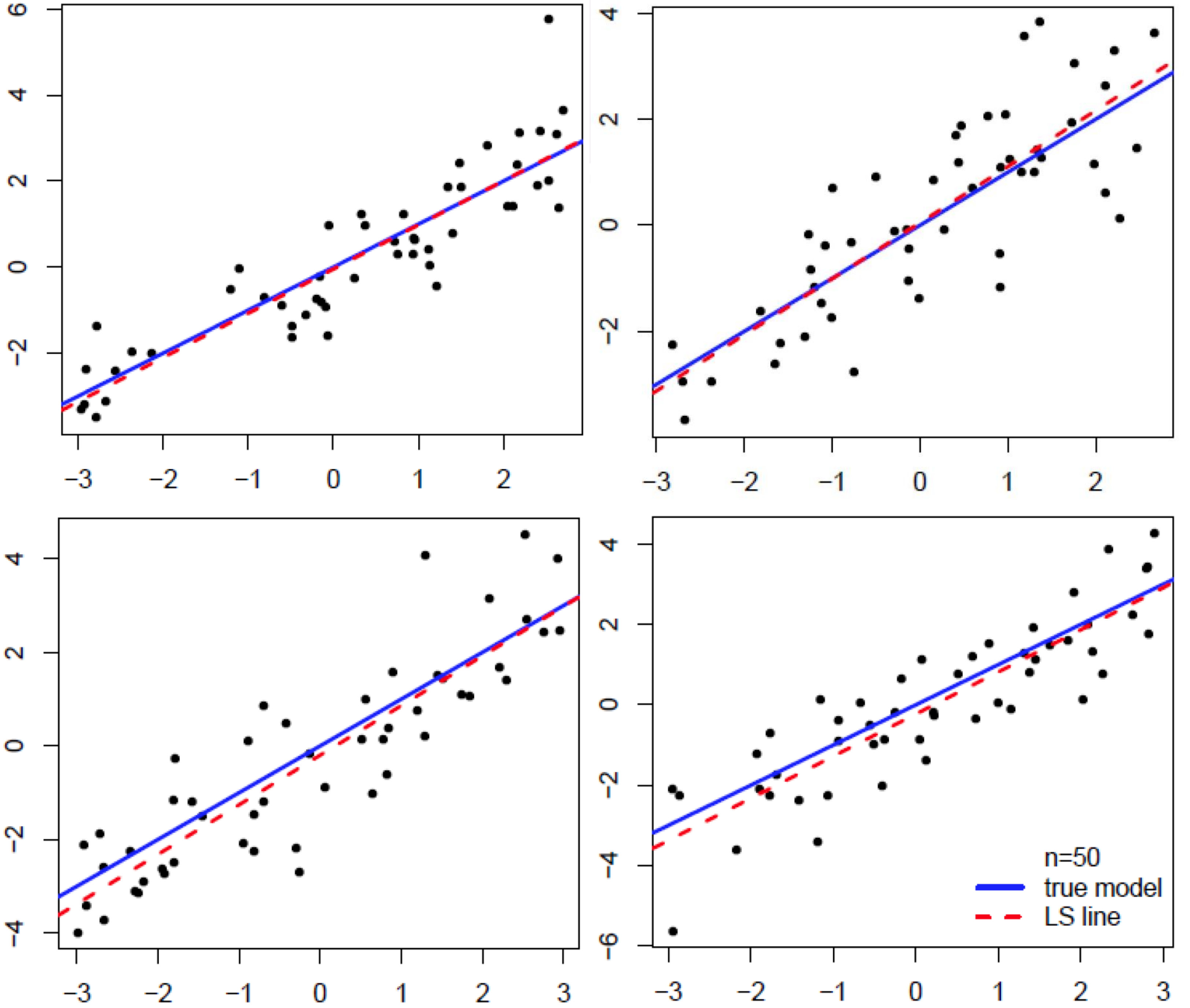
- Note that the “fitted” line may be fooled by particular realization of the residuals.
- Dashed line: fits the purple points only
- Solid line: fits all points
- Which line is better?



Sampling Distribution of LSE

- How much do our estimates depend on the particular random sample that we observe?
- Randomly draw different samples of the same size, and compute the estimated parameters for each sample.
- If the estimates do not vary much from sample to sample, then it does not matter which sample to choose. Vice versa.


n=5
true model
LS line


n=50
true model
LS line

Sampling Distribution of LSE

- The LS lines are much closer to the true line with larger sample size.
- When n=5, some lines are close, others are not. We need to get “lucky”.
- Also notice that the LS lines are, more often than not:
  - Closer to the true line near middle of the data cloud
  - They get farther apart away from the middle of the cloud

Q&A?