

24

Linear Regression

线性回归

以概率统计、几何、矩阵分解、优化为视角



我们必须承认，有多少数字，就有多少正方形。

We must say that there are as many squares as there are numbers.

—— 伽利略·伽利莱 (Galilei Galileo) | 意大利物理学家、数学家及哲学家 | 1564 ~ 1642



- ▶ matplotlib.pyplot.quiver() 绘制箭头图
- ▶ numpy.cov() 计算协方差矩阵
- ▶ seaborn.heatmap() 绘制热图
- ▶ seaborn.jointplot() 绘制联合分布/散点图和边际分布
- ▶ seaborn.kdeplot() 绘制 KDE 核概率密度估计曲线
- ▶ seaborn.pairplot() 绘制成对分析图
- ▶ statsmodels.api.add_constant() 线性回归增加一列常数 1
- ▶ statsmodels.api.OLS() 最小二乘法函数

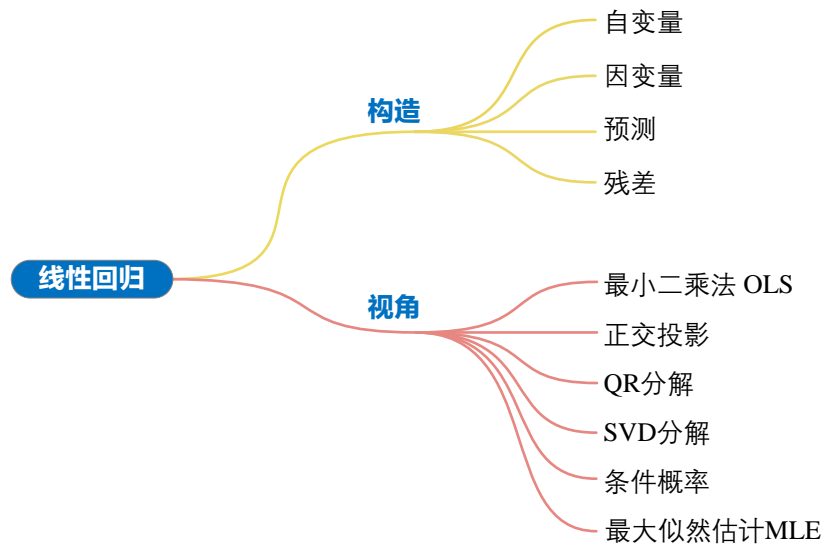
本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com



24.1 再聊线性回归

线性回归 (linear regression) 是最为常用的回归建模技术。它是利用线性关系建立因变量与一个或多个自变量之间的联系。线性回归模型相对简单，可解释性强，应用广泛。

本系列丛书从不同视角介绍过线性回归。比如，《数学要素》从代数、几何、优化角度讲过线性回归，《矩阵力量》则从线性代数、正交投影、矩阵分解视角分析线性回归。本章一方面总结这几个视角，另外一方面以条件概率、MLE 为视角再谈线性回归。

有监督学习

《矩阵力量》一书提过，线性回归是一种**有监督学习** (supervised learning)。有监督学习是一种机器学习方法，它利用已知的标签或输出值来训练模型，并用于预测未知的标签或输出值。在有监督学习中，我们通常会提供一组已知的训练样本，每个样本都包含一组输入特征和相应的输出标签。模型通过分析这些训练样本来学习如何将输入特征映射到输出标签，从而能够用于预测未知的输出值。

有监督学习通常分为两个主要的子类别：**分类** (classification) 和**回归** (regression)。在分类问题中，目标是将输入特征映射到有限的离散类别。在回归问题中，目标是将输入特征映射到连续的输出值。《数据有道》将介绍更多有关回归算法，而《机器学习》将关注常见分类算法。

简单线性回归

简单线性回归 (Simple Linear Regression, SLR) 也叫**一元线性回归** (univariate linear regression)，是指模型中只含有一个自变量和一个因变量，表达式如下：

$$y = \underbrace{b_0 + b_1 x}_{\hat{y}} + \varepsilon \quad (1)$$

其中， b_0 为**截距项** (intercept)， b_1 代表**斜率** (slope)。

x 又常被称作**自变量** (independent variable)、**解释变量** (explanatory variable) 或**回归元** (regressor)、**外生变量** (exogenous variables)、**预测变量** (predictor variables)；

y 常被称作**因变量** (dependent variable)、**被解释变量** (explained variable)、或**回归子** (regressand)、**内生变量** (endogenous variable)、**响应变量** (response variable) 等。

ε 为**残差项** (residuals)、**误差项** (error term)、**干扰项** (disturbance term) 或**噪音项** (noise term)。

图 1 所示为平面上的一个线性回归关系。

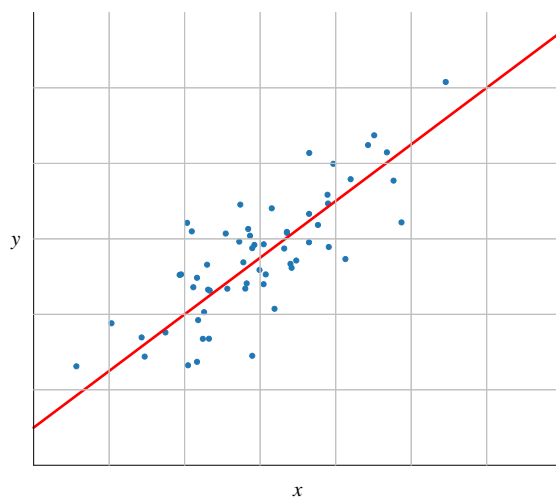


图 1. 平面上，一元线性回归

预测

利用 (1) 做预测，预测值 \hat{y} 为：

$$\hat{y} = b_0 + b_1 x \quad (2)$$

▲ 注意，“戴帽子”的 \hat{y} 表示预测值。

(2) 对应图 1 中的红色直线。

对于第 i 个数据点，预测值 $\hat{y}^{(i)}$ 可以通过下式计算得到：

$$\hat{y}^{(i)} = b_0 + b_1 x^{(i)} \quad (3)$$

残差

(1) 中残差项为：

$$\varepsilon = y - (b_0 + b_1 x) = y - \hat{y} \quad (4)$$

如图 2 所示，在平面上，残差项是 y 和 \hat{y} 之间的纵轴上的高度差。

▲ 注意，平面上，线性回归和主成分分析的结果看上去都是一条直线。但是两者差距甚远。线性回归是有监督学习，而主成分分析是无监督学习。从距离角度来看，线性回归关注的是沿纵轴的高度差，而主成分分析则是聚焦点到直线的距离。此外，从椭圆的角度来看，主成分分析对应椭圆的长轴、短轴，而线性回归则和与椭圆相切的矩形有关。“鸢尾花书”《编程不难》聊过这个话题，请大家回顾。

真实观察值 $y^{(i)}$ 和预测值 $\hat{y}^{(i)}$ 之差为第 i 个数据点的残差：

$$\varepsilon^{(i)} = y^{(i)} - \hat{y}^{(i)} \quad (5)$$

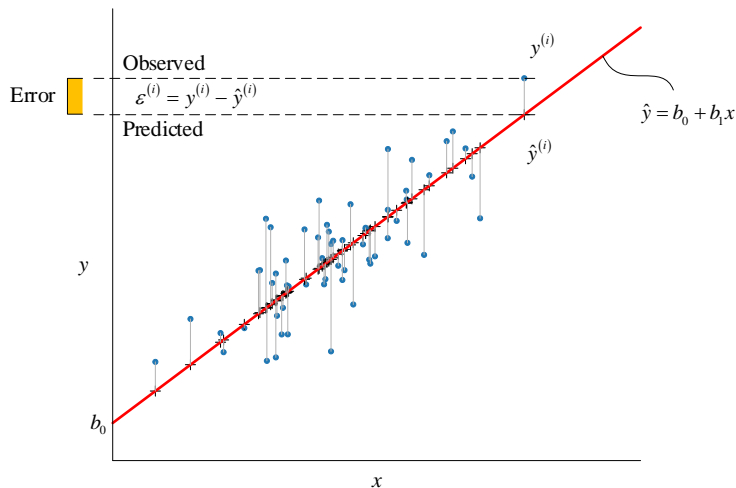


图 2. 简单线性回归中的残差项

矩阵形式

使用阵运算表达一元线性回归：

$$\mathbf{y} = b_0 \mathbf{I} + b_1 \mathbf{x} + \varepsilon \quad (6)$$

\mathbf{I} 为和 \mathbf{x} 形状相同的全 1 列向量；自变量数据 \mathbf{x} 、因变量数据 \mathbf{y} 和残差项 ε 分别包括 n 个样本对应的列向量为：

$$\mathbf{x} = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(n)} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon^{(1)} \\ \varepsilon^{(2)} \\ \vdots \\ \varepsilon^{(n)} \end{bmatrix} \quad (7)$$

图 3 解释 (6) 给出的矩阵运算。

$$\begin{array}{c} \mathbf{y} \\ \text{red column} \\ n \times 1 \end{array} = \begin{array}{c} b_0 \mathbf{I} \\ \text{grey column} \\ n \times 1 \end{array} + \begin{array}{c} b_1 \mathbf{x} \\ \text{blue column} \\ n \times 1 \end{array} + \begin{array}{c} \varepsilon \\ \text{yellow column} \\ n \times 1 \end{array}$$

图 3. 用矩阵运算表达一元回归

预测值构成的列向量 $\hat{\mathbf{y}}$ 为：

$$\hat{y} = b_0 I + b_1 x \tag{8}$$

如图 4 所示， \hat{y} 是 I 和 x 的线性组合。

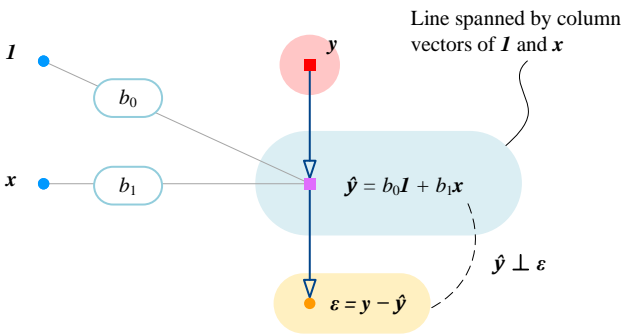


图 4. 一元最小二乘法线性回归数据关系

残差项列向量 ϵ 为：

$$\epsilon = y - \hat{y} \tag{9}$$

图 5 可视化求解残差项列向量 ϵ 过程。

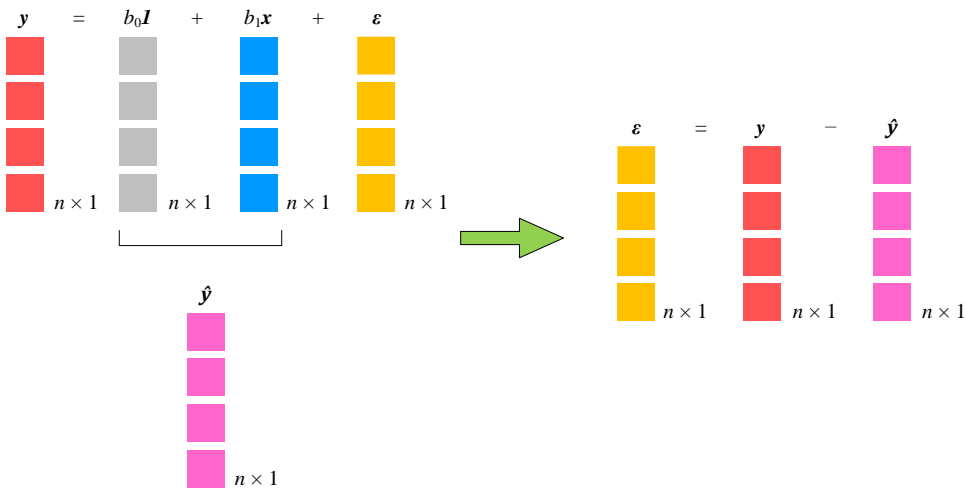


图 5. 求解残差项列向量

问题来了，如何确定参数 b_0 、 b_1 ？

24.2 最小二乘法

最小二乘法 (ordinary least squares, OLS) 通过**最小化残差值平方和** (sum of squared estimate of errors, SSE), 来计算得到最佳的拟合回归线参数:

$$\arg \min_{b_0, b_1} \text{SSE} = \arg \min_{b_0, b_1} \sum_{i=1}^n \left(\varepsilon^{(i)} \right)^2 \quad (10)$$

残差平方和 SSE 为:

$$\text{SSE} = \sum_{i=1}^n \left(\varepsilon^{(i)} \right)^2 = \sum_{i=1}^n \left(y^{(i)} - \hat{y}^{(i)} \right)^2 \quad (11)$$

⚠ 注意, “鸢尾花书”用 SSE 表达残差值平方和; 也有很多文献使用 RSS (residual sum of squares) 代表残差值平方和。

从几何角度, 图 6 中的每一个正方形的边长为 $\varepsilon^{(i)}$, 该正方形的面积代表一个残差平方项 $\left(\varepsilon^{(i)} \right)^2$; 图 6 所有正方形面积之和便是残差平方和 SSE。

➡ 我们在《数学要素》第 24 章聊过残差平方和 SSE 可以写成一个二元函数 $f(b_0, b_1)$ 。 $f(b_0, b_1)$ 对应的图像如图 7 所示。

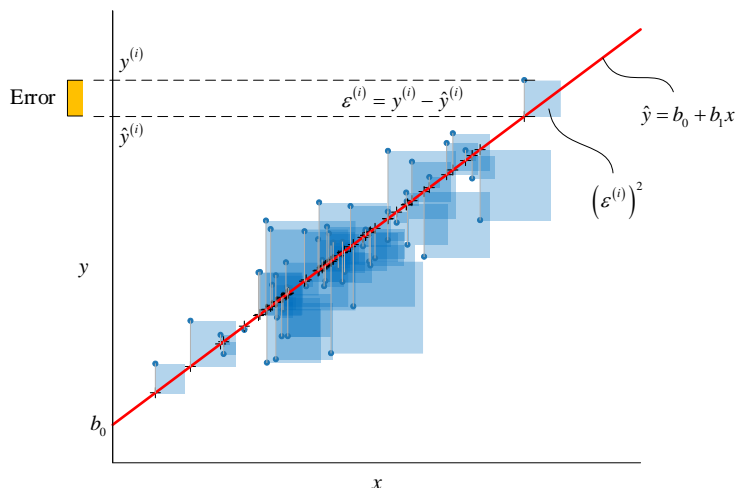
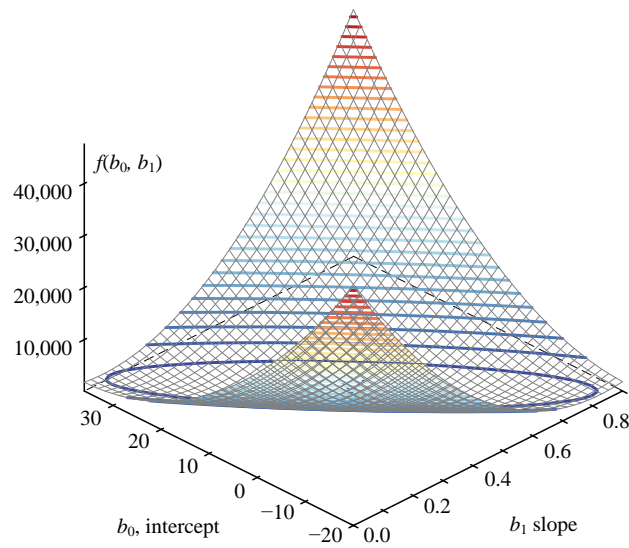


图 6. 残差平方和的几何意义

图 7. 误差平方和 SSE 随 b_0 、 b_1 变化构造的开口向上抛物曲面，图片来自《数学要素》第 24 章

24.3 优化问题

用线性代数工具构造 OLS 优化问题：

$$\arg \min_b \|y - Xb\| \quad (12)$$

也可以写成：

$$\arg \min_b \|\varepsilon\|^2 = \varepsilon^\top \varepsilon \quad (13)$$

令

$$b = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}, \quad X = \begin{bmatrix} I & x \end{bmatrix} \quad (14)$$

其中， X 又叫**设计矩阵** (design matrix)。

\hat{y} 可以写成：

$$\hat{y} = Xb \quad (15)$$

残差向量 ε 可以写成：

$$\varepsilon = y - b_0 I - b_1 x = y - Xb \quad (16)$$

定义 $f(b)$ 为：

$$f(b) = \varepsilon^\top \varepsilon = (y - Xb)^\top (y - Xb) \quad (17)$$

$f(\mathbf{b})$ 对 \mathbf{b} 求一阶导为 $\mathbf{0}$ 得到等式：

$$\frac{\partial f(\mathbf{b})}{\partial \mathbf{b}} = 2\mathbf{X}^T \mathbf{X} \mathbf{b} - 2\mathbf{X}^T \mathbf{y} = \mathbf{0} \quad (18)$$

如果 $\mathbf{X}^T \mathbf{X}$ 可逆，则 \mathbf{b} 为：

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (19)$$

24.4 投影视角



《矩阵力量》一本特别强调 OLS 的投影视角。

如图 8 所示，在 \mathbf{I} 和 \mathbf{x} 撑起平面 H 上，向量 \mathbf{y} 的投影为 $\hat{\mathbf{y}}$ ，而残差 $\boldsymbol{\varepsilon}$ 垂直于这个平面：

$$\begin{aligned} \boldsymbol{\varepsilon} \perp \mathbf{I} &\Rightarrow \mathbf{I}^T \boldsymbol{\varepsilon} = 0 \Rightarrow \mathbf{I}^T (\mathbf{y} - \hat{\mathbf{y}}) = 0 \\ \boldsymbol{\varepsilon} \perp \mathbf{x} &\Rightarrow \mathbf{x}^T \boldsymbol{\varepsilon} = 0 \Rightarrow \mathbf{x}^T (\mathbf{y} - \hat{\mathbf{y}}) = 0 \end{aligned} \quad (20)$$

以上两式合并：

$$\underbrace{\begin{bmatrix} \mathbf{I} & \mathbf{x} \end{bmatrix}}_{\mathbf{X}}^T (\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{0} \quad (21)$$

整理得到：

$$\mathbf{X}^T \mathbf{y} = \mathbf{X}^T \mathbf{X} \mathbf{b} \quad (22)$$

这和 (18) 一致。

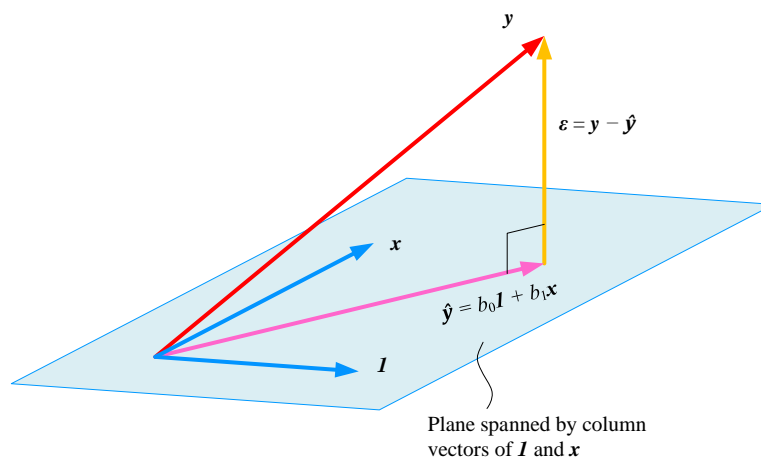


图 8. 几何角度解释一元最小二乘结果，二维平面

24.5 线性方程组：代数视角

实际上，下式就是一个**超定方程组** (overdetermined system)：

$$\mathbf{y} = \mathbf{X}\mathbf{b} \quad (23)$$

QR 分解

对 \mathbf{X} 进行 QR 分解得到：

$$\mathbf{X} = \mathbf{Q}\mathbf{R} \quad (24)$$

这样求得 \mathbf{b} 为：

$$\mathbf{b} = \mathbf{R}^{-1}\mathbf{Q}^T\mathbf{y} \quad (25)$$

奇异值分解

对 \mathbf{X} 进行完全型 SVD 分解得到：

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T \quad (26)$$

这样求得 \mathbf{b} 为：

$$\mathbf{b} = \mathbf{V}\mathbf{S}^{-1}\mathbf{U}^T\mathbf{y} \quad (27)$$

《矩阵力量》一册介绍过 $\mathbf{V}\mathbf{S}^{-1}\mathbf{U}^T$ 是 \mathbf{X} 的**摩尔-彭若斯广义逆** (Moore–Penrose inverse)。 \mathbf{S}^{-1} 的主对角线非零元素为 \mathbf{S} 的非零奇异值倒数， \mathbf{S}^{-1} 其余对角线元素均为 0。

24.6 条件概率

条件期望



本书第 12 章介绍过，线性回归还可以从条件概率视角来看。

如果随机变量 (X, Y) 服从二元高斯分布，给定 $X = x$ 条件下， Y 的条件期望为：

$$\mu_{Y|X=x} = \text{cov}(X, Y) \left(\sigma_X^2 \right)^{-1} (x - \mu_X) + \mu_Y = \rho_{X,Y} \frac{\sigma_Y}{\sigma_X} (x - \mu_X) + \mu_Y \quad (28)$$

这条回归直线的斜率为 $\rho_{X,Y}\sigma_Y/\sigma_X$ ，且通过点 (μ_X, μ_Y) 。

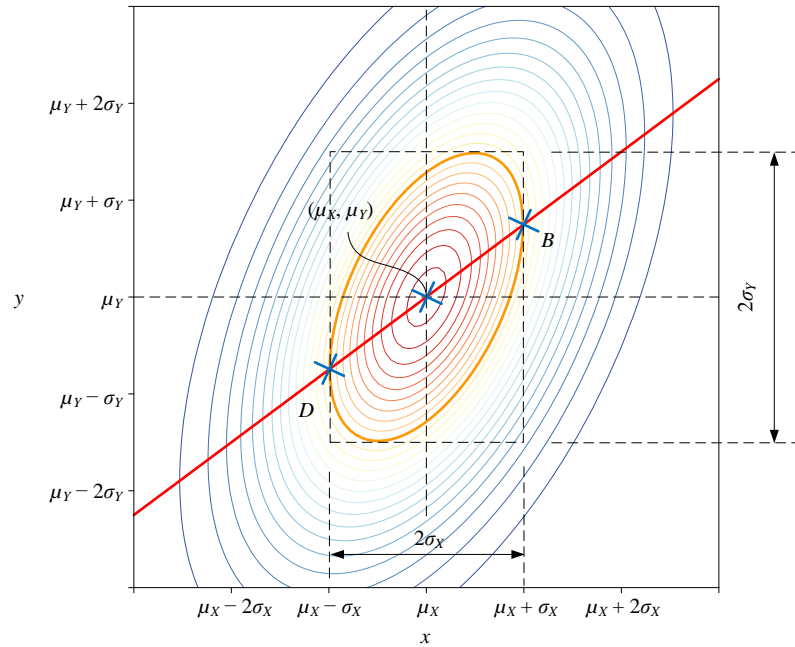


图 9. 给定 $X = x$ 的条件期望

图 10 所示为不同相关性系数条件下，回归直线和椭圆关系。

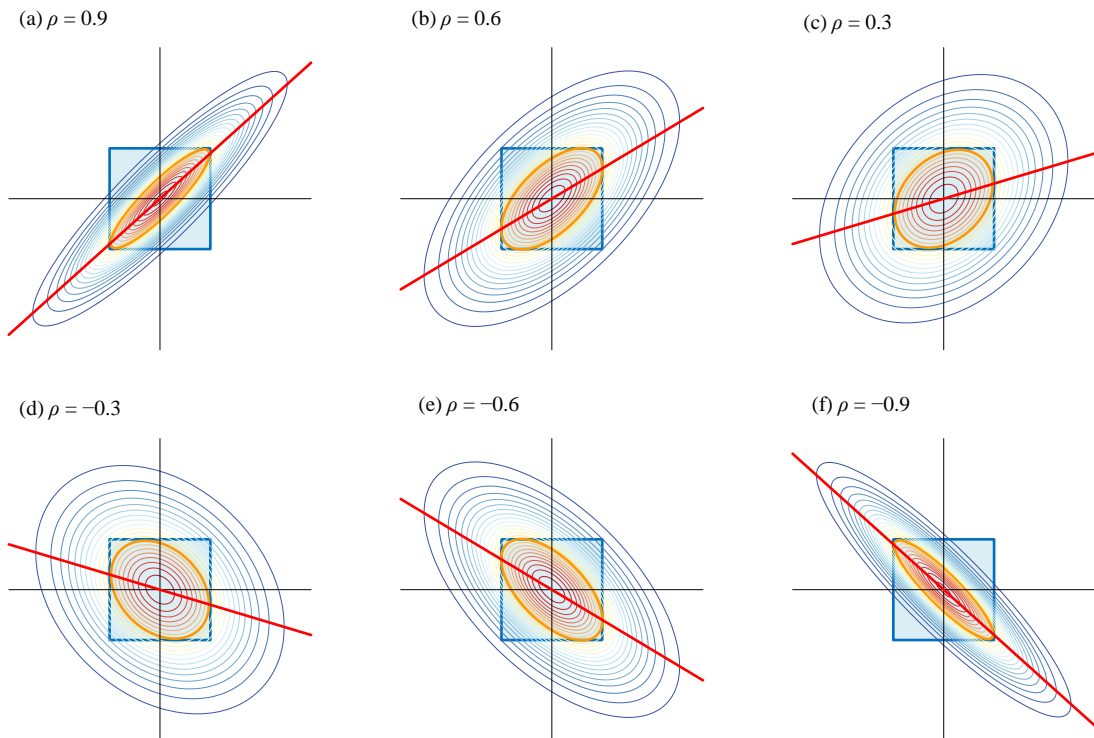


图 10. 条件期望直线位置和相关性系数关系， $\sigma_X = \sigma_Y$

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

以鸢尾花为例

定义鸢尾花花萼长度为 x ，鸢尾花花萼宽度为 y 。鸢尾花样本数据， x 和 y 的关系为：

$$y = 3.758 + 1.858 \left(x - 5.843 \right) \quad (29)$$

$\mu_X \quad \rho_{X,Y} \frac{\sigma_Y}{\sigma_X} \quad \mu_Y$

图 11 中散点为样本数据，其中直线代表花瓣长度、花萼长度之间回归关系。这幅图中，我们还绘制了马氏距离为 1 的椭圆。这个椭圆代表了花瓣长度、花萼长度的协方差矩阵。

图 12 所示为不考虑标签情况下，鸢尾花的成对特征图以及特征之间的回归关系。图 13 所示为考虑标签情况下，鸢尾花的成对特征图以及特征之间的回归关系。

⚠ 特别值得注意的是，两个随机变量之间的线性回归关系不代表两者存在“因果关系”。

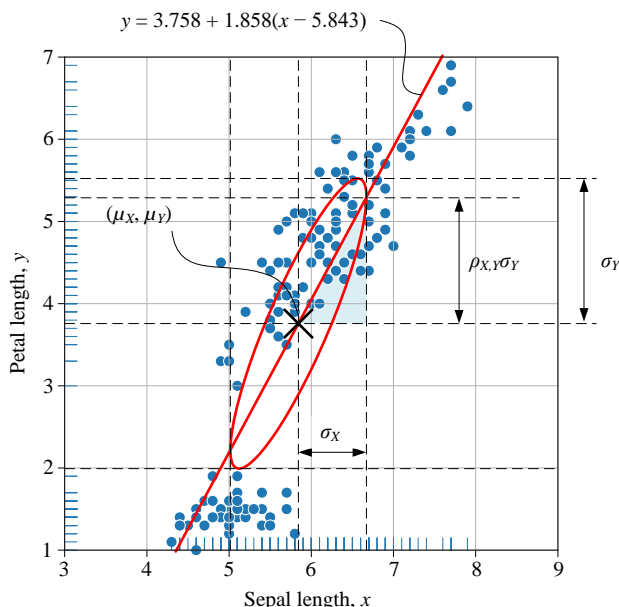


图 11. 花瓣长度、花萼长度之间回归关系



Bk5_Ch24_01.py 绘制图 11。

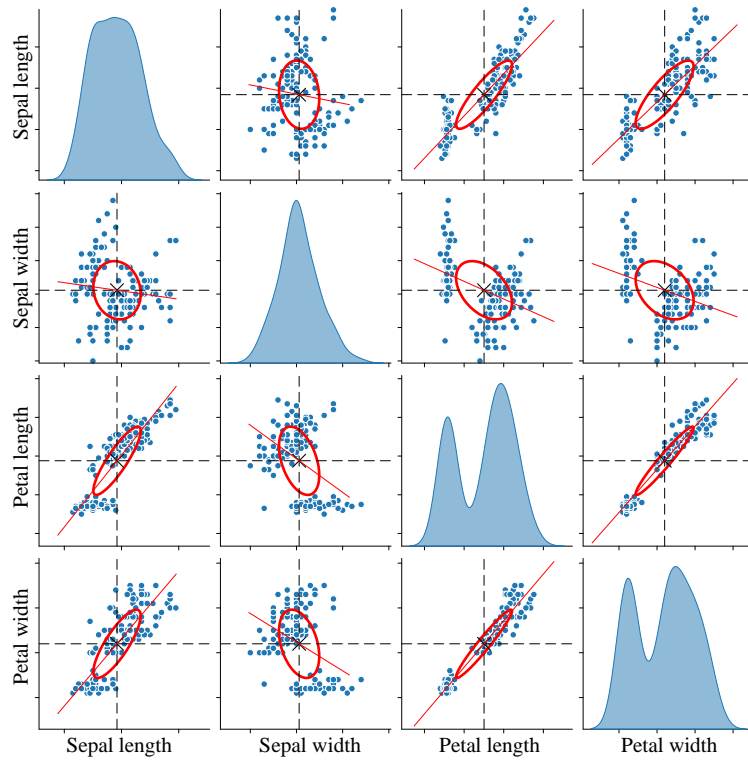


图 12. 成对特征图和回归关系

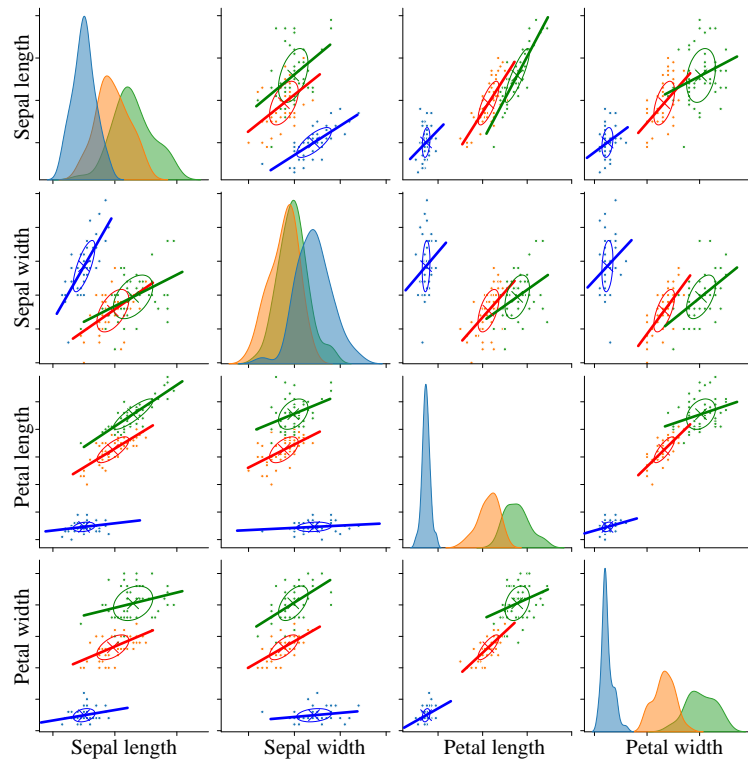
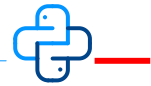


图 13. 成对特征图和回归关系，考虑分类标签



Bk5_Ch24_02.py 绘制图 12 和图 13。

24.7 最大似然估计 MLE

为了方便和本书前文有关最大似然估计内容对照阅读，本节中，线性回归解析式改写成：

$$y = \underbrace{\theta_0 + \theta_1 x}_{\hat{y}} + \varepsilon \quad (30)$$

对应的超定方程组写成：

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} \quad (31)$$

残差向量 $\boldsymbol{\varepsilon}$ 为：

$$\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\theta} \quad (32)$$

假设残差项服从正态分布：

$$\varepsilon \sim N(0, \sigma^2) \quad (33)$$

根据 (4)，也就是说 Y_i 服从：

$$Y_i \sim N(\theta_1 X_i + \theta_0, \sigma^2) \quad (34)$$

Y_i 的概率密度函数为：

$$f_{Y_i}(y_i; \theta_1 x_i + \theta_0, \sigma) = \frac{\exp\left(-\frac{(y_i - (\theta_1 x_i + \theta_0))^2}{2\sigma^2}\right)}{\sqrt{2\pi}\sigma} \quad (35)$$

似然函数可以写成：

$$L(\theta_0, \theta_1) = \prod_{i=1}^n \frac{\exp\left(-\frac{(y_i - (\theta_1 x_i + \theta_0))^2}{2\sigma^2}\right)}{\sqrt{2\pi}\sigma} \quad (36)$$

对数似然函数为：

$$\ln L(\theta_0, \theta_1) = -n \ln(\sqrt{2\pi}\sigma) - \frac{\sum_{i=1}^n (y_i - (\theta_1 x_i + \theta_0))^2}{2\sigma^2} \quad (37)$$

假设 σ 已知，最大化对数似然函数，等价于最小化 $\sum_{i=1}^n (y_i - (\theta_1 x_i + \theta_0))^2$ ，这和 (13) 优化问题一致。

$$\begin{aligned}\hat{\theta}_1 &= \frac{\sum_{i=1}^n (x^{(i)} - \mu_x)(y^{(i)} - \mu_y)}{\sum_{i=1}^n (x^{(i)} - \mu_x)^2} \\ \hat{\theta}_0 &= \mu_y - \hat{\theta}_1 \mu_x\end{aligned}\quad (38)$$

矩阵运算

假设残差服从正态分布 $N(0, \sigma^2)$ ，残差 $\varepsilon^{(i)}$ 对应的概率密度为：

$$f(\varepsilon^{(i)}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\varepsilon^{(i)})^2}{2\sigma^2}\right) \quad (39)$$

似然函数则可以写成：

$$L(\theta_0, \theta_1) = \prod_{i=1}^n \left\{ \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\varepsilon^{(i)})^2}{2\sigma^2}\right) \right\} = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^n (\varepsilon^{(i)})^2}{2\sigma^2}\right) \quad (40)$$

用矩阵运算表达上式得到：

$$L(\theta_0, \theta_1) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{\boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon}}{2\sigma^2}\right) \quad (41)$$

对数似然函数则可以写成：

$$\ln L(\theta_0, \theta_1) = -\frac{n}{2} \cdot \ln(2\pi\sigma^2) - \frac{\boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon}}{2\sigma^2} \quad (42)$$

对数似然函数进一步整理为：

$$\ln L(\theta_0, \theta_1) = -\frac{n}{2} \cdot \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) \quad (43)$$

对数似然函数对 $\boldsymbol{\theta}$ 求导为 $\mathbf{0}$ 得到等式：

$$\frac{1}{2\sigma^2} (2\mathbf{X}^T \mathbf{X}\boldsymbol{\theta} - 2\mathbf{X}^T \mathbf{y}) = \mathbf{0} \quad (44)$$

整理得到：

$$\mathbf{X}^T \mathbf{X}\boldsymbol{\theta} = \mathbf{X}^T \mathbf{y} \quad (45)$$

如果 $X^T X$ 可逆，则 θ 为：

$$\hat{\theta}_{MLE} = (X^T X)^{-1} X^T y \quad (46)$$

这和本章前文的优化解一致。

此外，线性回归还可以从最大后验概率估计 MAP 角度理解，这是《数据有道》要介绍的话题之一。

在代数、线性代数、优化、投影、QR 分解、SVD 分解几个视角基础上，这一章又提供理解线性回归两个新视角——条件概率、最大似然估计 MLE。

为了保证线性回归模型的有效性和精度，通常需要满足以下假设条件：a) 线性关系：自变量和因变量之间的关系必须是线性的。b) 独立性：观测值之间必须是独立的。c) 方差齐性：每个自变量的方差大小相近。d) 误差服从正态分布。e) 自变量之间不能有高度相关性或共线性，因为这会导致模型出现多重共线性，从而使得参数估计变得不稳定。

如果这些假设条件得到满足，那么线性回归模型将会给出较为准确和可靠的结果，否则模型的效果可能会受到影响。

“鸢尾花书”有关线性回归的内容并没有完全结束。图 14 所示为某个线性回归结果。给大家留个悬念，本系列丛书《数据有道》一册将讲解如何理解图 14 结果。

此外，《数据有道》将铺开介绍更多回归算法，比如多元回归分析、正则化、岭回归、套索回归、弹性网络回归、贝叶斯回归、多项式回归、逻辑回归，以及基于主成分分析的正交回归、主元回归等算法。

```

OLS Regression Results
=====
Dep. Variable:          AAPL      R-squared:                0.687
Model:                  OLS       Adj. R-squared:           0.686
Method:                 Least Squares   F-statistic:             549.7
Date:                  XXXXXXXXXX    Prob (F-statistic):       4.55e-65
Time:                  XXXXXXXXXX    Log-Likelihood:          678.03
No. Observations:      252          AIC:                     -1352.
Df Residuals:          250          BIC:                     -1345.
Df Model:               1
Covariance Type:       nonrobust
=====
               coef      std err          t      P>|t|      [0.025      0.975]
-----
const          0.0018      0.001      1.759      0.080      -0.000      0.004
SP500          1.1225      0.048     23.446      0.000       1.028      1.217
=====
Omnibus:                52.424    Durbin-Watson:           1.864
Prob(Omnibus):           0.000    Jarque-Bera (JB):        210.803
Skew:                    0.777    Prob(JB):                1.68e-46
Kurtosis:                7.203    Cond. No.                 46.1
=====

```

图 14. 线性回归结果