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

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。《矩阵力量》一书提过，线性回归是一种有监督学习 (supervised learning)。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》从优化角度讲过线性回归，《矩阵力量》从投影、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。

简单线性回归

简单线性回归 (Simple Linear Regression) 为一元线性回归模型 (univariate linear repression)，是指模型中只含有一个自变量和一个因变量，表达式如下：

$$y = \underbrace{b_0 + b_1 x}_{\hat{y}} + \varepsilon \tag{1}$$

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。图 1 所示为平面上的一个线性回归关系。

ε 为残差项 (residuals)、误差项 (error term)、干扰项 (disturbance term)或噪音项 (noise term)。

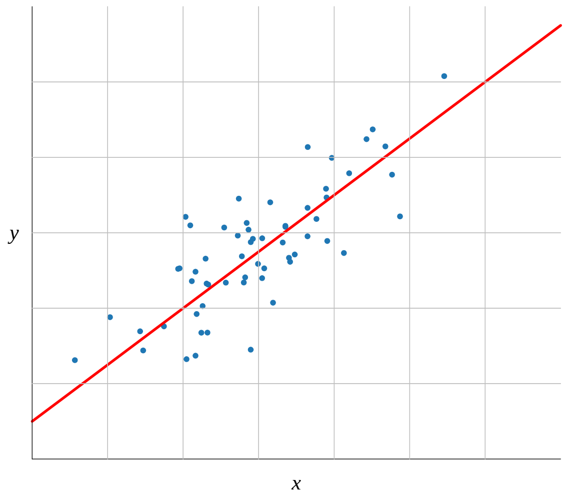

图 1. 平面上，一元线性回归

预测

利用 (1) 做预测，预测值 $\hat{y}$ 为：

$$\hat{y} = b_0 + b_1 x \tag{2}$$

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：

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

残差

(1) 中残差项为：

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

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。

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：

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

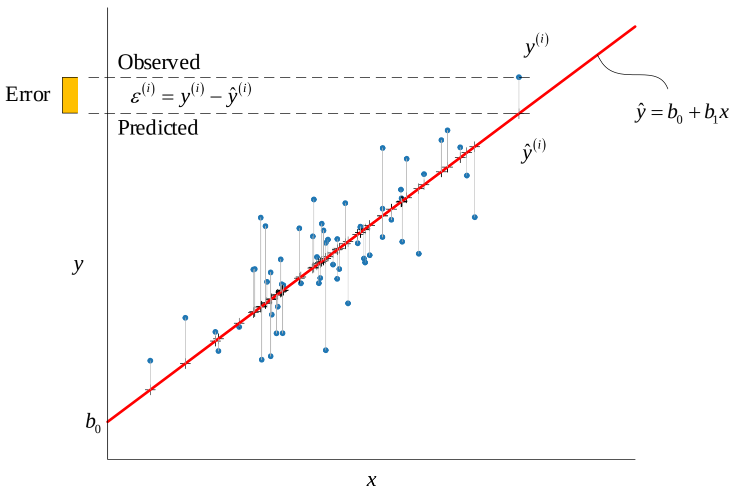

图 2. 简单线性回归中的残差项

矩阵形式

使用阵运算表达一元线性回归：

$$\boldsymbol{y} = b_0 \boldsymbol{1} + b_1 \boldsymbol{x} + \boldsymbol{\varepsilon} \tag{6}$$

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

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

图 3 解释 (6) 给出的矩阵运算。

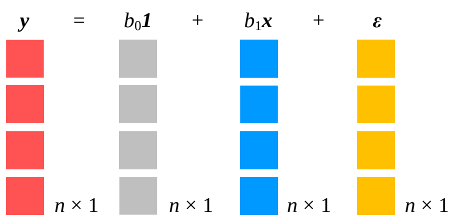

图 3. 用矩阵运算表达一元回归

预测值构成的列向量 $\hat{\boldsymbol{y}}$ 为：

$$\hat{\boldsymbol{y}} = b_0 \boldsymbol{1} + b_1 \boldsymbol{x} \tag{8}$$

$\hat{\boldsymbol{y}}$ 是 $\boldsymbol{1}$ 和 $\boldsymbol{x}$ 的线性组合。

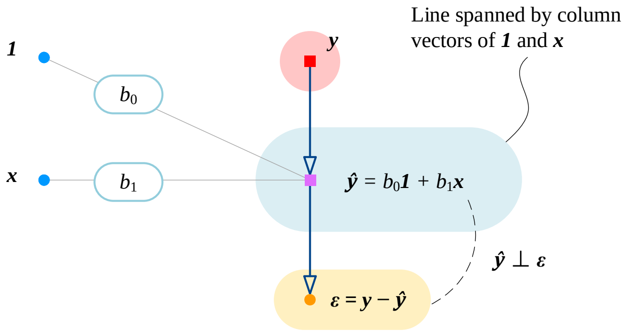

图 4. 一元最小二乘法线性回归数据关系

残差项列向量 $\boldsymbol{\varepsilon}$ 为：

$$\boldsymbol{\varepsilon} = \boldsymbol{y} - \hat{\boldsymbol{y}} \tag{9}$$

图 5 可视化求解残差项列向量 $\boldsymbol{\varepsilon}$ 过程。

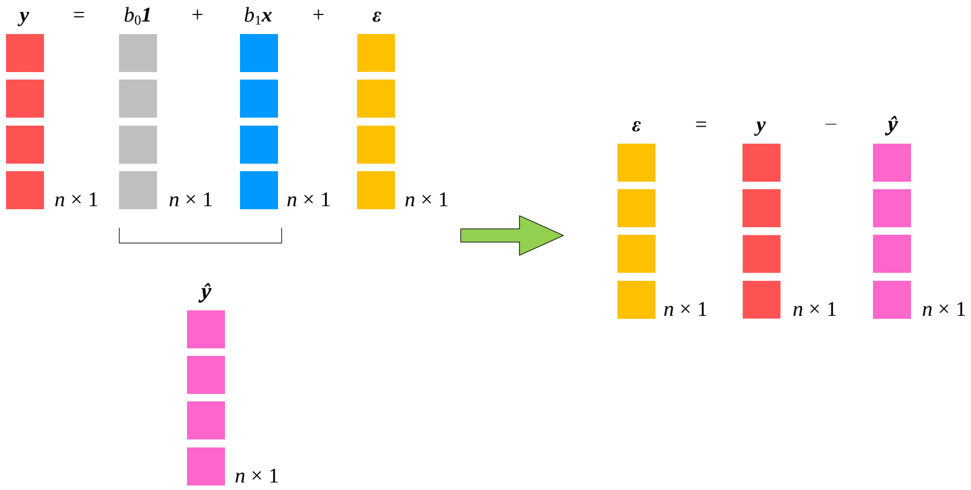

图 5. 求解残差项列向量

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：

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

残差平方和 SSE 为：

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

注意，丛书用 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。

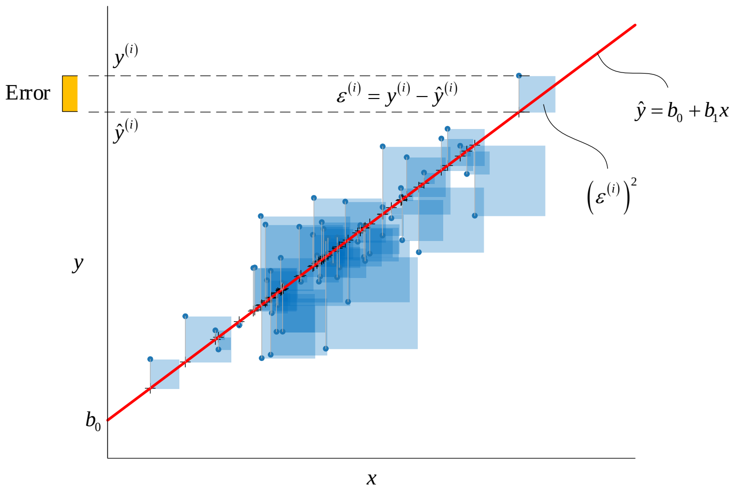

图 6. 残差平方和的几何意义

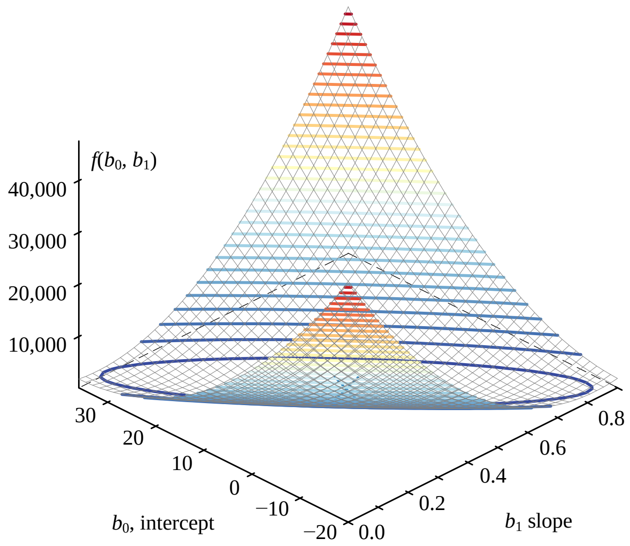

图 7. 误差平方和 SSE 随 b_0、b_1 变化构造的开口向上抛物面，图片来自《数学要素》第 24 章

24.3 优化问题

用线性代数工具构造 OLS 优化问题：

$$\arg\min_{\boldsymbol{b}} \left\| \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b} \right\| \tag{12}$$

也可以写成：

$$\arg\min_{\boldsymbol{b}} \ \left\| \boldsymbol{\varepsilon} \right\|^2 = \boldsymbol{\varepsilon}^{\mathrm{T}} \boldsymbol{\varepsilon} \tag{13}$$

令

$$\boldsymbol{b} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{1} & \boldsymbol{x} \end{bmatrix} \tag{14}$$

其中，$\boldsymbol{X}$ 又叫设计矩阵 (design matrix)。

$\hat{\boldsymbol{y}}$ 可以写成：

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{b} \tag{15}$$

残差向量 $\boldsymbol{\varepsilon}$ 可以写成：

$$\boldsymbol{\varepsilon} = \boldsymbol{y} - b_0\boldsymbol{1} - b_1\boldsymbol{x} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b} \tag{16}$$

定义 $f(\boldsymbol{b})$ 为：

$$f(\boldsymbol{b}) = \boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) \tag{17}$$

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

$$\frac{\partial f(\boldsymbol{b})}{\partial \boldsymbol{b}} = 2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{b} - 2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{0} \tag{18}$$

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

$$\boldsymbol{b} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{19}$$

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 $\boldsymbol{1}$ 和 $\boldsymbol{x}$ 撑起平面 H 上，向量 $\boldsymbol{y}$ 的投影为 $\hat{\boldsymbol{y}}$，而残差 $\boldsymbol{\varepsilon}$ 垂直于这个平面：

$$\begin{aligned} \boldsymbol{\varepsilon} \perp \boldsymbol{1} \quad &\Rightarrow \quad \boldsymbol{1}^{\mathrm{T}}\boldsymbol{\varepsilon} = 0 \quad \Rightarrow \quad \boldsymbol{1}^{\mathrm{T}}(\boldsymbol{y} - \hat{\boldsymbol{y}}) = 0 \\ \boldsymbol{\varepsilon} \perp \boldsymbol{x} \quad &\Rightarrow \quad \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\varepsilon} = 0 \quad \Rightarrow \quad \boldsymbol{x}^{\mathrm{T}}(\boldsymbol{y} - \hat{\boldsymbol{y}}) = 0 \end{aligned} \tag{20}$$

以上两式合并：

$$\underbrace{\begin{bmatrix} \boldsymbol{1} & \boldsymbol{x} \end{bmatrix}}_{X}{}^{\mathrm{T}}(\boldsymbol{y} - \hat{\boldsymbol{y}}) = \boldsymbol{0} \tag{21}$$

整理得到：

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{b} \tag{22}$$

这和 (18) 一致。

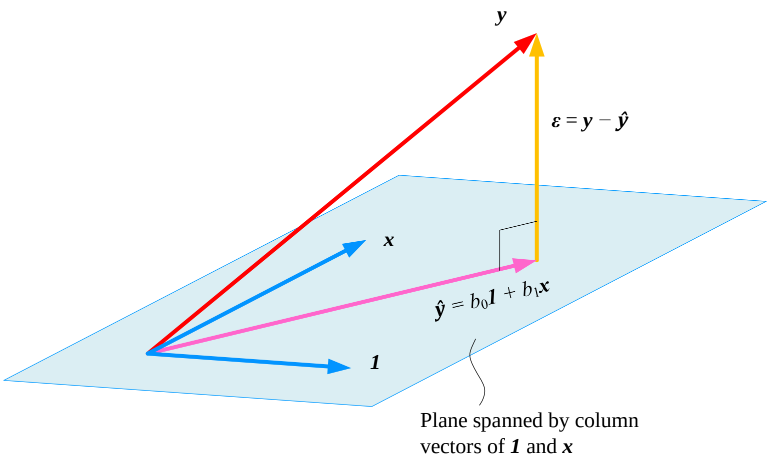

图 8. 几何角度解释一元最小二乘结果，二维平面

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)：

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} \tag{23}$$

QR 分解

对 $\boldsymbol{X}$ 进行 QR 分解得到：

$$\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R} \tag{24}$$

这样求得 $\boldsymbol{b}$ 为：

$$\boldsymbol{b} = \boldsymbol{R}^{-1}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{y} \tag{25}$$

奇异值分解

对 $\boldsymbol{X}$ 进行完全型 SVD 分解得到：

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}} \tag{26}$$

这样求得 $\boldsymbol{b}$ 为：

$$\boldsymbol{b} = \boldsymbol{V}\boldsymbol{S}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{y} \tag{27}$$

《矩阵力量》一册介绍过 $\boldsymbol{V}\boldsymbol{S}^{-1}\boldsymbol{U}^{\mathrm{T}}$ 是 $\boldsymbol{X}$ 的摩尔-彭若斯广义逆 (Moore–Penrose inverse)。$\boldsymbol{S}^{-1}$ 的主对角线非零元素为 $\boldsymbol{S}$ 的非零奇异值倒数，$\boldsymbol{S}^{-1}$ 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：

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

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)。

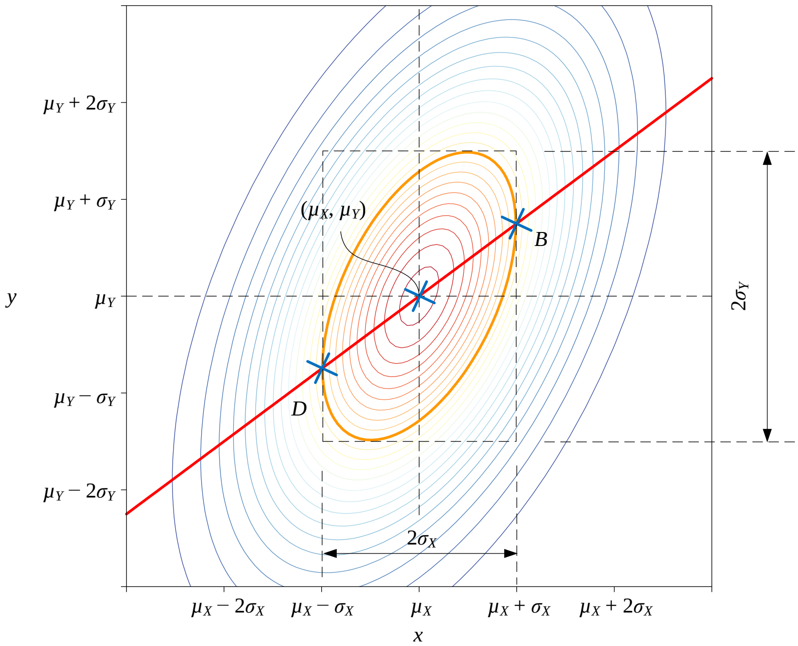

图 9. 给定 $X = x$ 的条件期望

图 10 所示为不同相关性系数条件下，回归直线和椭圆关系。

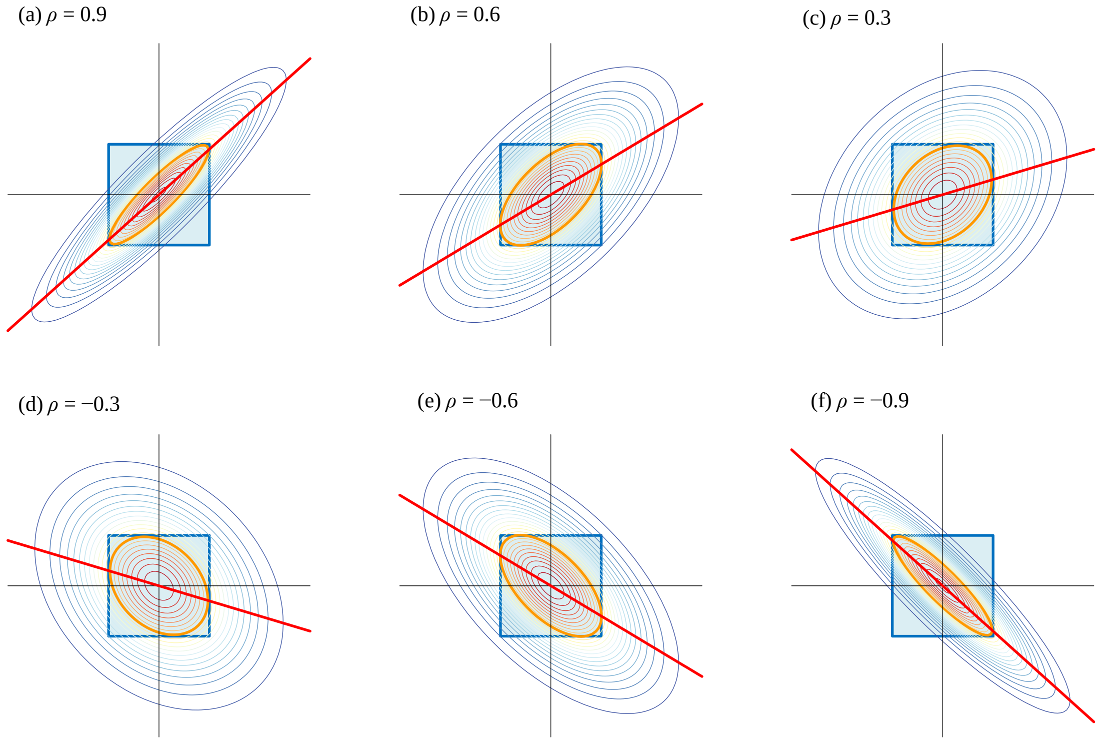

图 10. 条件期望直线位置和相关性系数关系，$\sigma_X = \sigma_Y$

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：

$$y = \underbrace{3.758}_{\mu_X} + \underbrace{1.858}_{\rho_{X,Y}\frac{\sigma_Y}{\sigma_X}} \left(x - \underbrace{5.843}_{\mu_Y} \right) \tag{29}$$

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”。

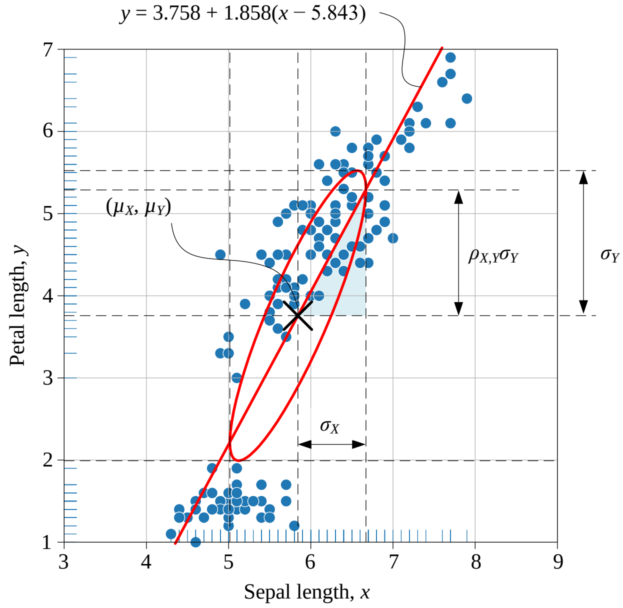

图 11. 花瓣长度、花萼长度之间回归关系

Bk5_Ch24_01.py 绘制图 11。

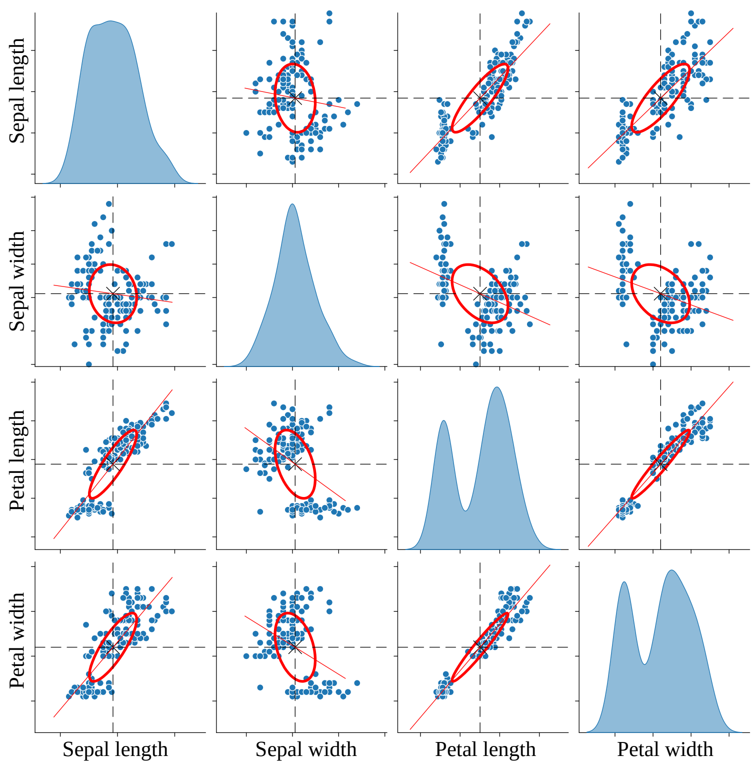

图 12. 成对特征图和回归关系

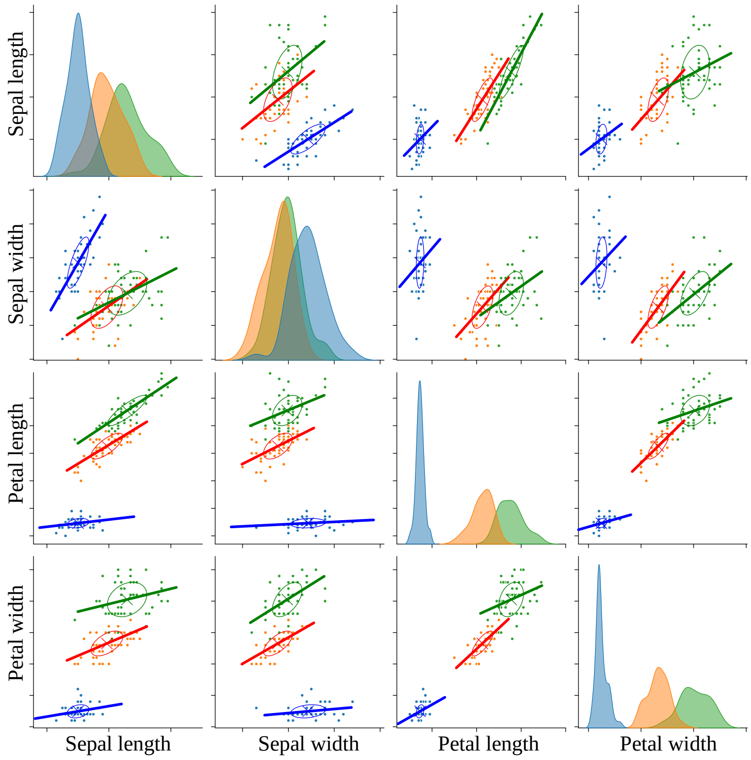

图 13. 成对特征图和回归关系，考虑分类标签

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比阅读，本节和下一节中，线性回归解析式改写成：

$$y = \underbrace{\theta_0 + \theta_1 x}_{\hat{y}} + \varepsilon \tag{30}$$

对应的超定方程组写成：

$$\boldsymbol{y} = \boldsymbol{X\theta} \tag{31}$$

残差向量 $\boldsymbol{\varepsilon}$ 为：

$$\boldsymbol{\varepsilon} = \boldsymbol{y} - \boldsymbol{X\theta} \tag{32}$$

假设残差项服从正态分布：

$$\varepsilon \sim N\left(0, \sigma^2\right) \tag{33}$$

根据 (4)，也就是说 Y_i 服从：

$$Y_i \sim N\left(\theta_1 X_i + \theta_0, \sigma^2\right) \tag{34}$$

Y_i 的概率密度函数为：

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

似然函数可以写成：

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

对数似然函数为：

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

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

$$\begin{aligned} \hat{\theta}_1 &= \frac{\sum_{i=1}^{n}\left(x^{(i)} - \mu_X\right)\left(y^{(i)} - \mu_Y\right)}{\sum_{i=1}^{n}\left(x^{(i)} - \mu_X\right)^2} \\ \hat{\theta}_0 &= \mu_Y - \hat{\theta}_1 \mu_X \end{aligned} \tag{38}$$

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：

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

似然函数则可以写成：

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

用矩阵运算表达上式得到：

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

对数似然函数则可以写成：

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

对数似然函数进一步整理为：

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

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

$$\frac{1}{2\sigma^2}\left(2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X\theta}-2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}\right)=0 \tag{44}$$

整理得到：

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X\theta}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{45}$$

如果 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 可逆，则 $\boldsymbol{\theta}$ 为：

$$\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{46}$$

这和本章前文的优化解一致。

$\ln L(\theta_0,\theta_1)$ 对 σ 求偏导为 0 得到：

$$-\frac{n}{\sigma}+\frac{1}{\sigma^3}(\boldsymbol{y}-\boldsymbol{X\theta})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{X\theta})=0 \tag{47}$$

进一步整理得到等式：

$$\sigma^2=\frac{1}{n}(\boldsymbol{y}-\boldsymbol{X\theta})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{X\theta})=\frac{\mathrm{SSE}}{n} \tag{48}$$

丛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，贝叶斯回归也是数据科学、机器学习重要话题之一，丛书后续将展开介绍这一话题。

```
                            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:                     XXXXXXXXXXX   Prob (F-statistic):           4.55e-65
Time:                     XXXXXXXXXXX   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