

3

Multivariate Linear Regression

多元线性回归

用多个解释变量来预测响应变量结果



科学不知道它对想象力的依赖。

Science does not know its debt to imagination.

—— 拉尔夫·沃尔多·爱默生 (Ralph Waldo Emerson) | 美国思想家、文学家 | 1942 ~ 2018



```

< matplotlib.pyplot.quiver() 绘制箭头图
< numpy.arccos() 反余弦函数
< numpy.cov() 计算协方差矩阵
< numpy.identity() 构造单位矩阵
< numpy.linalg.det() 计算矩阵的行列式值
< numpy.linalg.inv() 求矩阵逆
< numpy.linalg.matrix_rank() 计算矩阵的秩
< numpy.matrix() 构造矩阵
< numpy.ones() 构造全 1 矩阵或向量
< numpy.ones_like() 按照给定矩阵或向量形状构造全 1 矩阵或向量
< plot_wireframe() 绘制线框图
< scipy.stats.f.cdf() F 分布累积分布函数
< 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
< statsmodels.stats.outliers_influence.variance_inflation_factor() 计算方差膨胀因子

```

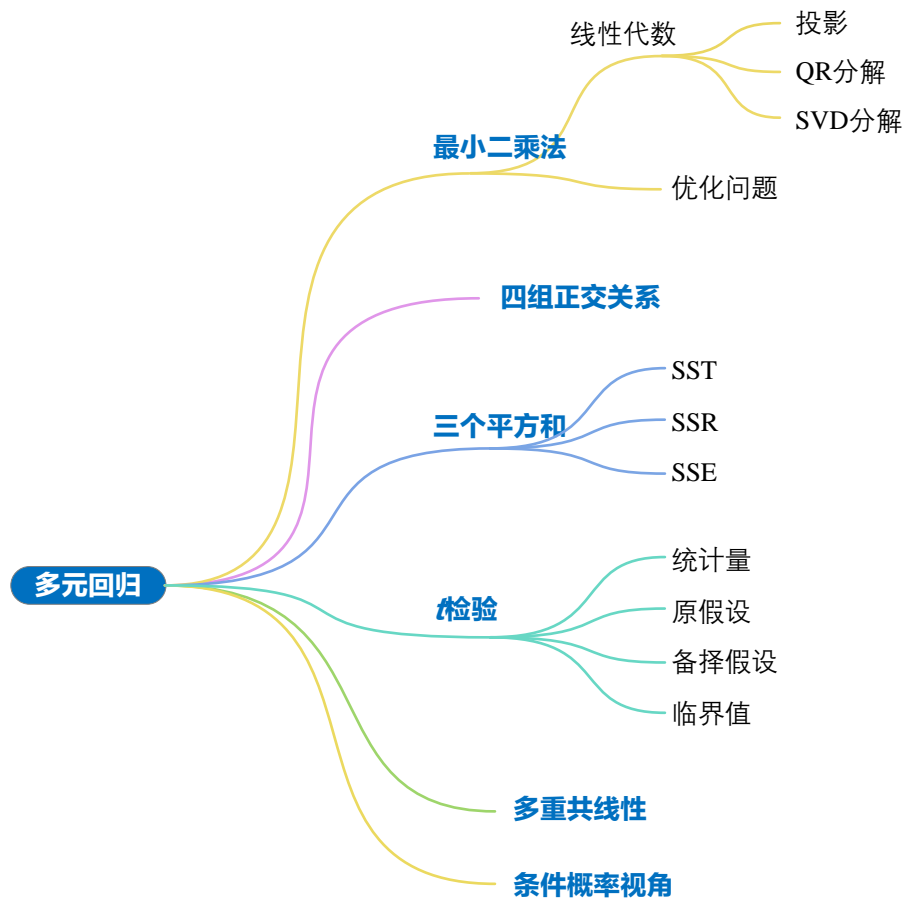
本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

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



3.1 多元线性回归

这一章将探讨多元线性回归。多元线性回归是一种统计分析方法，用于研究两个或多个自变量与一个因变量之间的关系。它通过拟合一个包含多个自变量的线性模型来预测因变量的值。

多元线性回归的表达式如下：

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_Dx_D + \varepsilon \quad (1)$$

其中， b_0 为截距项， b_1, b_2, \dots, b_D 代表自变量系数， ε 为残差项， D 为自变量个数。几何角度来看，多元线性回归得到一个**超平面** (hyperplane)。

用矩阵运算表达 (1)：

$$y = \underbrace{b_0\mathbf{1} + b_1x_1 + b_2x_2 + \dots + b_Dx_D}_{\hat{y}} + \varepsilon \quad (2)$$

其中， $\mathbf{1}$ 为全 1 列向量。

换一种方式来写 (2)：

$$y = \underbrace{X\mathbf{b}}_{\hat{y}} + \varepsilon \quad (3)$$

其中，

$$X_{n \times (D+1)} = [\mathbf{1} \quad x_1 \quad x_2 \quad \dots \quad x_D] = \begin{bmatrix} 1 & x_{1,1} & \dots & x_{1,D} \\ 1 & x_{2,1} & \dots & x_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,D} \end{bmatrix}_{n \times (D+1)}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_D \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon^{(1)} \\ \varepsilon^{(2)} \\ \vdots \\ \varepsilon^{(n)} \end{bmatrix} \quad (4)$$

矩阵 X 常被称作**设计矩阵** (design matrix)。图 1 所示矩阵运算对应 (3)。

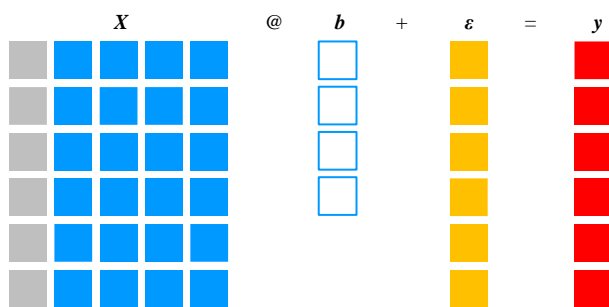


图 1. 多元线性回归模型矩阵运算

预测值构成的列向量 \hat{y} ，通过下式计算得到：

$$\hat{y} = Xb \quad (5)$$

残差向量的算式为：

$$\varepsilon = y - \hat{y} = y - Xb \quad (6)$$

如图 2 所示，第 i 个观测点的残差项，可以通过下式计算得到：

$$\varepsilon^{(i)} = y^{(i)} - \hat{y}^{(i)} = y^{(i)} - \mathbf{x}^{(i)}\mathbf{b}$$

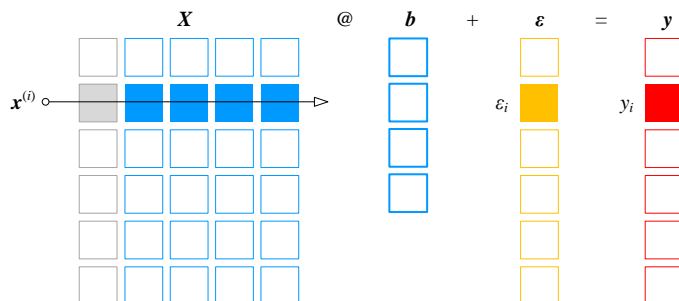


图 2. 计算第 i 个观测点的残差项

图 3 所示为多元 OLS 线性回归数据关系。也就是说， \hat{y} 可以看成设计矩阵 X 的列向量线性组合。

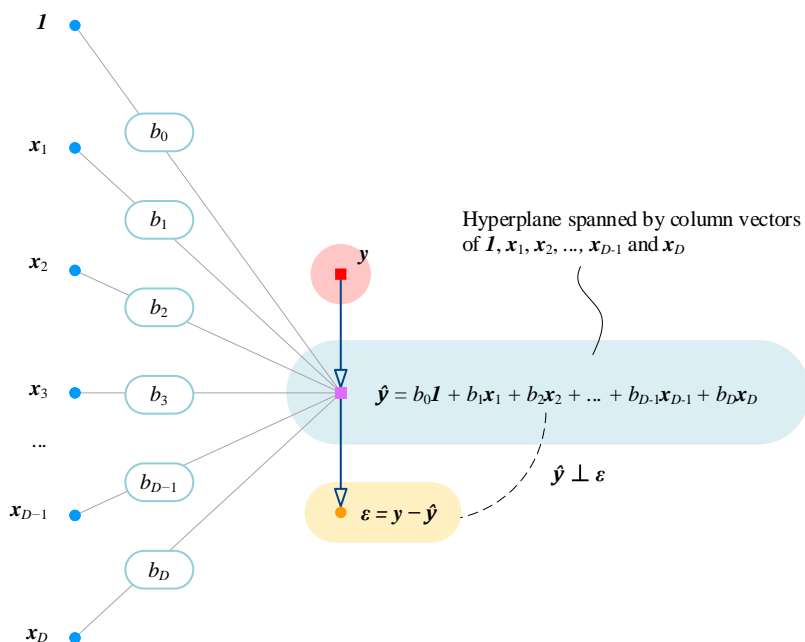


图 3. 多元 OLS 线性回归数据关系

⚠ 注意，矩阵 X 为 n 行， $D + 1$ 列，第一列为全 1 列向量；增加一列全 1 列向量目的是为了引入常数项。

如图 4 所示，如果数据都已经中心化 (去均值)，则可以不考虑常数项。

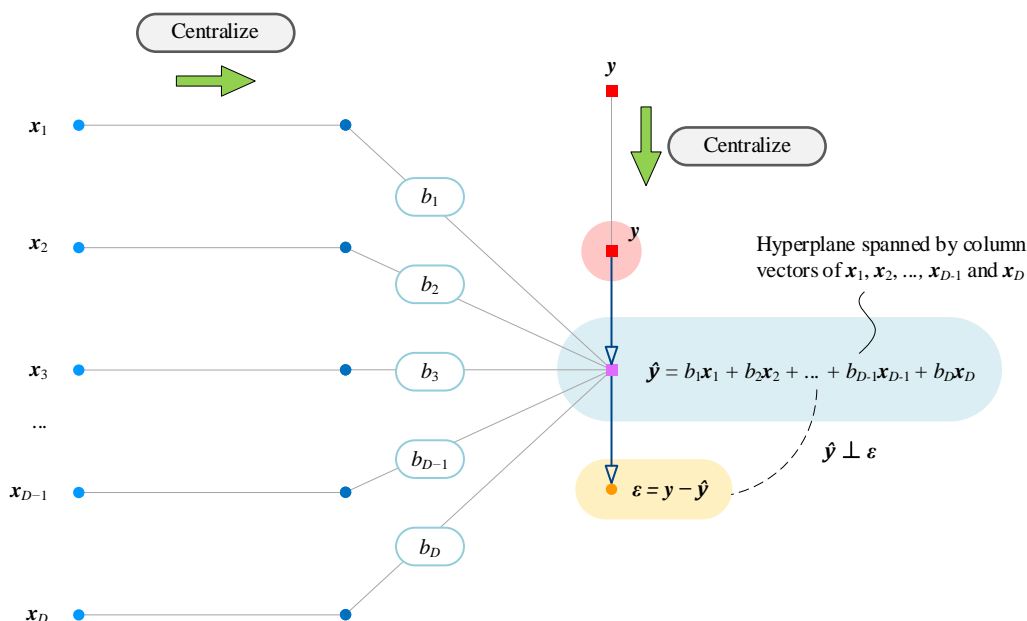


图 4. 多元 OLS 线性回归数据关系，中心化数据

3.2 优化问题：OLS

一般通过如下两种方式求得线性回归参数：

- ◀ **最小二乘法** (Ordinary Least Square, OLS)，因变量和拟合值之间的欧氏距离最小化；
- ◀ **最大似然概率估计** (Maximum Likelihood Estimation, MLE)，用样本数据反推最可能的模型参数值。

OLS 线性最小二乘法通过最小化残差值平方和 SSE 来计算得到最佳的拟合回归线参数：

$$\arg \min_b \text{SSE} \quad (7)$$

对于多元线性回归，残差平方和 SSE 为：

$$\text{SSE} = \sum_{i=1}^n (\varepsilon^{(i)})^2 = \varepsilon \cdot \varepsilon = \|\varepsilon\|_2^2 = \varepsilon^T \varepsilon = (y - Xb)^T (y - Xb) = \|y - Xb\|_2^2 \quad (8)$$

OLS 多元线性优化问题的目标函数可以写成：

$$f(b) = (y - Xb)^T (y - Xb) \quad (9)$$

$f(b)$ 可以整理为：

$$\begin{aligned} f(b) &= (y - Xb)^T (y - Xb) \\ &= (y^T - b^T X^T)(y - Xb) \\ &= y^T y - y^T Xb - b^T X^T y + b^T X^T Xb \\ &= \underbrace{b^T X^T Xb}_{\text{Quadratic term}} - \underbrace{2b^T X^T y}_{\text{Linear term}} + \underbrace{y^T y}_{\text{Constant}} \end{aligned} \quad (10)$$

观察上式，发现 $f(b)$ 可以看成是一个多元二次函数，含有二次项、一次项和常数项。

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

因此，对于二元回归，不考虑常数项系数 b_0 的话， b_1 和 b_2 构成的曲面 $f(b_1, b_2)$ 为椭圆抛物面，如图 5 所示。

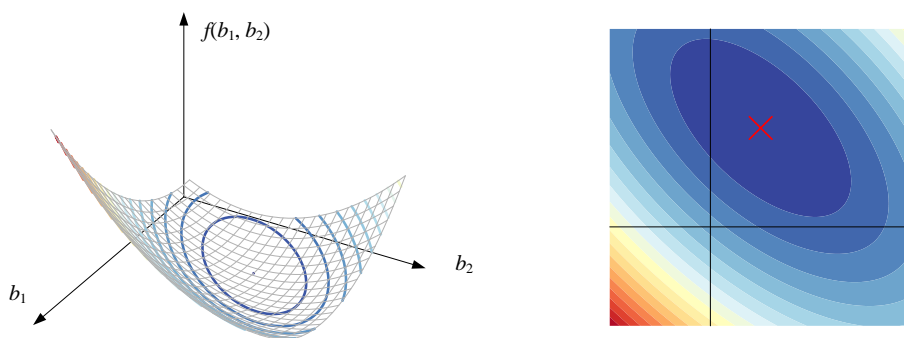


图 5. $f(b_1, b_2)$ 函数曲面

$f(\mathbf{b})$ 梯度向量如下：

$$\nabla f(\mathbf{b}) = \frac{\partial f(\mathbf{b})}{\partial \mathbf{b}} \quad (11)$$

$f(\mathbf{b})$ 为连续函数，取得极值时，梯度向量为零向量：

$$\nabla f(\mathbf{b}) = \mathbf{0} \Rightarrow \mathbf{X}^T \mathbf{X} \mathbf{b} - \mathbf{X}^T \mathbf{y} = \mathbf{0} \quad (12)$$

如果 $\mathbf{X}^T \mathbf{X}$ 可逆， \mathbf{b} 的解为：

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (13)$$



《矩阵力量》介绍过，如果 $\mathbf{X}^T \mathbf{X}$ 不可逆，可以用奇异值分解求伪逆。

$f(\mathbf{b})$ 的黑塞矩阵为：

$$\nabla^2 f(\mathbf{b}) = \frac{\partial^2 f(\mathbf{b})}{\partial \mathbf{b} \partial \mathbf{b}^T} = 2 \mathbf{X}^T \mathbf{X} \quad (14)$$

下面，判断 $f(\mathbf{b})$ 黑塞矩阵为正定矩阵，从而判定极值点为最小值点。

对于任意非零向量 \mathbf{a} ，下式恒大于等于 0：

$$\mathbf{a}^T (\mathbf{X}^T \mathbf{X}) \mathbf{a} = (\mathbf{X} \mathbf{a})^T (\mathbf{X} \mathbf{a}) = \|\mathbf{X} \mathbf{a}\|^2 \geq 0 \quad (15)$$

等号成立时，即 $\mathbf{X} \mathbf{a} = \mathbf{0}$ ，即当 \mathbf{X} 列向量线性相关，我们暂时不考虑这种情况。因此，对于 \mathbf{X} 为列满秩， $f(\mathbf{b})$ 黑塞矩阵为正定矩阵， $f(\mathbf{b})$ 在极值点处取得最小值。

模型拟合值向量 $\hat{\mathbf{y}}$ 为：

$$\hat{\mathbf{y}} = \mathbf{X} \mathbf{b} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (16)$$

残差向量 $\boldsymbol{\varepsilon}$ 为：

$$\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (17)$$

$\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ 为《矩阵力量》第 9 章介绍的**帽子矩阵** (hat matrix) \mathbf{H} ，它常出现在矩阵投影运算中。

令，

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \quad (18)$$

帽子矩阵 \mathbf{H} 为**幂等矩阵** (idempotent matrix)，幂等矩阵是指一个矩阵与自身相乘后仍等于它本身的矩阵，即满足 $\mathbf{H}^2 = \mathbf{H}$ 。幂等矩阵在线性代数中有广泛的应用，特别是在投影、几何变换等领域。在投影中，幂等矩阵可以用来描述一个向量在一个子空间上的投影；在几何变换中，幂等矩阵可以用来描述一个对象在进行相应变换后仍等于它本身。最简单的幂等矩阵就是单位矩阵 \mathbf{I} ，满足 $\mathbf{I}^2 = \mathbf{I}$ 。

利用帽子矩阵 \mathbf{H} ，

$$\begin{cases} \hat{\mathbf{y}} = \mathbf{H}\mathbf{y} \\ \boldsymbol{\varepsilon} = (\mathbf{I} - \mathbf{H})\mathbf{y} \end{cases} \quad (19)$$

3.3 几何解释：投影

图 6 所示为多维空间视角下的数据矩阵；矩阵 $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_D]$ 每一列代表一个特征，每一列可以看做一个向量。



鸢尾花书《矩阵力量》一书中，我们反复探讨过这一点。

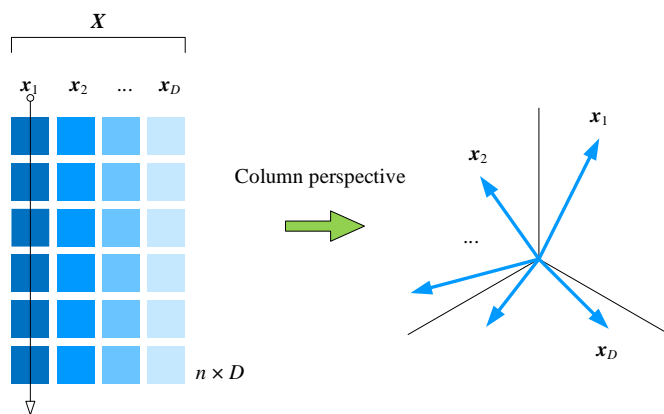


图 6. 多维空间视角下的矩阵 \mathbf{X}

不考虑常数项，预测值向量 $\hat{\mathbf{y}}$ 可以通过下式计算得到：

$$\hat{\mathbf{y}} = b_1 \mathbf{x}_1 + b_2 \mathbf{x}_2 + \dots + b_D \mathbf{x}_D \quad (20)$$

(20) 说明，预测值向量 $\hat{\mathbf{y}}$ 是自变量向量 $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_D$ 的线性组合。如果 $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_D$ 构成一个超平面 H ， $\hat{\mathbf{y}}$ 在 H 这个平面内。

有了这一思想，构造因变量向量 \mathbf{y} 和自变量向量 $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_D$ 的线性回归模型，相当于 \mathbf{y} 向 $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_D$ 构成的超平面 H 投影。如图 7 所示，预测值向量 $\hat{\mathbf{y}}$ 是因变量向量 \mathbf{y} 在 H 的投影结果：

$$\mathbf{y} = \hat{\mathbf{y}} + \boldsymbol{\varepsilon} \quad (21)$$

简单来说，从向量投影的角度来理解多元线性回归，可以将回归问题看作是将因变量向量在自变量向量所张成的子空间上的投影。

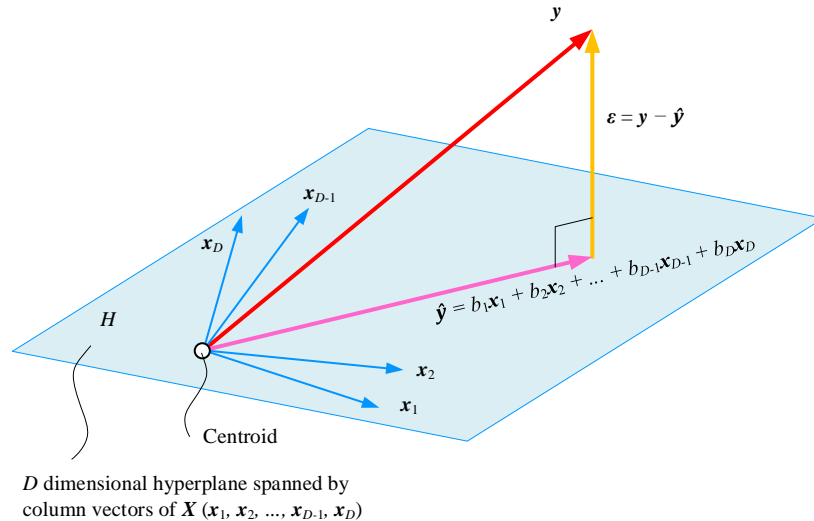


图 7. 几何角度解释多元最小二乘法线性回归

而残差项向量 $\boldsymbol{\varepsilon}$ 是预测值向量 $\hat{\mathbf{y}}$ 是因变量向量 \mathbf{y} 两者之差：

$$\boldsymbol{\varepsilon} = \mathbf{y} - \hat{\mathbf{y}} \quad (22)$$

残差项向量 $\boldsymbol{\varepsilon}$ 垂直于 $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_D$ 构成的超平面 H 。

由上所述，残差 $\boldsymbol{\varepsilon}$ ($\boldsymbol{\varepsilon} = \mathbf{y} - \hat{\mathbf{y}}$) 是无法通过 $(\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{D-1}, \mathbf{x}_D)$ 解释部分向量，垂直于超平面：

$$\boldsymbol{\varepsilon} \perp \mathbf{X} \Rightarrow \mathbf{X}^T \boldsymbol{\varepsilon} = 0 \quad (23)$$

得到

$$\mathbf{X}^T (\mathbf{y} - \mathbf{X}\mathbf{b}) = 0 \Rightarrow \mathbf{X}^T \mathbf{X}\mathbf{b} = \mathbf{X}^T \mathbf{y} \quad (24)$$

这和上一节得到的结果完全一致，但是从几何视角看 OLS，让求解过程变得非常简洁。

请大家再次注意，只有 \mathbf{X} 为列满秩时， $\mathbf{X}^T \mathbf{X}$ 才存在逆。

此外，我们可以很容易在 \mathbf{X} 最左侧加入一列全 1 向量 $\mathbf{1}$ ，残差项向量 $\boldsymbol{\varepsilon}$ 则垂直于 $\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_D$ 构成的超平面 H 。

《统计至简》介绍过 OLS 线性回归假设条件。OLS 线性回归的假设条件是用来保证模型的有效性和可靠性。简单来说，这些假设条件主要包括线性关系、正态分布、同方差性、独立性和残差之和为零。

首先，线性关系假设要求因变量和自变量之间的关系是线性的，即在自变量变化时，因变量的变化量是按照线性关系变化的。这个假设是 OLS 回归分析的前提条件，否则回归结果将会失真。

其次，正态分布假设要求模型的残差应该满足正态分布。正态分布是概率论和统计学中最为重要的分布之一，如果残差不满足正态分布，可能会导致回归结果失真。

同方差性假设要求残差的方差在各个自变量取值下都相等。如果残差的方差不相等，会导致回归结果的可靠性下降。

独立性假设要求各个观测值之间是独立的，即一个观测值的取值不受其他观测值的影响。如果存在相关性，回归结果可能会失真。

最后，残差之和为零要求模型的残差的总和为零，这是保证回归分析的正确性的必要条件。

总之，这些假设条件对于 OLS 线性回归的结果具有重要影响，需要在回归分析中进行检验和确认。

表 1 所示为用矩阵方式表达 OLS 线性回归假设。

表 1. 用矩阵运算表达 OLS 线性回归假设

假设	矩阵表达
线性模型	$\mathbf{y} = \mathbf{X}\mathbf{b} + \boldsymbol{\varepsilon}$
残差服从正态分布	$\boldsymbol{\varepsilon} \mathbf{X} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$
残差期望值为 0	$E(\boldsymbol{\varepsilon} \mathbf{X}) = \mathbf{0}$
残差同方差性	$\text{var}(\boldsymbol{\varepsilon} \mathbf{X}) = \begin{bmatrix} \text{var}(\varepsilon^{(1)}) & \text{cov}(\varepsilon^{(1)}, \varepsilon^{(2)}) & \cdots & \text{cov}(\varepsilon^{(1)}, \varepsilon^{(n)}) \\ \text{cov}(\varepsilon^{(2)}, \varepsilon^{(1)}) & \text{var}(\varepsilon^{(2)}) & \cdots & \text{cov}(\varepsilon^{(2)}, \varepsilon^{(n)}) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(\varepsilon^{(n)}, \varepsilon^{(1)}) & \text{cov}(\varepsilon^{(n)}, \varepsilon^{(2)}) & \cdots & \text{var}(\varepsilon^{(n)}) \end{bmatrix} = \sigma^2 \mathbf{I}$
矩阵 \mathbf{X} 不存在多重共线性	$\text{rank}(\mathbf{X}) = D + 1$ $\det(\mathbf{X}^T \mathbf{X}) \neq 0$

3.4 二元线性回归

为了方便大家理解，本节用实例讲解二元线性回归。

二元线性回归解析式为：

$$\hat{\mathbf{y}} = b_0 \mathbf{I} + b_1 \mathbf{x}_1 + b_2 \mathbf{x}_2 \quad (25)$$

图 8 所示为二元 OLS 线性回归数据关系。

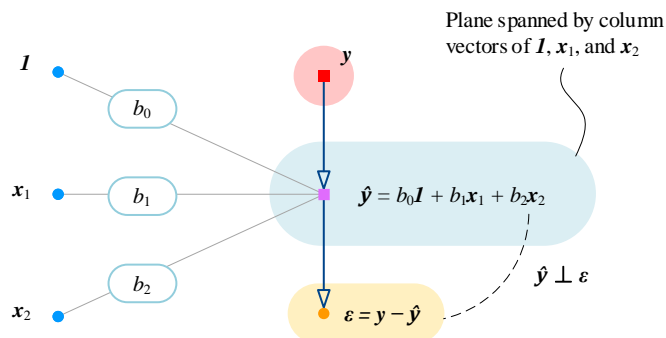


图 8. 二元 OLS 线性回归数据关系

本节介绍利用两个股票日收益率解释 S&P 500 日收益率。图 9 所示为参与回归数据 $[y, x_1, x_2]$ 的散点图。

图 10 所示为 $[y, x_1, x_2]$ 数据的成对特征分析图。

图 11 所示为 $[y, x_1, x_2]$ 数据的协方差矩阵、相关性和夹角热图。

图 12 所示为二元 OLS 线性回归结果。图 13 所示为三维数据散点图和回归平面。

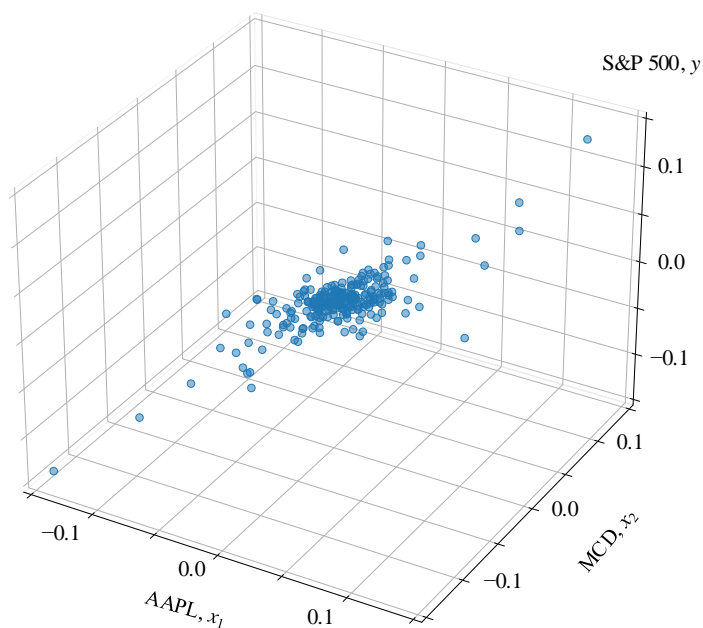
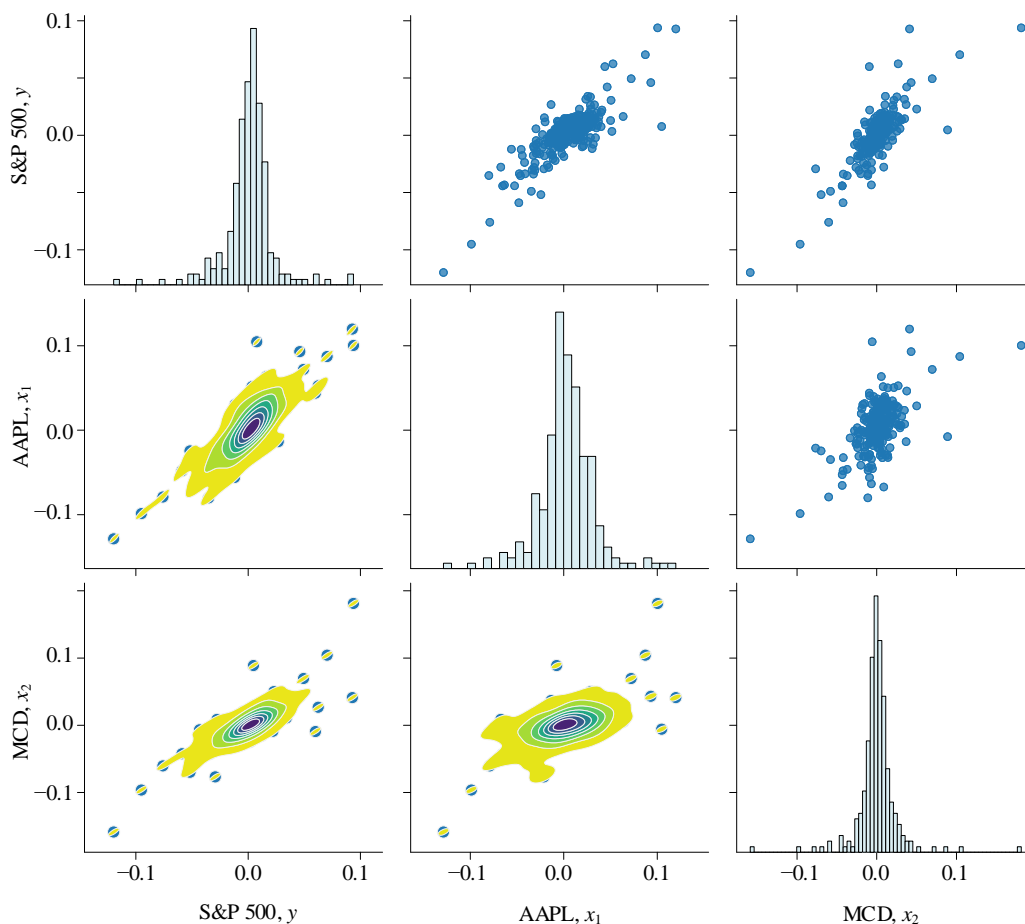
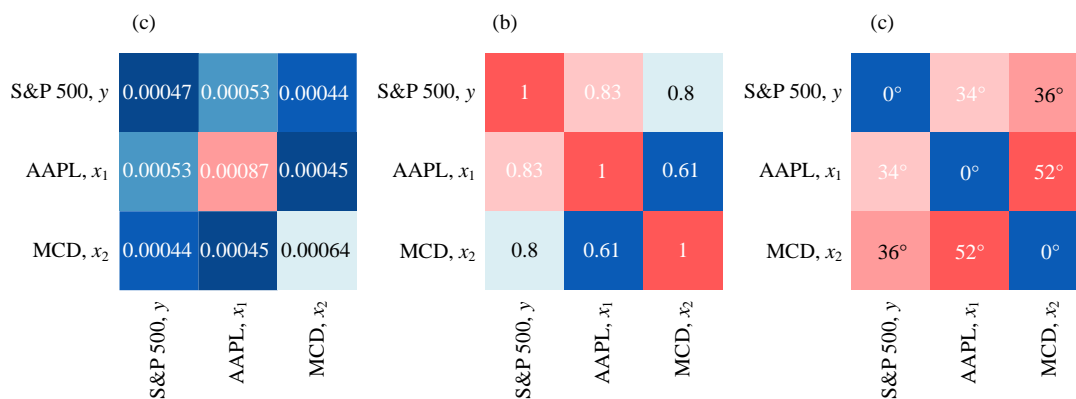


图 9. 二元线性回归数据

图 10. 二元线性回归数据 $[y, x_1, x_2]$ 成对特征分析图图 11. $[y, x_1, x_2]$ 数据的协方差矩阵、相关性和夹角热图

OLS Regression Results						
=====						
Dep. Variable:	SP500	R-squared:	0.830			
Model:	OLS	Adj. R-squared:	0.829			
Method:	Least Squares	F-statistic:	607.4			
Date:	XXXXXXXXXXXXXXXXXX	Prob (F-statistic):	1.69e-96			
Time:	XXXXXXXXXXXXXXXXXX	Log-Likelihood:	831.06			
No. Observations:	252	AIC:	-1656.			
Df Residuals:	249	BIC:	-1646.			
Df Model:	2					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

const	-0.0006	0.001	-0.984	0.326	-0.002	0.001
AAPL	0.3977	0.024	16.326	0.000	0.350	0.446
MCD	0.4096	0.028	14.442	0.000	0.354	0.465
=====						
Omnibus:	37.744	Durbin-Watson:	1.991			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	157.711			
Skew:	0.492	Prob(JB):	5.67e-35			
Kurtosis:	6.749	Cond. No.	59.4			

图 12. 二元 OLS 线性回归分析结果

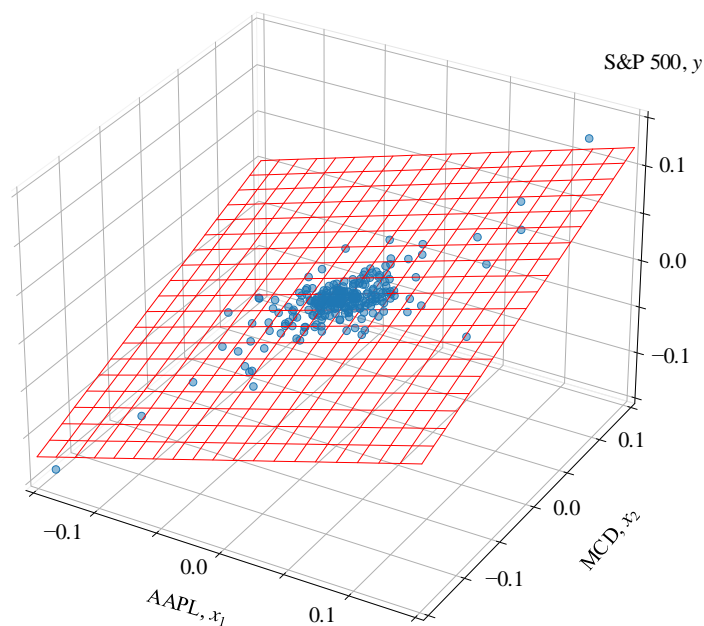


图 13. 三维空间，回归平面



Bk7_Ch03_01.ipynb 完成本节二元线性回归。

3.5 多元回归

本节介绍一个多元回归问题，构造多元 OLS 线性回归模型用 12 只股票日收益率预测 S&P 500 日收益率。图 14 所示股价数据。

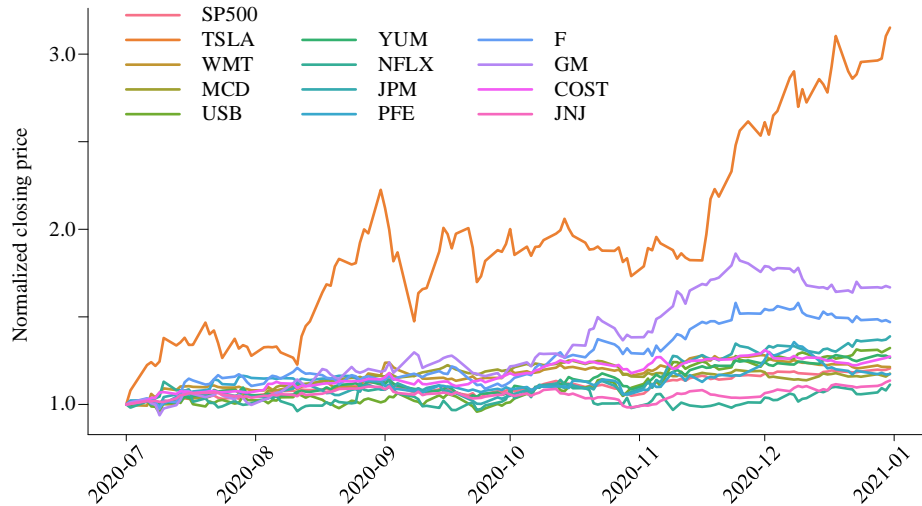
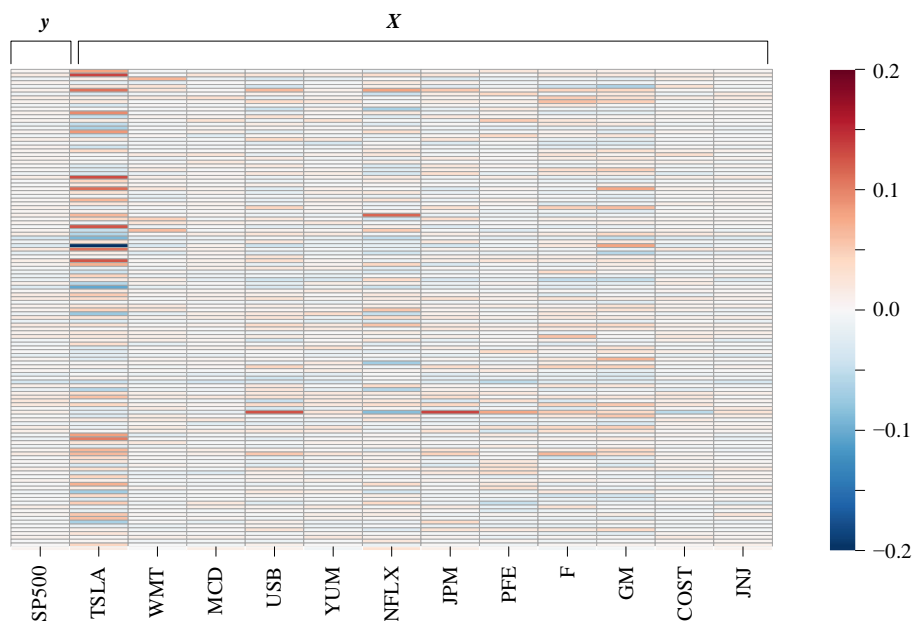


图 14. 股价数据，起始值归一化

根据股价水平计算得到的日收益率。图 15 所示为日收益率热图。图 16 所示为 $[y, X]$ 数据协方差矩阵。图 17 所示为均方差 (即波动率) 直方图。

图 18 所示为 $[y, X]$ 数据相关性系数矩阵热图。图 19 所示为几只不同股票股价收益率和 S&P 500 收益率相关性系数柱状图。利用余弦相似性，根据相关性系数矩阵，可以计算得到 $[y, X]$ 标准差向量夹角，矩阵热图如图 20 所示。图 21 所示为多元 OLS 线性回归解。



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

图 15. $[y, X]$ 日收益率热图

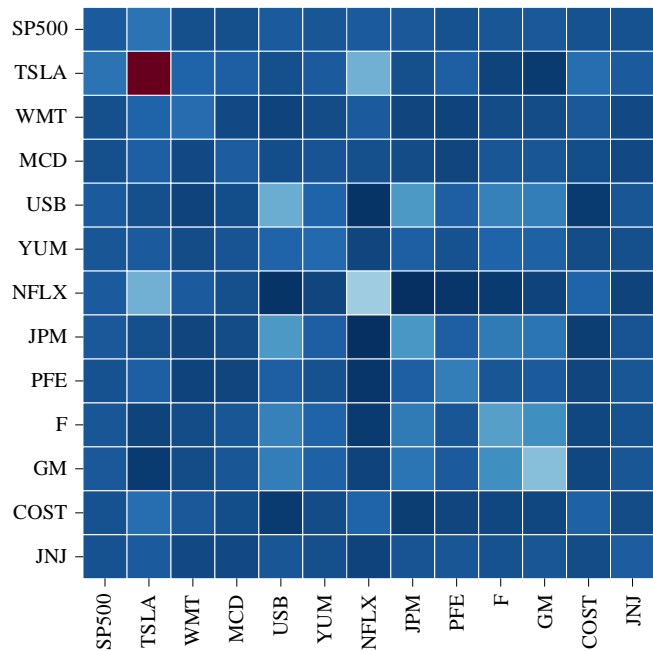


图 16. $[y, X]$ 数据协方差矩阵

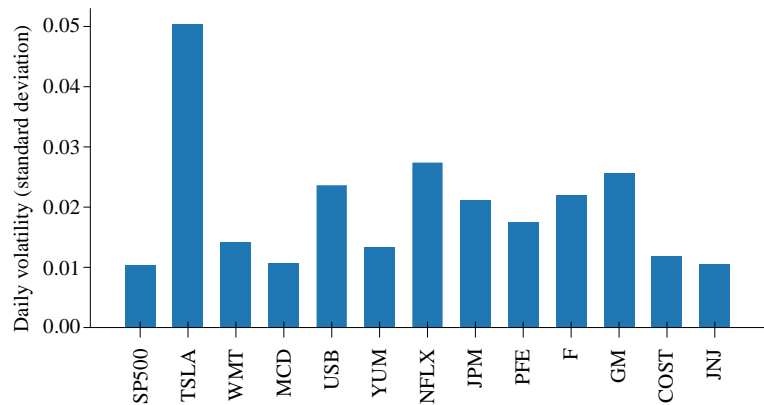


图 17. 日波动率柱状图

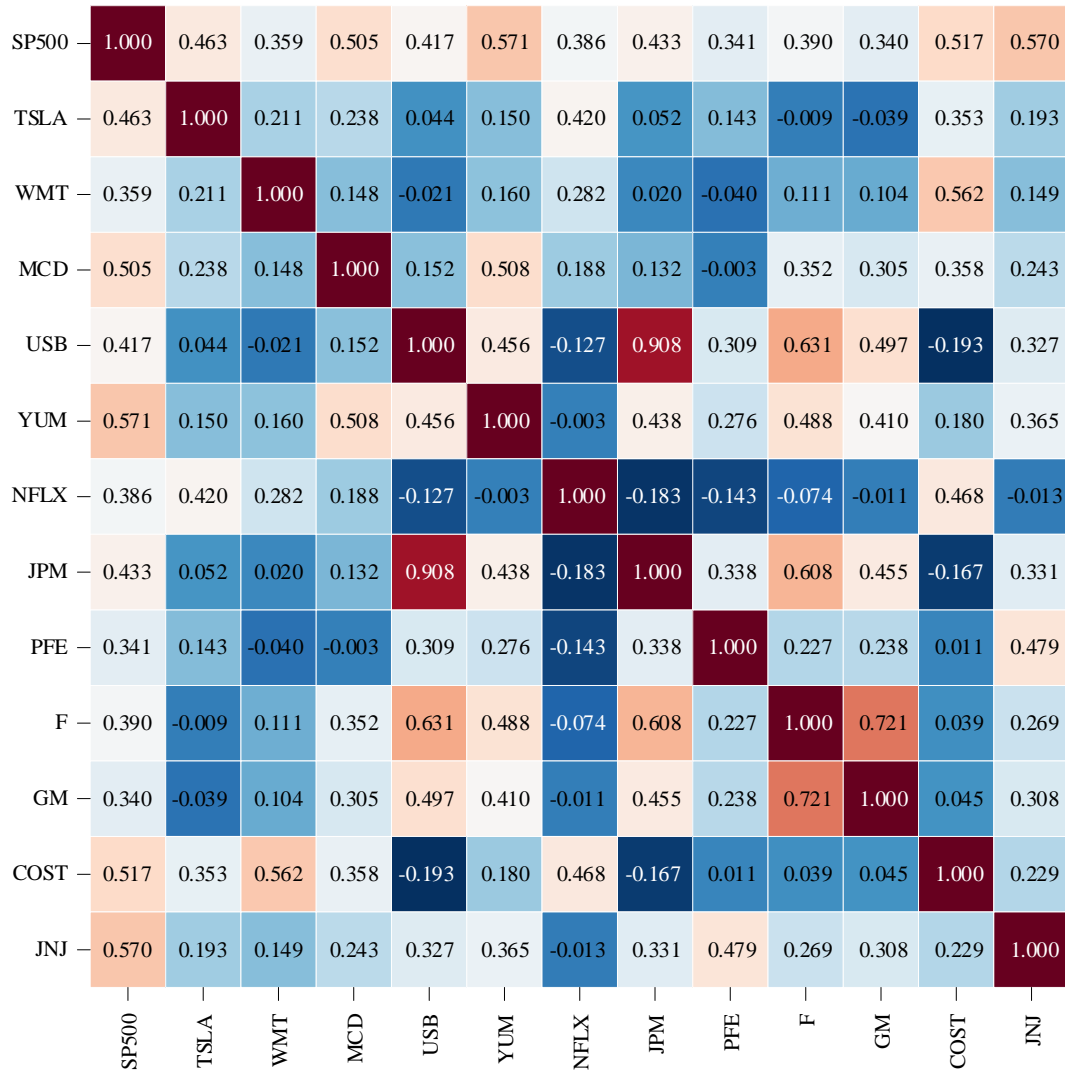
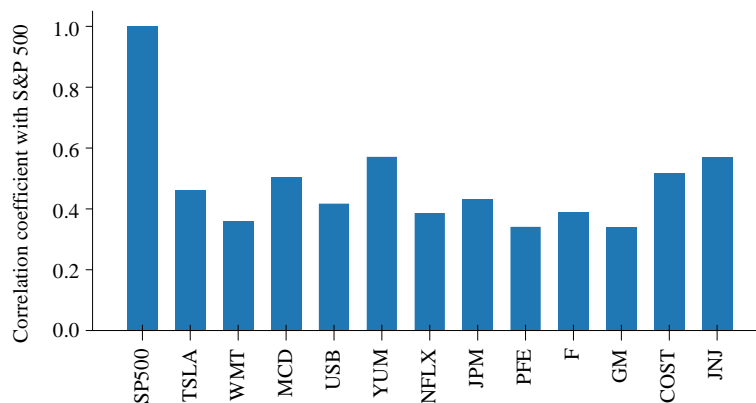
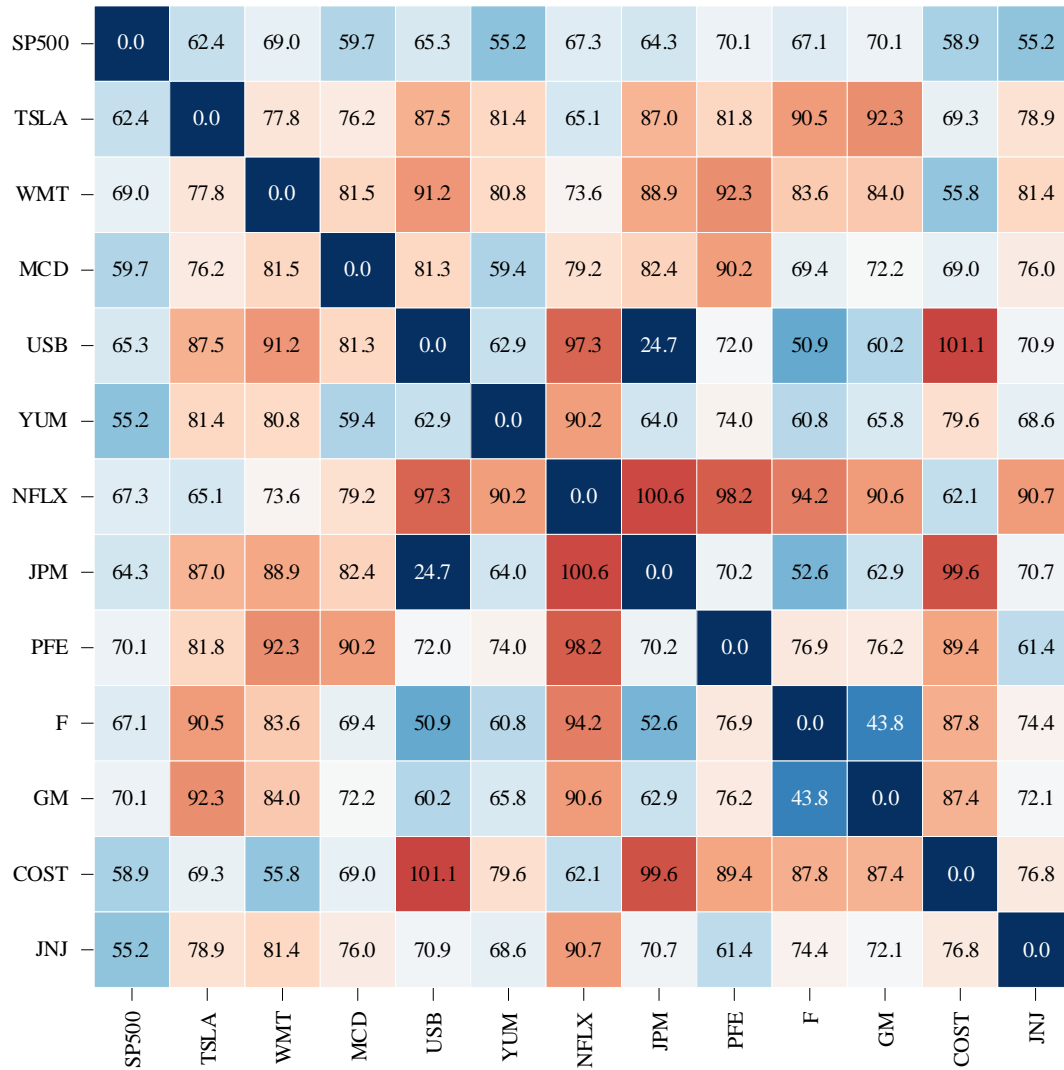
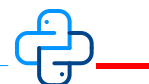
图 18. $[y, X]$ 数据相关性系数矩阵热图

图 19. 股价收益率和 S&P 500 收益率相关性系数柱状图

图 20. $[y, X]$ 标准差向量夹角矩阵热图，余弦相似性

OLS Regression Results						
Dep. Variable:	SP500	R-squared:	0.774			
Model:	OLS	Adj. R-squared:	0.750			
Method:	Least Squares	F-statistic:	32.48			
Date:	XXXXXXXXXXXXXXXXXX	Prob (F-statistic):	3.03e-31			
Time:	XXXXXXXXXXXXXXXXXX	Log-Likelihood:	493.88			
No. Observations:	127	AIC:	-961.8			
Df Residuals:	114	BIC:	-924.8			
Df Model:	12					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	-0.0005	0.000	-1.038	0.302	-0.001	0.000
TSLA	0.0248	0.011	2.248	0.027	0.003	0.047
WMT	0.0272	0.041	0.667	0.506	-0.054	0.108
MCD	0.1435	0.057	2.536	0.013	0.031	0.256
USB	0.0164	0.051	0.322	0.748	-0.084	0.117
YUM	0.1469	0.047	3.114	0.002	0.053	0.240
NFLX	0.0972	0.021	4.539	0.000	0.055	0.140
JPM	0.1415	0.055	2.583	0.011	0.033	0.250
PFE	0.0546	0.033	1.662	0.099	-0.010	0.120
F	-0.0068	0.036	-0.187	0.852	-0.078	0.065
GM	-0.0105	0.027	-0.388	0.699	-0.064	0.043
COST	0.2176	0.059	3.713	0.000	0.101	0.334
JNJ	0.2414	0.056	4.350	0.000	0.131	0.351
Omnibus:	7.561	Durbin-Watson:	1.862			
Prob(Omnibus):	0.023	Jarque-Bera (JB):	8.445			
Skew:	0.400	Prob(JB):	0.0147			
Kurtosis:	3.978	Cond. No.	156.			

图 21. 多元 OLS 线性回归分析结果



Bk7_Ch03_02.ipynb 完成本节多元线性回归。

3.6 正交关系

第一个直角三角形

通过上一章学习，大家都很清楚第一个勾股关系：

$$\underbrace{\|\mathbf{y} - \bar{\mathbf{y}}\mathbf{I}\|_2^2}_{\text{SST}} = \underbrace{\|\hat{\mathbf{y}} - \bar{\mathbf{y}}\mathbf{I}\|_2^2}_{\text{SSR}} + \underbrace{\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2}_{\text{SSE}} \quad (26)$$

具体如图 22 所示。上一章提到这一个直角三角形可以帮助我们解释 R^2 。

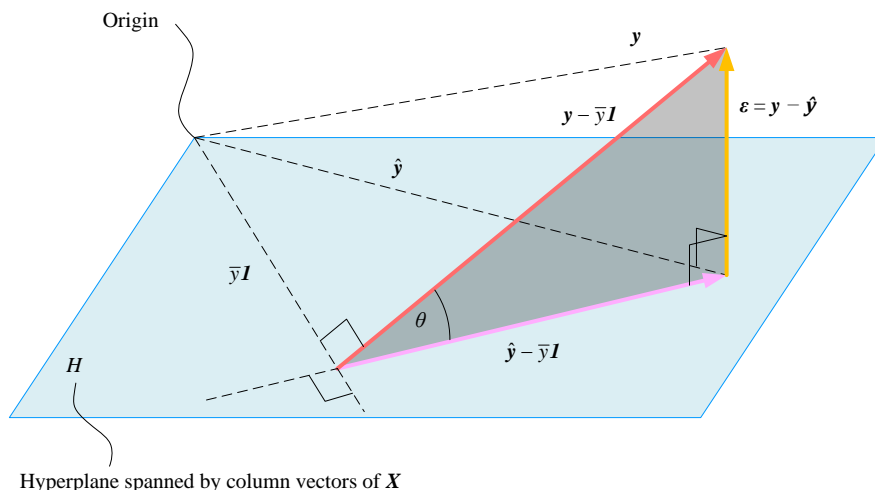


图 22. 第一个直角三角形

第二个直角三角形

除了 (26) 这个重要的直角三角形的勾股定理之外，还有另外一个重要的直角三角形勾股定理关系。

$$\|y\|_2^2 = \|\hat{y}\|_2^2 + \|y - \hat{y}\|_2^2 = \|\hat{y}\|_2^2 + \|\epsilon\|_2^2 \quad (27)$$

具体如图 23 所示。图 23 这个直角很容易理解。残差向量 ϵ 垂直于超平面 H 内的一切向量，显然 ϵ 垂直 \hat{y} 。

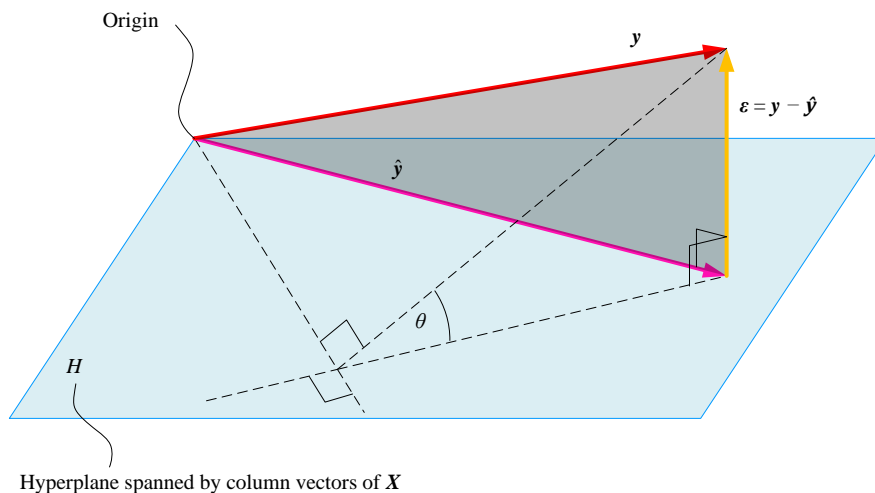


图 23. 第二个直角三角形

第三个直角三角形

此外，《矩阵力量》第 22 章介绍过，向量 $y - \bar{y}I$ 垂直于向量 $\bar{y}I$ ：

$$(\bar{y}I)^T (y - \bar{y}I) = 0 \quad (28)$$

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

具体如图 24 所示。上式体现的核心思想就是 y 中可以被均值解释的部分为 $\bar{y}I$ 。

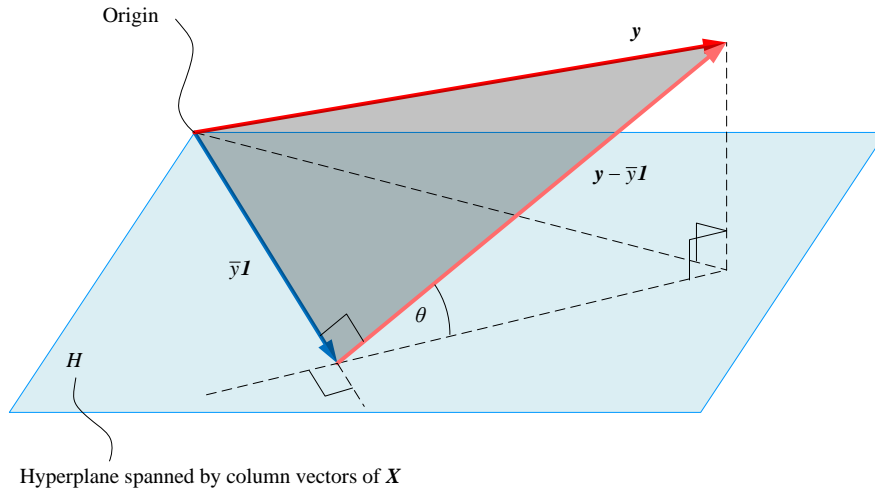


图 24. 第三个直角三角形

第四个直角三角形

OLS 假设残差之和为 0:

$$\sum_{i=1}^n \varepsilon^{(i)} = 0 \quad (29)$$

注意，如果总残差不为 0，就说明预测值的总和与实际观测值的总和不相等，这意味着模型存在偏差，不能很好地解释数据。

对应向量运算：

$$I^T \varepsilon = \varepsilon^T I = 0 \quad (30)$$

残差向量可以写成：

$$\varepsilon = y - \hat{y} = y - \bar{y}I - (\hat{y} - \bar{y}I) \quad (31)$$

上式左乘 I^T ，得到：

$$I^T \varepsilon = I^T (y - \bar{y}I) - \underbrace{I^T (\hat{y} - \bar{y}I)}_0 \quad (32)$$

即

$$I^T (\hat{y} - \bar{y}I) = 0 \quad (33)$$

也就是说，如图 25 所示， $\hat{y} - \bar{y}I$ 垂直于向量 $\bar{y}I$ ：

$$\bar{y}I^T (\hat{y} - \bar{y}I) = 0 \quad (34)$$

上式体现的核心思想就是 \hat{y} 的均值也是 \bar{y} 。

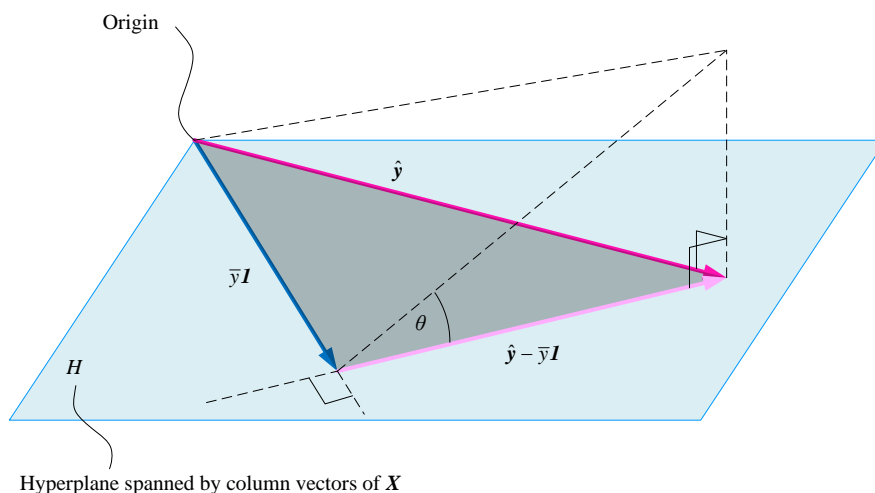


图 25. 第四个直角三角形

3.7 三个平方和

这一节介绍对于多元 OLS 线性回归，如何求解 SST、SSR 和 SSE 这三个平方和。

对于多元 OLS 线性回归模型，SST 可以通过矩阵运算求得：

$$SST = y^T \left(I - \frac{J}{n} \right) y \quad (35)$$

其中矩阵 J 为全 1 方阵，形状为 $n \times n$ ：

$$J_{n \times n} = \mathbf{1}\mathbf{1}^T = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} \quad (36)$$

SSR 可以通过矩阵运算求得：

$$SSR = y^T \left(H - \frac{J}{n} \right) y \quad (37)$$

其中矩阵 H 为本书前文所讲的帽子矩阵，形状为 $n \times n$ ：

$$H = X(X^T X)^{-1} X^T \quad (38)$$

同样，对于多元 OLS 线性回归模型，SSE 可以通过矩阵运算求得：

$$SSE = y^T (I - H) y \quad (39)$$

对于多元 OLS 线性回归模型，MSE 的矩阵运算为：

$$\begin{aligned}
 \text{MSE} &= \frac{\|(I-H)y\|_2^2}{n-k} \\
 &= \frac{y^T y - 2y^T H y + y^T H^2 y}{n-k} \\
 &= \frac{y^T y - y^T H y}{n-k} \\
 &= \frac{y^T (I-H)y}{n-k}
 \end{aligned} \tag{40}$$

上式推导过程采用帽子矩阵重要的性质。

3.8 t 检验

对于多元 OLS 线性回归模型，模型系数 $b_0, b_1, b_2 \dots b_D$ 的协方差矩阵 C 可以通过下式计算得到：

$$C = \hat{\sigma}^2 (X^T X)^{-1} \tag{41}$$

其中，

$$\hat{\sigma}^2 = \text{MSE} = \frac{\varepsilon^T \varepsilon}{n-k} \tag{42}$$

矩阵 C 的对角线元素 $C_{j+1, j+1}$ 为 \hat{b}_j 的方差，非对角线元素为 \hat{b}_j 和 \hat{b}_k 的协方差。

\hat{b}_j 的标准误 $\text{SE}(\hat{b}_j)$ 为：

$$\text{SE}(\hat{b}_j) = \sqrt{C_{j+1, j+1}} \tag{43}$$

对于多元线性回归，假设检验原假设和备择假设分别为：

$$\begin{cases} H_0 : b_j = b_{j,0} \\ H_1 : b_j \neq b_{j,0} \end{cases} \tag{44}$$

b_j 的 t 检验统计值：

$$T_j = \frac{\hat{b}_j - b_{j,0}}{\text{SE}(\hat{b}_j)} \tag{45}$$

类似地，如果下式成立，接受零假设 H_0 ：

$$-t_{1-\alpha/2, n-k} < T_j < t_{1-\alpha/2, n-k} \tag{46}$$

否则，则拒绝零假设 H_0 。

系数 b_j 的 $1-\alpha$ 置信区间为：

$$\hat{b}_j \pm t_{1-\alpha/2, n-k} \cdot \text{SE}(\hat{b}_j) \tag{47}$$

对于多元 OLS 线性模型，预测值 $\hat{y}^{(i)}$ ，的 $1 - \alpha$ 置信区间：

$$\hat{y}^{(i)} \pm t_{1-\alpha/2, n-2} \cdot \sqrt{\text{MSE}} \cdot \sqrt{\mathbf{x}^{(i)} (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{x}^{(i)})^T} \quad (48)$$

$\mathbf{x}^{(i)}$ 为矩阵 \mathbf{X} 的第 i 行：

$$\mathbf{x}^{(i)} = [1 \quad x_{i,1} \quad x_{i,2} \quad \cdots \quad x_{i,D}] \quad (49)$$

类似地，对于多元 OLS 线性回归模型， y_p 的预测区间估计为：

$$\hat{y}^{(i)} \pm t_{1-\alpha/2, n-2} \cdot \sqrt{\text{MSE}} \cdot \sqrt{1 + \mathbf{x}^{(i)} (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{x}^{(i)})^T} \quad (50)$$

3.9 多重共线性

线性回归模型的解释变量不满足相互独立的基本假设前提下，如果模型的解释变量存在多重共线性，将导致最小二乘法得到的模型参数估计量非有效且方差变大，参数估计量经济含义不合理等。

上一章介绍过采用**条件数** (Condition number) 来判定多重共线性。对 $\mathbf{X}^T \mathbf{X}$ 进行特征值分解，得到最大特征值 λ_{\max} 和最小特征值 λ_{\min} 。条件数的定义为两者的比值的平方根。条件数小于 30，可以不必担心多重共线性。

如果 $\mathbf{X}^T \mathbf{X}$ 可逆， $\mathbf{X}^T \mathbf{X}$ 的行列式值不为 0：

$$\det(\mathbf{X}^T \mathbf{X}) \neq 0 \quad (51)$$

这里再介绍一个评价共线性的度量指标，**方差膨胀因子** (variance inflation factor, VIF)，也称为**方差扩大因子**。

一个还有 n 个解释变量的矩阵 $\hat{\mathbf{X}}_i$ ，对于其中的任意解释变量 $\{X_{i,t}\}$ ，其对应的方差膨胀因子 VIF_i 可由下式计算：

$$\text{VIF}_i = \frac{1}{1 - R_i^2} \quad (52)$$

其中 R_i^2 是解释变量 $\{X_{i,t}\}$ 与其解释变量 $\{X_{j,t}, j \neq i\}$ 回归模型的决定系数：

$$X_{i,t} = \alpha_0 + \sum_{j=1, j \neq i}^n \alpha_j X_{j,t} + \varepsilon_t \quad (53)$$

当某个变量 $\{X_{i,t}\}$ 能被其他变量完全线性解释时， R_i^2 的值趋近于 1， VIF_i 的值将趋近于无穷大；所以，各个变量的 VIF 值越小，说明共线性越弱。最常用的 VIF 阈值是 10，即解释变量的 VIF 值都不大于 10 时，认为共线性在可接受范围内；此外， $\text{VIF} \leq 5$ 也是比较常见的、但相对而言更为严格的判断标准。

3.10 条件概率视角看多元线性回归



《统计至简》第 12 章介绍过，多元线性回归本质上就是条件概率中的条件期望值。

如果随机变量向量 χ 和 γ 服从多维高斯分布：

$$\begin{bmatrix} \chi \\ \gamma \end{bmatrix} \sim N \left(\begin{bmatrix} \mu_\chi \\ \mu_\gamma \end{bmatrix}, \begin{bmatrix} \Sigma_{\chi\chi} & \Sigma_{\chi\gamma} \\ \Sigma_{\gamma\chi} & \Sigma_{\gamma\gamma} \end{bmatrix} \right) \quad (54)$$

其中， χ 为随机变量 X_i 构成的列向量， γ 为随机变量 Y_j 构成的列向量：

$$\chi = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_D \end{bmatrix}, \quad \gamma = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_M \end{bmatrix} \quad (55)$$

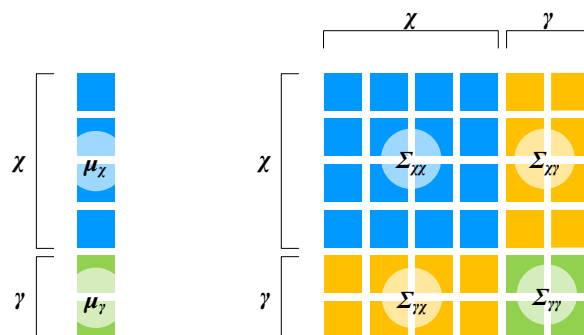


图 26. 均值向量、协方差矩阵形状，图片来自鸢尾花书《统计至简》第 12 章

如图 27 所示，给定 $\chi = x$ 的条件下 γ 的条件期望为：

$$E(\gamma | \chi = x) = \mu_{\gamma|\chi=x} = \Sigma_{\gamma\chi} \Sigma_{\chi\chi}^{-1} (x - \mu_\chi) + \mu_\gamma \quad (56)$$

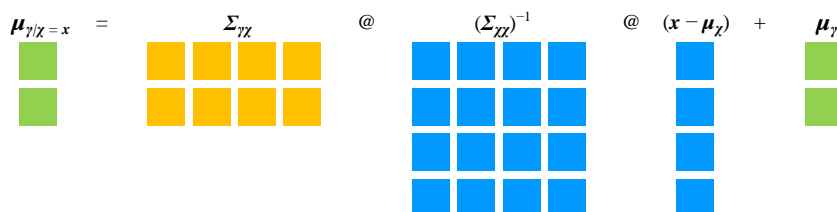


图 27. 给定 $\chi = x$ 的条件下 γ 的期望值的矩阵运算，图片来自鸢尾花书《统计至简》第 12 章

对于本例，我们对 (56) 进行转置得到：

$$\mu_{\gamma|x} = E(\gamma) + \underbrace{(x - E(X))(\Sigma_{XX})^{-1} \Sigma_{Xy}}_b \quad (57)$$

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

$[y, X]$ 对应的协方差矩阵如图 28 所示。图 29 为对 Σ_{XX} 求逆。

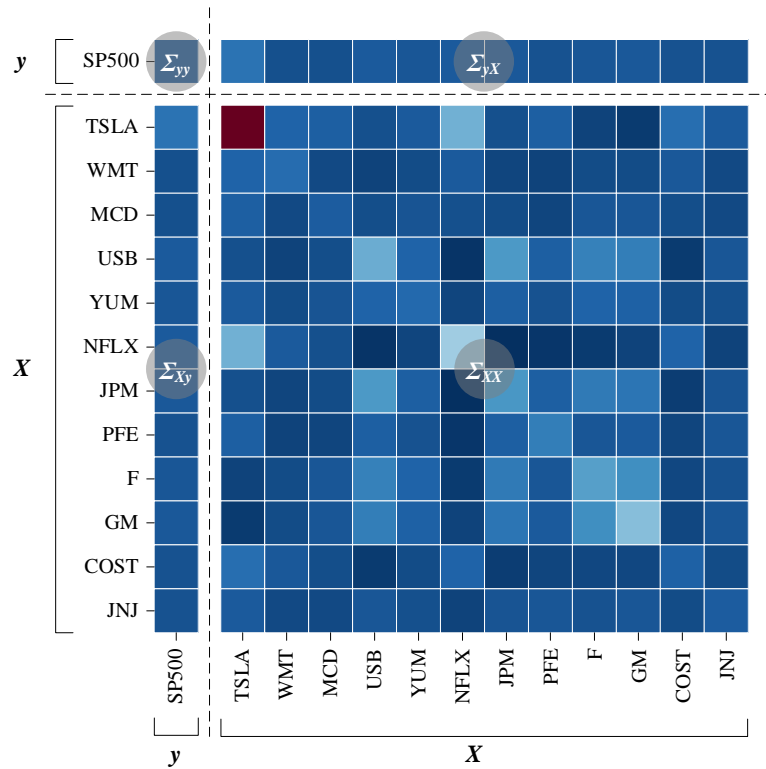


图 28. $[y, X]$ 协方差矩阵

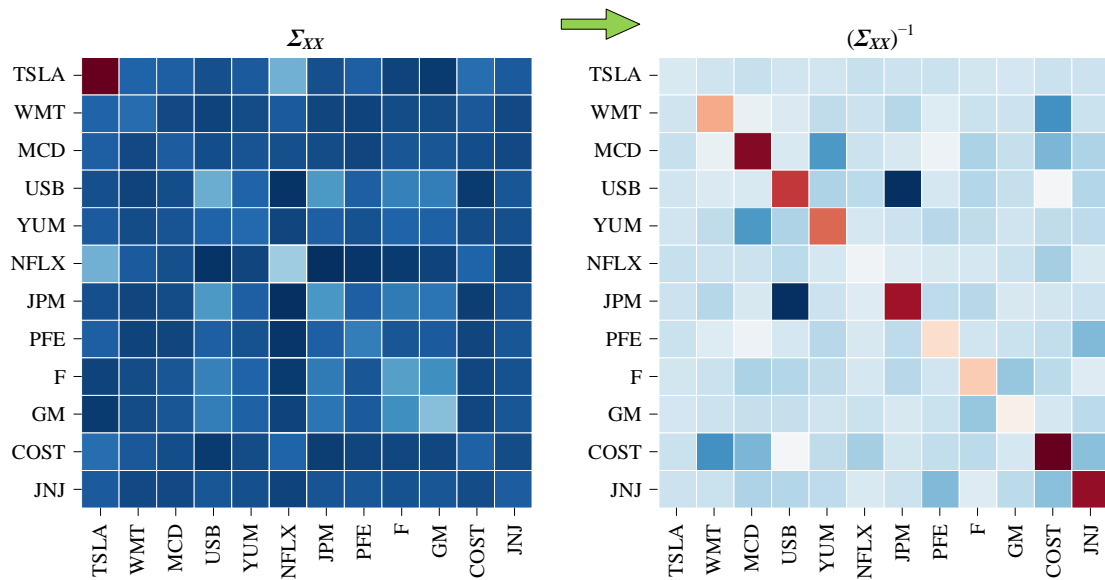


图 29. 分块协方差矩阵求逆

如图 30 所示，截距系数之外的多元线性回归系数向量为：

$$b_{1-D} = (\Sigma_{XX})^{-1} \Sigma_{Xy} \quad (58)$$

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

如图 31 所示, b_0 为:

$$b_0 = E(y) - E(X)b_{1-D} \quad (59)$$

其中, $E(X)$ 为行向量。

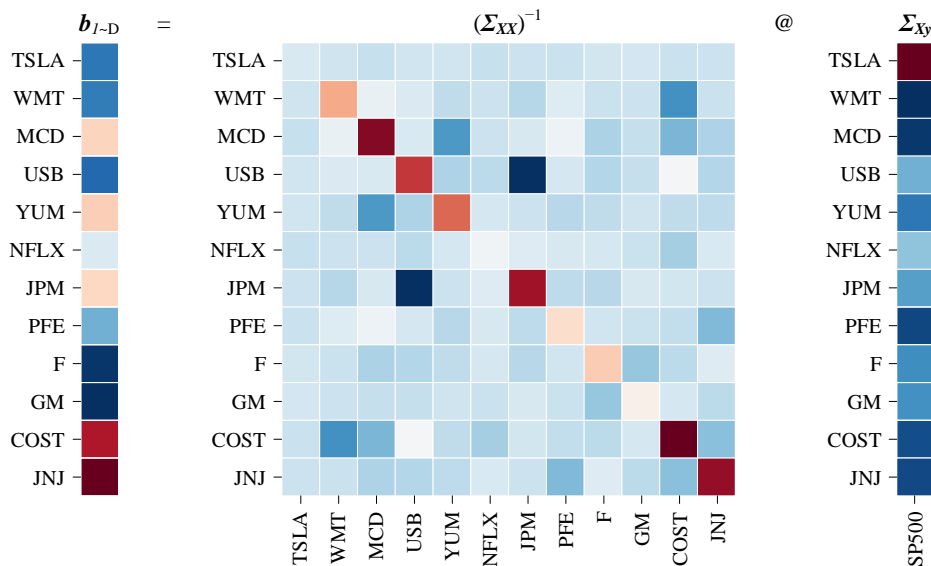


图 30. 求线性回归参数, 除截距以外

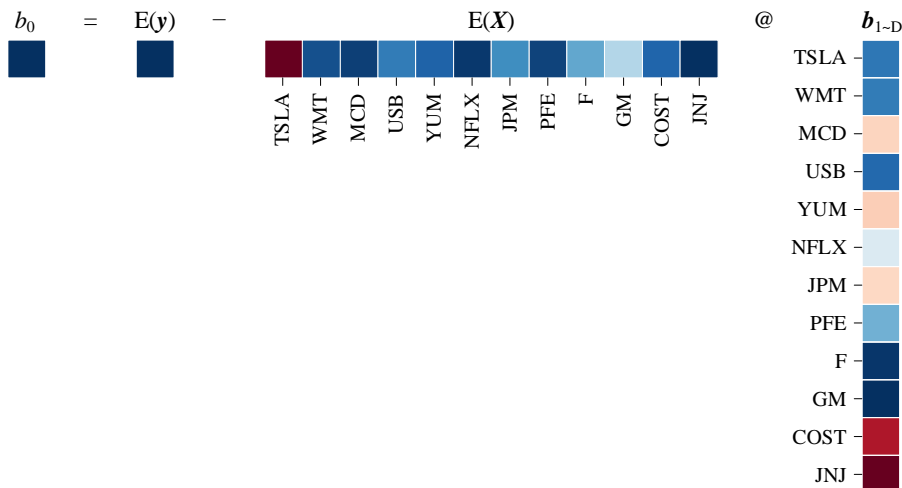
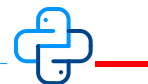
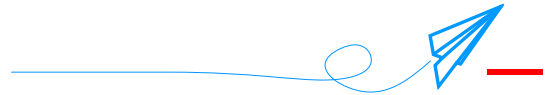


图 31. 求截距系数



Bk7_Ch03_03.ipynb 完成本节运算。



OLS 线性回归是一种在机器学习中常用的算法，它可以通过最小化残差平方和来建立线性模型，从而用于预测和分析因变量与自变量之间的关系。OLS 线性回归适用于数据分析、预测模型、异常检测、特征工程等多种机器学习任务。通过使用 OLS 线性回归，可以得出自变量对因变量的影响程度、探索自变量之间的关系、预测因变量的取值，以及识别异常值等。OLS 线性回归是一种简单但可靠的机器学习算法，为数据分析和预测建模提供了强大的工具和方法。

鸢尾花书从不同视角介绍过 OLS 线性回归。《数学要素》从代数、几何、优化角度讲过线性回归，《矩阵力量》从线性代数、正交投影、矩阵分解视角分析线性回归。《统计至简》又增加了条件概率、MLE 这两个视角。鸢尾花书有关 OLS 线性回归的讲解至此告一段落，本书后续将介绍回归中的正则化、贝叶斯回归、非线性回归等话题。