11 多元线性回归

Multivariate Linear Regression

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

11.1 多元线性回归

这一章将探讨多元线性回归。多元线性回归模型不止一个考虑自变量，而是多个自变量；即回归分析中引入多个相关解释因子。

多元线性回归的数学表达式如下：

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_D x_D + \varepsilon \tag{1}$$

其中，b_0 为截距项；$b_1, b_2, \ldots, b_D$ 代表自变量系数，ε 为残差项，D 为自变量个数。多元线性回归得到一个超平面 (hyperplane)。

用矩阵运算表达 (1)：

$$\boldsymbol{y} = \underbrace{b_0 \boldsymbol{1} + b_1 \boldsymbol{x}_1 + b_2 \boldsymbol{x}_2 + \cdots + b_D \boldsymbol{x}_D}_{\hat{\boldsymbol{y}}} + \boldsymbol{\varepsilon} \tag{2}$$

换一种方式表达 (2)：

$$\boldsymbol{y} = \underbrace{\boldsymbol{X}\boldsymbol{b}}_{\hat{\boldsymbol{y}}} + \boldsymbol{\varepsilon} \tag{3}$$

其中，

$$\boldsymbol{X}_{n\times(D+1)} = \begin{bmatrix} \boldsymbol{1} & \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_D \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,D} \\ 1 & x_{2,1} & \cdots & x_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,D} \end{bmatrix}_{n\times(D+1)}, \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_D \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon^{(1)} \\ \varepsilon^{(2)} \\ \vdots \\ \varepsilon^{(n)} \end{bmatrix} \tag{4}$$

矩阵 $\boldsymbol{X}$ 常被称作设计矩阵 (design matrix)。图 1 所示矩阵运算对应 (3)。

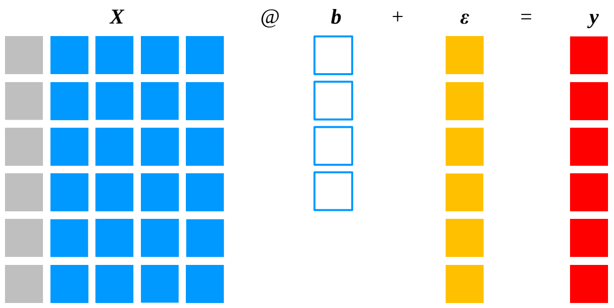


图 1. 多元线性回归模型矩阵运算

预测值构成的列向量 $\hat{\boldsymbol{y}}$，通过下式计算得到：

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{b} \tag{5}$$

残差项的计算式为：

$$\boldsymbol{\varepsilon} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b} \tag{6}$$

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：

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

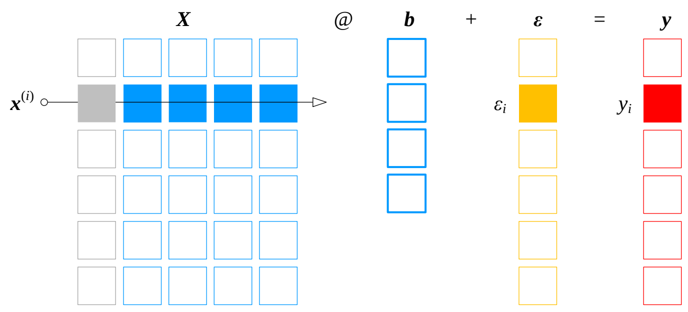


图 2. 计算第 i 个观测点的残差项

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

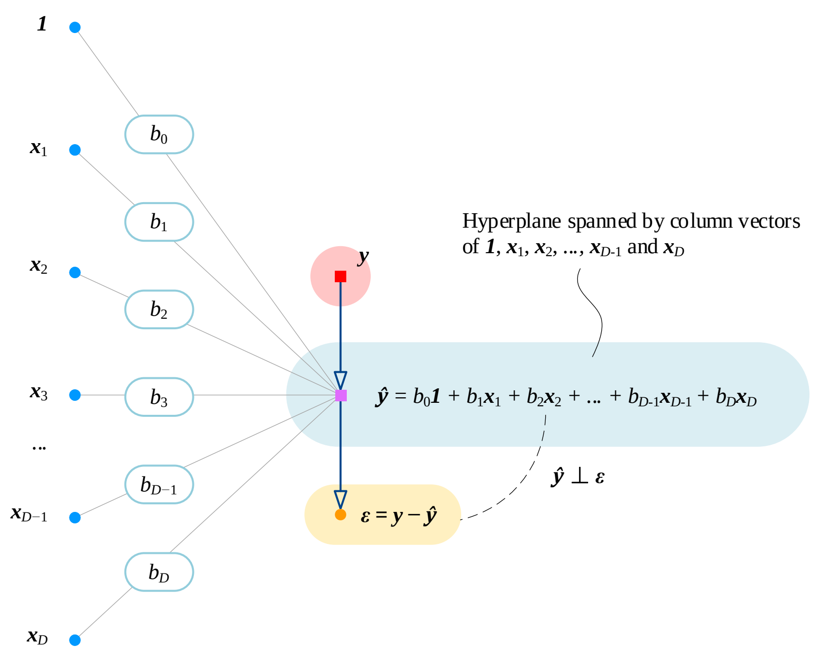


图 3. 多元 OLS 线性回归数据关系

注意，矩阵 $\boldsymbol{X}$ 为 n 行，$D+1$ 列，第一列为全 1；增加一列全 1 向量目的是为了考虑常数项。如图 4 所示，如果数据都已经中心化，则可以不必考虑常数项。

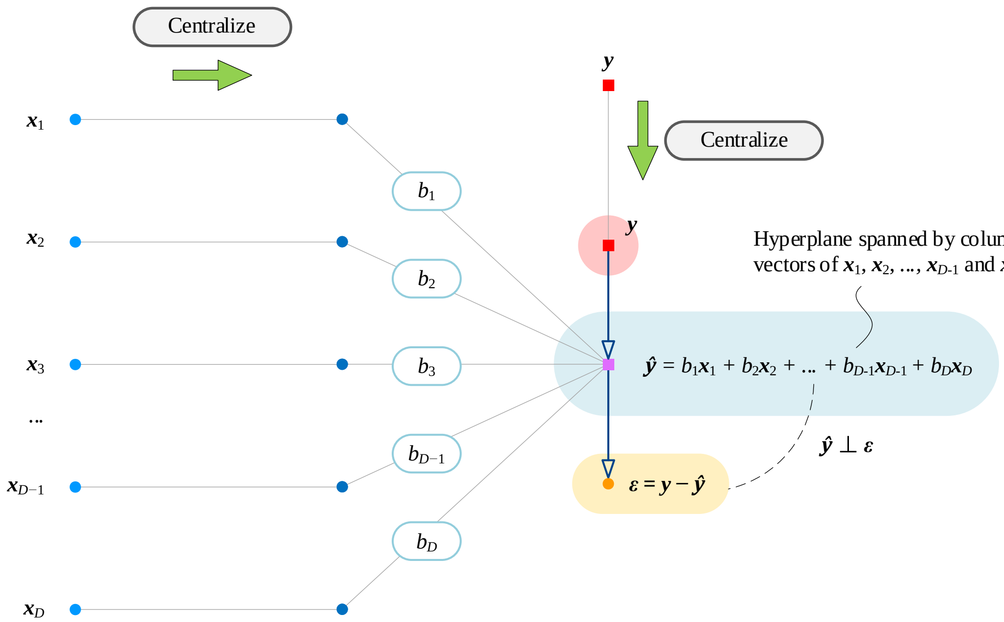


图 4. 多元 OLS 线性回归数据关系，中心化数据

11.2 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：

$$\arg\min_{\boldsymbol{b}} \ \text{SSE} \tag{7}$$

对于多元线性回归，残差平方和 SSE 为：

$$\text{SSE} = \sum_{i=1}^{n}\left(\varepsilon^{(i)}\right)^2 = \boldsymbol{\varepsilon}\cdot\boldsymbol{\varepsilon} = \|\boldsymbol{\varepsilon}\|_2^2 = \boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon} = \left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}\right)^{\mathrm{T}}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}\right) = \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}\|_2^2 \tag{8}$$

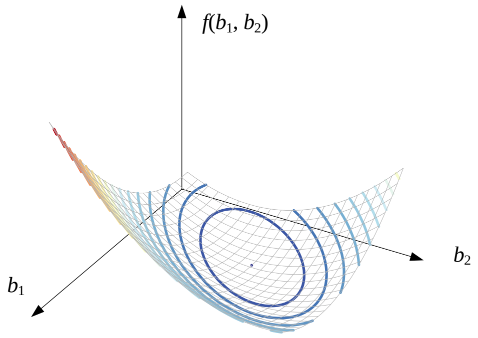
OLS 多元线性优化问题的目标函数可以写成：

$$f\left(\boldsymbol{b}\right) = \left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}\right)^{\mathrm{T}}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}\right) \tag{9}$$

$f(\boldsymbol{b})$ 可以整理为：

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

观察上式，发现 $f(\boldsymbol{b})$ 为多元二次函数，含有二次项、一次项和常数项。因此，对于二元回归，不考虑常数项系数 b_0，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。



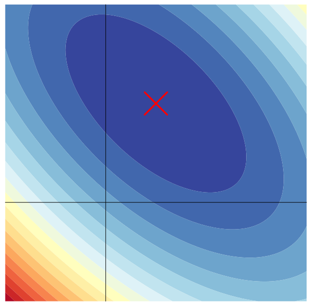


图 5. $f(b_1, b_2)$ 函数曲面

$f(\boldsymbol{b})$ 梯度向量如下：

$$\nabla f(\boldsymbol{b}) = \frac{\partial f(\boldsymbol{b})}{\partial \boldsymbol{b}} = \left(2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{b}-2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}\right)^{\mathrm{T}} \tag{11}$$

$f(\boldsymbol{b})$ 为连续函数，取得极值时，梯度向量为零向量：

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

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

$$\boldsymbol{b} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{13}$$

$f(\boldsymbol{b})$ 的黑塞矩阵为：

$$\nabla^2 f(\boldsymbol{b}) = \frac{\partial^2 f(\boldsymbol{b})}{\partial \boldsymbol{b}\partial \boldsymbol{b}^{\mathrm{T}}} = 2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} \tag{14}$$

下面，判断 $f(\boldsymbol{b})$ 黑塞矩阵为正定矩阵，这样极值点为最小值点。

对于任意非零向量 $\boldsymbol{a}$，下式恒大于等于 0：

$$\boldsymbol{a}^{\mathrm{T}}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)\boldsymbol{a} = (\boldsymbol{X}\boldsymbol{a})^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{a}) = \|\boldsymbol{X}\boldsymbol{a}\|^2 \geq 0 \tag{15}$$

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

模型拟合值向量 $\hat{\boldsymbol{y}}$ 为：

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{16}$$

残差向量 $\boldsymbol{\varepsilon}$ 为：

$$\boldsymbol{\varepsilon} = \boldsymbol{y} - \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{17}$$

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

$$\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}} \tag{18}$$

帽子矩阵 $\boldsymbol{H}$ 为幂等矩阵 (idempotent matrix)，即满足 $\boldsymbol{H}^2 = \boldsymbol{H}$。

利用 $\boldsymbol{H}$，

$$\begin{cases} \hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y} \\ \boldsymbol{\varepsilon} = \left(\boldsymbol{I} - \boldsymbol{H}\right)\boldsymbol{y} \end{cases} \tag{19}$$

11.3 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 $\boldsymbol{X}$ $[\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_D]$ 每一列为一个特征，每一列可以看做一个向量。从书《矩阵力量》一书中，我们反复探讨过这一点。

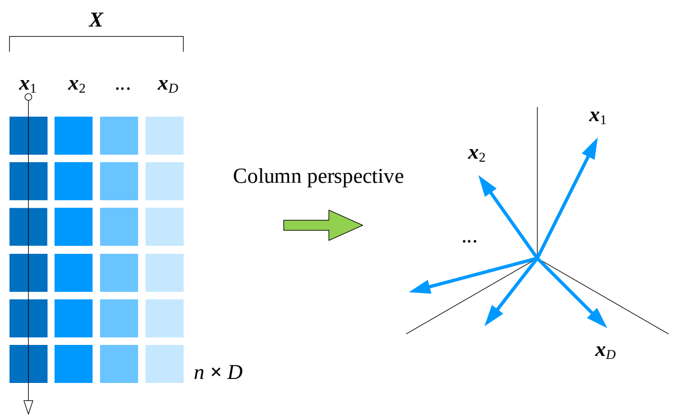


图 6. 多维空间视角下的矩阵 $\boldsymbol{X}$

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

$$\hat{\boldsymbol{y}} = b_1\boldsymbol{x}_1 + b_2\boldsymbol{x}_2 + \cdots + b_D\boldsymbol{x}_D \tag{20}$$

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

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

$$\boldsymbol{y} = \hat{\boldsymbol{y}} + \boldsymbol{\varepsilon} \tag{21}$$

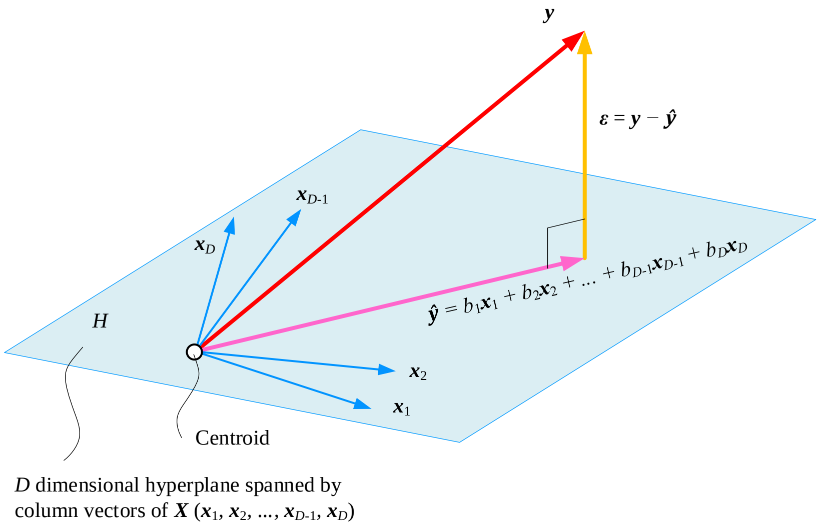


图 7. 几何角度解释多元最小二乘法线性回归

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

$$\boldsymbol{\varepsilon} = \boldsymbol{y} - \hat{\boldsymbol{y}} \tag{22}$$

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

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

$$\boldsymbol{\varepsilon} \perp \boldsymbol{X} \quad \Rightarrow \quad \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon} = 0 \tag{23}$$

得到

$$\boldsymbol{X}^{\mathrm{T}}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\right) = 0 \quad \Rightarrow \quad \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{24}$$

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 $\boldsymbol{X}$ 为列满秩时，$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 才存在逆。

此外，我们可以很容易在 $\boldsymbol{X}$ 最左侧加入一列全 1 向量 $\boldsymbol{1}$，b 的解也可以通过 **Error! Reference source not found.** 计算得到。

表 1 所示为用矩阵方式再次表达 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	$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{b}+\boldsymbol{\varepsilon}$
残差服从正态分布	$\boldsymbol{\varepsilon}\vert\boldsymbol{X}\sim N\left(0,\hat{\sigma}^2\boldsymbol{I}\right)$
残差期望值为 0	$\mathrm{E}\left(\boldsymbol{\varepsilon}\vert\boldsymbol{X}\right)=\boldsymbol{0}$ $\mathrm{E}\left(\boldsymbol{\varepsilon}\right)=\boldsymbol{0}$
残差同方差性	$\mathrm{var}\left(\boldsymbol{\varepsilon}\vert\boldsymbol{X}\right)=\begin{bmatrix}\mathrm{var}\left(\varepsilon^{(1)}\right) & \mathrm{cov}\left(\varepsilon^{(1)},\varepsilon^{(2)}\right) & \cdots & \mathrm{cov}\left(\varepsilon^{(1)},\varepsilon^{(n)}\right)\\ \mathrm{cov}\left(\varepsilon^{(2)},\varepsilon^{(1)}\right) & \mathrm{var}\left(\varepsilon^{(2)}\right) & \cdots & \mathrm{cov}\left(\varepsilon^{(2)},\varepsilon^{(n)}\right)\\ \vdots & \vdots & \ddots & \vdots\\ \mathrm{cov}\left(\varepsilon^{(n)},\varepsilon^{(1)}\right) & \mathrm{cov}\left(\varepsilon^{(n)},\varepsilon^{(2)}\right) & \cdots & \mathrm{var}\left(\varepsilon^{(n)}\right)\end{bmatrix}=\hat{\sigma}^2\boldsymbol{I}$
矩阵 $\boldsymbol{X}$ 不存在多重共线性	$\mathrm{rank}\left(\boldsymbol{X}\right)=D+1$ $\det\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)\neq 0$

11.4 二元回归

二元线性回归解析式为：

$$\hat{\boldsymbol{y}}=b_0\boldsymbol{1}+b_1\boldsymbol{x}_1+b_2\boldsymbol{x}_2 \tag{25}$$

图 8 所示为二元 OLS 线性回归数据关系。

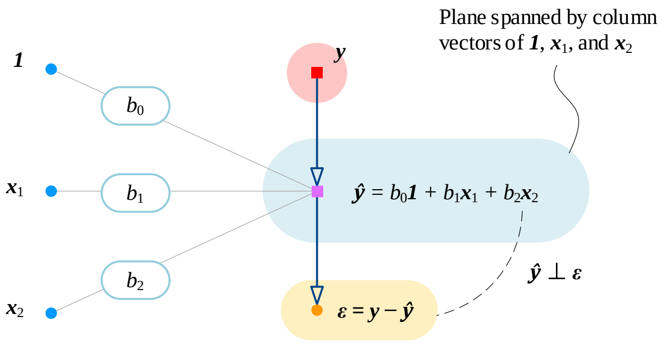


图 8. 二元 OLS 线性回归数据关系

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 [$\boldsymbol{y}$, $\boldsymbol{x}_1$, $\boldsymbol{x}_2$] 的散点图。图 10 所示为 [$\boldsymbol{y}$, $\boldsymbol{x}_1$, $\boldsymbol{x}_2$] 数据的成对特征分析图。图 11 所示为 [$\boldsymbol{y}$, $\boldsymbol{x}_1$, $\boldsymbol{x}_2$] 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。

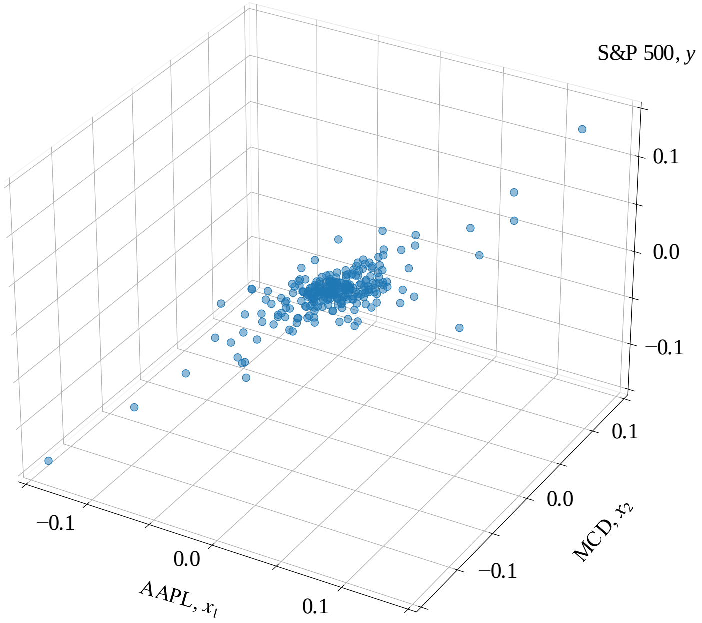


图 9. 二元线性回归数据

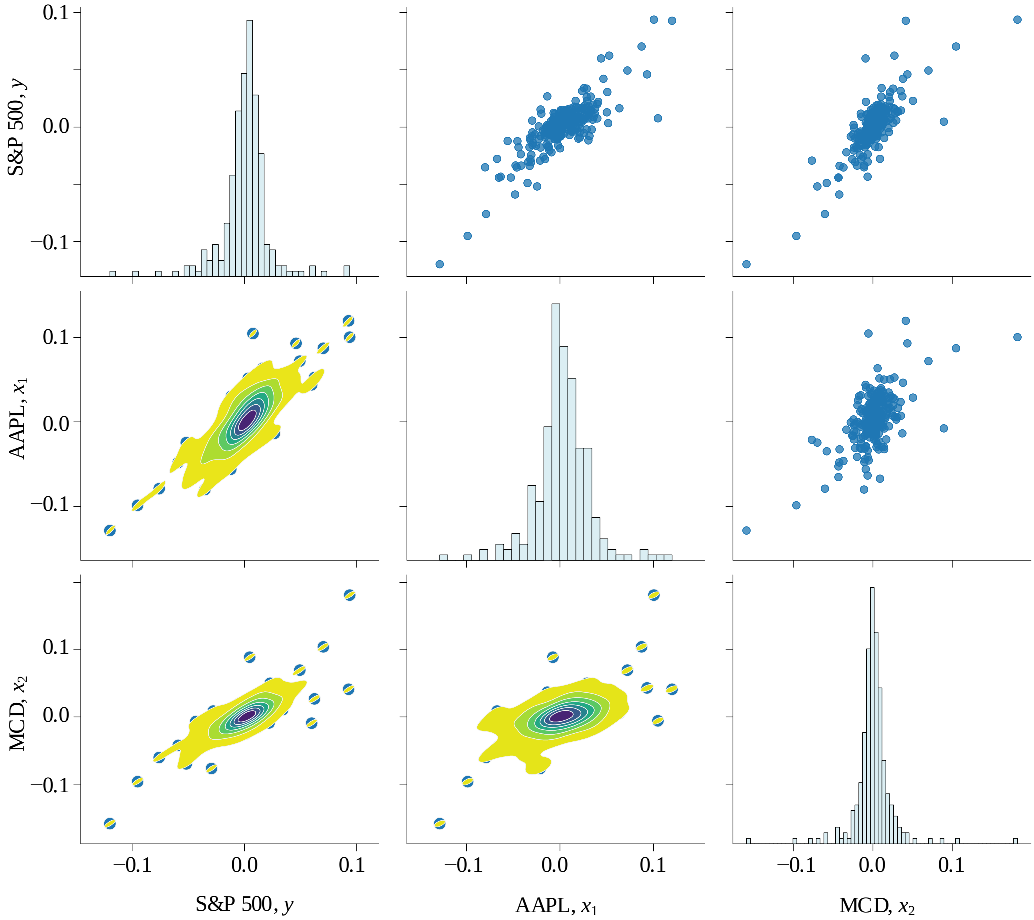


图 10. 二元线性回归数据 $[\boldsymbol{y}, \boldsymbol{x}_1, \boldsymbol{x}_2]$ 成对特征分析图

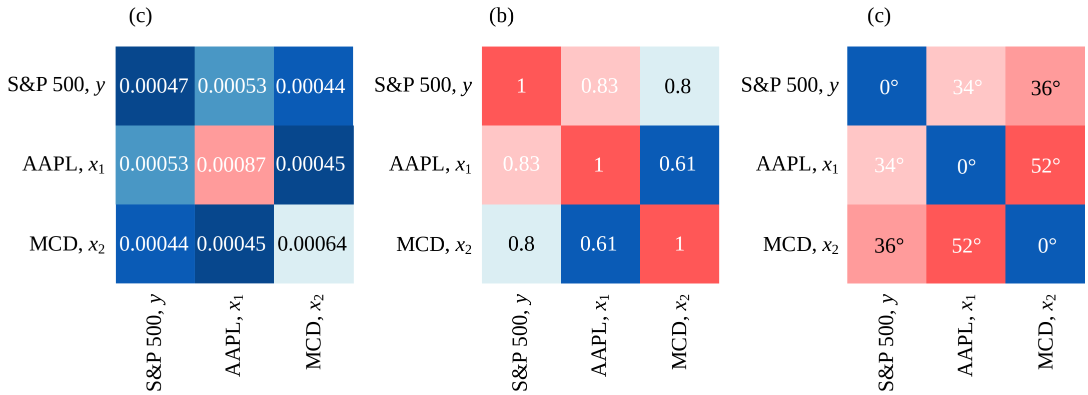


图 11. $[\boldsymbol{y}, \boldsymbol{x}_1, \boldsymbol{x}_2]$ 数据的协方差矩阵、相关性和夹角热图

```
                            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:                XXXXXXXXXXXXXXXX   Prob (F-statistic):           1.69e-96
Time:                XXXXXXXXXXXXXXXX   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

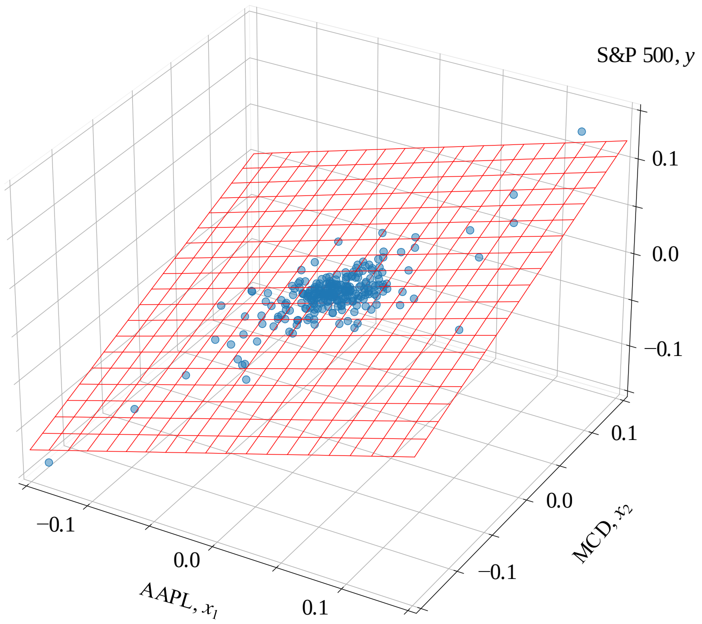


图 13. 三维空间，回归平面

Bk6_Ch11_01.py 完成本节二元线性回归。

11.5 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。

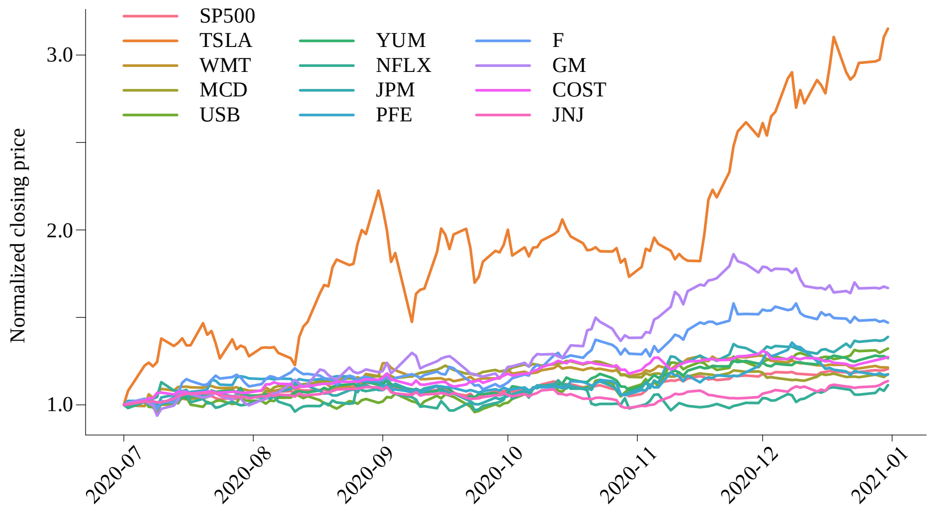


图 14. 股价数据，起始值归一化

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[$\boldsymbol{y}$, $\boldsymbol{X}$] 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[$\boldsymbol{y}$, $\boldsymbol{X}$] 数据相关性系数矩阵热图。图 19 所示为股价收益率和 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[$\boldsymbol{y}$, $\boldsymbol{X}$] 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。

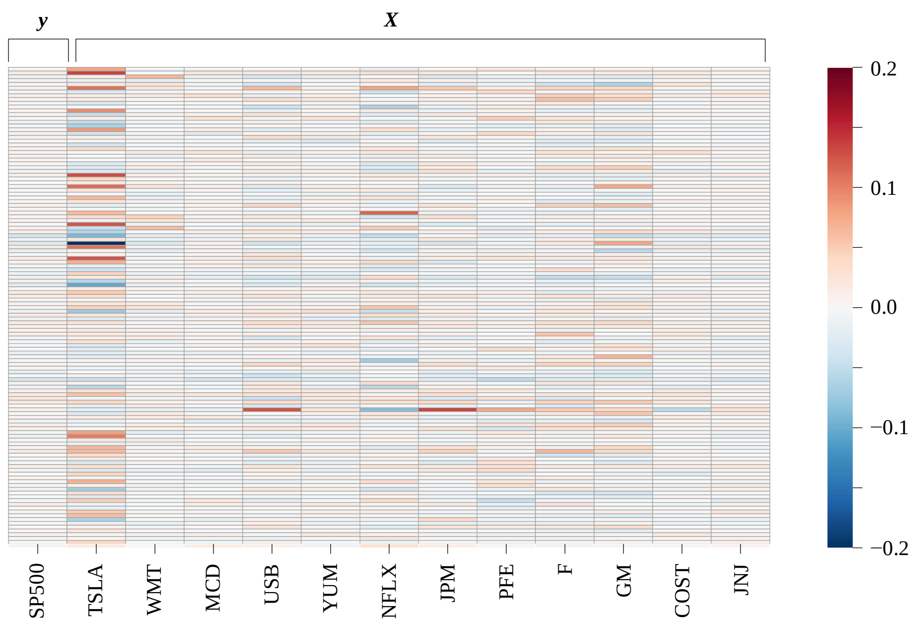


图 15. [$\boldsymbol{y}$, $\boldsymbol{X}$] 日收益率热图

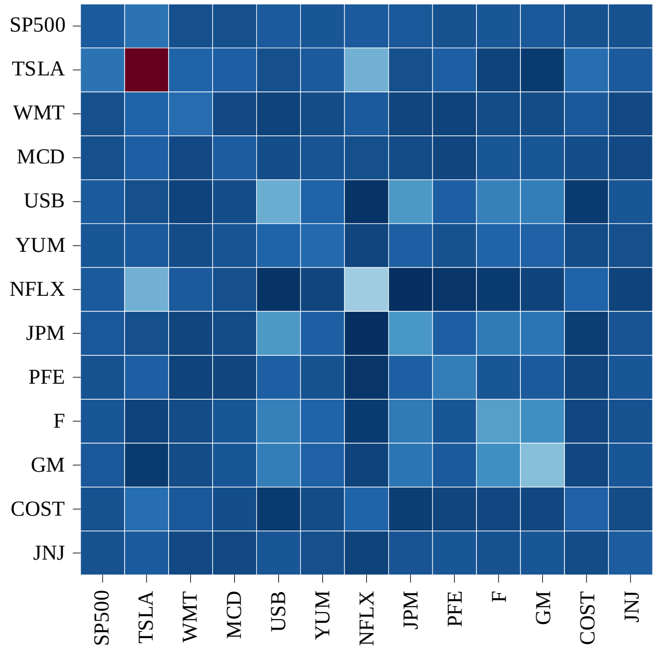


图 16. [$\boldsymbol{y}$, $\boldsymbol{X}$] 数据协方差矩阵

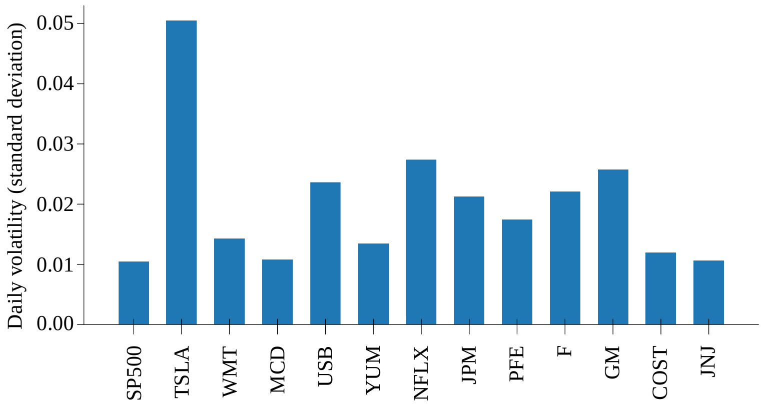


图 17. 日波动率柱状图

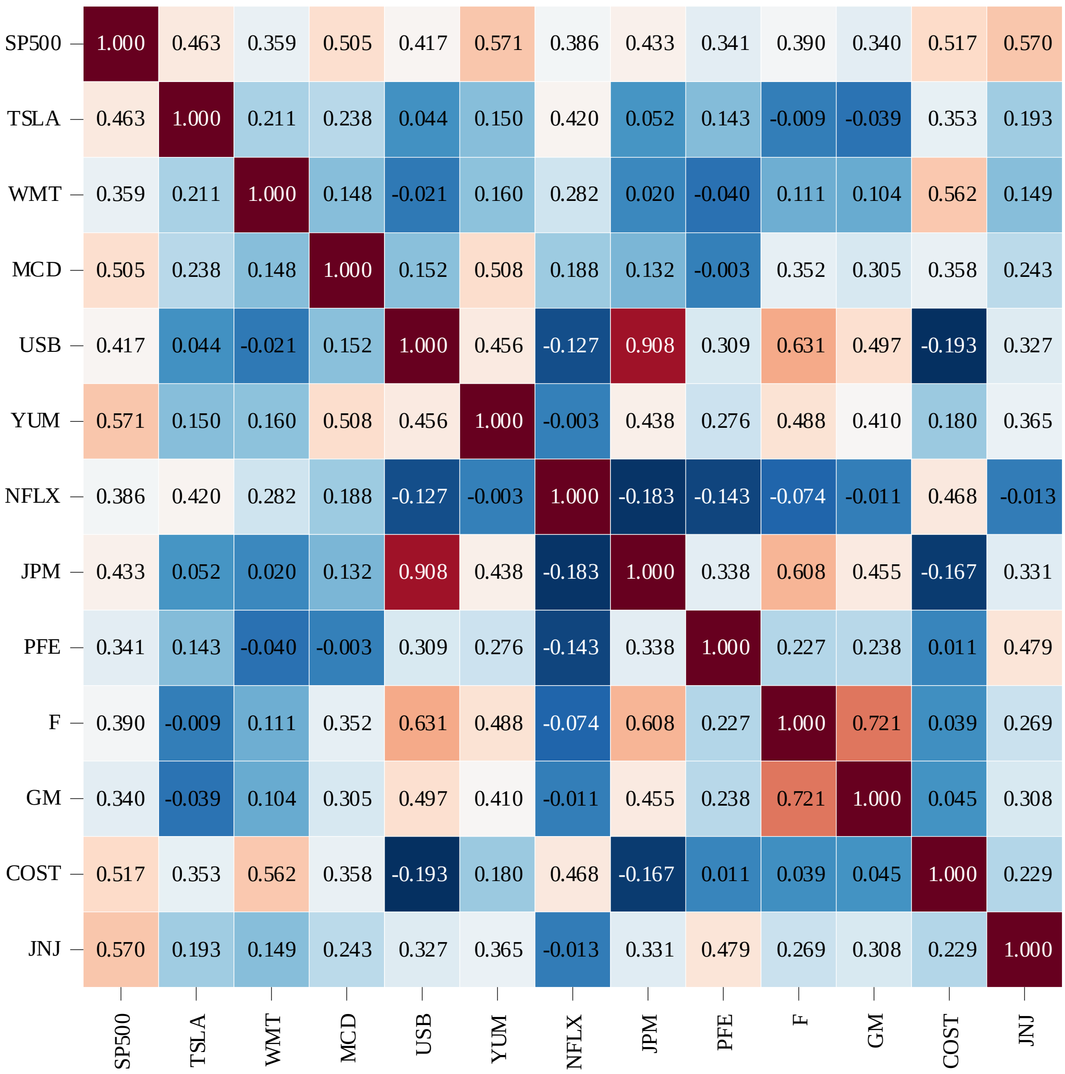


图 18. [y, X] 数据相关性系数矩阵热图

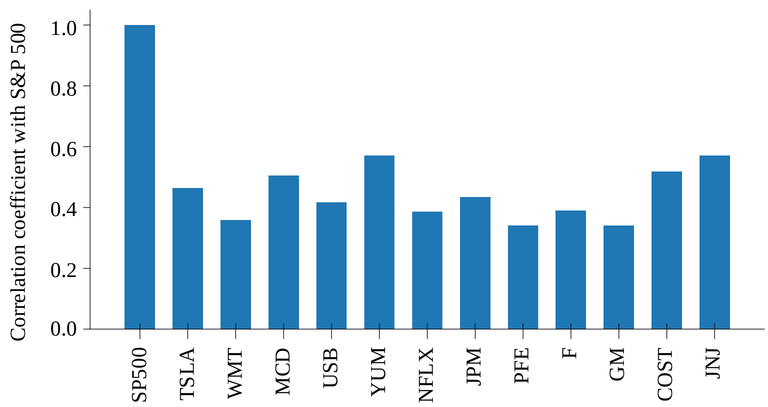


图 19. 股价收益率和 S&P 500 收益率相关性系数柱状图

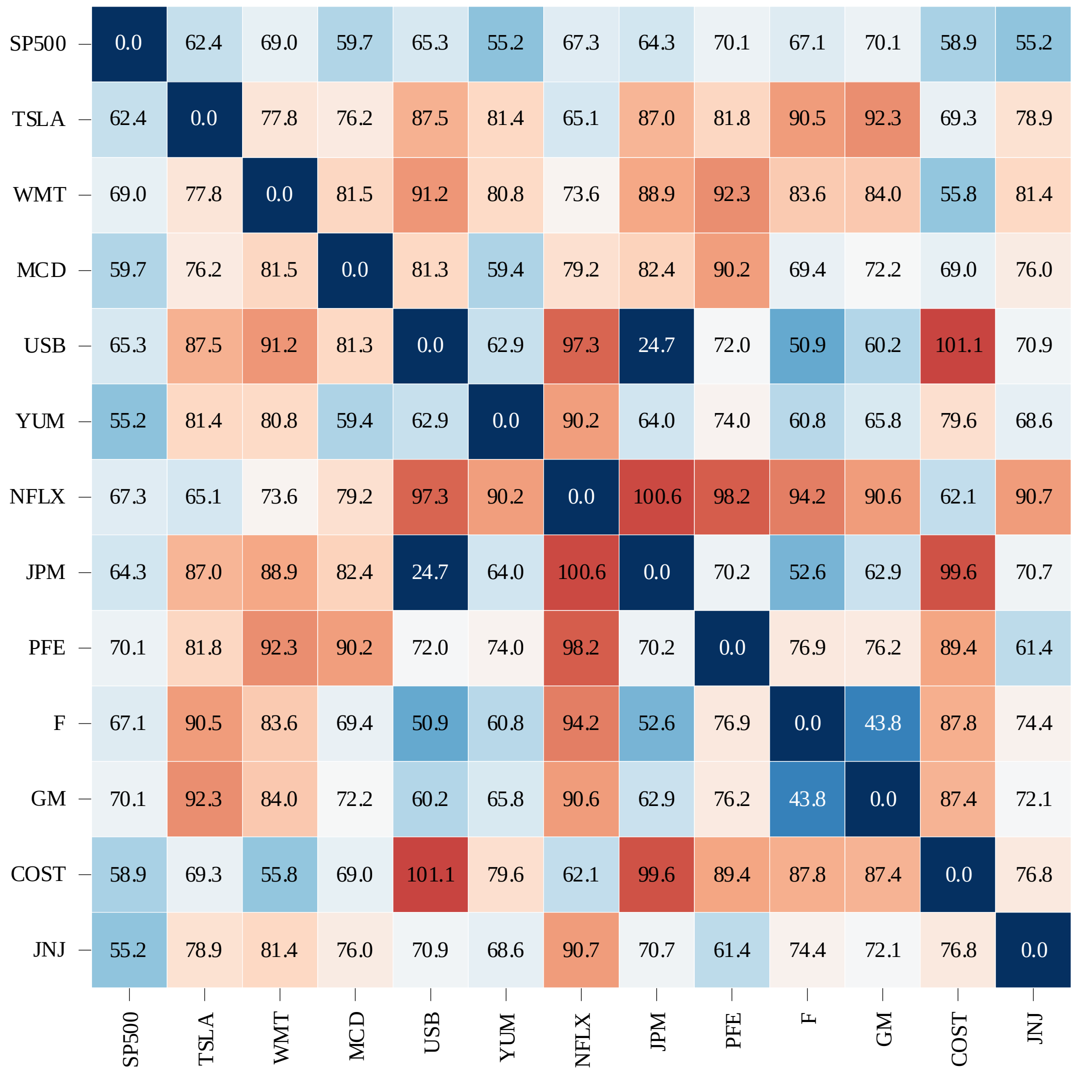


图 20. [$\boldsymbol{y}$, $\boldsymbol{X}$] 标准差向量夹角矩阵热图，余弦相似性

```
                            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:                XXXXXXXXXXXXXXXX   Prob (F-statistic):           3.03e-31
Time:                XXXXXXXXXXXXXXXX   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

Bk6_Ch11_02.py 完成本节多元线性回归。

11.6 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：

$$\underbrace{\left\|\boldsymbol{y}-\bar{y}\boldsymbol{1}\right\|_2^2}_{\text{SST}}=\underbrace{\left\|\hat{\boldsymbol{y}}-\bar{y}\boldsymbol{1}\right\|_2^2}_{\text{SSR}}+\underbrace{\left\|\boldsymbol{y}-\hat{\boldsymbol{y}}\right\|_2^2}_{\text{SSE}} \tag{26}$$

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。

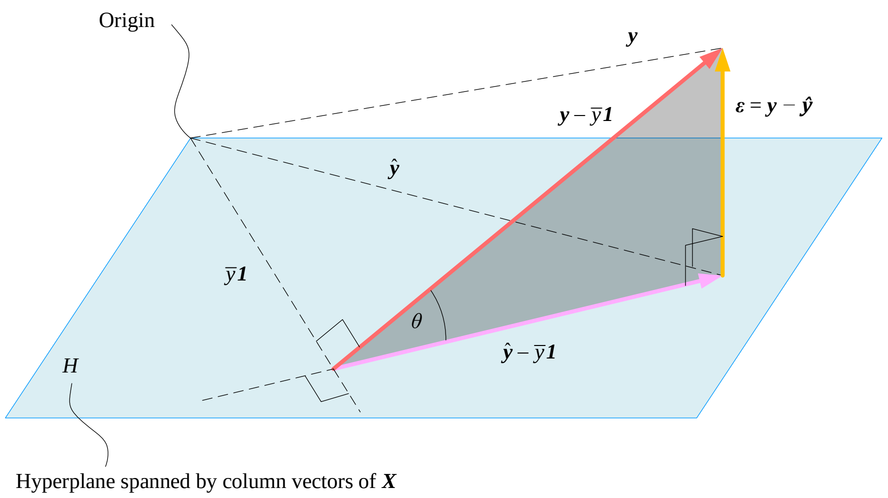


图 22. 第一个直角三角形

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。

$$\|\boldsymbol{y}\|_2^2=\|\hat{\boldsymbol{y}}\|_2^2+\|\boldsymbol{y}-\hat{\boldsymbol{y}}\|_2^2=\|\hat{\boldsymbol{y}}\|_2^2+\|\boldsymbol{\varepsilon}\|_2^2 \tag{27}$$

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

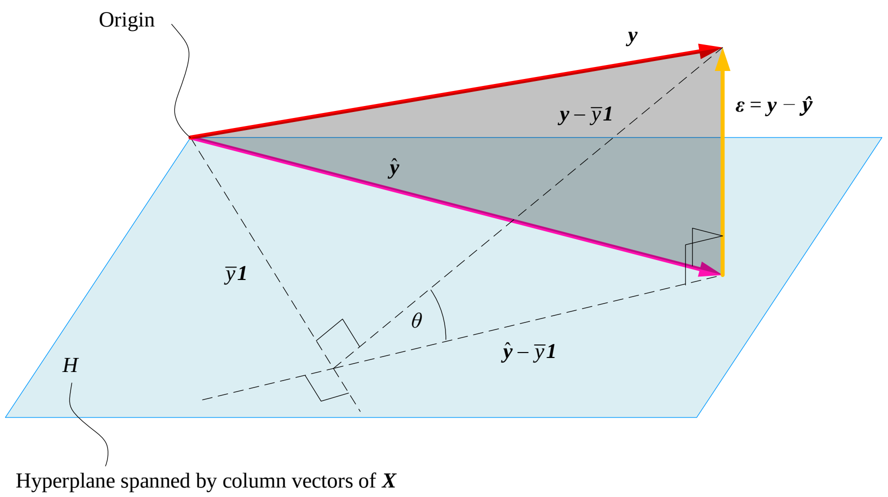


图 23. 第二个直角三角形

第三个直角三角形

此外，《矩阵力量》第 22 章介绍过，向量 $\boldsymbol{y}-\bar{y}\boldsymbol{1}$ 垂直于向量 $\bar{y}\boldsymbol{1}$：

$$(\bar{y}\boldsymbol{1})^{\mathrm{T}}(\boldsymbol{y}-\bar{y}\boldsymbol{1})=0 \tag{28}$$

具体如图 24 所示。上式体现的核心思想就是 $\boldsymbol{y}$ 的均值为 $\bar{y}$。

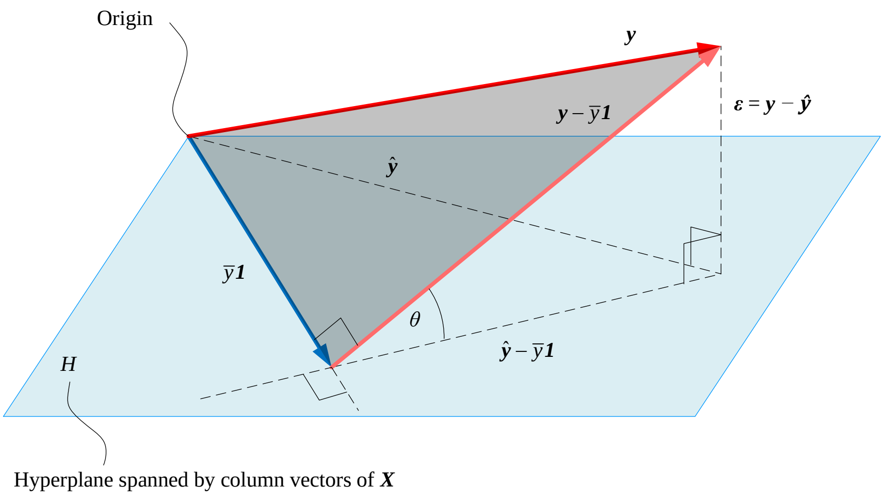


图 24. 第三个直角三角形

第四个直角三角形

OLS 假设残差之和为 0：

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

对应向量运算：

$$\boldsymbol{1}^{\mathrm{T}}\boldsymbol{\varepsilon}=\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{1}=0 \tag{30}$$

残差向量可以写成：

$$\boldsymbol{\varepsilon}=\boldsymbol{y}-\hat{\boldsymbol{y}}=\boldsymbol{y}-\bar{y}\boldsymbol{1}-(\hat{\boldsymbol{y}}-\bar{y}\boldsymbol{1}) \tag{31}$$

上式左乘 $\boldsymbol{1}^{\mathrm{T}}$，得到：

$$\underbrace{\boldsymbol{1}^{\mathrm{T}}\boldsymbol{\varepsilon}}_{0}=\underbrace{\boldsymbol{1}^{\mathrm{T}}(\boldsymbol{y}-\bar{y}\boldsymbol{1})}_{0}-\boldsymbol{1}^{\mathrm{T}}(\hat{\boldsymbol{y}}-\bar{y}\boldsymbol{1}) \tag{32}$$

即

$$\mathbf{1}^{\mathrm{T}}\left(\hat{\boldsymbol{y}}-\bar{y}\mathbf{1}\right)=0 \tag{33}$$

也就是说，如图 25 所示， $\hat{\boldsymbol{y}}-\bar{y}\mathbf{1}$ 垂直于向量 $\bar{y}\mathbf{1}$：

$$\bar{y}\mathbf{1}^{\mathrm{T}}\left(\hat{\boldsymbol{y}}-\bar{y}\mathbf{1}\right)=0 \tag{34}$$

上式体现的核心思想就是 $\hat{\boldsymbol{y}}$ 的均值为 $\bar{y}$ 。

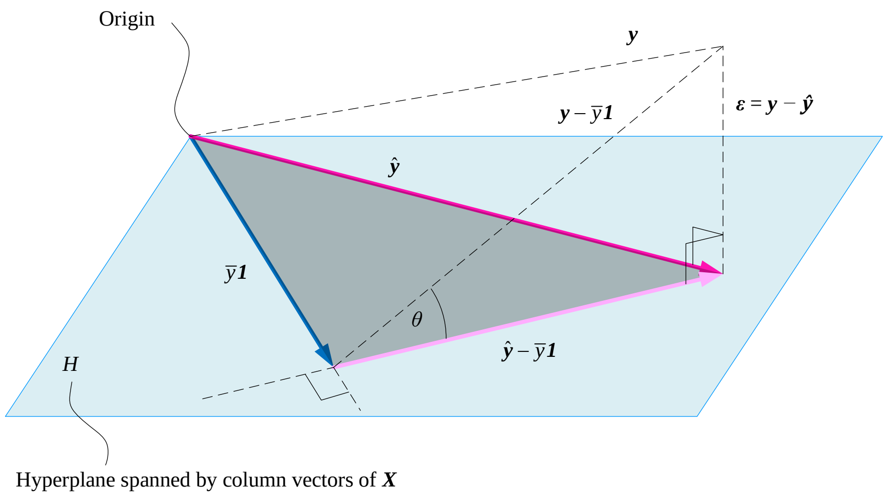


图 25. 第四个直角三角形

11.7 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：

$$\text{SST} = \boldsymbol{y}^{\mathrm{T}}\left(\boldsymbol{I}-\frac{\boldsymbol{J}}{n}\right)\boldsymbol{y} \tag{35}$$

其中矩阵 $\boldsymbol{J}$ 为全 1 方阵，形状为 $n \times n$：

$$\boldsymbol{J}_{n\times n}=\mathbf{1}\mathbf{1}^{\mathrm{T}}=\begin{bmatrix}1 & 1 & \cdots & 1\\ 1 & 1 & \cdots & 1\\ \vdots & \vdots & \ddots & \vdots\\ 1 & 1 & \cdots & 1\end{bmatrix} \tag{36}$$

SSR 可以通过矩阵运算求得：

$$\text{SSR} = \boldsymbol{y}^{\mathrm{T}}\left(\boldsymbol{H}-\frac{\boldsymbol{J}}{n}\right)\boldsymbol{y} \tag{37}$$

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

$$\boldsymbol{H}=\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}} \tag{38}$$

同样，对于多元 OLS 线性回归模型，SSE 可以通过矩阵运算求得：

$$\mathrm{SSE}=\boldsymbol{y}^{\mathrm{T}}\left(\boldsymbol{I}-\boldsymbol{H}\right)\boldsymbol{y} \tag{39}$$

对于多元 OLS 线性回归模型，MSE 的矩阵运算为：

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

上式推导过程采用帽子矩阵重要的性质，帽子矩阵 $\boldsymbol{H}$ 为幂等矩阵 (idempotent matrix)，即满足 $\boldsymbol{H}^2 = \boldsymbol{H}$。

11.8 t 检验

对于多元 OLS 线性回归模型，模型系数 b_0、b_1、b_2 … b_D 的方差协方差矩阵 $\boldsymbol{C}$ 可以通过下式计算得到：

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

其中，

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

矩阵 $\boldsymbol{C}$ 的对角线元素 $\boldsymbol{C}_{j+1,j+1}$ 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 $\mathrm{SE}\left(\hat{b}_j\right)$ 为：

$$\mathrm{SE}\left(\hat{b}_j\right)=\sqrt{\boldsymbol{C}_{j+1,j+1}} \tag{43}$$

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：

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

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：

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

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

$$\hat{y}^{(i)} \pm t_{1-\alpha/2,\, n-2} \cdot \sqrt{\mathrm{MSE}} \cdot \sqrt{\boldsymbol{x}^{(i)}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{x}^{(i)}\right)^{\mathrm{T}}} \tag{48}$$

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

$$\boldsymbol{x}^{(i)} = \begin{bmatrix} 1 & x_{i,1} & x_{i,2} & \cdots & x_{i,D} \end{bmatrix} \tag{49}$$

类似地，对于多元 OLS 线性回归模型，y_{p} 的预测区间估计为：

$$\hat{y}^{(i)} \pm t_{1-\alpha/2,\, n-2} \cdot \sqrt{\mathrm{MSE}} \cdot \sqrt{1 + \boldsymbol{x}^{(i)}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{x}^{(i)}\right)^{\mathrm{T}}} \tag{50}$$

11.9 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 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 进行特征值分解，得到最大特征值 $\lambda_{\max}$ 和最小特征值 $\lambda_{\min}$。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 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 可逆，$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 的行列式值不为 0：

$$\det\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right) \neq 0 \tag{51}$$

这里再介绍一个共线性的度量指标，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 $\hat{\boldsymbol{X}}_t$，对于其中的任意解释变量 $\left\{X_{i,t}\right\}$，其对应的方差膨胀因子 VIF_i 可由下式计算：

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2} \tag{52}$$

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：

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

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，$\mathrm{VIF} \le 5$ 也是比较常见的、但相对而言更为严格的判断标准。

11.10 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 $\boldsymbol{\chi}$ 和 $\boldsymbol{\gamma}$ 服从多维高斯分布：

$$\begin{bmatrix} \boldsymbol{\chi} \\ \boldsymbol{\gamma} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu}_\chi \\ \boldsymbol{\mu}_\gamma \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{\chi\chi} & \boldsymbol{\Sigma}_{\chi\gamma} \\ \boldsymbol{\Sigma}_{\gamma\chi} & \boldsymbol{\Sigma}_{\gamma\gamma} \end{bmatrix} \right) \tag{54}$$

其中，$\boldsymbol{\chi}$ 为随机变量 X_i 构成的列向量，$\boldsymbol{\gamma}$ 为随机变量 Y_j 构成的列向量：

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

如图 26 所示，给定 $\boldsymbol{\chi} = \boldsymbol{x}$ 的条件下 $\boldsymbol{\gamma}$ 的条件**期望**为：

$$\mathrm{E}(\boldsymbol{\gamma} \mid \boldsymbol{\chi} = \boldsymbol{x}) = \boldsymbol{\mu}_{\gamma|\chi=x} = \boldsymbol{\Sigma}_{\gamma\chi} \boldsymbol{\Sigma}_{\chi\chi}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_\chi) + \boldsymbol{\mu}_\gamma \tag{56}$$

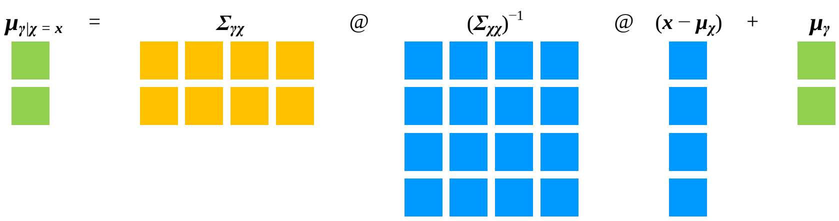


图 26. 给定 $\boldsymbol{\chi} = \boldsymbol{x}$ 的条件下 $\boldsymbol{\gamma}$ 的**期望**值的矩阵运算，图片来自《统计至简》

对于本例，我们对 (56) 进行转置得到：

$$\mu_{y|x} = \mathrm{E}(y) + \left(x - \mathrm{E}(X)\right)\underbrace{\left(\Sigma_{XX}\right)^{-1}\Sigma_{Xy}}_{b} \tag{57}$$

[y, X] 对应的协方差矩阵如图 27 所示。图 28 为对 Σ_{XX} 求逆。

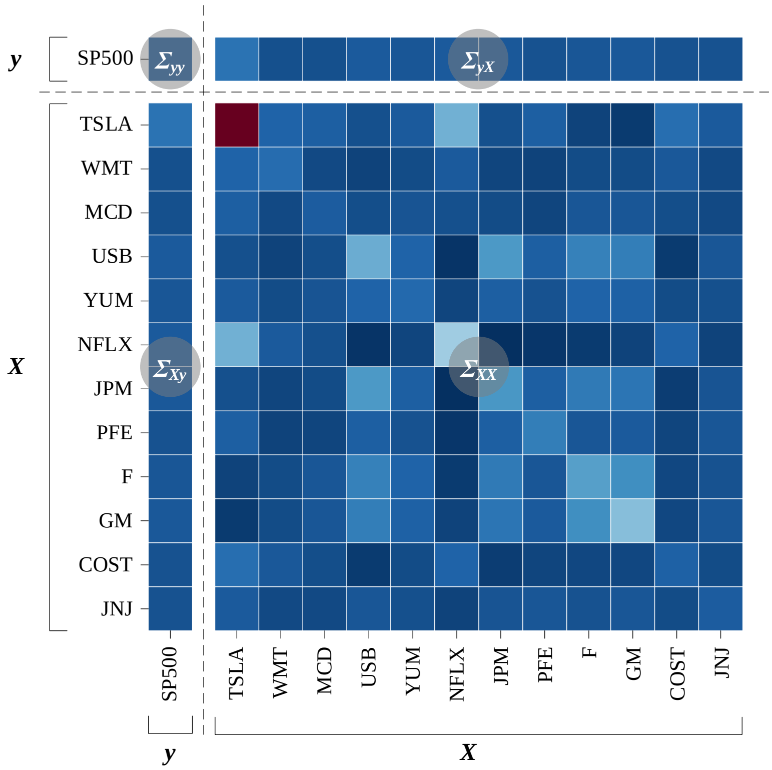


图 27. [y, X] 协方差矩阵

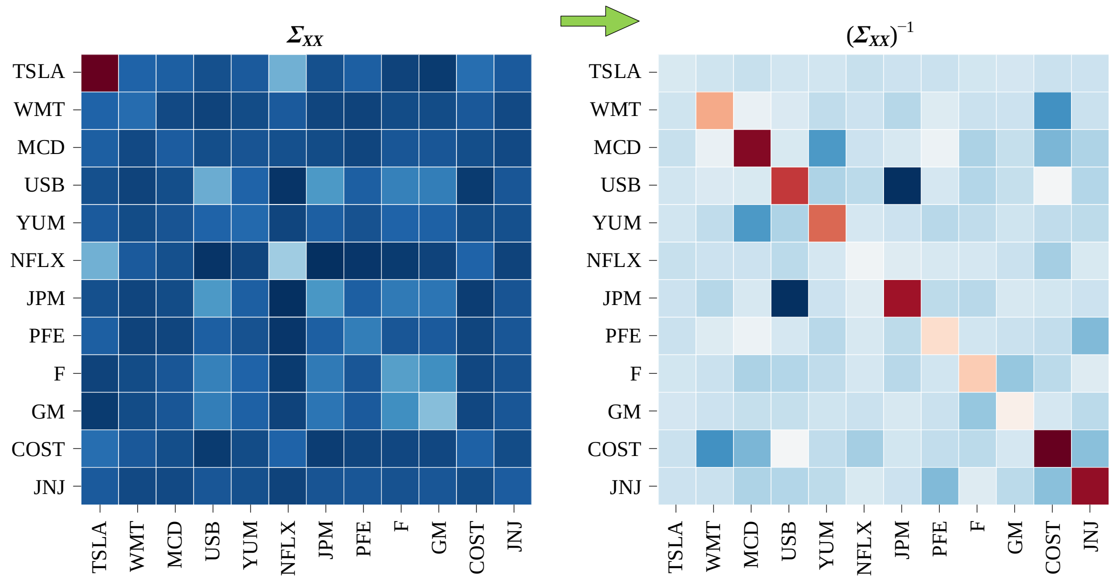


图 28. 分块协方差矩阵求逆

如图 29 所示，截距系数之外的多元线性回归系数向量为：

$$\boldsymbol{b}_{1\sim D} = \left(\boldsymbol{\Sigma}_{XX}\right)^{-1} \boldsymbol{\Sigma}_{Xy} \tag{58}$$

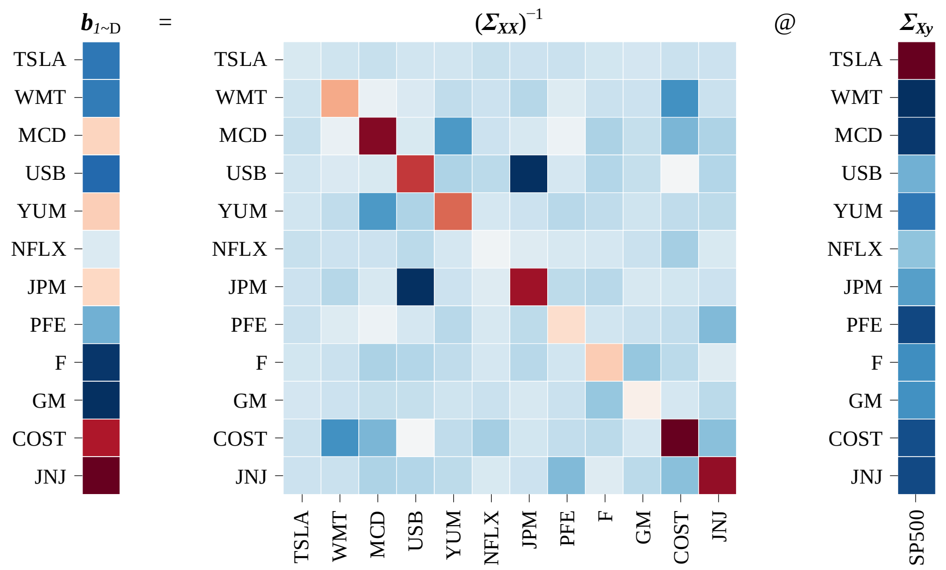


图 29. 求线性回归参数，除截距以外

如图 30 所示，b_0 为：

$$b_0 = \mathrm{E}\left(\boldsymbol{y}\right) - \mathrm{E}\left(\boldsymbol{X}\right)\boldsymbol{b}_{1\sim D} \tag{59}$$

其中，$\mathrm{E}(\boldsymbol{X})$ 为行向量。

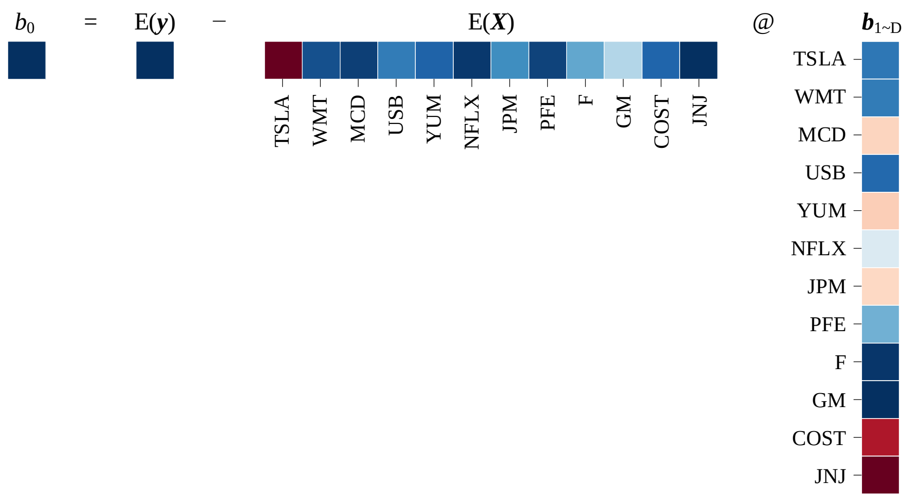


图 30. 求截距系数

Bk6_Ch11_03.py 完成本节运算。