线性回归

陈飞宇

fchen@cqu.edu.cn

办公室：软件学院529

实例1

	检验项目	测定结果		单位	参考范围
1.	白细胞计数	6.30		10^9/L	4.00 - 10.0
2.	中性粒细胞百分比	62.4		%	50.0 - 70.0
3.	淋巴细胞百分比	33.1		%	20.0 - 40
4.	单核细胞百分比	4.5		%	3.0 - 10.0
5.	中性粒细胞计数	4.10		10^9/L	2.00 - 7.0
6.	淋巴细胞计数	2.0		10^9	0.8 - 4.00
7.	单核细胞计数	0.20		10^9/L	0.12 - 0.8
8.	红细胞计数	4.33		10^12/L	4.09 - 5.74
9.	血红蛋白	117	↓	g/L	120 - 172
10.	红细胞压积	34.7	↓	%	38.0 - 50.8
11.	平均红细胞体积	80.0	↓	fL	83.9 - 99.1
12.	平均血红蛋白量	27.1	↓	pg	27.8 - 33.8
13.	平均血红蛋白浓度	338		g/L	320 - 355
14.	红细胞分布宽度CV	13.1		%	0.0 - 14.6
15.	血小板	287.0		10^9/L	85.0 - 303.0
16.	血小板平均分布宽度	10.9	↓	fL	12.0 - 22.0

图片来源于百度

回归分析（Regression Analysis）

职业：人类学家、气象学家、地理学家、统计学家、探险家

评价：比 10 个生物学家中的 9 个更懂数学和物理，

比 20 个 数学家中的 19 个更懂生物和医学。

Galton 在研究身高的遗传关系时发现了回归效应：

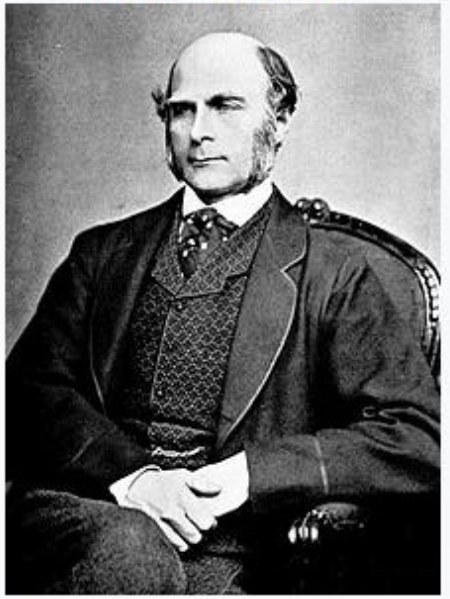
总的趋势是父亲的身高增加时，儿子的身高也倾向于增加。

1.父亲高于平均身高时，他们的儿子身高比他更高的概率要小于比他更矮的概率；

2.父亲矮于平均身高时，他们的儿子身高比他更矮的概率要小于比他更高的概率。

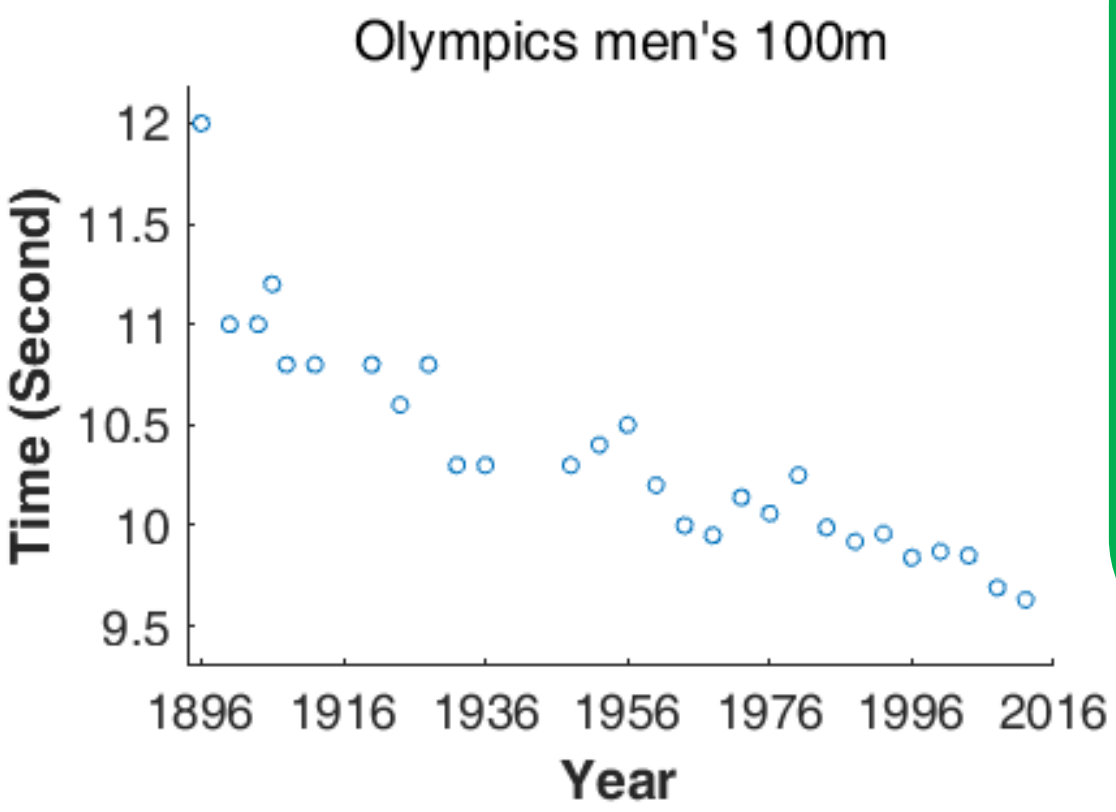
世界上最早研究大数据及其交叉学科的科学家之一

Sir Francis Galton



Born	16 February 1822 Birmingham, West Midlands, England
Died	17 January 1911 (aged 88) Haslemere, Surrey, England
Residence	England
Nationality	British
Alma mater	King's College London Trinity College, Cambridge

Olympics men’s 100m



用直线来描述年份与时间的关系：

$$y = w_1 x + w_0$$

哪条直线是最佳的呢？

如何确定最优的参数w1和w0？

损失函数

平方损失函数（Squared Loss Function）：

$$\mathcal{L}_i(w_0, w_1) = (y_i - (w_0 + w_1 x_i))^2$$

绝对损失函数（Absolute Loss Function）：

$$\mathcal{L}_i(w_0, w_1) = |y_i - (w_0 + w_1 x_i)|$$

平均损失函数（Average Loss Function）：

$$\mathcal{L}(w_0, w_1) = \frac{1}{n}\sum_i (y_i - (w_0 + w_1 x_i))^2$$

线性回归模型

线性回归模型 **(Linear Regression):**

$$\underset{w_0,w_1}{\operatorname{argmin}} \mathcal{L}(w_0, w_1) = \frac{1}{n}\sum_i (y_i - (w_0 + w_1 x_i))^2$$

把目标函数$\mathcal{L}(w_0, w_1)$展开，有

$$\mathcal{L} = \frac{1}{n}\sum_i (y_i^2 - 2y_i(w_0 + w_1 x_i) + w_0^2 + 2w_0 w_1 x_i + w_1^2 w_i^2)$$

线性回归模型

$\mathcal{L}(w_0, w_1)$关于w_0求导:

$$\mathcal{L} = \frac{1}{n}\sum_i (y_i^2 - 2y_i(w_0 + w_1 x_i) + w_0^2 + 2w_0 w_1 x_i + w_1^2 w_i^2)$$

$$\frac{\partial \mathcal{L}}{\partial w_0} = -2\left(\frac{1}{n}\sum_i y_i\right) + 2w_0 + 2w_1\left(\frac{1}{n}\sum_i x_i\right)$$

$$w_0^* = \left(\frac{1}{n}\sum_i y_i\right) - w_1^*\left(\frac{1}{n}\sum_i x_i\right)$$

线性回归模型

$\mathcal{L}(w_0, w_1)$关于w_1求导:

$$\mathcal{L} = \frac{1}{n}\sum_i (y_i^2 - 2y_i(w_0 + w_1 x_i) + w_0^2 + 2w_0 w_1 x_i + w_1^2 w_i^2)$$

$$\frac{\partial \mathcal{L}}{\partial w_1} = \frac{2}{n}\sum_i x_i(w_0 - y_i) + 2w_1\left(\frac{1}{n}\sum_i x_i^2\right)$$

$$w_1^* = -\frac{1}{n}\sum_i x_i(w_0^* - y_i) \Big/ \frac{1}{n}\sum_i x_i^2$$

线性回归模型

把w_0^*代入w_1^*的方程，化简后有:

$$w_1^* = \frac{\sum_i y_i(x_i - \frac{1}{n}\sum_i x_i)}{\sum_i x_i^2 - \frac{1}{n}(\sum_i x_i)^2}$$

$$w_0^* = \left(\frac{1}{n}\sum_i y_i\right) - w_1^*\left(\frac{1}{n}\sum_i x_i\right)$$

线性回归模型

令 $\bar{x}=\frac{1}{n}\sum_i x_i$, $\bar{y}=\frac{1}{n}\sum_i y_i$, $\overline{xy}=\frac{1}{n}\sum_i x_i y_i$, $\overline{x^2}=\frac{1}{n}\sum_i x_i^2$,则

$$w_1^*=\frac{\sum_i y_i\left(x_i-\frac{1}{n}\sum_i x_i\right)}{\sum_i x_i^2-\frac{1}{n}(\sum_i x_i)^2}=\frac{\frac{1}{n}(\sum_i y_i\left(x_i-\frac{1}{n}\sum_i x_i\right))}{\frac{1}{n}(\sum_i x_i^2-\frac{1}{n}(\sum_i x_i)^2)}=\frac{\overline{xy}-\bar{x}\,\bar{y}}{\overline{x^2}-(\bar{x})^2}$$

$$w_0^*=\left(\frac{1}{n}\sum_i y_i\right)-w_1^*\left(\frac{1}{n}\sum_i x_i\right)=\bar{y}-w_1^*\bar{x}$$

线性回归模型

对于给定数据，线性回归模型

$$\underset{w_0,w_1}{\operatorname{argmin}} \frac{1}{n}\sum_i (y_i - (w_0 + w_1 x_i))^2$$

的解为：

$$w_1^* = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - (\overline{x})^2}, \quad w_0^* = \overline{y} - w_1^* \overline{x}$$

人工例子

对下面一组数据进行线性回归分析：

x	1	3	5
y	4.8	11.3	17.2

解：

i	xi	yi	xi * yi	xi * xi
1	1	4.8	1 x 4.8	1 x 1
2	3	11.3	3 x 11.3	3 x 3
3	5	17.2	5 x 17.2	5 x 5
平均值				

例子取自 A first Course in Machine Learning– S. Rogers and M. Girolami.

人工例子

对下面一组数据进行线性回归分析：

x	1	3	5
y	4.8	11.3	17.2

解：

i	xi	yi	xi * yi	xi * xi
1	1	4.8	1 x 4.8	1 x 1
2	3	11.3	3 x 11.3	3 x 3
3	5	17.2	5 x 17.2	5 x 5
平均值	3	11.1	41.57	11.67

例子取自 A first Course in Machine Learning– S. Rogers and M. Girolami.

人工例子

对下面一组数据进行线性回归分析：

x	1	3	5
y	4.8	11.3	17.2

解：

i	xi	yi	xi * yi	xi * xi
平均值	3	11.1	41.57	11.67

$$w_1^* = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - (\overline{x})^2} = \frac{41.57 - 3 \times 11.1}{11.67 - (3)^2} = 3.1$$

$$w_0^* = \bar{y} - w_1^* \bar{x} = 11.1 - 3.1 \times 3 = 1.8$$

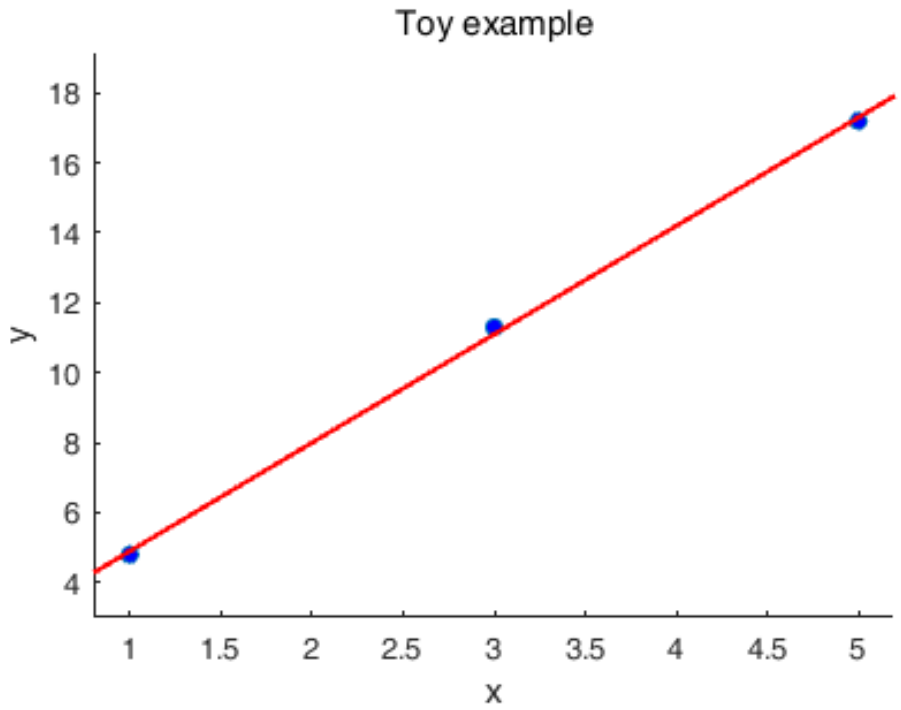
所求直线方程为：$y = 3.1\,x + 1.8$

例子取自 A first Course in Machine Learning– S. Rogers and M. Girolami.

人工例子

数据：

x	1	3	5
y	4.8	11.3	17.2



例子取自 A first Course in Machine Learning– S. Rogers and M. Girolami.

Olympics men's 100m

经过计算得到：

i	xi	yi	xi * yi	xi * xi
平均值	1954.5	10.36	20236.2	3.82×10^6

因此：

$$w_1^* = \frac{\overline{xy}-\overline{x}\,\overline{y}}{\overline{x^2}-(\overline{x})^2} = \frac{20236.2-1954.5\times 10.36}{3.82\times 10^6-(1954.5)^2} = -0.0133$$

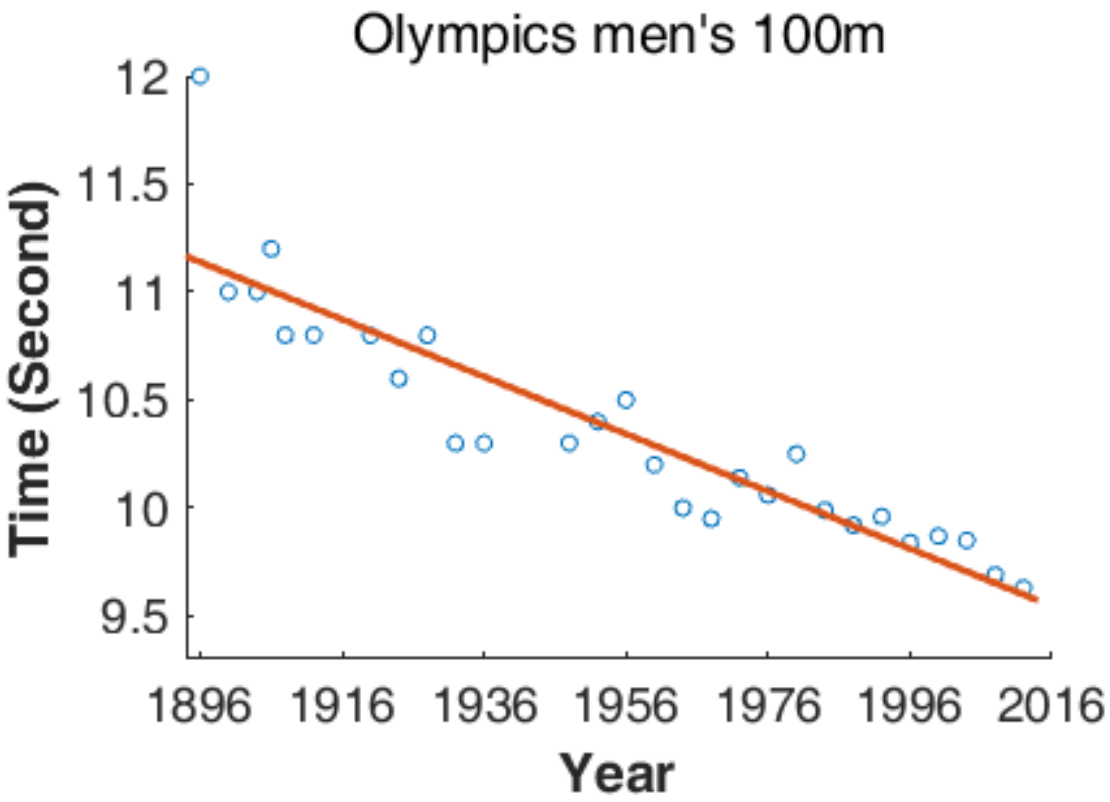
$$w_0^* = \overline{y} - w_1^*\overline{x} = 10.36 - 0.0133\times 1954.5 = 36.309$$

所求直线方程为： $y = -0.0133\,x + 36.309$

例子取自 A first Course in Machine Learning– S. Rogers and M. Girolami.

Olympics men's 100m

$$y = -0.0133\,x + 36.309$$



例子取自 A first Course in Machine Learning– S. Rogers and M. Girolami.

预测（Prediction）

留一法（Leave one out）：

把 $x = 2016, x = 2020$分别代入直线方程，

$$y = -0.0133\,x + 36.309$$

得到 $y = 9.55, y = 9.49$。

然而，2016年里约奥运会男子100米最好成绩是9.81s（博尔特）。

说明：

- 线性回归算法学习了一条直线，或者说两个参数(Paremeters)
- 我们可以利用线性回归算法对未知数据进行预测(Prediction)
- 线性回归的效果主要取决于数据本身的分布情况(Distribution)

多元线性回归

线性回归：寻求最优参数w_0, w_1，使得

$$y_i \approx w_0 + w_1 x_i = \hat{w}^{\mathrm{T}} \hat{x}_i$$

其中$\hat{w} = (w_0, w_1)^{\mathrm{T}}$, $\hat{x}_i = (1, x_i)^{\mathrm{T}}$.

考虑d维数据点 $x_i = (x_{i1}, x_{i2}, \dots, x_{id})$，

$$y_i \approx w_0 + (w_1, w_2, \dots, w_d) \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{pmatrix} = \hat{w}^{\mathrm{T}} \hat{x}_i$$

其中$\hat{w} = (w_0, w_1, \dots, w_d)^{\mathrm{T}}, \hat{x}_i = (1, x_{i1}, \dots, x_{id})^{\mathrm{T}}$.

多元线性回归

令 $y = (y_1, y_2, \cdots y_n)^{\mathrm{T}}$,

$$X = \begin{pmatrix} \widehat{x_1}^{\mathrm{T}} \\ \widehat{x_2}^{\mathrm{T}} \\ \vdots \\ \widehat{x_n}^{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1d} \\ 1 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix} \in \mathbb{R}^{n \times (d+1)}$$

$$y - X\hat{w} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} - \begin{pmatrix} \widehat{x_1}^{\mathrm{T}}\hat{w} \\ \widehat{x_2}^{\mathrm{T}}\hat{w} \\ \vdots \\ \widehat{x_n}^{\mathrm{T}}\hat{w} \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} - \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1d} \\ 1 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{pmatrix}$$

多元线性回归

因此，单变量线性回归模型

$$\underset{w_0,w_1}{\operatorname{argmin}} \frac{1}{n}\sum_i (y_i - (w_0 + w_1 x_i))^2$$

可以推广到多变量回归模型

$$\underset{\hat{w}}{\operatorname{argmin}} \mathcal{L}(\hat{w}) = \|y - X\hat{w}\|_2^2$$

该问题的最优解：

$$\hat{w}^* = (X^T X)^{-1} X^T y$$

人工例子（回顾）：

对下面一组数据进行线性回归分析：

x	1	3	5
y	4.8	11.3	17.2

解： $y = 3.1x + 1.8$

首先构造矩阵

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \end{pmatrix}, \qquad y = \begin{pmatrix} 4.8 \\ 11.3 \\ 17.2 \end{pmatrix}$$

例子取自 A first Course in Machine Learning– S. Rogers and M. Girolami.

人工例子（回顾）：

计算矩阵乘积$\boldsymbol{X^{\mathrm{T}}X}$及其逆矩阵$(\boldsymbol{X^{\mathrm{T}}X})^{-1}$：

$$\boldsymbol{X^{\mathrm{T}}X} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 5 \end{pmatrix} \begin{pmatrix} \mathbf{1} & \mathbf{1} \\ \mathbf{1} & \mathbf{3} \\ \mathbf{1} & \mathbf{5} \end{pmatrix} = \begin{pmatrix} \mathbf{3} & \mathbf{9} \\ \mathbf{9} & \mathbf{35} \end{pmatrix}$$

$$(\boldsymbol{X^{\mathrm{T}}X})^{-1} = \frac{\mathbf{1}}{\mathbf{3 \times 35 - 9 \times 9}} \begin{pmatrix} \mathbf{35} & \mathbf{-9} \\ \mathbf{-9} & \mathbf{3} \end{pmatrix}$$

利用公式 $\hat{\boldsymbol{w}}^* = (\boldsymbol{X^{\mathrm{T}}X})^{-1}\boldsymbol{X^{\mathrm{T}}y}$：

$$\hat{\boldsymbol{w}}^* = \frac{\mathbf{1}}{24} \begin{pmatrix} \mathbf{35} & \mathbf{-9} \\ \mathbf{-9} & \mathbf{3} \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 5 \end{pmatrix} \begin{pmatrix} \mathbf{4.8} \\ \mathbf{11.3} \\ \mathbf{17.2} \end{pmatrix} = \begin{pmatrix} 1.8 \\ 3.1 \end{pmatrix}$$

例子取自 A first Course in Machine Learning– S. Rogers and M. Girolami.

多元线性回归

对于给定数据，线性回归模型

$$\underset{w_0,w_1}{\operatorname{argmin}} \frac{1}{n}\sum_i (y_i - (w_0 + w_1 x_i))^2$$

的解为：公式1.

$$w_1^* = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - (\overline{x})^2}, \quad w_0^* = \overline{y} - w_1^* \overline{x}$$

公式2.

$$\hat{w}^* = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$$

概率解释（Probabilistic interpretation）

假设变量 y_i和变量x_i满足：

$$y_i = \widehat{w}^{\mathrm{T}}\widehat{x_i} + \varepsilon_i$$

其中误差 ε_i 服从高斯分布 $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$：

$$p(\varepsilon_i) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(\varepsilon_i)^2}{2\sigma^2}\right)$$

那么 $$p(y|x; \widehat{w}) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y_i-\widehat{w}^{\mathrm{T}}\widehat{x_i})^2}{2\sigma^2}\right)$$

似然函数（Likelihood function）

概率 $p(y_i|\hat{x}_i;\hat{w}) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(y_i-\hat{w}^{\mathrm{T}}\hat{x}_i)^2}{2\sigma^2})$ 被看做输出变量y_i关于输入变量$\hat{x}_i$和固定参数$\hat{w}$的函数。

同时，这个量也可以被当做在已知变量y_i和变量$\hat{x}_i$的前提下，关于参数$\hat{w}$的函数，即似然函数（Likelihood function）。

$$\mathcal{L}_i(\hat{w}) = p(y_i|\hat{x}_i;\hat{w}) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(y_i-\hat{w}^{\mathrm{T}}\hat{x}_i)^2}{2\sigma^2})$$

极大似然法（Maximum Likelihood）

假设所有数据都是独立同分布（independent and identically distributed，简称 i.i.d.）的，则

$$\mathcal{L}(\widehat{\boldsymbol{w}}) = \prod_{i=1}^{m} \mathcal{L}_i(\widehat{\boldsymbol{w}}) = \prod_{i=1}^{m} p(\boldsymbol{y}_i|\widehat{\boldsymbol{x}}_i; \widehat{\boldsymbol{w}})$$

极大似然法：令每个样本 $\widehat{\boldsymbol{x}}_i$输出为$\boldsymbol{y}_i$的概率越大越好。

$$\boxed{\underset{\widehat{\boldsymbol{w}}}{\operatorname{argmax}}\, \mathcal{L}(\widehat{\boldsymbol{w}})}$$

极大似然法（Maximum Likelihood）

考虑对数似然函数（Log-likelihood function）：

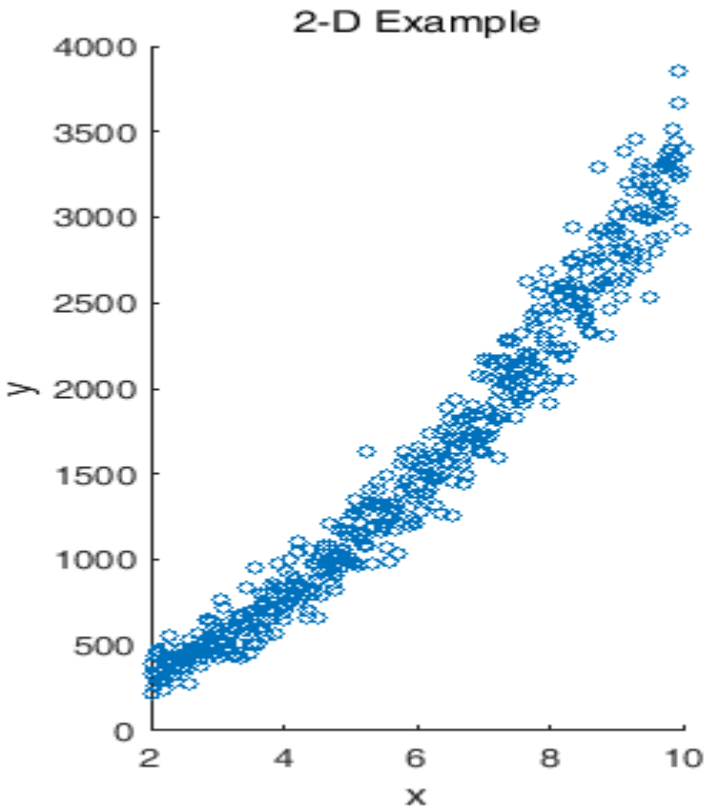
$$\ell(\widehat{\boldsymbol{w}}) = \ln \mathcal{L}(\widehat{\boldsymbol{w}}) = \ln \prod_{i=1}^{n} p(\boldsymbol{y}_i | \widehat{\boldsymbol{x}}_i; \widehat{\boldsymbol{w}})$$

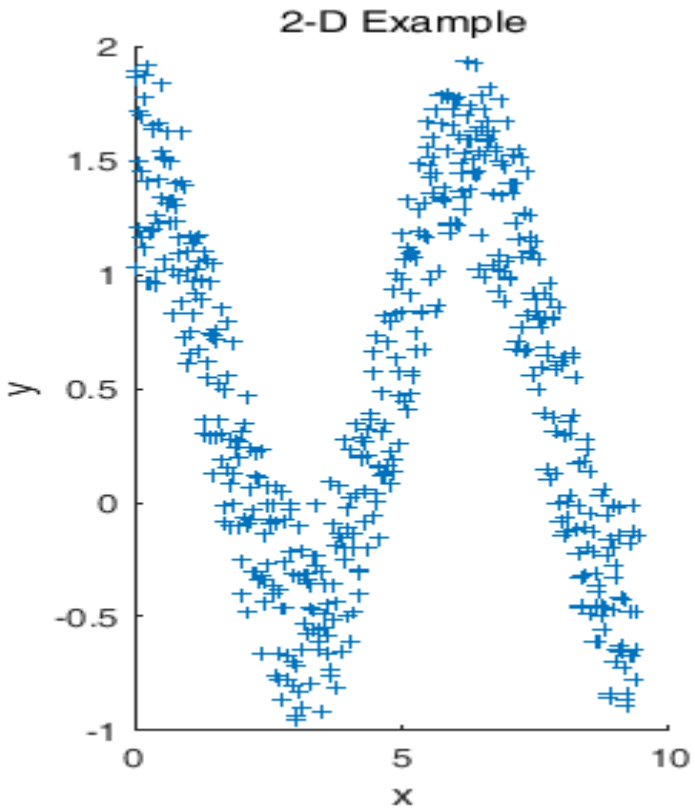
$$= \sum_{i=1}^{n} \ln p(\boldsymbol{y}_i | \widehat{\boldsymbol{x}}_i; \widehat{\boldsymbol{w}}) = \ln \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(\boldsymbol{y}_i - \widehat{\boldsymbol{w}}^{\mathrm{T}} \widehat{\boldsymbol{x}}_i\right)^2$$

因此，下面两个优化模型等价：

$$\underset{\widehat{\boldsymbol{w}}}{\operatorname{argmax}} \sum_{i=1}^{n} \ln p(\boldsymbol{y}_i | \widehat{\boldsymbol{x}}_i; \widehat{\boldsymbol{w}}) \quad \Leftrightarrow \quad \underset{\widehat{\boldsymbol{w}}}{\operatorname{argmin}} \sum_{i=1}^{n} \left(\boldsymbol{y}_i - \widehat{\boldsymbol{w}}^{\mathrm{T}} \widehat{\boldsymbol{x}}_i\right)^2$$

线性回归模型（延伸）


2-D Example
y
x
0
500
1000
1500
2000
2500
3000
3500
4000
2
4
6
8
10


2-D Example
y
x
-1
-0.5
0
0.5
1
1.5
2
0
5
10

广义线性回归模型

线性回归模型：

$$y = w^{\mathrm{T}}x + b$$

广义线性回归模型（Generalized linear model）：

$$g(y) = w^{\mathrm{T}}x + b$$

或者

$$y = g^{-1}(w^{\mathrm{T}}x + b)$$

其中 g 被称为联系函数（link function）。

线性回归小结 I

- 我们介绍了回归任务的思想以及如何建立模型。

  分析变量之间的关系、线性回归模型

- 我们定义了损失函数来评估线性回归模型的好坏。

  平方损失函数、极大似然函数

- 我们推导了线性回归模型中两个参数的显式表达。

  最小二乘法（向量与矩阵形式）

线性回归小结 II

- 我们使用了线性回归算法对两组数据进行分析。

人工例子、奥运会男子100米

- 我们应用了线性回归算法进行预测，并检验结果。

2016、2020奥运会男子100米

- 我们描述了线性回归模型的概率解释。

极大似然法

练习题

对于线性回归问题，给定

$$w_0^* = -\left(\frac{1}{n}\sum_i y_i\right) + w_1^*\left(\frac{1}{n}\sum_i x_i\right)$$

$$w_1^* = -\sum_i x_i(w_0^* - y_i)\Big/\frac{1}{n}\sum_i x_i^2$$

试推导：

$$w_1^* = \frac{\sum_i y_i(x_i - \frac{1}{n}\sum_i x_i)}{\sum_i x_i^2 - \frac{1}{n}(\sum_i x_i)^2}$$

练习题

对于一维数据，试证明下列两种解法的等价性。

公式1. $$w_1^* = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2}, \quad w_0^* = \bar{y} - w_1^* \bar{x}$$

公式2. $$\hat{w}^* = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$$

练习题

用 Matlab 实现线性回归的算法：

1. 构造人工数据。 提示：（x,y）要呈线性分布。
2. 利用公式1和公式2求出直线方程。
3. 评价两种方法的优劣（运行时间、目标函数等）
4. 画图。（画出原始数据点云、直线）