11

Multivariate Linear Regression

3 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.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—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—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:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_D x_D + \varepsilon$$
 (1)

其中, b_0 为截距项; $b_1, b_2, ..., b_D$ 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):

$$\mathbf{y} = \underbrace{b_0 \mathbf{1} + b_1 \mathbf{x}_1 + b_2 \mathbf{x}_2 + \dots + b_D \mathbf{x}_D}_{\hat{\mathbf{y}}} + \boldsymbol{\varepsilon}$$
 (2)

换一种方式表达(2):

$$y = Xb + \varepsilon \tag{3}$$

其中.

$$\boldsymbol{X}_{n \times (D+1)} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{x}_{1} & \boldsymbol{x}_{2} & \cdots & \boldsymbol{x}_{D} \end{bmatrix} = \begin{bmatrix} 1 & \boldsymbol{x}_{1,1} & \cdots & \boldsymbol{x}_{1,D} \\ 1 & \boldsymbol{x}_{2,1} & \cdots & \boldsymbol{x}_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \boldsymbol{x}_{n,1} & \cdots & \boldsymbol{x}_{n,D} \end{bmatrix}_{n \times (D+1)}, \quad \boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_{1} \\ \boldsymbol{y}_{2} \\ \vdots \\ \boldsymbol{y}_{n} \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}_{0} \\ \boldsymbol{b}_{1} \\ \vdots \\ \boldsymbol{b}_{D} \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \boldsymbol{\varepsilon}^{(1)} \\ \boldsymbol{\varepsilon}^{(2)} \\ \vdots \\ \boldsymbol{\varepsilon}^{(n)} \end{bmatrix}$$
(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)。

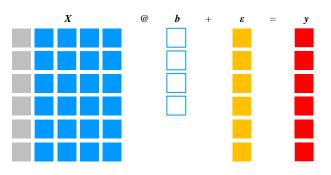


图 1. 多元线性回归模型矩阵运算

预测值构成的列向量 ŷ, 通过下式计算得到:

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

残差项的计算式为:

$$\varepsilon = y - \hat{y} = y - Xb \tag{6}$$

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:

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

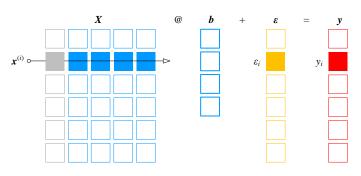


图 2. 计算第 i 个观测点的残差项

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。

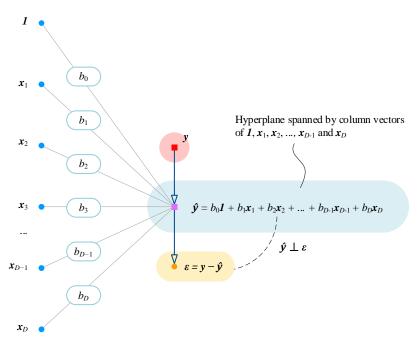


图 3. 多元 OLS 线性回归数据关系

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

本书配套徽课视频均发布在 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

注意,矩阵 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。

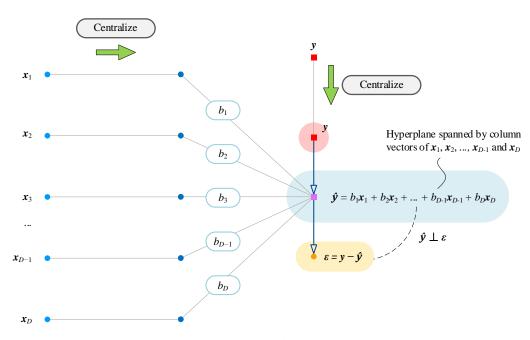


图 4. 多元 OLS 线性回归数据关系,中心化数据

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:

$$\underset{b}{\operatorname{arg\,min}} \text{ SSE} \tag{7}$$

对于多元线性回归,残差平方和 SSE 为:

$$SSE = \sum_{i=1}^{n} \left(\varepsilon^{(i)} \right)^{2} = \varepsilon \cdot \varepsilon = \left\| \varepsilon \right\|_{2}^{2} = \varepsilon^{T} \varepsilon = \left(y - Xb \right)^{T} \left(y - Xb \right) = \left\| y - Xb \right\|_{2}^{2}$$
(8)

OLS 多元线性优化问题的目标函数可以写成:

$$f(b) = (y - Xb)^{\mathsf{T}} (y - Xb) \tag{9}$$

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

f(b) 可以整理为:

$$f(b) = (y - Xb)^{T} (y - Xb)$$

$$= (y^{T} - b^{T} X^{T}) (y - Xb)$$

$$= y^{T} y - y^{T} Xb - b^{T} X^{T} y + b^{T} X^{T} Xb$$

$$= \underbrace{b^{T} X^{T} Xb}_{Quadratic term} \underbrace{-2b^{T} X^{T} y}_{Linear term} + y^{T} y$$

$$= \underbrace{b^{T} X^{T} Xb}_{Quadratic term} \underbrace{-2b^{T} X^{T} y}_{Constant} + y^{T} y$$

$$= \underbrace{b^{T} X^{T} Xb}_{Quadratic term} \underbrace{-2b^{T} X^{T} y}_{Constant} + y^{T} y$$

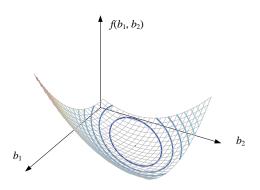
$$= \underbrace{b^{T} X^{T} Xb}_{Quadratic term} \underbrace{-2b^{T} X^{T} y}_{Constant} + y^{T} y$$

$$= \underbrace{b^{T} X^{T} Xb}_{Quadratic term} \underbrace{-2b^{T} X^{T} y}_{Constant} + y^{T} y$$

$$= \underbrace{b^{T} X^{T} Xb}_{Quadratic term} \underbrace{-2b^{T} X^{T} y}_{Constant} + y^{T} y$$

$$= \underbrace{b^{T} X^{T} Xb}_{Quadratic term} \underbrace{-2b^{T} X^{T} y}_{Constant} + y^{T} y$$

观察上式,发现 $f(\mathbf{b})$ 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。



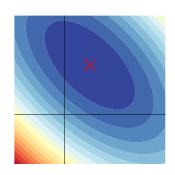


图 5. f(b1, b2) 函数曲面

f(b) 梯度向量如下:

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

f(b) 为连续函数,取得极值时,梯度向量为零向量:

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

如果 X^TX 可逆,b的解为:

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

f(b) 的黑塞矩阵为:

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

下面,判断 $f(\mathbf{b})$ 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a,下式恒大于等于0:

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

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

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

模型拟合值向量 ŷ 为:

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

残差向量 ε 为:

$$\varepsilon = y - X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}y \tag{17}$$

 $X(X^TX)^{-1}X^T$ 为《矩阵力量》第 9 章介绍的帽子矩阵 (hat matrix) 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 H 为幂等矩阵 (idempotent matrix),即满足 $H^2 = H$ 。

利用H,

$$\begin{cases} \hat{y} = Hy \\ \varepsilon = (I - H)y \end{cases}$$
 (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 $X[x_1,x_2,...,x_D]$ 每一列为一个特征,每一列可以看做一个向量。丛书《矩阵力量》一书中,我们反复探讨过这一点。

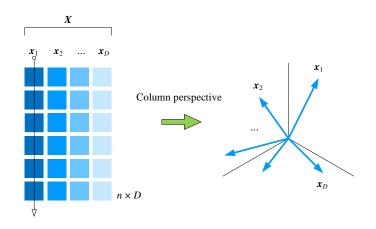


图 6. 多维空间视角下的矩阵 X

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

不考虑常数项,预测值向量 \hat{y} 可以通过下式计算得到:

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

(20) 说明,预测值向量 \hat{y} 是自变量向量 $x_1, x_2, ..., x_D$ 的线性组合。如果 $x_1, x_2, ..., x_D$,构成一个超平面 H; \hat{y} 在 H 这个平面内。

有了这一思想,构造因变量向量y和自变量向量 $x_1, x_2, ..., x_D$ 的线性回归模型,相当于y向 $x_1, x_2, ..., x_D$ 构成的超平面H投影。如图7所示,预测值向量 \hat{y} 是因变量向量y在H的投影结果:

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

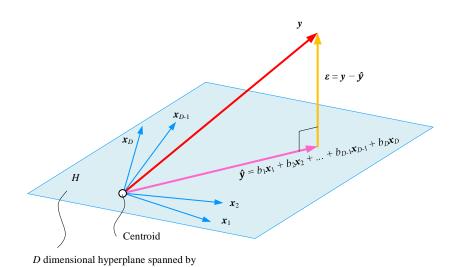


图 7. 几何角度解释多元最小二乘法线性回归

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

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

残差项向量 ε 垂直于 $x_1, x_2, ..., x_D$ 构成的超平面 H。

column vectors of $X(x_1, x_2, ..., x_{D-1}, x_D)$

由上所述,残差 ε ($\varepsilon = y - \hat{y}$) 是无法通过 ($x_0, x_1, ..., x_{D-1}, x_D$) 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

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

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 X 为列满秩时, X^TX 才存在逆。

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

此外,我们可以很容易在X最左侧加入一列全1向量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。

ı Z	矩阵表达

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

表 1. 用矩阵运算表达 OLS 线性回归假设

二元线性回归解析式为:

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

图 8 所示为二元 OLS 线性回归数据关系。

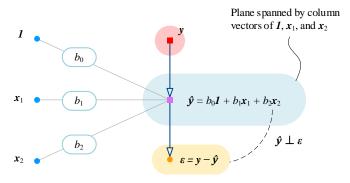


图 8. 二元 OLS 线性回归数据关系

本 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

本书配套徽课视频均发布在 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

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。

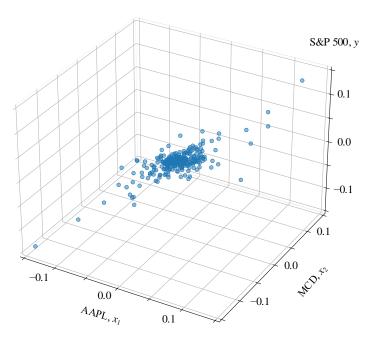


图 9. 二元线性回归数据

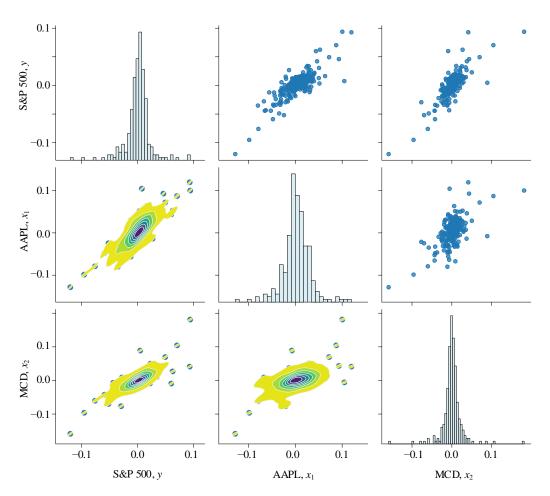


图 10. 二元线性回归数据 $[y,x_1,x_2]$ 成对特征分析图

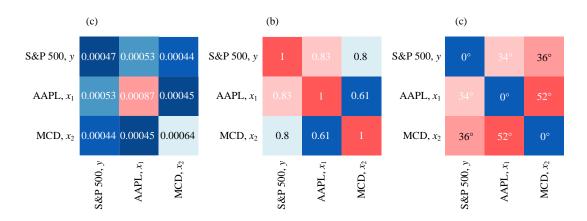


图 $11. [y, x_1, x_2]$ 数据的协方差矩阵、相关性和夹角热图

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: ht —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

OLS Regression Results

	=======						
Dep. Variable:				R-squared:			
	(OLS .	Adj.	0.829			
	Least Squa	res	F-sta		607.4		
XXX	XXXXXXXXXX	XXX	Prob	(F-statistic):		1.69e-96	
Time: XXXXX No. Observations:				ikelihood:		831.06	
ns:	:	252	AIC:	-1656.			
	:	249	BIC:			-1646.	
		2					
e:	nonrob	ust					
coef	std err		===== t	P> t	[0.025	0.975]	
-0.0006	0.001	-0.	 984	 0.326	-0.002	0.001	
0.3977	0.024	16.	326	0.000	0.350	0.446	
0.4096	0.028	14.	442	0.000	0.354	0.465	
=======	37.	===== 744	===== Durbi	======== n-Watson:		 1.991	
	0.	000	Jarqu	e-Bera (JB):		157.711	
	0.	492	Prob(JB):		5.67e-35	
	6.	749	Cond.	No.		59.4	
	xxx xxx ns: e: 	Least Squa	OLS Least Squares XXXXXXXXXXXXXXXX XXXXXXXXXXXXXXX ns: 252 249 2 e: nonrobust -0.0006 0.001 -0. 0.3977 0.024 16. 0.4096 0.028 14. 37.744 0.000 0.492	OLS Adj. Least Squares F-sta XXXXXXXXXXXXXXXX Prob XXXXXXXXXXXXXXX Log-L ns: 252 AIC: 249 BIC: 2 e: nonrobust coef std err t -0.0006 0.001 -0.984 0.3977 0.024 16.326 0.4096 0.028 14.442 37.744 Durbi 0.000 Jarqu 0.492 Prob(OLS Adj. R-squared: Least Squares F-statistic: XXXXXXXXXXXXXXXX Prob (F-statistic): XXXXXXXXXXXXXXX Log-Likelihood: ns: 252 AIC: 249 BIC: 2 e: nonrobust coef std err t P> t -0.0006 0.001 -0.984 0.326 0.3977 0.024 16.326 0.000 0.4096 0.028 14.442 0.000 37.744 Durbin-Watson: 0.000 Jarque-Bera (JB): 0.492 Prob(JB):	OLS Adj. R-squared: Least Squares F-statistic: XXXXXXXXXXXXXXXX Prob (F-statistic): XXXXXXXXXXXXXXX Log-Likelihood: ns: 252 AIC: 249 BIC: 2 e: nonrobust coef std err t P> t [0.025] -0.0006 0.001 -0.984 0.326 -0.002 0.3977 0.024 16.326 0.000 0.350 0.4096 0.028 14.442 0.000 0.354 37.744 Durbin-Watson: 0.000 Jarque-Bera (JB): 0.492 Prob(JB):	

图 12. 二元 OLS 线性回归分析结果

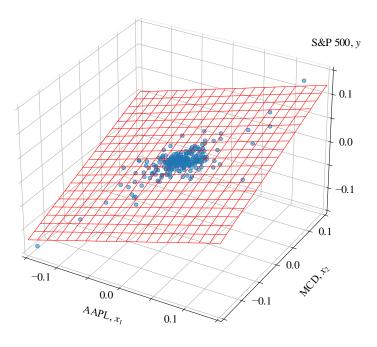


图 13. 三维空间, 回归平面



Bk6_Ch11_01.py 完成本节二元线性回归。

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

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。

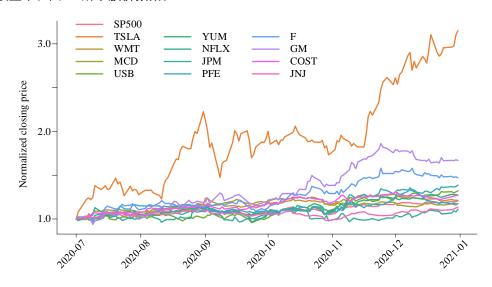
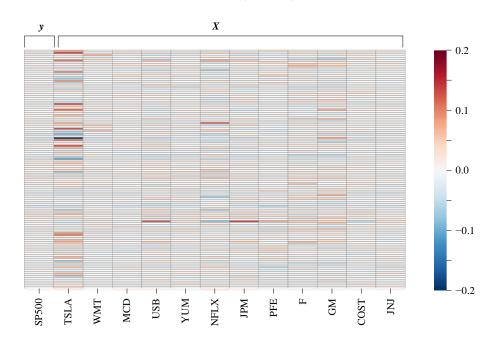


图 14. 股价数据, 起始值归一化

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[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

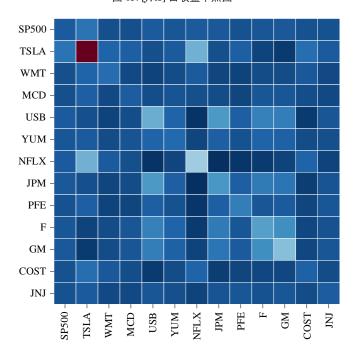


图 16. [y, X] 数据协方差矩阵

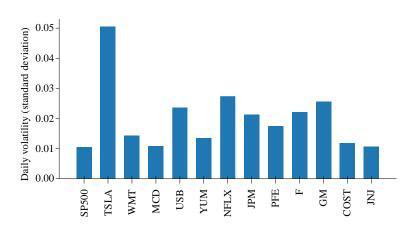


图 17. 日波动率柱状图

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

SP500 -	1.000	0.463	0.359	0.505	0.417	0.571	0.386	0.433	0.341	0.390	0.340	0.517	0.570
TSLA –	0.463	1.000	0.211	0.238	0.044	0.150	0.420	0.052	0.143	-0.009	-0.039	0.353	0.193
WMT –	0.359	0.211	1.000	0.148	-0.021	0.160	0.282	0.020	-0.040	0.111	0.104	0.562	0.149
MCD -	0.505	0.238	0.148	1.000	0.152	0.508	0.188	0.132	-0.003	0.352	0.305	0.358	0.243
USB –	0.417	0.044	-0.021	0.152	1.000	0.456	-0.127	0.908	0.309	0.631	0.497	-0.193	0.327
YUM –	0.571	0.150	0.160	0.508	0.456	1.000	-0.003	0.438	0.276	0.488	0.410	0.180	0.365
NFLX –	0.386	0.420	0.282	0.188	-0.127	-0.003	1.000	-0.183	-0.143	-0.074	-0.011	0.468	-0.013
ЈРМ —	0.433	0.052	0.020	0.132	0.908	0.438	-0.183	1.000	0.338	0.608	0.455	-0.167	0.331
PFE –	0.341	0.143	-0.040	-0.003	0.309	0.276	-0.143	0.338	1.000	0.227	0.238	0.011	0.479
F –	0.390	-0.009	0.111	0.352	0.631	0.488	-0.074	0.608	0.227	1.000	0.721	0.039	0.269
GM –	0.340	-0.039	0.104	0.305	0.497	0.410	-0.011	0.455	0.238	0.721	1.000	0.045	0.308
COST –	0.517	0.353	0.562	0.358	-0.193	0.180	0.468	-0.167	0.011	0.039	0.045	1.000	0.229
JNJ –	0.570	0.193	0.149	0.243	0.327	0.365	-0.013	0.331	0.479	0.269	0.308	0.229	1.000
	SP500 -	TSLA -	WMT -	MCD -	USB -	YUM -	NFLX -	– Mdt	PFE -	<u>г</u> т	- W5	COST –	L

图 18. [y, X] 数据相关性系数矩阵热图

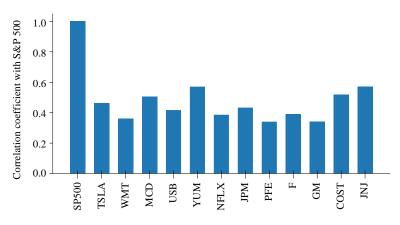


图 19. 股价收益率和 S&P 500 收益率相关性系数柱状图

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: ht —_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

SP500 -	0.0	62.4	69.0	59.7	65.3	55.2	67.3	64.3	70.1	67.1	70.1	58.9	55.2
TSLA -	62.4	0.0	77.8	76.2	87.5	81.4	65.1	87.0	81.8	90.5	92.3	69.3	78.9
WMT -	69.0	77.8	0.0	81.5	91.2	80.8	73.6	88.9	92.3	83.6	84.0	55.8	81.4
MCD -	59.7	76.2	81.5	0.0	81.3	59.4	79.2	82.4	90.2	69.4	72.2	69.0	76.0
USB –	65.3	87.5	91.2	81.3	0.0	62.9	97.3	24.7	72.0	50.9	60.2	101.1	70.9
YUM –	55.2	81.4	80.8	59.4	62.9	0.0	90.2	64.0	74.0	60.8	65.8	79.6	68.6
NFLX -	67.3	65.1	73.6	79.2	97.3	90.2	0.0	100.6	98.2	94.2	90.6	62.1	90.7
ЈРМ –	64.3	87.0	88.9	82.4	24.7	64.0	100.6	0.0	70.2	52.6	62.9	99.6	70.7
PFE -	70.1	81.8	92.3	90.2	72.0	74.0	98.2	70.2	0.0	76.9	76.2	89.4	61.4
F -	67.1	90.5	83.6	69.4	50.9	60.8	94.2	52.6	76.9	0.0	43.8	87.8	74.4
GM -	70.1	92.3	84.0	72.2	60.2	65.8	90.6	62.9	76.2	43.8	0.0	87.4	72.1
COST -	58.9	69.3	55.8	69.0	101.1	79.6	62.1	99.6	89.4	87.8	87.4	0.0	76.8
JNJ –	55.2	78.9	81.4	76.0	70.9	68.6	90.7	70.7	61.4	74.4	72.1	76.8	0.0
	SP500 -	TSLA -	WMT -	MCD -	USB -	YUM -	NFLX -	- Mdf	PFE -	ГТ 	- WD	COST -	- INI

图 20. [y, X] 标准差向量夹角矩阵热图,余弦相似性

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

						0.774			
Dep. Varial	ole:		_	R-squared:					
Model:			_	Prob (F-statistic):					
Method:		Least Squa							
Date:		XXXXXXXXXXX							
Time:		XXXXXXXXXXX	KXX Log-Li 127 AIC:	kelihood:		493.88			
No. Observa						-961.8 -924.8			
Df Model:	ıs:		114 BIC: 12			-924.6			
Covariance	Time:	nonrob							
========		:========	.============						
	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			
TMW	0.0272	0.041	0.667	0.506	-0.054	0.108			
MC D	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 Durbir	n-Watson:		1.862			
Prob (Omnib	us):	0.	023 Jarque	e-Bera (JB):		8.445			
Skew:		0.	400 Prob(J	лв):		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:

$$\underbrace{\left\|\mathbf{y} - \overline{\mathbf{y}}\mathbf{I}\right\|_{2}^{2}}_{\text{SST}} = \underbrace{\left\|\hat{\mathbf{y}} - \overline{\mathbf{y}}\mathbf{I}\right\|_{2}^{2}}_{\text{SSR}} + \underbrace{\left\|\mathbf{y} - \hat{\mathbf{y}}\right\|_{2}^{2}}_{\text{SSE}} \tag{26}$$

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 \mathbb{R}^2 。

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

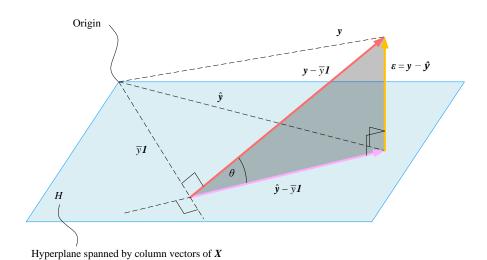


图 22. 第一个直角三角形

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。

$$\|\mathbf{y}\|_{2}^{2} = \|\hat{\mathbf{y}}\|_{2}^{2} + \|\mathbf{y} - \hat{\mathbf{y}}\|_{2}^{2} = \|\hat{\mathbf{y}}\|_{2}^{2} + \|\mathbf{\varepsilon}\|_{2}^{2}$$
 (27)

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

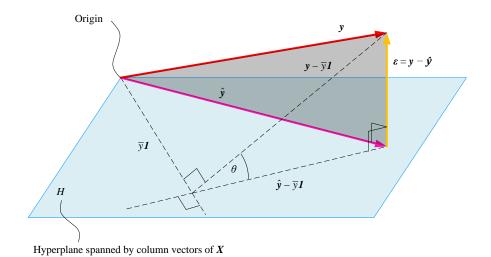


图 23. 第二个直角三角形

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

本书配套徽课视频均发布在 B 站——生姜 DrGinger: https://space.bilibili.com/513194466

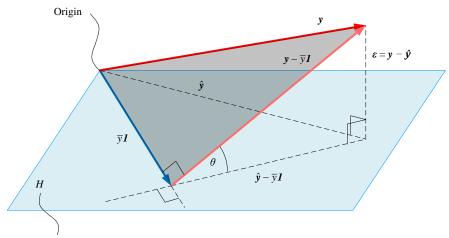
欢迎大家批评指教,本书专属邮箱: jiang.visualize.ml@gmail.com

第三个直角三角形

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

$$\left(\overline{y}I\right)^{\mathrm{T}}\left(y-\overline{y}I\right)=0\tag{28}$$

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



Hyperplane spanned by column vectors of \boldsymbol{X}

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:

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

残差向量可以写成:

$$\varepsilon = y - \hat{y} = y - \overline{y}I - (\hat{y} - \overline{y}I) \tag{31}$$

上式左乘 I^T , 得到:

$$I^{\mathsf{T}}_{0} \varepsilon = \underbrace{I^{\mathsf{T}} \left(y - \overline{y} I \right)}_{0} - I^{\mathsf{T}} \left(\hat{y} - \overline{y} I \right)$$
(32)

即

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

本书配套徽课视频均发布在 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

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

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

$$\overline{y} \mathbf{I}^{\mathrm{T}} \left(\hat{\mathbf{y}} - \overline{y} \mathbf{I} \right) = 0 \tag{34}$$

上式体现的核心思想就是 \hat{y} 的均值为 \bar{y} 。

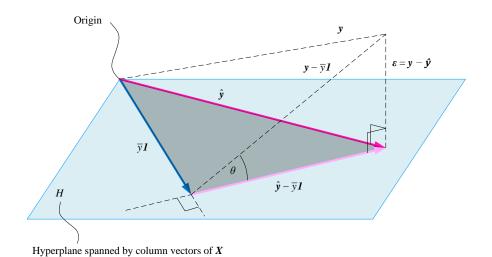


图 25. 第四个直角三角形

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:

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

其中矩阵 J 为全 1 方阵,形状为 $n \times n$:

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

SSR 可以通过矩阵运算求得:

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

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

其中矩阵 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:

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

对于多元 OLS 线性回归模型, MSE 的矩阵运算为:

$$MSE = \frac{\left\| (I - H) y \right\|_{2}^{2}}{n - k}$$

$$= \frac{y^{\mathsf{T}} y - 2 y^{\mathsf{T}} H y + y^{\mathsf{T}} H^{2} y}{n - k}$$

$$= \frac{y^{\mathsf{T}} y - y^{\mathsf{T}} H y}{n - k}$$

$$= \frac{y^{\mathsf{T}} (I - H) y}{n - k}$$

$$(40)$$

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

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 C 可以通过下式计算得到:

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

其中,

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

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 $SE(\hat{b}_{j})$ 为:

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

对于多元线性回归, 假设检验原假设和备择假设分别为:

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

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

 b_i 的 t 检验统计值:

$$T_{j} = \frac{\hat{b}_{j} - b_{j,0}}{\operatorname{SE}(\hat{b}_{j})} \tag{45}$$

类似地,如果下式成立,接受零假设 H_0 :

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

否则,则拒绝零假设 H_0 。

系数 b_i 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:

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

 $x^{(i)}$ 为矩阵 X 的第 i 行:

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

类似地,对于多元 OLS 线性回归模型,yp的预测区间估计为:

$$\hat{y}^{(i)} \pm t_{1-\alpha/2, n-2} \cdot \sqrt{\text{MSE}} \cdot \sqrt{1 + x^{(i)} (X^{\mathrm{T}} X)^{-1} (x^{(i)})^{\mathrm{T}}}$$
(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 $X^{\mathsf{T}}X$ 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 X^TX 可逆, X^TX 的行列式值不为 0:

$$\det\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)\neq0\tag{51}$$

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 $\hat{X}_{t,r}$ 对于其中的任意解释变量 $\left\{X_{t,r}\right\}$,其对应的方差 膨胀因子 VIF; 可由下式计算:

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

$$VIF_i = \frac{1}{1 - R_i^2} \tag{52}$$

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

$$X_{i,t} = \alpha_0 + \sum_{j=1, j \neq i}^{n} \alpha_j X_{j,t} + \varepsilon_t$$
 (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,VIF ≤ 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 χ 和 γ 服从多维高斯分布:

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

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

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

如图 26 所示,给定 $\chi = x$ 的条件下 γ 的条件期望为:

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

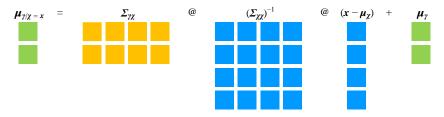


图 26. 给定 $\chi = x$ 的条件下 γ 的期望值的矩阵运算,图片来自《统计至简》

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

对于本例, 我们对 (56) 进行转置得到:

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

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

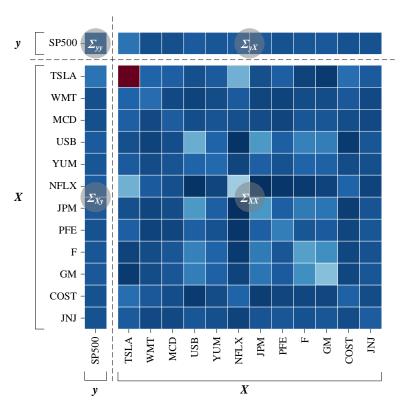
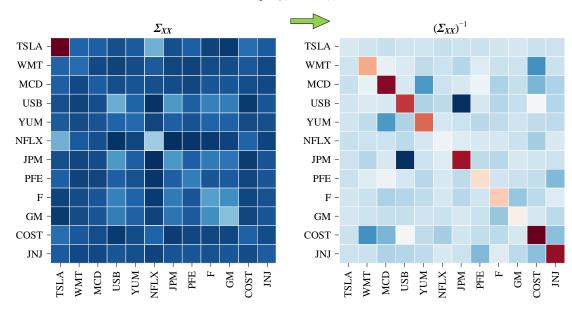


图 27. [y, X] 协方差矩阵



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 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:

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

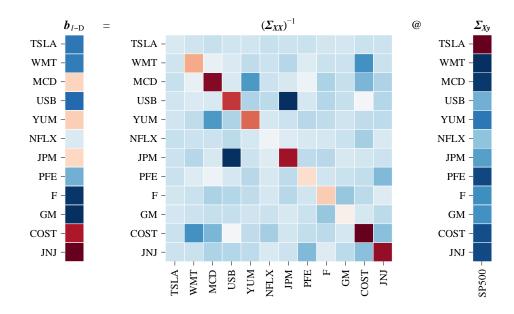


图 29. 求线性回归参数, 除截距以外

如图 30 所示, b₀ 为:

$$b_0 = \mathbf{E}(\mathbf{y}) - \mathbf{E}(\mathbf{X})\mathbf{b}_{1\sim D} \tag{59}$$

其中, E(X) 为行向量。

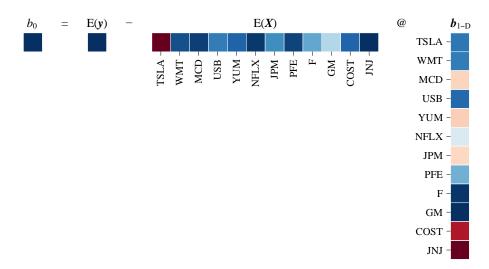


图 30. 求截距系数

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



Bk6_Ch11_03.py 完成本节运算。